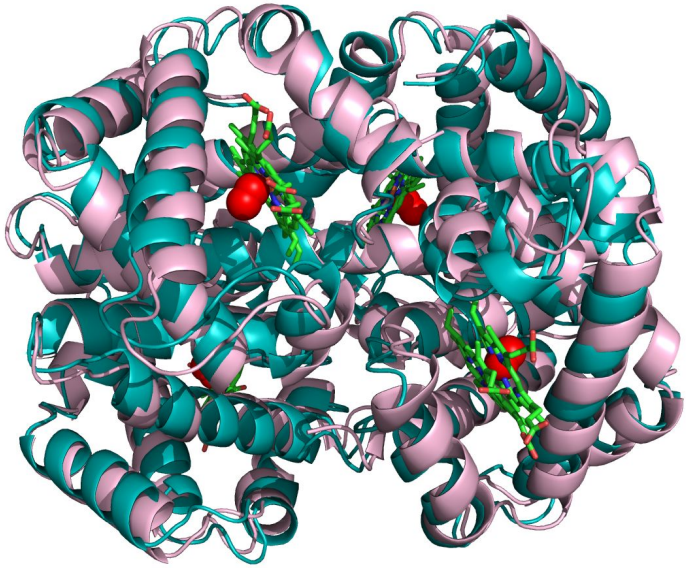
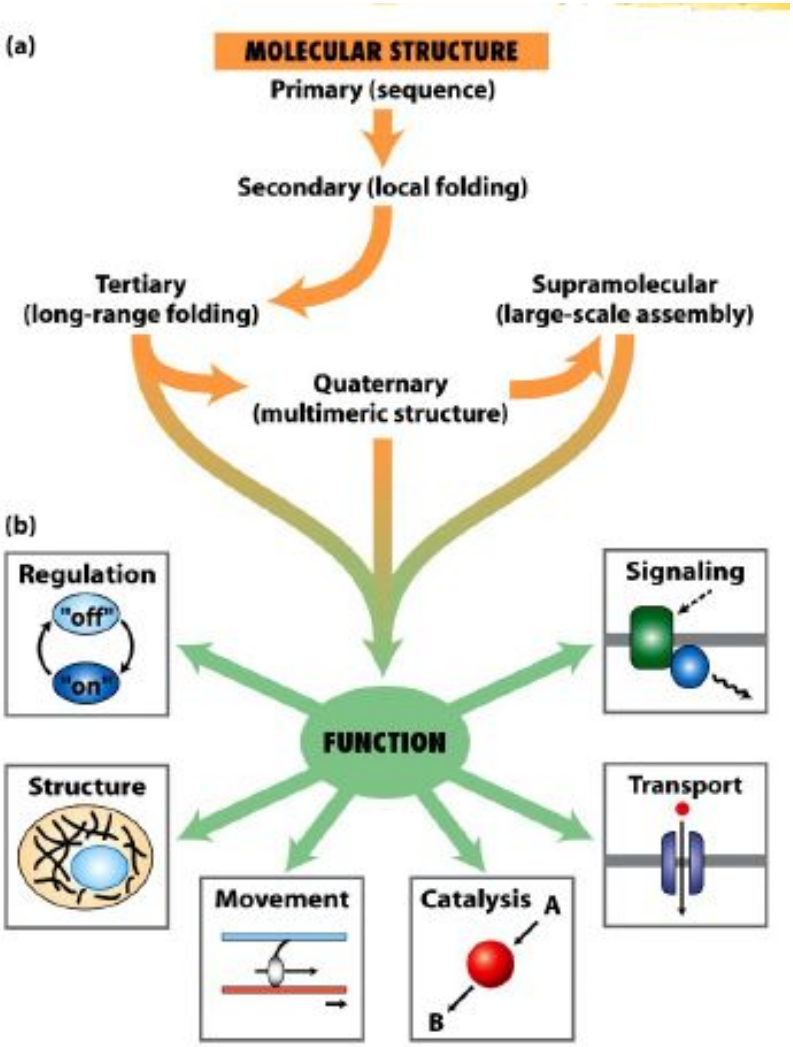


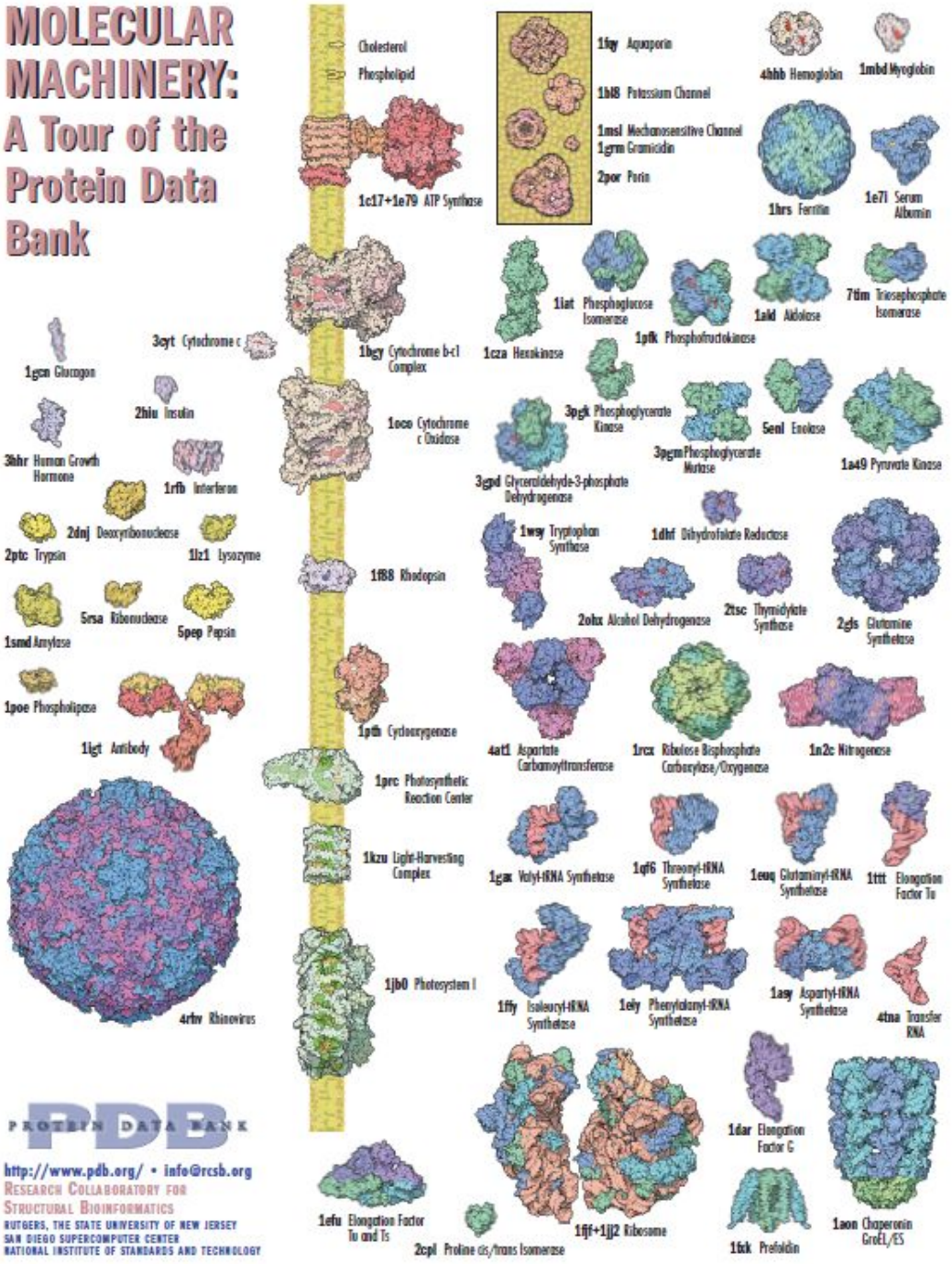
Estructuras de proteínas

Bioinformática
UNSAM - 2016

Formas y funciones de las proteínas



MOLECULAR MACHINERY: A Tour of the Protein Data Bank

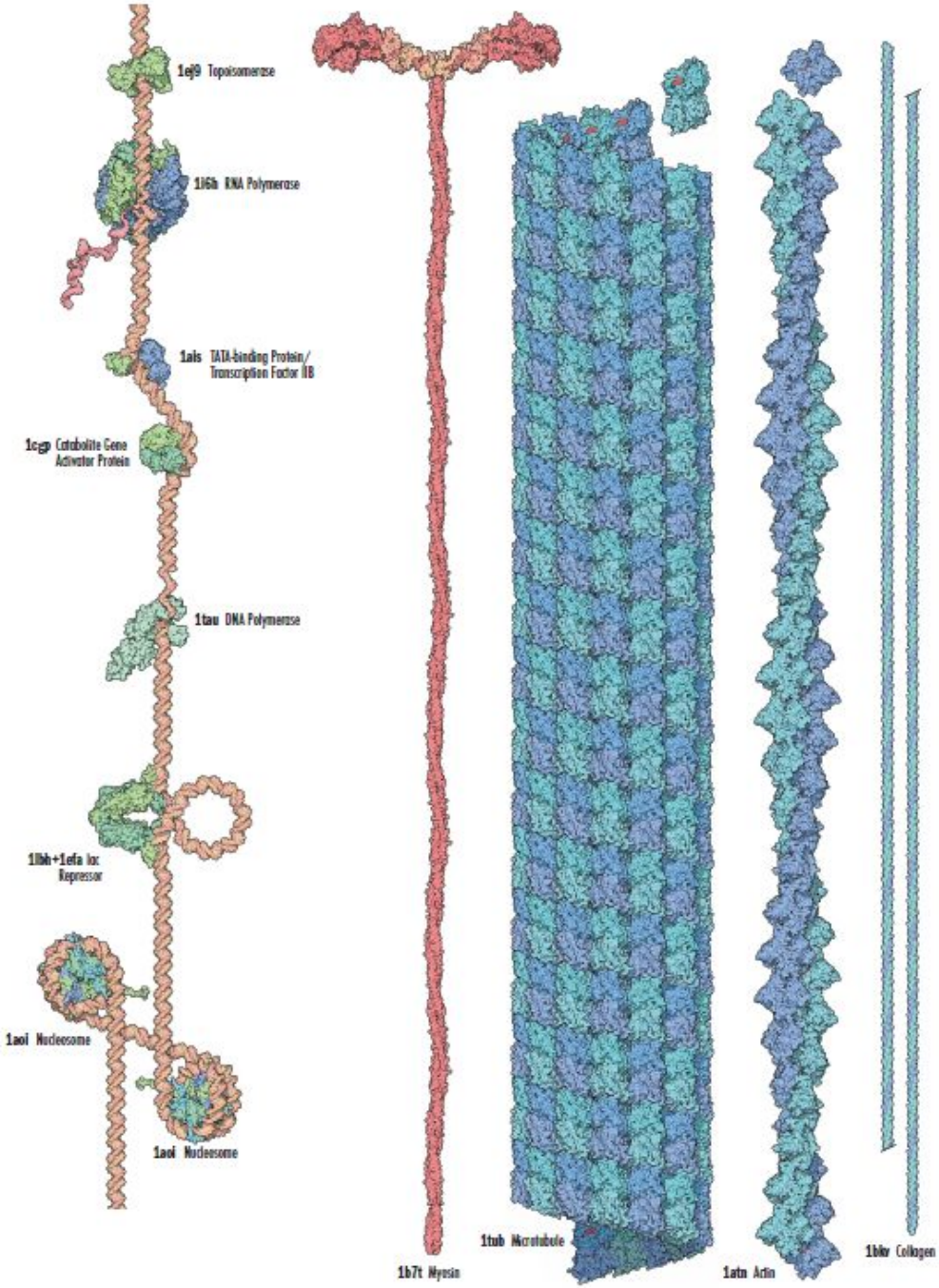


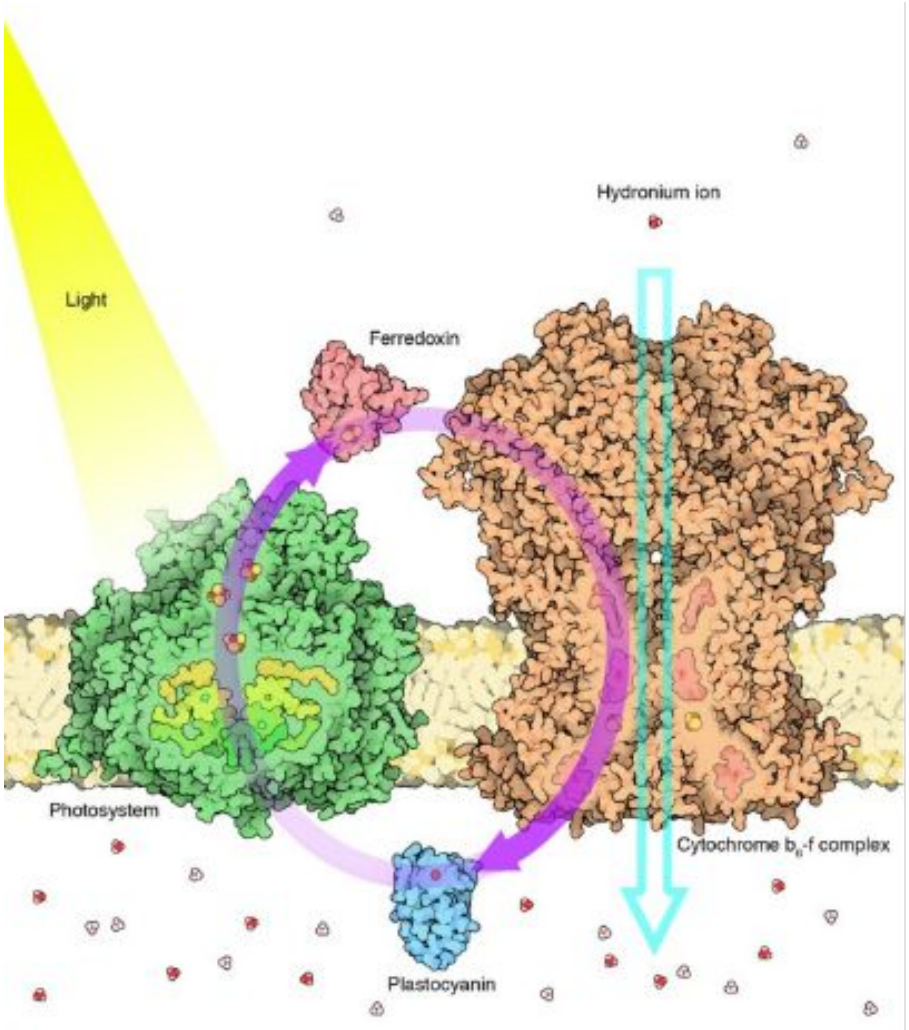
PROTEIN DATA BANK

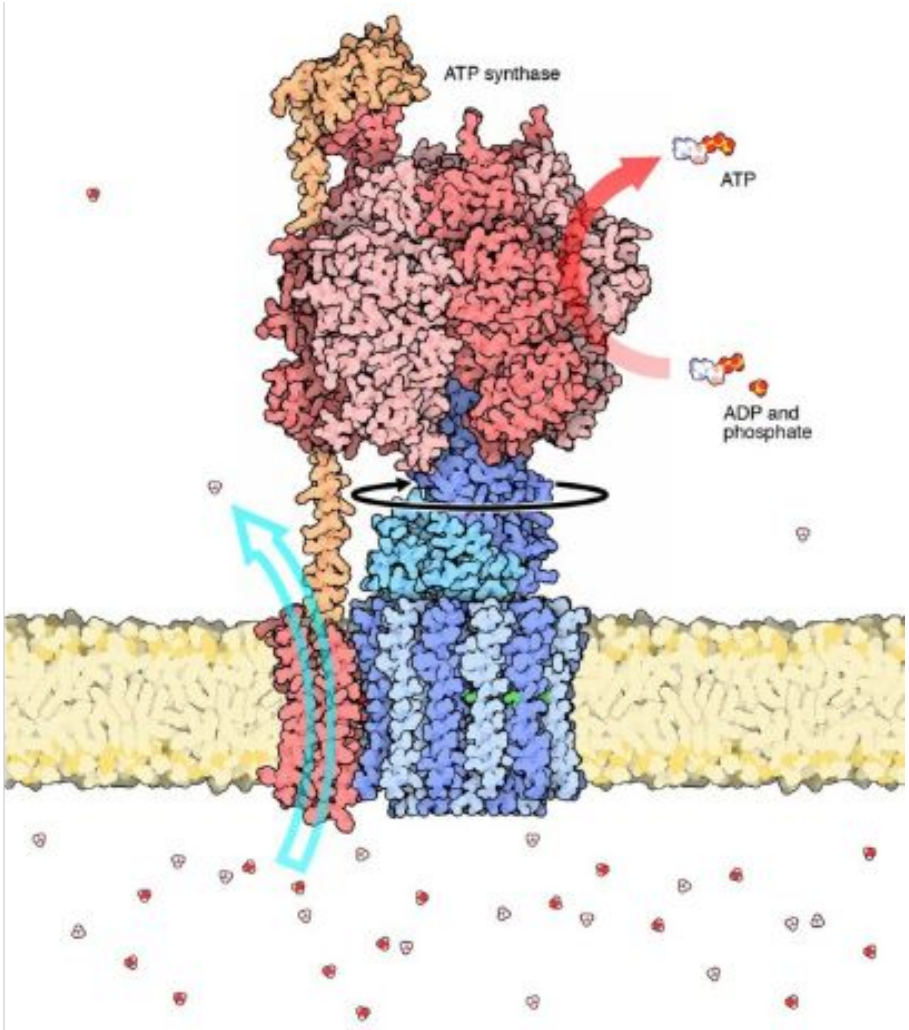
http://www.pdb.org/ • info@rcsb.org

RESEARCH COLLABORATORY FOR STRUCTURAL BIOINFORMATICS

RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY
SAN DIEGO SUPERCOMPUTER CENTER
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY


1ej9 Topoisomerase
1l6h RNA Polymerase
1ais TATA-binding Protein/
Transcription Factor IIB
1cgp Catabolite Gene
Activator Protein
1tau DNA Polymerase
1lbh+1efa lac
Repressor
1aoi Nucleosome
1aoi Nucleosome
1b7t Myosin
1tub Microtubule
1atn Actin
1bkv Collagen


Light
Hydronium ion
Ferredoxin
Photosystem
Cytochrome b_6-f complex
Plastocyanin


ATP synthase
ATP
ADP and phosphate

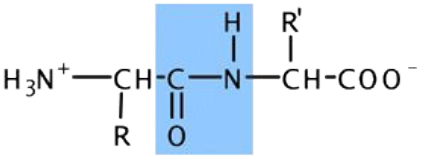
ON THE STRUCTURE OF NATIVE, DENATURED, AND COAGULATED PROTEINS

By A. E. Mirsky* and Linus Pauling

Gates Chemical Laboratory, California Institute of Technology, Pasadena, California

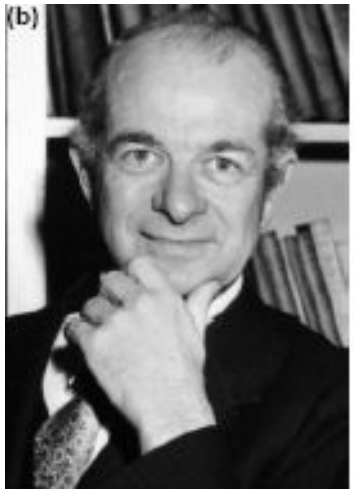
Communicated June 1, 1936

"Our conception of a ***native protein molecule*** (showing specific properties) is the following. The molecule consists of one polypeptide chain which continues without interruption throughout the molecule (or, in certain cases, of two or more such chains), this chain is ***folded into a uniquely defined configuration***, in which it is held by hydrogen bonds between the peptide nitrogen and oxygen atoms and also between the free amino and carboxyl groups of the diamino and dicarboxyl amino acids residues"



"The characteristic specific properties of native proteins we attribute to their **uniquely defined configurations**"

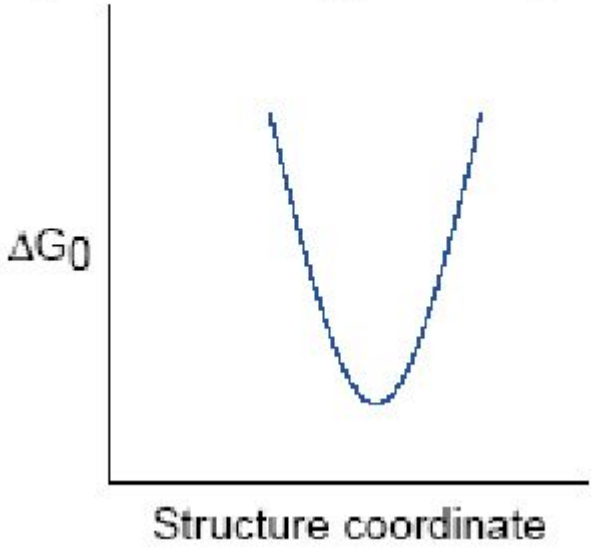
"The denature protein molecule we consider to be characterized by the ***absence*** of a uniquely defined configuration"



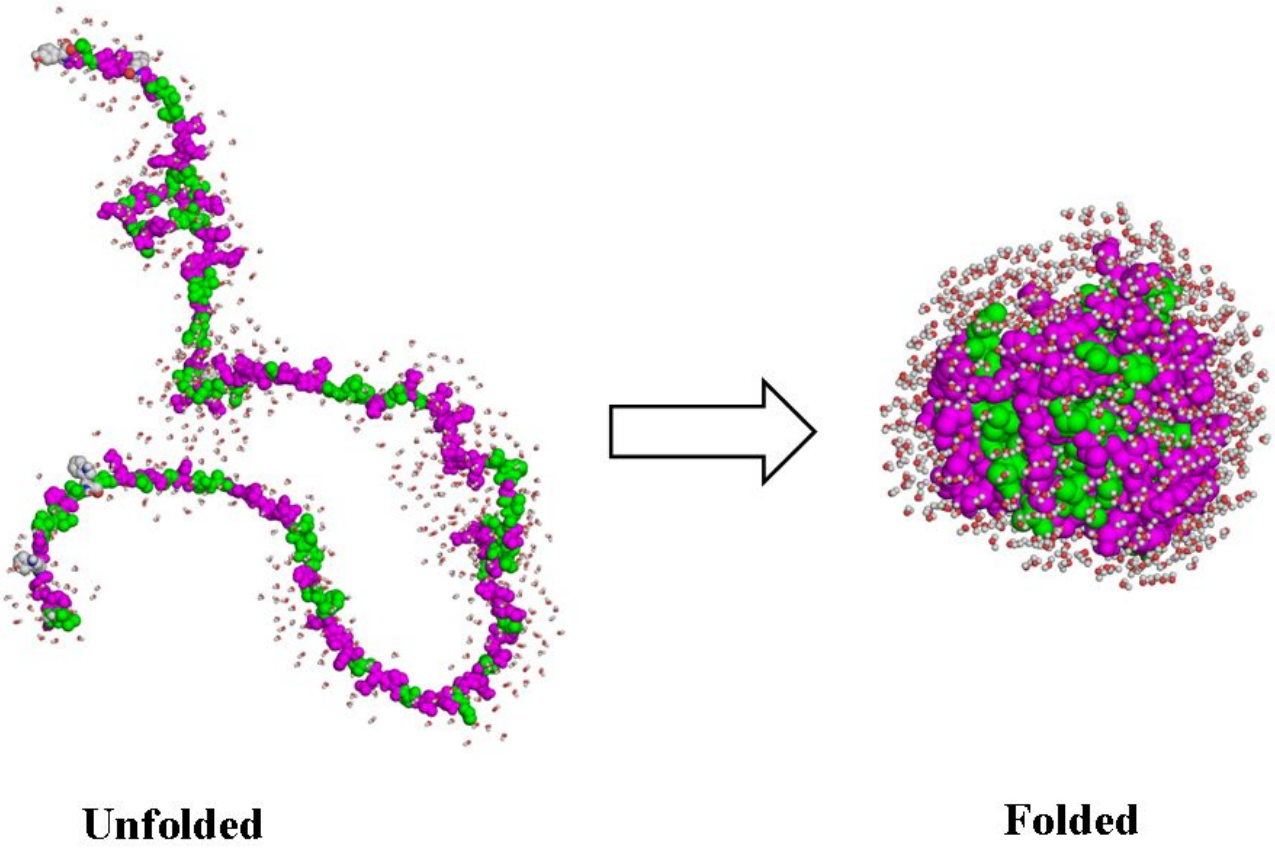
Linus Pauling
1904-1994

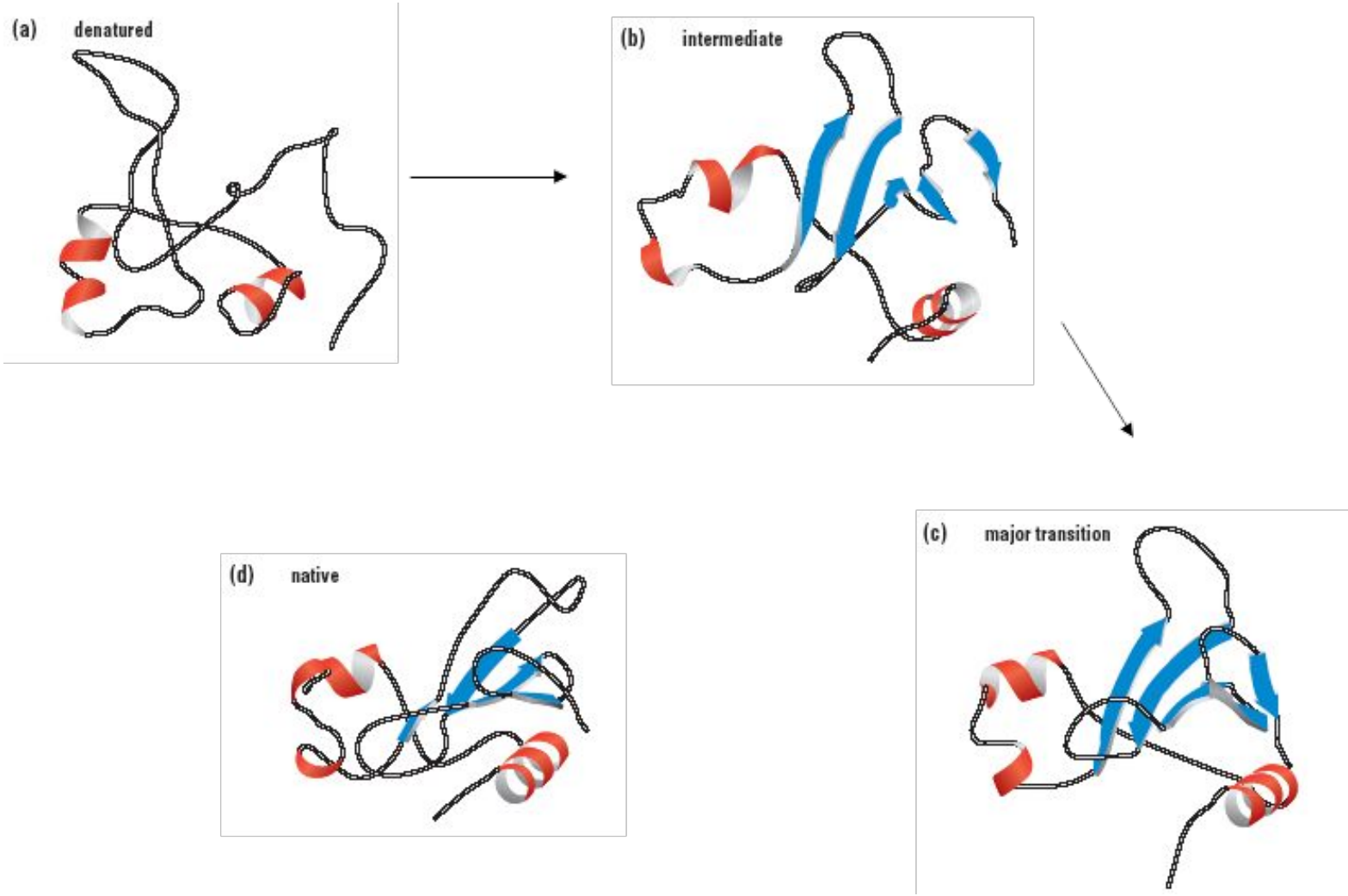
(a) 'Simplistic view'

