Estructura de Proteínas

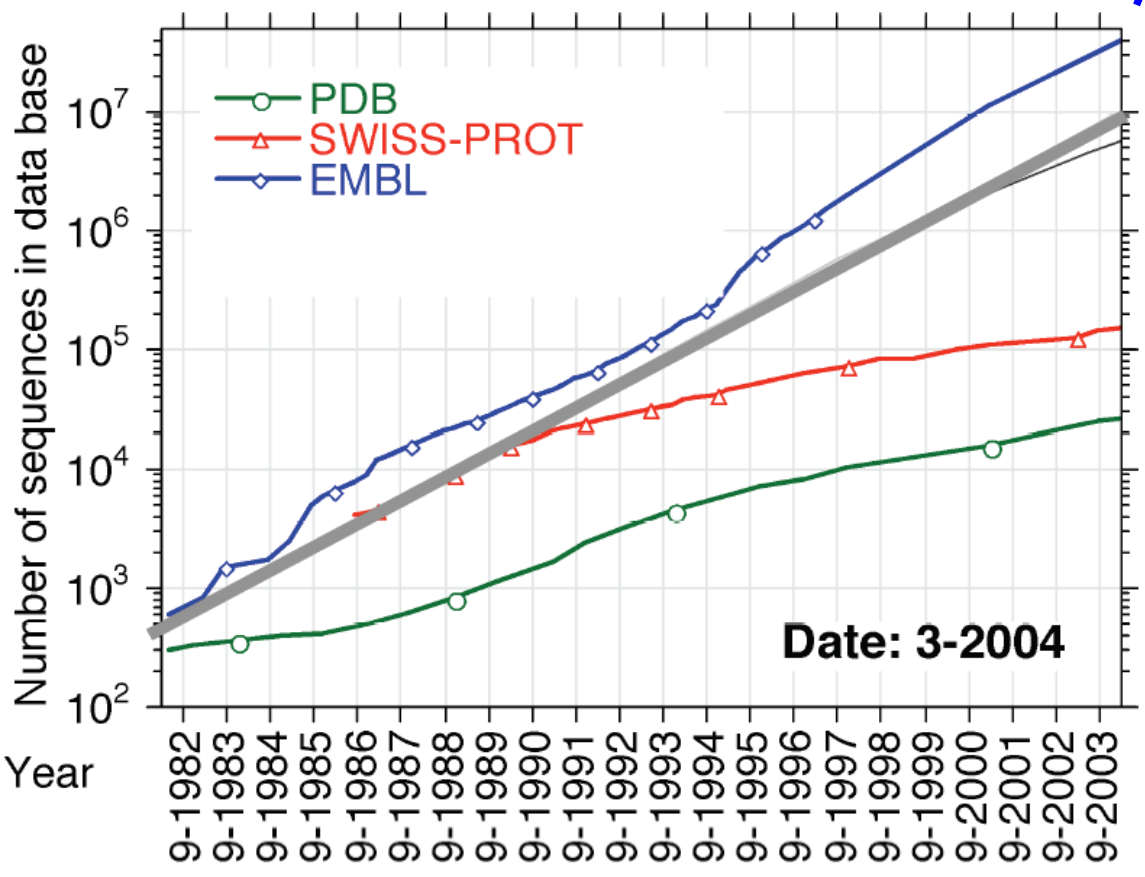
Predicción de Estructura Secundaria y otras Características 1D

Florencio Pazos (CNB-CSIC)

Florencio Pazos Cabaleiro
Computational Systems Biology Group (CNB-CSIC)
pazos@cnb.csic.es

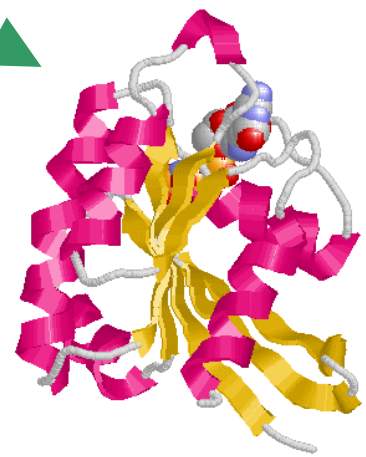
CNB – Marzo 2008

Conocimiento experimental de secuencias, funciones y estructuras de proteínas

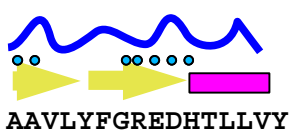
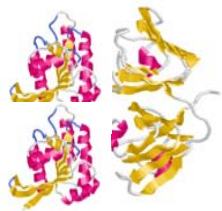


MREYKLVVLGSGGVGKSALTVQFVQGIFVDEYDPTIEDSY
RKQVEVDCQQCMLEILDTAGTEQFTAMRDLYMKNGQGFA
VYSITAQSTFNDLQDLREQILRVKDTEDVPMILVGNKCDL
EDERVVGKEQGQNLARQWCNCAFLESSAKSKINVNEIFYD
LVRQINR

MLEILDTAGTEQFTAMRDLYMKNGQGFAL
VYSITAQSTFNDLQDLREQILRVKDTEDVPMIL
VGNKCDLEDERV



Predicción de estructura de proteínas
Clasificación de los métodos de predicción

Nivel estructura proteínas / Representación de la proteína / Uso de información extra	Secundaria	--------	terciaria	cuaternaria
	1D 	2D 	3D 	4D 
Ab Initio	pred. str. secundaria	mutaciones correlacionadas	- dinámica molecular - minimización de energía	*docking*
No Ab-Initio	pred. str. secundaria		- modelado por homología - *threading*	*docking con filtros*

Predicción de estructura de proteínas

Características 1D

Características 1D de una secuencia: Características que pueden ser representadas por un único valor asociado a cada aminoácido (B. Rost).

Estos valores suelen tomar la forma de etiquetas de estado, como por ejemplo en el caso de la estructura secundaria (H->hélice, E->lámina, T->giro). También pueden tomar valores continuos (% superficie accesible, ...)

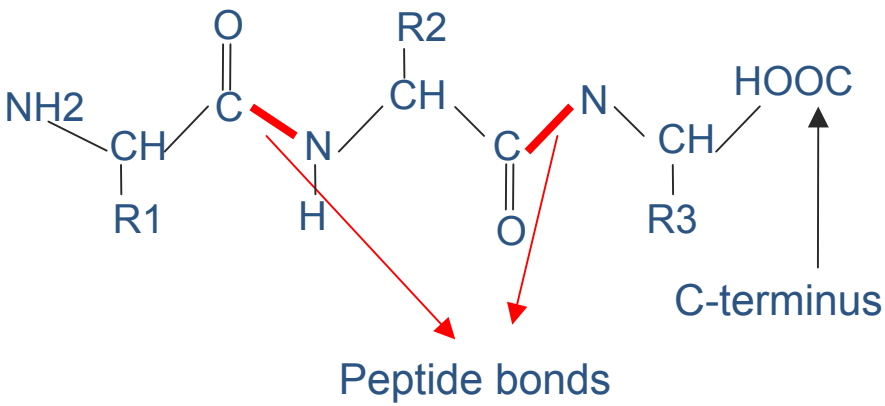
Algunas características 1D:

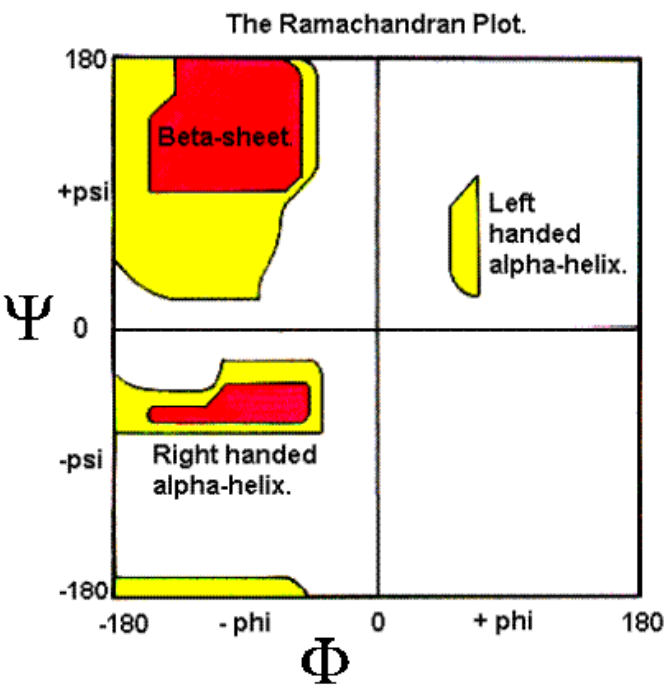
- Estructura secundaria
- Accesibilidad al solvente
- Modificaciones post-transcripcionales
- Péptidos señal
- *Coiled-coils*
- Regiones desordenadas
- etc.

