FACULTAD DE INGENIERÍA Y CIENCIAS EXACTAS
DEPARTAMENTO DE BIOTECNOLOGÍA Y TECNOLOGÍA ALIMENTARIA
UNIVERSIDAD ARGENTINA DE LA EMPRESA

Bioinformática

ANÁLISIS COMPUTACIONAL DE SECUENCIAS

Dr. Lucas L. Maldonado (PhD)

Lic. Biotechnologist and Molecular Biologist

Bioinformatics and genomics specialist

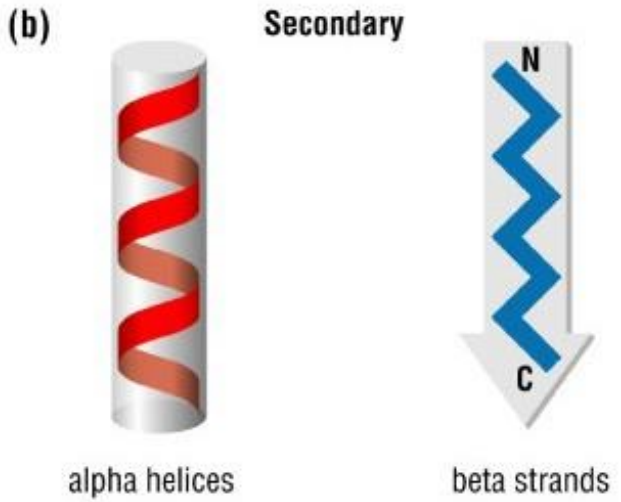
CONICET
Fac. de Medicina - UBA
Fac. de Ciencias Exactas y Naturales – UBA

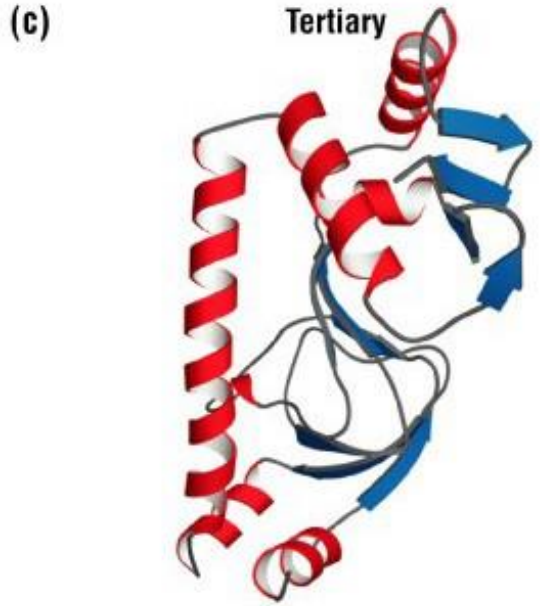
lucamaldonado@uade.edu.ar
lmaldonado@fmed.uba.ar
lucas.l.maldonado@gmail.com.ar

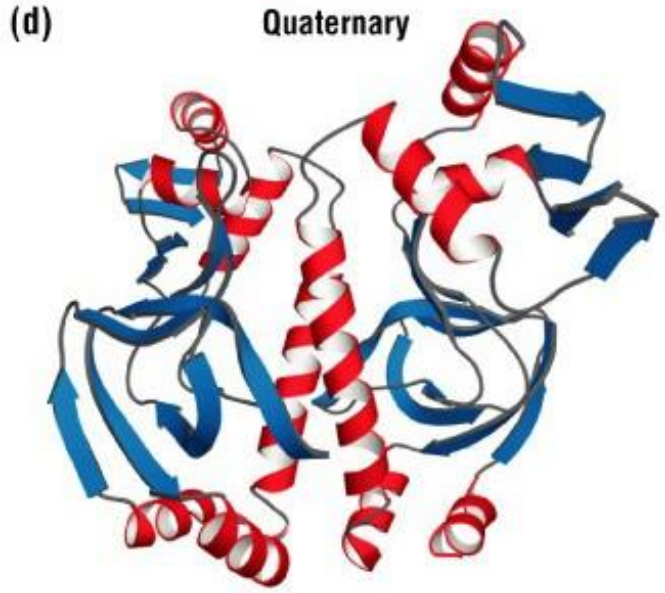
Estructura de proteínas

Modelado de proteínas

Los cuatro niveles de la estructura proteíca

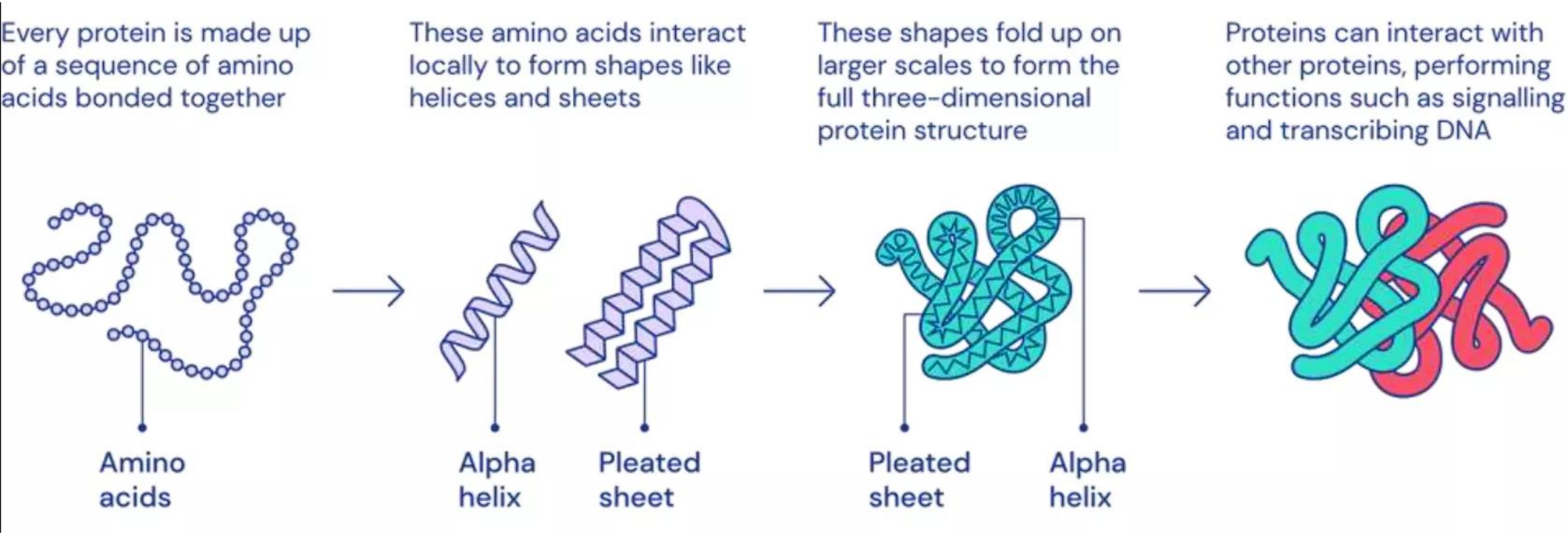






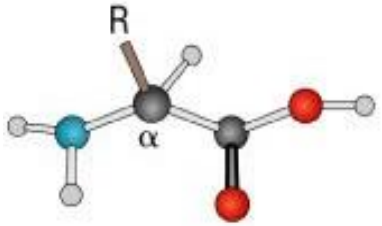
Los cuatro niveles de la estructura proteíca

Folding

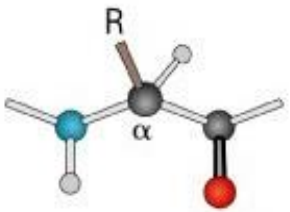


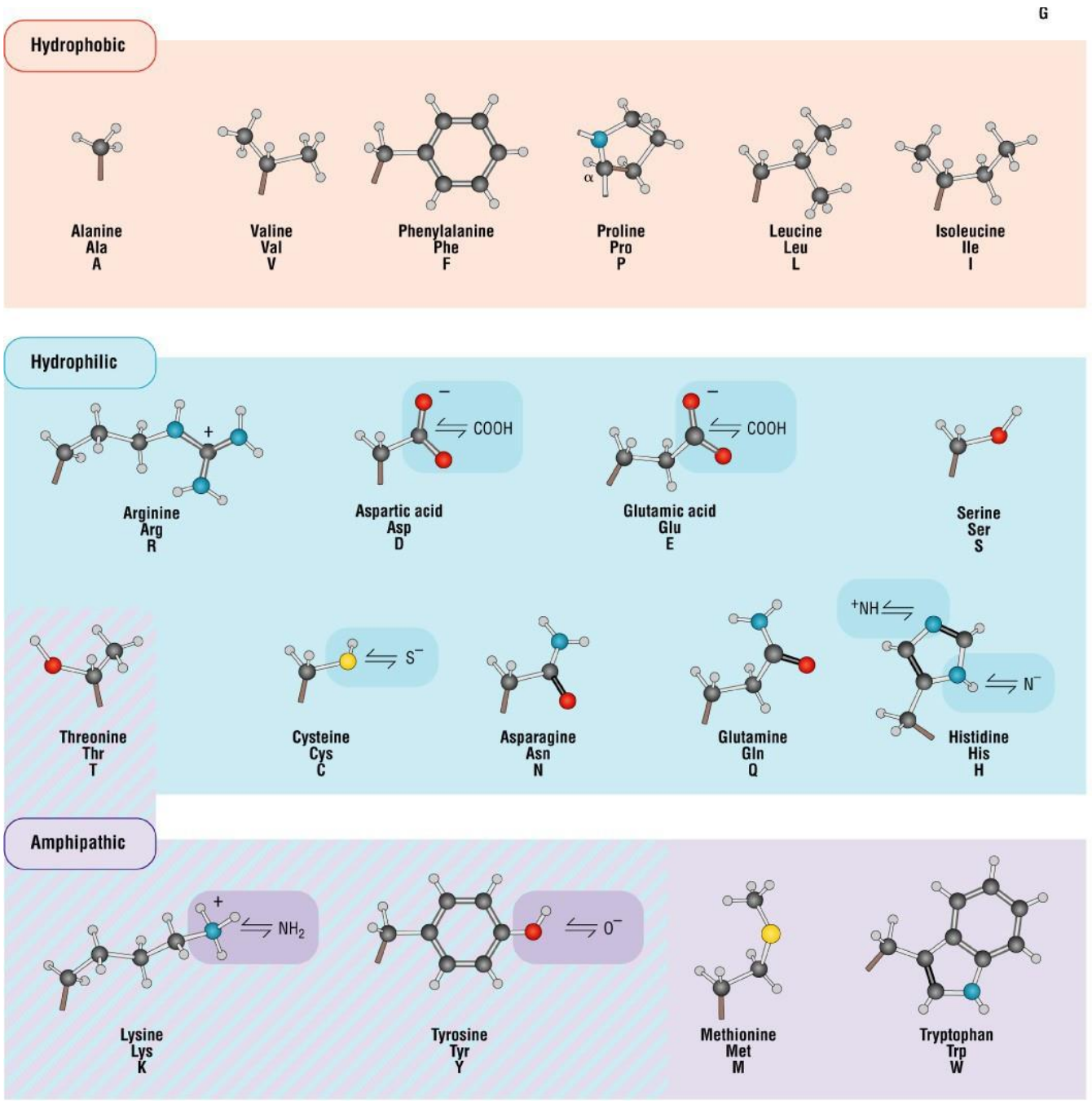
Los amino ácidos: estructura y carácter químico

Estructura general de un aminoácido

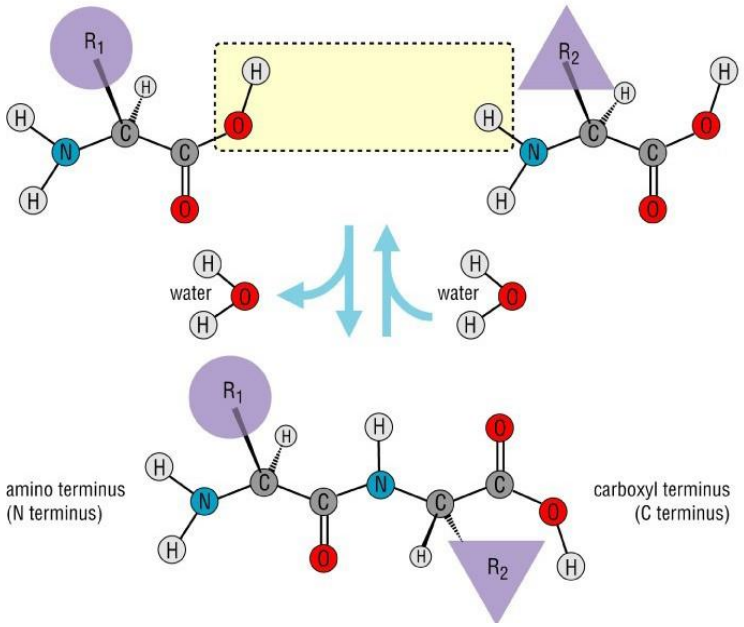


Formando enlace peptídico



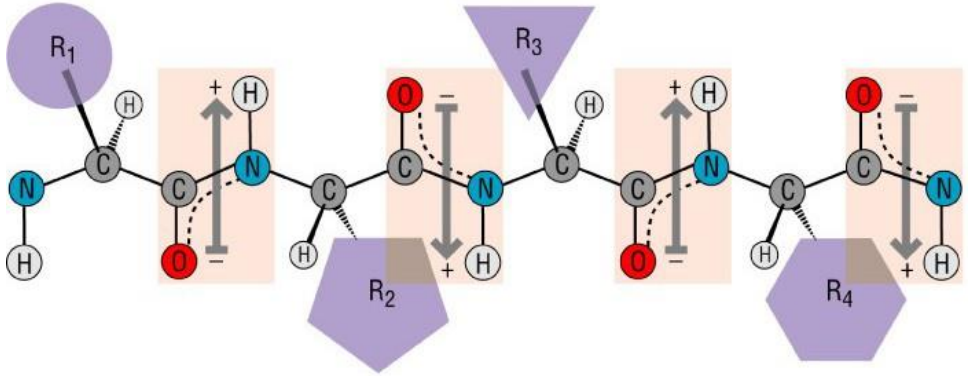


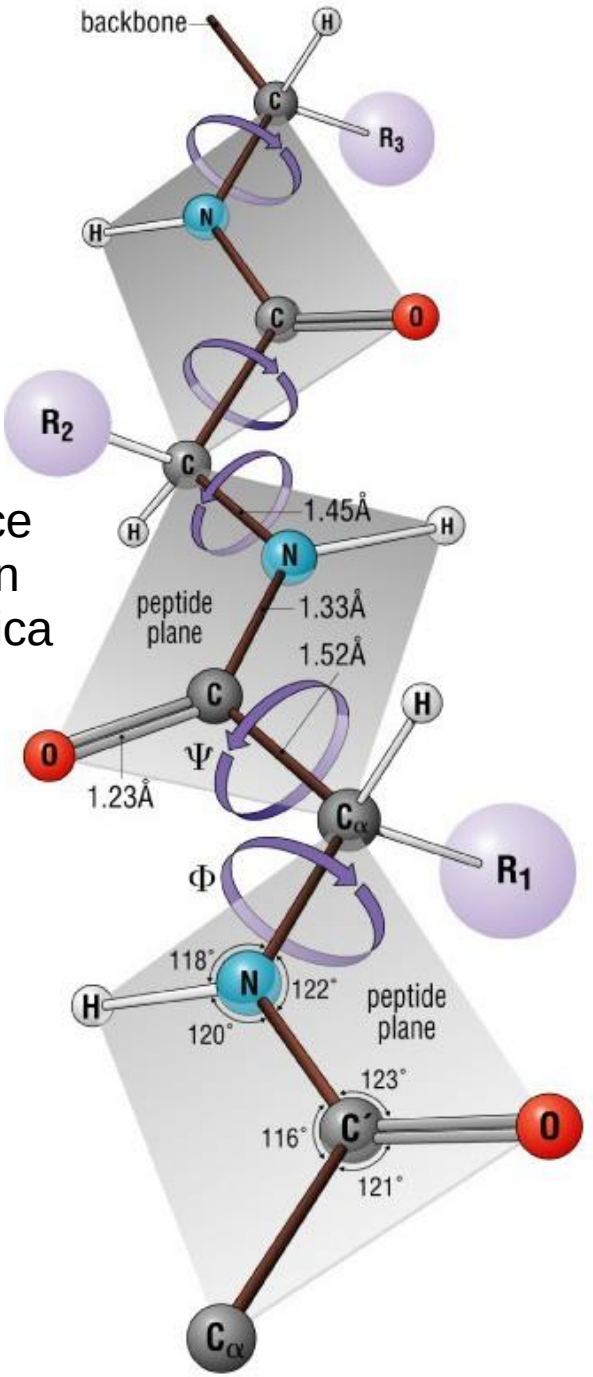
La propiedades del enlace peptídico afectan la estabilidad y flexibilidad de las proteínas



Los enlaces tipo amida son muy estables

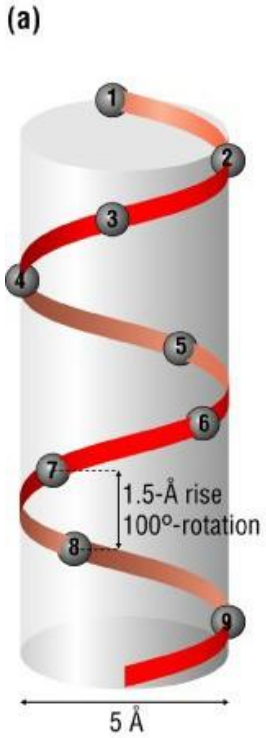
Caracter de doble enlace parcial afecta la rotación de la cadena polipeptídica



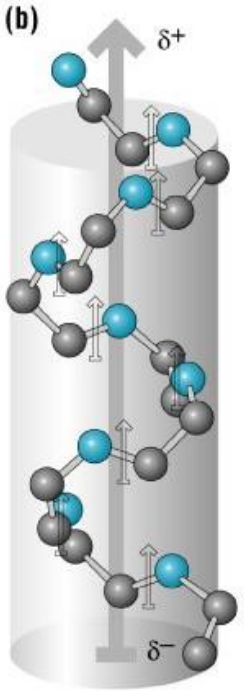


Resonancia de los enlaces tiene dos efectos: incremento de la estabilidad y momento dipolar

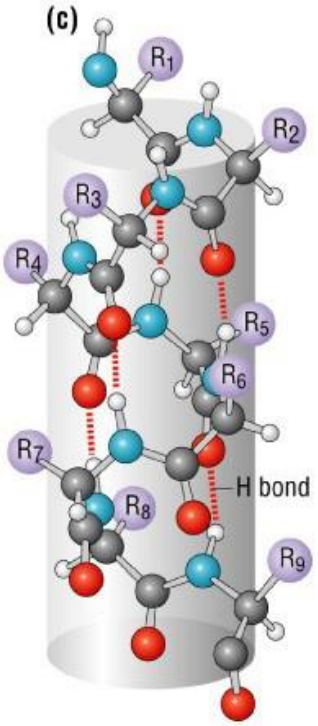
Estructura secundaria: Alfa hélices



Distancias

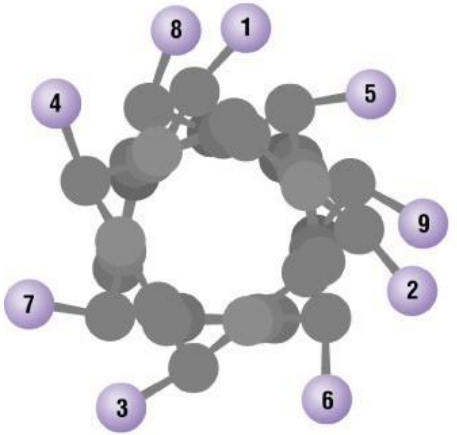


Dipolo



Puentes de hidrógeno

Vista superior



Las cadenas laterales determinan el carácter hidrofílico, hidrofóbico o anfipático de una alfa hélice

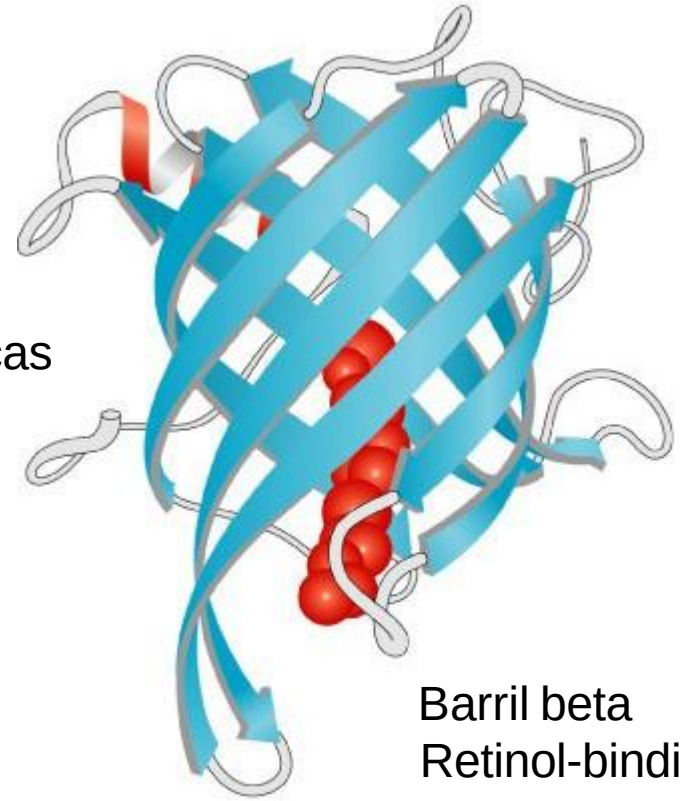
Variantes poco frecuentes de alfa hélices

Average Conformational Parameters of Helical Elements

Conformation	Phi	Psi	Omega	Residues per turn	Translation per residue
Alpha helix	–57	–47	180	3.6	1.5
3-10 helix	–49	–26	180	3.0	2.0
Pi-helix	57	–70	180	4.4	1.15
Polyproline I	–83	+158	0	3.33	1.9
Polyproline II	–78	+149	180	3.0	3.12
Polyproline III	–80	+150	180	3.0	3.1

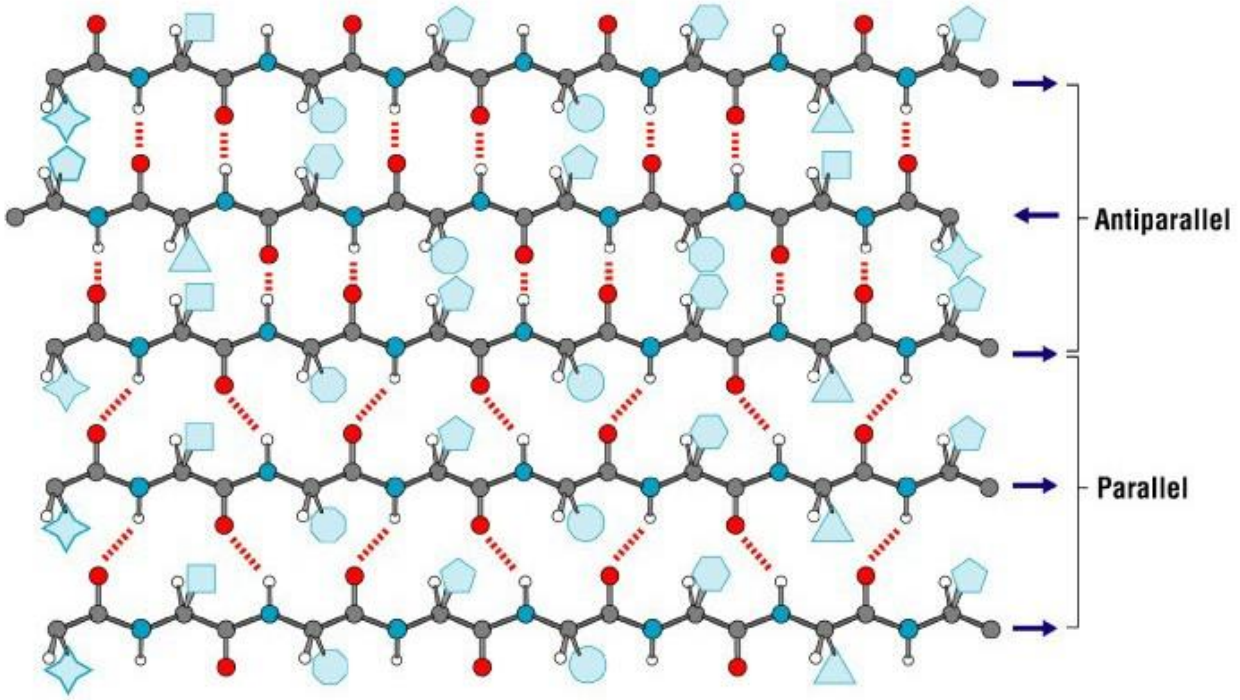
Estructura secundaria: Hojas beta

Hojas beta anfipáticas



Barril beta
Retinol-binding protein

Estructura hoja beta



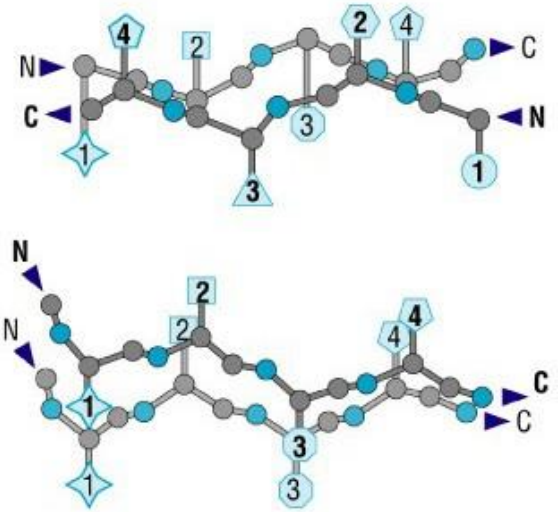


Gráfico de Ramachandran

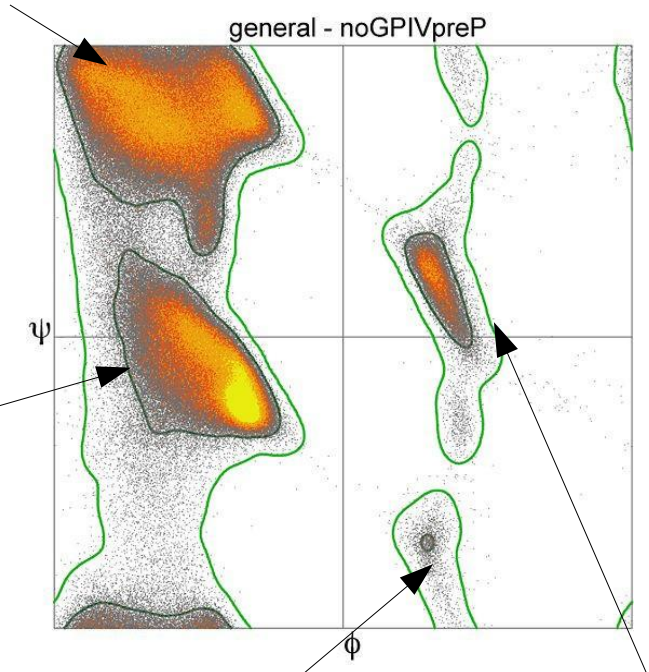
Hoja beta

Más de 1.000.000 de datos de alta calidad

Predicción de estructura secundaria

Conformational Preferences of the Amino Acids

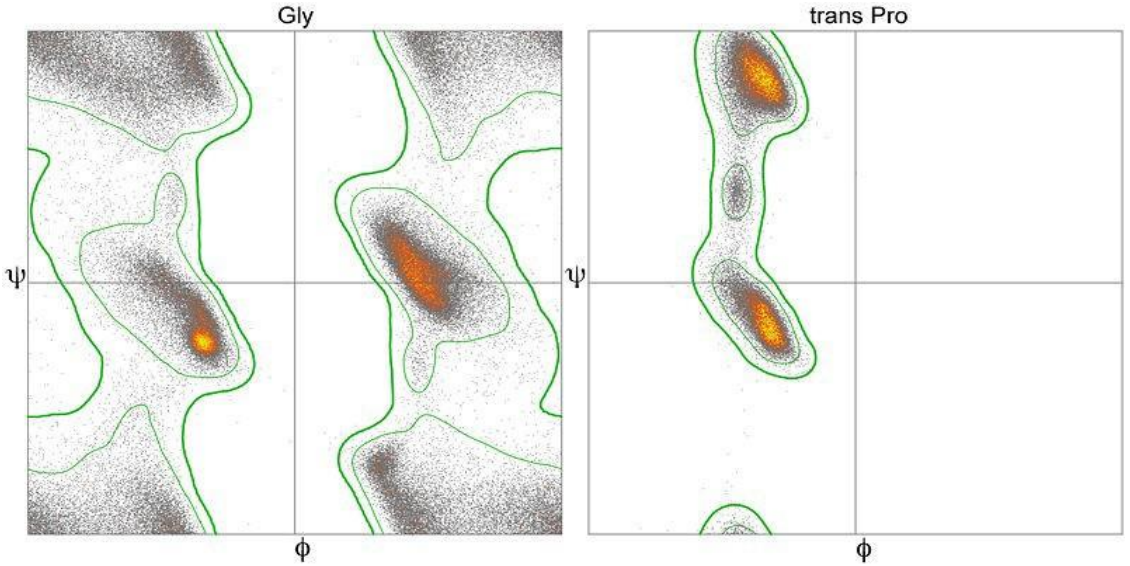
Amino acid	Preference α-helix	β-strand	Reverse turn
Glu	**1.59**	0.52	1.01
Ala	**1.41**	0.72	0.82
Leu	**1.34**	1.22	0.57
Met	**1.30**	1.14	0.52
Gln	**1.27**	0.98	0.84
Lys	**1.23**	0.69	1.07
Arg	**1.21**	0.84	0.90
His	**1.05**	0.80	0.81
Val	0.90	**1.87**	0.41
Ile	1.09	**1.67**	0.47
Tyr	0.74	**1.45**	0.76
Cys	0.66	**1.40**	0.54
Trp	1.02	**1.35**	0.65
Phe	1.16	**1.33**	0.59
Thr	0.76	**1.17**	0.90
Gly	0.43	0.58	**1.77**
Asn	0.76	0.48	**1.34**
Pro	0.34	0.31	**1.32**
Ser	0.57	0.96	**1.22**
Asp	0.99	0.39	**1.24**