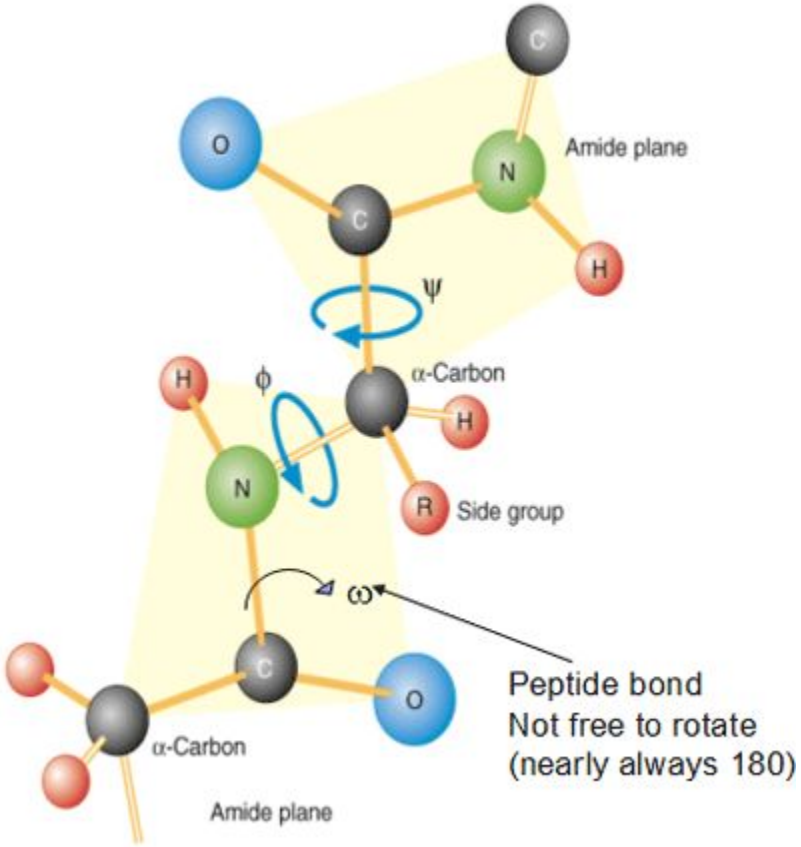
(i) Smooth energy landscape

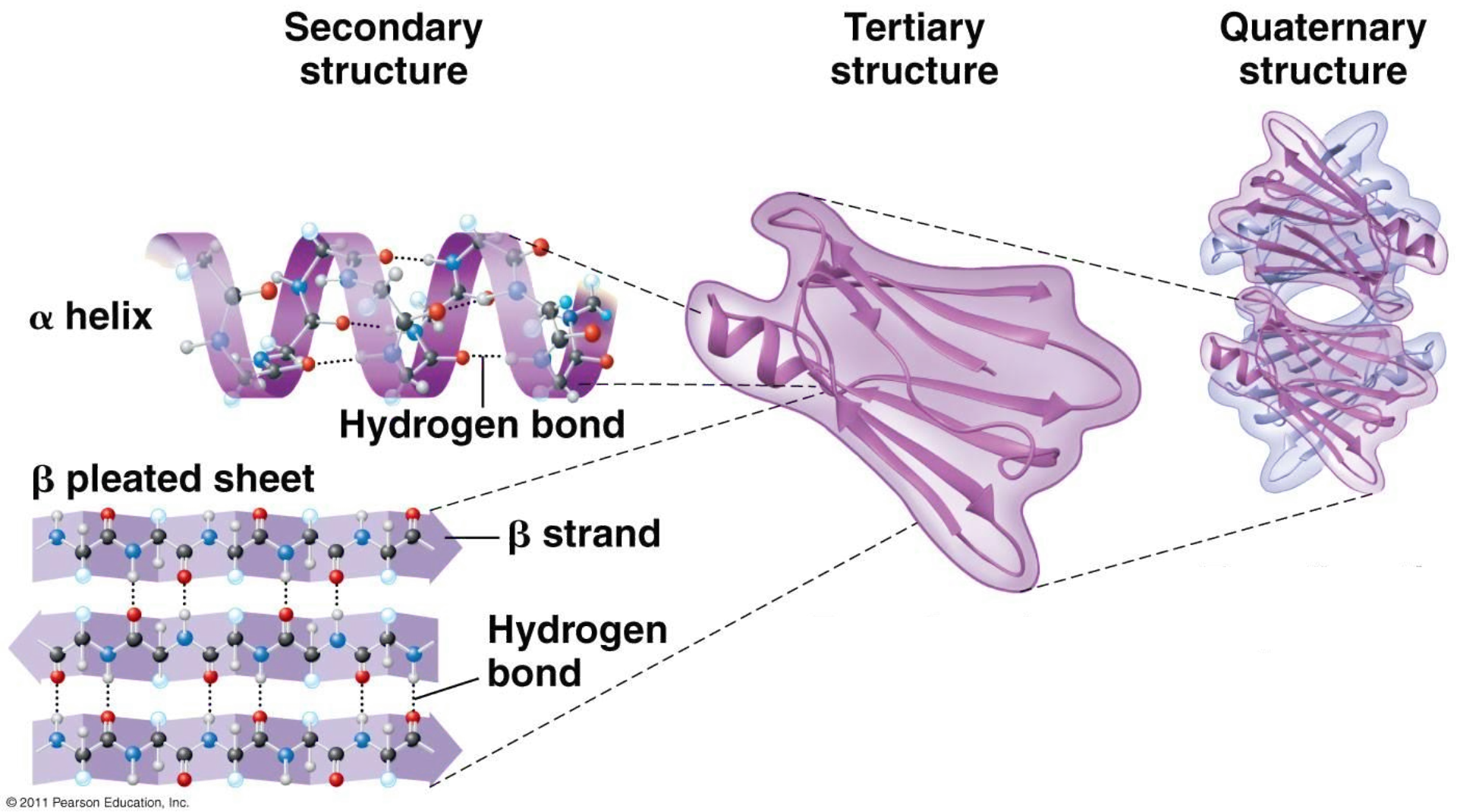


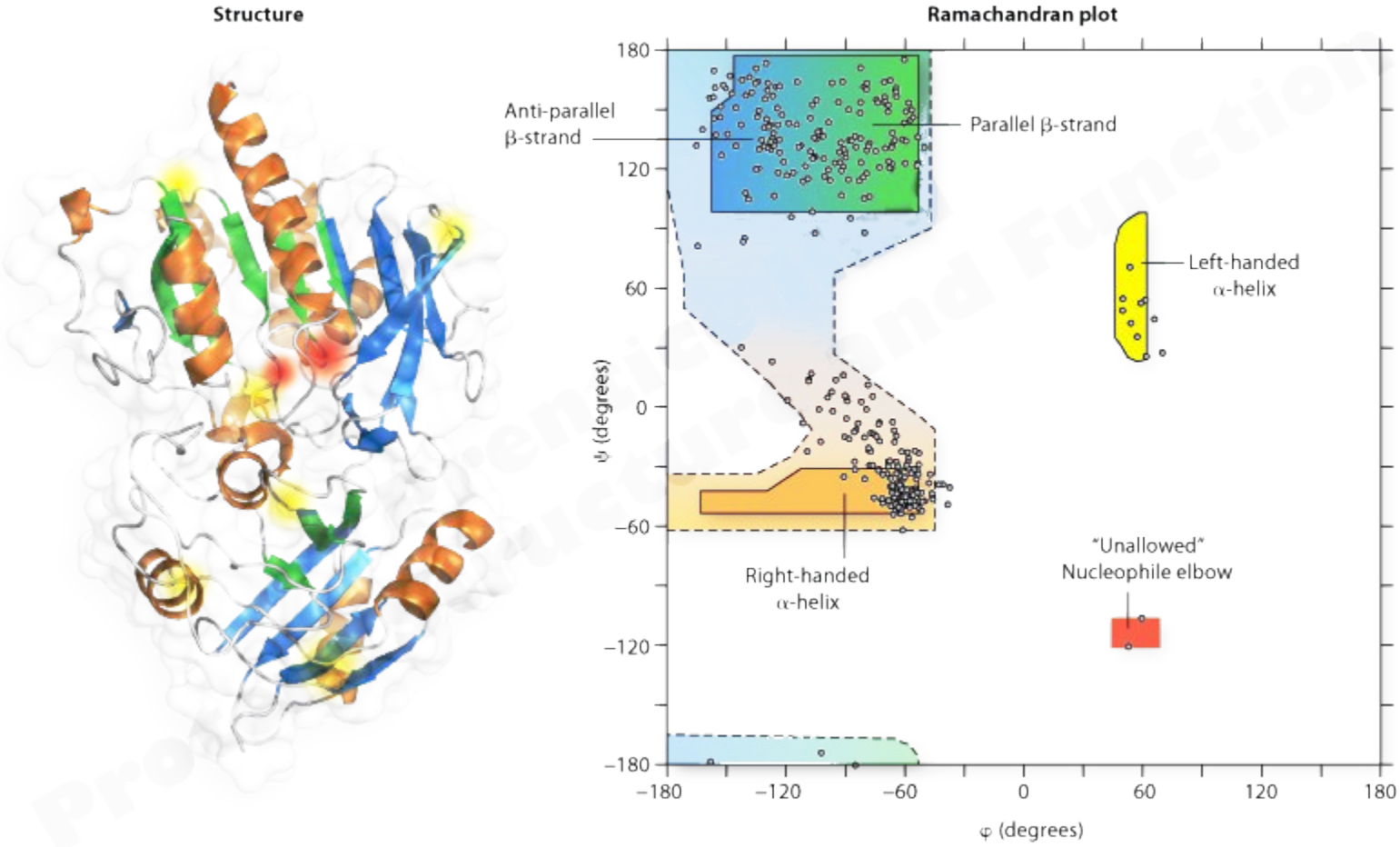
Las proteínas para funcionar se pliegan (la mayoría)




(a) denatured
(b) intermediate
(c) major transition
(d) native


C
O
N
Amide plane
C
H
Ψ
α-Carbon
H
ϕ
H
N
R
Side group
ω
C
O
α-Carbon
Amide plane
Peptide bond
Not free to rotate
(nearly always 180)




Structure
Ramachandran plot
Anti-parallel
β-strand
Parallel β-strand
Left-handed
α-helix
Right-handed
α-helix
"Unallowed"
Nucleophile elbow
180
120
60
0
−60
−120
−180
ψ (degrees)
−180
−120
−60
0
60
120
180
φ (degrees)

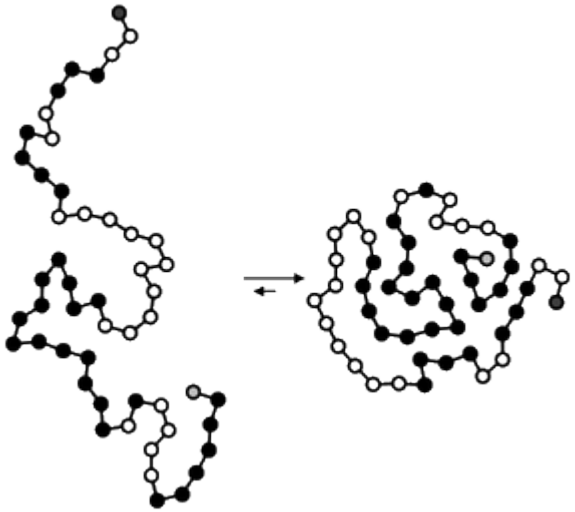
Secuenciación: estructura primaria

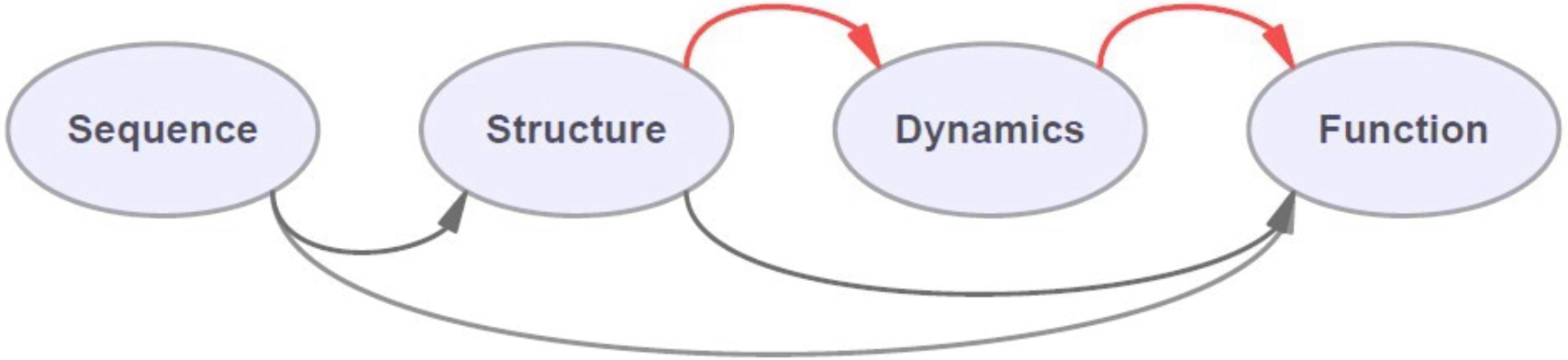
Sanger and Thompson, Biochem. J. 53:353-366. 1953

It would thus seem that no general conclusions can be drawn from these results concerning the general principles which govern the arrangement of the amino-acid residues in protein chains. In fact, it would seem more probable that there are no such principles, but that each protein has its own unique arrangement; an arrangement which endows it with its particular properties and specificities and fits it for the function that it performs in nature.

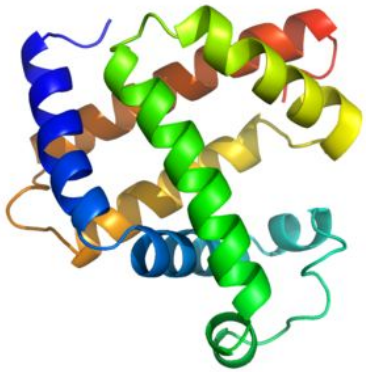
Frederick Sanger
1918-2013

Nobel Prize (Chemistry, 1958)
Nobel Prize (Chemistry, 1980)



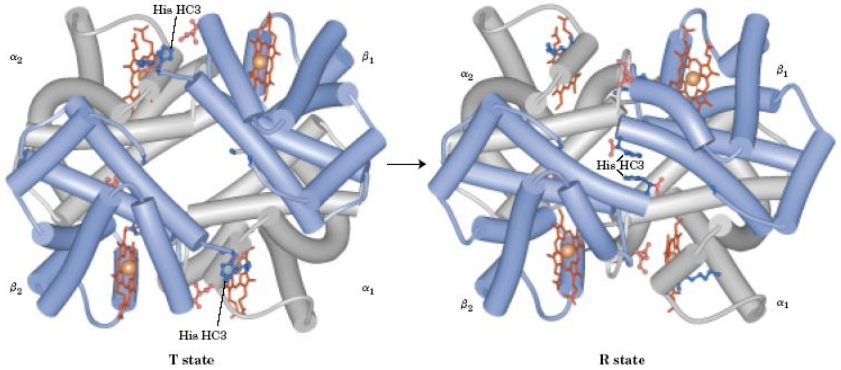

Sequence
Structure
Dynamics
Function

John Kendrew
1917-1997



Mioglobina

Max Perutz
1914-2002



Oxihemoglobina

Desoxihemoglobina

Teoría "Termodinámica"

El experimento de **Anfinsen** de 1973 muestra cómo la **información** estructural está contenida en la secuencia proteica

Desnaturalizando la *RNasa A* con *urea*, permitiendo la reducción de los puentes disulfuro con *b-mercaptoethanol (2 ME)*

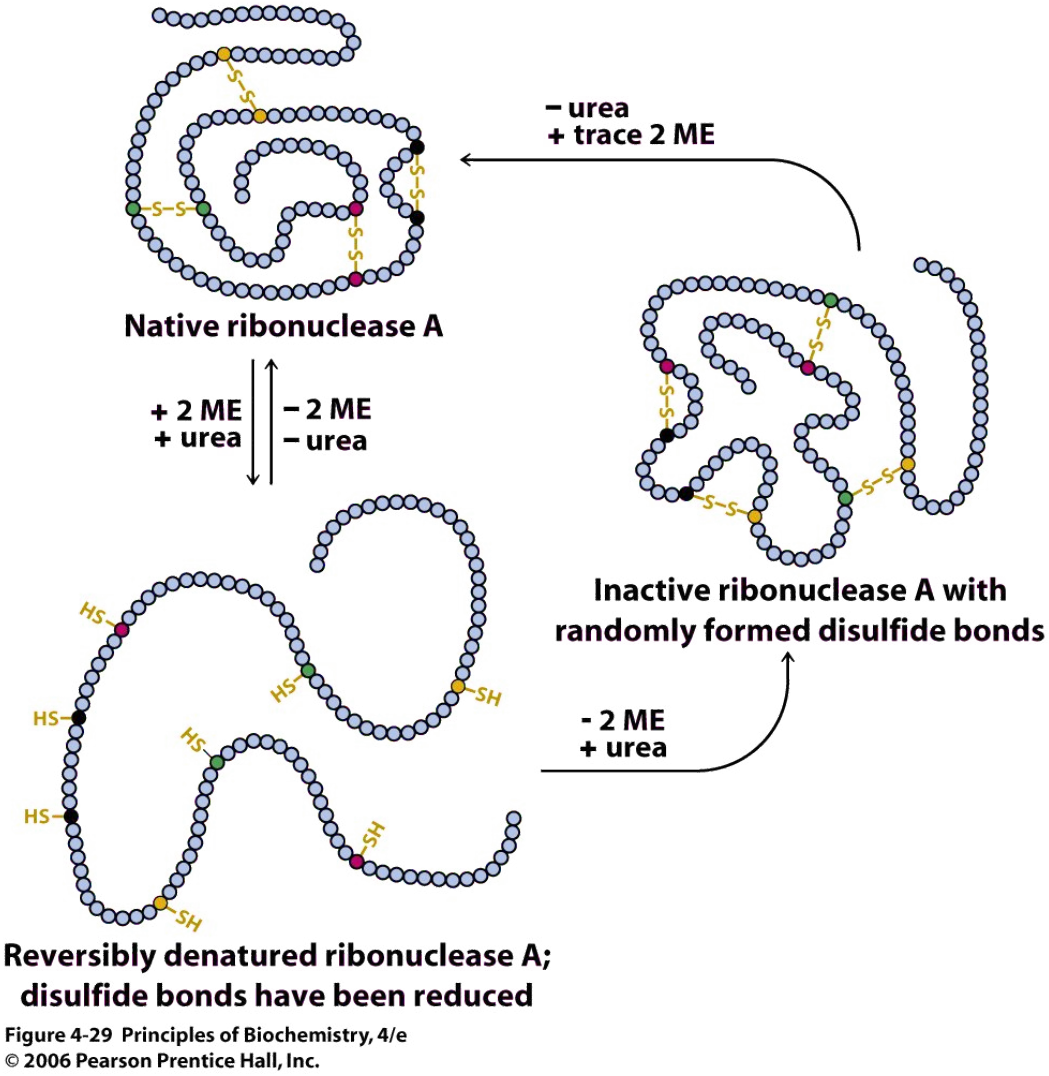


Figure 4-29 Principles of Biochemistry, 4/e

Teoría “Termodinámica”

Derivaciones del experimento de **Anfinsen**

1. Las proteínas se pueden plegar y desplegar repetidas veces
2. La “fuerza impulsora” es termodinámica
3. Toda la información para el plegado de una proteína está en su secuencia
4. El estado nativo de las proteínas está en un mínimo de energía

Teoría “Termodinámica”

Derivaciones del experimento de **Anfinsen**

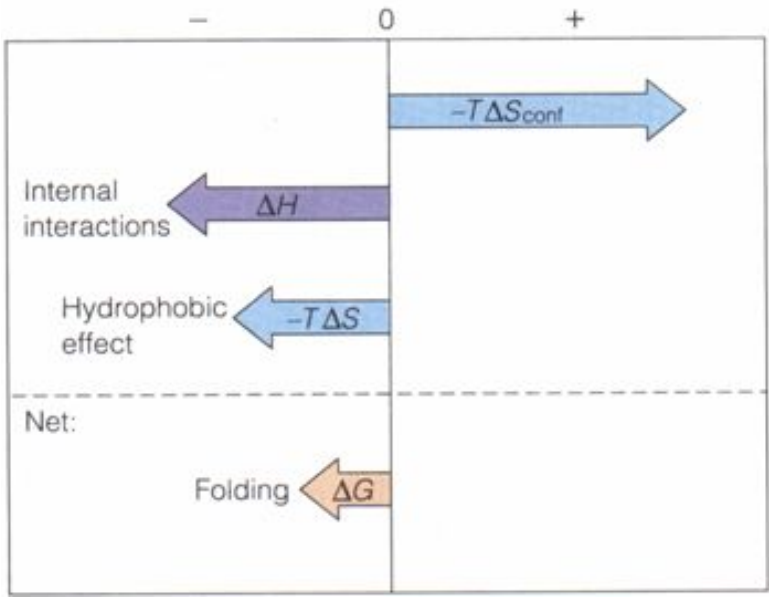
1. Las proteínas se pueden plegar y desplegar repetidas veces
2. **La “fuerza impulsora” es termodinámica**
3. Toda la información para el plegado de una proteína está en su secuencia
4. El estado nativo de las proteínas está en un mínimo de energía

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

H q-q, dipolo-dipolo, q-dipolo, q-dipolo inducido, dipolo-dipolo inducido, dipolo inducido-dipolo inducido, catión-π, π-π

S Efecto hidrofóbico

"Biology is dominated by the chemistry of the noncovalent bond"
Alan Fersht, 1992


–
0
+
$-T\Delta S_{cont}$
Internal interactions
ΔH
Hydrophobic effect
$-T\Delta S$
Net:
Folding
ΔG

Paradoja de Levinthal

Para describir la posición de un átomo en el espacio uno necesita definir **3 coordenadas**

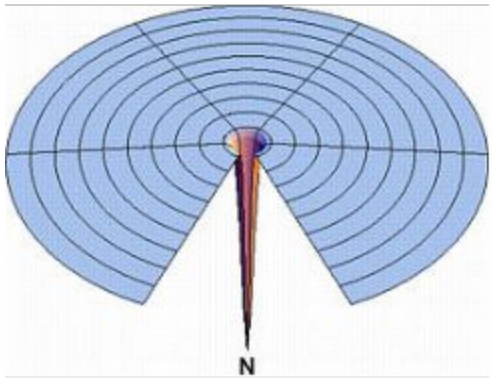
Para una proteína de **100 aminoácidos** necesitamos 6000 coordenadas (consideramos que un AA contiene en promedio 20 átomos)

Si cada una de esas coordenadas puede tomar al menos 2 valores distintos tendríamos que el **espacio conformacional de la proteína sería de 2^{6000}**

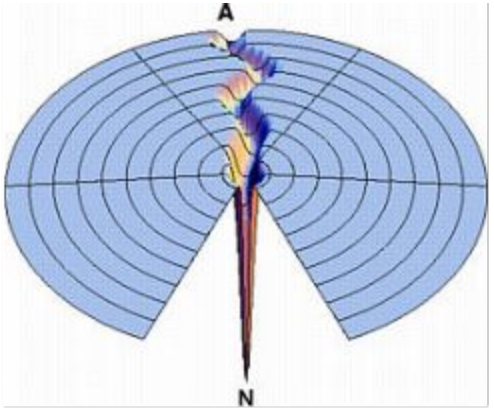
Si cada transición entre cada una de estas conformaciones distintas requiere de un tiempo de, por ejemplo, **110^{-13} segundos**...

Plegar una proteína requeriría **2^{5987} segundos!!!**

El problema del golfista ciego



Energy

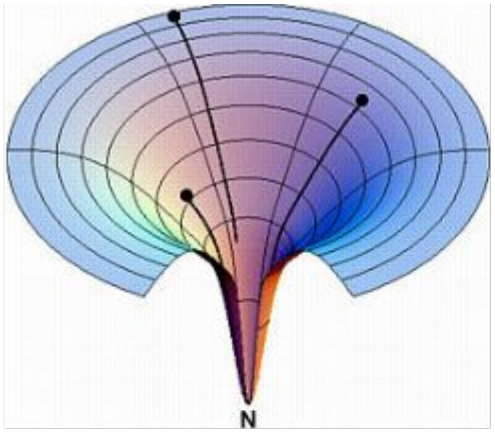


Grooved golf course

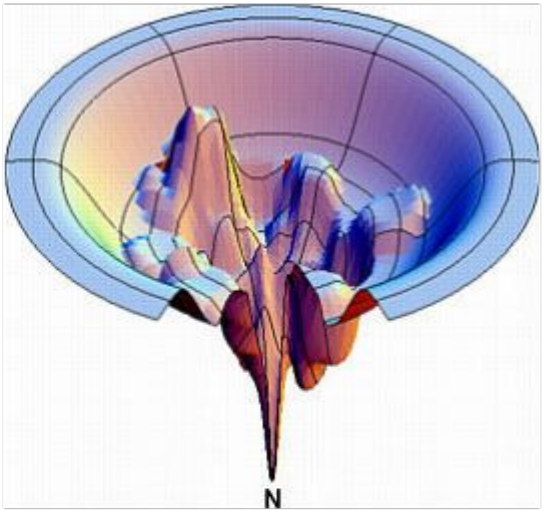
Paradoja de levinthal:
el campo de golf plano

Paradoja de levinthal:
El campo de golf con canales o caminos

El problema del golfista ciego



Solución del camino.
El fondo del embudo es el mínimo termodinámico.
Varios diferentes caminos cinéticos alcanzan el fondo



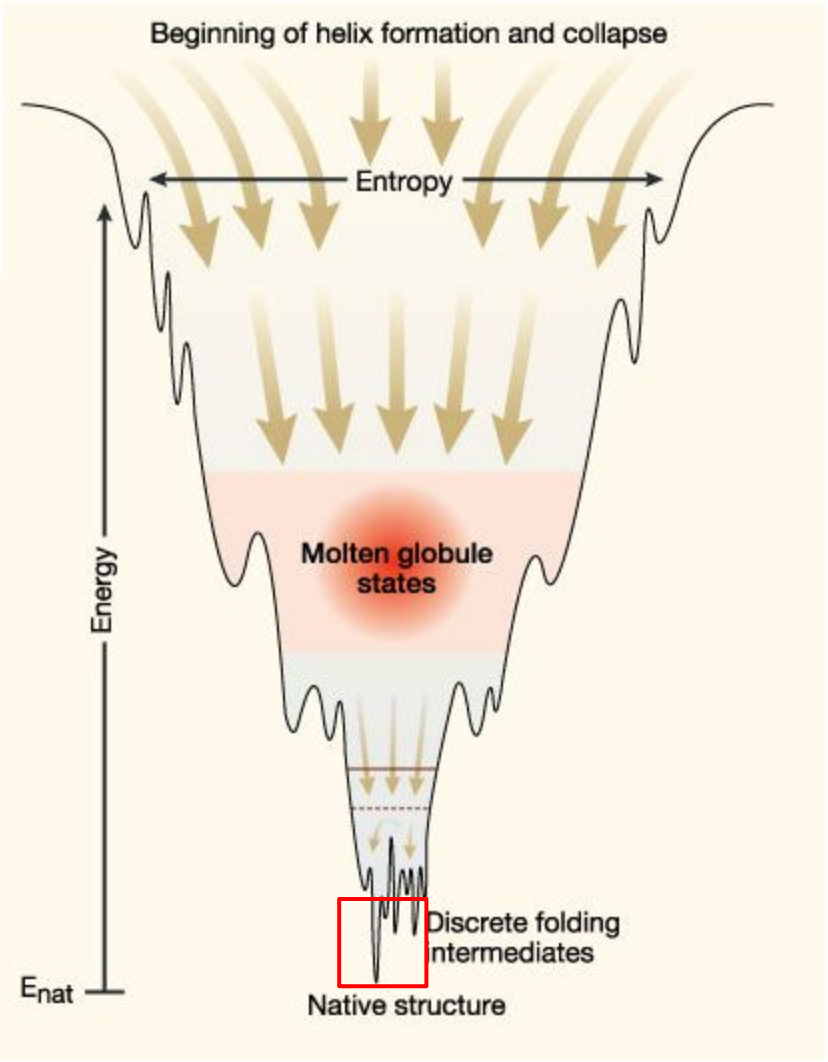
Más realístico, embudo áspero

THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective

José Nelson Onuchic
Department of Physics, University of California at San Diego, La Jolla, California 92093-0319

Zaida Luthey-Schulten and Peter G. Wolynes
School of Chemical Sciences, University of Illinois, Urbana, Illinois 61801

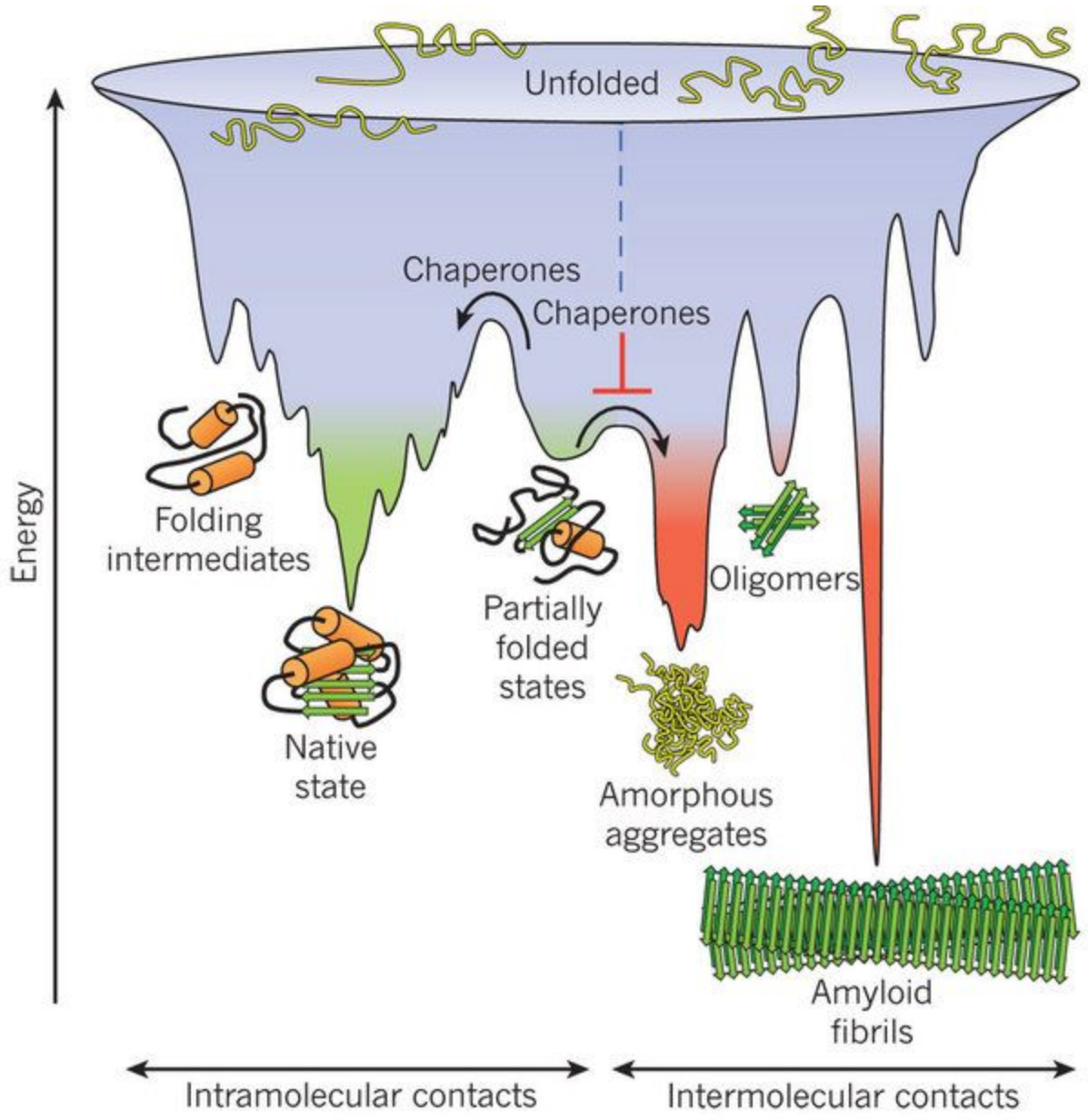
El espacio conformacional se reduce cada vez que se produce un buen contacto o lo que se llama un contacto “nativo”

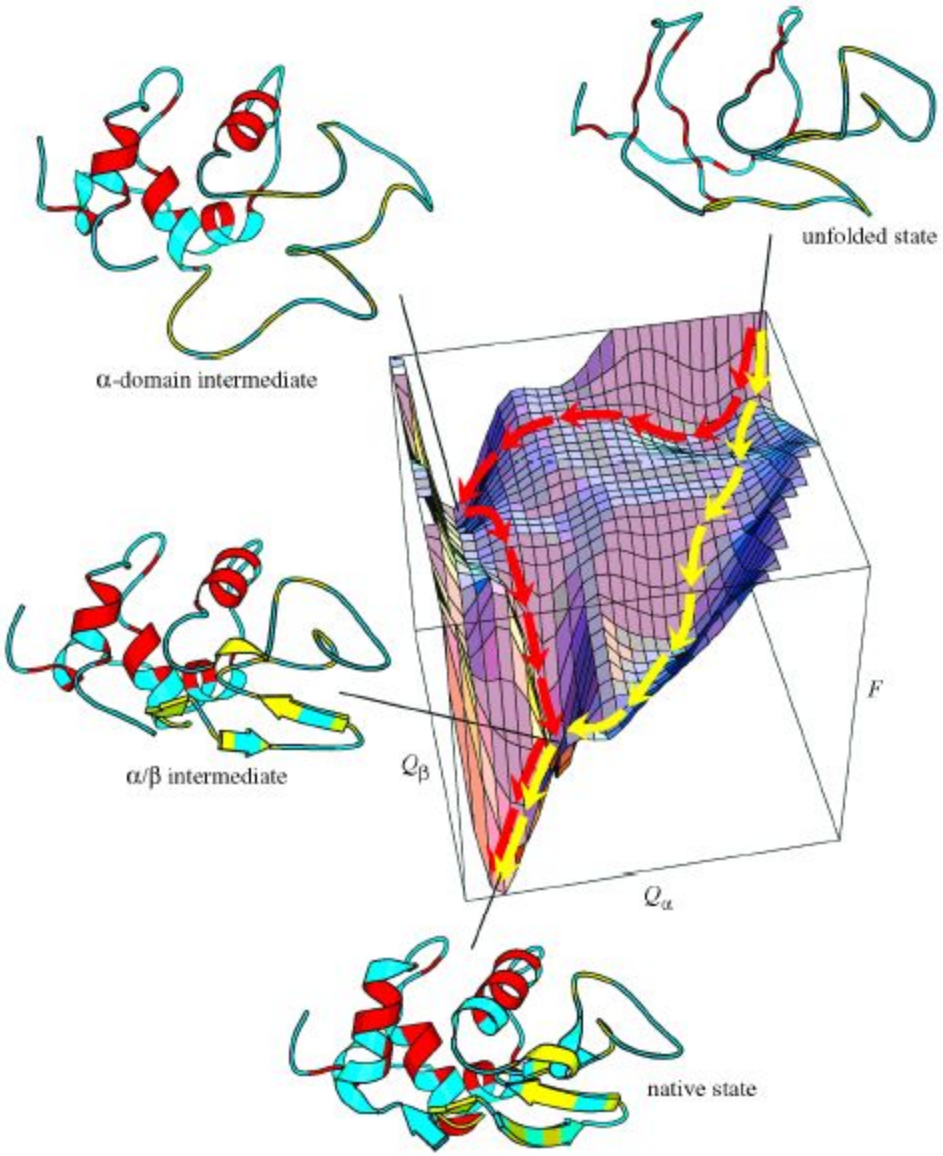


Teoría “Termodinámica”