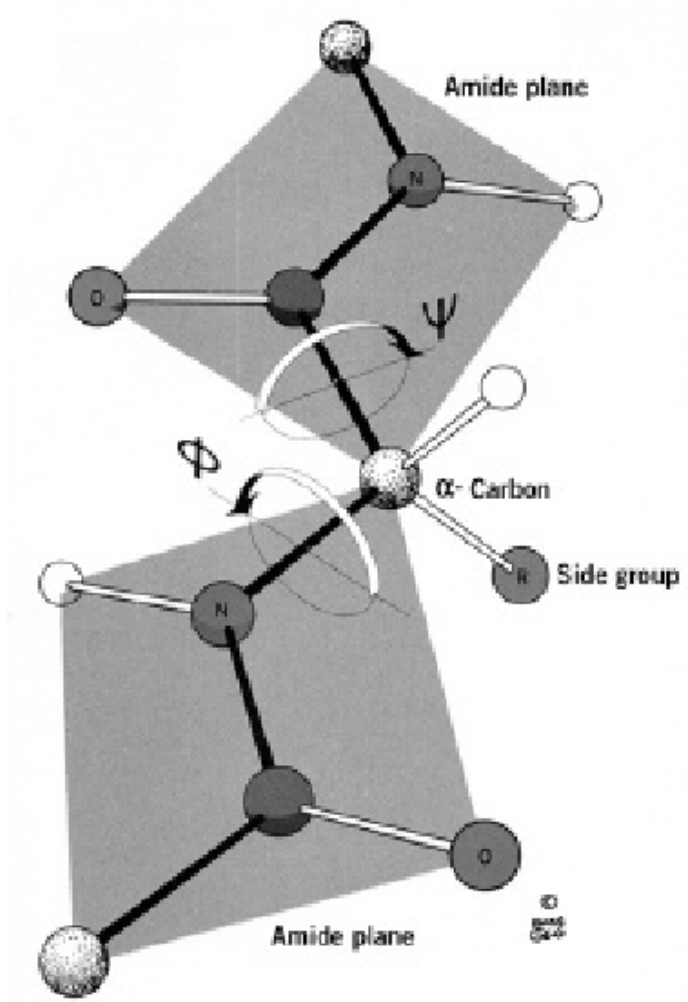
¿Por qué predecir estructura secundaria y otras características 1D?

- No siempre es posible generar un modelo 3D (fiable).
- Ayuda en la predicción de plegamiento 3D (restringe plegamientos posibles)
- Predicción de función: Motivos de estructura secundaria peculiares
- El mapeo de toda las predicciones 1D a lo largo de una secuencia da mucha información sobre los posibles dominios estructurales y funcionales, sitios activos, zonas diferenciadas, ...

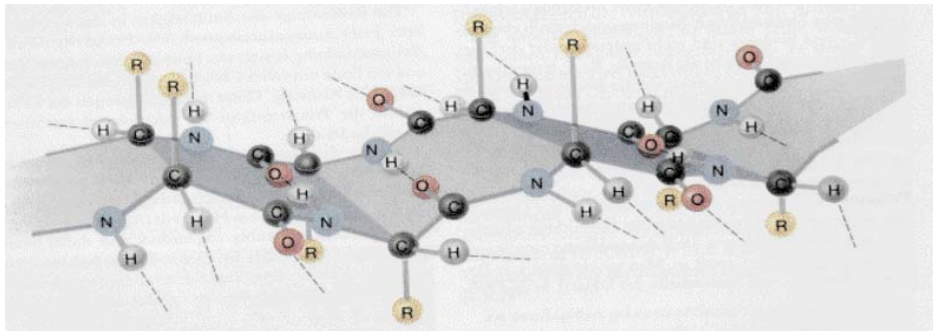
Características 1D
Estructura Secundaria

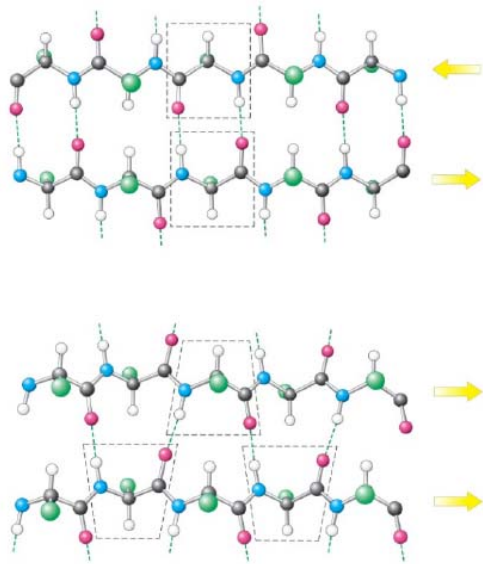




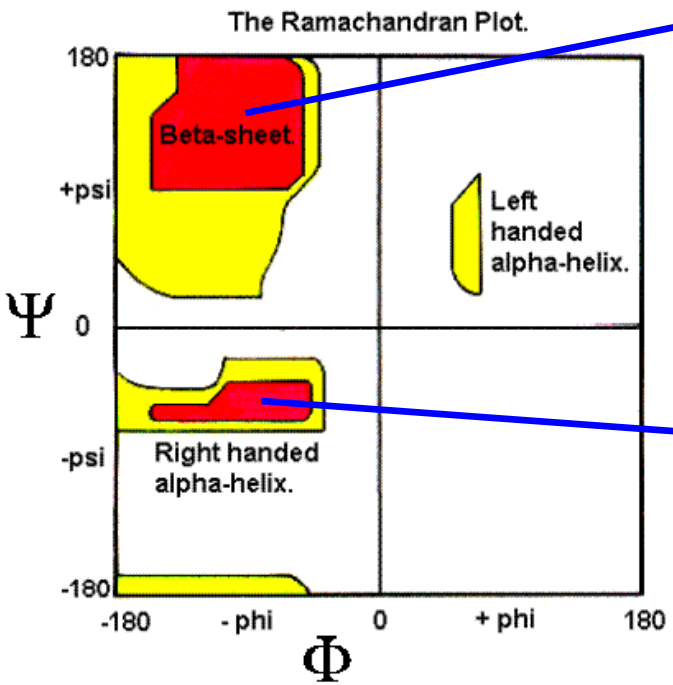


Características 1D
Estructura Secundaria

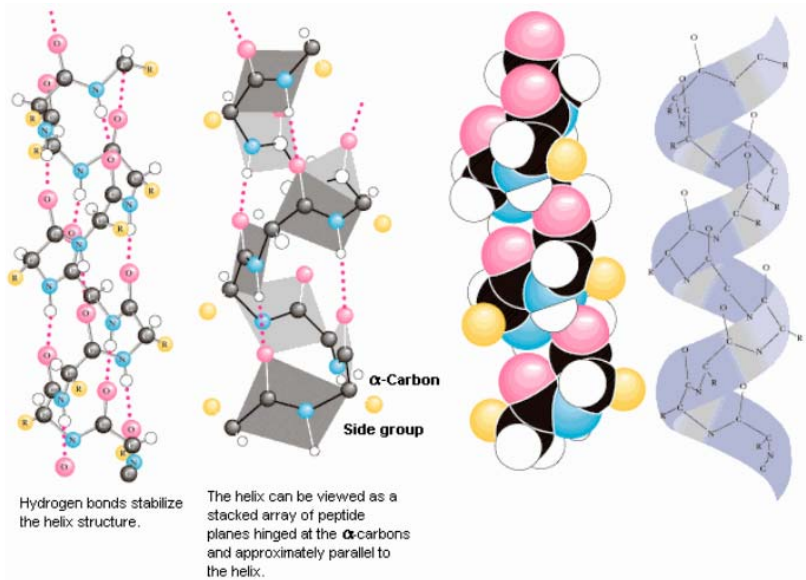




β-strand



α-helix



Características 1D
Estructura Secundaria

```
  1 ASKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTT
      TTGGGGSSEEEEEEEEEEEETTEEEEEEEEEEEETTTTEEEEEEEETT

 51 GKLPVPWPTLVTTFSYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTIFF
    SS SS GGGGHHHHSSS GGG B  GGGGGG HHHHTTTT EEEEEEEEE

101 KDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNV
    TTS EEEEEEEEEEETTEEEEEEEEEE   TTSTTTTT B S    EEE

151 YIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHY
    EEEEEGGGIEEEEEEEEEEETTS EEEEEEEEEEESSSS      SEE

201 LSTQSALSKDPNEKRDHMVLLEFVTAAGIT HGMDELYK
    EEEEEEEE  TT  SSEEEEEEEEEES
```