Alfa hélice
Mano derecha

Beta turn

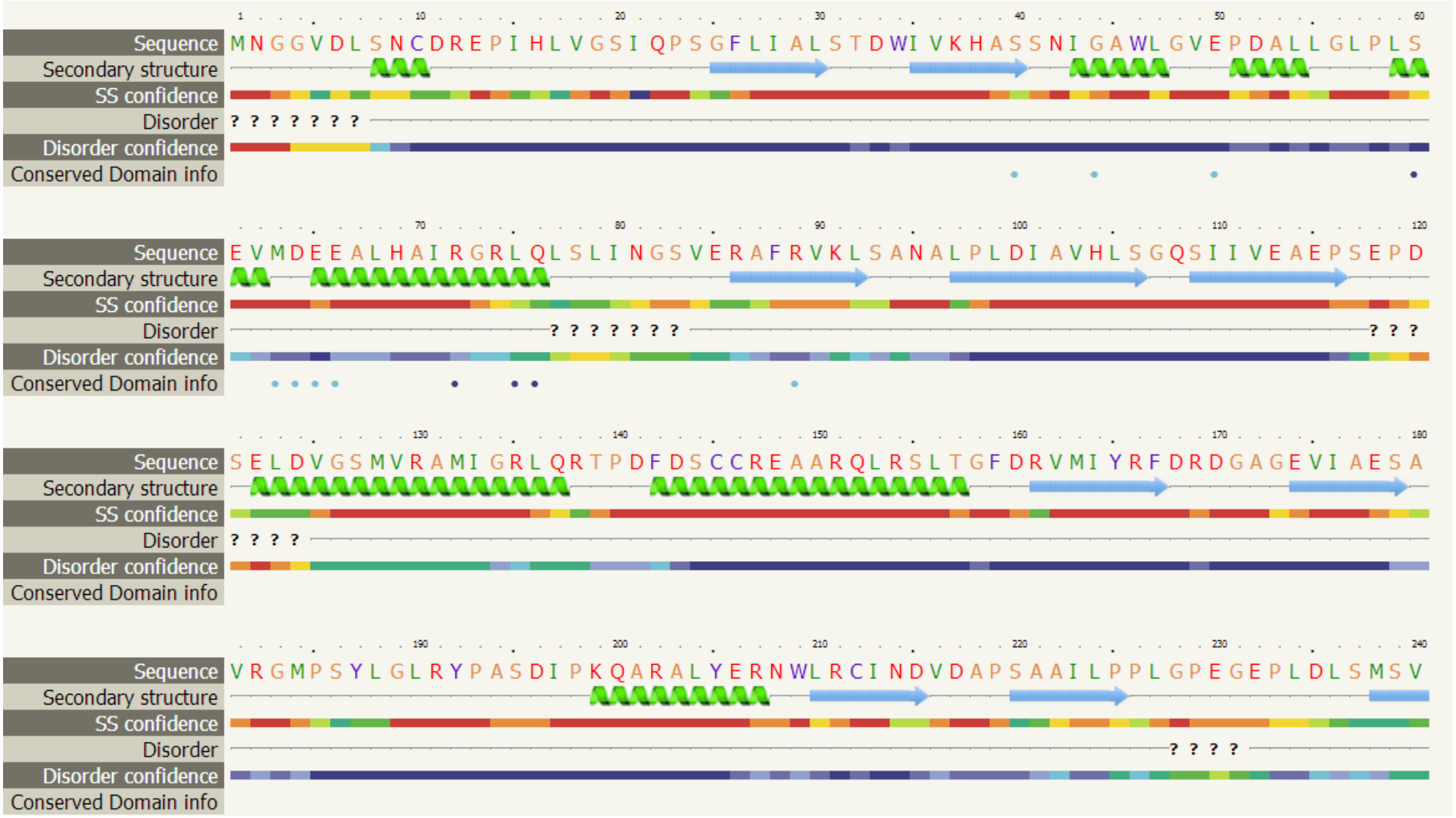
Alfa hélice
Mano izquierda



Prefencia por ciertos aminoácidos de acuerdo a la estructura

http://proteopedia.org/wiki/index.php/Ramachandran_Plots

Predicción de estructura secundaria

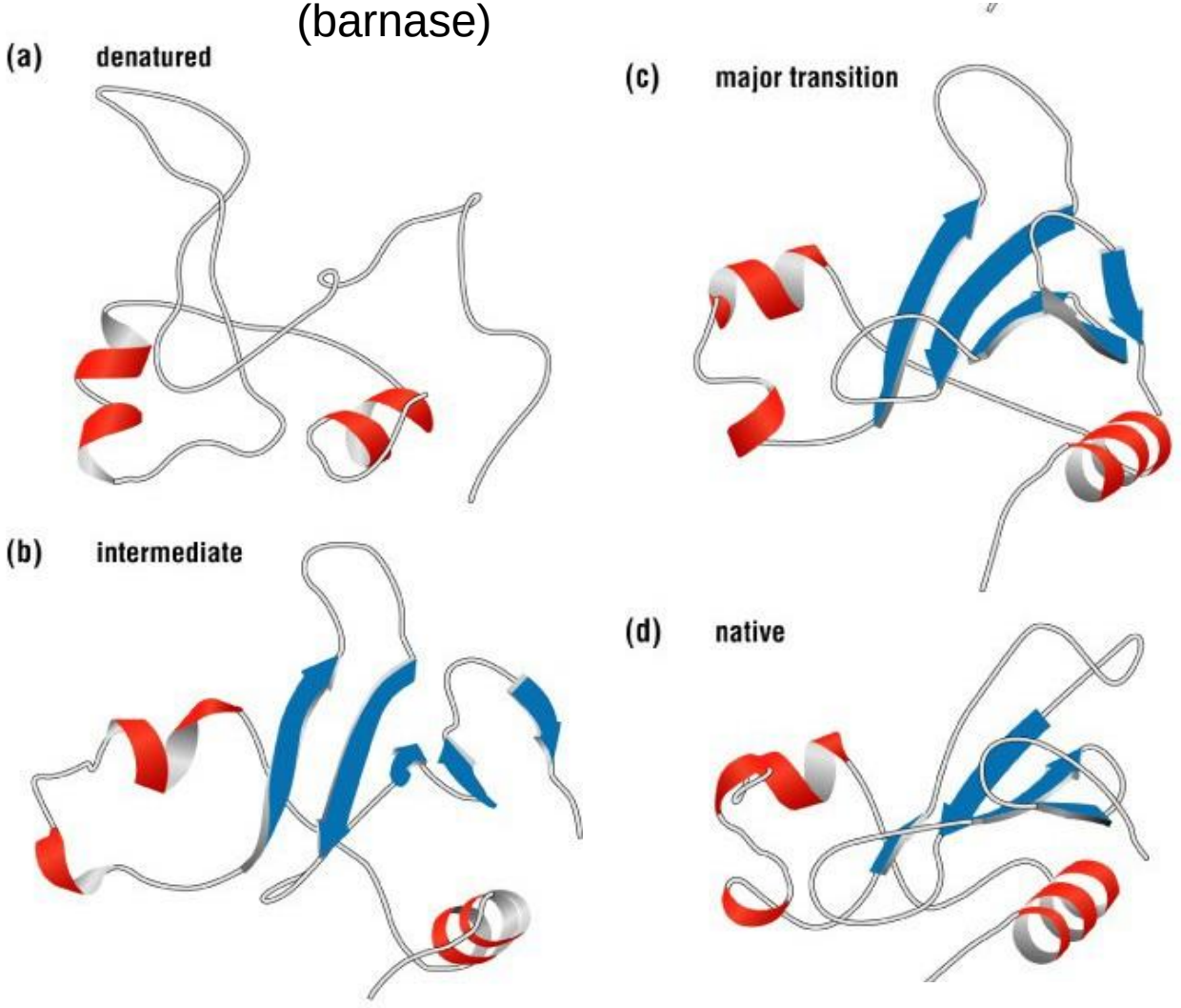


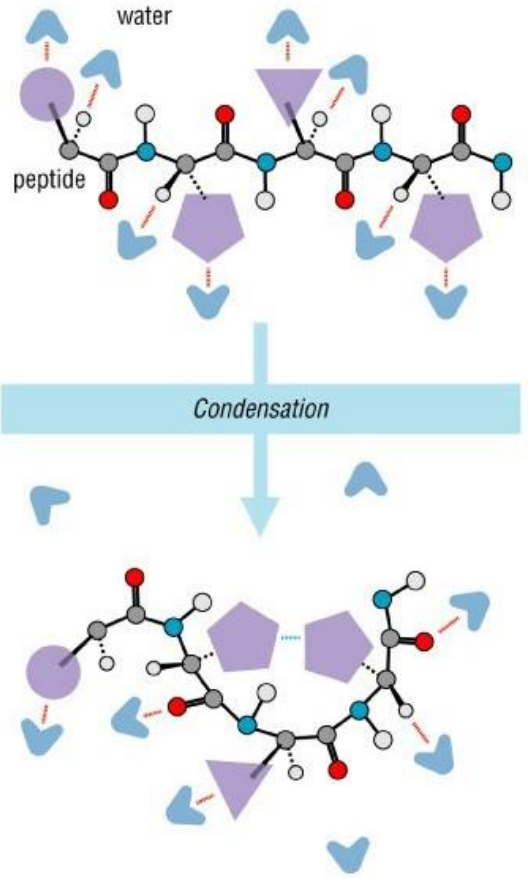
Phyre server www.sbg.bio.ic.ac.uk/phyre2/

Plegamiento de proteínas (Julio Caramelo)

La estructura primaria determina el plegamiento

Intermediarios de plegamiento (barnase)



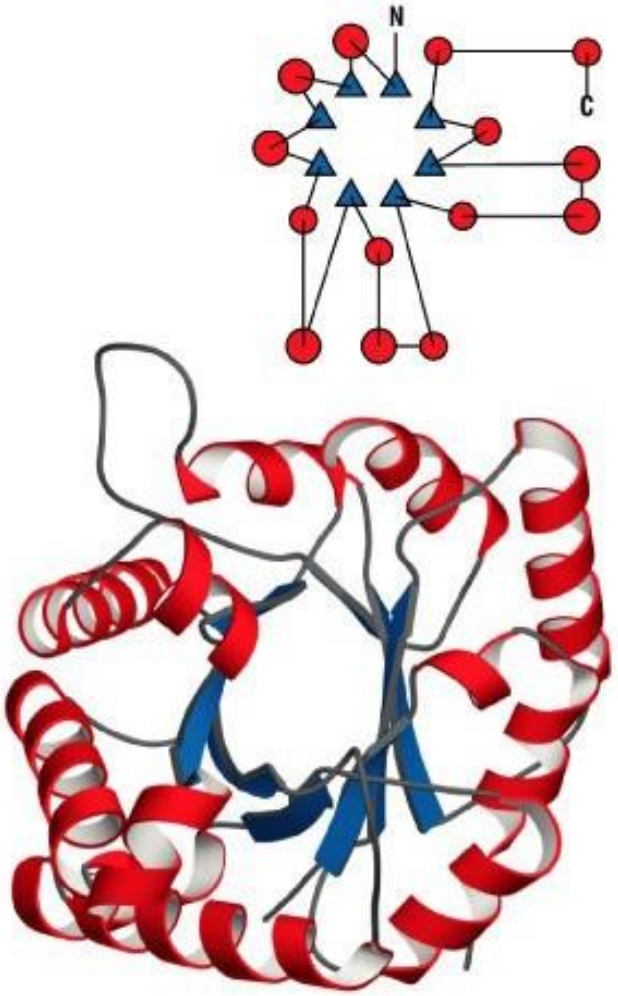


La competencia entre las interacciones internas y el agua controlan el plegamiento

Estructura terciaria

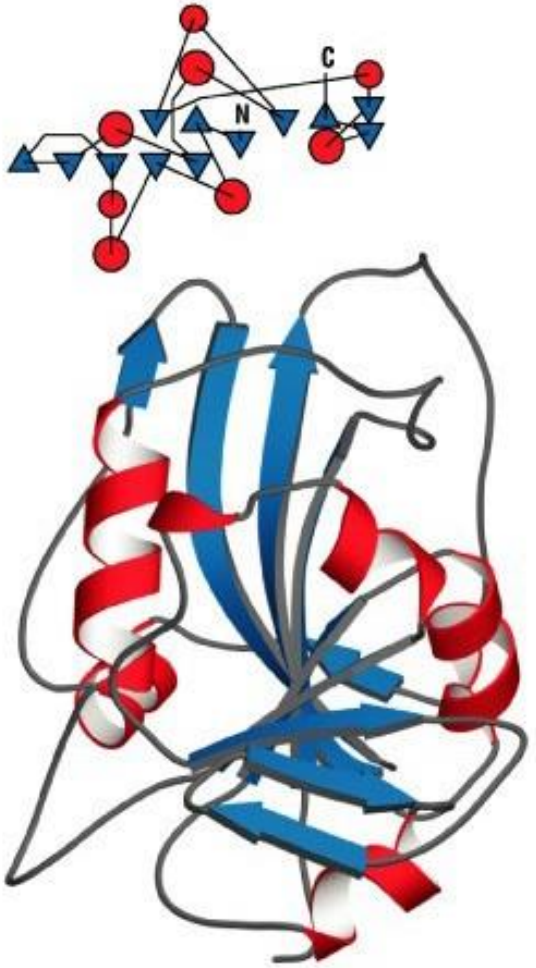
La condensación de múltiples elementos de estructura secundaria conduce a la estructura terciaria

Barril alfa/beta paralelo



TIM

Dominio alfa/beta con hojas mixtas

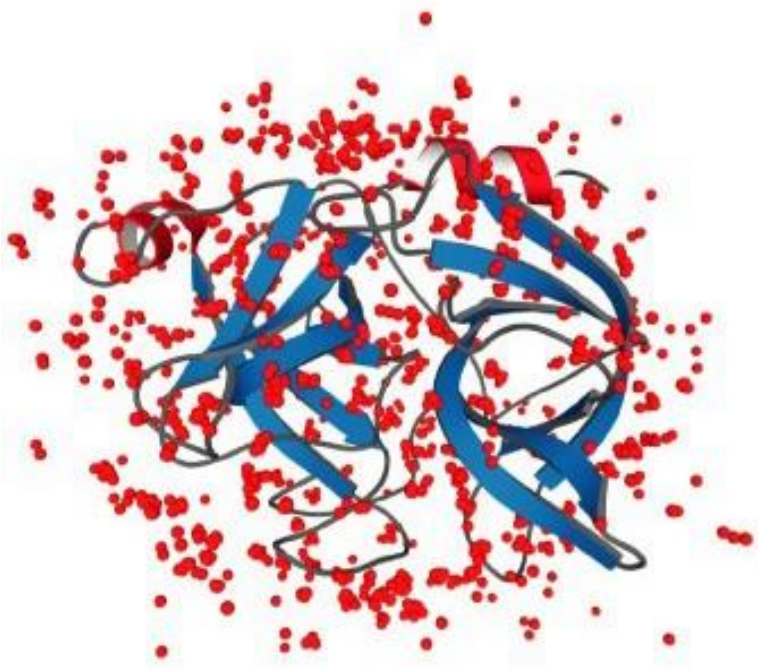


DHFR

Ambas tienen 8 hebras beta conectadas por alfa hélices

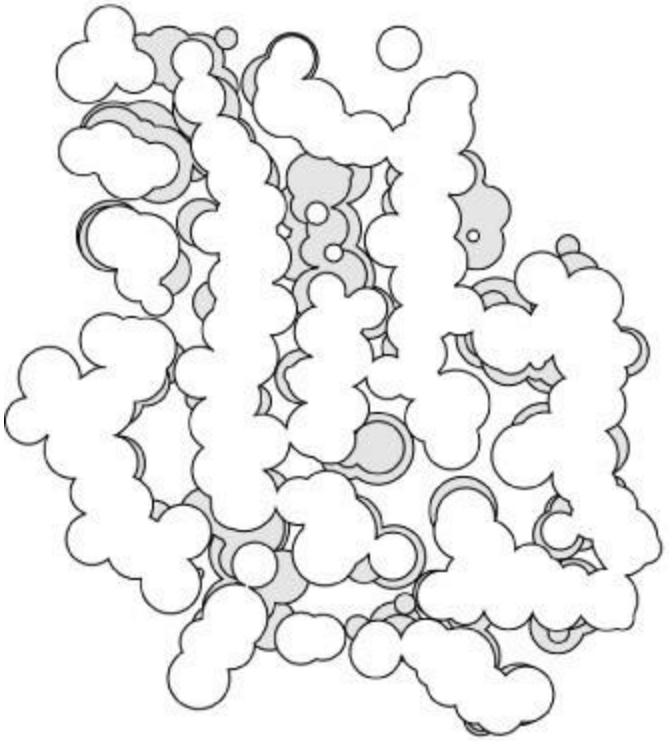
Estructura terciaria

Primera capa de hidratación de
la elastasa pancreatica porcina.



Las moléculas de agua unidas a la superficie
son una parte importante de la estructura
y se consideran parte de la estructura terciaria

Corte del interior de una proteína

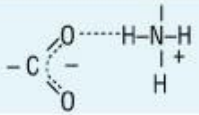
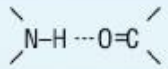
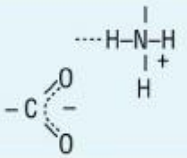
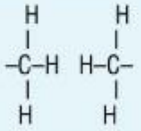


La estructura terciaria es estabilizada
por el empaquetamiento de los átomos

La proteínas plegadas son estabilizadas mediante interacciones débiles no covalentes

Débiles

Chemical Interactions that Stabilize Polypeptides

Interaction	Example	Distance dependence	Typical distance	Free energy (bond dissociation enthalpies for the covalent bonds)
Covalent bond	$-C_\alpha-C-$	-	1.5 Å	356 kJ/mole (610 kJ/mole for a C=C bond)
Disulfide bond	–Cys–S–S–Cys–	-	2.2 Å	167 kJ/mole
Salt bridge		Donor (here N), and acceptor (here O) atoms <3.5 Å	2.8 Å	12.5–17 kJ/mole; may be as high as 30 kJ/mole for fully or partially buried salt bridges (see text), less if the salt bridge is external
Hydrogen bond		Donor (here N), and acceptor (here O) atoms <3.5 Å	3.0 Å	2–6 kJ/mole in water; 12.5–21 kJ/mole if either donor or acceptor is charged
Long-range electrostatic interaction		Depends on dielectric constant of medium. Screened by water. 1/r dependence	Variable	Depends on distance and environment. Can be very strong in nonpolar region but very weak in water
Van der Waals interaction		Short range. Falls off rapidly beyond 4 Å separation. $1/r^6$ dependence	3.5 Å	4 kJ/mole (4–17 in protein interior) depending on the size of the group (for comparison, the average thermal energy of molecules at room temperature is 2.5 kJ/mole)

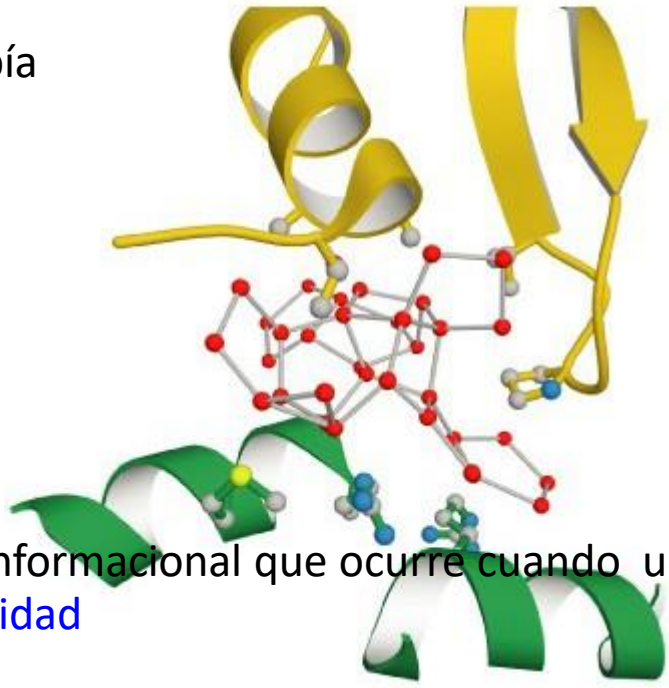
Estructura terciaria

El plegamiento proteíco es un compromiso termodinámico
Entalpía: calor liberado por la formación de las interacciones
Entropía: contribuído por el agua. El plegamiento incrementa la entropía del sistema. El **efecto hidrofóbico** contribuye a la entropía

Las aguas que rodean residuos hidrofóbicos están más ordenadas que las aguas líquidas. Cuando los residuos condesan, expulsan el agua incrementando la entropía

La estabilidad es definida con la **energía libre**, una función que combina tanto la entalpía como la entropía

La diferencia de energía libre entre los estados desplegado y plegado es de entre 21-42 Kj/mol



La energía liberada por la formación de los enlaces débiles es contrabalanceada por la enorme pérdida de estabilidad conformacional que ocurre cuando un polipéptido se pliega.

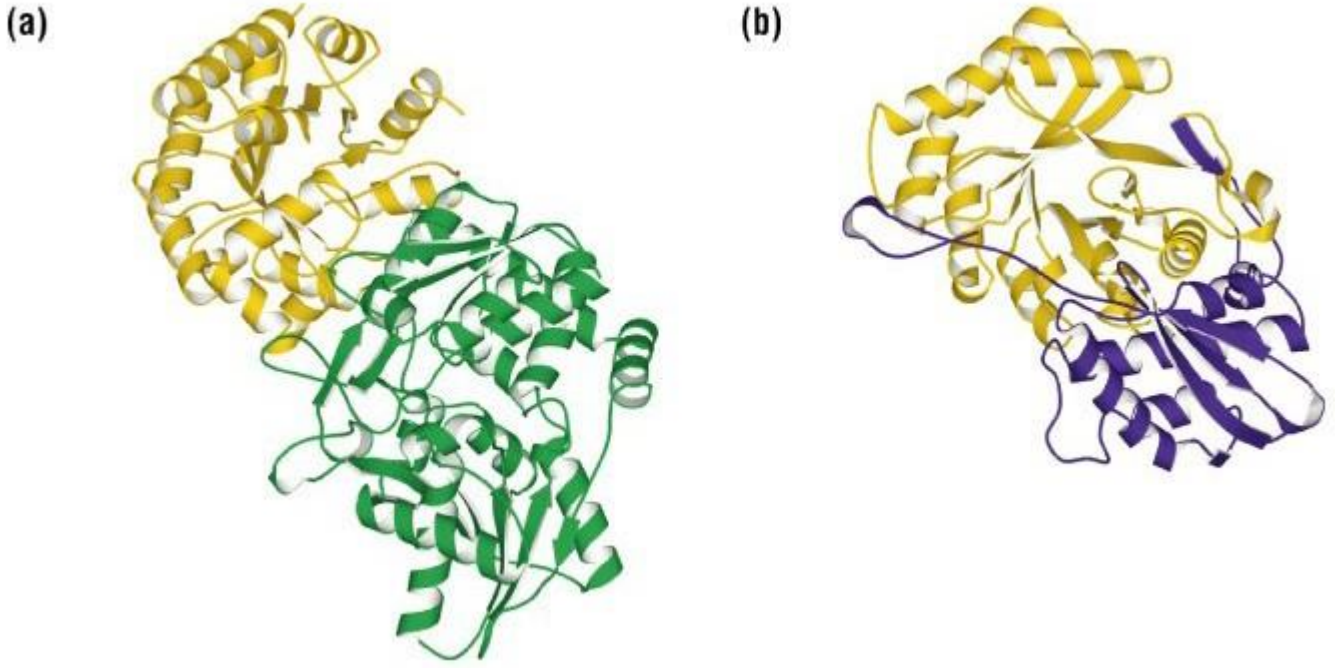
Consecuencia: Flexibilidad

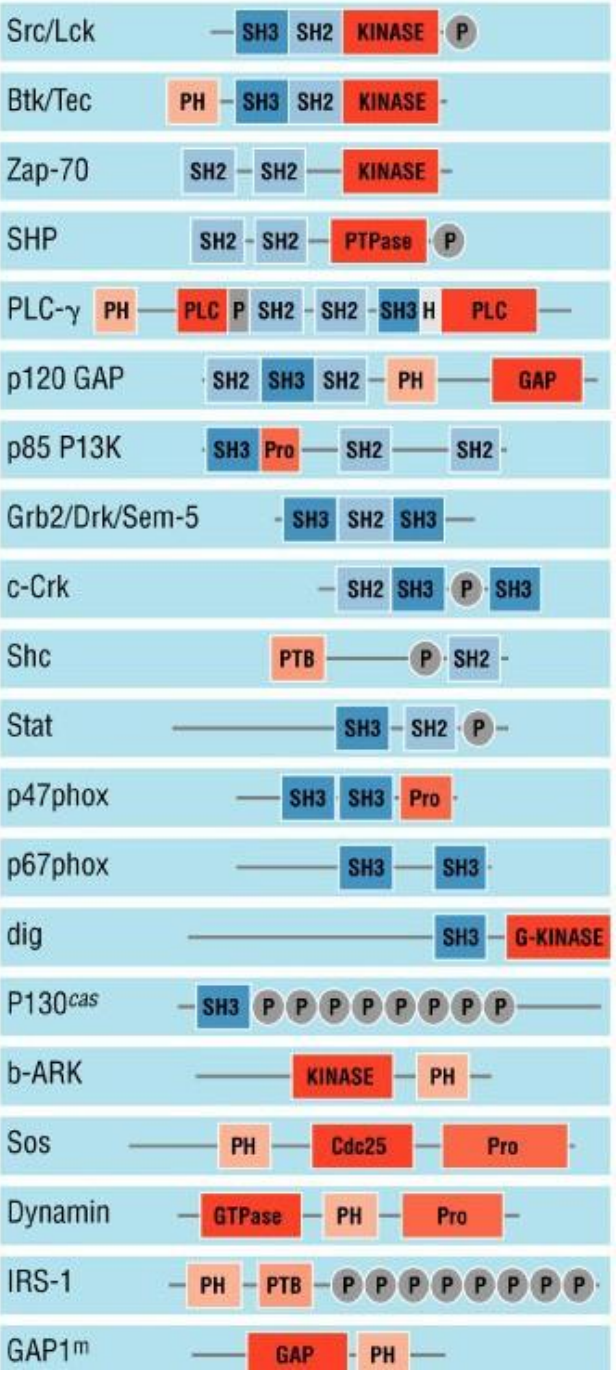
Los dominios proteícos son una región compacta de la proteína que, generalmente, esta formada por segmento continuo de amino ácidos y puede plegarse de manera estable por si misma en solución

Las proteínas multidominio evolucionaron por fusión de genes que codifican para proteínas separadas

Tryptophan synthase

Galactonate dehydratase





Los dominios de color amarillo son similares aunque no tienen similitud de secuencia o relación funcional

La proteínas son modulares, es decir, estan formadas por dominios que se intercambian (LEGO proteins)

La mayoría de las estructuras nuevas pueden dividirse en dominios previamente conocidos.

Los motivos estructurales tienen residuos claves conservados

Table 2.2 Amino acid sequences of calcium-binding EF motifs in three different proteins

Parvalbumin	V	K	K	A	F	A	I	I	D	Q	D	K	S	G	F	I	E	E	D	E	L	K	L	F	L	Q	N	F
Calmodulin	F	K	E	A	F	S	L	F	D	K	D	G	D	G	T	I	T	T	K	E	L	G	T	V	M	R	S	L
Troponin-C	L	A	D	C	F	R	I	F	D	K	N	A	D	G	F	I	D	I	E	E	L	G	E	I	L	R	A	T

Calcium-binding residues are orange, and residues that form the hydrophobic core of the motif are light green. The helix-loop-helix region shown underneath is colored as

¿Dónde están los AA?

- No todos los aminoácidos son iguales:
  - Cada uno tiene sus preferencias
    - Estructura secundaria
    - De primeros vecinos
    - De interior exterior
- Búsqueda de motivos estructurales:
  - Tema abierto en bioinformática
  - Métodos de búsqueda:
    - Grafos
    - Hashing
    - Manuales

Table 6.3 *The Packing of Residues in the Interior of Proteins*

Residue	Average volume of buried residues (Å³)[a]	Fraction of residues at least 95% buried[b]	Relative free energy of residue in interior to that on surface (kcal/mol)[c]
Gly	66	0.36	0
Ala	92	0.38	−0.14
Val	142	0.54	−0.55
Leu	168	0.45	−0.59
Ile	169	0.60	−0.68
Ser	99	0.22	0.40
Thr	122	0.23	0.32
Asp	125	0.15	0.78
Asn	125	0.12	0.75
Glu	155	0.18	1.15
Gln	161	0.07	0.80
Lys	171	0.03	2.06
Arg	202	0.01	1.40
His	167	0.17	0.02
Phe	203	0.50	−0.61
Tyr	204	0.15	0.28
Trp	238	0.27	−0.39
Cys	106[d]	0.40[d]	−0.61[d,e]
	118[e]	0.50[e]	—
Met	171	0.40	−0.65
Pro	129	0.18	0.50

[a] From C. Chothia, *Nature* 254:304–308 (1975).

[b] Average for 12 proteins. From C. Chothia, *J. Mol. Biol.* 105:1–14 (1976).

[c] Calculated as $-RT \log_e f$, where f is the ratio of the occurence of this amino acid residue on the interior to that on the surface. The values were normalized with that for Gly set to zero. From S. Miller et al., *J. Mol. Biol.* 105:641–656 (1987).

[d] When in disulfide form.

[e] When in thiol form.

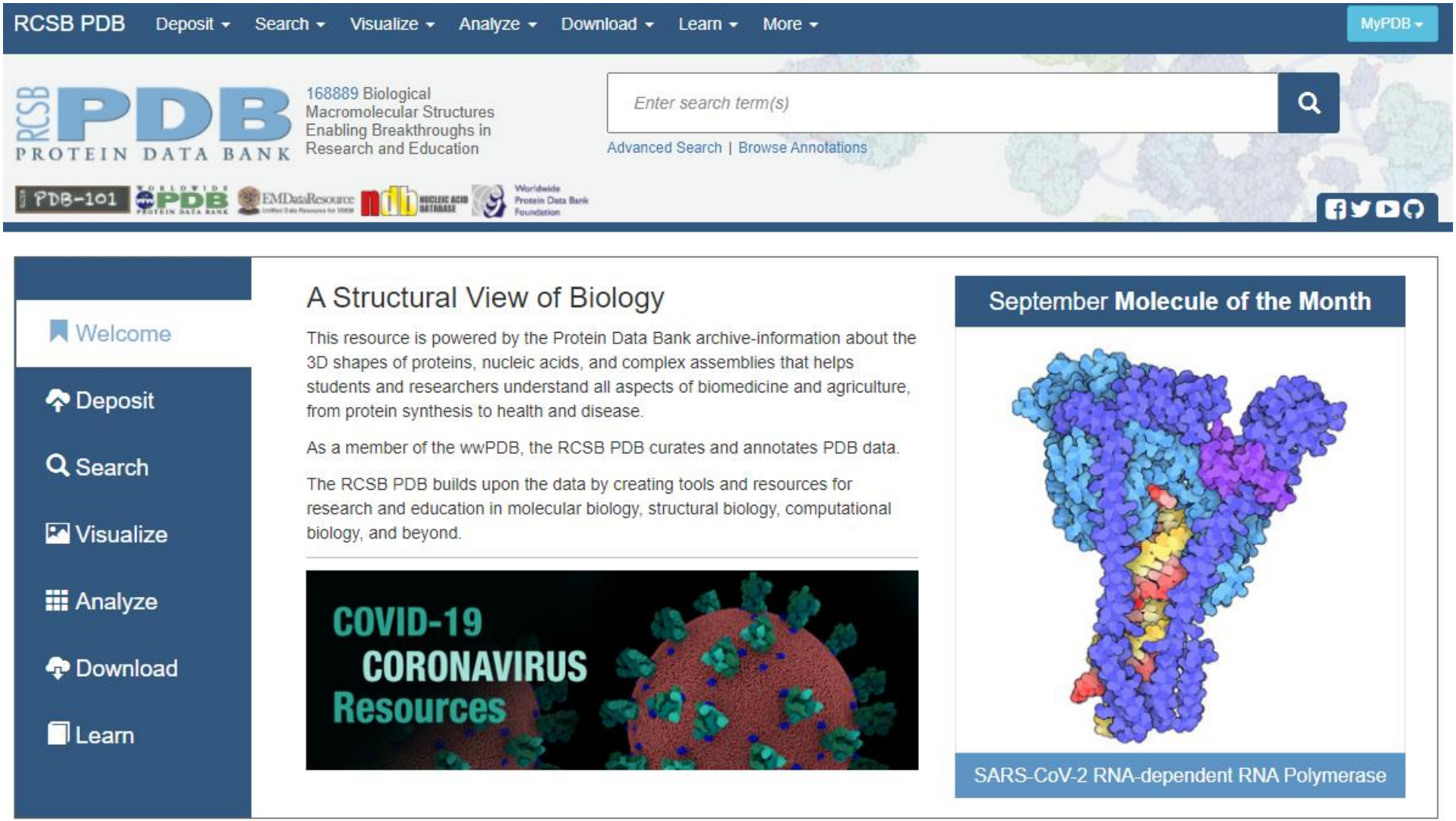
¿Cómo esta guardada la información estructural?

- Protein Data Bank (PDB)

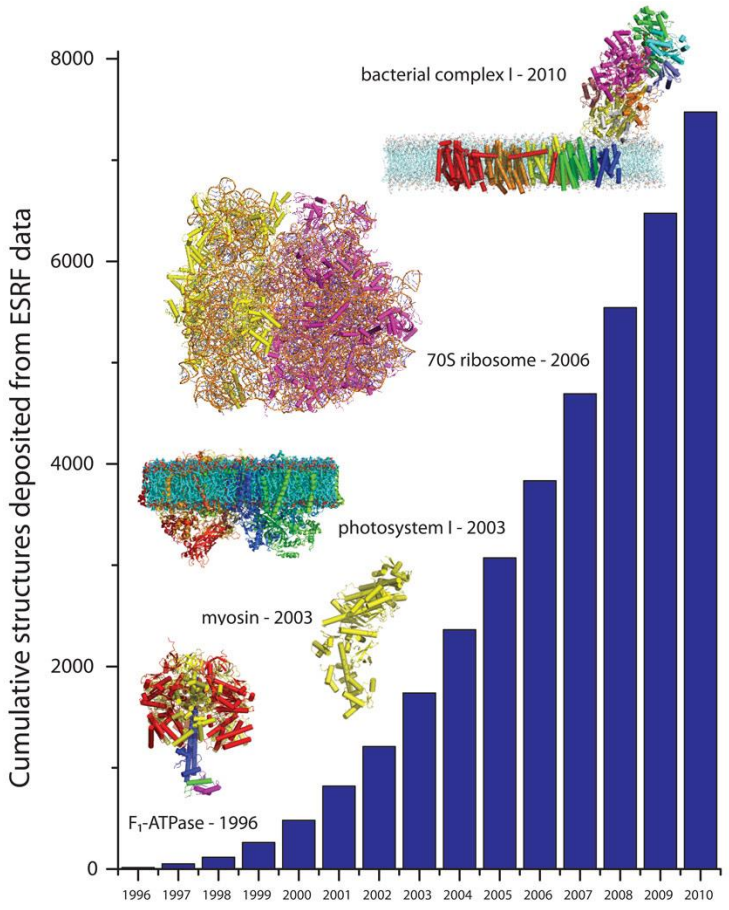
https://proteinstructures.com/Structure/Structure/proteinstructure-databases2.html#:~:text=In%20PDB%20files%20these%20conformations,(mean%2Dsquare%20displacement).

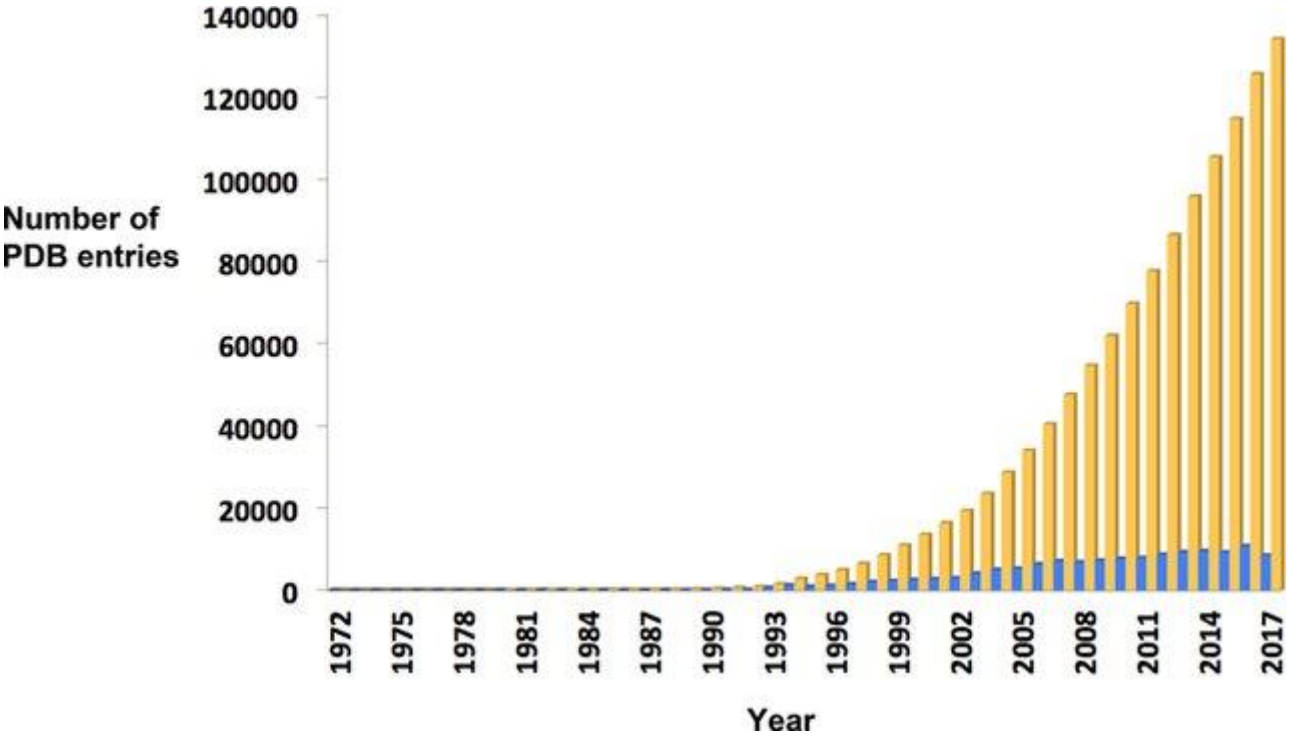
¿Qué es el PDB?