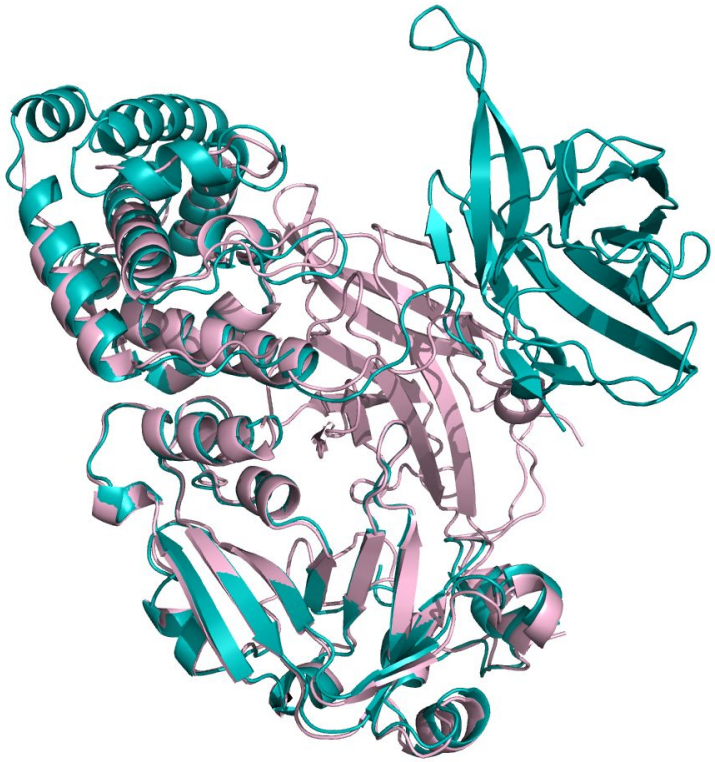
Derivaciones del experimento de **Anfinsen**

1. Las proteínas se pueden plegar y desplegar repetidas veces
2. La “fuerza impulsora” es termodinámica
3. **Toda la información para el plegado de una proteína está en su secuencia**
4. **El estado nativo de las proteínas está en un mínimo de energía**

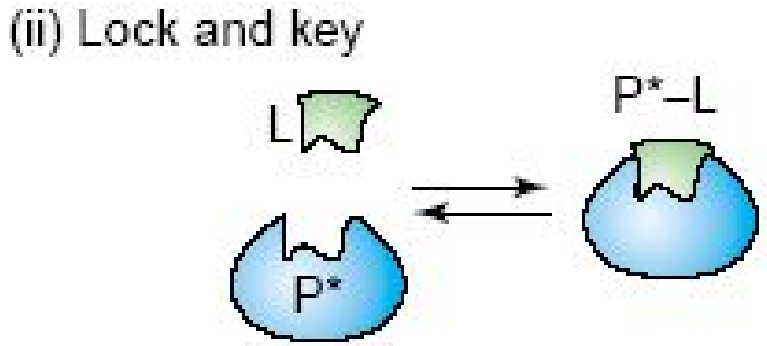

Unfolded
Chaperones
Chaperones
Energy
Folding intermediates
Native state
Partially folded states
Oligomers
Amorphous aggregates
Amyloid fibrils
Intramolecular contacts
Intermolecular contacts


unfolded state
α-domain intermediate
α/β intermediate
native state
Q_β
Q_α
F

Relación estructura-función



Modelo "Llave-cerradura" (1894)



H. E. Fischer

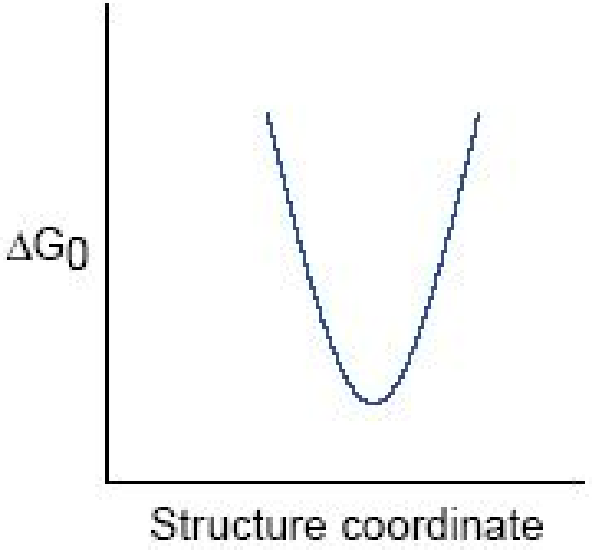
The restricted effects of the enzymes may therefore be explained by the assumption that the approach of the molecules that cause the chemical process can occur only in the case of a similar geometric shape.

To use a picture, I would like to say that enzyme and glucoside have to fit to each other like a lock and key in order to exert a chemical effect on each other.

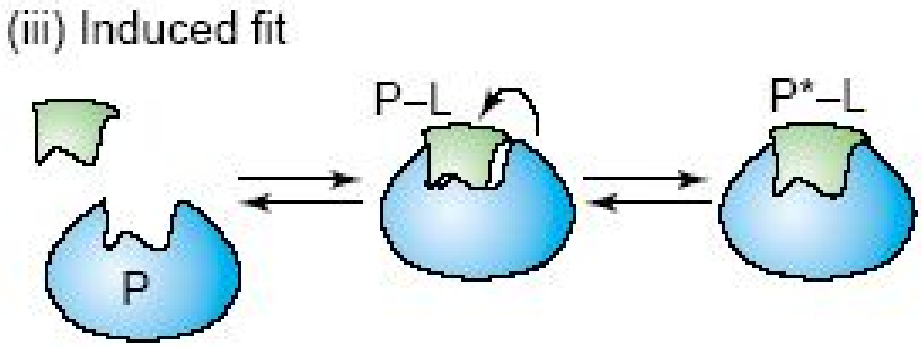
E. Fischer, *Ber.*, 27 (1894) 3189 – 3232.

(a) 'Simplistic view'

(i) Smooth energy landscape

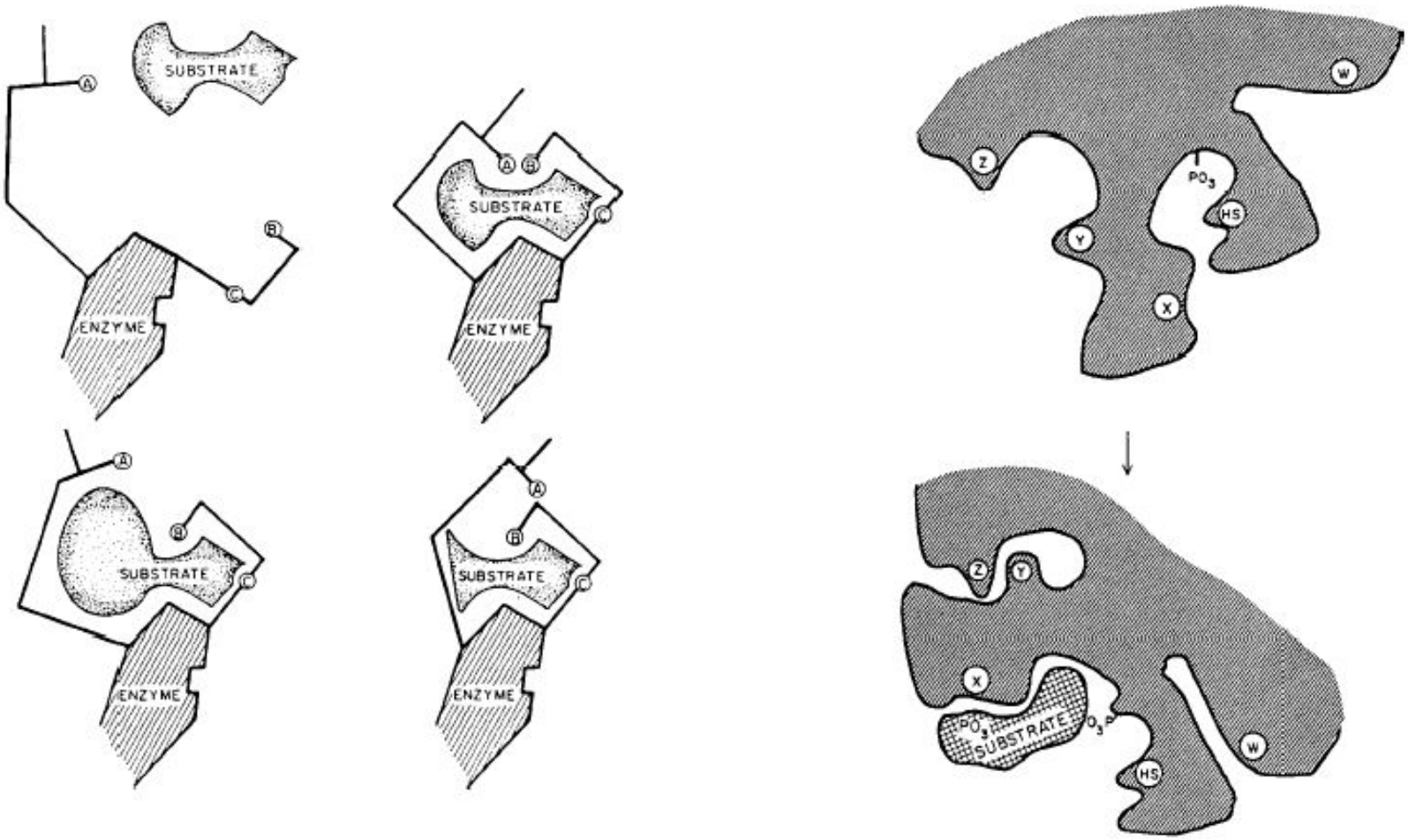


Modelo del Ajuste Inducido (1958)

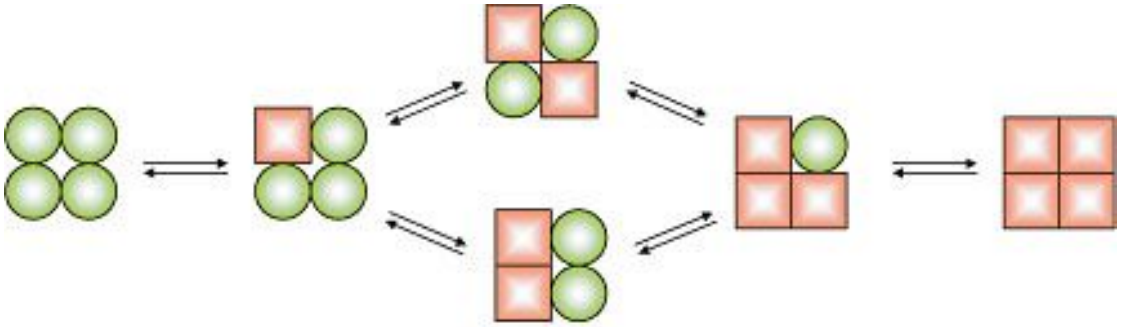


Daniel E. Koshland, Jr
1921-

Modelo del Ajuste Inducido (1958)

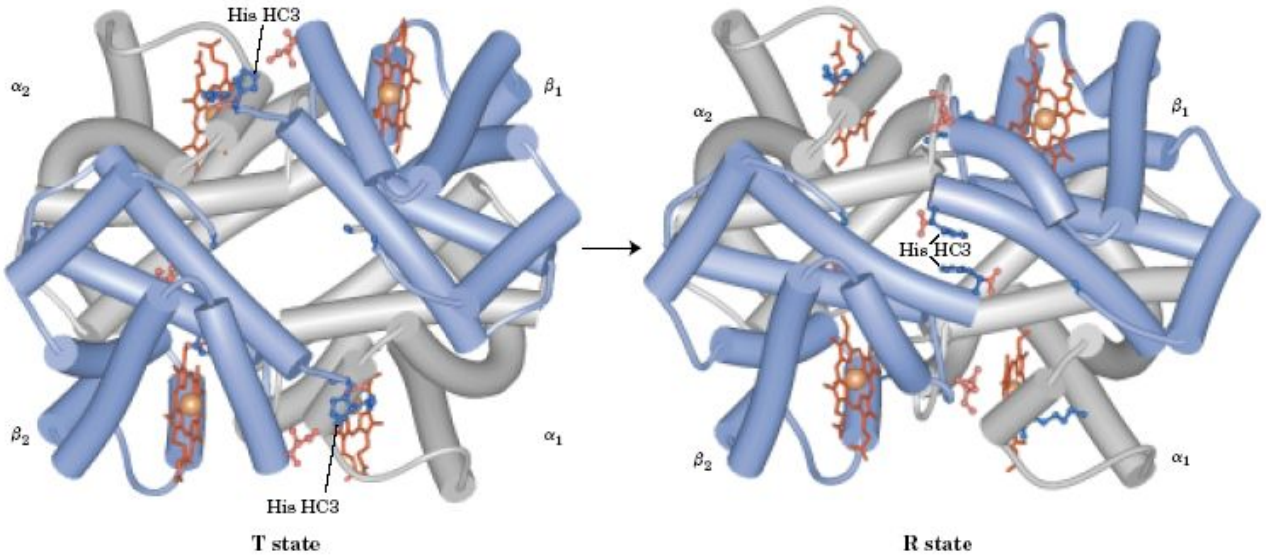


Este modelo de regulación alostérica de enzimas sugiere que las subunidades de proteínas multiméricas tienen dos estados conformacionales. La unión del ligando provoca el cambio conformacional. Aunque las subunidades pasan a través de cambios conformacionales de forma independiente

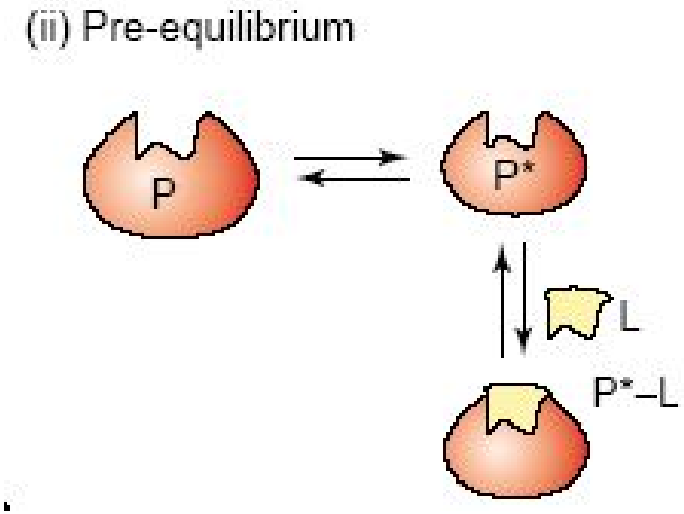


Koshland-Némethy-Filmer's Square model

Koshland, Nemethy and Filmer (KNF)

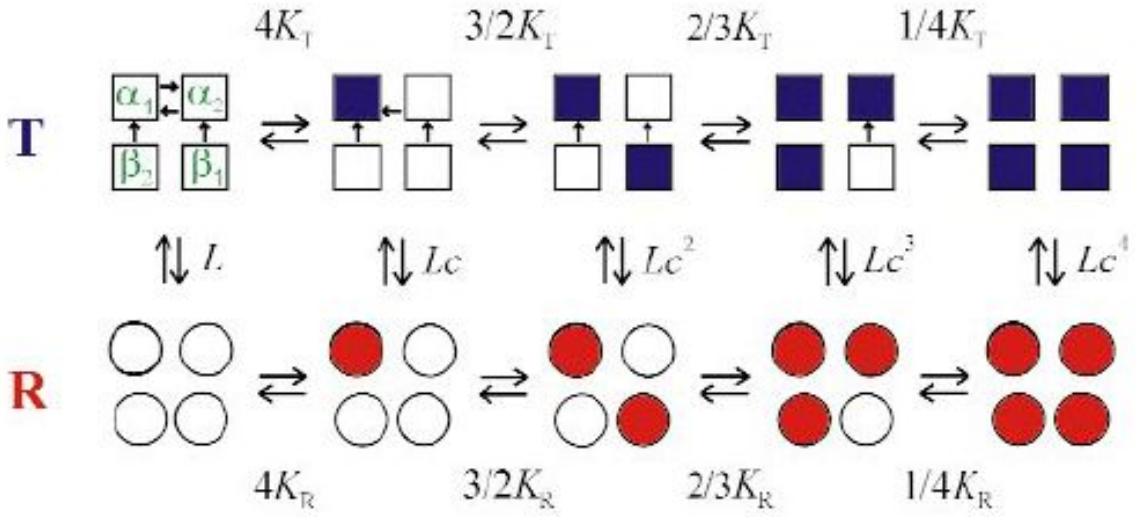

His HC3
α_2
β_1
β_2
α_1
His HC3
T state
α_2
β_1
His HC3
β_2
α_1
R state

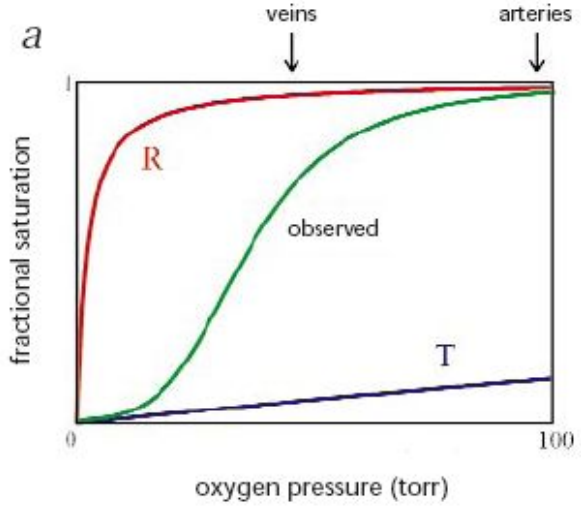
Diversidad conformacional en el estado nativo



Monod, Wyman and Changeux, 1965

Jacques Monod
1910-1976





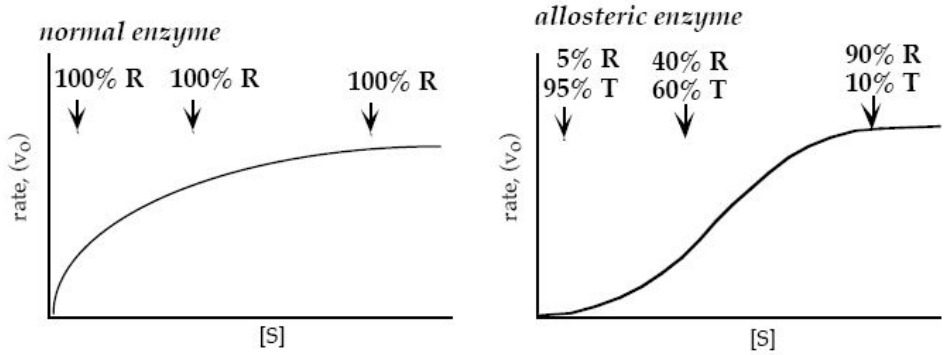
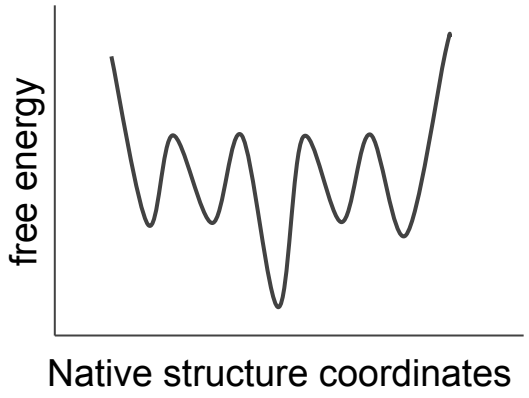
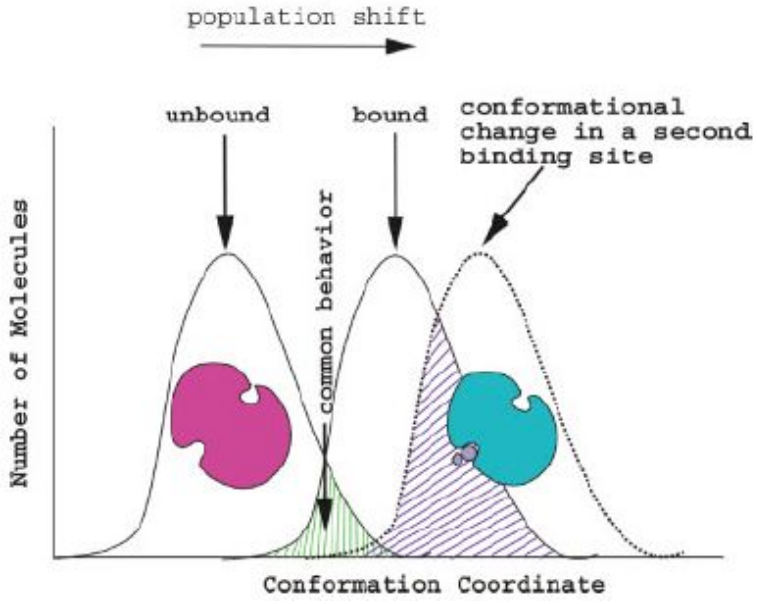


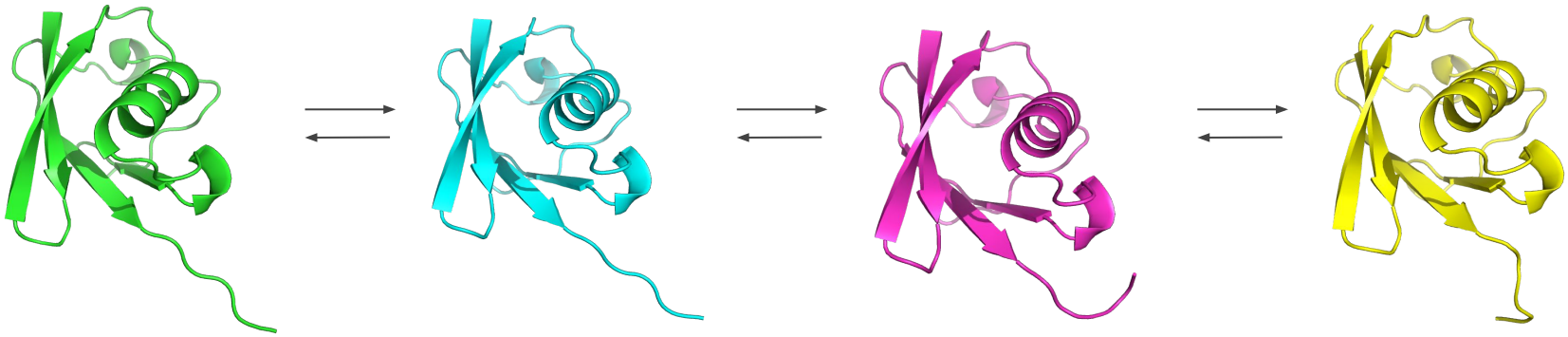
Fig. 18 Per cent of enzyme in the active form.

‘New view’

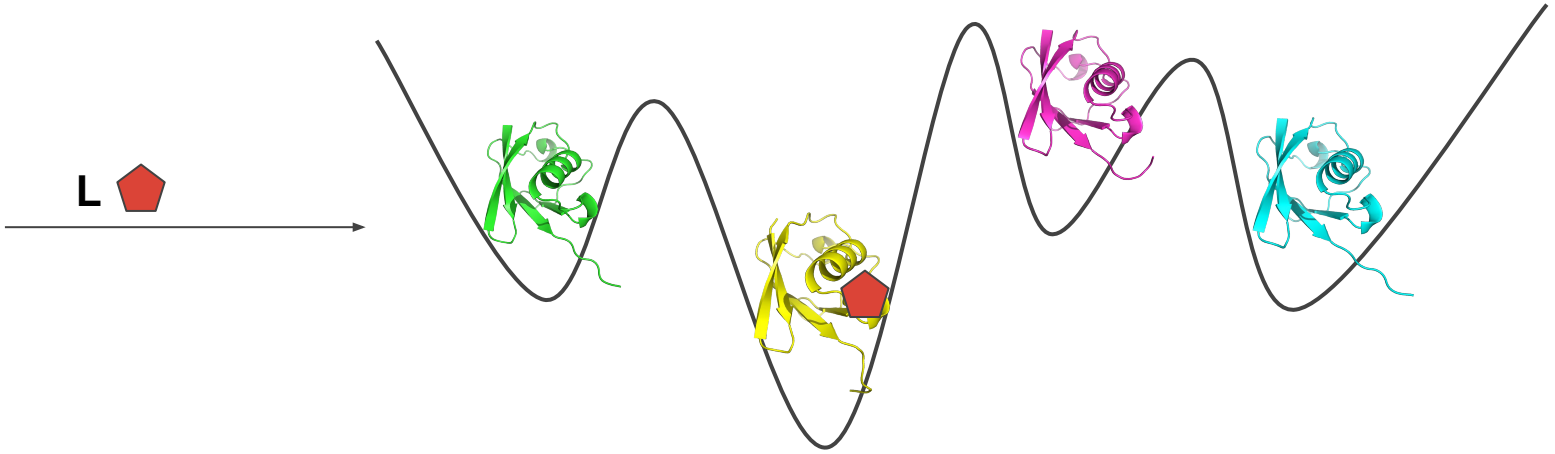

free energy
Native structure coordinates


population shift
unbound
bound
conformational change in a second binding site
common behavior
Number of Molecules
Conformation Coordinate

La diversidad conformacional hace referencia a que los confórmeros de una proteína se encuentran en un pre-equilibrio dinámico

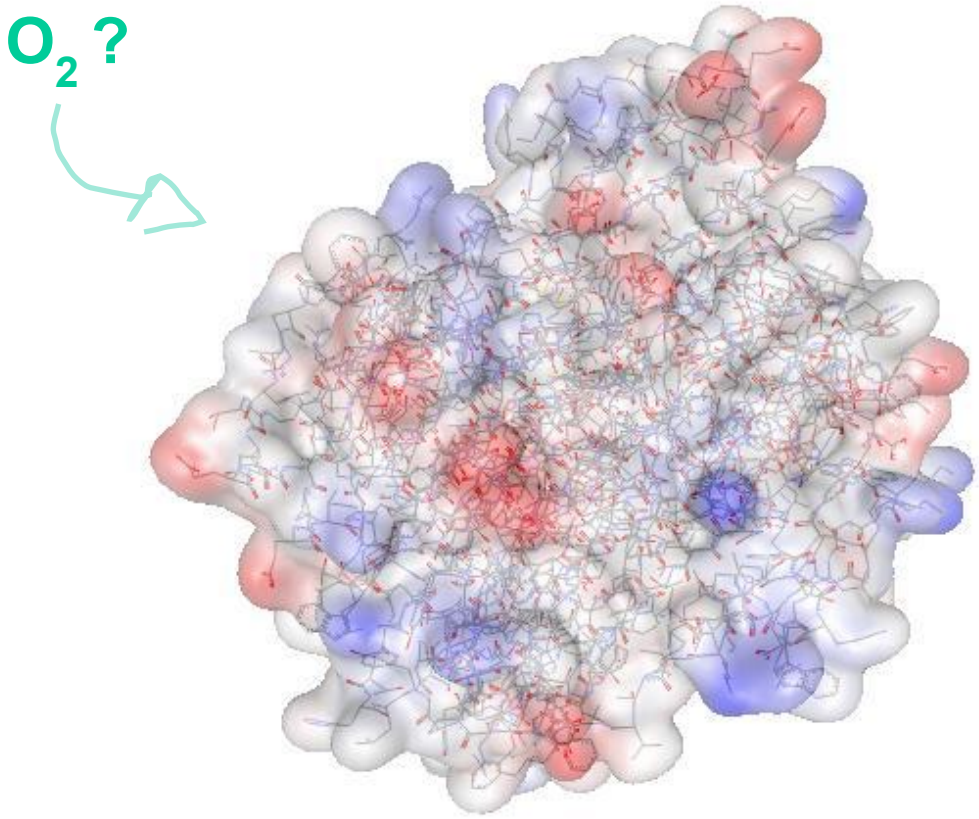


La presencia del ligando cambia el equilibrio hacia la conformación más favorable