Definición: T=hydrogen bond turn, H=helix, G=310 helix, I=phi helix, B=residue in isolated beta bridge, E=strand, and S=bend

Predicción: H/E/T (3 states only)

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

Estructura Secundaria
Primera Generación de Métodos

Métodos estadísticos basados simplemente en la tendencia de cada aminoácido a formar cada uno de los elementos de estructura secundaria

- Chou y Fasman en 1974, propusieron el primero de estos métodos. Emplearon estadísticas extraídas de las **15 estructuras** resueltas por cristalografía de rayos-X en aquella época. Estas probabilidades fueron calculadas para cada resíduo por separado. Más adelante este método mostró una exactitud del 57% sobre 62 proteínas.

- Garnier (1978). Estimó las probabilidades para interacciones de pares de resíduos significativas, obteniendo una mayor fiabilidad (~60%)

Chou, P.Y. and Fasman, G.D. (1974) Prediction of protein conformation. *Biochemistry*, **13**, 222-244/225.

Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, **120**, 97-120.

Estructura Secundaria
Primera Generación de Métodos

Name	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	142	83	66	0.06	0.076	0.035	0.058
Arginine	98	93	95	0.070	0.106	0.099	0.085
Aspartic Acid	101	54	146	0.147	0.110	0.179	0.081
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Cysteine	70	119	119	0.149	0.050	0.117	0.128
Glutamic Acid	151	037	74	0.056	0.060	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.190	0.152
Histidine	100	87	95	0.140	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.070
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.120	0.139	0.125	0.106
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Valine	106	170	50	0.062	0.048	0.028	0.053

Glu, Met Ala y Leu : tendencia a formar helices hélices.
Val, Ile y Tyr: tendencia a formar láminas beta.
Gly, Pro, ...: tendencia a formar giros.

Chou, P.Y. and Fasman, G.D. (1974) Prediction of protein conformation. *Biochemistry*, **13**, 222-244/225.

Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, **120**, 97-120.

Estructura Secundaria
Segunda Generación de Métodos

- La principal característica de estos métodos es la utilización de ventanas de resíduos adyacentes en secuencia, incluyendo así información de contexto a la predicción.

- Un gran número de algoritmos de predicción se usaron en esta generación de métodos:
  - Redes Neuronales Artificiales, Teoría de Grafos, Métodos basados en reglas, Estadística multivariable, ...
  - Esta innovación acercó la predicción de estructura secundaria a la barrera del 70% de fiabilidad.

- Limitaciones
  - Fiabilidad (prediccciones 3-estados < 70%)
  - Se obtienen bajas fiabilidades para cadenas-β
  - La hélices y láminas predichas tienden a ser demasiado cortas.
  - Debido a:
  - El número de estructuras disponibles sigue siendo demasiado pequeño para extrapolar al espacio de secuencias. Difiriendo a veces entre distintos cristales para la misma secuencia.
  - NO se tienen en cuenta los efectos provocados por resíduos situados a grandes distancias en secuencia (pero no en el espacio)

Garnier, J. and Robson, B. (1989) The GOR method for predicting secondary structure in proteins. In D., F.G. (ed.), *Prediction of protein structure and the principles of protein conformation*. Plenum Press, New York, pp. 417-465

Estructura Secundaria
Tercera Generación de Métodos

Iniciada por Levin (~69%) y Rost y Sander en 1994 (PHD 72%)

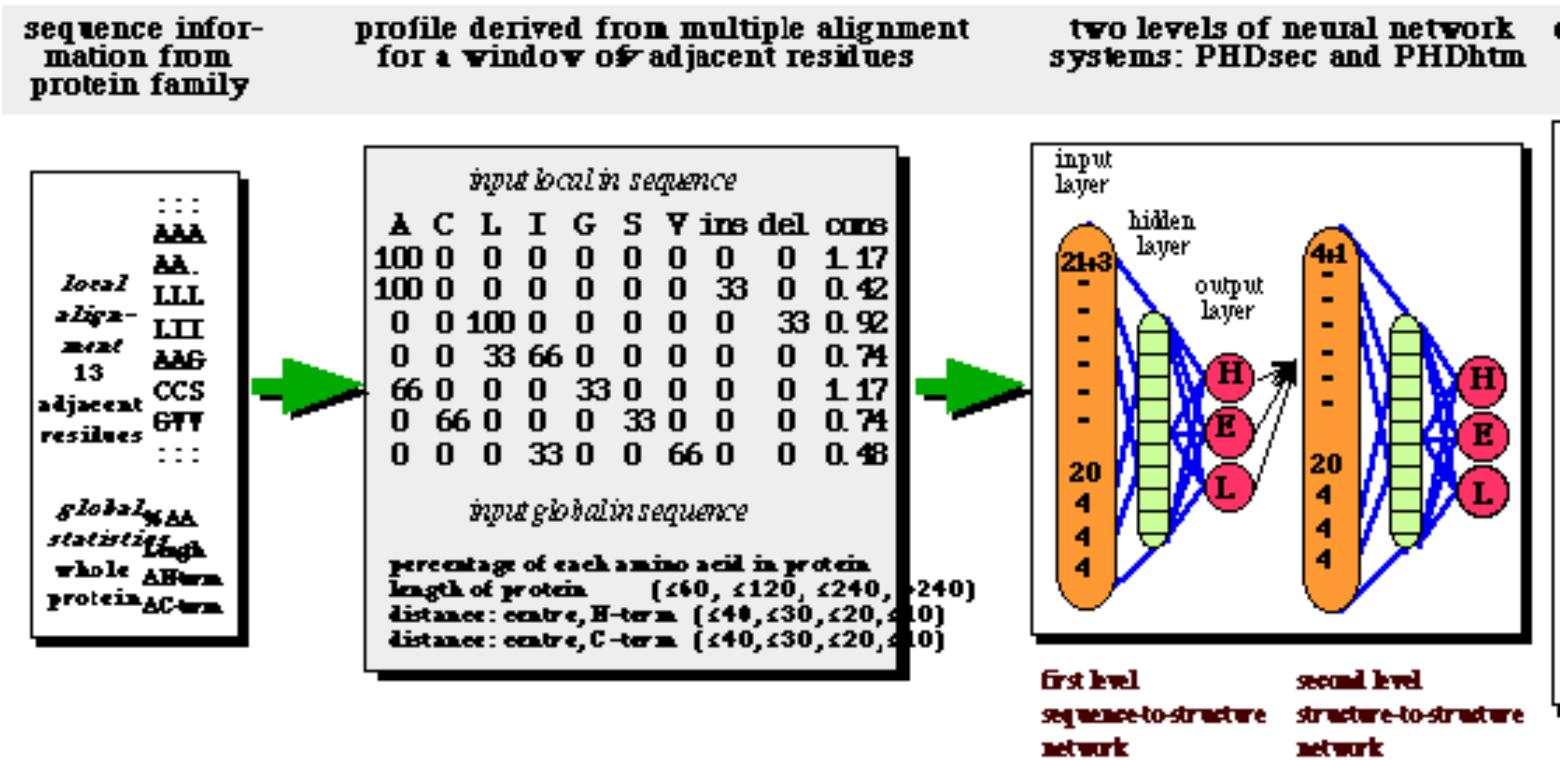
- La principal innovación de esta tercera generación es la inclusión de información evolutiva adicional en forma de alineamientos múltiples (perfiles) (Levin, 1993).
- Además, se resuelve el sesgo en las predicciones de cadenas-β balanceando el conjunto de entrenamiento (dado que las estructuras contienen más hélices que láminas; Rost y Sander, 1994)
- "Suavizado" de predicciones mediante una segunda red.
- Rompen el limite del 70%

Levin JM, Pascarella S, Argos P, Garnier J. (1993). Quantification of secondary structure prediction improvement using multiple alignments. *Protein Eng.* **6(8):**849-54.