- Es el único repositorio Mundial sobre datos estructurales de Macromoléculas: Proteínas, ADN, ARN.
- en http://www.pdb.org



¿Cuántas Estructuras hay?





¿Qué información contiene un PDB?

- http://www.wwpdb.org/documentation/format23/v2.3.html
- Cada archivo es una colección de registros (de 1 o más líneas) divididos en campos
  - Cada línea es auto identificatoria, indicando con los primeros 6 caracteres el registro al que pertenece
- Los registros se pueden clasificar en secciones

¿Qué información contiene un PDB?

The file contains important information on the method used to solve the structure, on the various parameters related to the quality of the X-ray data (like resolution, the R-factor etc.).
The R-factor and resolution are the two most central parameters for the assessment of the quality of the structure. Higher resolution of the X-ray data and lower R-factor values ensure a better fit of the structure to the experimental electron density.

The PDB file also contains the symmetry operations for the specific space group of the crystal, data on the quality of the geometry of the model:

- deviations of bond lengths
- bond angles and torsion angles from ideal values
- secondary structure content
- description of missing regions in the structure (a result of weak electron density due to the flexibility of the structure

```
HEADER    IMMUNOGLOBULIN                          10-JUL-92   1IGM
TITLE     THREE DIMENSIONAL STRUCTURE OF AN FV FROM A HUMAN IGM
COMPND    2 MOLECULE: IGM-KAPPA POT FV (LIGHT CHAIN);
KEYWDS    IMMUNOGLOBULIN
EXPDTA    X-RAY DIFFRACTION
AUTHOR    Z.-C.FAN,L.W.GUDDAT,A.B.EDMUNDSON
REVDAT    2   01-APR-03 1IGM    1       JRNL
JRNL      REF    J.MOL.BIOL.                   V. 228   188 1992
JRNL      REFN   ASTM JMOBAK  UK ISSN 0022-2836
REMARK    2 RESOLUTION. 2.30 ANGSTROMS.
DBREF     1IGM L    1   115  GB     5524145  AAD44145         1    115
SEQADV    1IGM ALA L   34  GB   5524145   ASN    34 CONFLICT
SEQRES    1 L  115  ASP ILE GLN MET THR GLN SER PRO SER SER LEU SER ALA
SEQRES    2 L  115  SER VAL GLY ASP ARG VAL THR ILE THR CYS GLN ALA SER
FORMUL    3  HOH   *147(H2 O)
HELIX     1    1 GLN L   79  ILE L   83  5                                   5
SHEET     3    4 6 ALA H  92  HIS H  99 -1  O  ALA H  92   N  VAL H 117
SSBOND    1 CYS L   23    CYS L   88
SSBOND    2 CYS H   22    CYS H   96
CISPEP    1 SER L    7    PRO L    8          0       -16.82
ATOM      1  N   ASP L   1      10.690  42.371  25.139  1.00  8.79      N
ATOM      2  CA  ASP L   1      11.865  42.110  24.277  1.00  9.46      C
ATOM      3  C   ASP L   1      11.968  43.306  23.320  1.00  9.32      C
```

Identificación del registro

Campos

Sección Título:

- Contiene la información que describe a que experimento y que molécula corresponde el archivo
- Posee los siguientes registros: HEADER, OBSLTE, TITLE, CAVEAT, COMPND, SOURCE, KEYWDS, EXPDTA, AUTHOR, REVDAT, SPRSDE, JRNL, y REMARK
- El registro HEADER identifica unívocamente al pdb: 3 campos: código pdb (#XYZ) , frase identificatoria, fecha de ingreso de deposito.

```
HEADER    HYDROLASE (CARBOXYLIC ESTER)            08-APR-93   2PHI
```

- El registro TITLE describe le registro como lo hace el titulo de un paper
- El registro COMPND describe la macromolécula con mayor detalle. (nombre de la proteína, cadena, etc.)

```
COMPND   2 MOLECULE: HEMOGLOBIN;
COMPND   3 CHAIN: A, B, C, D;
COMPND   4 ENGINEERED: YES;
COMPND   5 MUTATION: YES
COMPND   6 OTHER_DETAILS: DEOXY FORM
```

Sección de Estructura Primaria

- Contiene la información sobre la secuencia de la Macromolecula , y LNKS a otras bases de datos

- El registro DBREF provee la conexión a otras bases de datos (Gene Bank, Swiss Prot, Trembl etc.)

  DBREF 3HSV A 1 92 SWS P22121 HSF_KLULA 193 284

  pdbid-chain data-base DB-cod DB-Id-Code

- El registro MODRES posee información sobre modificaciones post-traduccionales. Glycosylación, Fosforilación, bloqueo Nt, Aminación Ct, Configuración-D, etc.

  MODRES 3ABC DAL A 32 ALA POST-TRANSLATIONAL MODIFICATION,D-ALANINE

- El registro HET se utiliza para describir residuos no Standard como grupos prostéticos, inhibidores, moléculas de solvente etc.

  3 tipos: HETNAM, HETSYN, FORMUL

Part of a PDB file header

```
REMARK 1
REMARK 2
REMARK 2 RESOLUTION. 2.10 ANGSTROMS.
REMARK 3
REMARK 3 REFINEMENT.
REMARK 3 PROGRAM : CNS 1.0
REMARK 3 AUTHORS : BRUNGER,ADAMS,CLORE,DELANO,GROS,GROSSE-
REMARK 3 : KUNSTLEVE,JIANG,KUSZEWSKI,NILGES, PANNU,
REMARK 3 : READ,RICE,SIMONSON,WARREN
REMARK 3
REMARK 3 REFINEMENT TARGET : ENGH & HUBER
REMARK 3
REMARK 3 DATA USED IN REFINEMENT.
REMARK 3 RESOLUTION RANGE HIGH (ANGSTROMS) : 2.10
REMARK 3 RESOLUTION RANGE LOW (ANGSTROMS) : 29.55
REMARK 3 DATA CUTOFF (SIGMA(F)) : 0.000
REMARK 3 DATA CUTOFF HIGH (ABS(F)) : 312841.620
REMARK 3 DATA CUTOFF LOW (ABS(F)) : 0.0000
REMARK 3 COMPLETENESS (WORKING+TEST) (%) : 97.9
REMARK 3 NUMBER OF REFLECTIONS : 22179
REMARK 3
REMARK 3 FIT TO DATA USED IN REFINEMENT.
REMARK 3 CROSS-VALIDATION METHOD : THROUGHOUT
REMARK 3 FREE R VALUE TEST SET SELECTION : RANDOM
REMARK 3 R VALUE (WORKING SET) : 0.214
REMARK 3 FREE R VALUE : 0.247
REMARK 3 FREE R VALUE TEST SET SIZE (%) : 10.000
REMARK 3 FREE R VALUE TEST SET COUNT : 2207
REMARK 3 ESTIMATED ERROR OF FREE R VALUE : 0.005
REMARK 3
```

Part of a PDB file header showing some of the data like resolution (2.1 Å), resolution range of the data, No of reflections collected from the crystal, R-factor, etc. After that we can read about the start and end residues of the helices and β-strands in the structure. And last in this view are the coordinates of the atoms:

- Sección de Estructura Secundaria
  - Describe las zonas de estructura secundaria. HELIX, SHEET, TURN

    HELIX 1 HA GLY A 86 GLY A 94 1

- Sección de transformaciones asociadas al cristal
  - Describe las transformaciones geometricas asociadas al experimento cristalografico. CRYST1, ORIGXn, SCALEm, MATRIXn, TVEC

Sección de Estructura Secundaria

```
HELIX 1 1 PRO A 22 ILE A 26 5 5
HELIX 2 2 GLN A 29 ASP A 42 1 14
HELIX 3 3 PRO A 43 GLY A 46 5 4
HELIX 4 4 ASP A 53 GLY A 57 5 5
HELIX 5 5 SER A 59 LEU A 69 1 11
HELIX 6 6 ASN A 84 ILE A 88 5 5
HELIX 7 7 SER A 114 GLY A 120 1 7
HELIX 8 8 ASP A 123 GLY A 131 1 9
HELIX 9 9 GLY A 138 ASN A 144 1 7
HELIX 10 10 GLU A 152 LEU A 156 5 5
HELIX 11 11 GLU A 157 GLY A 171 1 15
HELIX 12 12 ARG A 202 ASP A 207 1 6
HELIX 13 13 ASP A 220 ASP A 237 1 18
HELIX 14 14 ASP A 237 LEU A 263 1 27
HELIX 15 15 PRO A 264 VAL A 266 5 3
HELIX 16 16 PRO A 269 LEU A 283 1 15
HELIX 17 17 GLY A 287 GLU A 305 1 19
HELIX 18 18 GLY A 311 SER A 324 1 14
HELIX 19 19 HIS A 325 LEU A 327 5 3
HELIX 20 20 VAL A 341 LEU A 349 1 9
SHEET 1 A 5 VAL A 106 LEU A 109 0
SHEET 2 A 5 GLY A 146 ILE A 150 1 O TYR A 147 N VAL A 107
SHEET 3 A 5 PHE A 188 GLY A 194 1 O VAL A 189 N LEU A 148
SHEET 4 A 5 VAL A 48 PHE A 51 1 N VAL A 48 O LEU A 190
SHEET 5 A 5 LEU A 211 GLU A 214 1 O LEU A 211 N LEU A 49
SHEET 1 B 2 ILE A 72 VAL A 75 0
SHEET 2 B 2 VAL A 99 LYS A 102 -1 N ILE A 100 O ALA A 74
SHEET 1 C 2 ALA A 121 LEU A 122 0
SHEET 2 C 2 PHE A 135 GLU A 136 -1 N GLU A 136 O ALA A 121
SHEET 1 D 2 GLU A 172 VAL A 175 0
SHEET 2 D 2 ILE A 182 PRO A 185 -1 O ILE A 182 N VAL A 175
CRYST1 90.259 90.259 83.716 90.00 90.00 120.00 P 65 6
```

Further down there is a list of the secondary structure elements within the structure, also showing the first and last residue in each element:

Sección de Coordenadas

- El registro ATOM contiene las coordenadas anotadas por átomo, cada línea corresponde a un átomo. Los átomos de residuos no Standard poseen registro HETATM con el mismo formato.

```
ATOM    149  CB AVAL A  25      30.385  17.437  57.230  0.28 13.88           C
ATOM    150  CB BVAL A  25      30.166  17.399  57.373  0.72 15.41           C
```

- Campos: Número del átomo, Nombre del átomo, (Conformación alternativa), Nombre del aminoácido, cadena, Número del aminoacido, coords: X-Y-Z, ocupancia, Factor-B, símbolo

- El registro SIGATM presenta la desviación estandard de la posición del átomo
- El registro ANISOU presenta los valores de Temperatura anisotropicos (simil B-fact)
- Los registros MODEL-ENDMDL marcan el inicio y fin de un modelo resuelto por RMN

Sección de Coordenadas

```
ATOM 1 N ARG A 18 14.699 61.369 62.050 1.00 39.19 N
ATOM 2 CA ARG A 18 14.500 62.241 60.856 1.00 38.35 C
ATOM 3 C ARG A 18 13.762 61.516 59.729 1.00 36.05 C
ATOM 4 O ARG A 18 14.354 60.740 58.982 1.00 34.91 O
ATOM 5 CB ARG A 18 15.850 62.753 60.334 1.00 42.36 C
ATOM 6 CG ARG A 18 16.537 63.770 61.247 1.00 46.92 C
ATOM 7 CD ARG A 18 17.825 64.314 60.629 1.00 51.24 C
ATOM 8 NE ARG A 18 18.442 65.347 61.462 1.00 54.15 N
```

Identificación del registro

Campos

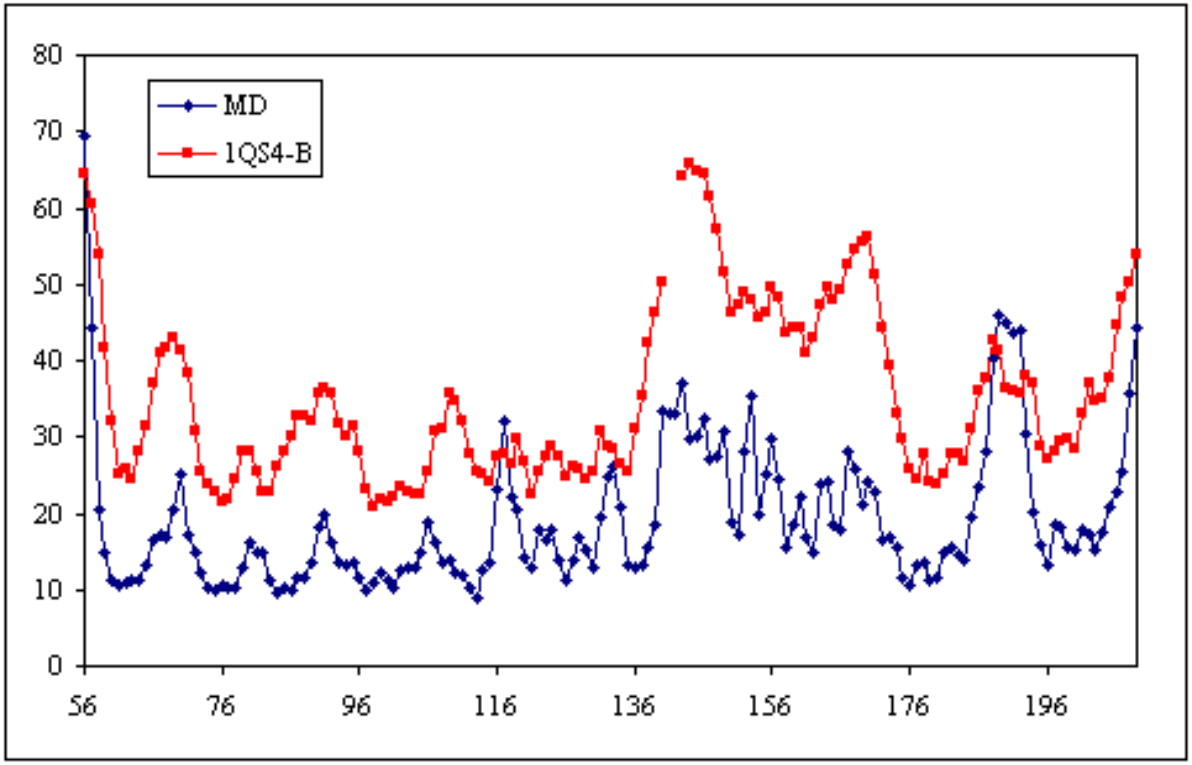
After the general informational part, the x,y,z coordinates of the atoms are listed:

notice that the structure starts at amino acid Arg 18! Amino acids from 1 to 17 are missing. The reason is that there was no electron density for these residues to build them into the model (see for example the discussion on structure quality in homology modeling).

B-factor

El B-factor describe la fluctuación del átomo en el cristal, esta relacionado directamente con la fluctuación media de un átomo en una dinámica Molecular

The numbers in the last column in the file are called the temperature factors, or B-factor, for each atom in the structure. The B-factor describes the displacement of the atomic positions from an average (mean) value (mean-square displacement). Higher flexibility results in larger displacements, and eventually lower electron density.



The values of the B-factors are normally between 15 to 30 (sq. Angstroms), but often much higher than 30 for flexible regions.

Modelado de Proteínas

Introducción:

para que predecir estructuras?

Etapas en la predicción?

Modelado por Homologia

Alineamiento

Cadena Carbonada

Loops

Rotameros

Chequeo

Ab-initio

Potenciales estadisticos

Dinamica Molecular.

Threading

Phyre

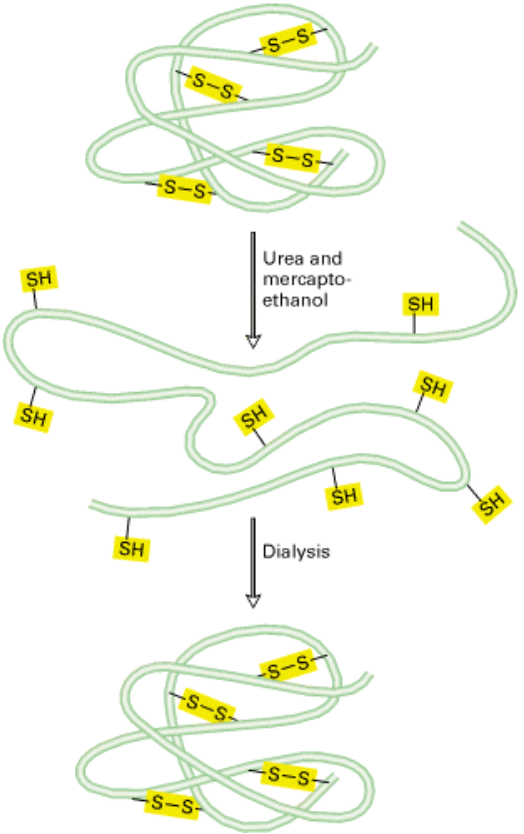
De la secuencia a la estructura

Toda la información de la estructura nativa de la proteína esta codificada en la secuencia + la solución o entorno en el que se encuentra.

Afinsen!

Anfinsen wanted to show that the information for protein folding resided entirely within the amino acid sequence of the protein. He choose ribonuclease A as his model for folding but he couldn't completely denature the protein unless he treated it with the denaturant urea **plus** 2ME to break the disulfide bridges.

Under those conditions, the protein unfolded. It would refold spontaneously once he removed urea and 2ME from the folding solution. Ribonuclease A regained biological activity under those conditions. This demonstrated that refolding could take place *in vitro*.

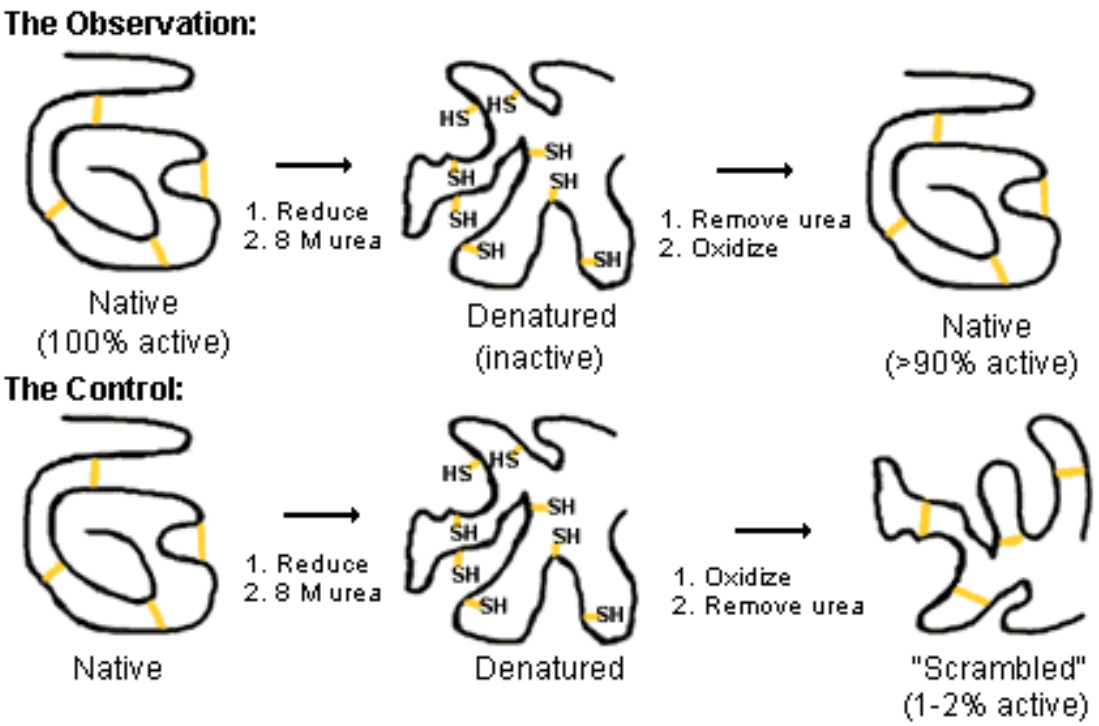


Anfinsen, 1973

De la secuencia a la estructura

Anfinsen discovered that removing 2ME but not urea led to recovery of 1% of the activity. This is attributed to the formation of random disulfide bridges between the 8 cysteines present in the protein. There are 105 different possibilities (7x5x3x1) so the 1% recovery makes sense.

It also shows that the correct three-dimensional conformation must be achieved fairly rapidly when urea is removed since most of the protein under those conditions becomes active.

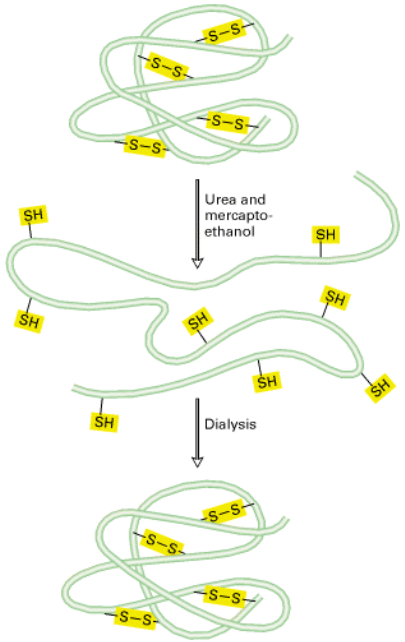


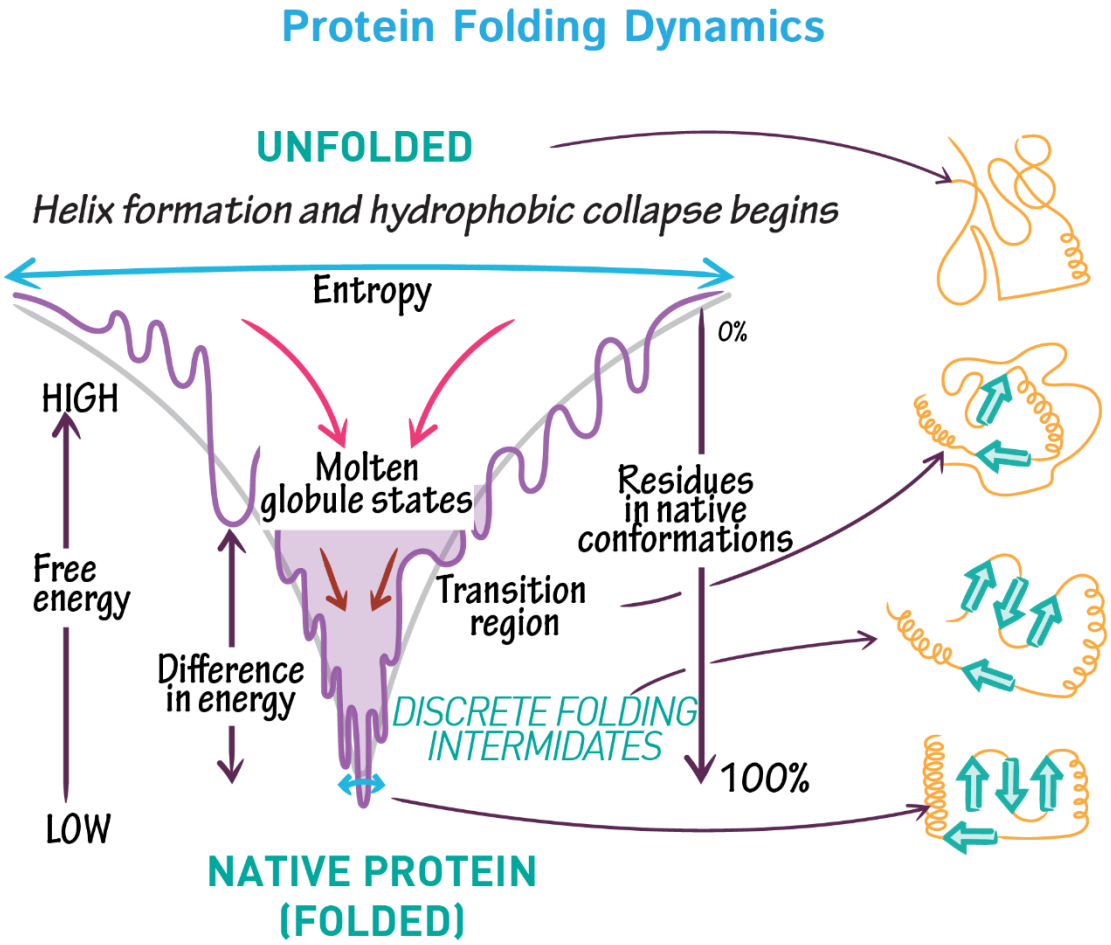
Anfinsen, 1973

De la secuencia a la estructura

"Thermodynamic Hypothesis", which states that the native conformation of a protein is adopted spontaneously.

El estado nativo es el mínimo de energía libre!





His summary of the experiments was presented as a Nobel Prize Lecture and published in: Anfinsen, C.B. (1973) "Principles that govern the folding of protein chains." Science 181 223-230. *Anfinsen, 1973*

Paradoja de Levinthal

1969 Cyrus Levinthal :

- Encontrar el Estado nativo visitando conformaciones al azar NO ES POSIBLE.

- Cadena polipeptídica corta tiene 10^{143} conformaciones

  Conformaciones son probadas en nanosegundos o picosegundos.
  Tiempo mayor a la edad del universo!!!!

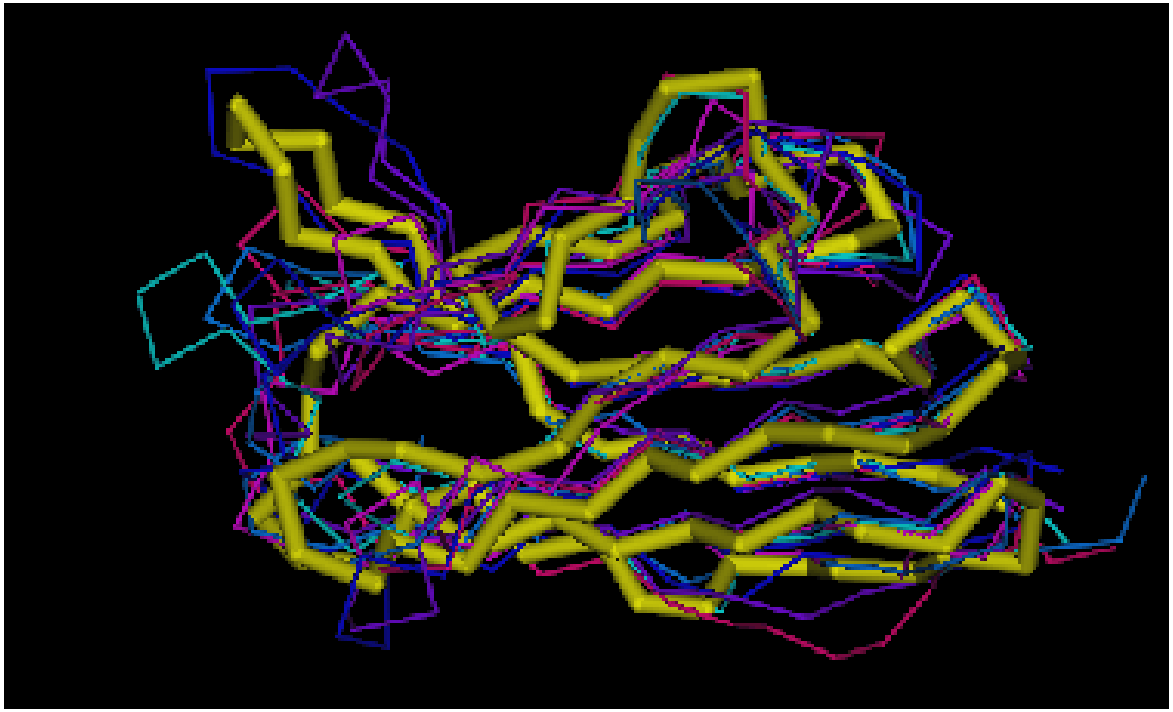
Si no podemos desde primeros principios entonces:

- Podemos usar la información existente
- Comparar e inferir

Métodos de predicción usando información almacenada en bases de datos

- Modelado por Homología
- Threading o Enebrado

Modelado por homologia o comparativo



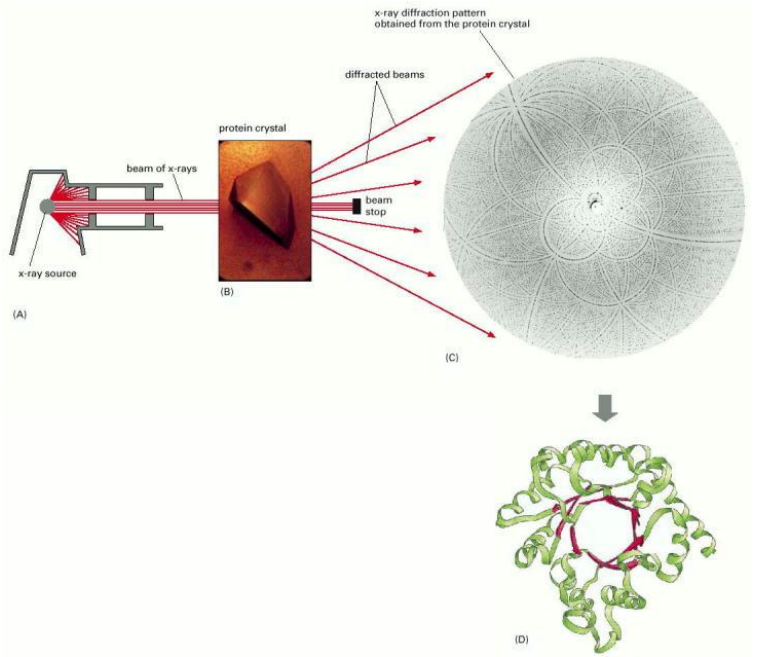
Es un método que permite predecir la Estructura terciaria (3D) de una proteína deseada (Target) conociendo su secuencia de aminoácidos (Estructura primaria)

La estructura terciaria de una homologa (template) resuelta por rayos-X o NMR

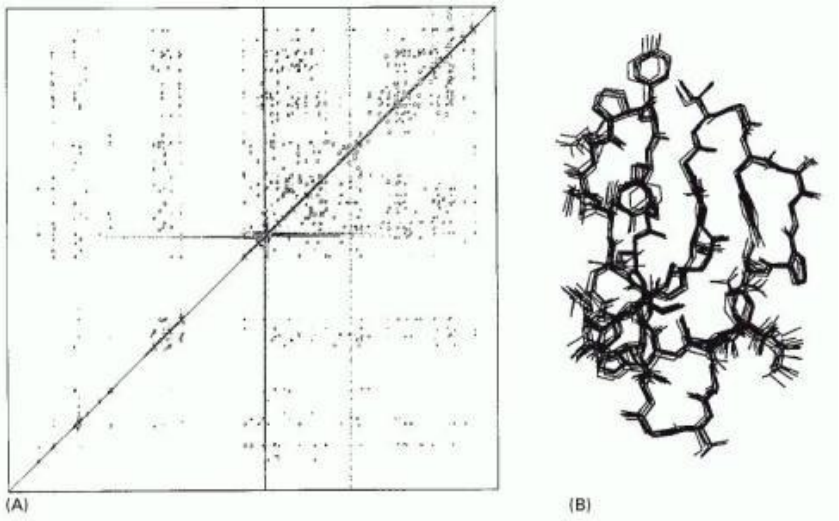
```
ACDEFGHIKLMNPQRST--FGHQWERT-----TYREWYEGHADS
ASDEYAHLRILDPQRSTVAYAYE--KSFAPPGSFKWEYEAHADS
MCDEYAHIRLMNPERSTVAGGHQWERT----GSFKEWYAAHADD
```

Si comparten similitud de secuencia seguramente tendrán estructuras parecidas

Determining structure: X-ray crystallography



Determinacion de la estructura: NMR spectroscopy



- Las espectroscopias de X-Ray NMR permiten determinar las estructuras de proteinas y sus complejos.
- Estos metodos son caros y requieren de mucho trabajo y tiempo.
- Por suerte! Toda la informacion esta centralizada en una base de datos(PDB)

METHOD	X-ray	NMR
hydrogen atoms	✘	✓
well defined –$CONH_2$	✘	✓
water molecules	✓	✘
Metal ions, cofactors	✓	✓

¿Por qué predecir?

- Muchas veces Nuestra proteína de interés no tiene estructura 3D conocida
- HM es el mejor de los método de predicción de estructura terciaria a partir de la secuencia
- El universo de los “folds” posibles es chico:

Se estima que entre 1000-7000

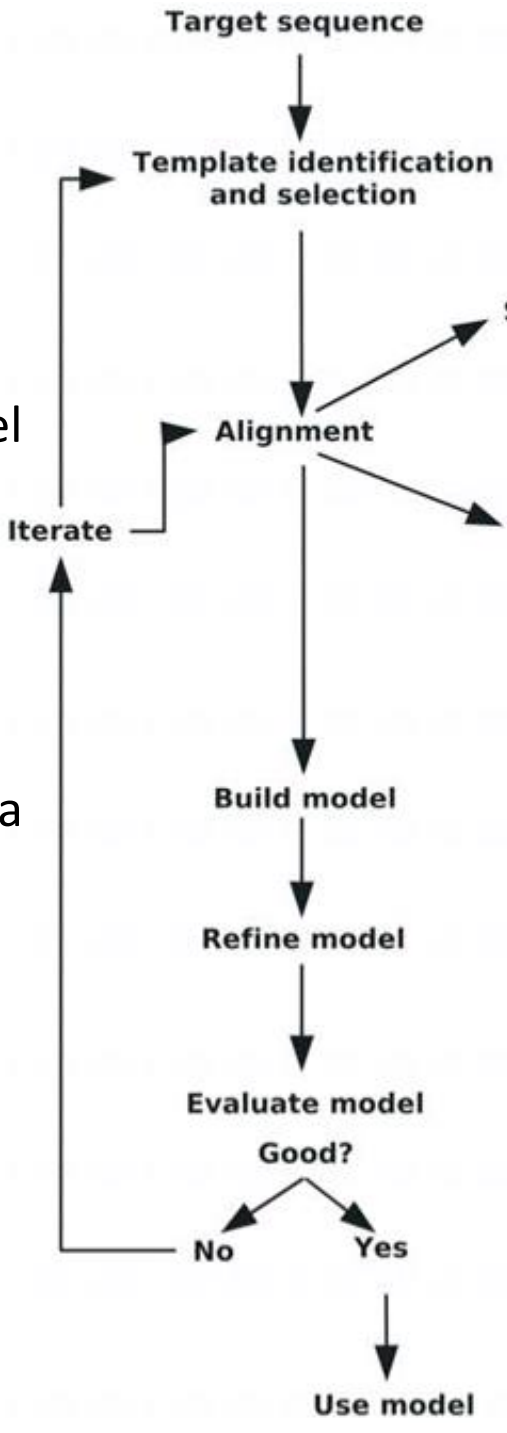
¿Por qué funciona la predicción?