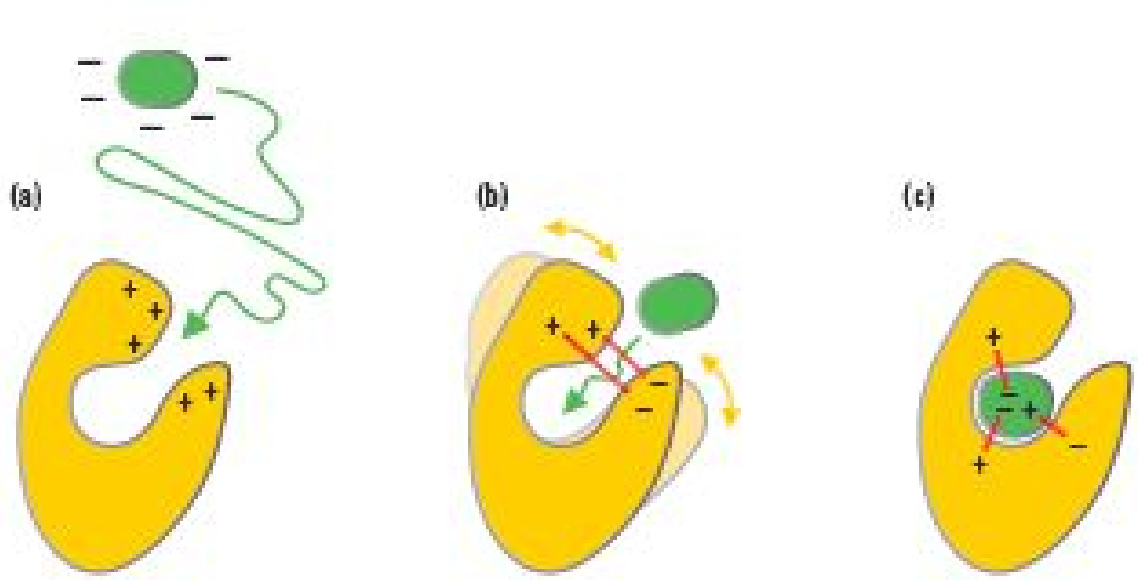


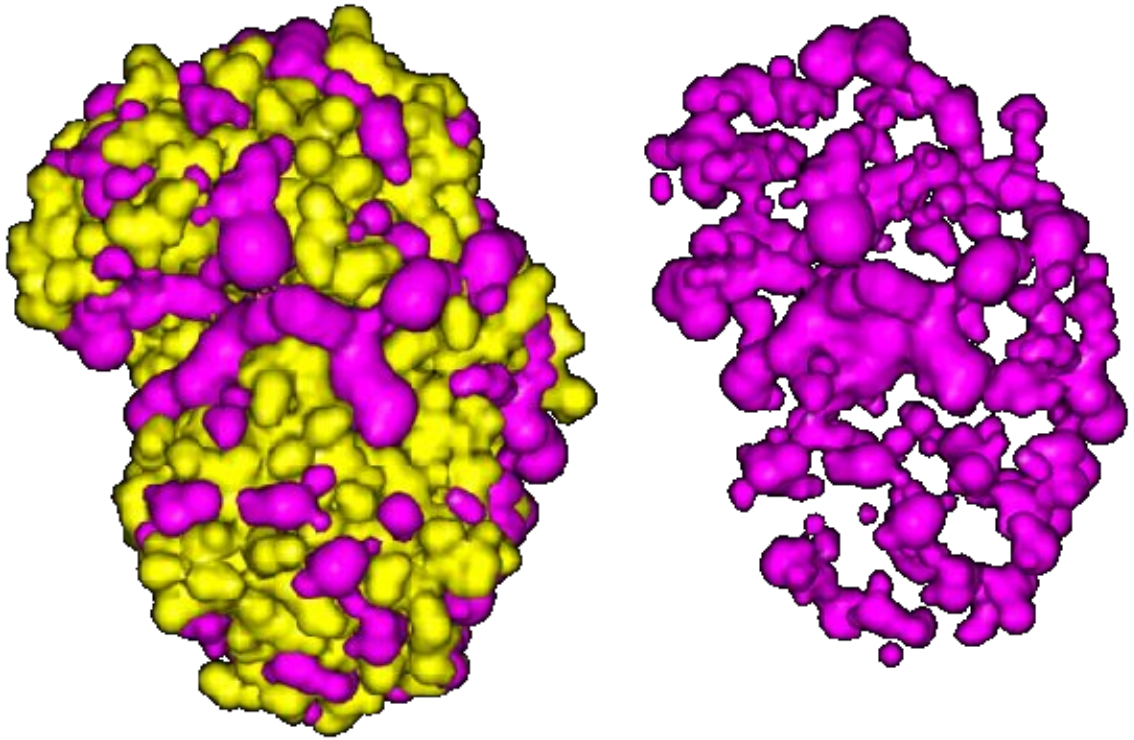
Conformational selection and population-shift

Superficie



Túneles, Cavidades





Estado Nativo

•Las proteínas existen como un *ensamble* de distintas conformaciones o estructuras

• La existencia de esta multiplicidad conformacional es esencial para la actividad/función de las proteínas

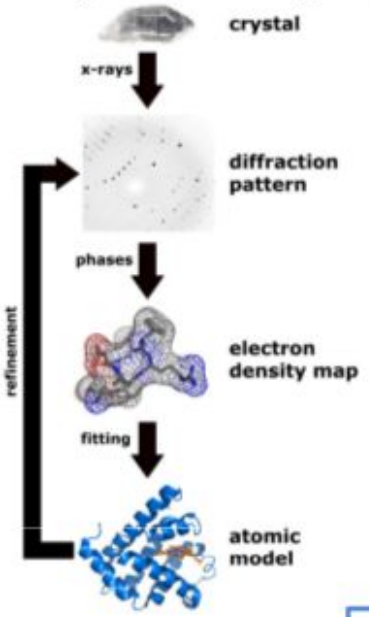
• El estado nativo no es único

• El alosterismo no es una propiedad exclusiva de “enzimas” sino que se aplica a casi cualquier proteína globular

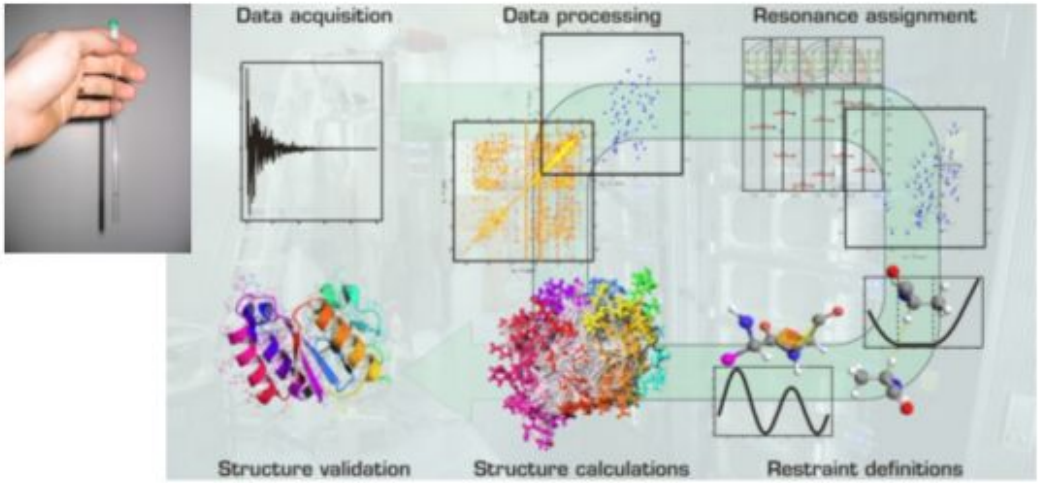
• La función de una proteína no depende solamente de los residuos en el “sitio activo”, sino también de residuos próximos y alejados del mismo

Determinación experimental

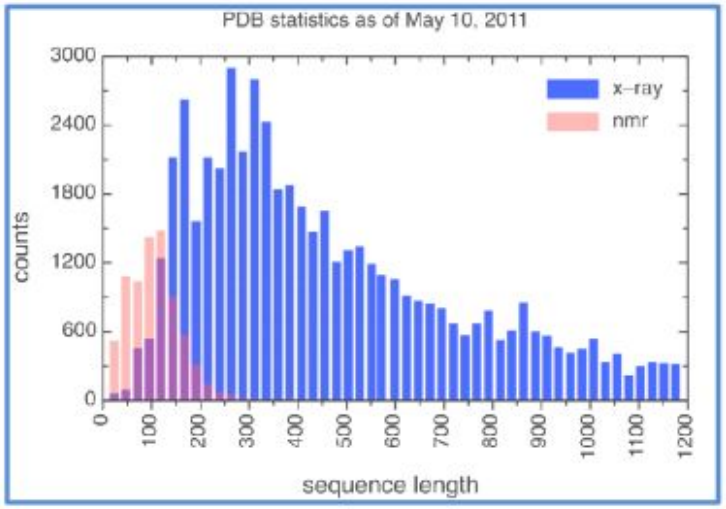
X-Ray Crystallography



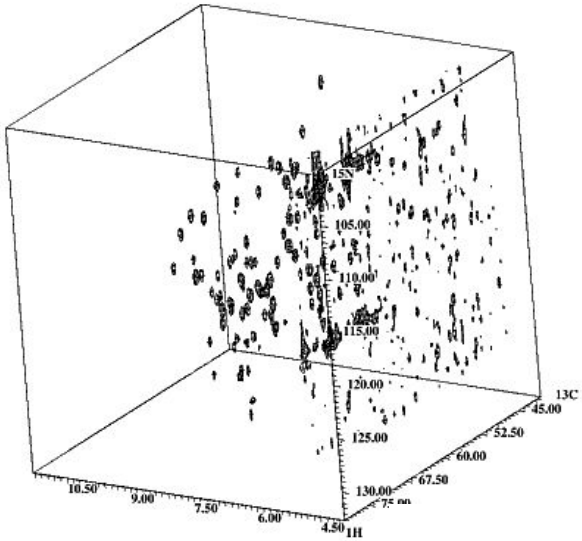
NMR Spectroscopy



- Distances
- Dihedrals
- ...



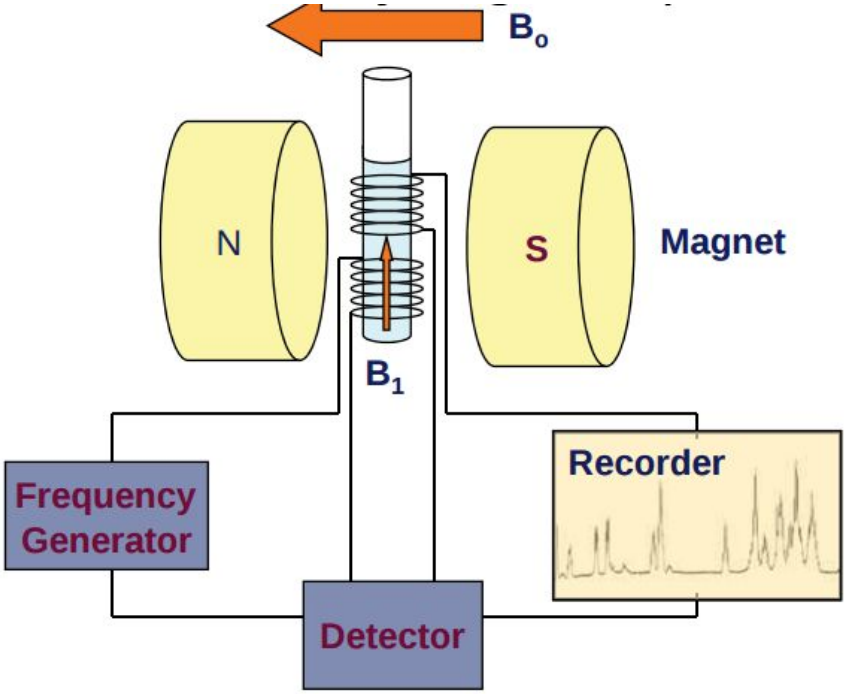
RMN (resonancia magnética nuclear)



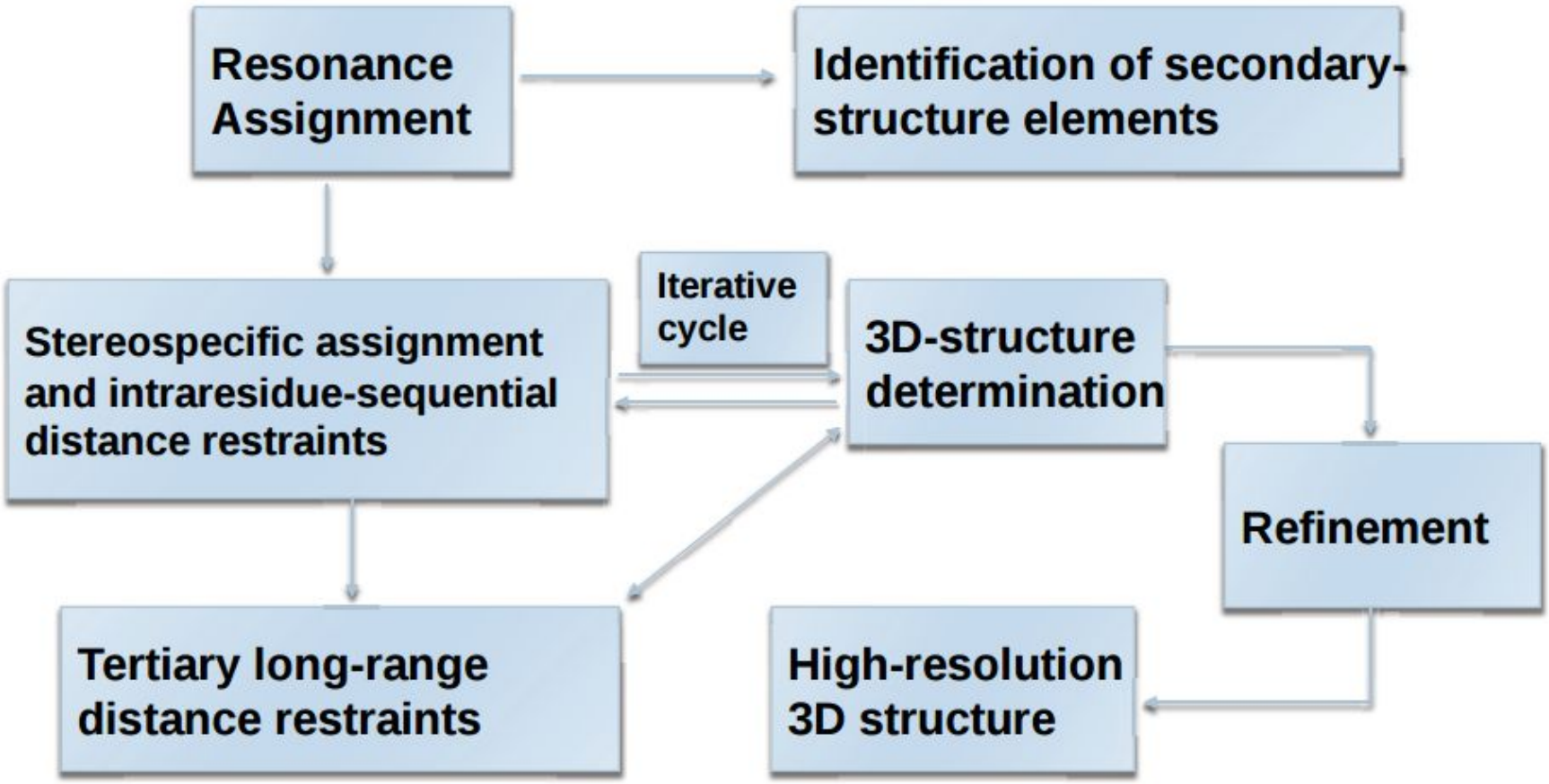
900Mhz

RMN (resonancia magnética nuclear)

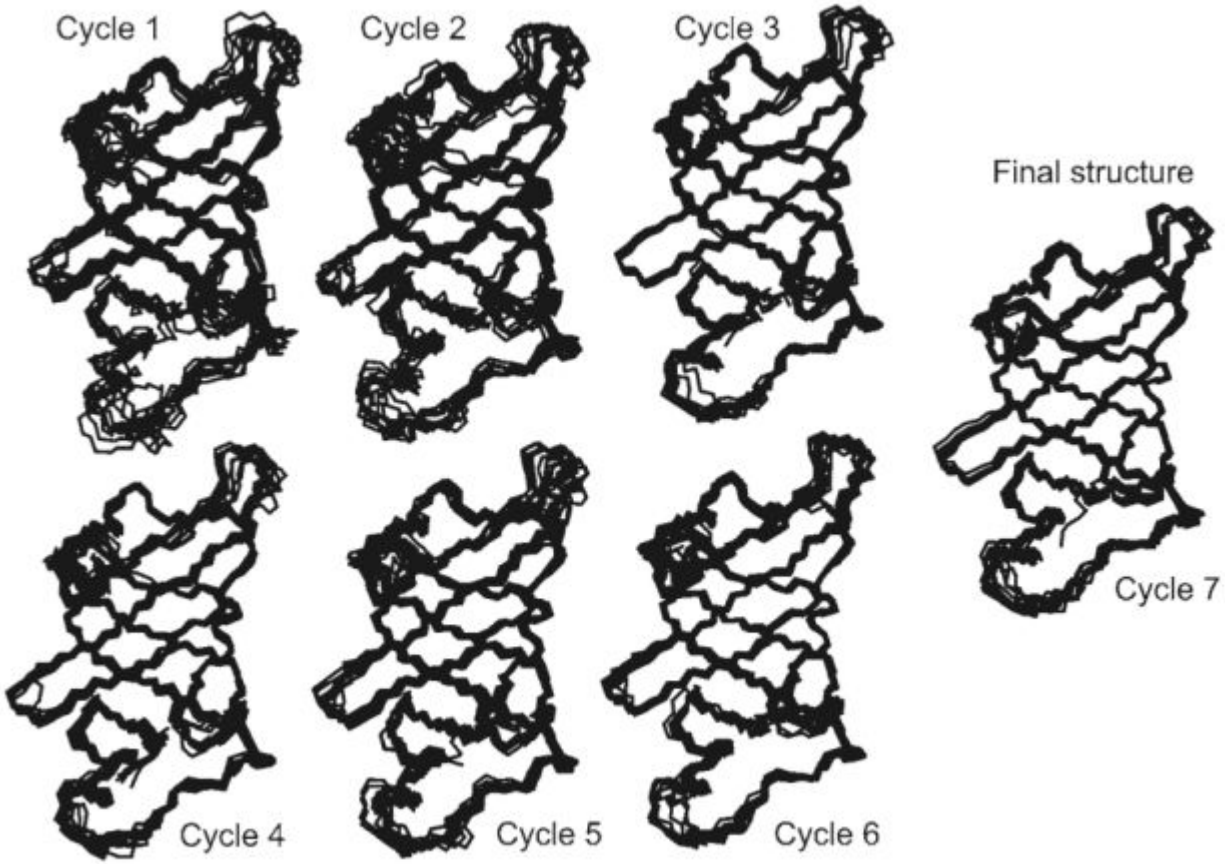
- En NMR se miden transiciones de spin nuclear
- Las diferencias de energía son bajas
  - Bajar relación señal-ruido.
  - Alta precisión
- En proteínas trabajamos con núcleos de spin : 1H (natural), 13C y 15N (precisan enriquecimiento isotópico).
- Cada núcleo resuena a una frecuencia particular.



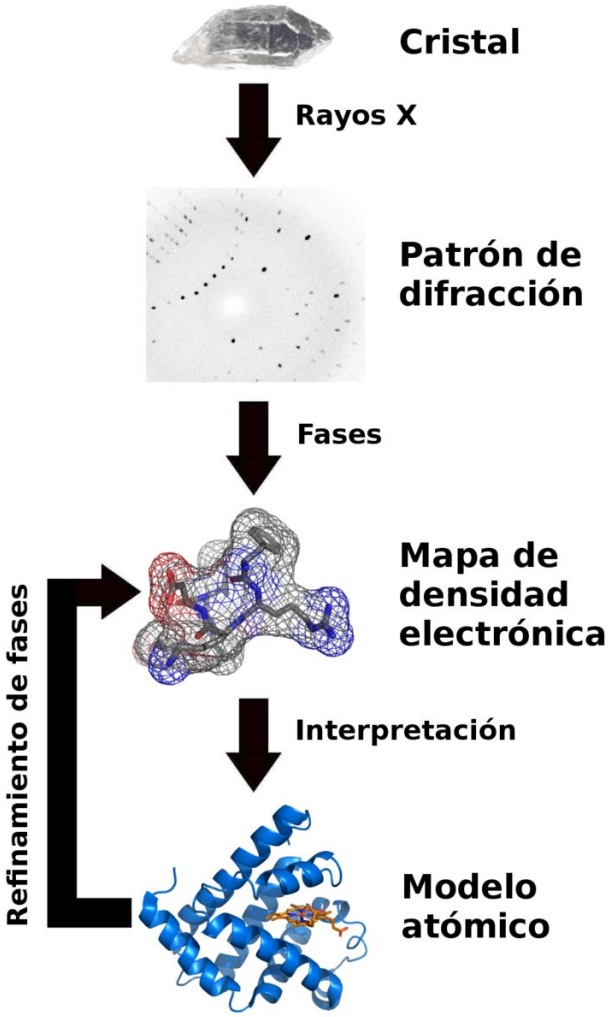
RMN (resonancia magnética nuclear)

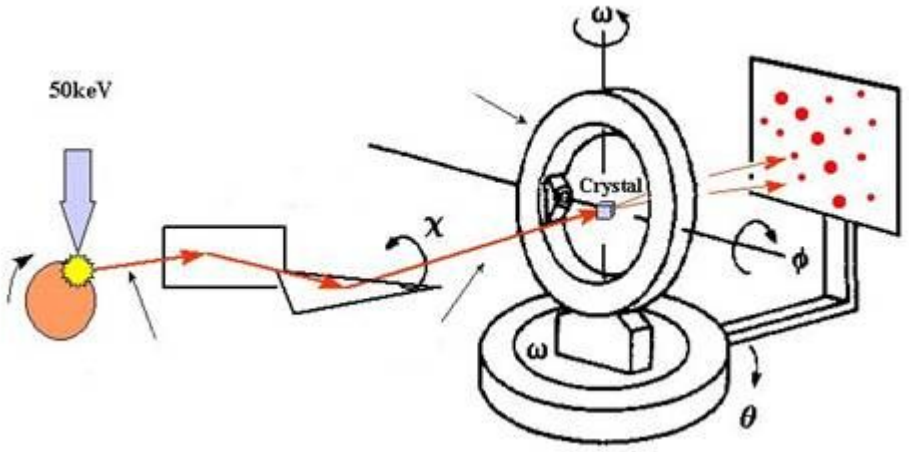


RMN (resonancia magnética nuclear)

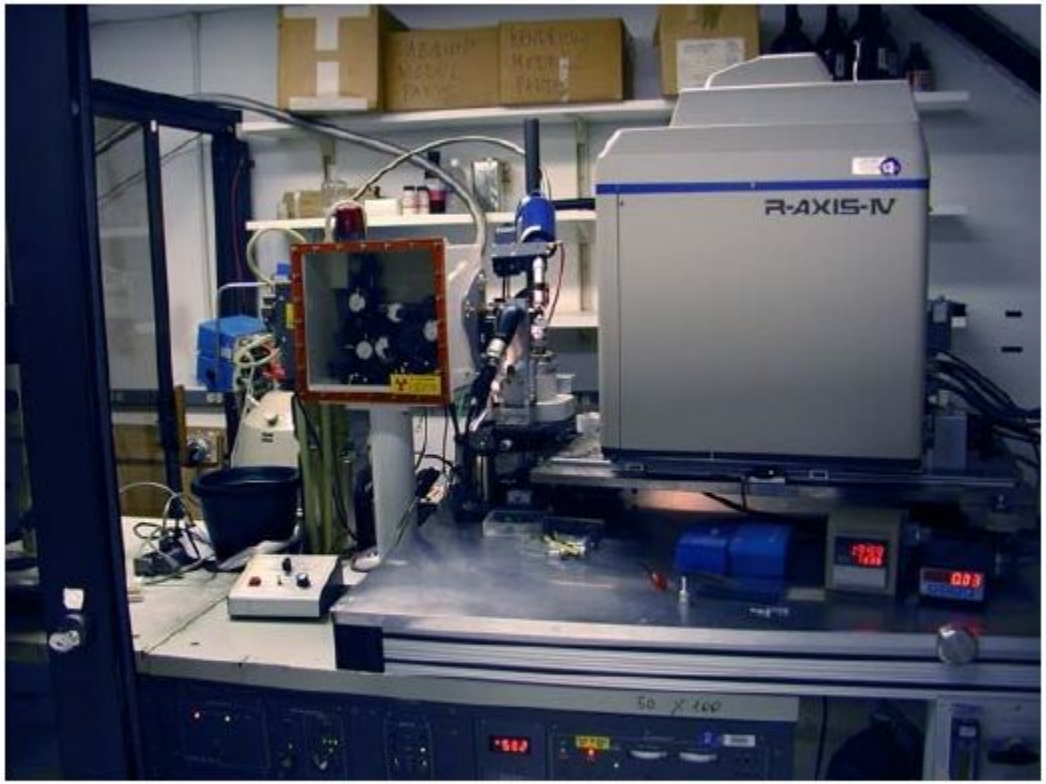


Difracción por rayos-X

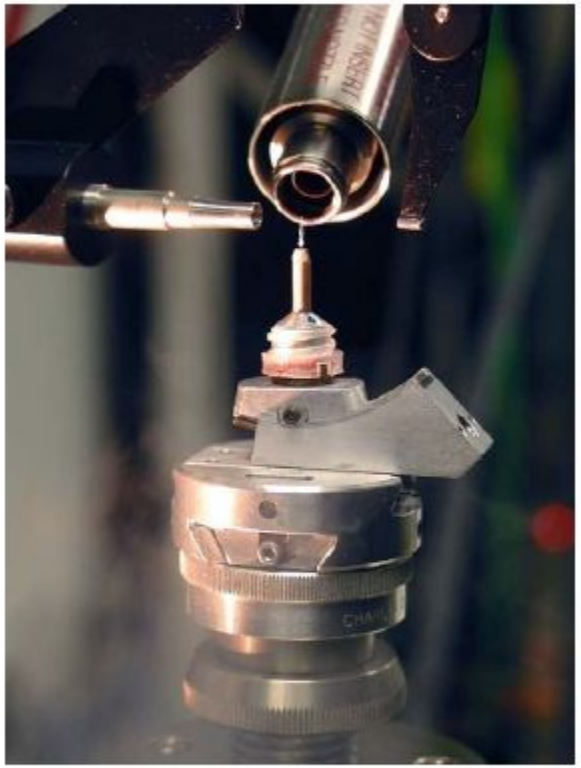




Difracción por rayos-X

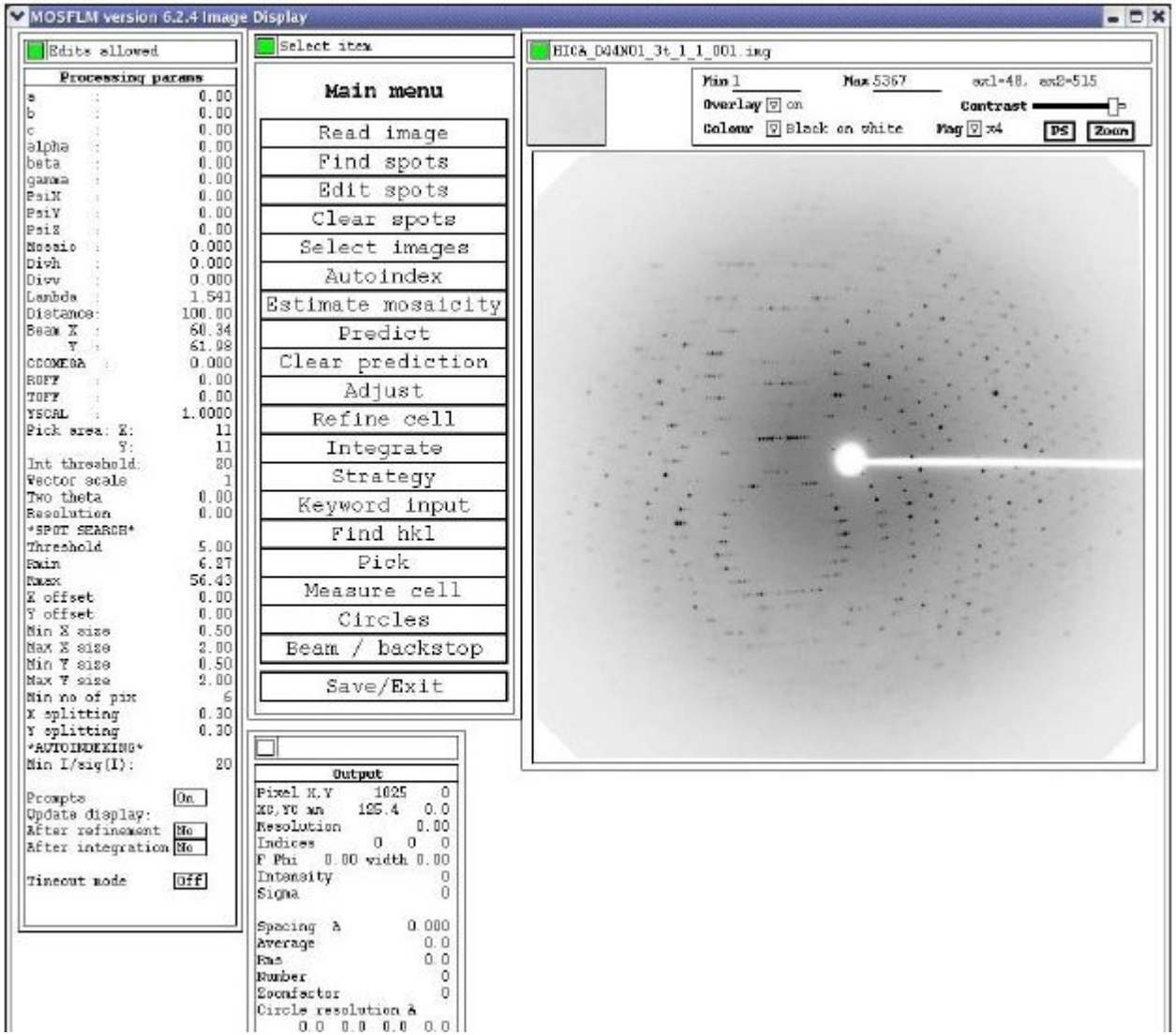


Equipo para emisión de rayos X sobre cristales de proteína

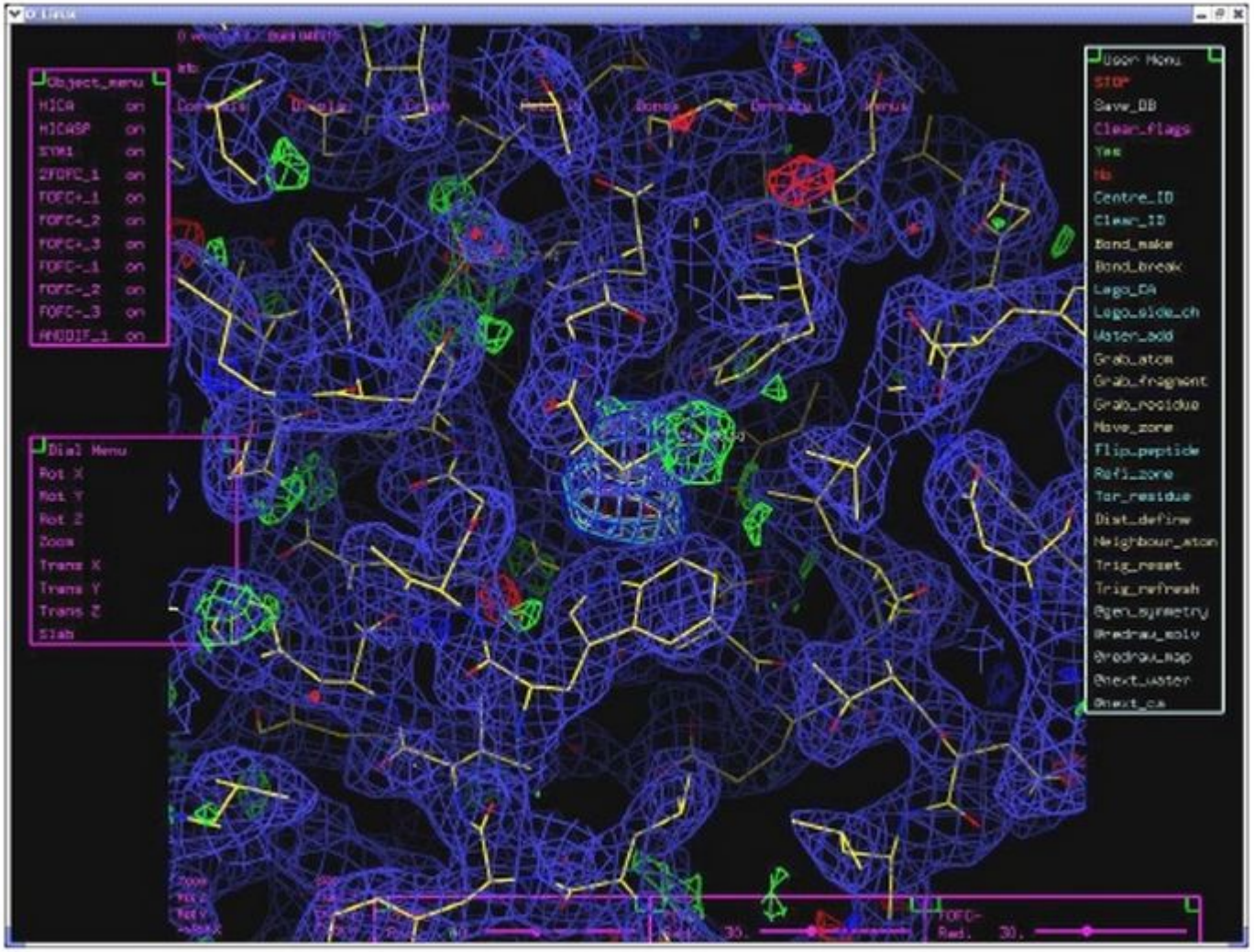


Cristal de proteína montado en el goniómetro (mide ángulos) y enfriado con el Cryostream