Rost, B. and Sander, C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A*, **90**, 7558-7562.

Rost, B., Sander, C. and Schneider, R. (1994) PHD - A mail server for protein secondary structure prediction. *Comp. Applic. Biosci.*, **10**, 53-60.

Estructura Secundaria. Tercera Generación de Métodos



Rost, B. and Sander, C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A*, **90**, 7558-7562.

Rost, B., Sander, C. and Schneider, R. (1994) PHD - A mail server for protein secondary structure prediction. *Comp. Applic. Biosci.*, **10**, 53-60.

Estructura Secundaria. Tercera Generación de Métodos

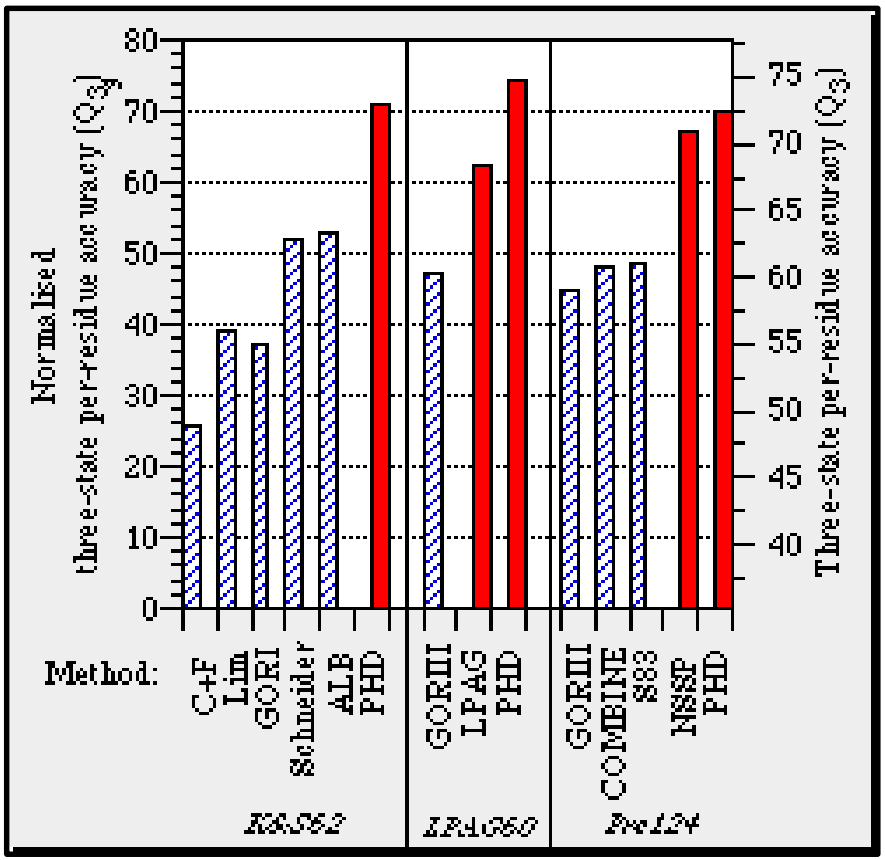
- Varios métodos han seguido estrategias similares a PHD, mejorando sus resultados a través del prefiltrado de los alineamientos de entrada y la extensión de los perfiles mediante PSIBLAST introducido por David Jones en PSIPRED (1999) con fiabilidades próximas al 77%, o mediante HMMs usados por Kevin Karplus *et al.* en SAMT99sec (1999).

- Otros métodos siguen una estrategia diferente, buscando el consenso de diferentes métodos, como es el caso de Jpred2 (Cuff y Barton, 2000).

Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, **292**, 195-202.

Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. (1998). JPred: a consensus secondary structure prediction server. *Bioinformatics*. **14(10):**892-3.

Estructura Secundaria



Métodos de Primera generación: Chou & Fasman, Lim, GORI

Métodos de Segunda generación : Schneider, ALB, GORIII

Métodos de Tercera generación: LPAG, COMBINE, S83, NSSP, PHD

76-78%

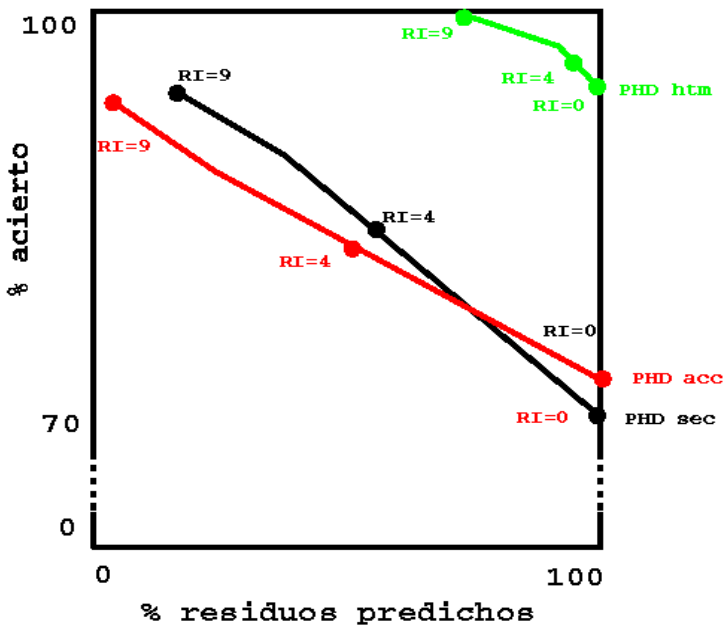
Limite de fiabilidad?

- Límite en la propia definición de estructura Secundaria (DSSP vs. Otros)
- Límite en información local

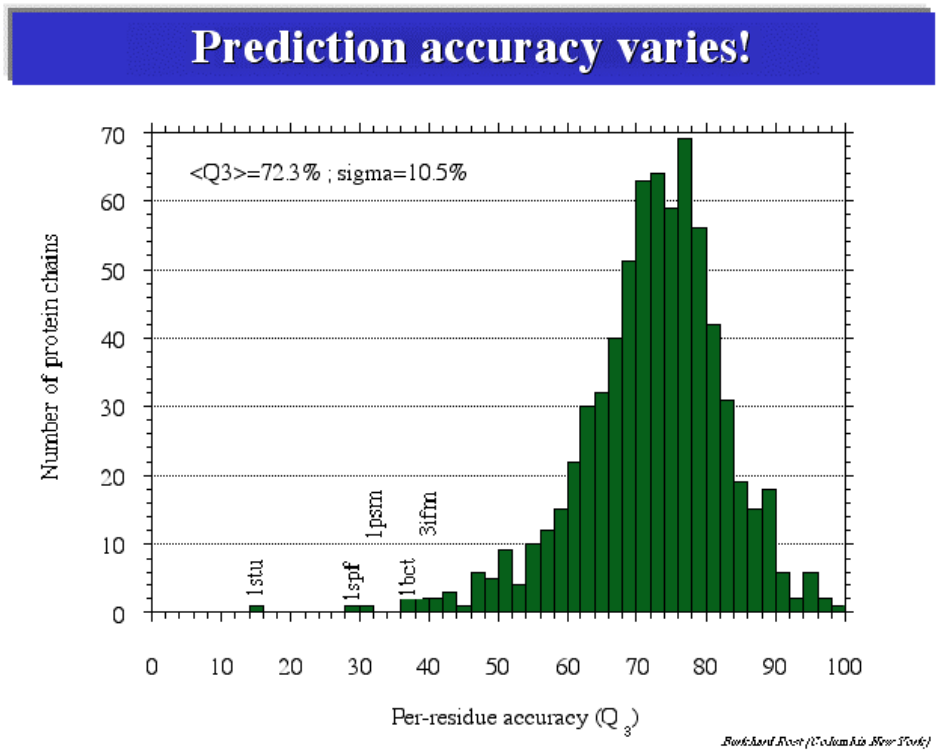
Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