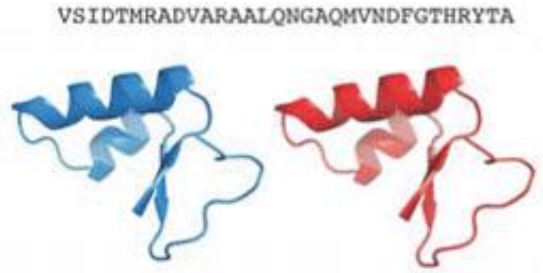
- Se sabe que secuencias similares adoptan estructuras similares
- En una familia los residuos conservados en la secuencia, conservan la misma estructura
- La topología de los residuos del sitio activo también es conservada
- El proceso de evolución por selección natural “tolera” variaciones de la secuencia

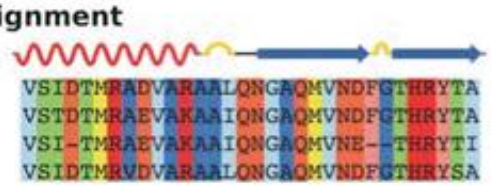
Los pasos a seguir…

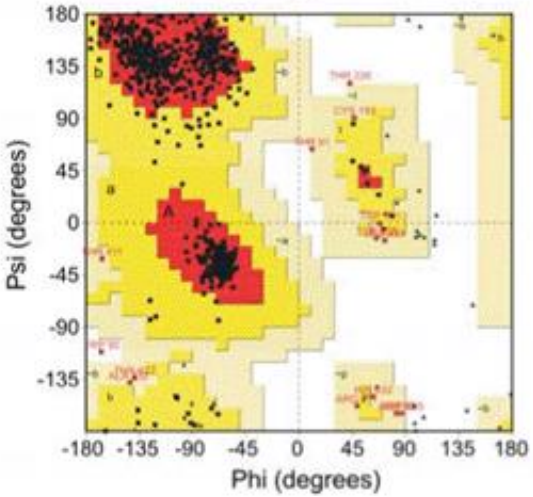
Etapas generales modelado por homología

1. Encontrar un templado adecuado (Ej: BLAST pdb).
2. Alineamiento Target-Template (Es el paso más importante).
    1) Corregir el alineamiento
3. Construcción del modelo.
    1) Modelo de la cadena carbonada
    2) Modelado de los loops
    3) Cadenas Laterales
4. Evaluación
5. Refinamiento del modelo



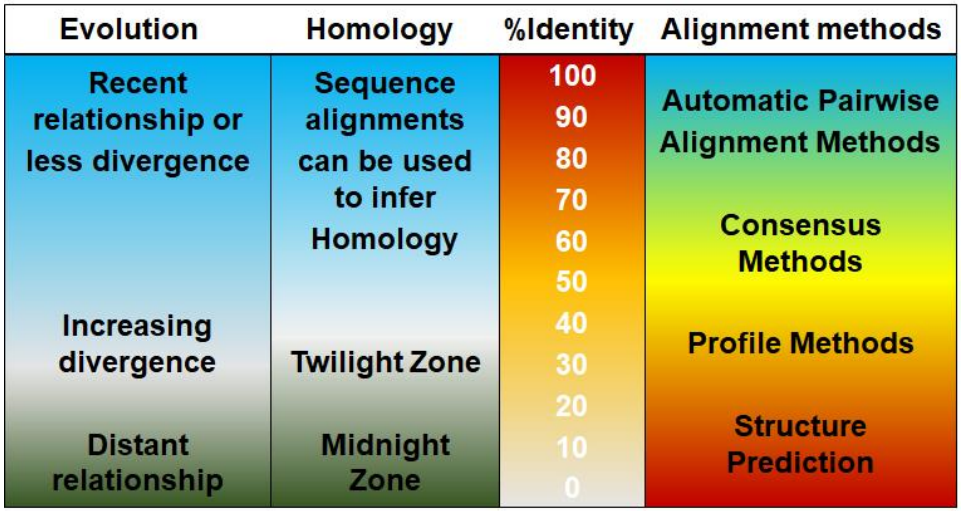






1. Búsqueda del templado

- Se busca por alineamiento con Estructuras conocidas:
  - Como primera aprox. BLAST-pdb
  - Sin templado NO hay modelo
  - Se pueden utilizar más de un templado



2. Alineamiento

- **En realidad ya alineamos al buscar el templado!!!**
  - La preguntas son:

¿Cuándo esta bien?

- Idealmente identidad > 40%
- Entre 25-40% zona gris
- alinear a lo largo de toda la secuencia “target”
- Pequeños errores en alineamiento grandes errores en el modelo

¿Cómo mejoro?

- Alineamiento múltiple:
  Más de un template puede ayudar a identificar zonas.
- Estructura secundaria:
  Evito GAPS en zonas de estructura secundaria
- Aminoácidos conservados:
  Conocimento de la proteína en estudio.

3. Construcción del modelo.

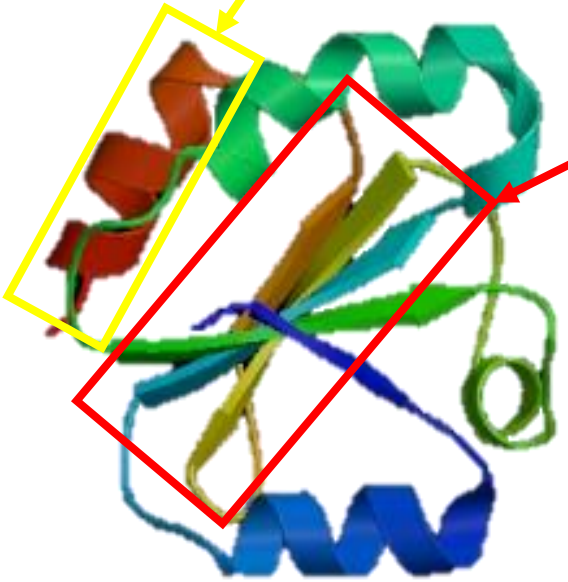
1) Modelo de la cadena carbonada **Generar coordenadas del modelo (x,y,z)**

2) Modelado de los loops

3) Cadenas Laterales

3a-Búsqueda de regiones conservadas estructuralmente

```
Target  ACDEFGHIKLMNPQRST--FGHQWERT-----TYREWYEG
Tmpl.1  ASDEYAHLRILDPQRSTVAYAYE--KSFAPPGSFKWEYEA
Tmpl.2  MCDEYAHIRLMNPERSTVAGGHQWERT----GSFKEWYAA
        HHHHHHHHHHHHHCCCCCCCCCCCCCCCCCBBBBBBBBB
```



- Corresponden a las regiones estructuralmente mas “estables” de la proteína (Por gral. el interior).
- Altamente conservadas, y por gral. sin GAPS.
- Usualmente corresponden a elementos de estructura secundaria.

Transferir la cadena carbonada:

- Zonas de estructura secundaria conservadas.
- Unir estas zonas:
- Cuando sea directo caso anterior
- Generar Loops

3. Construcción del modelo.

1) Modelo de la cadena carbonada

2) Modelado de los loops

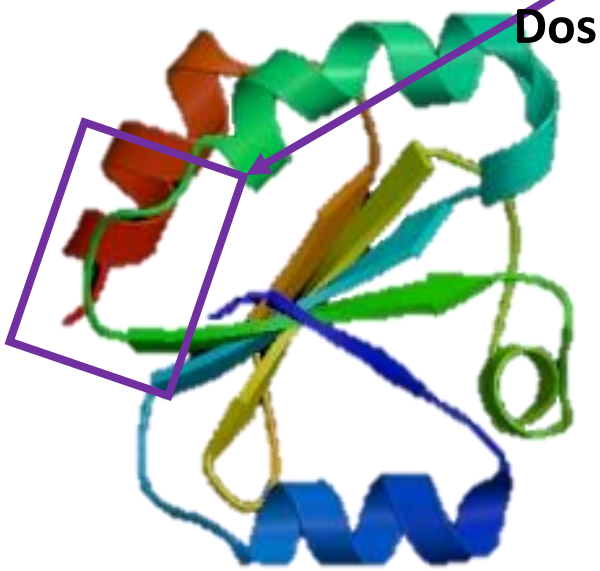
3) Cadenas Laterales

Problemas:

Son flexibles!
Cambian de conformación en solución!
Cambian de conformación al interactuar!

3a-Búsqueda de regiones conservadas estructuralmente

```
Target  ACDEFGHIKLMNPQRST--FGHQWERT-----TYREWYEG
Tmpl.1  ASDEYAHLRILDPQRSTVAYAYE--KSFAPPGSFKWEYEA
Tmpl.2  MCDEYAHIRLMNPERSTVAGGHQWERT----GSFKEWYAA
        HHHHHHHHHHHHHCCCCCCCCCCCCCCCCCBBBBBBBBB
```



Dos grandes posibilidades:

1. Dinámica Molécular/Sampleado conformacional

2. templates Librería de loops
   a) Loops extraídos de estructuras de alta resolución (< 2 A) por RX.
   b) Incluyen longitud, secuencia, coordenadas + estructura 2ria de los #aa anteriores y posteriores.
   c) Los aa anteriores y posteriores deben coincidir estructuralmente.

3. Construcción del modelo.

1) Modelo de la cadena carbonada

2) Modelado de los loops

3) Cadenas Laterales

3a-Búsqueda de regiones conservadas estructuralmente

```
Target  ACDEFGHIKLMNPQRST--FGHQWERT-----TYREWYEG
Tmpl.1  ASDEYAHLRILDPQRSTVAYAYE--KSFAPPGSFKWEYEA
Tmpl.2  MCDEYAHIRLMNPERSTVAGGHQWERT----GSFKEWYAA
        HHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCBBBBBBBBB
```

Aminoácidos idénticos: transferir todas las coordenadas del mismo.

Aminoácidos similares: transferir "backbone", reemplazar cadena lateral respetando las torsiones (χ).

Aminoácidos diferentes: transferir solo el "backbone" y evaluar rotámeros.

4. Evaluación

Los programas de evaluación verifican si el modelo tiene: “Todo aquello que una buena estructura debe tener”.

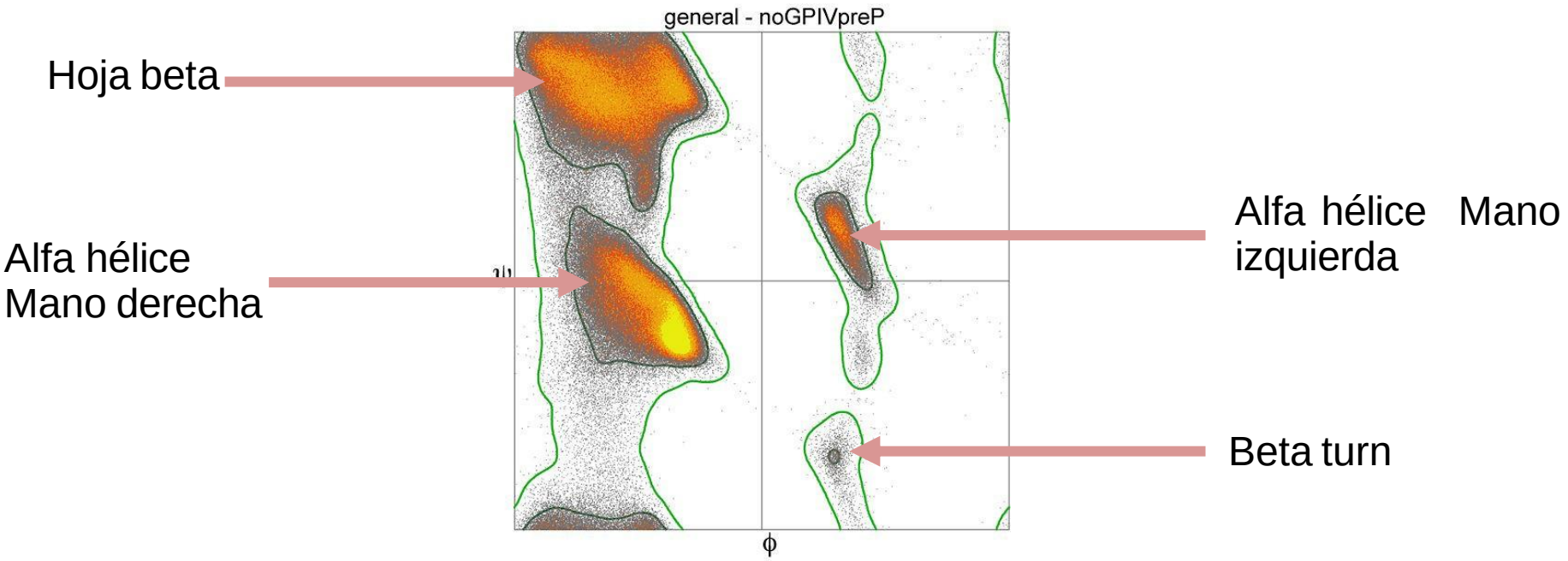
- Posición correcto de aa. Interior Exterior.
- Conformación más probable
- Distancias y ángulos
- Interacciones
- Etc….

Se pueden buscar estadisticamente:

PROCHECK

WHATCHECK

Una buena estructura posee el Mínimo numero de ángulos de torsión no permitidos (Ramachandran plot)



5. Refinamiento del modelo

Refinar el modelo “optimizando la geometria”

Se busca minimizar la función que relaciona la energia con las coordenadas atómicas:

- Potencial clásico ALL ATOM
  - AMBER, GROMACS, CHARMM

Objetivo:

- Eliminar superposiciones de átomos
- Ajustar distancias y ángulos de enlaces a valores típicos (Estereoquímica de la estructura)

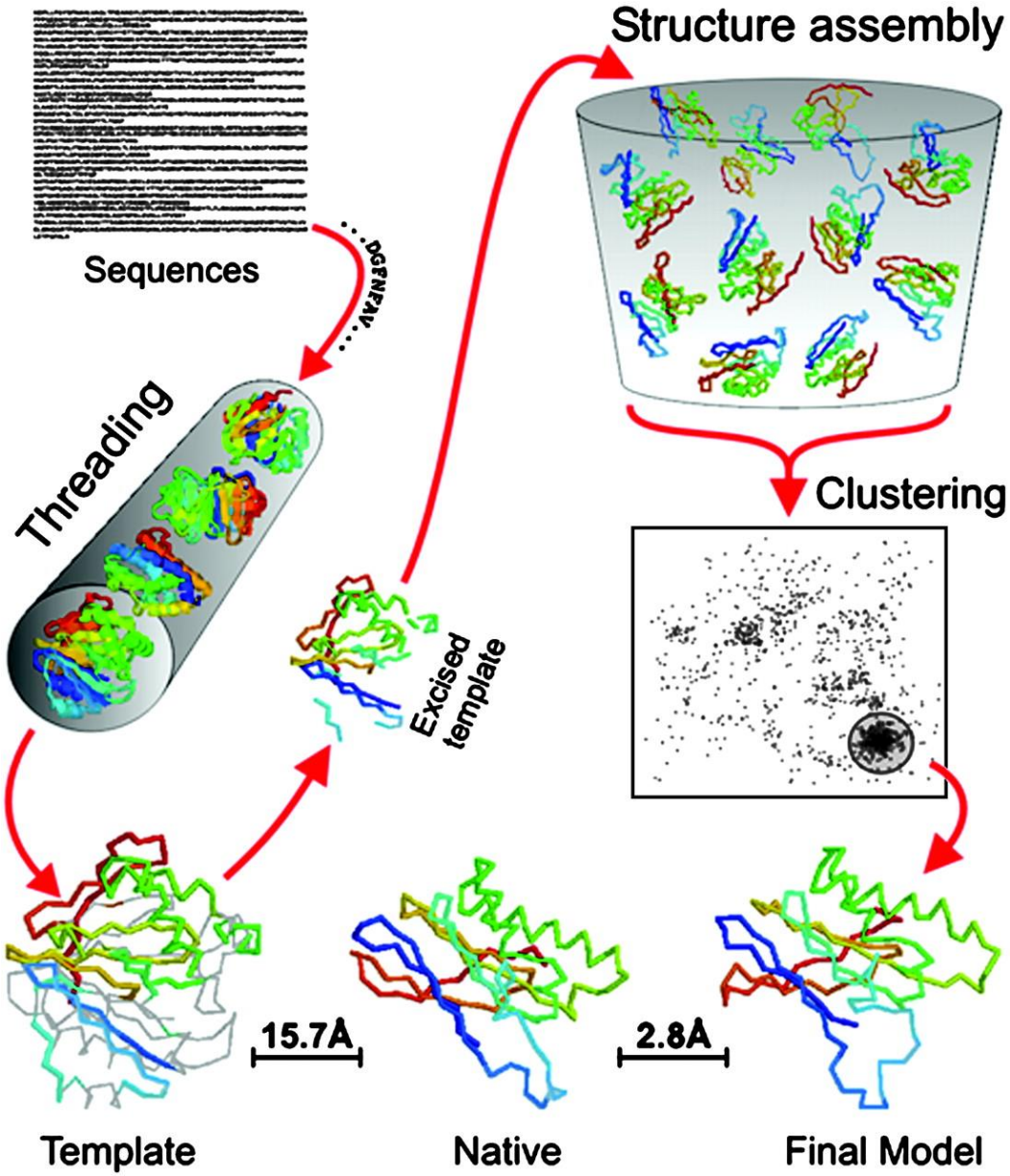
Aplicaciones

- Drug discovery
- Ligand-Protein interaction
- Protein function
- Protein-protein interaction

Programas…….

COMPOSER	P	www-cryst.bioc.cam.ac.uk
CONGEN	P	www.congenomics.com/congen/congen.html
CPH models	S	www.cbs.dtu.dk/services/CPHmodels/
DRAGON	P	www.nimr.mrc.ac.uk/~mathbio/a-aszodi/dragon.html
ICM	P	www.molsoft.com
InsightII	P	www.msi.com
MODELLER	P	guitar.rockefeller.edu/modeller/modeller.html
LOOK	P	www.mag.com
QUANTA	P	www.msi.com
SYBYL	P	www.tripos.com
SCWRL	P	www.cmpharm.ucsf.edu/~bower/scrwl/scrwl.html
SWISS-MOD	S	www.expasy.ch/swissmod
WHAT IF	P	www.sander.embl-heidelberg.de/whatif/

Métodos de Threading



Es una visión más física del problema pero combinan la información de los alineamientos con las estructures

Muy útiles para obtener modelos estructurales cuando la identidad con los modelos de estructura conocida es baja.

Threading o enebrar. SE ENEBRA una secuencia sobre otra.

Problema inverso de plegado:

> ¿Cuál de los plegamientos conocidos es "parecido" al plegamiento de la secuencia incognita?

Intentan adaptar la secuencia de la proteína a plegamientos (folds) de referencia

Emplea tipos de folds canónicos (SCOP, CATH).

Se evalua la "estabilidad" de cada uno de los folds teóricos en los que se ha plegado la proteína

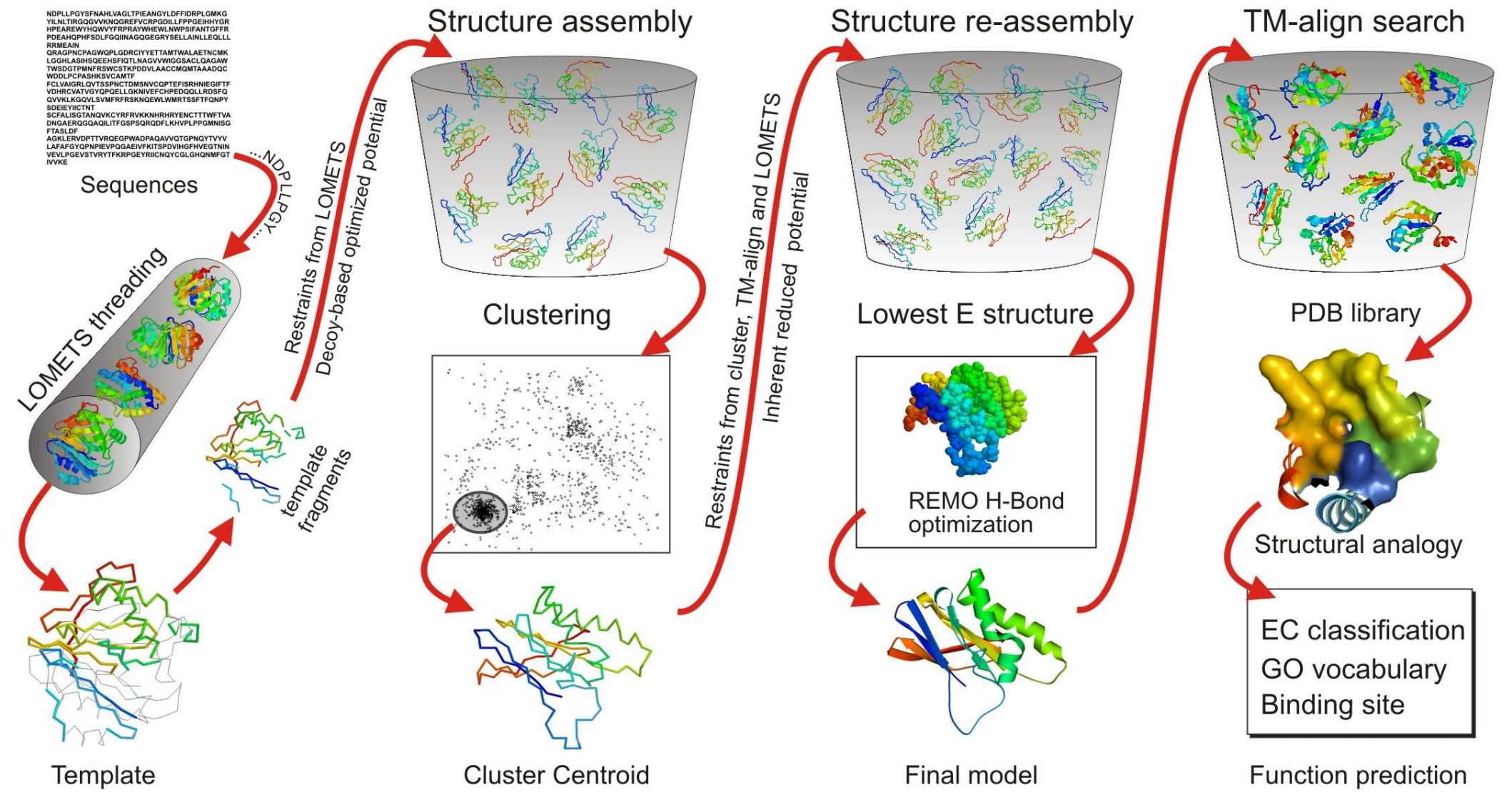
Se escoge el de mayor puntuación.

LOMETS Online

Meta Server Based Protein Fold Recognitions

LOMETS (Local Meta-Threading Server) is meta-threading method for template-based protein structure prediction. For a given sequence, it generates 3D models by collecting high-scoring structural templates from 11 locally-installed threading programs (CEthreader FFAS3D, HHpred, HHsearch, MUSTER, Neff-MUSTER, PPAS, PRC, PROSPECT2, SP3, and SparksX). A detailed description of the server can be seen in the About LOMETS page. Please post your questions and comments about LOMETS at the Service Syster Discussion Board.

Updating Notes: LOMETS has been updated to LOMETS2 with major updates including: (i) template liberary: while template libraries in former LOMETS are generated separately for different threading programs, which can result in inconsistent update and completenes of template structures, an unified and comprehensive template library is now created and weekly updated for all threading programs; (ii) MSA profile: a deep multiple sequence alignment (MSA) approach is developed to create deep sequence profiles for all templat proteins, which significantly improves the accuracy of almost all the profile- and HHM-based threading alignments; (iii) threading programs: more than half of the old threading programs were renewed and/or replaced by the state-of-the-art methods, including the cutting edge contact-based threading algorithms; (iv) function annotations: completely redesigned the output page which contains now structure-based function annotations derived from threading templates.



https://zhanglab.ccmb.med.umich.edu/LOMETS/

Phyre2

Protein Homology/analogY Recognition Engine V 2.0

Subscribe to Phyre at Google Groups
Email:
Subscribe
Visit Phyre at Google Groups

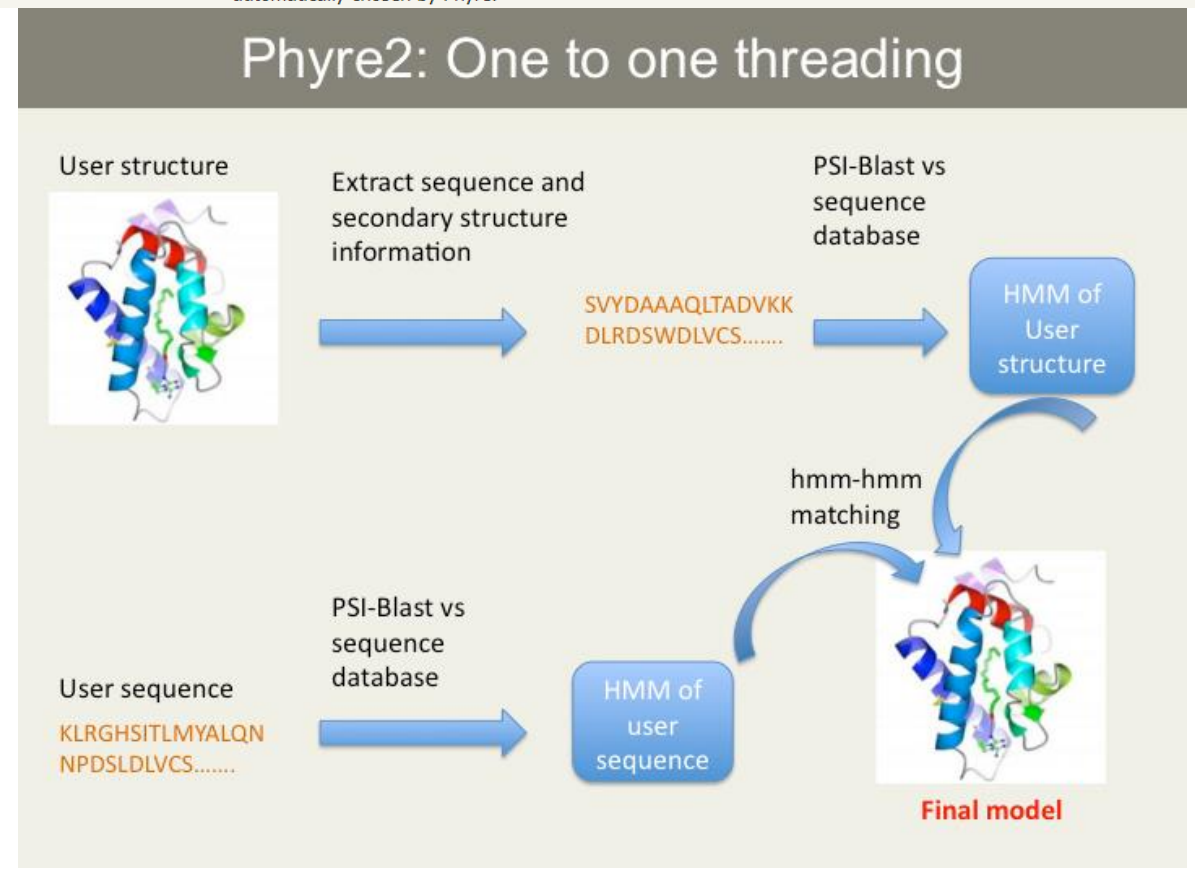
Help

Home | What's New | Video Tutorials | Interpreting Results | Advanced features | Slides for teaching | FAQ | About

One 2 One threading

What it is

Sometimes you may have a protein sequence and you want to model it on a specific template of your choosing. Perhaps you have a newly solved structure that's not in the Phyre database or you have some biological information that indicates your chosen template would produce a more accurate model than the one(s) automatically chosen by Phyre.



Se trabaja sobre una secuencia incognita por similitud de secuencia

Se genera un profile 5 iteraciones de PSI-BLAST se combinan en un solo alineamiento.

La estructura secundaria se predice con: PSI-PRED, SSPRO, JNET Disopred

Profile y la estrutura son alineados usando algoritmos profile-profile

Del alineamiento y ranking, se eligen los 10 mejores

Se genera un modelo 3D como Homology Modeling.

Éxito en 15% a 25% de identidad

Programas de Threading

- 3D-pssm (ICNET) now PHYRE. Based on sequence profiles, solvatation potentials and secondary structure.
- TOPITS (PredictProtein server) (EMBL). Based on coincidence of secondary structure and accesibility.
- UCLA-DOE Structure Prediction Server (UCLA). Executes various *threading* programs and report a consensus.
- 123D+ Combines substitution matrix, secondary structure prediction, and contact capacity potentials.
- SAM/HMM (UCSC). Basen on Markov models of alignments of crystalized proteins.
- FAS (Burnham Institute). Based on profile-profile matching algorithms of the query sequence with sequences from clustered PDB database.
- PSIPRED-GenThreader (Brunel)
- THREADER2 (Warwick). Based on solvatation potentials and contacts obtained from crystalized proteins.
- ProFIT CAME (Salzburg)

Métodos ab-initio

Plegar la proteína con potenciales basados en la *FÍSICA* sin comparar con una *ESTRUCTURA CONOCIDA*.

Primeros principios:

Son transferibles: No dependen de un caso particular.
Varían en cuanta FÍSICA usan:

Ab-initio reales: AMBER, CHARMM (All Atom)
- Lentos
- muchos mínimos cercanos
- Dificil sampleo

Ab-initio estadísticos: Grano grueso
- Rápidos
- Pocos mínimos
- Fácil sampleo

- Intentan plegar proteínas pequeñas a partir de potenciales estadísticos, sin recurrir “a priori” al conocimiento previo del plegamiento de proteínas similares.
- Emplean métodos muy sencillos de muestreo del espacio conformacional de las proteínas.
- Muy poco precisos.
- Aplicables solo a proteínas pequeñas.

Docking chemicals into **protein** cavities

Docking is a structure-based technique which attempts to find the “best” match, between two molecules.

In the field of molecular modeling**, docking is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex (pose).**

• Knowledge of the preferred orientation in turn may be used to predict the **strength of association** or **binding affinity** between two molecules using for example **scoring functions**.

During the course of the docking process, the ligand and the protein adjust their conformation to achieve an overall "best-fit" and this kind of conformational adjustment resulting in the overall binding is referred to as **"induced-fit“.**

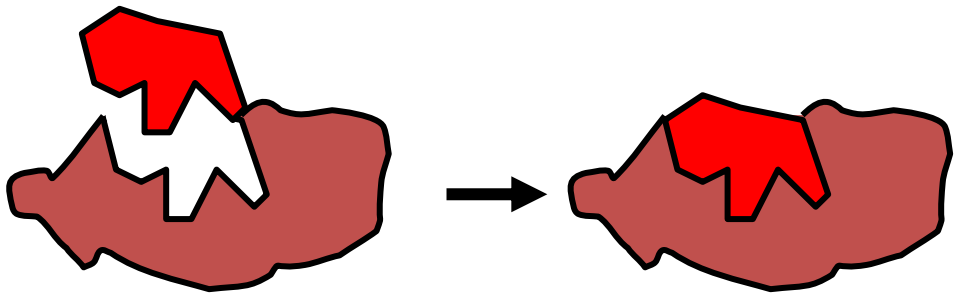
SHAPE COMPLEMENTARITY:- One approach uses a matching technique that **describes the protein and the ligand as complementary surfaces.**
SIMULATION:- The second approach simulates the actual docking process in which the **ligand-protein pairwise interaction energies are calculated**. Both approaches have significant advantages as well as some limitations.

Docking: the lock & key principle

Geometrical Complementarity

substrate

enzyme

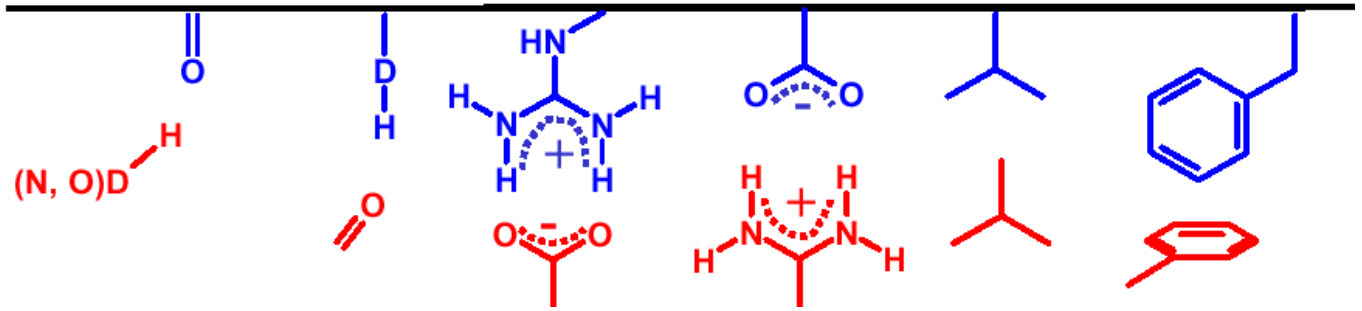


One can think of molecular docking as a problem of *"lock-and-key"*, in which one wants to find the correct relative orientation of the *"key"* which will open up the *"lock"* (where on the surface of the lock is the key hole, which direction to turn the key after it is inserted, etc.)