Difracción por rayos-X



Difracción por rayos-X



Difracción por rayos-X

- Se debe elegir un método para resolver el PROBLEMA DE FASE:
  - Métodos de Patterson
  - Métodos directos
  - Dispersión anómala
  - Reemplazamiento isomorfo
  - Reemplazamiento molecular
- El objetivo es obtener el mapa de densidad electrónica.
- La densidad se relaciona con las ondas de difracción mediante la fórmula de Transformadas de Fourier.

$$\rho(x,y,z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}|\mathbf{F}_{hkl}|\exp(-2\pi i(xh+yk+zl))$$

El módulo del factor de estructura es simplemente la raíz cuadrada de la intensidad del punto de difracción

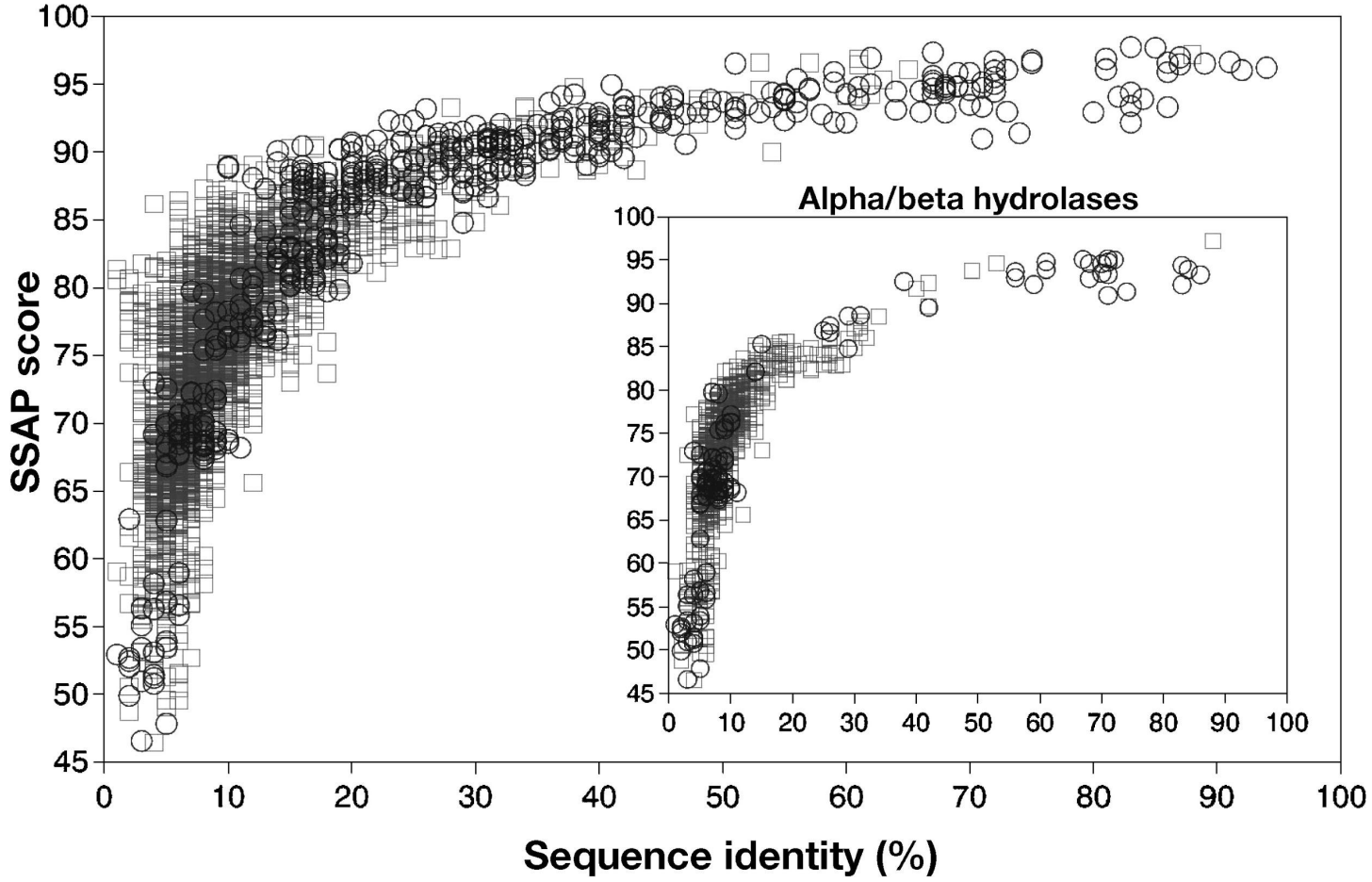
Clasificación estructural de proteínas

Annu. Rev. Biochem. 2005. 74:867–900
doi: 10.1146/annurev.biochem.74.082803.133029

PROTEIN FAMILIES AND THEIR EVOLUTION—A STRUCTURAL PERSPECTIVE

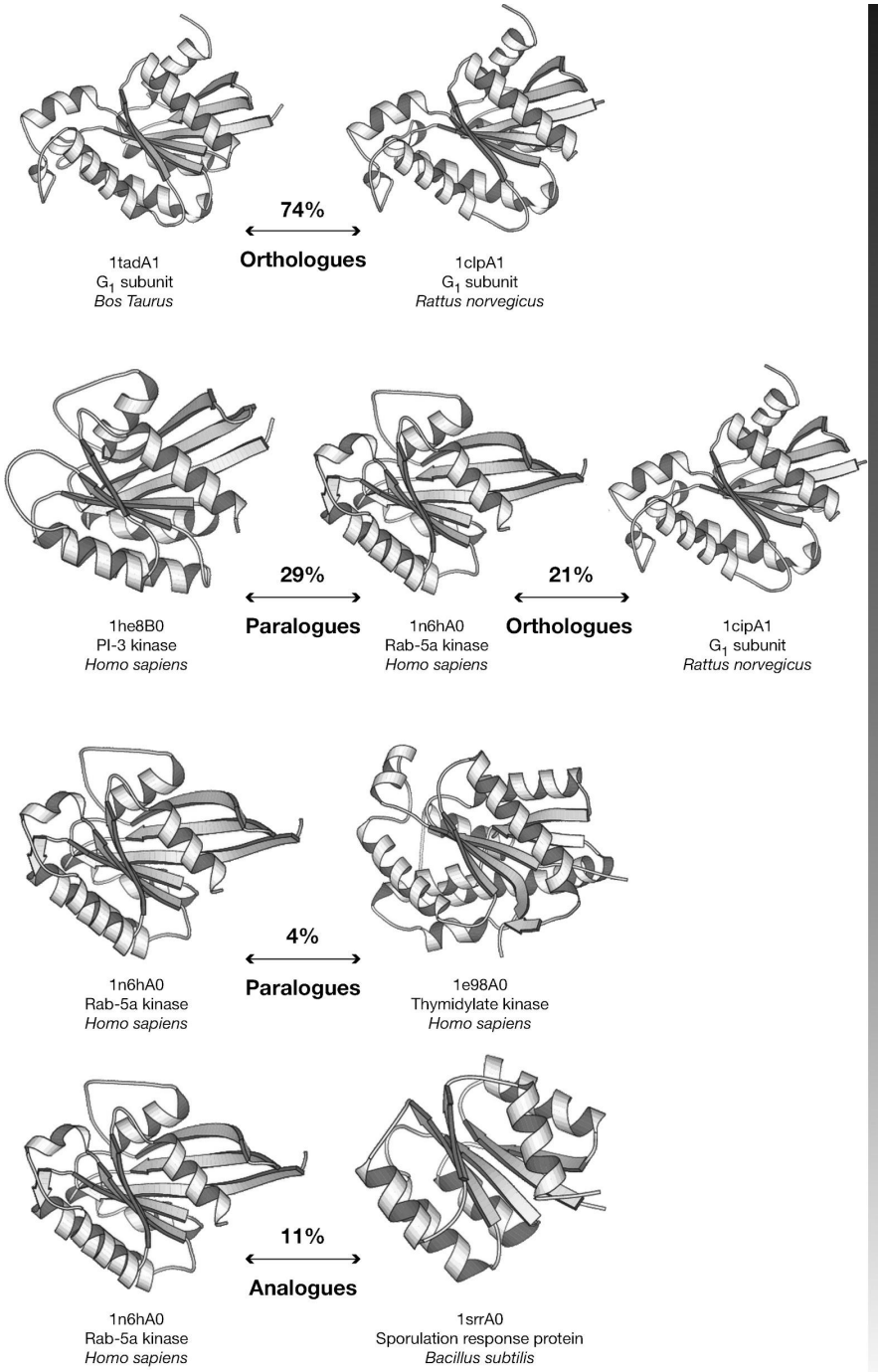
Christine A. Orengo[1] and Janet M. Thornton[2]
[1]*Department of Biochemistry and Molecular Biology, University College, London WC1E 6BT, United Kingdom; email: orengo@biochemistry.ucl.ac.uk*
[2]*European Bioinformatics Institute, Hinxton Campus, Cambridge CB10 1SD, United Kingdom; email: thornton@ebi.ac.uk*

Key Words protein classifications, comparative genomics, bioinformatics



Correlation between structure similarity (measured by the SSAP structure comparison algorithm, 0–100) and sequence similarity (measured by sequence identity) for pairs of homologous domain structures in the CATH domain database.
Homologous proteins possessing the same function are labeled as circles. Squares indicate homologous relatives with different functions.

Schematic representation of the progression from close homologues, through more remote (*twilight zone*) and very remote (*midnight zone*) homologues and finally analogous structural relatives.



Close Homologues

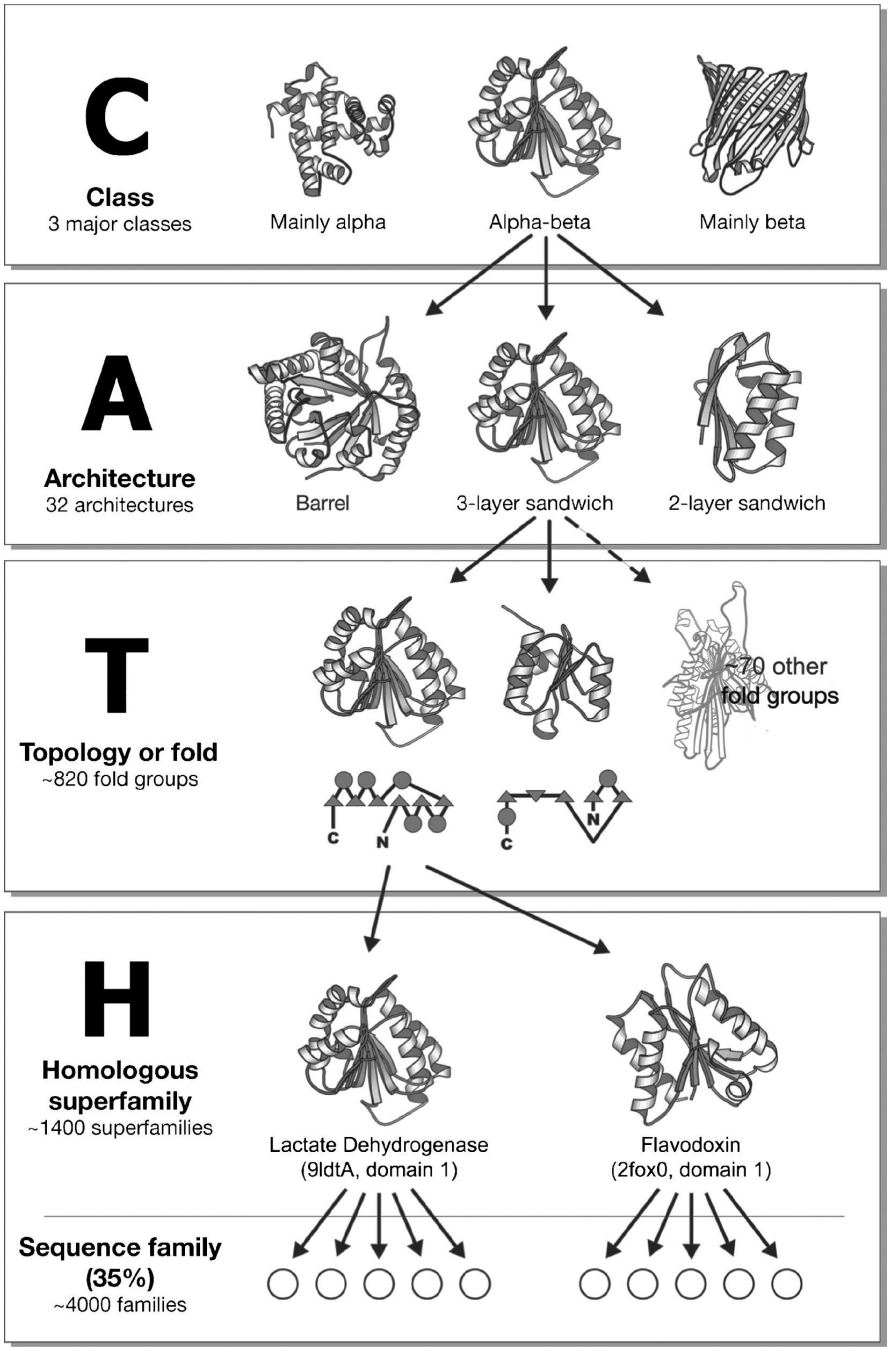
Remote Homologues

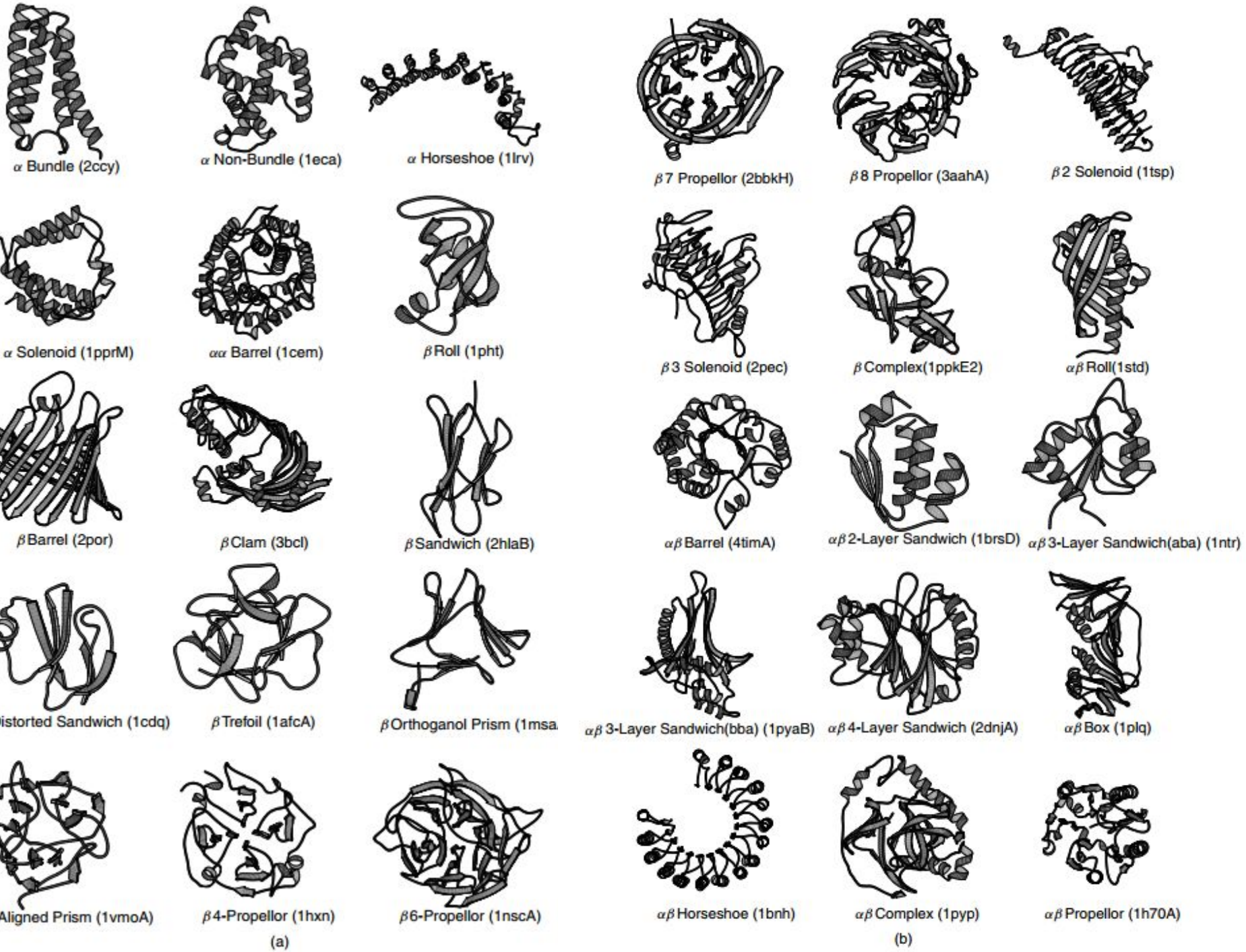
Distant Homologues

Analogues

CATH

The four major hierarchical levels in the **CATH** structural classification: **c**lass, **a**rchitecture, **t**opology or fold level, and **h**omologous superfamily. Three of the most highly populated architectures in the classification are illustrated.




α Bundle (2ccy)
α Non-Bundle (1eca)
α Horseshoe (1lrv)
β7 Propellor (2bbkH)
β8 Propellor (3aahA)
β2 Solenoid (1tsp)
α Solenoid (1pprM)
αα Barrel (1cem)
β Roll (1pht)
β3 Solenoid (2pec)
β Complex(1ppkE2)
αβ Roll(1std)
β Barrel (2por)
β Clam (3bcl)
β Sandwich (2hlaB)
αβ Barrel (4timA)
αβ 2-Layer Sandwich (1brsD)
αβ 3-Layer Sandwich(aba) (1ntr)
β Distorted Sandwich (1cdq)
β Trefoil (1afcA)
β Orthoganol Prism (1msa
αβ 3-Layer Sandwich(bba) (1pyaB)
αβ 4-Layer Sandwich (2dnjA)
αβ Box (1plq)
β Aligned Prism (1vmoA)
β4-Propellor (1hxn)
β6-Propellor (1nscA)
αβ Horseshoe (1bnh)
αβ Complex (1pyp)
αβ Propellor (1h70A)
(a)
(b)

CATH

Table 7.3 Description of each level in hierarchical structural classifications

Level in hierarchy	Description
Protein class	The class of a protein structure reflects the proportion of α-helices or β-strands within the three-dimensional structure. The major classes are mainly-α, mainly-β, alternating α/β and α+β. In CATH the α/β and α+β classes are merged
Protein architecture	This is the description of the gross arrangement of secondary structures (α-helices and β-strands) in three-dimensional space, independent of their connectivity
Protein fold/topology	This is the description of the gross arrangement of secondary structures in three-dimensional space, and is dependent on the orientation of secondary structures and the connectivities between them
Homologous superfamily	A homologous superfamily is a group of proteins whose structures and functions suggest a common evolutionary origin
Homologous family	Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater. However, in some cases similar functions and structures provide definitive evidence of common descent in the absence of high sequence identity

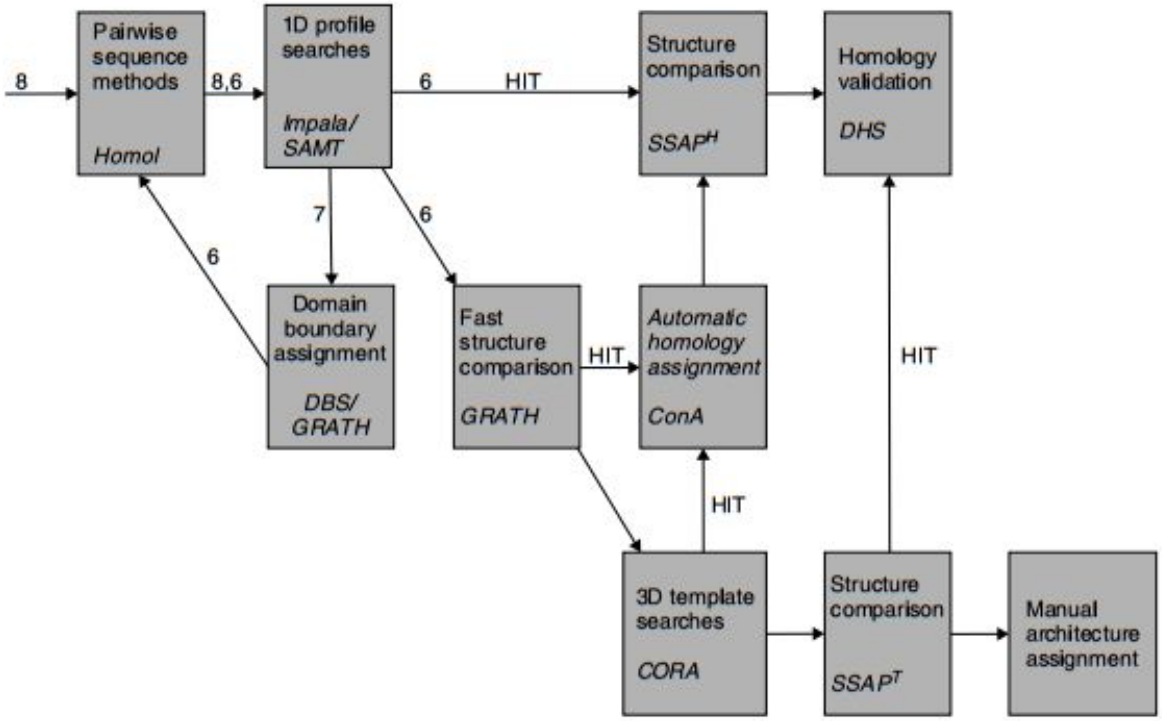
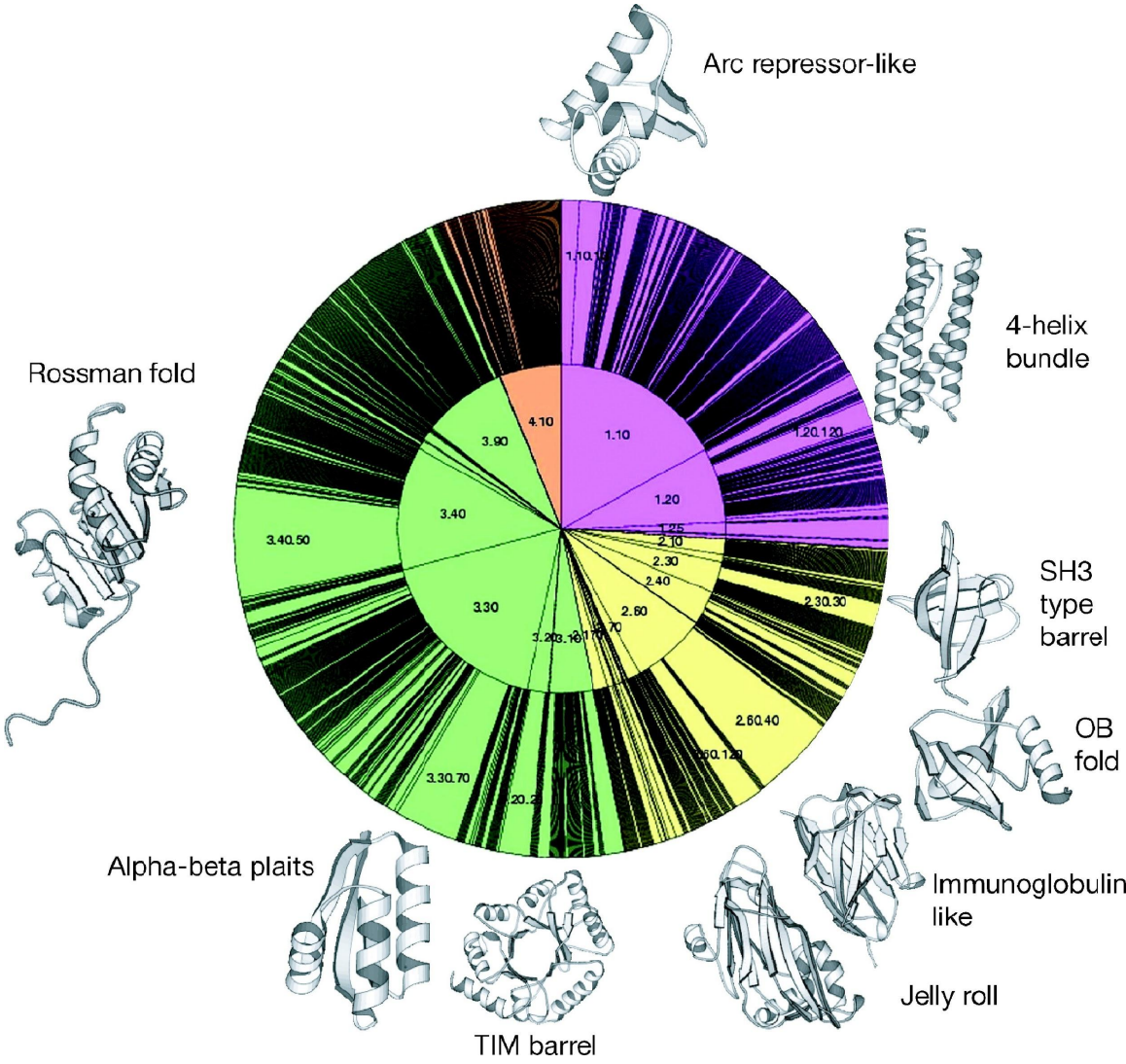


Figure 13.3. Flowchart describing the CATH classification procedure. New sequences are first compared pairwise, by sequence (HOMOL), with each other and all the entries in the database. Those that have not been identified as a sequence match are then compared using profile-based methods. If a homologue is found, the structure is compared structurally with all the members of the homologous superfamily (SSAP) and the DHS family data are updated. Those that are unmatched are assigned domain boundaries using DBS. The resulting single domains are again compared by pairwise and profile-based sequence methods. If no homologous relative is found, a fast structural comparison program (GRATH) is used to compare the domain with all sequence families within the CATH database. Structural templates (CORA) are then used to compare the structure with representatives from all the top scoring fold groups to assign homology using ConA. If GRATH does not find a significant hit, then the structure is compared against CORA templates from all superfamilies in the same class to identify a match. Finally, pairwise SSAPs against the database are run on any remaining unclassified structures as a final validation. If there is no significant fold match the architecture is assigned manually.

The strategy used in classifying new structures into the database can thus be broken down into five major steps: (1) Close relatives are identified first using pairwise sequence methods. (2) Sequence profiles and structure comparison protocols are used to detect more distant relatives. (3) Structures unclassified at this stage are then examined using both automatic and manual procedures to determine domain boundaries. (4) Unclassified domain structures are recompared using the methods employed in steps 2 and 3. (5) Finally, any structures remaining unclassified are manually assigned to architectures within CATH or new architectures are described (see Fig. 13.3). The algorithms and manual validation protocols, used at different stages of the classification, are described below.

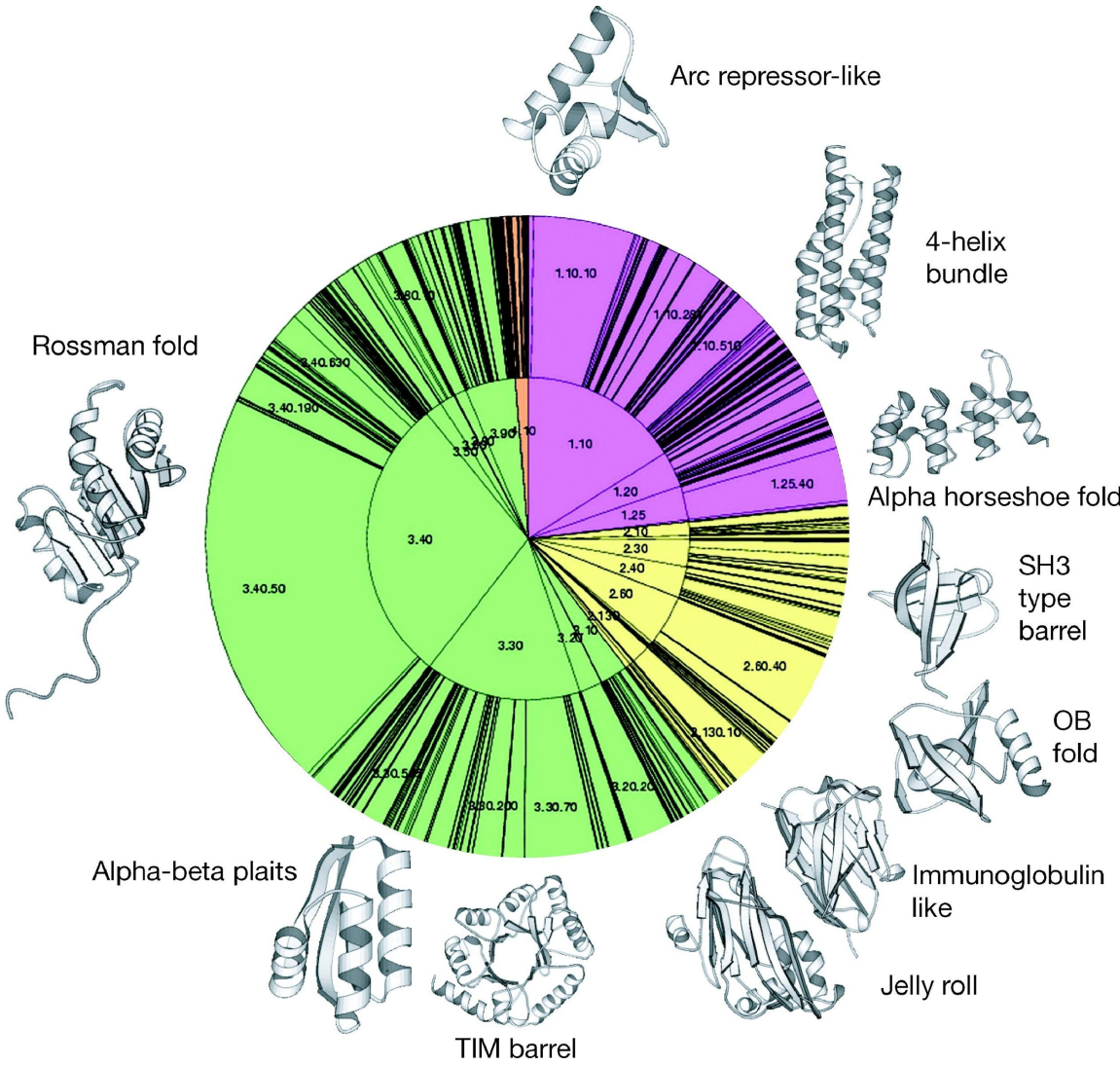
CATH

Illustration of the distribution of domain structures from the PDB among the different levels in the CATH hierarchy. The three classes are illustrated in color: mainly α, pink; mainly β, yellow; and α-β, green. The inner wheel corresponds to different architectures in the classification and the outer wheel to different fold groups. Each fold group has been subdivided according to the numbers and populations of different homologous superfamilies adopting that fold.