Predicción Estructura Secundaria
(factores a tener en cuenta)

Equilibrio exactitud/%predicho

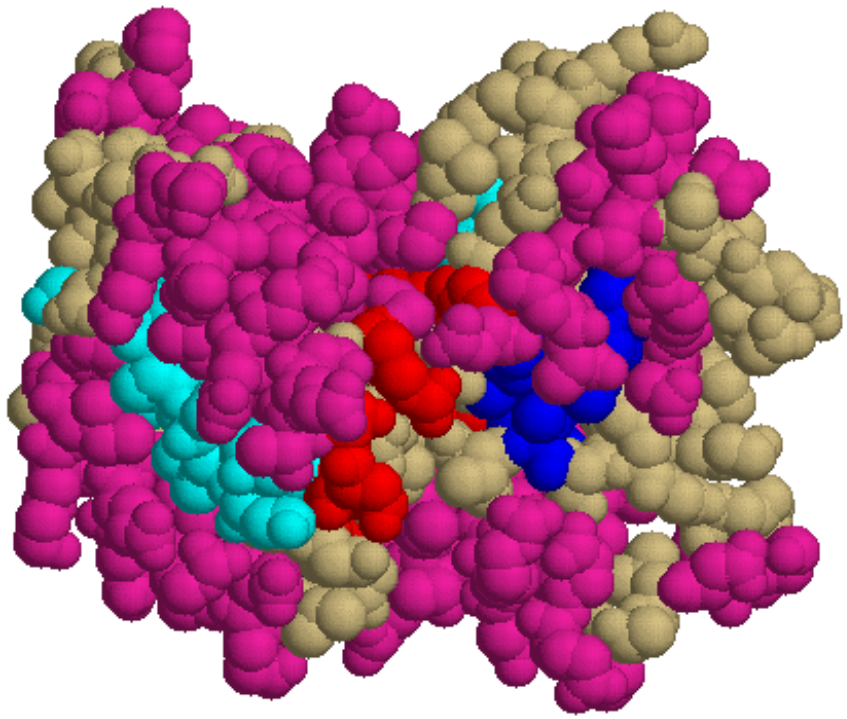


Variabilidad en exactitud media por proteína

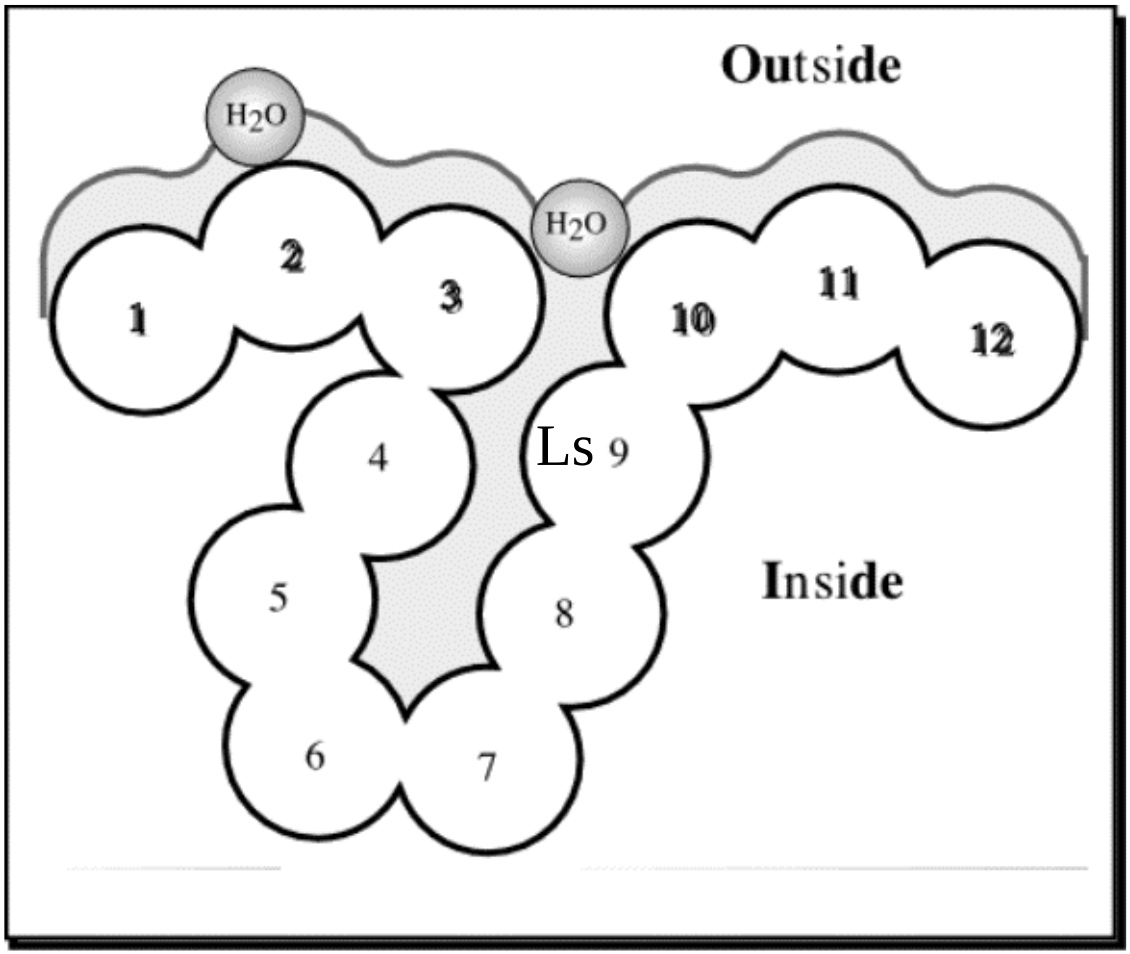


Métodos 1D
Accesibilidad al solvente

- Discriminación de modelos
- Sitios funcionales y de interacción
- Diseño de mutantes, proteinas marcadas, etc.



Accesibilidad al solvente



Los programas de definición de accesibilidad (a partir de la estructura 3D) reportan, para cada residuo, la superficie accesible en Å^2.