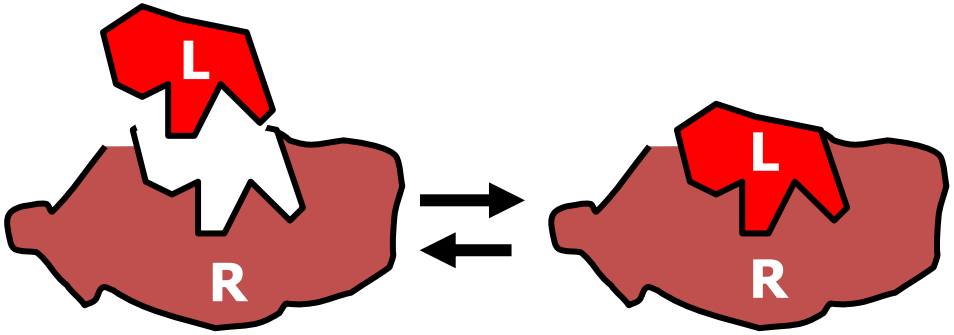
non-covalent inter-molecular interactions

receptor

ligand



Thermodynamics of association



- Hydrophilic and hydrophobic interactions.
- Hydrogen bond donated.
- Hydrogen bond accepted.
- Ligand orientation with best complementarity score.
- Binding affinities.
- Ionic interaction.
- Aromatic Interaction.
- Vander Waals' forces
- Electrostatic forces
- Free energies.

Association constant (equilibrium)

$$K_A = \frac{[LR]}{[L][R]}$$

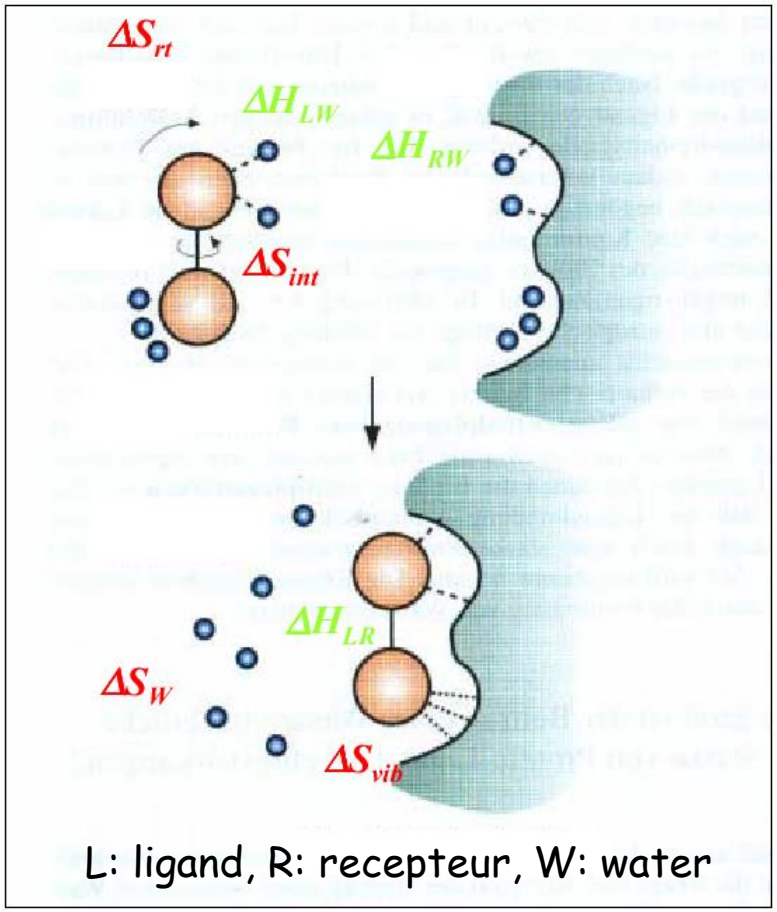
no unit

$$K_A = \exp - \frac{\Delta G^0}{RT}$$

ΔG^0; RT in J/mole

K_D (M)	$\Delta G°$ (kcal/mol) at 298K
10^{-15}	-20.4
10^{-12}	-16.4
10^{-9}	-12.3
10^{-6}	-8.2
10^{-3}	-4.1

Thermodynamics of association



L: ligand, R: recepteur, W: water

Free energy $\Delta G = \Delta H - T\Delta S$

Enthalpy H ~ sum of interactions

Entropy S ~ order

What is gained/lost upon binding?

TYPES OF INTERACTIONS

- **Electrostatic forces** - Forces with electrostatic origin are due to the charges residing in the matter.
- **Electrodynamics forces** - The most widely known is probably the van der Waals interaction.
- **Steric forces** - These are caused by entropy. For example, in cases where entropy is limited, there may be forces to minimize the free energy of the system.
- **Solvent-related forces** – These are due to the structural changes of the solvent. These structural changes are generated, when ions, colloids, proteins etc, are added into the structure of solvent. The most commonly are Hydrogen bond and hydrophobic interactions

Targeting Protein by chemicals

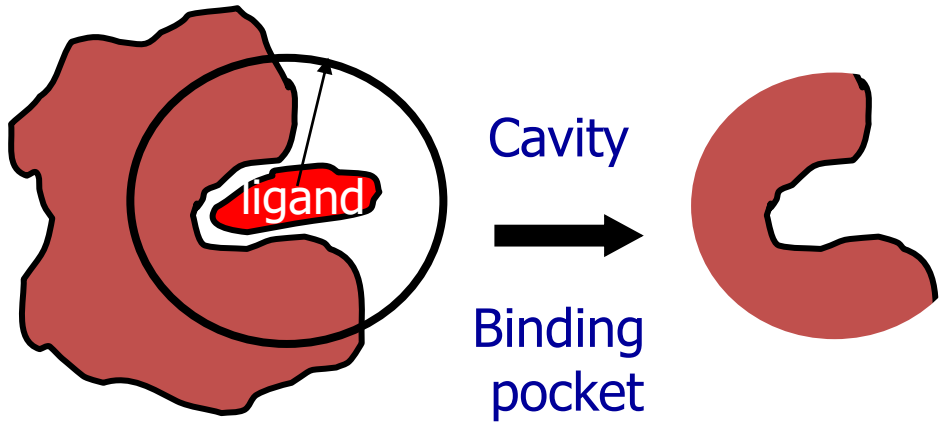
protein function

- **biochemical function** = interaction with other molecules
- **biological function** = consequence of these interactions

"druggable" binding site

cavity

- depth
- volume
- lipophilic surface



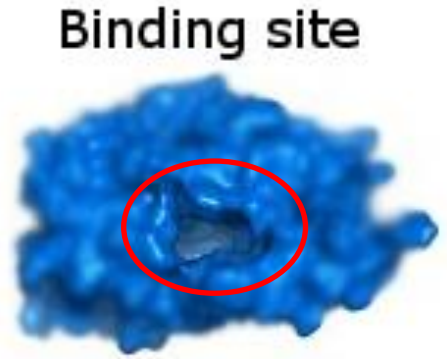
about 6000 « druggable » sites in PDB protein X-ray structures

http://bioinfo-pharma.u-strasbg.fr/scPDB/

Pose vs. binding site

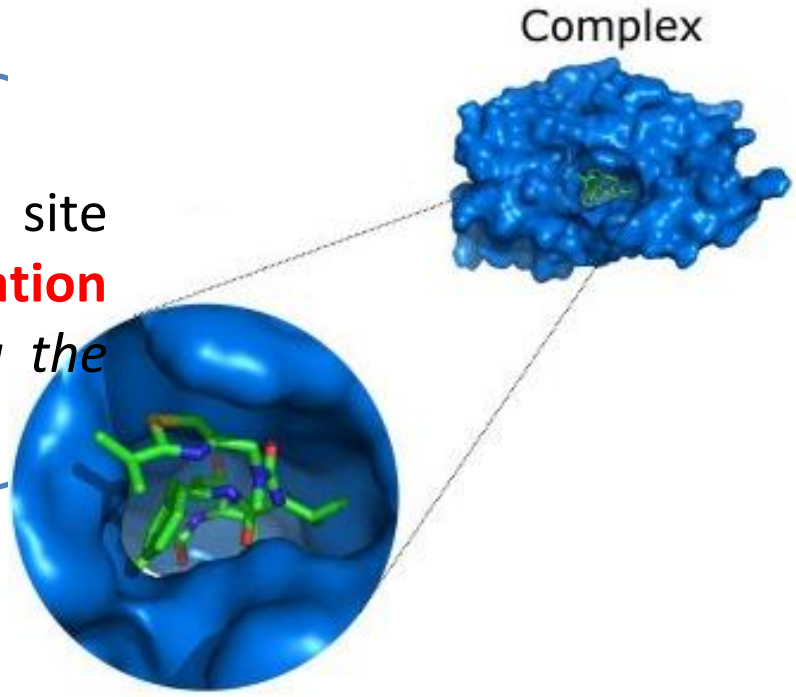
Binding site (or “active site”)

the part of the protein where the ligand binds generally a cavity on the protein surface can be identified by looking at the crystal structure of the protein bound with a known inhibitor



Pose (or “binding mode”)

The *geometry* of the ligand in the binding site
Geometry = **location, orientation and conformation**
*Protein-ligand docking is **not** about identifying the binding site*



STEPS OF DOCKING

Step 1: Start with Crystal Co-ordinates with target receptor

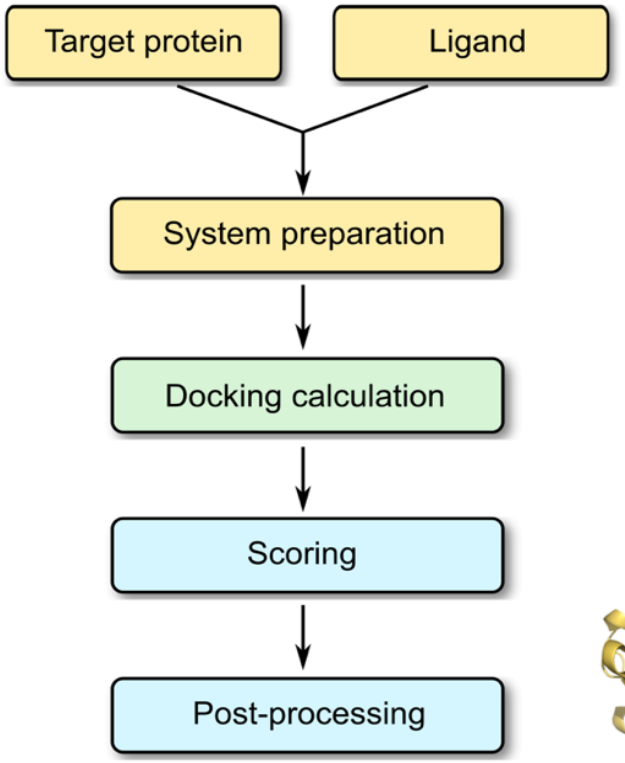
Step 2: Generate molecular surface for receptors

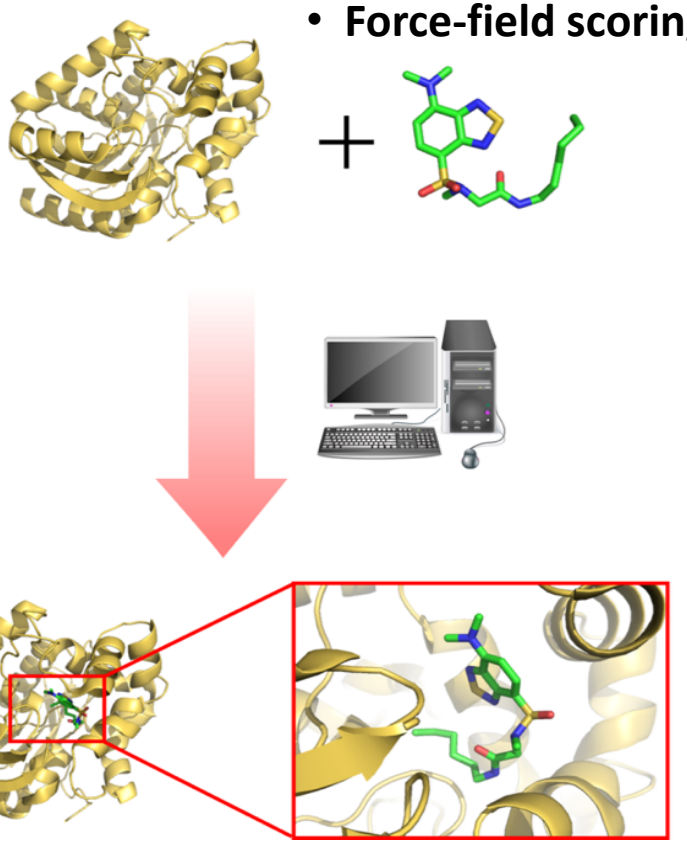
Step 3: Generate spheres to fill the active site of the receptor: The spheres become potential locations for ligand atoms.

Step 4: Sphere centres are then matched with the ligand atoms, to determine possible orientations for the ligand.

Step 5: Find the top scoring or the best ranking:

- **Shape scoring,**
- **Electrostatic scoring and**
- **Force-field scoring.**





TYPES OF DOCKING

Rigid Docking (Lock and Key)

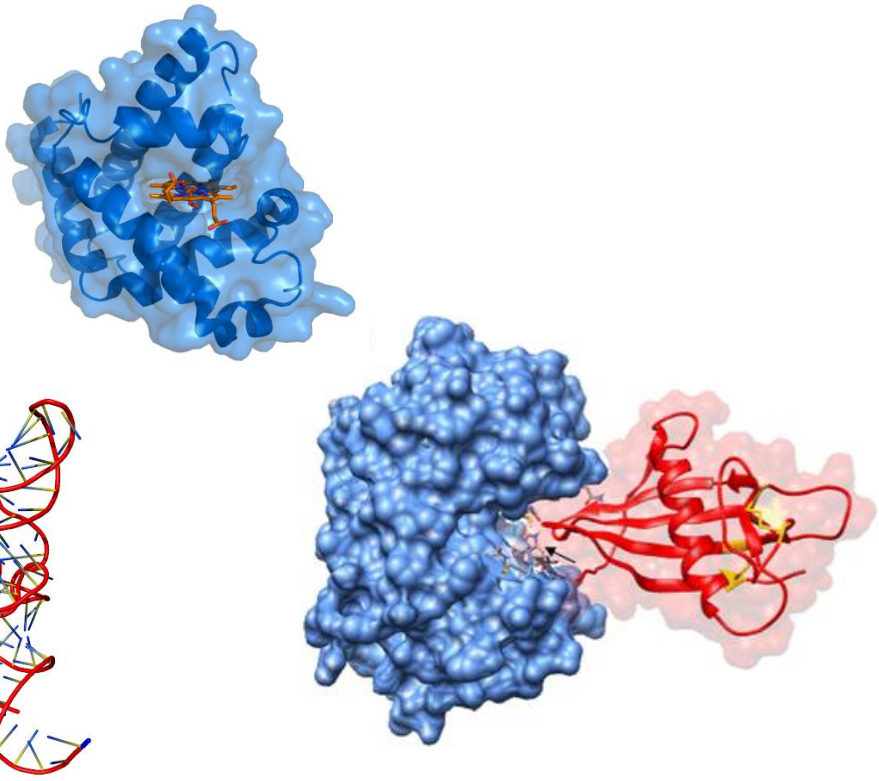
In rigid docking, the internal geometry of both the receptor and ligand are treated as rigid.

Flexible Docking (**Induced fit**)

An enumeration on the rotations of one of the molecules (usually smaller one) is performed. Every rotation the energy is calculated; later the most optimum pose is selected.

Docking can be between….

- Protein - Ligand
- Protein – Protein
- Protein – Nucleotide

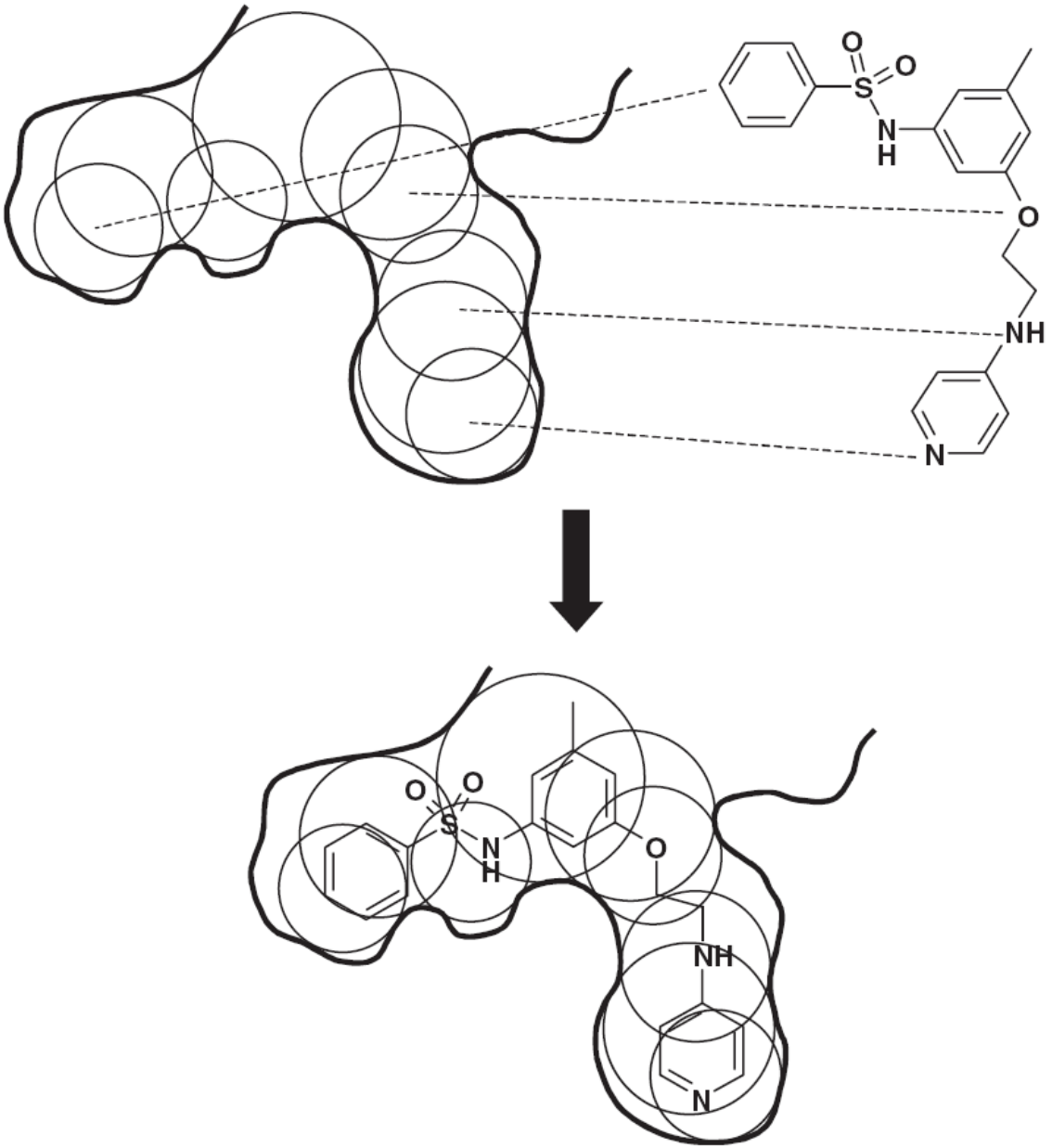


THE DOCK ALGORITHM – RIGID DOCKING

The DOCK algorithm developed by Kuntz and co-workers is generally considered one of the major advances in protein–ligand docking [Kuntz et al., *JMB,* **1982**, *161*, 269]

The earliest version of the DOCK algorithm only considered rigid body docking and was designed to identify molecules with a high degree of shape complementarity to the protein binding site.

The first stage of the DOCK method involves the construction of a “negative image” of the binding site consisting of a series of overlapping spheres of varying radii, derived from the molecular surface of the protein

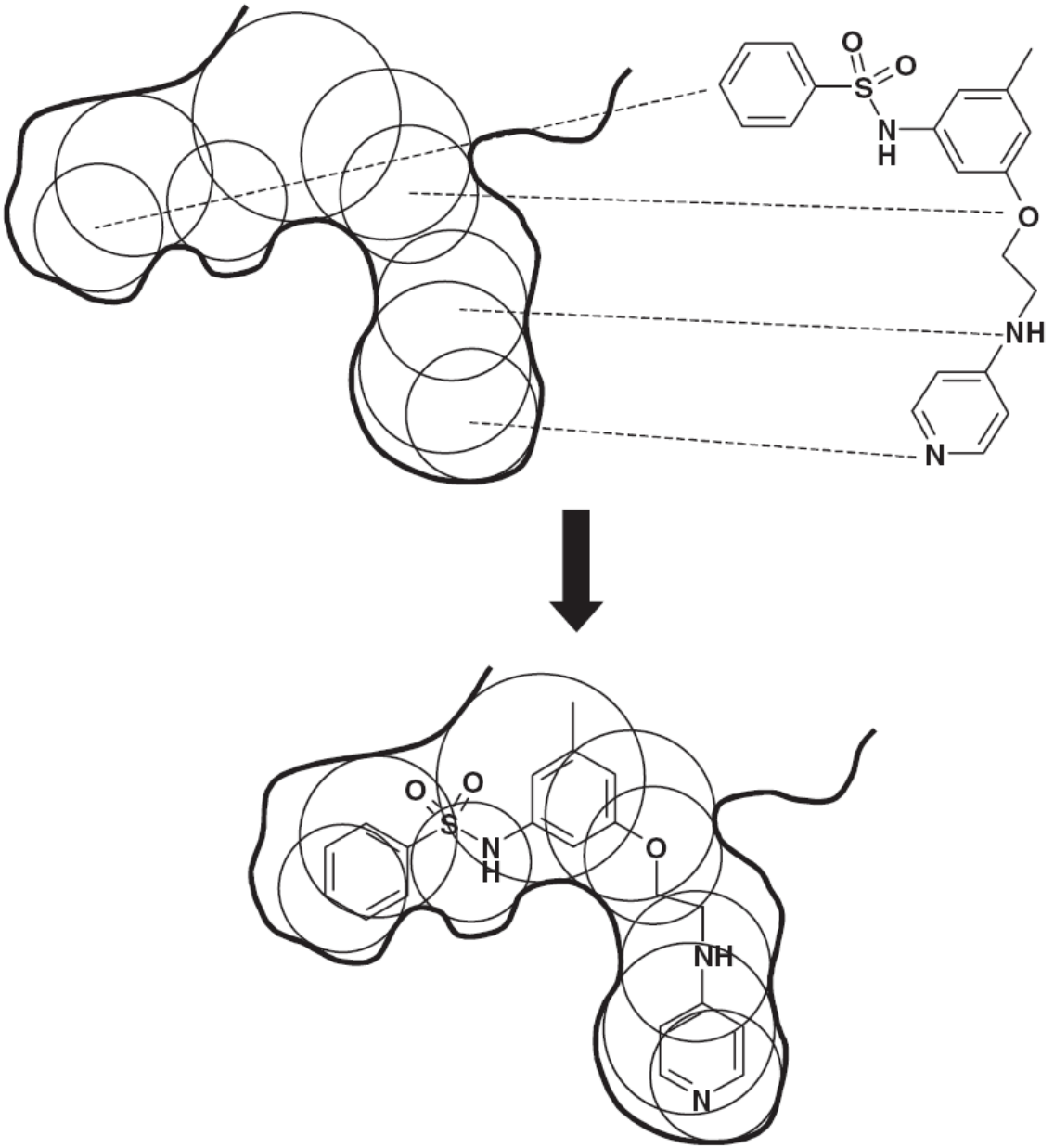


THE DOCK ALGORITHM – RIGID DOCKING

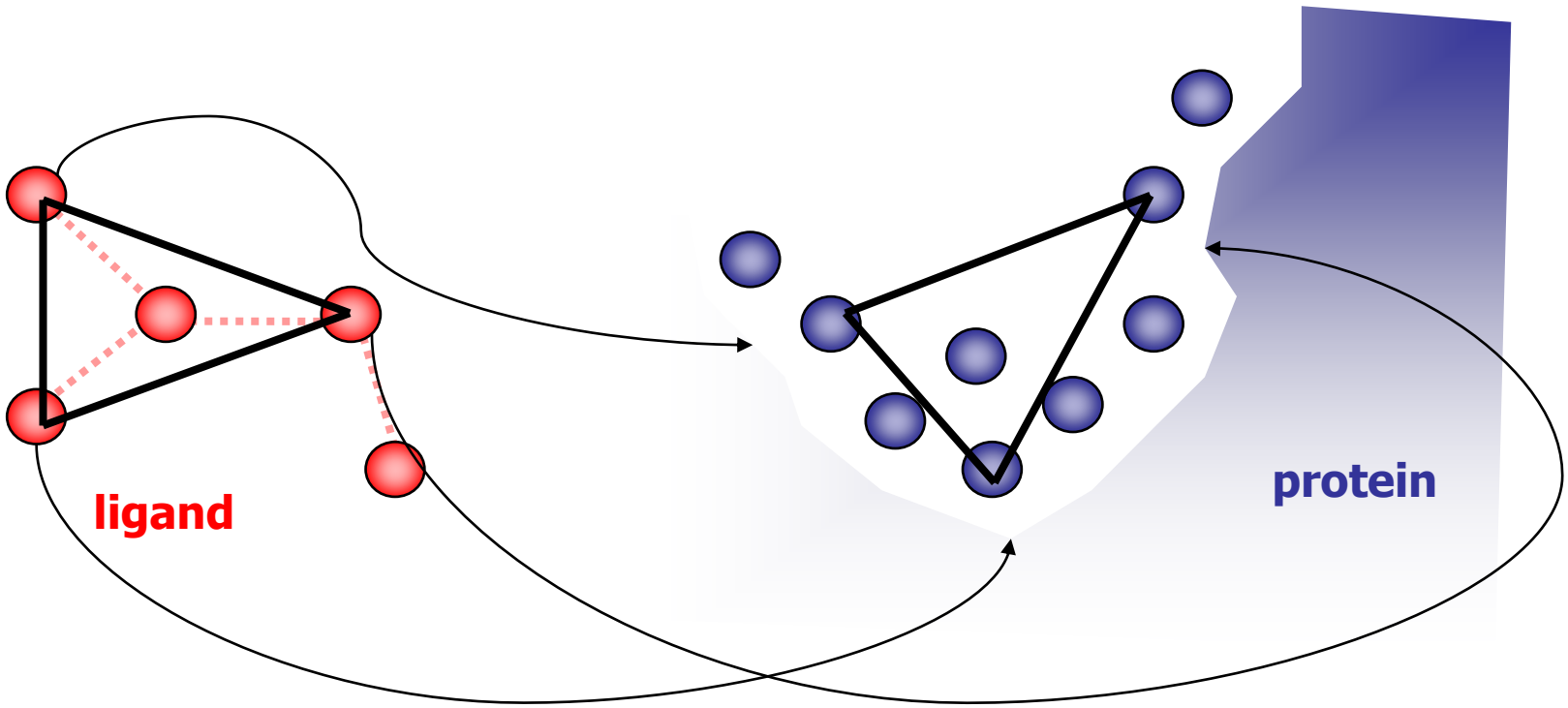
Ligand atoms are then matched to the sphere centres so that the distances between the atoms equal the distances between the corresponding sphere centres, within some tolerance.

The ligand conformation is then oriented into the binding site. After checking to ensure that there are no unacceptable steric interactions, it is then scored.

New orientations are produced by generating new sets of matching ligand atoms and sphere centres. The procedure continues until all possible matches have been considered.



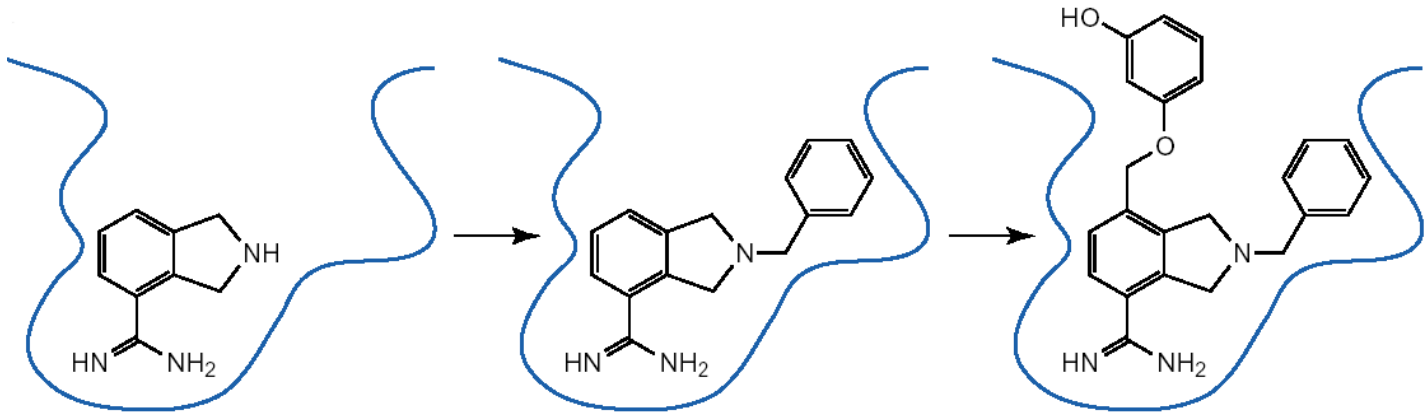
Geometric matching of triangles



Ligand flexibility in geometric algorithms

Incremental construction

(FlexX, DOCK, Surflex)



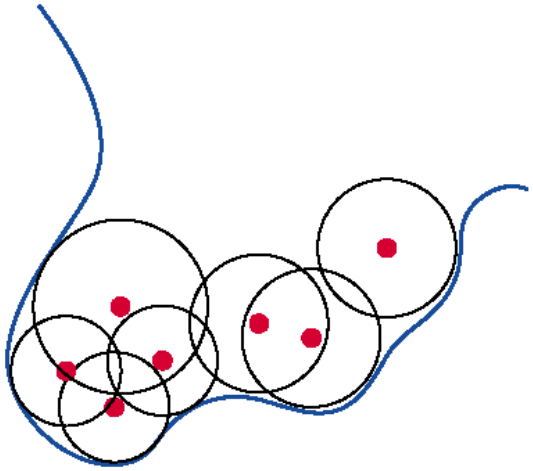
Library of conformers

- rigid docking of all conformers (DOCK, MOE)
- conformer generation usually not included in the docking tool
- in MOE-Dock, the conformational search is coupled to docking, using a library of preferred torsion values for rotatable bonds

Examples of geometric algorithm

DOCK

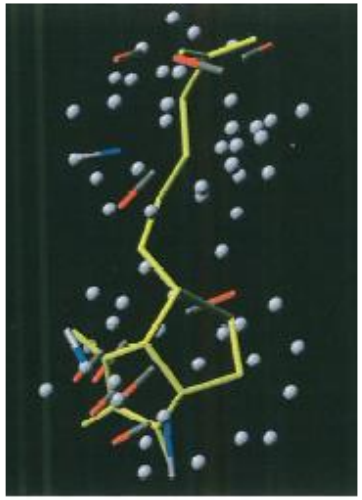
Kuntz, I.D et al. (1982) J. Mol. Biol



- Protein cavity filled with overlapping spheres (variable radius).
- Feature points: sphere center colored according to physico-chemical properties

Surflex, MOE

Jain A (2003) J Med Chem

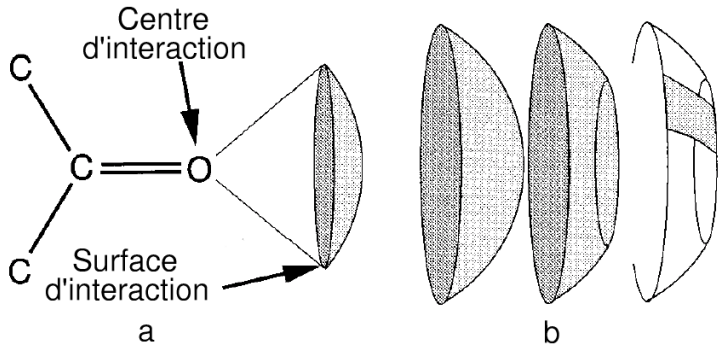


Cluster of probes

- steric (apolar)
- Polar
  (NH, CO in Surflex, polar in MOE)

Flexx

M. Rarey et al. (1996) J. Comp.-Aid. Mol. Design



Interaction centers and interaction surfaces identified on both receptor (a) and ligand (b)

- H bond
- Salt bridges
- Aromatics
- methyl-aromatics
- amide-aromatics

FLEXIBLE OR SEMI-FLEXIBLE DOCKING

Flexible docking is the most common form of docking today

Conformations of each molecule are generated on-the-fly by the search algorithm during the docking process - The algorithm can avoid considering conformations that do not fit.