CATH

Illustration of the distribution of CATH domains among the sequences from 150 completed genomes, in Gene3D. In this case, the fold groups labeled in the outer circle have been divided according to the number and size of close sequence families within each fold group.



Cinco grandes “group folds” contienen cerca de un quinto de todas las superfamilias de homólogos de CATH

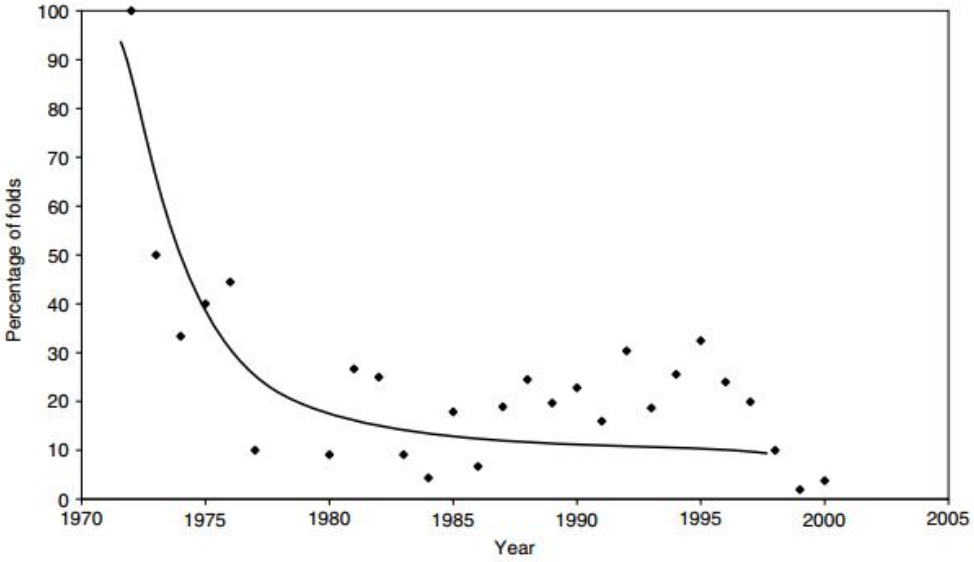


Figure 13.7. Number of unique folds identified annually as a percentage of the number of new structures determined.

The popularity of these folds may be a result of divergent or convergent evolution. Divergent evolution gives rise to families of proteins in which the structure is generally well conserved but sequences may have changed to the extent that no significant similarity remains. In paralogues, which arise from duplication of the gene within an organism, the function of protein may also have been modified or changed. Thus, apparently diverse superfamilies within these superfolds may in fact be extremely distant relatives whose relationships cannot easily be verified from the available sequence or functional data.

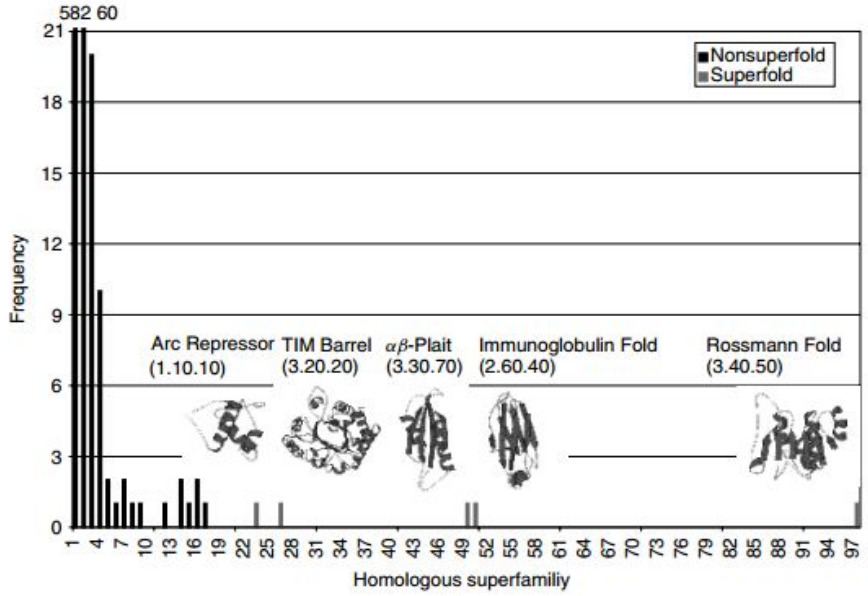


Figure 13.8. Populations of different levels in the CATH hierarchy; homologous superfamilies within fold groups. The October 2000 version of CATH was used to generate the histograms.

Are Folds Distinct or Is There a Structural Continuum?

Even more speculative are suggestions that perhaps the earliest evolutionary unit corresponded to a much smaller structural motif than a domain (68, 69) (e.g., a supersecondary structural motif such as an $\alpha\beta$-motif, β-hairpin, α-hairpin). This idea lends support to the notion of a structural continuum proposed by several researchers (15, 70) to explain the observation that some regions of fold space are very densely populated so that distinguishing between different folds becomes difficult and subjective (71).

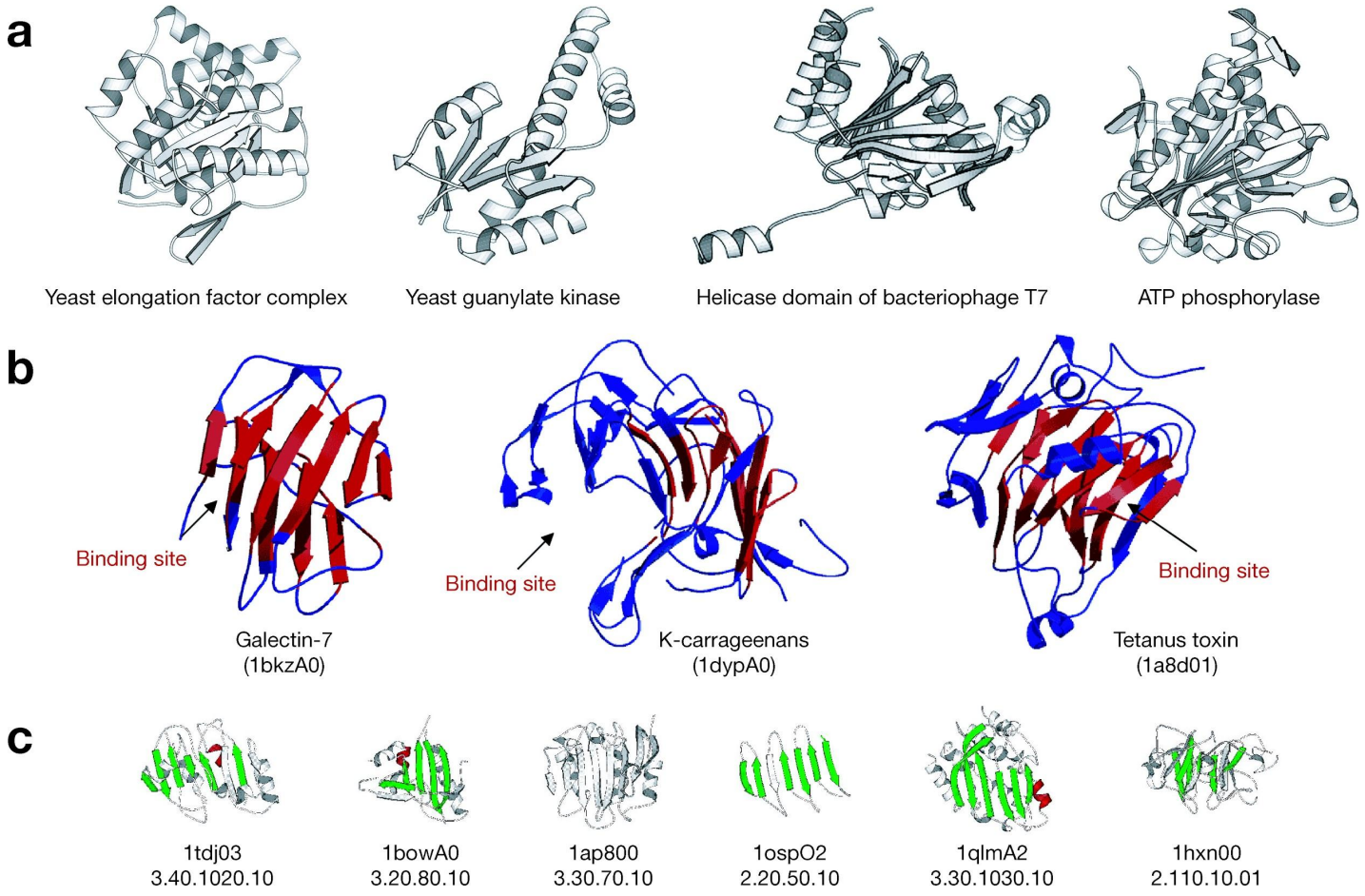


Illustration of structural overlaps between proteins **(a)** from the same homologous superfamily, P-loop hydrolases, **(b)** from the same homologous superfamily, the galactin-type carbohydrate recognition domain superfamily, and **(c)** between proteins adopting different folds in the mainly β-sandwich architecture.

Gene3D

Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis

Jonathan Lees*, Corin Yeats, James Perkins, Ian Sillitoe, Robert Rentzsch, Benoit H. Dessailly and Christine Orengo

Institute of Structural and Molecular Biology, University College London, Darwin Building, Gower St, London, WC1E 6BT, UK

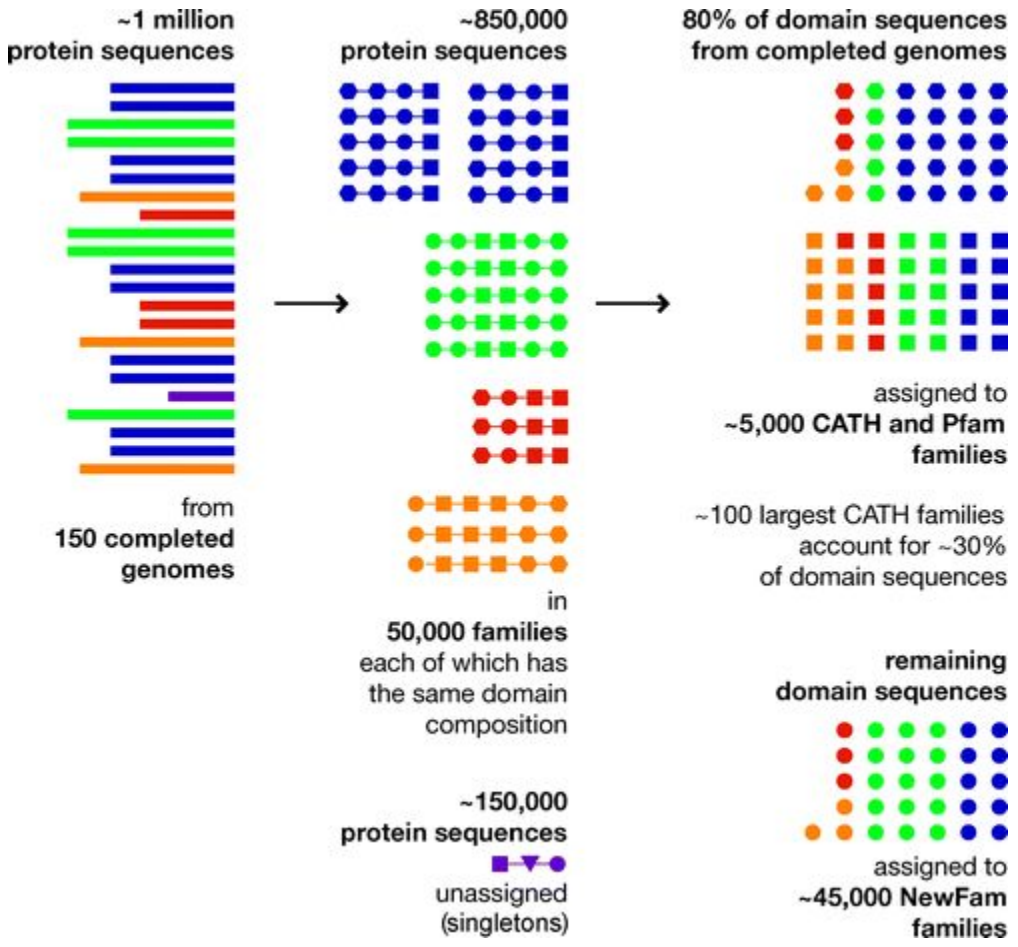
Received September 26, 2011; Revised November 12, 2011; Accepted November 14, 2011

ABSTRACT

Gene3D http://gene3d.biochem.ucl.ac.uk is a comprehensive database of protein domain assignments for sequences from the major sequence databases. Domains are directly mapped from structures in the CATH database or predicted using a library of representative profile HMMs derived from CATH superfamilies. As previously described, Gene3D integrates many other protein family and function databases. These facilitate complex associations of molecular function, structure and evolution. Gene3D now includes a domain functional family (FunFam) level below the homologous superfamily level assignments. Additions have also been made to the interaction data. More significantly, to help with the visualization and interpretation of multi-genome scale data sets, we have developed a new, revamped website. Searching has been simplified with more sophisticated filtering of results, along with new tools based on Cytoscape Web, for visualizing protein–protein interaction networks, differences in domain composition between genomes and the taxonomic distribution of individual superfamilies.

database (5) uses a combination of manual curation and automated evidence gathering to generate a superfamily classification of such structures in the PDB (6). An accurate HMM and graph theory-based method, DomainFinder (Yeats, Redfern and Orengo, manuscript in revision), is used to identify and resolve the boundaries of predicted domains. The new release of Gene3D (v10.2) provides over 16 million predicted domains from 2549 CATH superfamilies in 60% of approximately 15 million scanned sequences. This is an increase of 5% in domain annotation coverage compared with our last review in NAR (1). Gene3D domain annotations are provided via the Gene3D website (http://gene3d.biochem.ucl.ac.uk), the CATH-Gene3D DAS (http://gene3d.biochem.ucl.ac.uk/Gene3D/Das), RESTful web services at http://gene3d.biochem.ucl.ac.uk/WebServices/ (7) and InterPro (8).

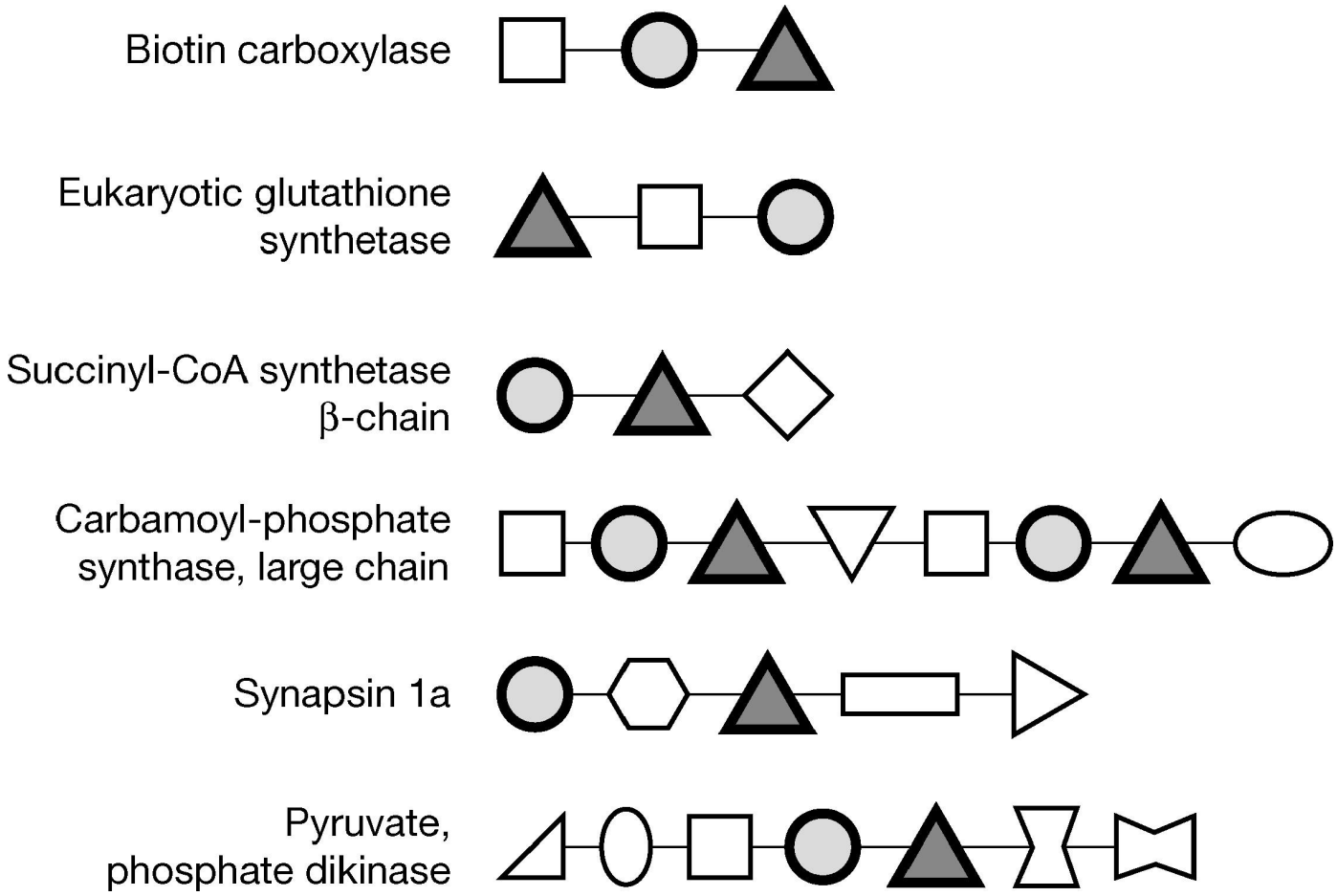
Protein domains, and distinct combinations of them, are considered the primary building blocks of protein function evolution. The assignment of domains to a protein can help identifying functionally important residues from distant homologues (9) provide mechanistic explanations for the effects of sequence polymorphisms (10) and enable the 'inheritance' of interactions from homologues (11,12). To enhance the domain annotations generated by Gene3D, we also integrate many other complementary data sources. These include molecular and pathway function annotations from GO (13), taxonomic



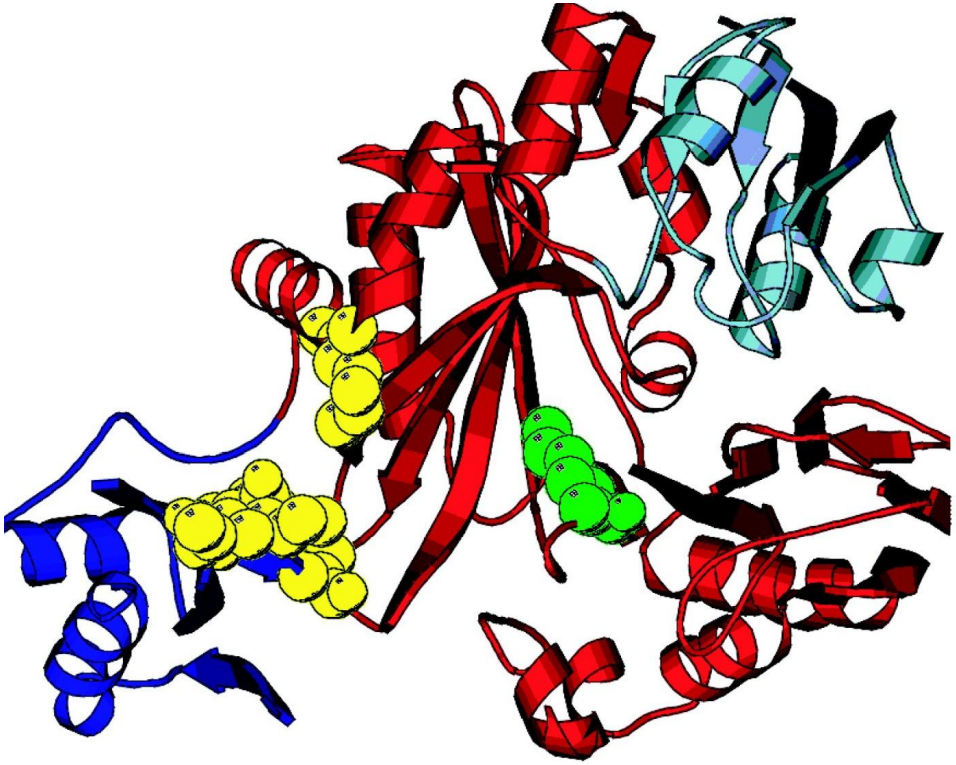
Schematic representation of the classification of sequences, from complete genomes, into protein and domain families.

The numbers of domain families identified are given together with the proportion of domain sequences in completed genomes that can be assigned to each type of domain family (*CATH, Pfam*).

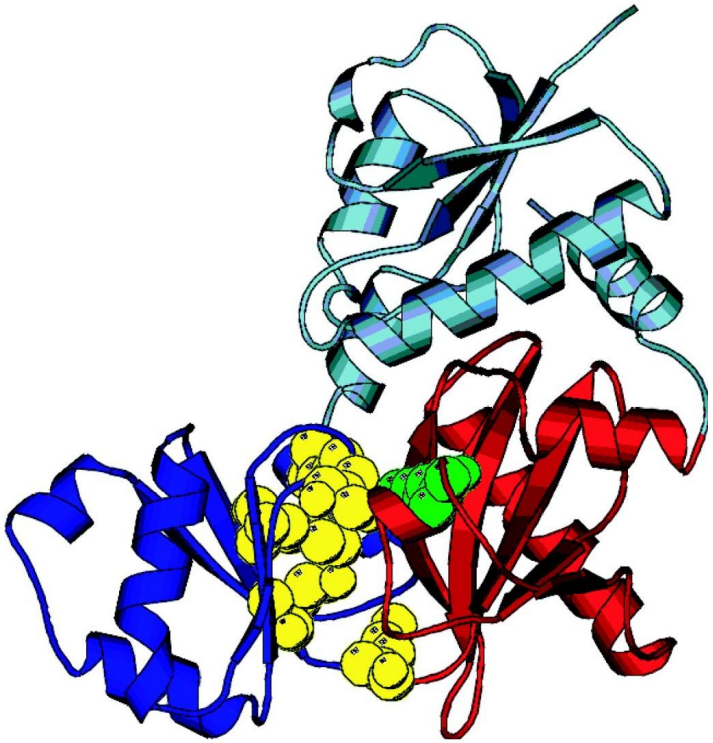
Domain sequences that are not assigned to *CATH* or *Pfam* are assigned to a new type of family, called *NewFam*.



Domain compositions of proteins all possessing the supradomain pair (shaded), comprising the ATP grasp domains that are involved in binding ATP



D-alanine ligase

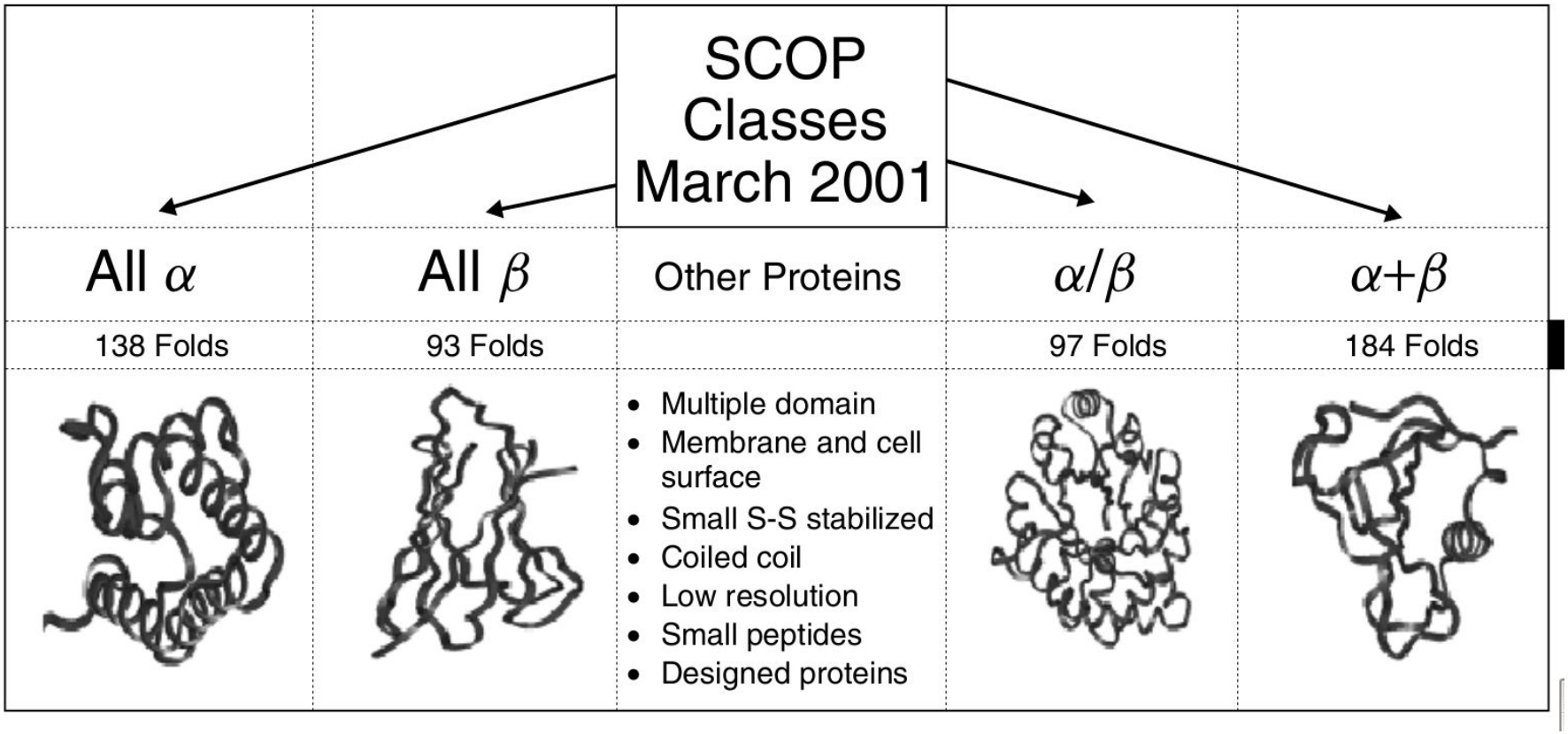


Biotin carboxylase

The ATP-dependent carboxylase-amine/thiol ligase superfamily is large and structurally and functionally diverse. All relatives comprise three domains and bind Mg2+ATP in a cleft between a large and small domain. Substrate specificity varies between relatives, and the enzymes act on a vast array of donor and acceptor substrates. However, the chemistry is conserved throughout involving the ATP-dependent ligation of a substrate carboxylate to an amine or thiol group of a second substrate. In all relatives, the substrate-binding site is located on the large domain. However, some relatives (e.g., d-alanine d-alanine ligase) contain a large structural embellishment, comprising an additional β-sheet that encompasses the active site enclosing it in a box-like conformation that severely restricts access.