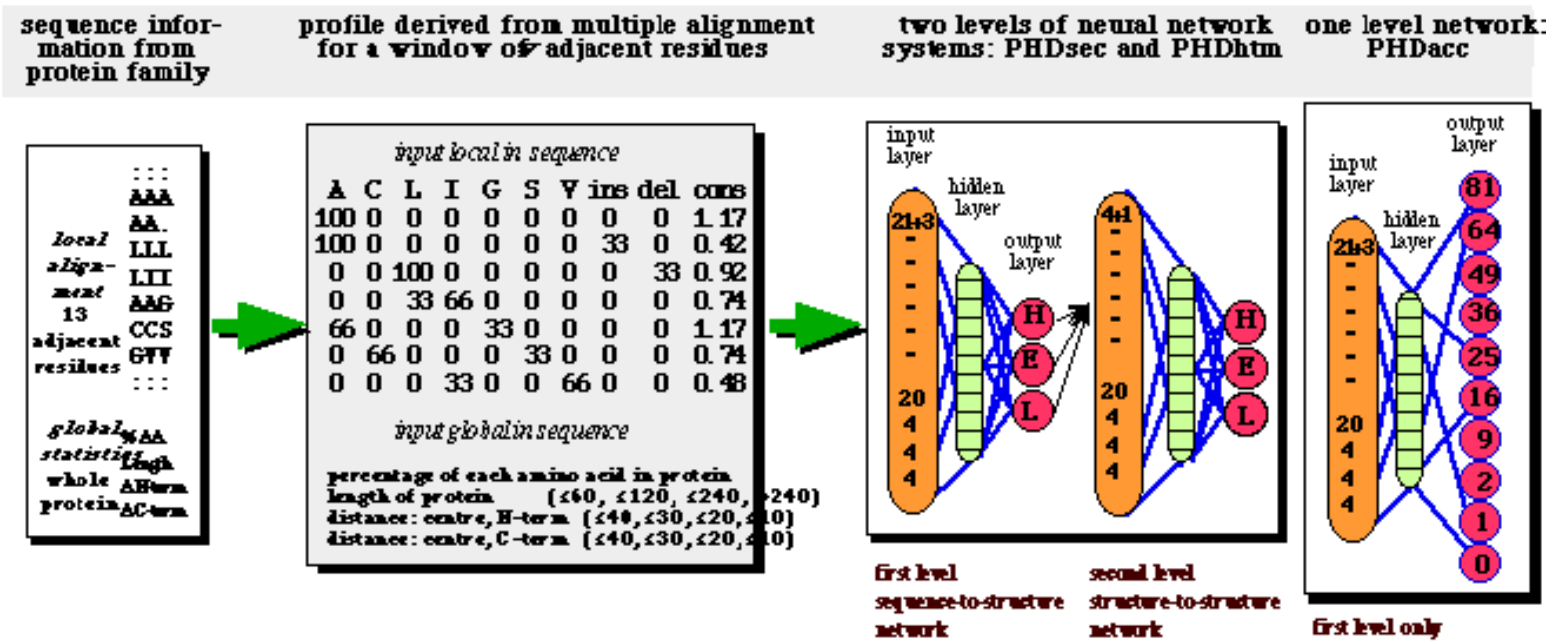
La mayoría de los métodos de predicción reducen el problema a la predicción de dos estados oculto (accs. relativas. <16%, abs <50 Å^2) o expuesto (accs. relativas >= 16%, abs >=50 Å^2).

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

Accesibilidad al solvente

- Misma historia que estructura secundaria: frecuencias -> ventanas -> redes neuronales + información evolutiva (alns.).

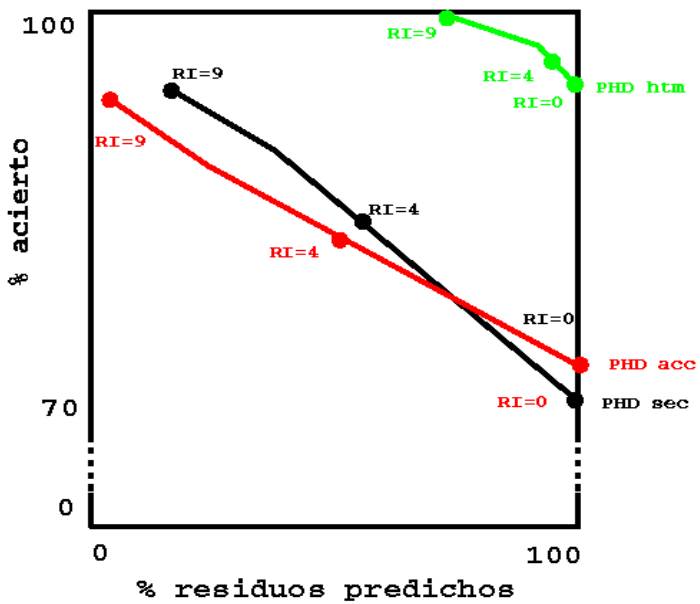
- Los programas suelen ser los mismos que para estructura secundaria, con pequeñas adaptaciones de la NN al caso concreto de la predicción de accesibilidad.

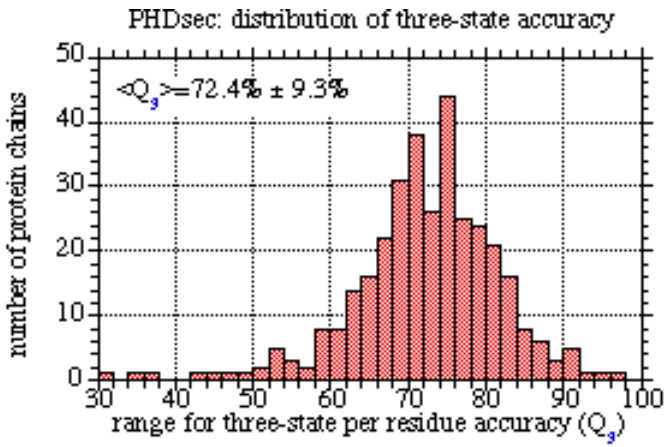


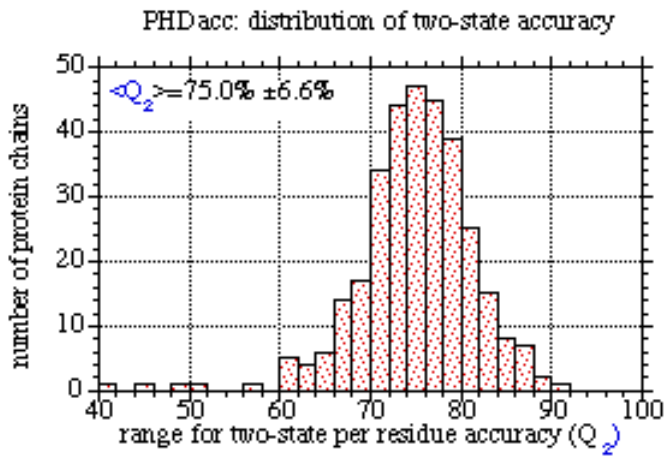
Rost, B. and Sander, C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A*, **90**, 7558-7562.

Rost, B., Sander, C. and Schneider, R. (1994) PHD - A mail server for protein secondary structure prediction. *Comp. Applic. Biosci.*, **10**, 53-60.