Exhaustive (systematic) searching computationally too expensive as the search space is very large

One common approach is to use stochastic search methods

These don't guarantee optimum solution, but good solution within reasonable length of time

Stochastic means that they incorporate a degree of randomness

Such algorithms include genetic algorithms (GOLD), simulated annealing (AutoDock)

An alternative is to use incremental construction methods

These construct conformations of the ligand within the binding site in a series of stages

First one or more "base fragments" are identified which are docked into the binding site

The orientations of the base fragment then act as anchors for a systematic conformational analysis of the remainder of the ligand

Example: FlexX

HANDLING PROTEIN CONFORMATIONS

Most docking software treats the protein as rigid

Rigid Receptor Approximation

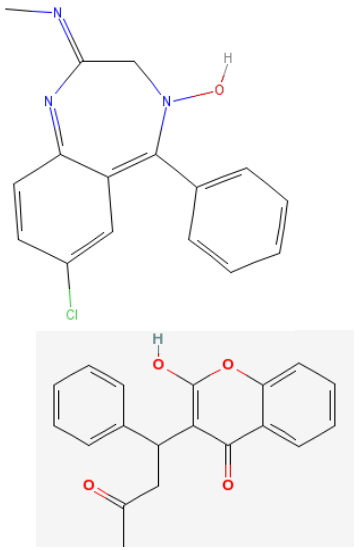
This approximation may be invalid for a particular protein-ligand complex as...

the protein may deform slightly to accommodate different ligands (ligand-induced fit)

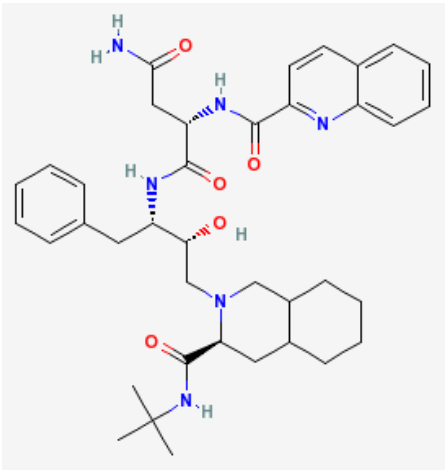
protein side chains in the active site may adopt different conformations

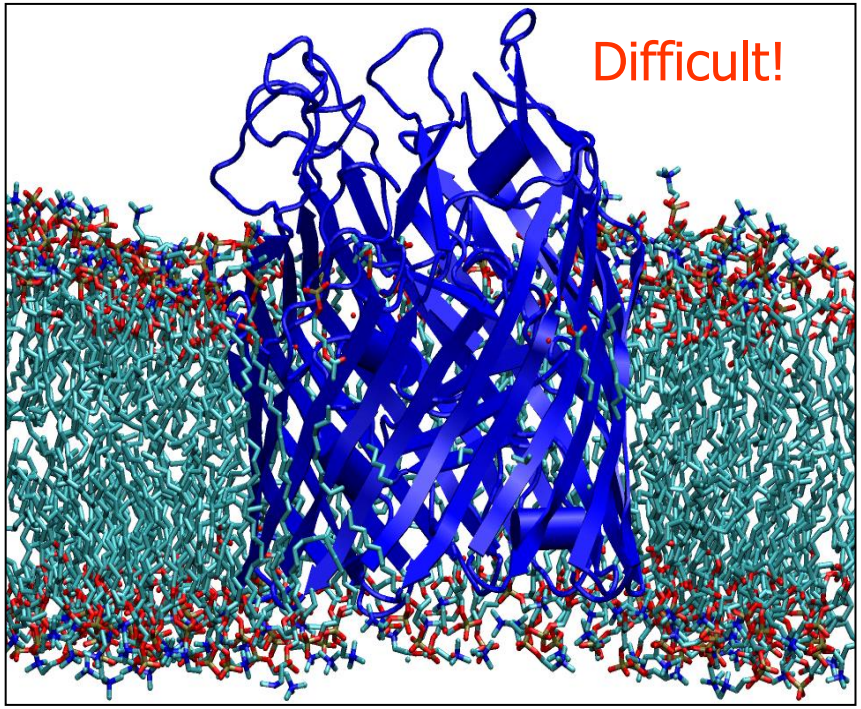
- Some docking programs allow protein side-chain flexibility
  - For example, selected side chains are allowed to undergo torsional rotation around acyclic bonds
  - Increases the search space
- Larger protein movements can only be handled by separate dockings to different protein conformations
  - Ensemble docking (e.g. GOLD 5.0)

Semi-flexible docking



feasible





Difficult!

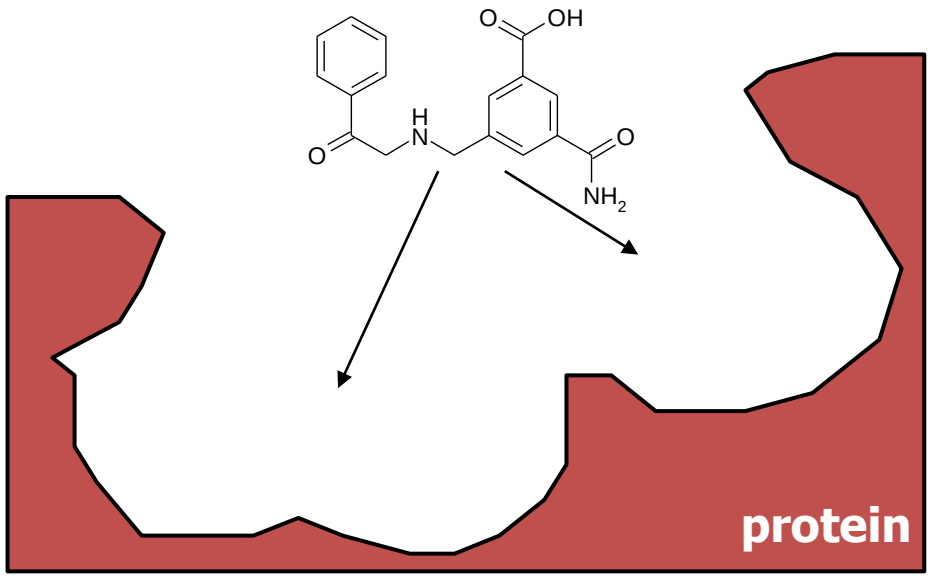
< 25-30	number of rotatable bonds	≥ 3 per amino acids
➔ exhaustive	conformational search	partial
seconds to hours	cpu time	huge!

search for the best pose of *flexible* ligand into *rigid* protein site

Pose = Orientation & conformation

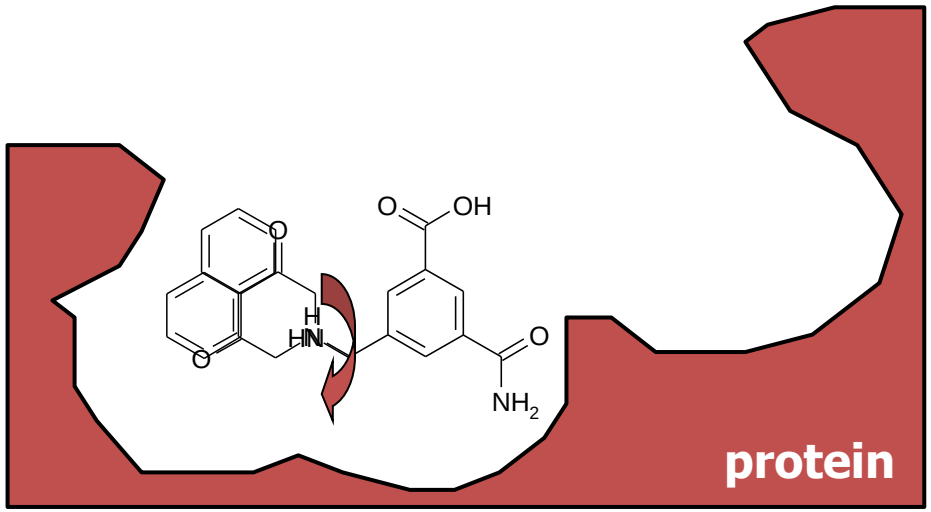
Orientation

- Position of the ligand in the protein
- Rigid body motions
- Translations + rotations of the whole molecule



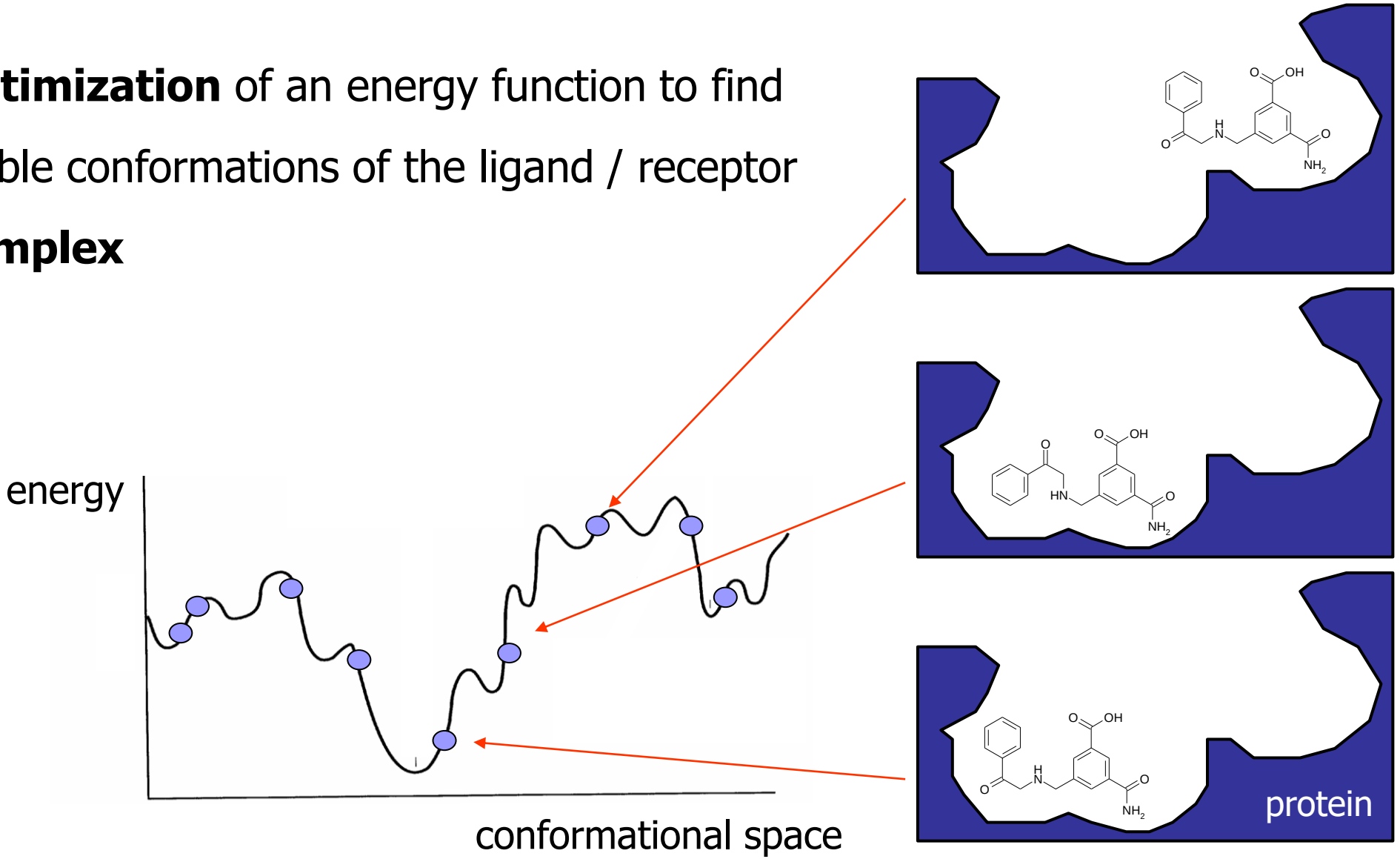
Conformation

- 3D structure of the molecule
- Molecular flexiblity

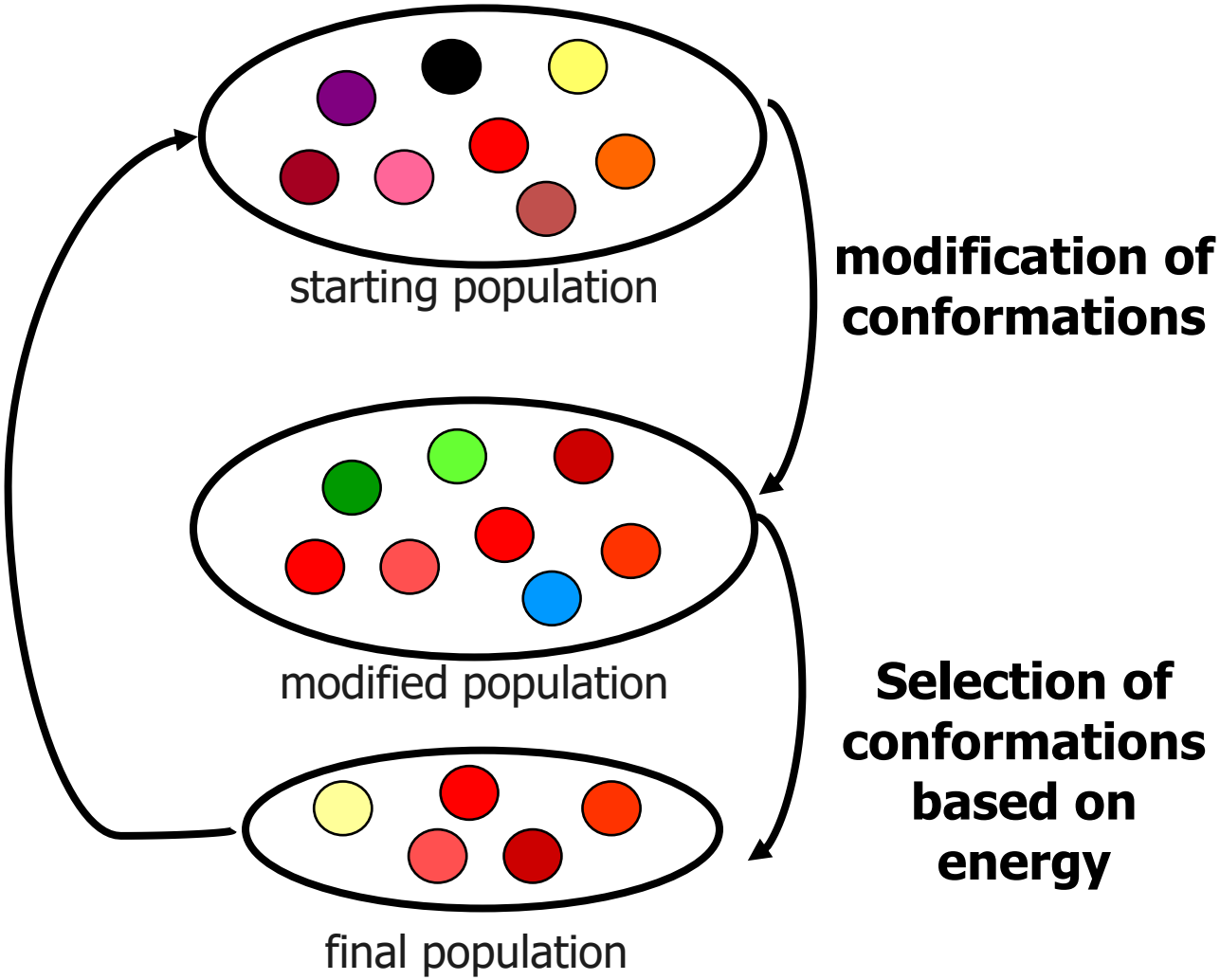


energy-based algorithms

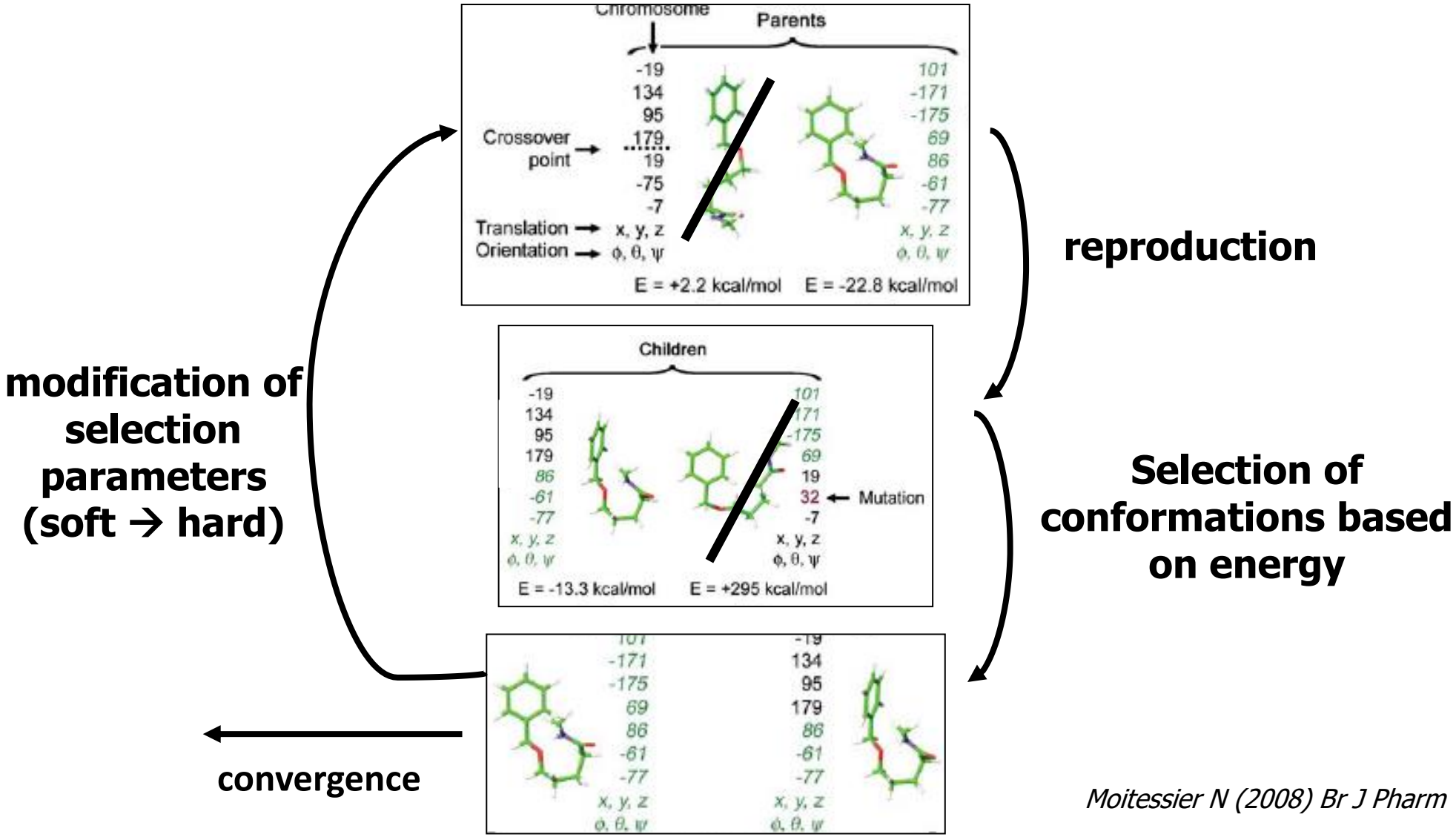
Optimization of an energy function to find stable conformations of the ligand / receptor **complex**



iterative generation of populations of conformers

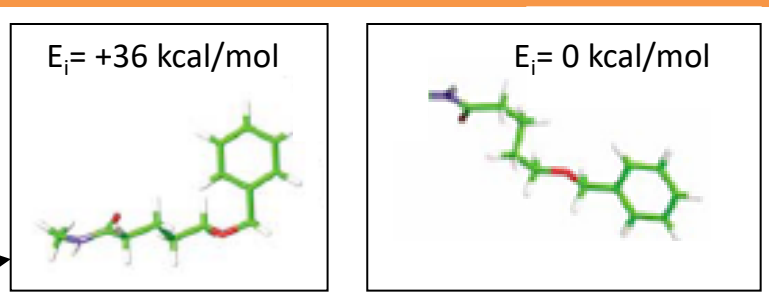


Genetic algorithm (gold, autodock)

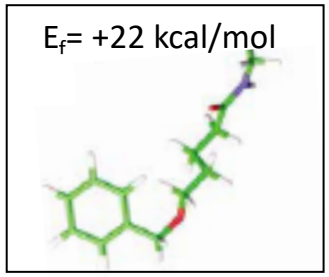


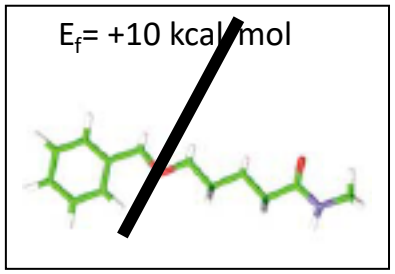
Moitessier N (2008) Br J Pharm

Monte Carlo (ICM, RosettaLigand)



Bond rotation
Tranlation
Rotation





Selection of conformations based on energy

if $E_f < E_i$ accept

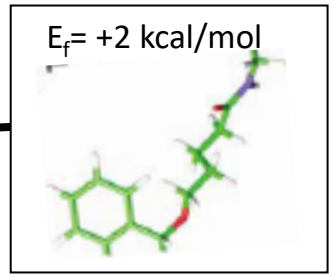
if $E_f > E_i$ metropolis

$$exp\left(-\frac{Ef - Ei}{K_B T}\right) > z$$

accept

x cooling cycles (T decrease)

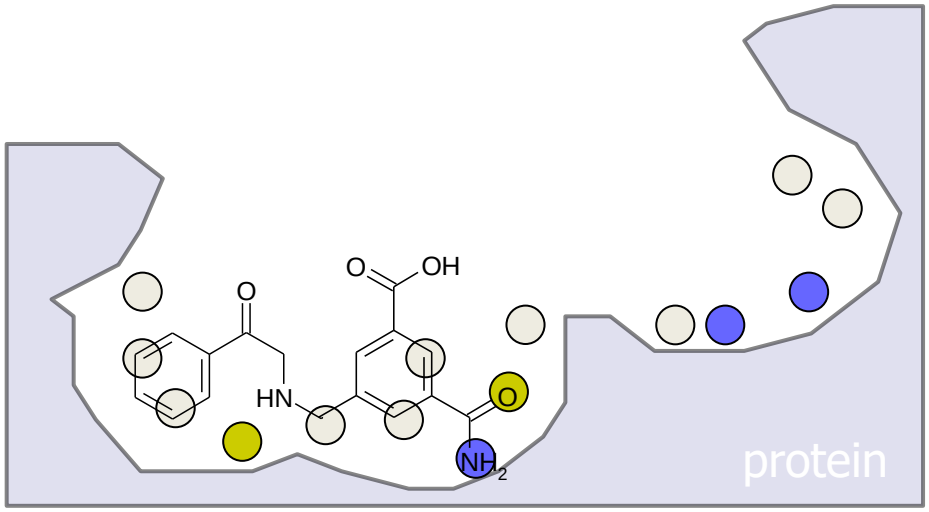
…
n iterations



geometry-based *vs* energy-based algorithms

geometry-based

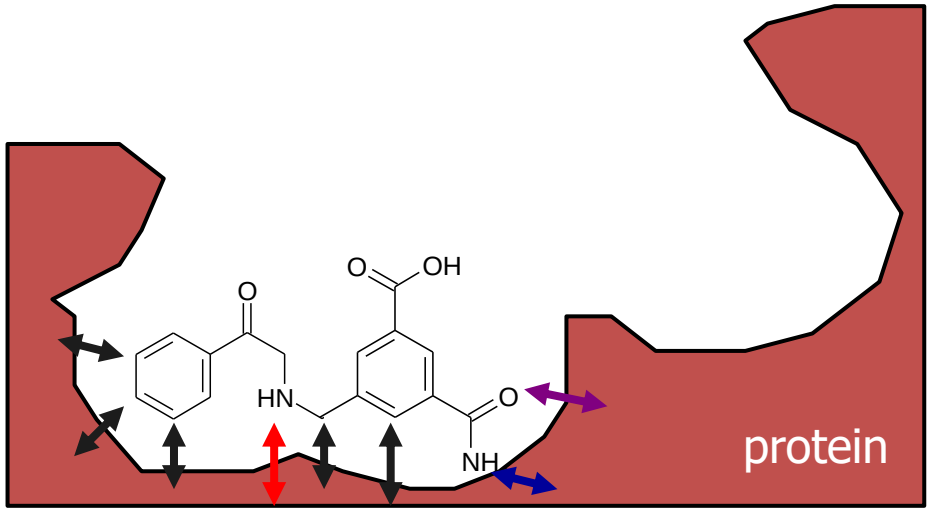
determinist methods



- Protein: feature points **in the cavity**
- Ligand: atoms or feature points
- Docking: translations + rotations of **rigid** ligand to **superpose matching points**

energy-based

stochastic methods

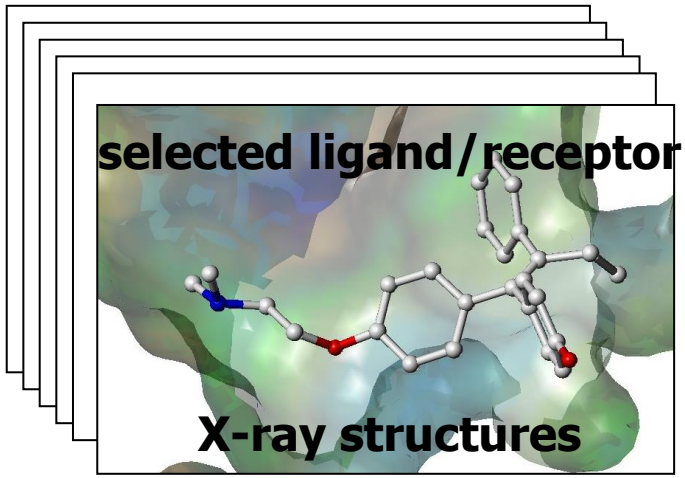


- Protein : surface atoms / feature points
- Ligand : atoms or feature points
- Docking: modification of ligand **position/conformation** to **optimize** an "**energy" function** that describes **molecular interactions**

Scoring ligand/receptor interaction

empirical *vs* force field scoring functions (SF)

Empirical SF



non-bonded interactions

distance between ligand - receptor pairs of atoms

experimental ΔG

empirical SF

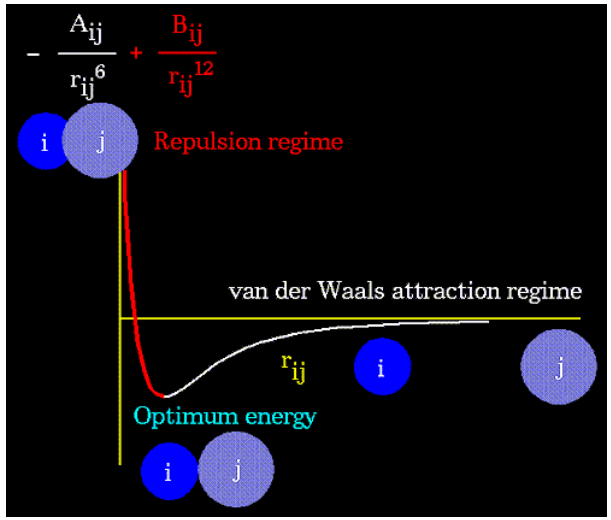
knowledge-based SF

Force Field SF

$S = E_{ligand} + E_{complexe}$

E_{ligand} = internal energy

$$E_{complexe} = \sum_{i}^{lig}\sum_{j}^{rec}\left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + 332.0\frac{q_i q_j}{\varepsilon r_{ij}}\right]$$



Prediction of free energy by Empirical SF

$$S \approx \Delta G = \sum k_i * F_i$$

F_i : **function** which describes protein – ligand interaction (H-bond, salt bridge..)

computed using geometry predicted by docking

k_i : **constant** adjusted using the **training set**

MOE *affinity dG* SF

$$dG_{bind} = k_{hb}\sum_{hb} f_{hb} + k_{ml}\sum_{ml} f_{ml} + k_{hh}\sum_{hh} f_{hh} + k_{hp}\sum_{hp} f_{hp} + k_{aa}\sum_{aa} f_{aa}$$

k: constant

f: count of atomic contact

hb: H-bond donor – H-bond acceptor
ml: ionic metal - ligand
hh: hydrophobic atom – hydrophobic atoms
hp: hydrophic atom – polar atom
aa: any atom - any atom

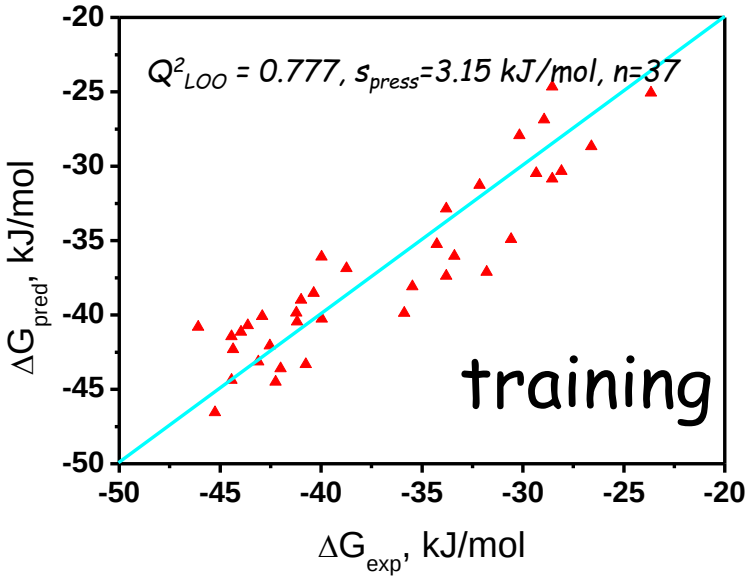
Example of empirical SF performance

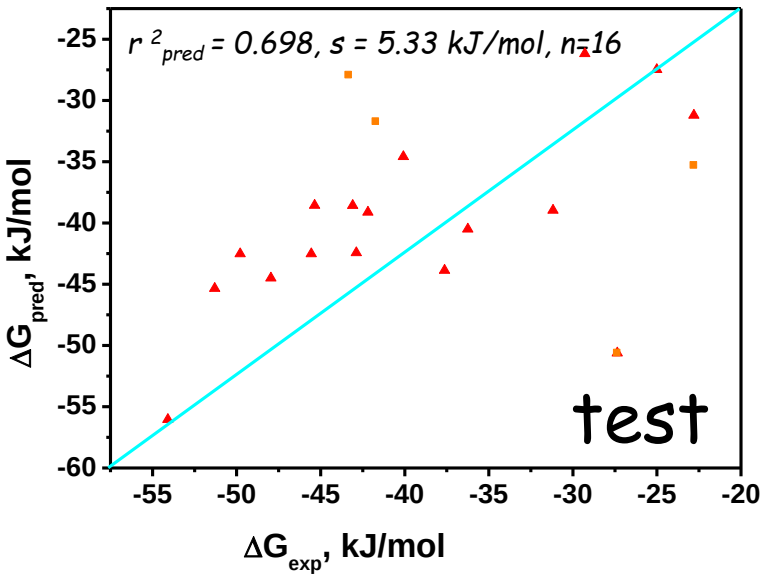
Böhm (1994)

H-bond *ionic* *lipophilic* *rotation*

$$\Delta G_{bind} = \Delta G_0 + \Delta G_{hb} \sum_{hb} f(\Delta R, \Delta\alpha) + \Delta G_{ionic} \sum_{ionic} f(\Delta R) + \Delta G_{lipo} \sum_{hb} |A_{lipo}| + \Delta G_{rot} NROT$$

+5.4 kJ/mol -4.7 kJ/mol -8.3 kJ/mol -0.17 kJ/mol.Å^2 +1.4 kJ/mol/rot





Strength and weakness of SF

Empirical SF

- fast calculation
- adaptable to custom target

- training set (incompleteness, inaccuracy of data)
- binding mode (if few polar interactions, underestimation of score)
- missing penalty (steric clashes, polar/apolar match, internal ligand energy, lost of entropy, geometry of directional interaction, local environment of interaction)

Force Field SF

- independant of training set

- strongly depends on ligand size
- force field accuracy, difficulty to set parameters
- no account of entropy
- Often includes empirical terms !

The state of the art

- docking accuracy *(Warren et al. 2006 Proteins, Kellenberger et al. 2004 Proteins)*
  - many programs able to reproduce x-ray conformation
  - but performance is highly dependend on the studied protein
- cpu time:
  - few seconds to several minutes to dock one compound
  - app. screening rate: 1,500 compounds/day/processor
- hit rate (true positives in hit list) in screening by high throughput docking
  - ~ 50 % retrospective studies
  - from 10% to 30% in prospective studies using X-ray structures
  - lower rates for docking using homology models

The limitations

Pre-processing of the protein

Lacking hydrogens, hydrogen bonds networks, protonation states of his, lys, asp, glu

Water molecule(s) involved in the binding mode

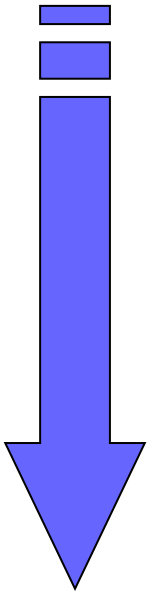
Pre-processing of the ligands

Protonation states, tautomers

Flexibility of the ligand — ***no serious problem***

Flexibility of the protein /binding site — ***a difficult problem***

Fuzzy scoring functions — ***the biggest problem***



APPLICATIONS OF MOLECULAR DOCKING

- Determination of the lowest free energy structures for the receptor-ligand complex.
- Calculate the differential binding of a ligand to two different macromolecular receptors.
- Study the geometry of a particular complex.
- Propose modification of lead molecules to optimize potency or other properties.
- De novo design for lead generation.
- Library design.
- Screening for the side effects that can be caused by interactions with other molecules.
- To check the specificity of the potential drug against homologous proteins through docking.
- Docking is also a widely used tool for predicting protein-protein interaction.
- Knowledge of the molecular associations aid in understanding a variety of pathways taking place in the living and in revealing of the possible pharmacological targets.
- Protein-ligand docking can also be used to predict pollutants that can be degraded by enzymes.

AlphaFold

An overview

BBC

AI breakthrough could spark medical revolution

By Paul Rincon
Science editor, BBC News website

22 July | Comments

The New York Times

La inteligencia artificial predice las formas de las moléculas del futuro

nature

NEWS | 30 November 2020

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

EL PAÍS

GOOGLE DEEPMIND >

La inteligencia artificial revela la forma de los ladrillos básicos de la vida y abre una nueva era en la ciencia

AlphaFold

What is it?

- AF is an **Artificial intelligence** program
  - Google's DeepMind

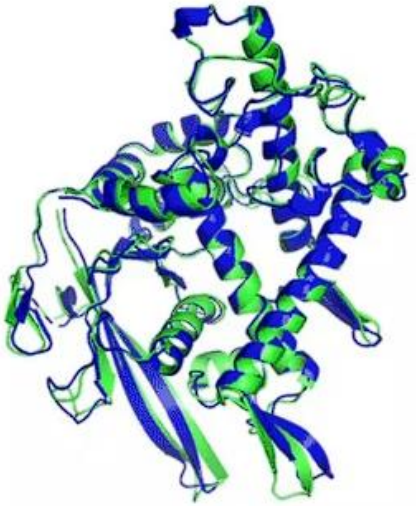
The Goal:

- Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence

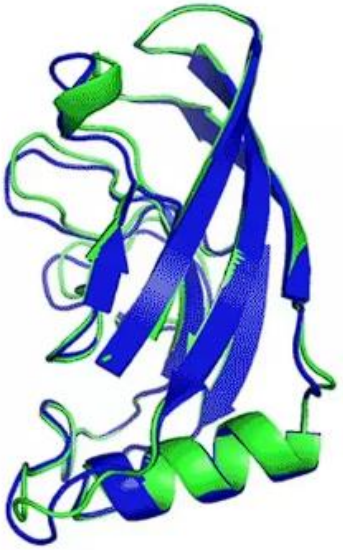
MGAFGHGFG
TYHKLAALED
GTLKHHAKLQ
PHLSLLCMF...