SCOP

The scop classification of proteins has been constructed manually by visual inspection and comparison of structures, but with the assistance of tools to make the task manageable and help provide generality. The job is made more challenging--and theoretically daunting--by the fact that the entities being organized are not homogeneous: sometimes it makes more sense to organize by individual domains, and other times by whole multi-domain proteins.1



SCOP

Díficil 30%>= **<<%**

Class	Folds	Superfamilies	Families
All alpha proteins	138	224	337
All beta proteins	93	171	276
α/β proteins	97	167	374
$\alpha + \beta$ proteins	184	263	391
Multidomain proteins	23	28	35
Membrane and cell surface proteins	11	17	28
Small proteins	54	77	116
Total	605	947	1557

Major structural similarity

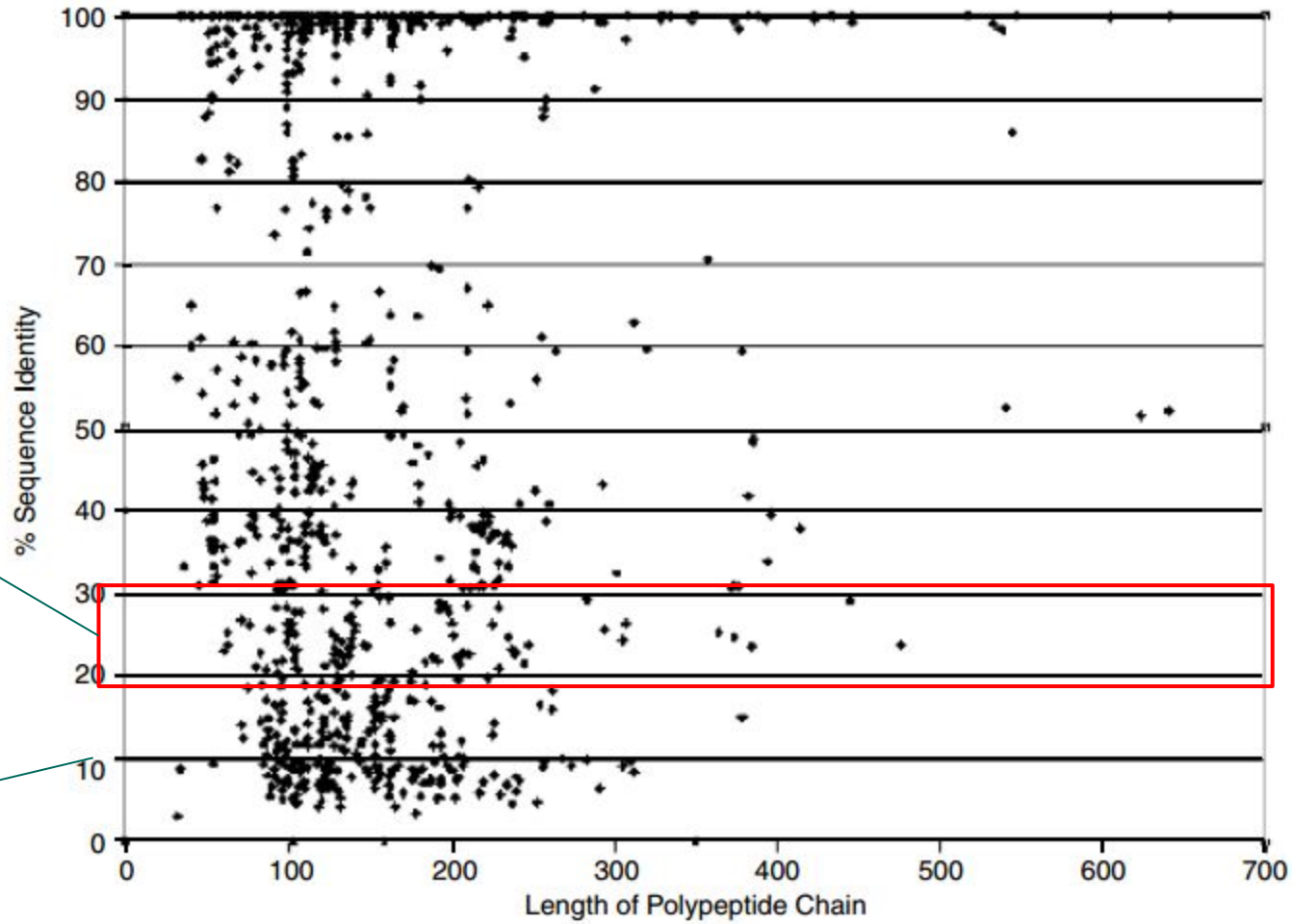
Probable common evolutionary origin

Clear evolutionary relationship

Alineamiento Estructural

Comparación de Estructuras

- Cambios conformacionales
- Caracterización de la familia estructural
- Detección de relaciones evolutivas lejanas
- Variación estructural en una familia por adaptación funcional
- Identificación de estructuras nativas comunes. Super plegamientos
- Clasificación estructural de proteínas
- Comparación de modelos estructurales



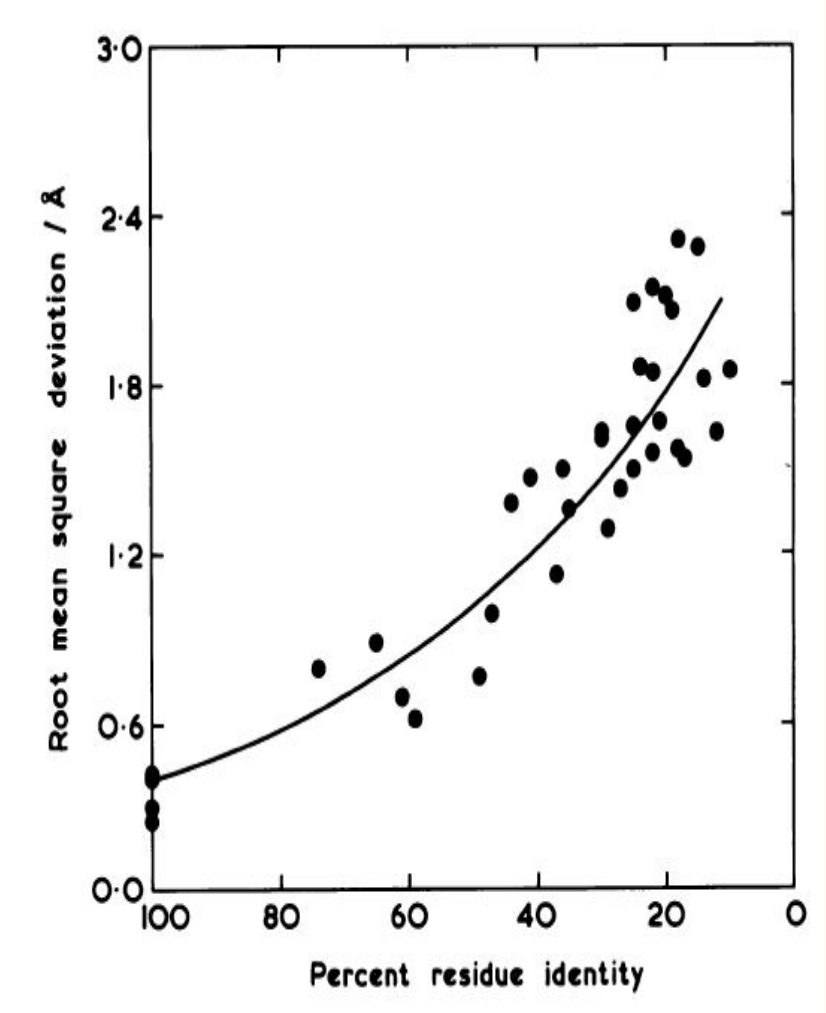
Twilight zone

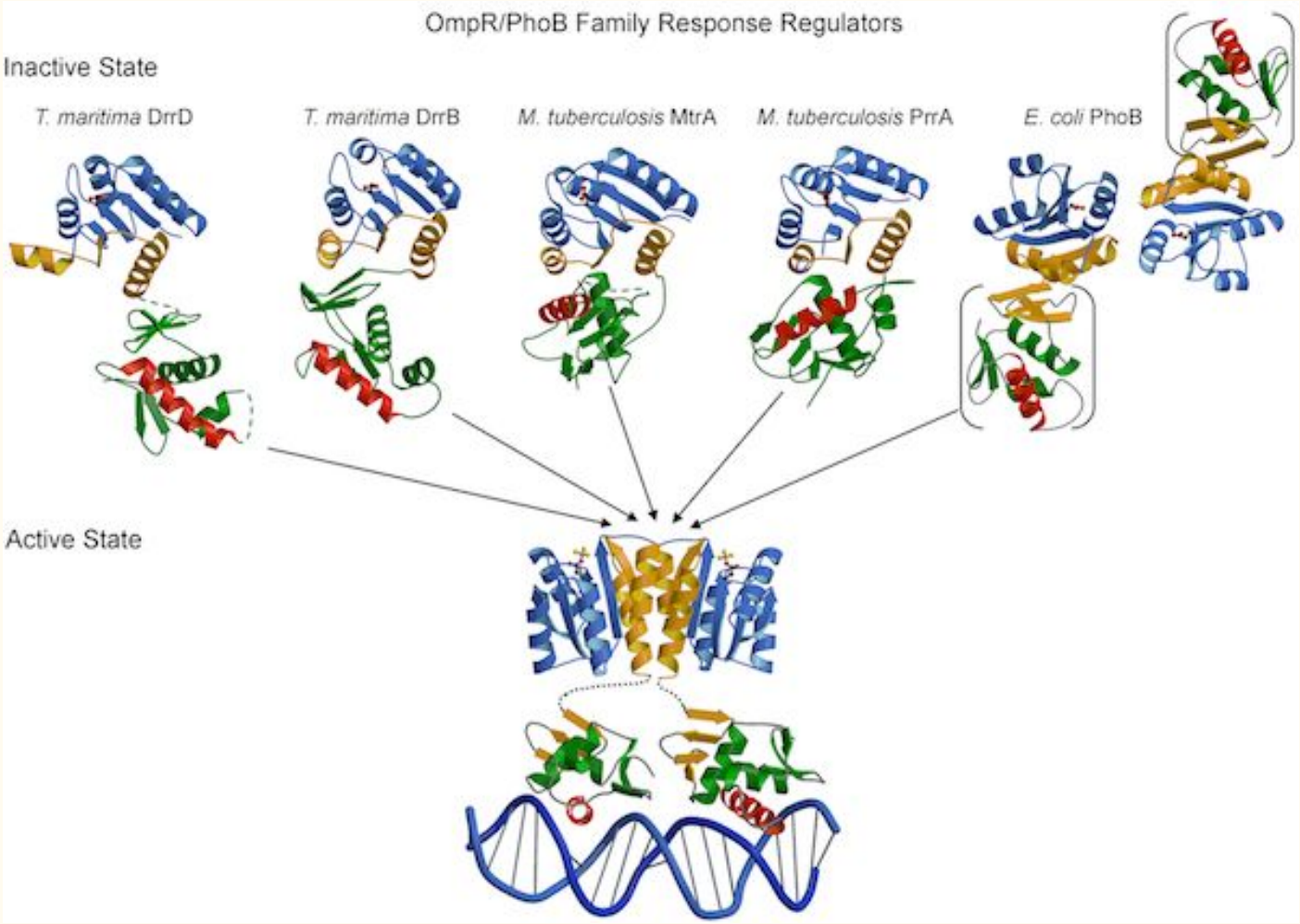
Midnight zone

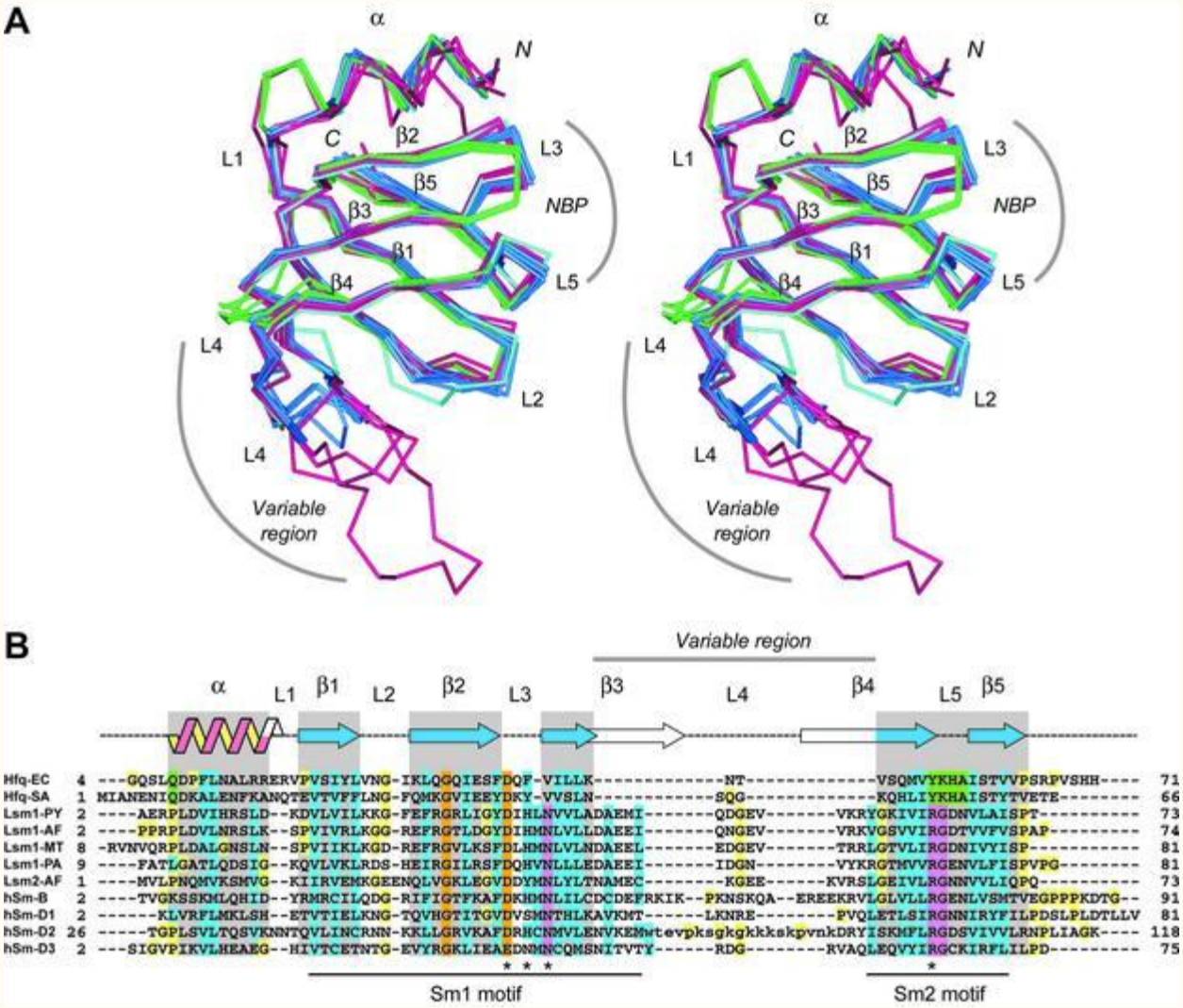
Figure 16.1. Structure similarity versus sequence similarity. Each data point represents one of 1000 randomly selected polypeptide chains from the PDB showing structure similarity as measured by CE with a z-score > 4.5.

The EMBO Journal vol.5 no.4 pp.823–826, 1986

The relation between the divergence of sequence and structure in proteins

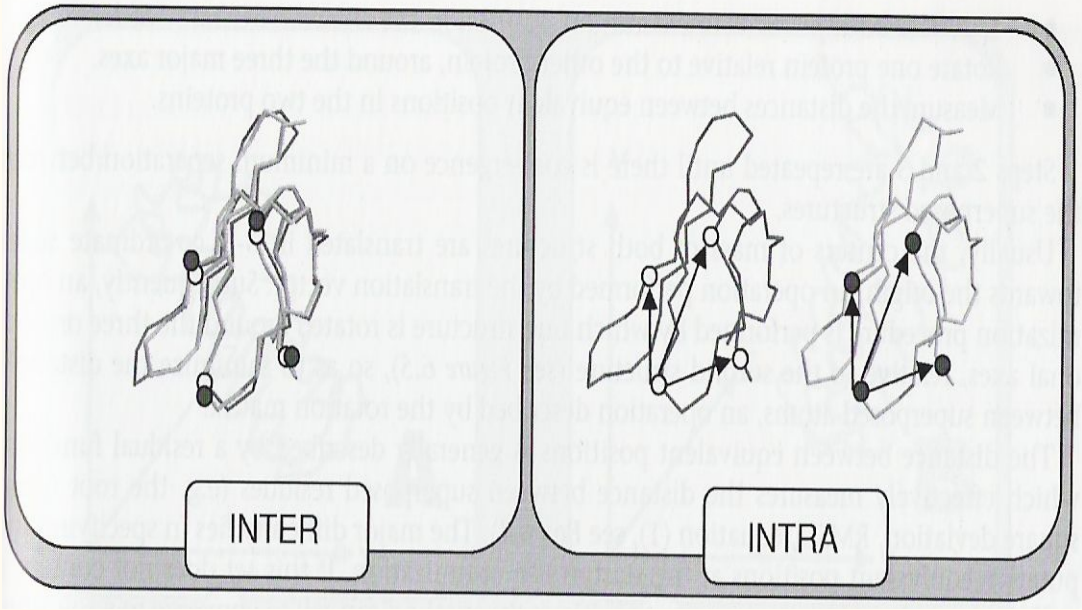



OmpR/PhoB Family Response Regulators
Inactive State
T. maritima DrrD
T. maritima DrrB
M. tuberculosis MtrA
M. tuberculosis PrrA
E. coli PhoB
Active State


A
α
N
C
β2
L1
L3
β5
NBP
β3
β1
β4
L5
L4
L2
L4
Variable region
α
N
C
β2
L1
L3
β5
NBP
β3
β1
β4
L5
L4
L2
L4
Variable region
B
Variable region
α
L1
β1
L2
β2
L3
β3
L4
β4
L5
β5
Hfq-EC 4 ---GQSLQDPFLNALRRERVPVSIYLVNG-IKLQGQIESFDQF-VILLK----------NT----------VSQMVYKHAISTVVPSRPVSHH---- 71
Hfq-SA 1 MIANENIQDKALENFKANQTEVTVFFLNG-FQMKGVIEEYDKY-VVSLN----------SQG----------KQHLIYKHAISTYTVETE-------- 66
Lsm1-PY 2 ----AERPLDVIHRSLD--KDVLVILKKG-FEFRGRLIGYDIHLNVVLADAEMI-------QDGEV-------VKRYGKIVIRGDNVLAISPT---------- 73
Lsm1-AF 2 ----PPRPLDVLNRSLK--SPVIVRLKGG-REFRGTLDGYDIHMNLVLLDAEEI-------QNGEV-------VRKVGSVVIRGDTVVFVSPAP--------- 74
Lsm1-MT 8 -RVNVQRPLDALGNSLN--SPVIIKLKGD-REFRGVLKSFDLHMNLVLNDAEEL-------EDGEV-------TRRLGTVLIRGDNIVYISP---------- 81
Lsm1-PA 9 ----FATLGATLQDSIG--KQVLVKLRDS-HEIRGILRSFDQHVNLLLEDAEEI-------IDGN--------VYKRGTMVVRGENVLFISPVPG-------- 81
Lsm2-AF 1 ----MVLPNQMVKSMVG--KIIRVEMKGEENQLVGKLEGVDDYMNLYLTNAMEC--------KGEE------KVRSLGEIVLRGNNVVLIQPQ---------- 73
hSm-B 2 ----TVGKSSKMLQHID--YRMRCILQDG-RIFIGTFKAFDKHMNLILCDCDEFRKIK--PKNSKQA--EREEKRVLGLVLLRGENLVSMTVEGPPPKDTG-- 91
hSm-D1 2 ------KLVRFLMKLSH--ETVTIELKNG-TQVHGTITGVDVSMNTHLKAVKMT------LKNRE--------PVQLETLSIRGNNIRYFILPDSLPLDTLLV 81
hSm-D2 26 -----TGPLSVLTQSVKNNTQVLINCRNN-KKLLGRVKAFDRHCNMVLENVKEMwtevpksgkgkkkskpvnkDRYISKMFLRGDSVIVVLRNPLIAGK---- 118
hSm-D3 2 ---SIGVPIKVLHEAEG--HIVTCETNTG-EVYRGKLIEAEDNMNCQMSNITVTY------RDG--------RVAQLEQVYIRGCKIRFLILPD--------- 75
* * *
*
Sm1 motif
Sm2 motif

Comparación de Estructuras

- Superposición de estructuras y medición de distancias intermoleculares basados principalmente en características geométricas
- Comparación de medidas intramoleculares
- Combinación de ambas



Sequence-based structural alignment

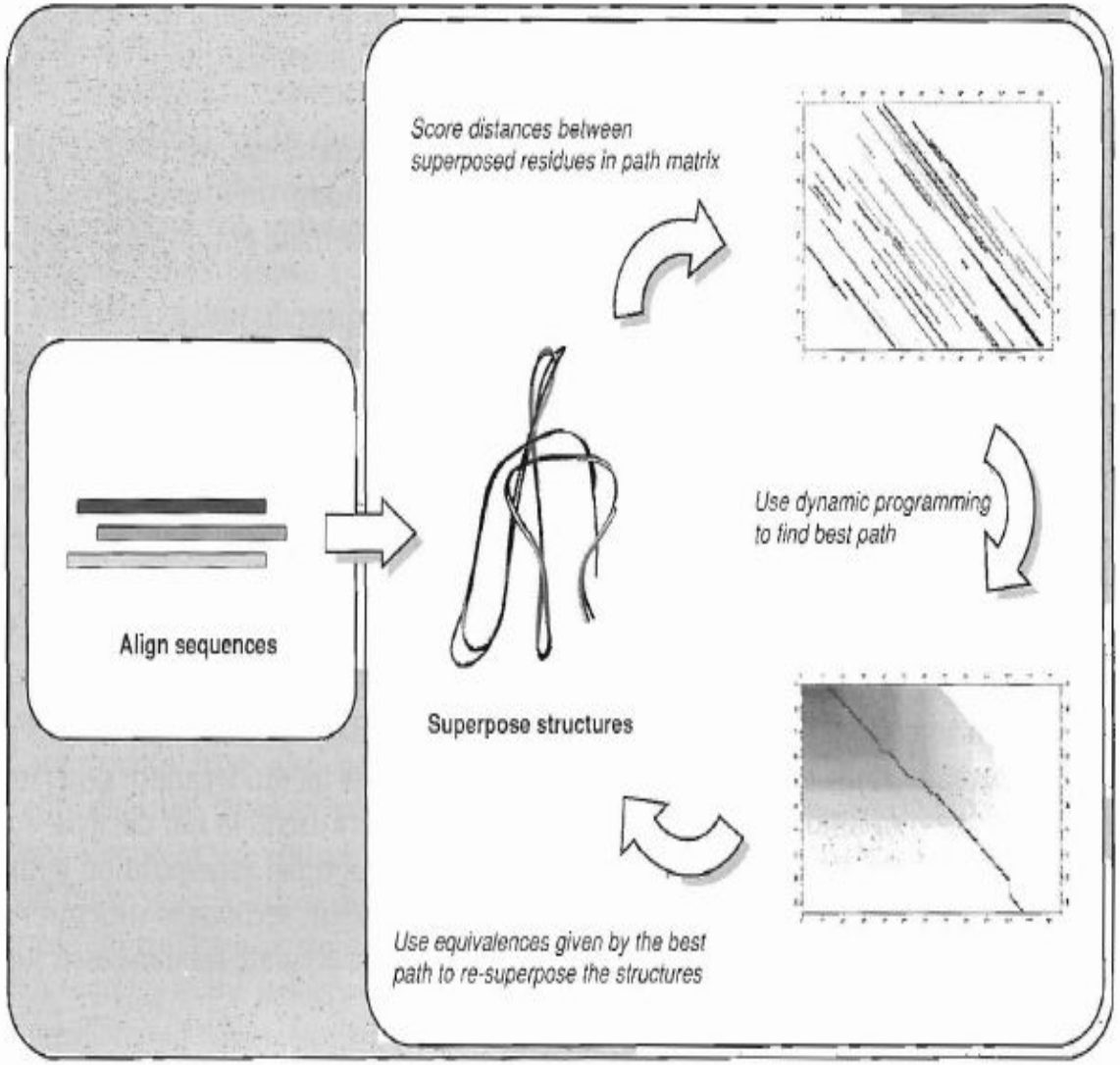
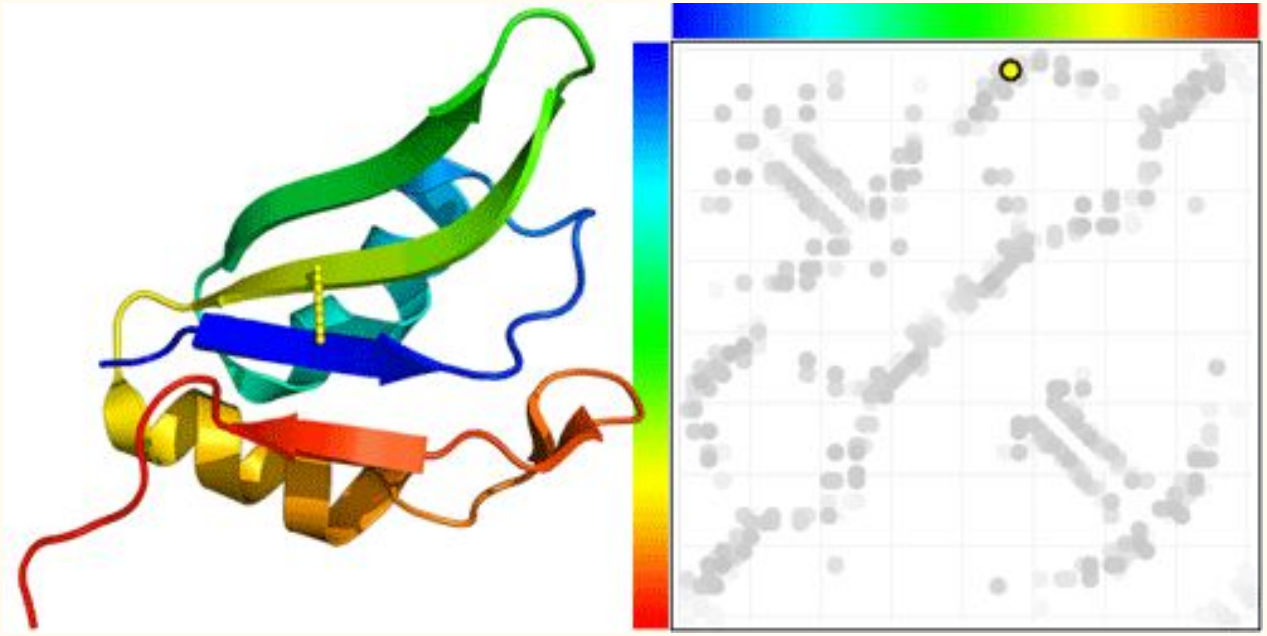


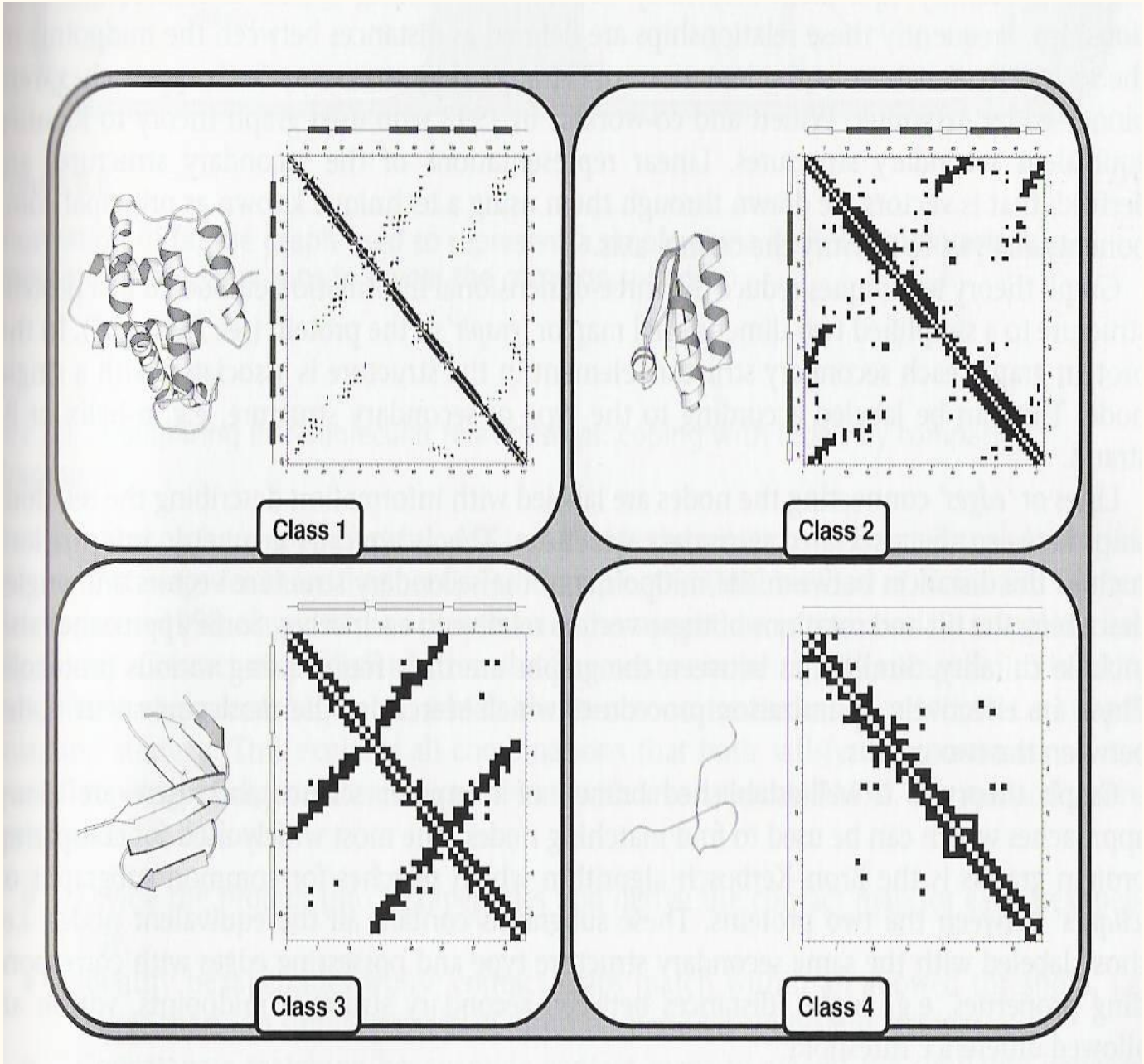
Figure 6.6

Schematic representation of the STAMP method devised by Russell and Barton. Sequence alignment is first used to determine putative equivalent positions in order to superpose the structures. Distances between the superposed positions are then used to score a 2-D score matrix within a specified window which is analyzed by dynamic programming (see Chapter 3) to obtain a better set of equivalent positions on which the structures are re-superposed.

Mapas de contactos



Mapas de contactos



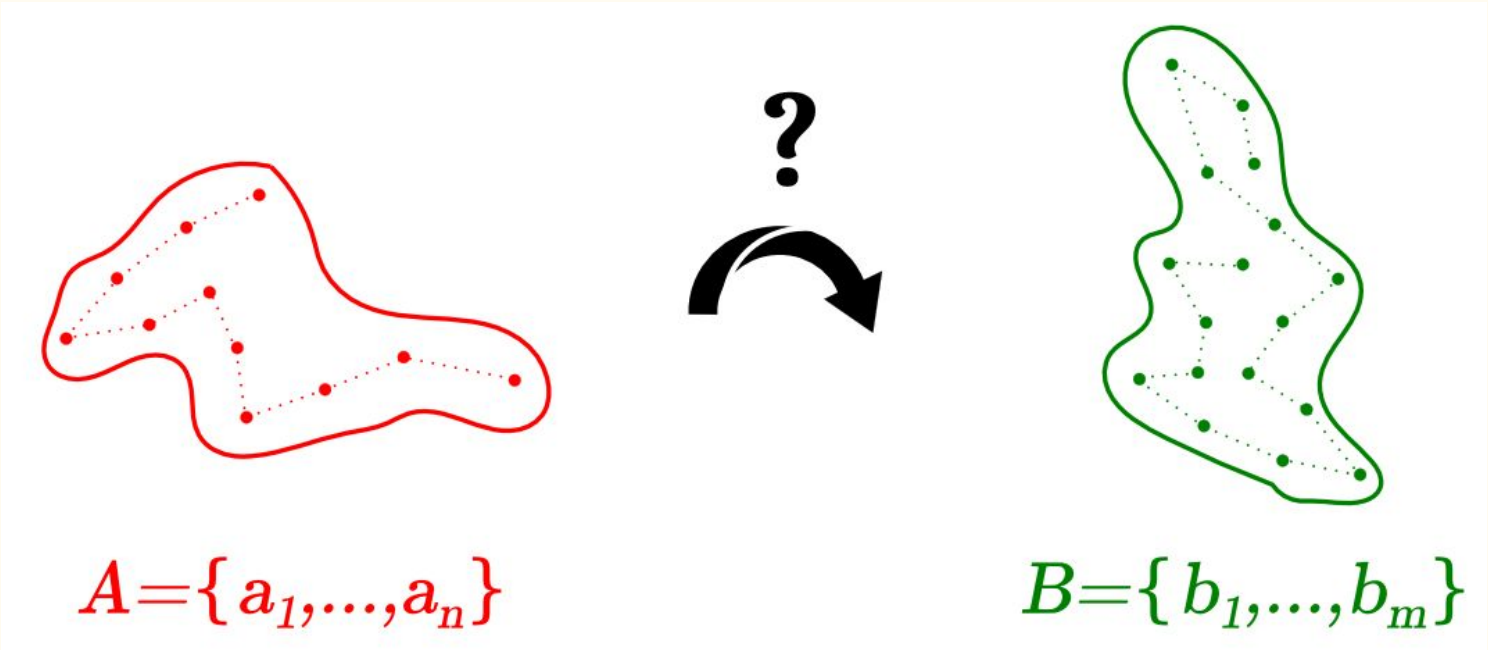
Alineamiento Estructural

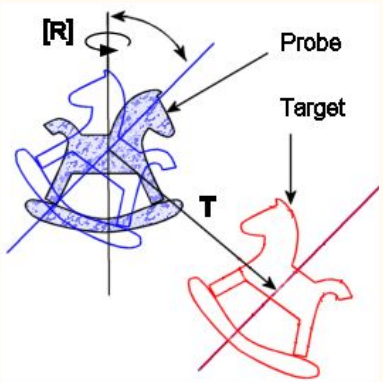
Rigid: Only rigid-body transformations are considered between the structures being compared

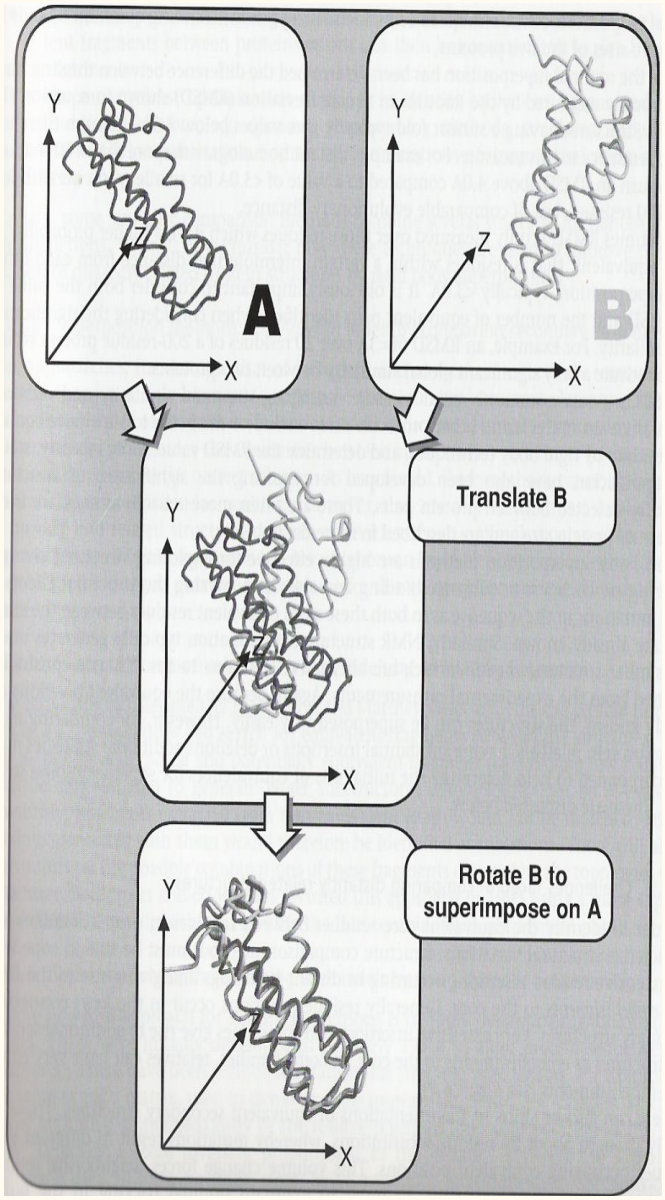
Flexible: The method allows for some flexibility within the structures being compared, such as movements around hinge regions

Establishing an alignment between these sets is equivalent to two steps:

1. Establish a match M = {(ai,bj) | ai ∈ A, bj ∈ B}
2. Rotate and translate A onto B so that equivalent atoms are as close as possible.