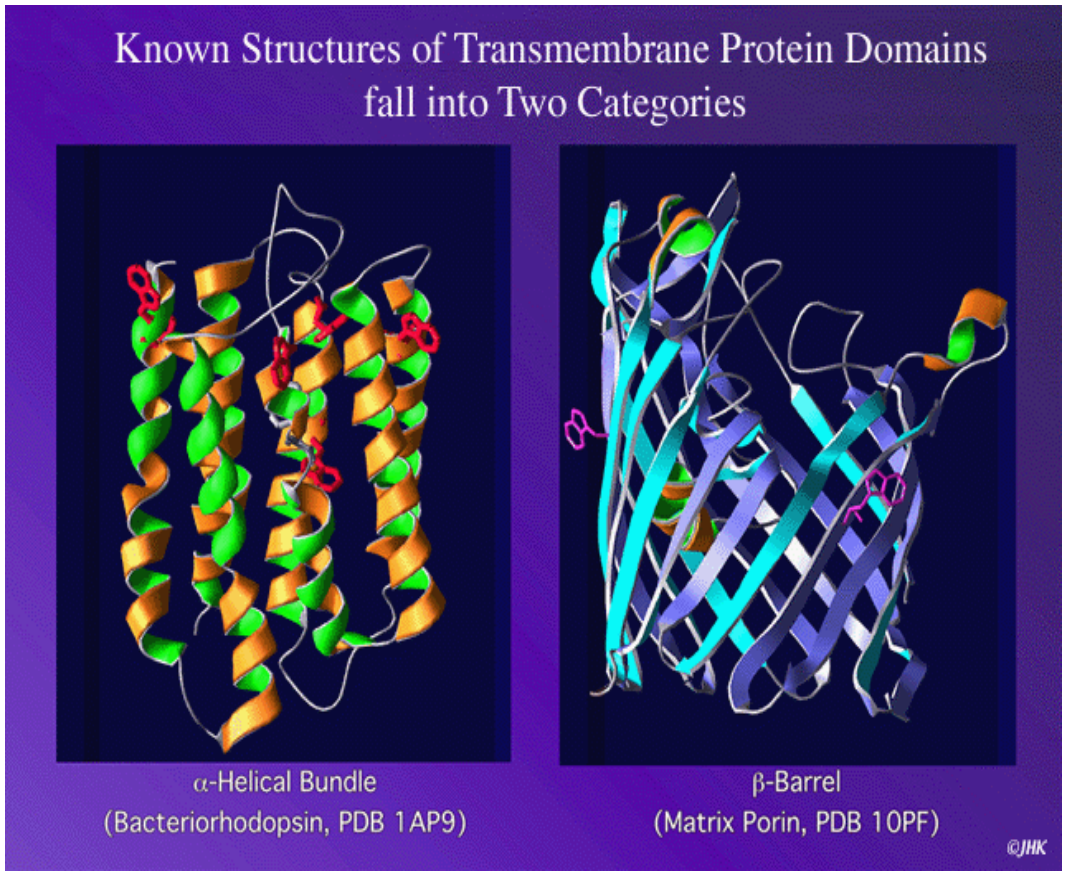
Accesibilidad al solvente







Métodos 1D
Segmentos Transmembrana



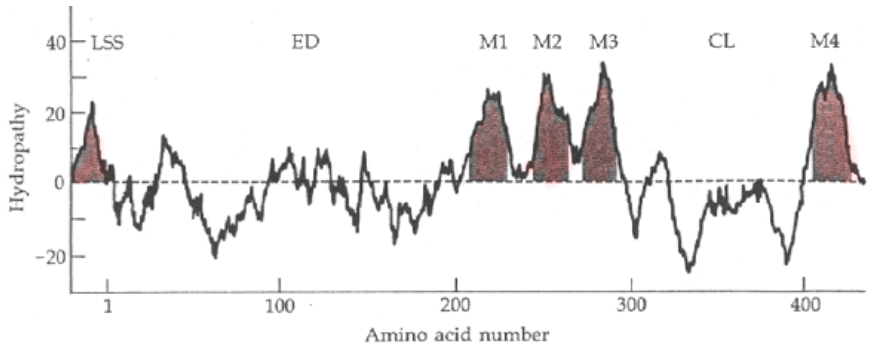
-Difíciles de cristalizar.
Pocas estructuras.

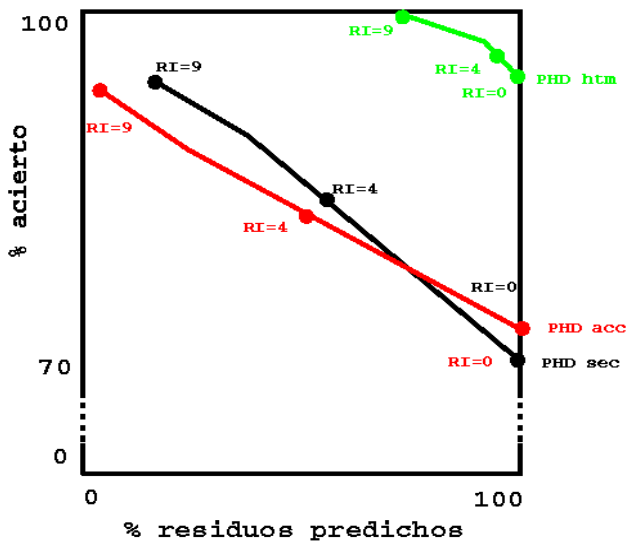
- Primera información sobre dominios, zonas funcionales, etc.

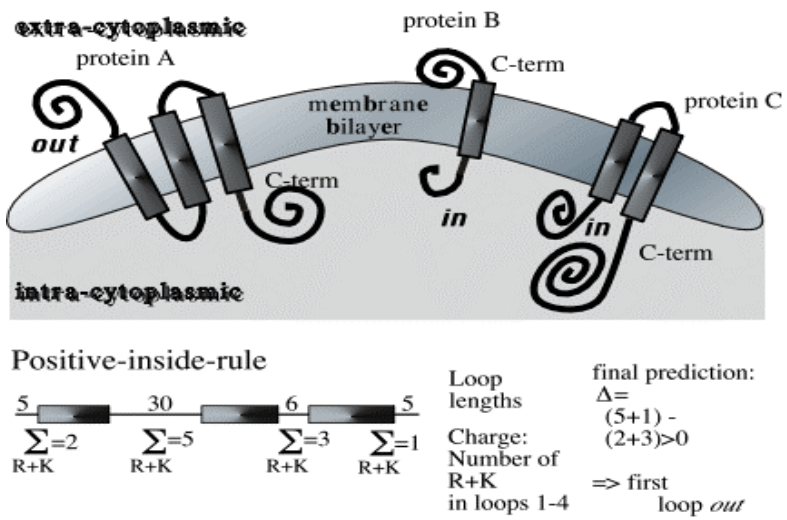
Segmentos Transmembrana
Hélices transmembrana

(1) Las hélices transmembrana tienden a tener una logitud de 20-30 resíduos con una hidrofobidad total alta.

(2) Las regiones de conexión entre hélices del interior del citoplasma tienen una carga positiva mayor que las del exterior