It "solves" two main problems:

1. Sequence-Structure gap
2. Protein folding



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

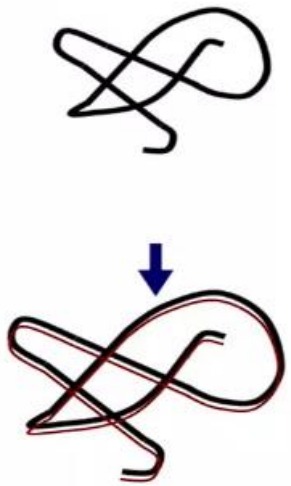
Why solving these problems?

Jumper, J., Evans, R., Pritzel, A. et al. Nature 596, 583–589 (2021).

Homology modeling

INPUT: query sequence Q

MGAFGHGFGTYHKLAALEDIIG



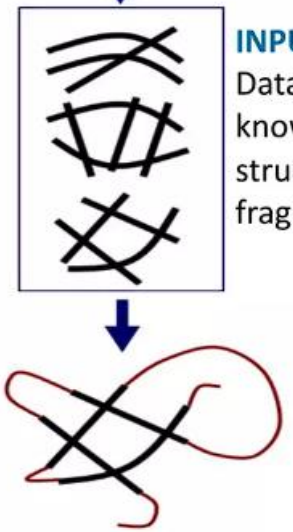
INPUT:
Database of protein structures

1. find protein P **high sequence similarity** to Q
2. return P's **structure as an approximation** to Q's structure

Threading & Fragment assembly

INPUT: query sequence Q

MGAFGHGFGTYHKLAALEDIIG



INPUT:
Database of known folds or structure fragments

1. find a **set of fragments that Q** can be aligned with
2. return **F as an approximation to Q's** structure

Molecular dynamics

INPUT: query sequence Q

MGAFGHGFGTYHKLAALEDIIG

- Force field
- Molecular mechanics

1. **Laws of physics to simulate folding** of Q

CASP before AlphaFold

Homology modeling

Molecular dynamics

Threading & Fragment assembly

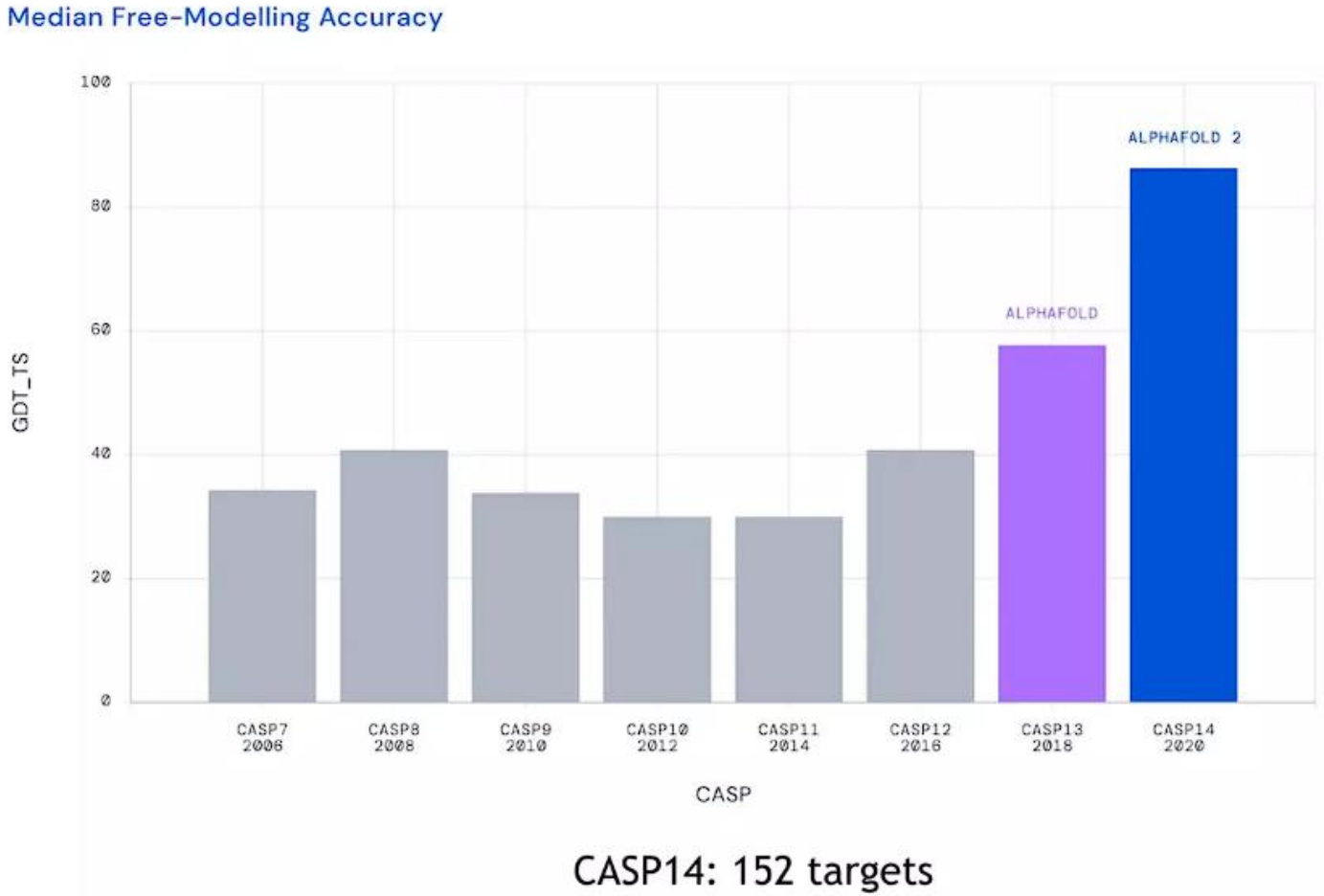
The metric:

- *How well is the prediction compared with the experimental data?*

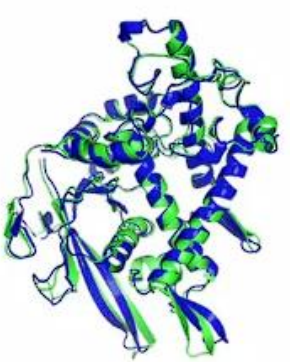
GDT: Global Distance Test

- Compares two structures
- From 0 to 100 (%)
- Greater is better
- Uses distance cutoffs
- Uses alpha Carbons
- More accurate than RMSD

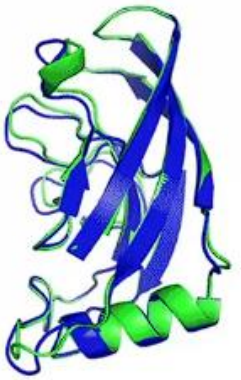
CASP and AlphaFold



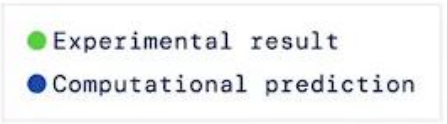
CASP14: 152 targets



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

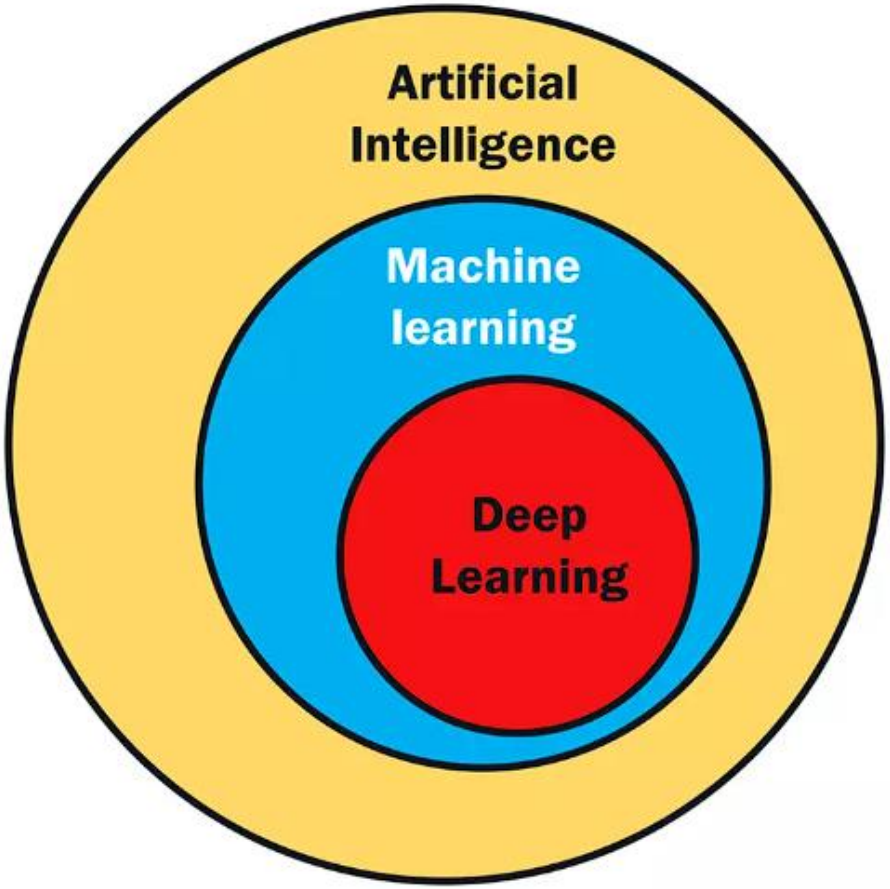


T1049 / 6y4f
93.3 GDT
(adhesin tip)



Jumper, J., Evans, R., Pritzel, A. et al. Nature 596, 583–589 (2021).

How does it work? AlphaFold uses Deep Learning



Machine learning:

Learn from data

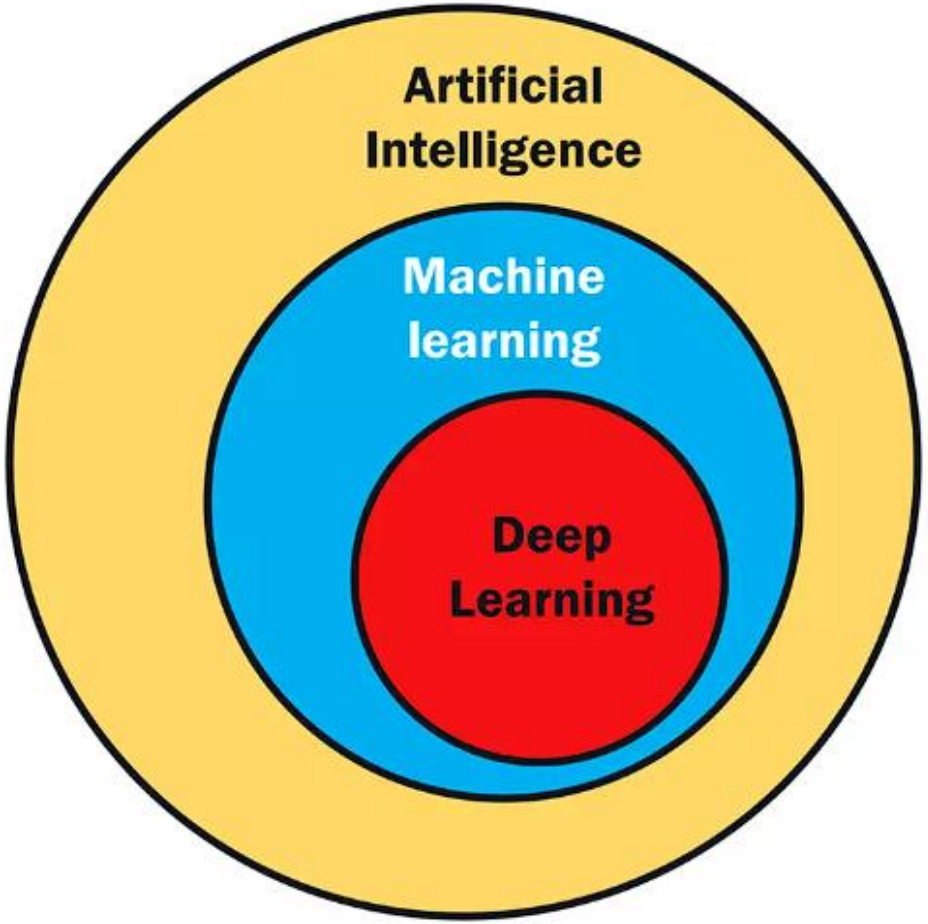
*"The field of study that gives computers the ability to learn without being **explicitly programmed**"*

Traditional Approach

Machine Learning Approach

Grokking Deep Learning/, by **Andrew** W. **Trask**, Manning Publications, 2019

How does it work? AlphaFold uses Deep Learning



Machine learning:

Learn from data

"The field of study that gives computers the ability to learn without being explicitly programmed"

f a true relationship between two variables

$$\mathbf{X} \longrightarrow \mathbf{y}$$

ML: ***approximates*** f ***using data*** $(\mathbf{X}, \mathbf{y})$

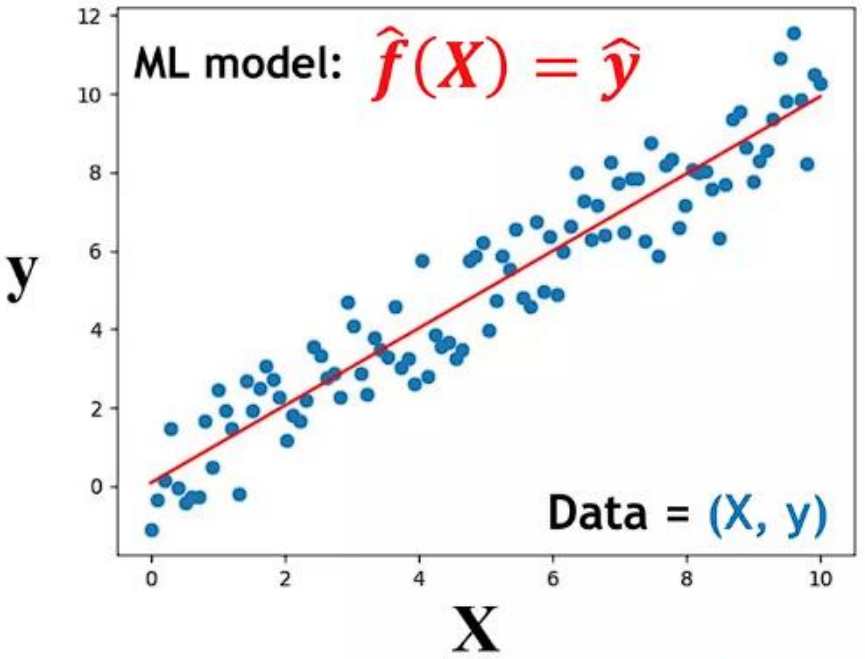
$$f \approx \hat{f} + \mathcal{E}$$

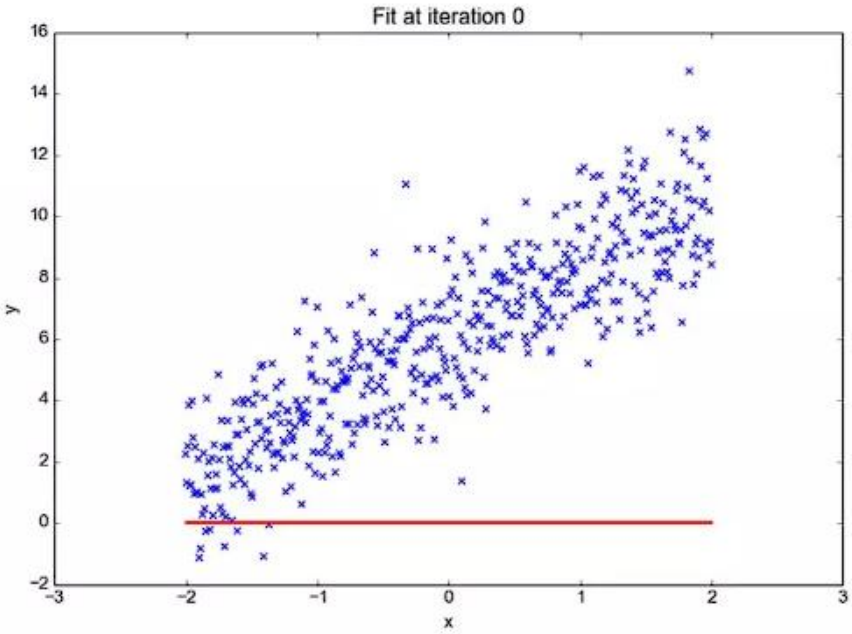
↖ The ML model

Grokking Deep Learning/, by **Andrew** W. **Trask**, Manning Publications, 2019

Machine Learning

A linear **regression** model





1. The ML model (blueprint): $\hat{y} = w * x + b$
2. A training algorithm
   1. Data (training set)
   2. Loss function (error) $\hat{y} \approx y$
   3. Optimization algorithm
3. A validation and a test set

The goal: Minimize the error

1. Training set
2. Test set (data never seen by the model)

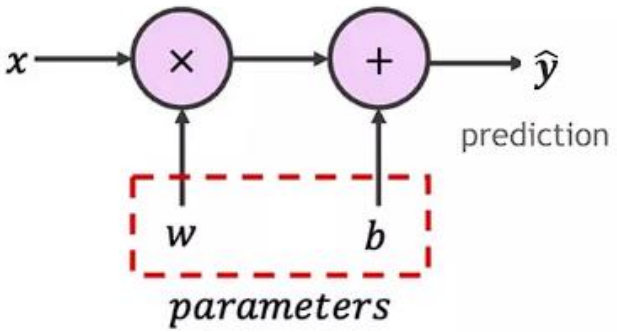
↓

Generalization

Machine Learning → Deep Learning

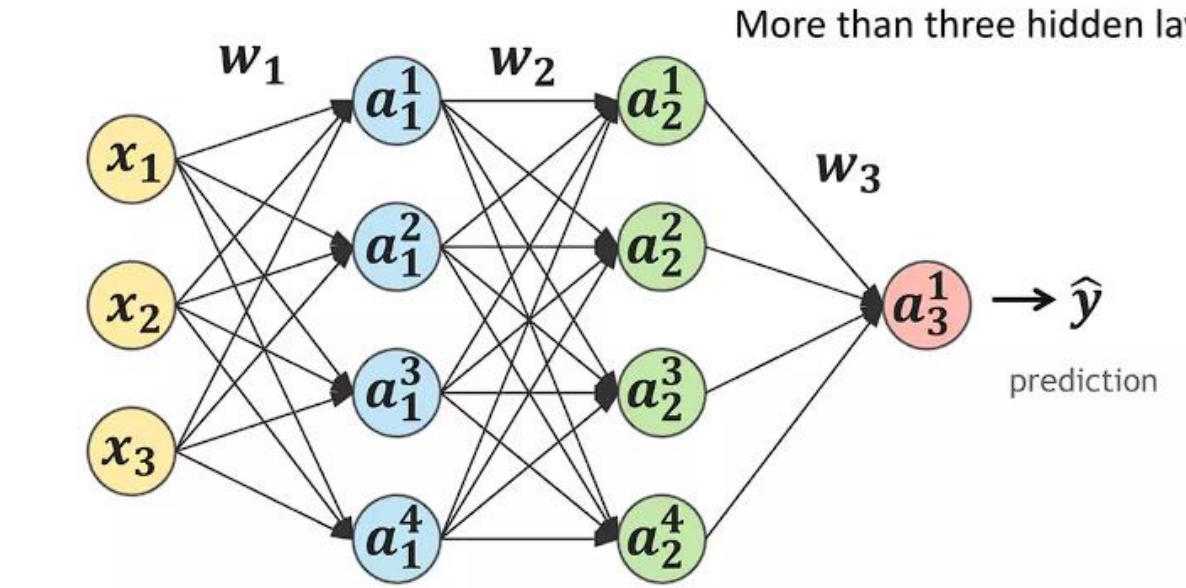
A linear **regression** model

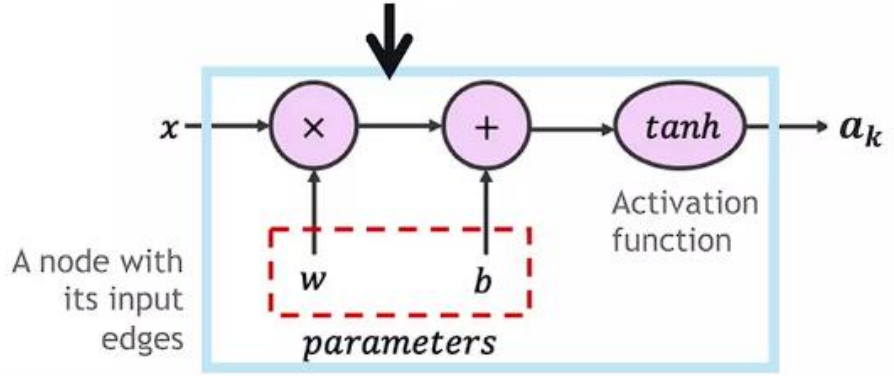
$$\hat{y} = w * x + b$$



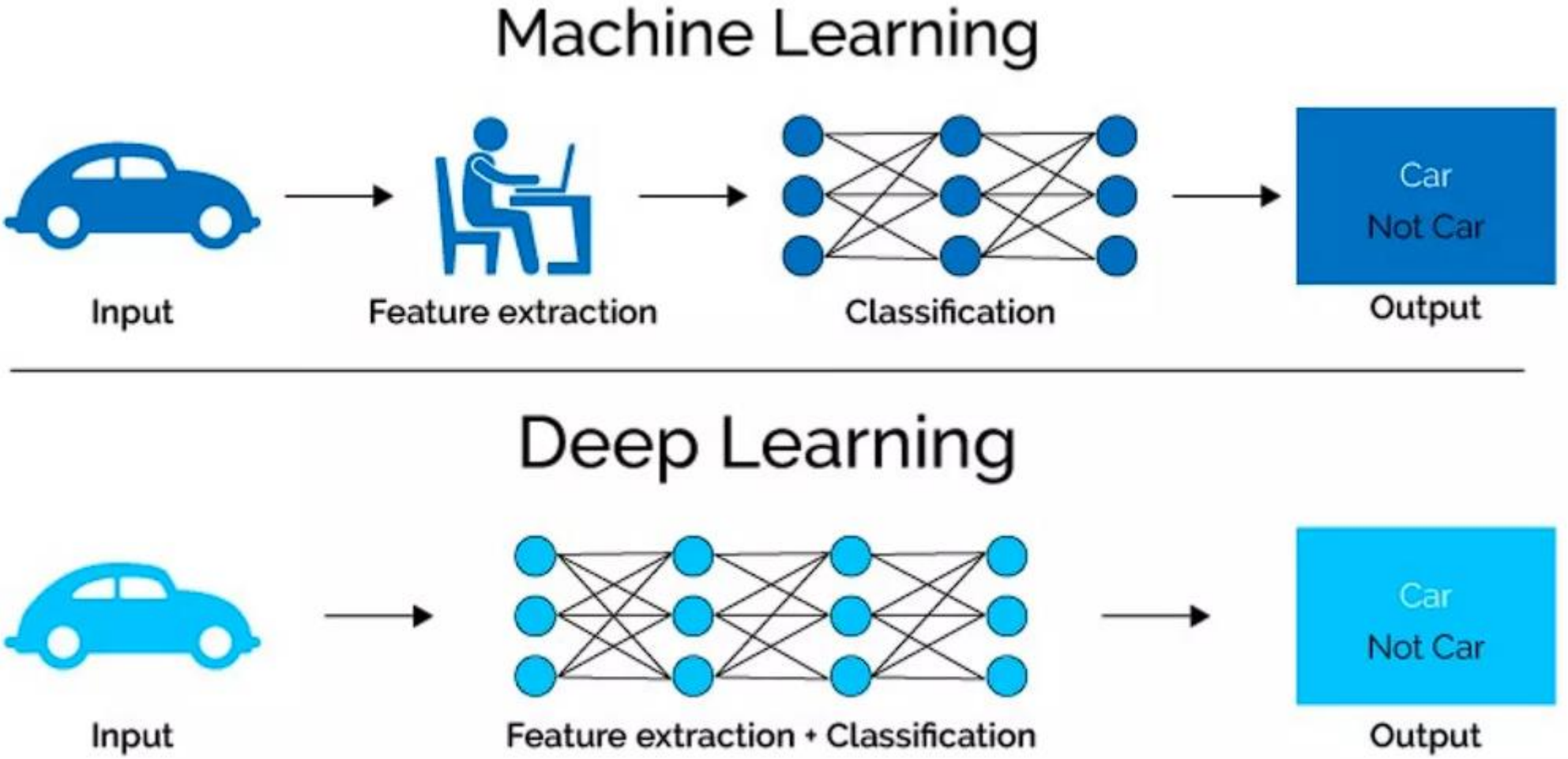
- A linear relationship between ***x*** and ***y***

A Neural Network (Feed Forward)



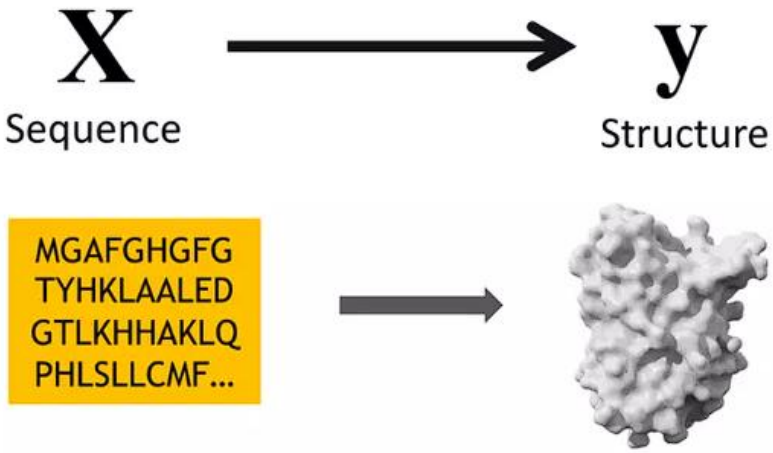


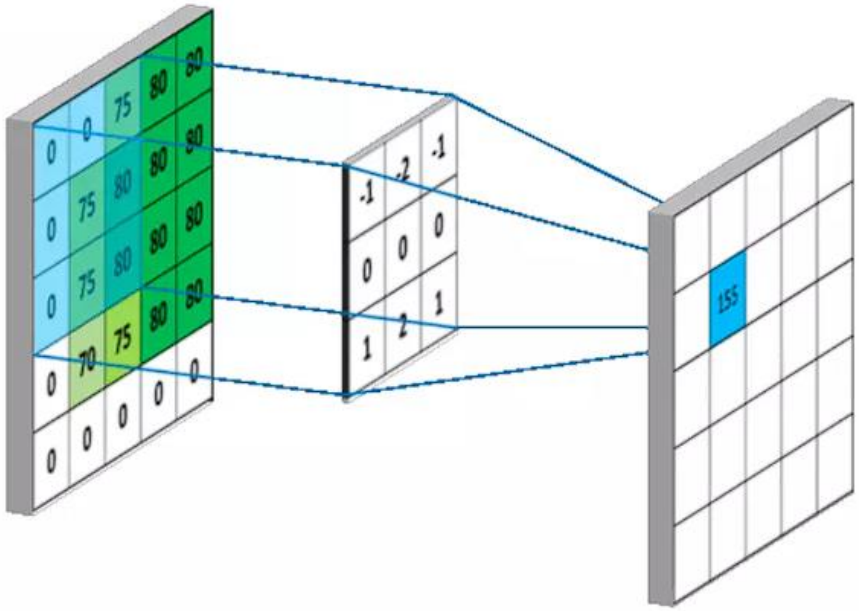
- More complex models
- Learns no linear relationships
- Learns interactions between features (***X***)
  - **Feature extraction**



Key: Feature Extraction

AlphaFold 1





The model:

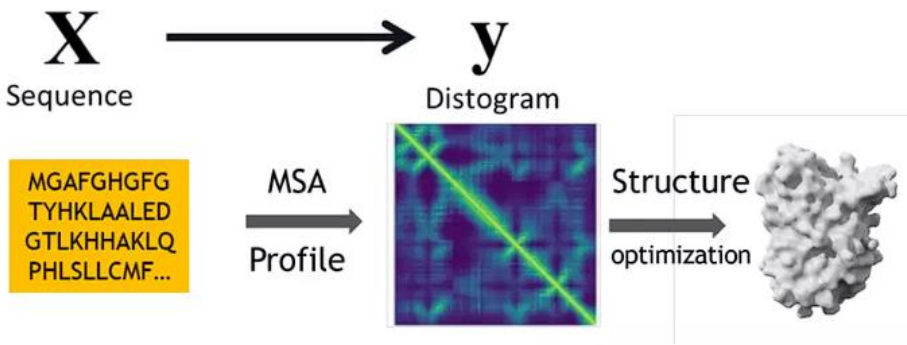
- **CASP13 (2018)**
- **Convolutional-based Neural Network**

Training:

- **Structures:** 31,247 domains
- **Sequences:** UniClust30

Senior, et al. (2020). **Nature**, 577(7792), 706-710.

AlphaFold 1



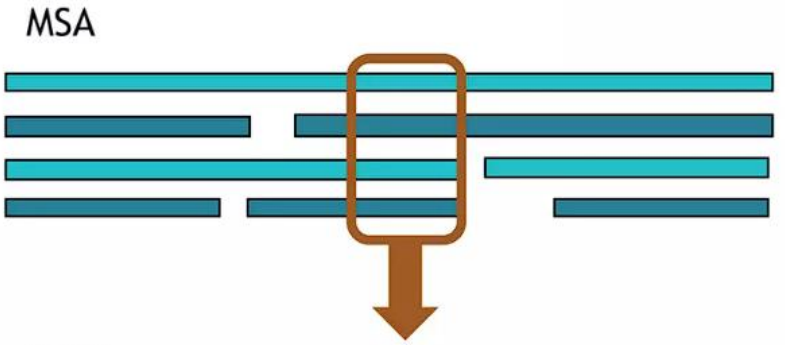
- **Input:**
  - Protein amino acid sequence

↓

- **Multiple Sequence Alignments (MSA):**
  - Profile features

MSA to Profiles - PSSM:

→ Familes and Domains



PSSM

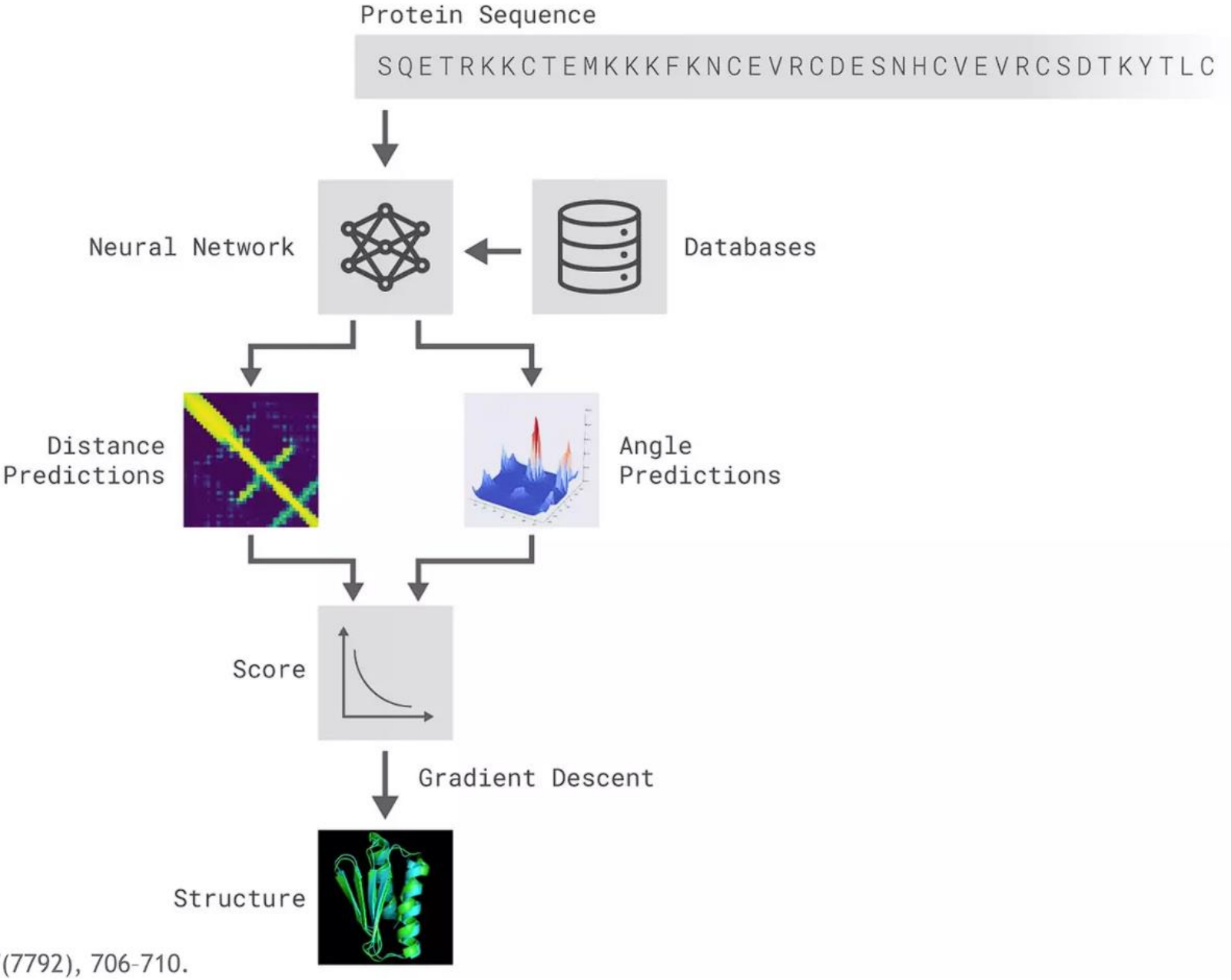
sequence positions

amino acids	1	2	3	4	5	6	7	.. .	n
A	-	-	0.2	-	-	-	-	-	-
R	-	-	0.3	-	-	-	-	-	-
F	-	-	0.5	-	-	-	-	-	-
G	-	-	0	-	-	-	-	-	-
..	-	-	0.8	-	-	-	-	-	-
Y	-	-	1.2	-	-	-	-	-	-

position-specific scoring matrix

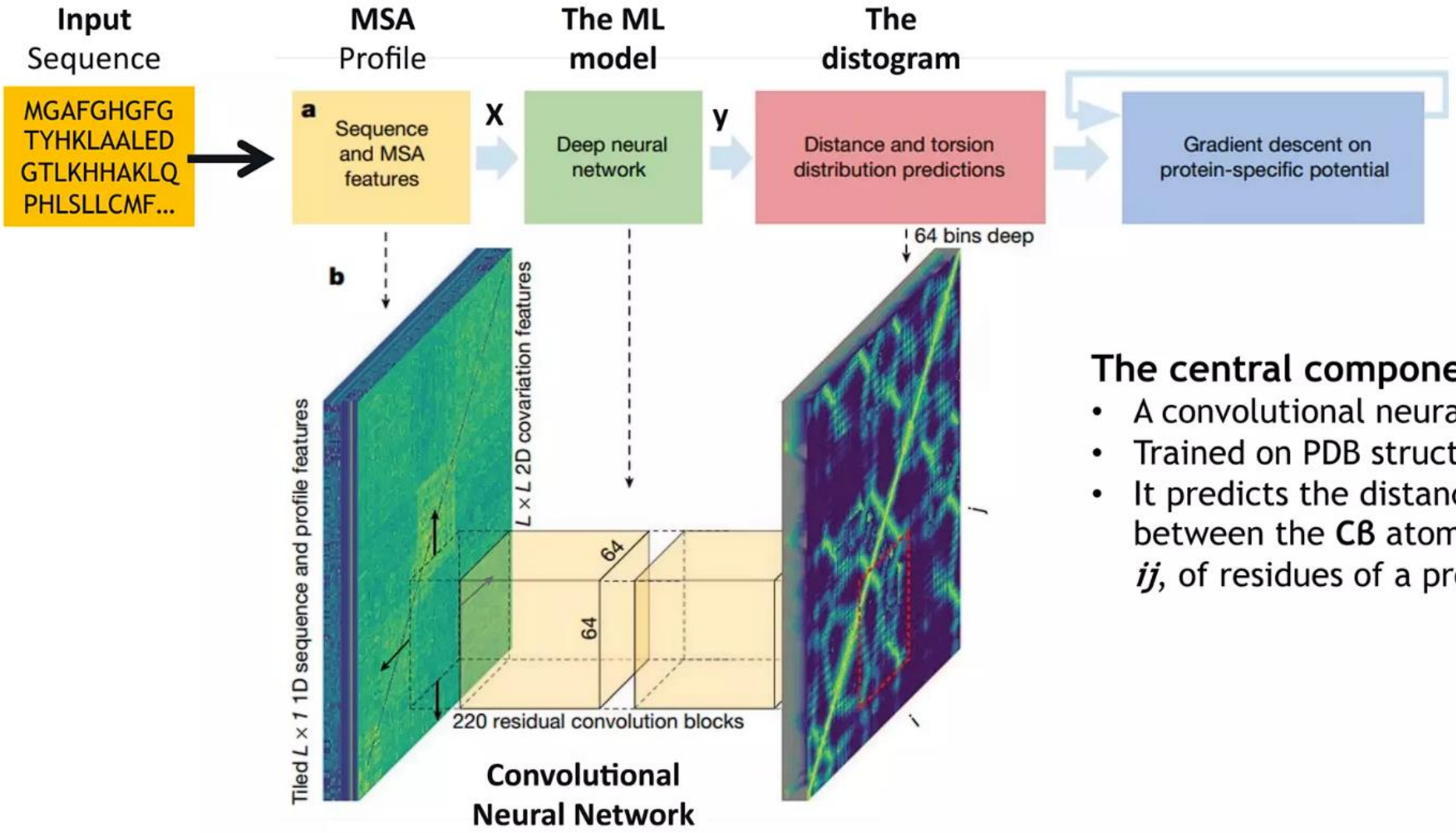
Senior, et al. (2020). **Nature**, 577(7792), 706–710.