Rossman y Argos, 1970s. **Superposición de Cuerpo Rígido**

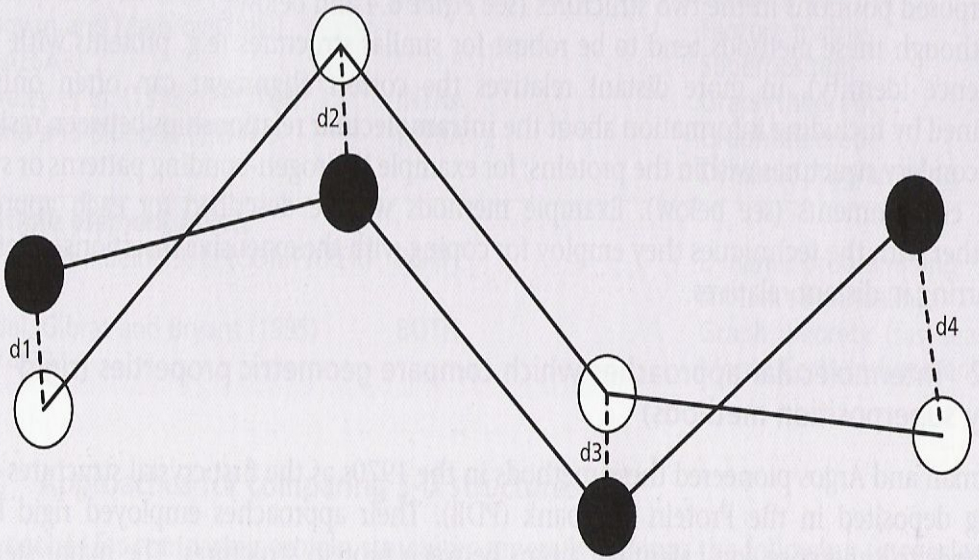
Box 6.1 Root Mean Square Deviation (RMSD)

The residual function used to measure the similarity between two protein structures following rigid body superposition is typically the root mean square deviation between the structures (RMSD).

The RMSD between two structures is quite simply the square root of the average squared distance between equivalent atoms, defined by the equation:

$$R = \sqrt{\frac{\sum_{i=1}^{N} d_i^2}{N}} \qquad (1)$$

The distances (d_i) can be visualized as:



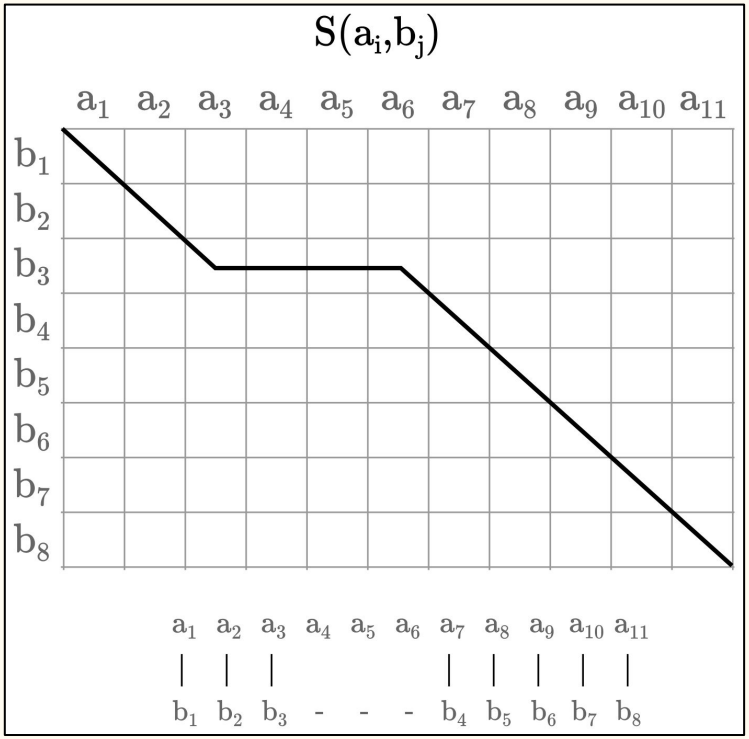
RMSD

- En general se usa el *RMSD* de *C alpha*
- Idéntica proteína cristalizada en las mismas condiciones varias veces puede tener un *RMSD* promedio de 0.5 Å
- Idéntica proteína cristalizada con/sin sustrato/inhibidor/efector alostérico/cofactor etc puede tener *RMSD* hasta 25 Å (diversidad conformacional)
- En general la variación en el *RMSD* debido a cambios conformacionales "comunes" es aproximadamente 3-5 Å. A estos movimientos se los llama *"native-like"*

MAMMOTH (Matching Molecular Models Obtained from Theory)

Step 1: Pairing up residues (similarity matrix)

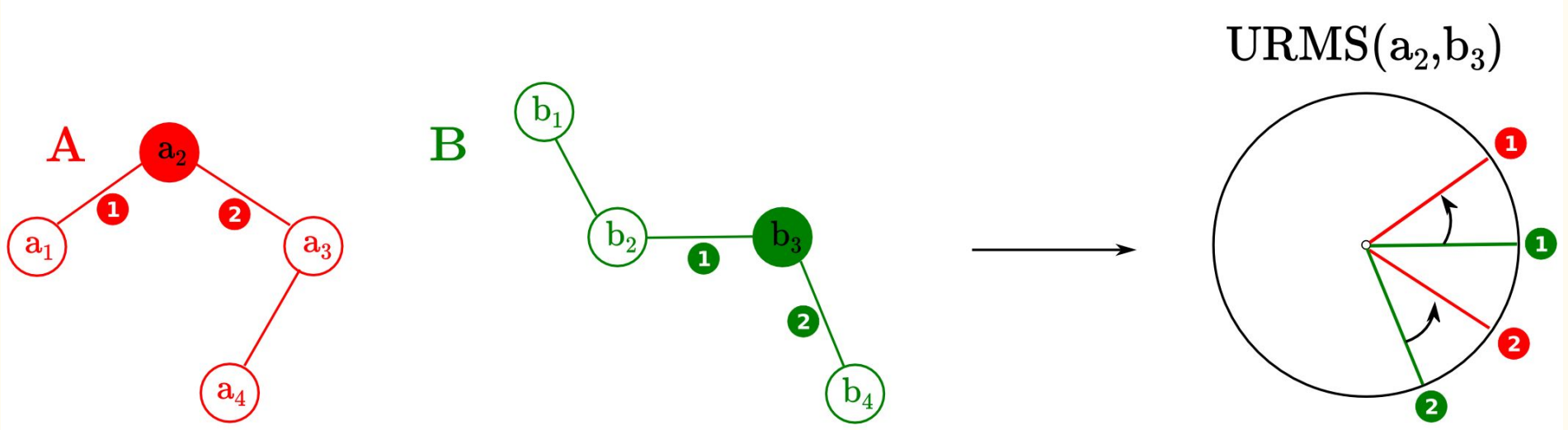
In order to establish a match between equivalent atoms in A and B, MAMMOTH, like several other structural alignment algorithms, uses a well-established alignment technique: a similarity matrix (often inferred from and referenced as a distance matrix). A similarity matrix for alignment is an n x m array where each entry S(ai,bj) represents the pairwise similarity score between residues ai and bj.



The *optimal* alignment is simply the alignment which maximises this score.

MAMMOTH (Matching Molecular Models Obtained from Theory)

For sequence alignments similarity scores can be assigned to residues from substitution tables like BLOSUM. However, it is not immediately clear of an appropriate equivalent for structures. MAMMOTH, like several other algorithms, defines the similarity between different residues by examining their local structural landscape. Specifically, this means comparing fragments of each backbone, centred on the residue of interest. MAMMOTH uses the URMS distance between heptapeptide fragments. This distance is illustrated below using 2D chains and tripeptide fragments.



The optimal rotation is found, superposing equivalent vectors as best as possible, and then the RMSD of the endpoints on the surface of the sphere is calculated as URMS(ai,bj).

MAMMOTH (Matching Molecular Models Obtained from Theory)

Step 2: Global superposition (MaxSub)

The above alignment results in a match M' optimising the local structural similarity of residues in each structure, however, their is no guarantee that this will result in a set of coordinates close in global space. In order to finalise the match set $M \subseteq M'$ as well as calculating the optimal superposition of the paired residues of A onto their equivalent points in B, MAMMOTH use theMaxSub algorithm.

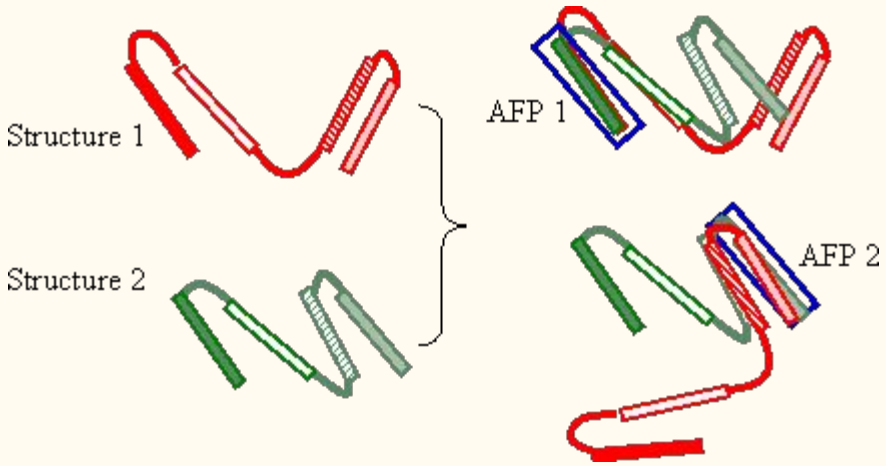
Scoring the alignment

n order to summarise the significance of this alignment, the algorithm generates the PSI score: the percentage structural identity (which is simply the size of the maximum subset divided by the length of the shortest protein). As a global measure of the strength of similarity the PSI score is poorly constructed and scales with protein length. In order to adjust for this bias, MAMMOTH fits a Gumbel distribution to PSI scores obtained from random structure comparisons between unrelated proteins at bins of different lengths. This results in a z-score measuring, instead of the PSI of an alignment, the likelihood of obtaining a PSI score as good as that by chance between any two proteins of the same lengths.

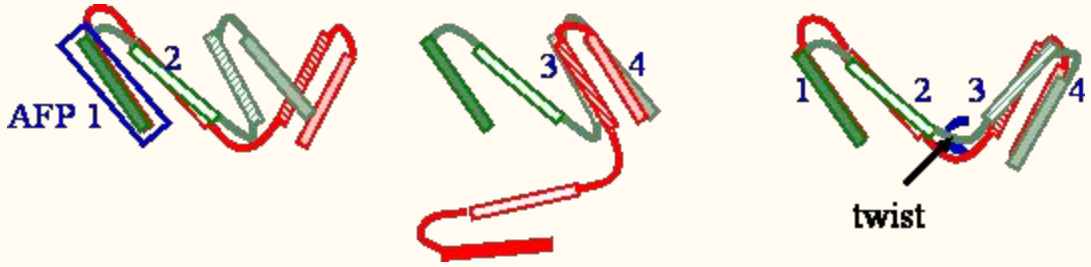
Alineamiento Estructural

Rigid: Only rigid-body transformations are considered between the structures being compared

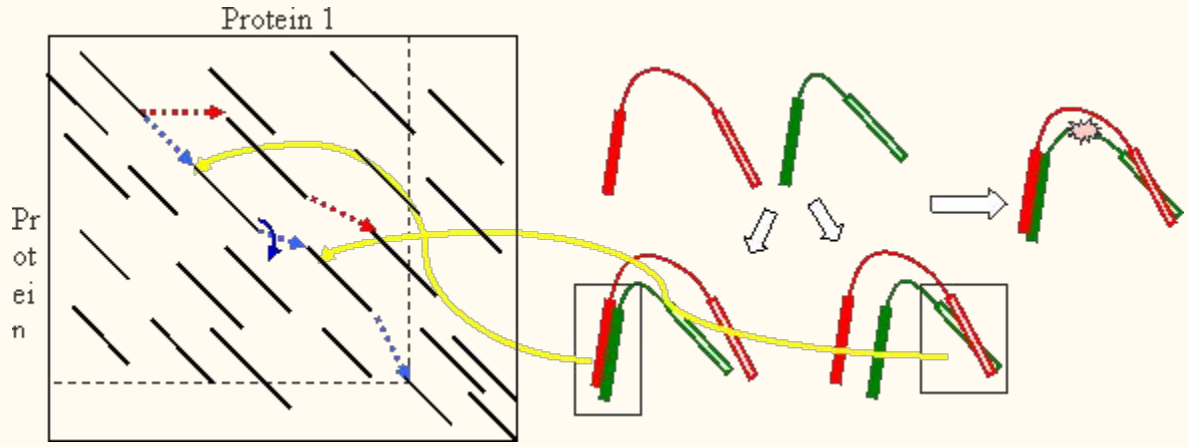
Flexible: The method allows for some flexibility within the structures being compared, such as movements around hinge regions



Flexible: *FATCAT*



Flexible: *FATCAT*



Flexible: *FATCAT*

The **FATCAT report** for 2spcA and 1aj3 is shown below as an example:

FATCAT reports the overall similarity between the two structures (probability, RMSD and equivalent positions), and the alignment represented in three ways. Alignment blocks seperated by twists are shown in different colors (in this cases, red, pink and yellow).

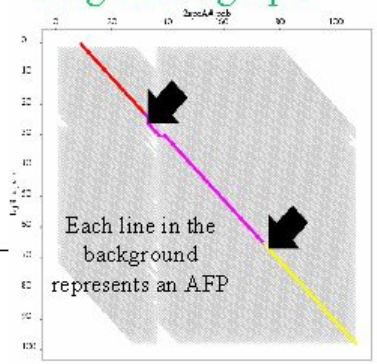
1. The two structures are significantly similar with P-value of 1.05e-06 (raw score is 220.61)
2. The structure alignment has 94 equivalent positions with an RMSD of 2.02, with 2 twists. View the alignment file
3. View the graph of FATCAT chaining result (download the postscript file)
4. View the superposition of the two structures (a pdb file in which structure 2spcA is in chain A, and twisted structure 1aj3 is in chain B)
   - View the superposition by Chime online
   - Download the pdb file
   - Download the rasmol script for the above pdb file

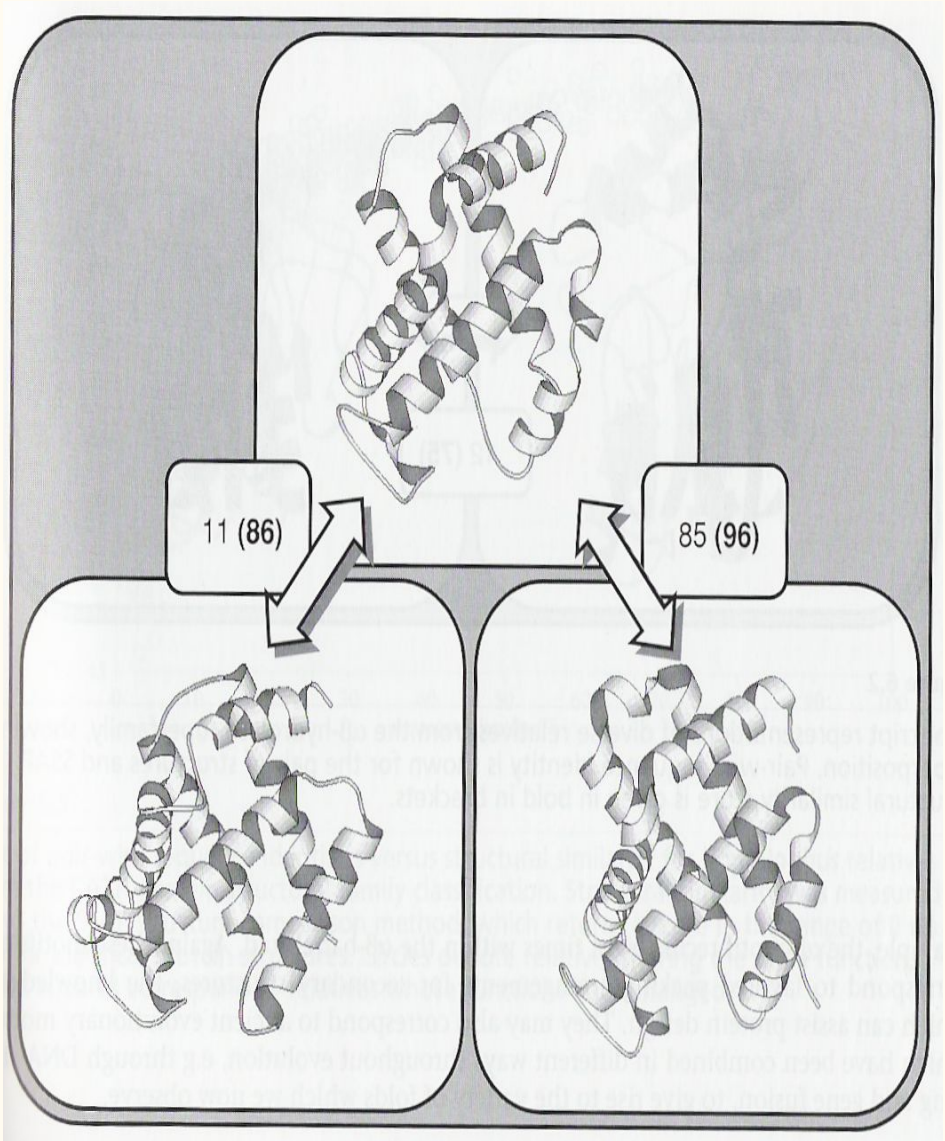
Structure superposition

Alignment file

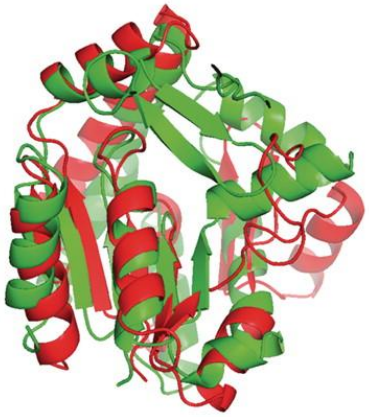
```
Align 2spcA.pdb 107 with 1aj3.pdb 98
Twists 2 ini-len 96 ini-rmsd 4.88 opt-equ 94 opt-rmsd 2.02 chain-rmsd 25.78 Score 220.61 align-len 100
P-value 1.05e-06 Afp-num 7648 Identity 6.00% Similarity 21.00%
Block  0 afp  3 score 70.23 rmsd  0.74 gap 0 (0.00%)
Block  1 afp  5 score 106.92 rmsd  3.51 gap 0 (0.00%)
Block  2 afp  4 score 78.88 rmsd  3.44 gap 0 (0.00%)

Chain 1:   10 RDCELAESWMSAREAFLNADDDAN--AGGNVEALIKKHEDFDKAINGHEQKIAALQTVADQLIAQNHYAS
              1111111111111111111111111  2222  22222222222222222222222222222222  3
Chain 2:    1 HQFFRDMDDRESWIKKKKLLVSSEDYGRDL--TGVQNLRKKHKRLEAELAAHEPAIQGVLDTGKKLSDDN

Chain 1:   78 NLVDEKPKQVLEPWRHLKEGLIEKRSRLGI
              333333333333333333333333333333
Chain 2:   69 TIGKEEIQQRLAQFVDHWKELKQLAAARGC

Note: positions are from PDB; the numbers between alignments are block index
```

Alignment graph

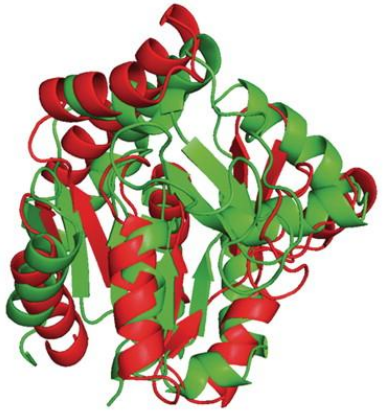



11 (86)
85 (96)

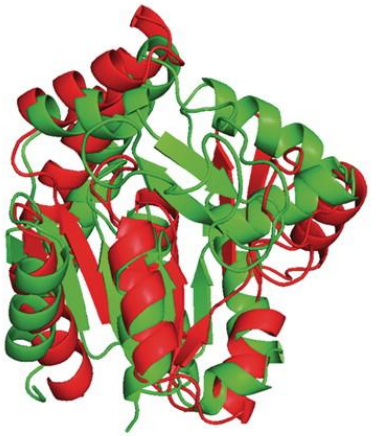
iPBA: **2.33**/124



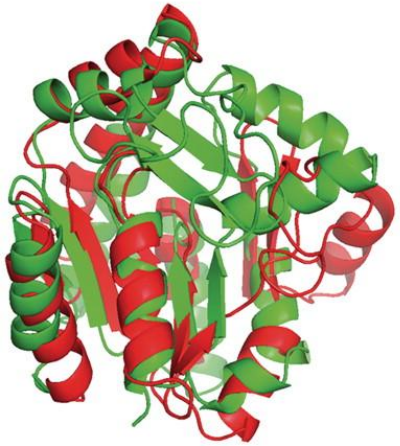
CE: **4.00**/151



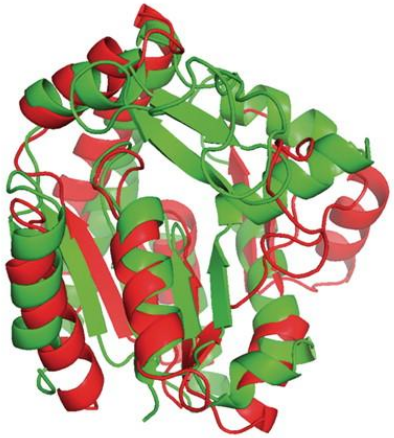
DALI: **3.70**/147



TM-Align: **3.43**/152



GANGSTA+: **2.93**/116



ALADYN: **3.30**/113

Table 6.1 Structure comparison algorithms

Authors	Type of method *(Intermolecular/ Intramolecular)*	Algorithms
Residue level		
Rao and Rossmann (1973)	INTER	Superposition
Padlan (1975)	INTRA	Distance matrices
Sutcliffe *et al.* (1987) (MNYFIT)	INTER	Superposition
Taylor and Orengo (1989) (SSAP family)	INTRA	Dynamic programming
Subbarao and Haneef (1991)	BOTH	Graph theoretic
Rose and Eisenmenger (1991) (Comp3-D)	INTER	Dynamic programming Superposition
Russell and Barton (1992) (STAMP)	INTER	Sequence alignment Superposition Dynamic programming
Nussinov group (1993)	INTRA	Geometric hashing
Fisher *et al.* (1993)	INTRA	Geometric hashing
Subbiah *et al.* (1993)	INTER	Superposition Dynamic programming
Yee and Dill (1993) (CONGENEAL)	INTRA	Distance matrices
May and Johnson (1994) (GA_FIT)	INTER	Genetic algorithm Dynamic programming Superposition
Boutonnet *et al.* (1995)	INTER	Multiple linkage clustering
Residue fragment level		
Remington and Matthews (1980)	INTER	Superposition
Rackovsky and Goldstein (1988)	INTER	Superposition
Vrend and Sander (1991)	INTER	Superposition Clustering
Holm and Sander (1993) (DALI)	INTRA	Distance matrices Combinatorics Monte Carlo optimization
Lessel and Schomburg (1994) (in BRAGI)	BOTH	Distance matrices Superposition Clustering
Feng and Sippl (1996) (ProSup)	INTER	Dynamic programming
Shindyalov and Bourne (1998) (CE)	BOTH	Combinatorial extension
Secondary structure level		
Richards and Kundrot (1988)	INTRA	Distance matrices
Abagyan and Maiorov (1988) (FAESAR)		Feature matrix Superposition
Grindley *et al.* (1993) (PROTEP)	INTRA	Graph theoretic
Rufino and Blundell (1994)	INTRA	Graph theoretic Dynamic programming
Multiple element levels		
Sali and Blundell (1990) (COMPARER)	BOTH	Dynamic programming Simulated annealing
Madej, Gibrat and Bryant (1995) (VAST)	BOTH	Graph theoretic (fast search) Monte Carlo (refinement)

NAME	Description	Class	Type	Flexible
DeepAlign[3]	Protein structure alignment beyond spatial proximity (evolutionary information and hydrogen-bonding are taken into consideration)	Cα + Seq	Pair	No
MICAN[6]	MICAN can handle **M**ultiple-chains, **I**nverse alignments, **C** α only models, **A**lternative alignments, and **N**on-sequential alignments	Cα	Pair	No
TopMatch[2]	Protein structure alignment and visualization of structural similarities; alignment of multiprotein complexes	Cα	Pair	No
TS-AMIR	**T**opology **S**tring **A**lignment **M**ethod for **I**ntensive **R**apid comparison of protein structures	Geometry	Pair	nil
msTALI	**m**ultiple **s**Tructure **ALI**gnment	Cα & Dihed & SSE & Surf	Multi	Yes
mulPBA	**mul**tiple **PB** sequence alignment	PB	Multi	Yes
SAS-Pro	**S**imiltaneous **A**lignment and **S**uperimposition of **PRO**teins	???	Pair	Yes
MIRAGE-align	**M**atch **I**ndex based structural alignment method	SSE & PPE	Pair	No
SPalign	**S**tructure **P**airwise **align**ment	Cα	Pair	No
Kpax	Fast Alignments using Gaussian Overlap	Other	Pair	No
3DCOMB[4]	extension of DeepAlign	Cα	Multi	No
TS-AMIR[5]	A topology string alignment method for intensive rapid protein structure comparison	SSE & Cα	Pair	No
EpitopeMatch	discontinuous structure matching; induced fit consideration; flexible geometrical and physicochemical specificity definition; transplantation of similar spatial arrangements of amino acid residues	Cα-AllA	Multi	Yes
CLICK	Topology-independent 3D structure comparison	SSE & Cα & SASA	Pair	Yes
3D-Blast	Comparing three-dimensional shape-density	Density	Pair	No
DEDAL	**DE**scriptor **D**efined **AL**ignment	SSE & Cα & C-Map	Pair	Yes
deconSTRUCT	Database search on substructural level and pairwise alignment.	SSE	Multi	No
ProBiS	Detection of Structurally Similar **Pro**tein **Bi**nding **S**ites by Local Structural Alignment	Surf	Pair	Yes
ALADYN	**Dyn**amics-based **A**lignment: superposing proteins by matching their collective movements	Cα	Pair	No
SA Tableau Search	Fast and accurate protein substructure searching with simulated annealing and GPUs	SSE	Pair	No
RCSB PDB Protein Comparison Tool	Provides CE, FATCAT, CE variation for Circular Permutations, Sequence Alignments	Cα	Pair	yes
Smolign	**S**patial **mo**tifs based protein structural **align**ment	SSE & C-Map	Multi	Yes
STON	NA	Cα	Pair	No
MAX-PAIRS	NA	Cα	Pair	No
QP Tableau Search	Tableau-based protein substructure search using quadratic programming	SSE	Pair	No
MISTRAL	Energy-based multiple structural alignment of proteins	Cα	Multi	No
GANGSTA+	Combinatorial algorithm for nonsequential and gapped structural alignment	SSE/C-Map	Pair	No
BLOMAPS	Conformation-based alphabet alignments	Cα	Multi	nil
CLEPAPS	Conformation-based alphabet alignments	Cα	Pair	nil
Matt	**M**ultiple **A**lignment with **T**ranslations and **T**wists	Cα	Multi	Yes
RAPIDO	**R**apid **A**lignment of **P**rotein structures **I**n the presence of **D**omain m**O**vements	Cα	Pair	Yes
ComSubstruct	Structural Alignment based on Differential Geometrical Encoding	Geometry	Pair	Yes
Fr-TM-align	**Fr**agment-**TM**-score based protein structure **align**ment	Cα	Pair	no
TOPS+ COMPARISON	Comparing topological models of protein structures enhanced with ligand information	Topology	Pair	Yes
TOPS++FATCAT	**F**lexible Structure **A**lignmen**T** by **C**haining **A**ligned Fragment Pairs Allowing **T**wists derived from **TOPS+** String Model	Cα	Pair	Yes
TABLEAUSearch	Structural Search and Retrieval using a Tableau Representation of Protein Folding Patterns	SSE	Pair	No
MSVNS for MaxCMO	A simple and fast heuristic for protein structure comparison	C-Map	Pair	No
PyMOL	"super" command does sequence-independent 3D alignment	Protein	Hybrid	No
EXPRESSO	Fast Multiple Structural Alignment using T-Coffee and Sap	Cα	Multi	nil
CAALIGN	Cα Align	Cα	Multi	nil
Matchprot	Comparison of protein structures by growing neighborhood alignments	Cα	Pair	No
ProCKSI	**Pro**tein (Structure) **C**omparison, **K**nowledge, **S**imilarity and **I**nformation	Other	Pair	No
SARST	**S**tructure similarity search **A**ided by **R**amachandran **S**equential **T**ransformation	Cα	Pair	nil
MolLoc	**Mo**lecular **Loc**al Surface Alignment	Surf	Pair	No
SABERTOOTH	Protein Structural Alignment based on a vectorial Structure Representation	Cα	Pair	Yes
SALIGN	Sequence-Structure Hybrid Method	Seq	Multi	No
ProSMoS	**Pro**tein **S**tructure **Mo**tif **S**earch	SSE	Pair	No
MUSTANG	**MU**ltiple **ST**ructural **A**lig**N**ment Al**G**orithm	Cα & C-Map	Multi	nil

Sigue la teoría?....ZZzzzZZ

La bioinformática estructural abarca...

- Todo lo anterior.
- Base de datos (como casi toda rama de la bioinformática)
- Predicción de estructura secundaria a partir de una secuencia (GORI-IV, PHD, PSIPRED, JPRED..)
- Asignamiento de fold para búsqueda de homólogos remotos y predicción de estructura (HHPRED, PHYRE,Threading, Profiles…)
- Predicción de estructura terciaria (Modelado por homología y *ab initio*)
- Métodos de dinámica molecular.
- Docking de biomoléculas y compuestos químicos para el descubrimiento de nuevas drogas.
- Interacciones proteína-proteína.