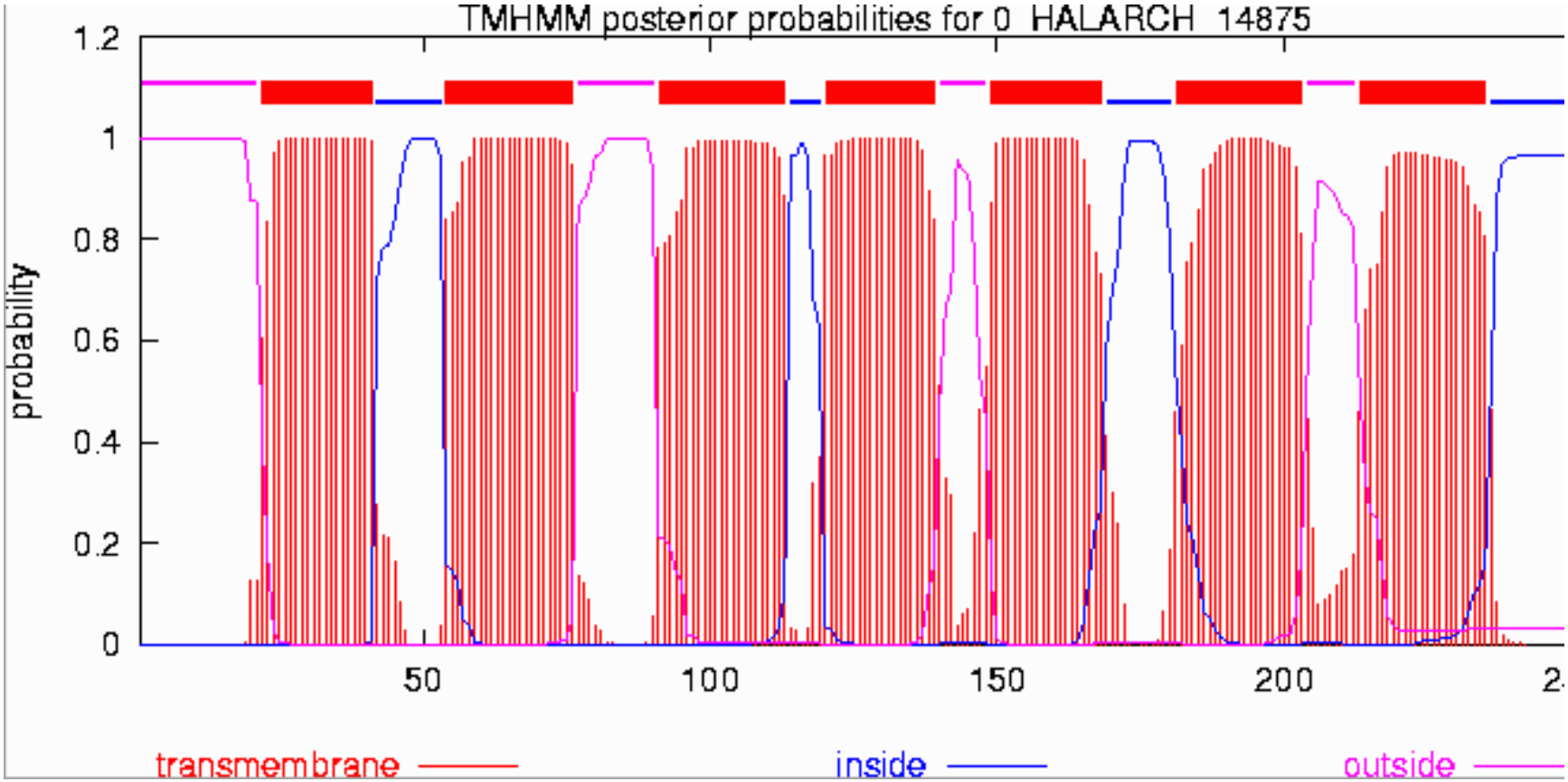




Segmentos Transmembrana
Hélices transmembrana

MEMSAT - *http://bioinf.cs.ucl.ac.uk/psipred/*
Algoritmo de programación dinámica que hace predicciones basadas en tablas estadísticas compiladas de los datos de proteínas de membrana.
TMAP - http://www.mbb.ki.se/tmap/index.html
Usa estadíticas extraídas de perfiles de secuencia.
TopPred2 - http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html
Promedia los valores de hidropatía con una ventana trapezoidal
HMMTOP - http://www.enzim.hu/hmmtop/
Se definen 5 estados estructurales y mediante HMMs para generar fragmentos de secuencia que maximizen la frecuencia de cada estado.
PHDhtm - *http://www.embl-heidelberg.de/predictprotein/*
Combina redes neuronales, alineamientos múltiples y programación dinámica (proporciona un índice de fiabilidad).
DAS - *http://www.enzim.hu/DAS/DAS.html*
Utiliza alineamientos múltiples de un conjunto no redundante de proteínas de membrana.
TMHMM - *http://www.cbs.dtu.dk/services/TMHMM/*
Métodos estadísticos y HMMs que ayudan a mejorar la localización y orientación de hélices trans-membrana.

Segmentos Transmembrana
Hélices transmembrana



Predicciones 1D
Otros

ExPASy Proteomics tools: *http://www.expasy.ch/tools/*
CBS prediction servers: *http://www.cbs.dtu.dk/services/*

COIL – Coiled-coil regions.
PSORT - prediction of signal proteins and localisation sites
SignalP - prediction of signal peptides

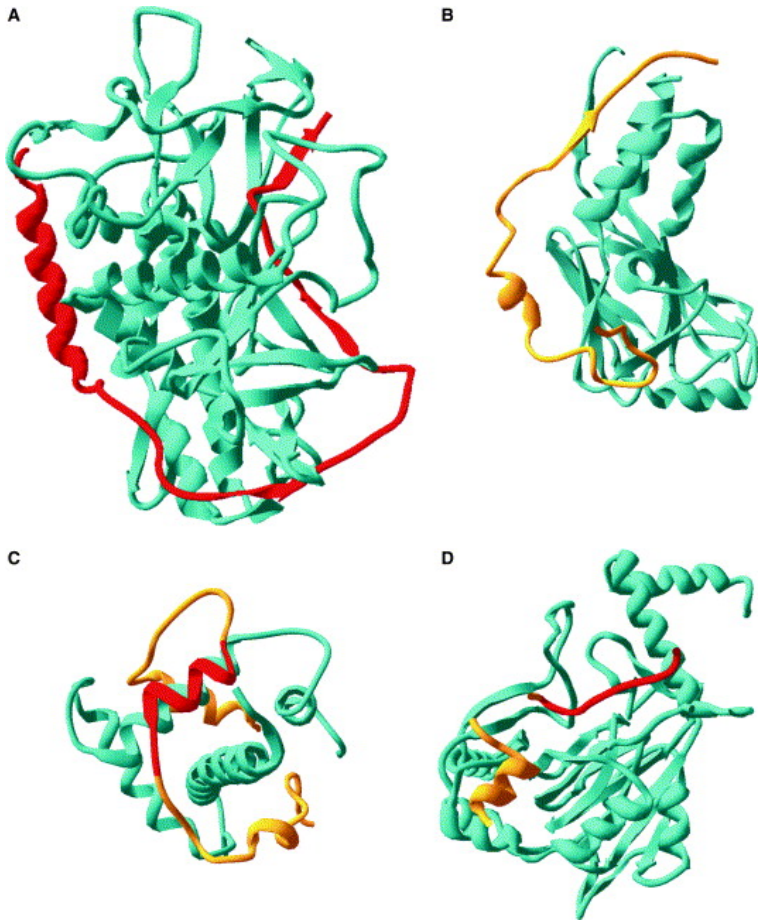
ChloroP - prediction of chloroplast peptides
NetOGlyc - prediction of O-glycosilation sites in mammalian proteins
Big-PI - prediction of glycosil -phosphatidyl inositol modification sites
DGPI - prediction of anchor and breakage sites for GPI

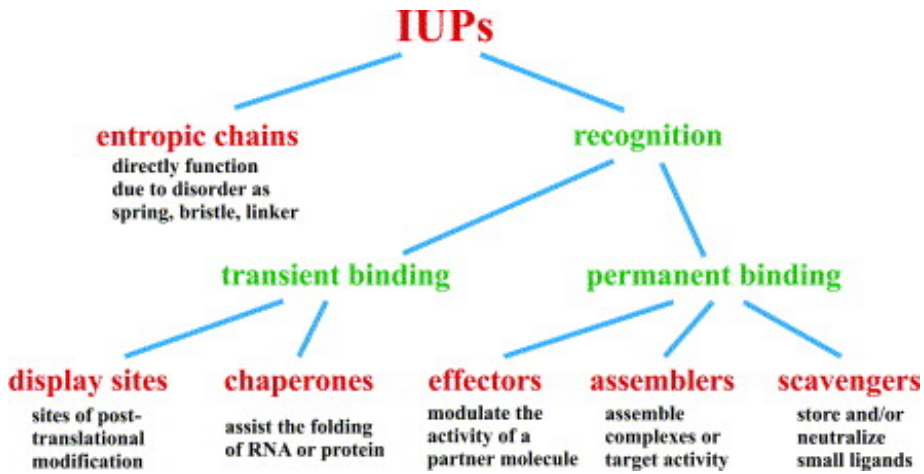
NetPhos - prediction of phosphorylation sites (Ser, Thr, Tyr) in eukaryotes
NetPicoRNA - prediction of cleavage sites for proteases in the picornavirus
NMT - prediction of N-miristoilation of N-terminals
Sulfinator - predicts sulphattation sites in tyrosines
.....
...

$[\mathbf{a}bc\mathbf{d}efg]_n$

Lupas, A., Dyke, M.v. and Stock, J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162-1164.

Proteínas y regiones desestructuradas





Tompa, P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett*, **579**, 3346-3354.

Vucetic, S., Brown, C. J., Dunker, A. K. & Obradovic, Z. Flavors of protein disorder. *Proteins* **52**, 573-84. (2003).

Regiones desestructuradas
Métodos de predicción

Compositionally biased regions. Wootton et al (*SEG*).

Specific for disorder. 003 Jones UCL (David Jones, University College London) support vector machines (*DISOPRED*)

		CASP6			
Group	N	Spec.	Sens.	Prod.	Score
193	66	0.715	0.828	0.593	6.57
96	65	0.507	0.955	0.485	5.07
3	66	0.496	0.949	0.471	4.84
347	66	0.509	0.915	0.466	4.66
676	58	0.450	0.952	0.428	4.31
18	23	0.358	0.990	0.354	4.20
60	66	0.398	0.965	0.384	3.65
675	59	0.584	0.715	0.418	3.43
461	65	0.422	0.885	0.373	3.11
536	66	0.344	0.983	0.338	3.09
633	64	0.549	0.713	0.391	3.00
686	57	0.323	0.964	0.312	2.81
472	61	0.390	0.891	0.348	2.62
667	59	0.326	0.903	0.295	2.20
673	59	0.459	0.743	0.341	2.15
19	44	0.244	0.987	0.240	1.81
674	59	0.178	0.980	0.175	1.15
679	55	0.163	0.995	0.162	1.00
545	64	0.406	0.691	0.280	0.80
245	60	0.060	0.942	0.057	-0.55

Wootton, J.C. and Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases. *Meth in Enzym*, **266**, 554-571

Ward, J. J., McGuffin, L. J., Bryson K., Buxton, B. F. & Jones, D. T. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, **20:**2138-2139.

http://pdg.cnb.uam.es/pazos/cursos/protstr_cnb