AlphaFold 1



Senior, et al. (2020). **Nature**, 577(7792), 706–710.

AlphaFold 1



The central component:

- A convolutional neural network
- Trained on PDB structures
- It predicts the distances d_{ij} between the **Cβ** atoms of pairs, ij, of residues of a protein.

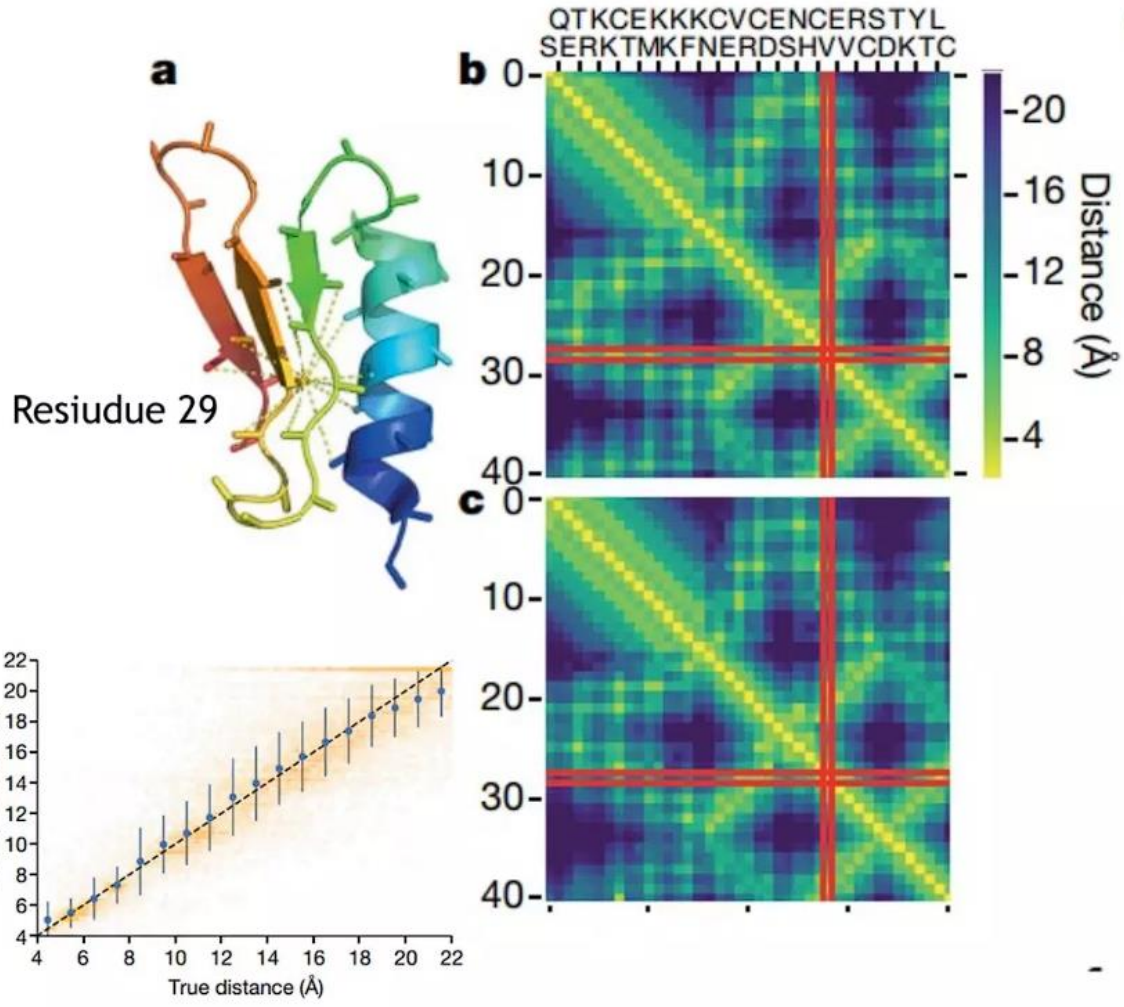
Senior, et al. (2020). **Nature**, 577(7792), 706-710.

Input MSA

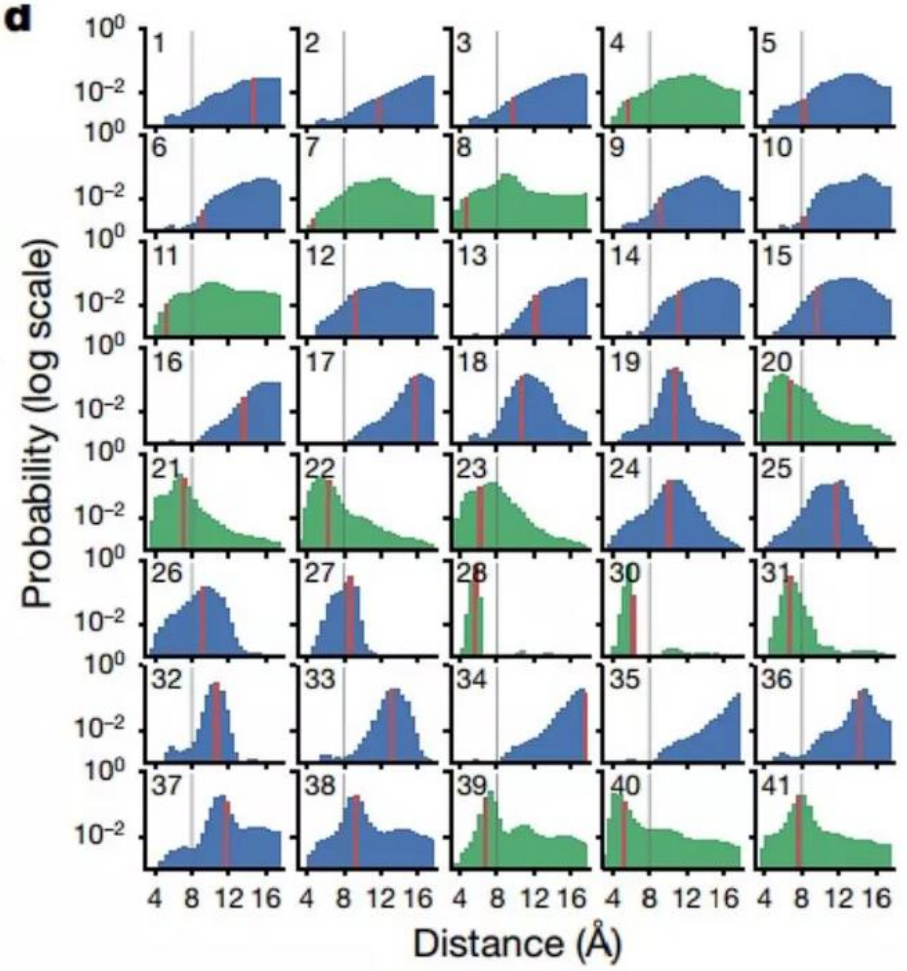
- DB Search
  - JackHMMER > Mgnify (Metagenome DB)
  - JackHMMER > Uniref90
  - Hhblits > Uniclust30 + BFD (Metagenome DB)
  - Above are simply stacked.
- MSA processing (to reduce the memory usage)
  - MSA block deletion (MSA-level dropout)
  - Nseq sequences are randomly selected as MSA cluster centers
  - 15% are masked
  - Among masked: 10%, replaced to random amino acids; 10%, replaced from MSA profile; 10%, unchanged; 70%: targets for prediction.
  - Nextra_Seq selected for extra MSA sequences

AlphaFold 1

The distogram

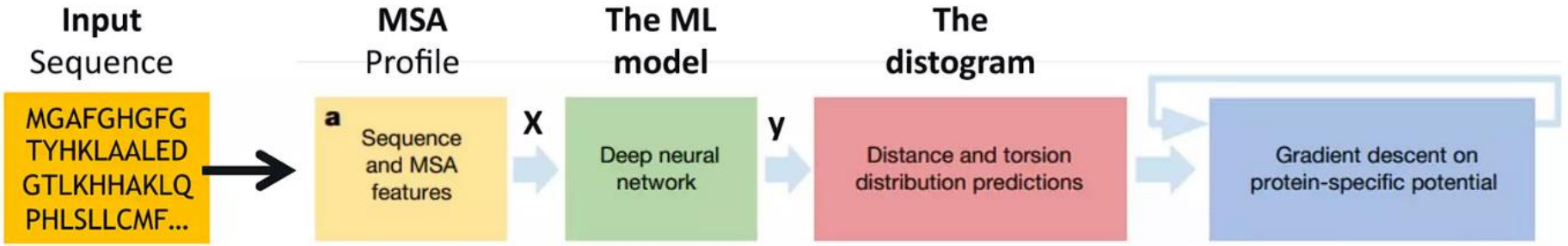


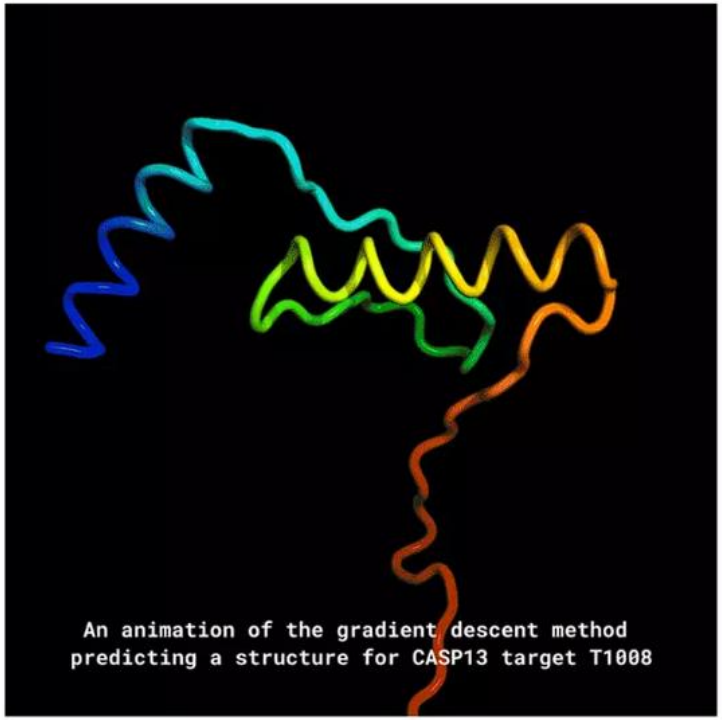
The predicted probability distributions for distances of residue 29 to all other residues (41)

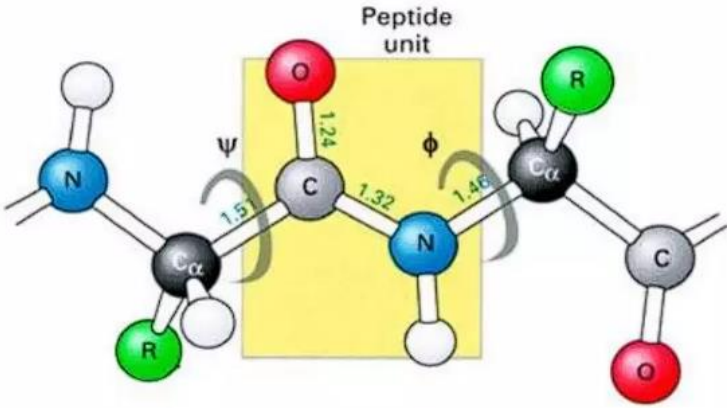


Senior, et al. (2020). **Nature**, 577(7792), 706–710.

AlphaFold 1 Protein folding



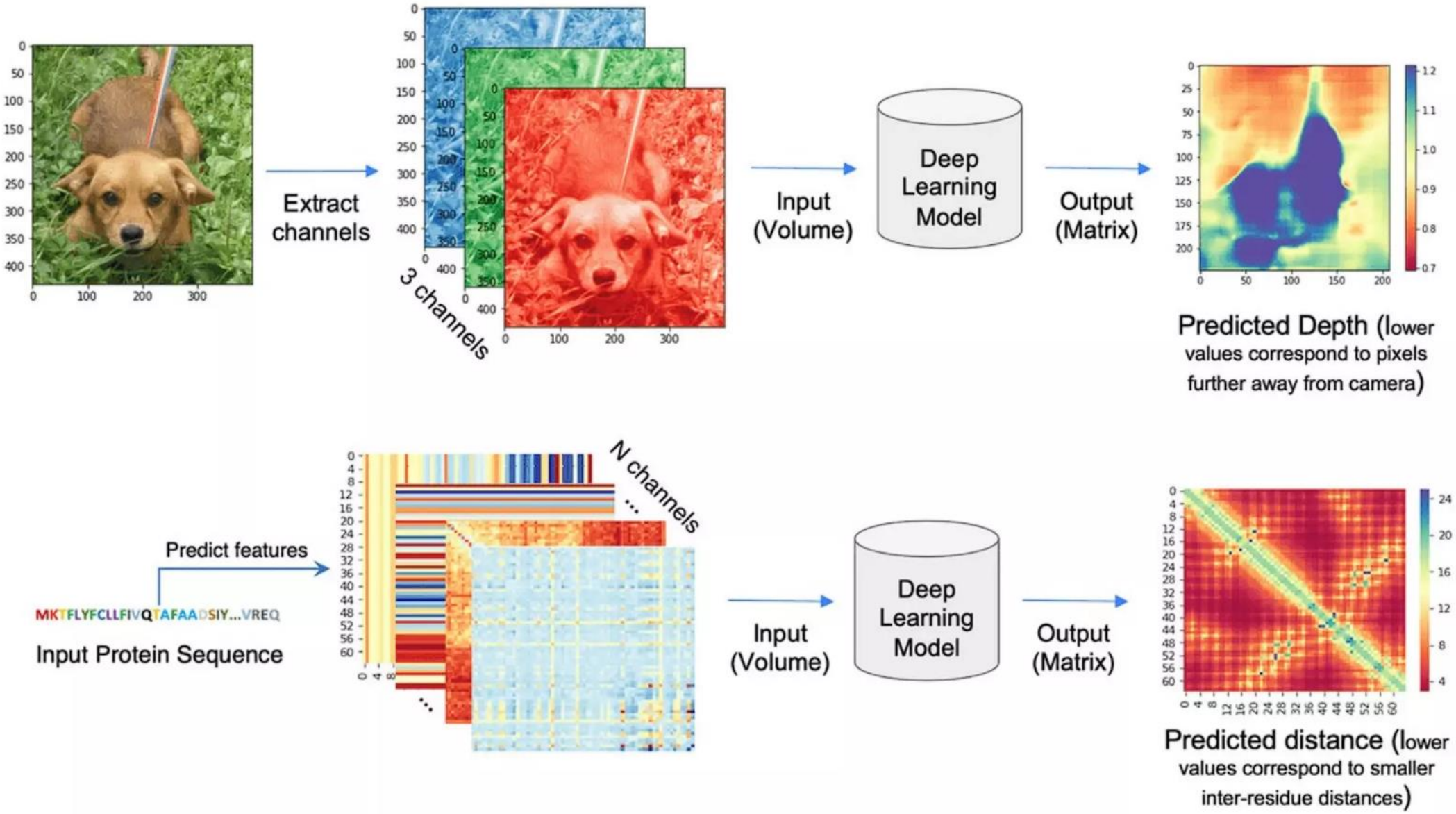


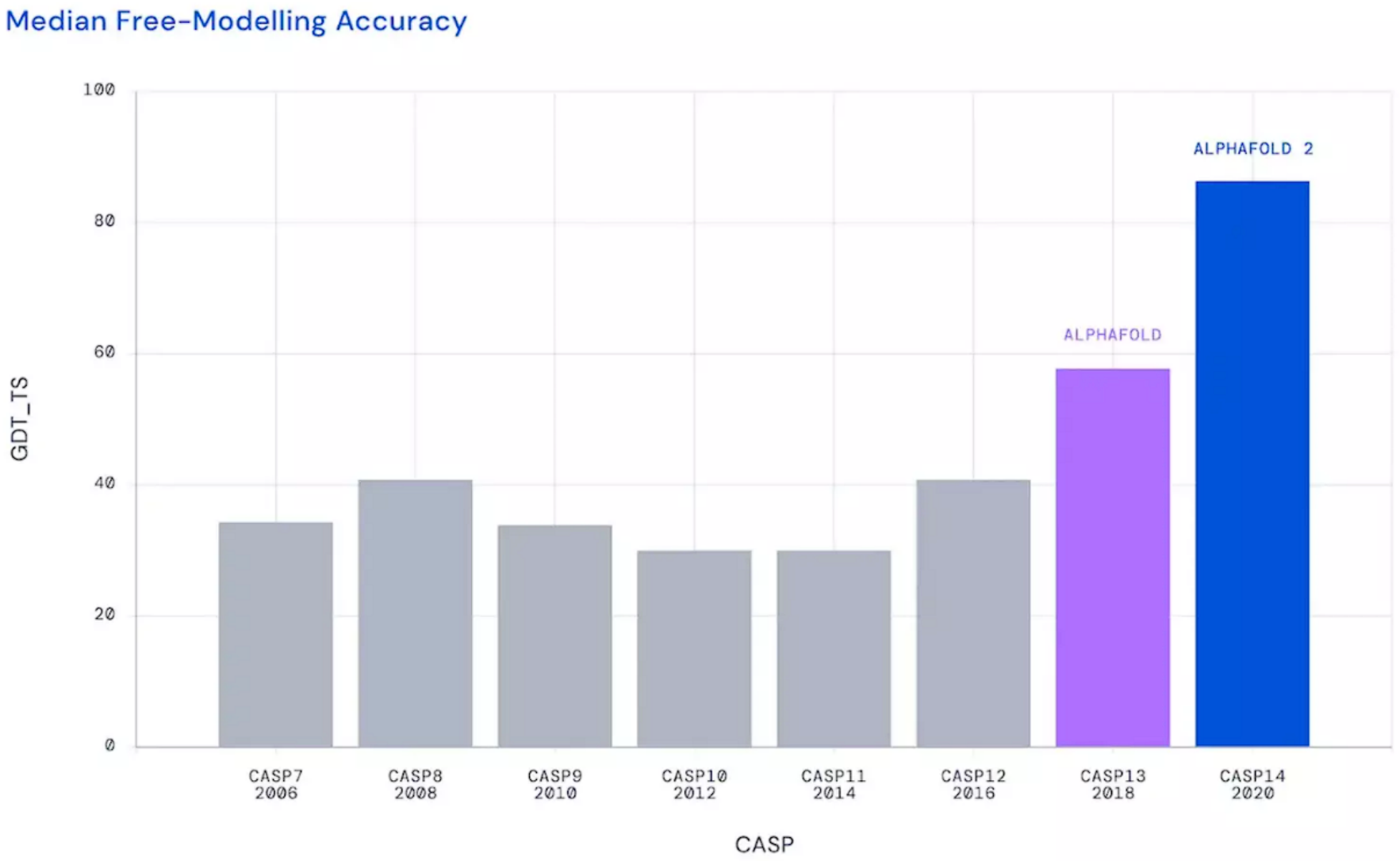


Gradient descent:

- Rotate the phi and psi angles
- Match the predicted **Cβ** atoms distances

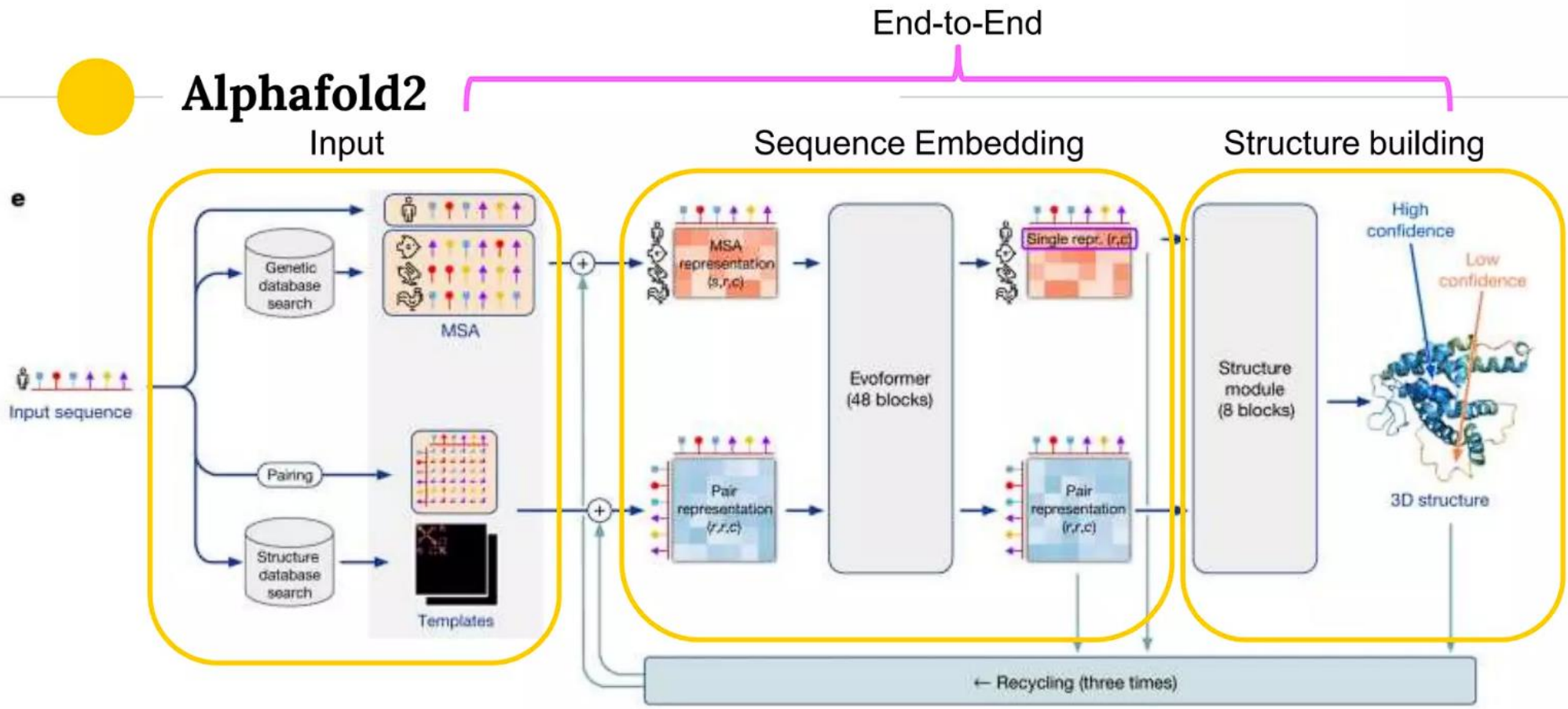
Senior, et al. (2020). **Nature**, 577(7792), 706-710.


Extract channels
3 channels
Input (Volume)
Deep Learning Model
Output (Matrix)
Predicted Depth (lower values correspond to pixels further away from camera)
Predict features
MKTFLYFCLLFIVQTAFAADSIY...VREQ
Input Protein Sequence
N channels
Input (Volume)
Deep Learning Model
Output (Matrix)
Predicted distance (lower values correspond to smaller inter-residue distances)



Jumper, J., Evans, R., Pritzel, A. et al. Nature 596, 583–589 (2021).

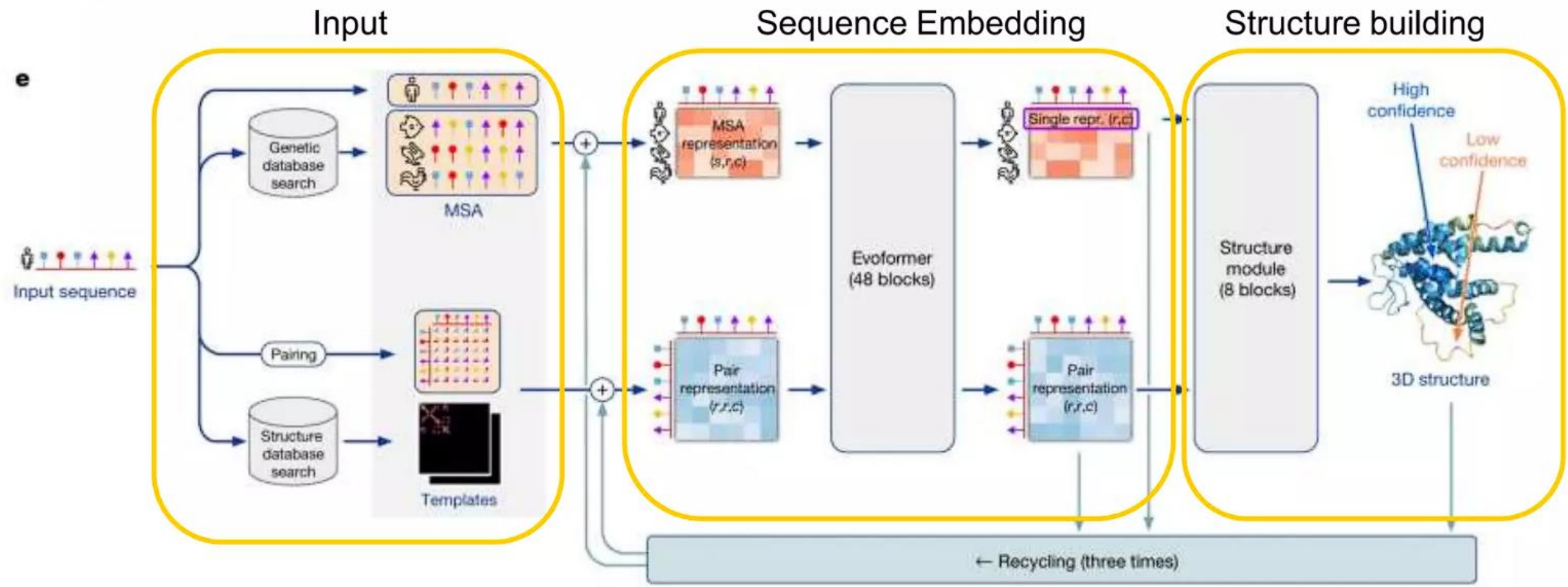
Alphafold2



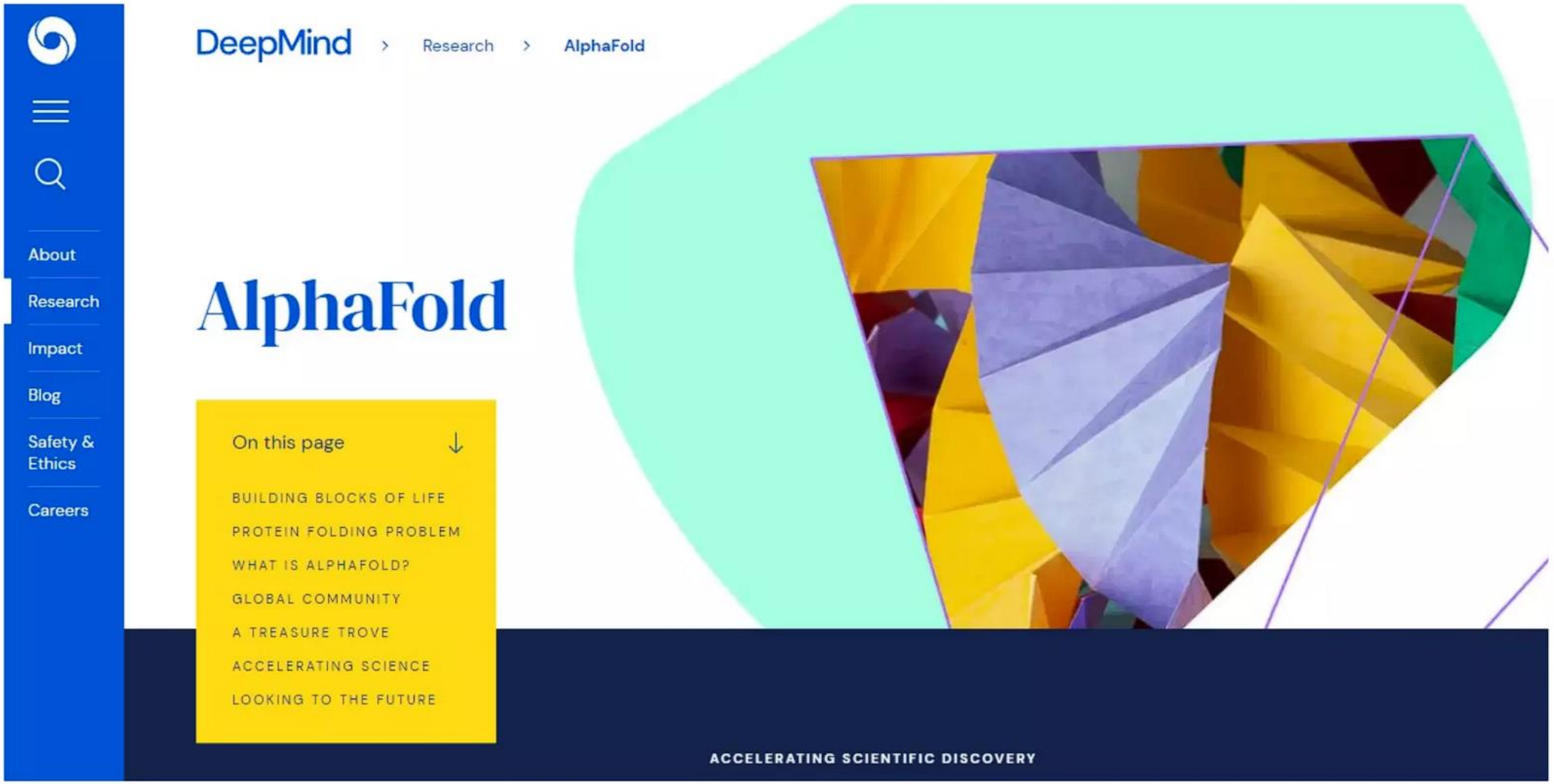
(Inference) code at Github & well-documented Suppl. Info. Released

(Jumper J et al., Nature 2021)

Alphafold2 Paper & Code Release



(Jumper J et al., Nature 2021)



https://github.com/deepmind/alphafold