

# MODELISATION MOLECULAIRE

## INTRODUCTION: un peu d'histoire

- Expérimentalement
  - *Début cinquante:*
    - *Séquence primaire d'une protéine : l'insuline par Sanger*
  - *Début soixante:*
    - *Structure 3D d'une protéine: hemoglobine (Kendrew et al) et myoglobine Perutz et al.*
    - *1965 :le lysozyme.*
- Théoriquement
  - Pauling et Corey:
    - Structure secondaire
  - 1965: Ramachandran
    - Carte  $[\varphi, \psi]$  acides aminés modèle de sphères dures.

# ***MODELISATION MOLECULAIRE***

Expérimentalement

Depuis :

15000 structures de macromolécules biologiques : protéines, fragments d'ADN, complexes ADN-protéine, complexes membranaires. Ribosome

Théoriquement:

Depuis:

Etude de macromolécules: repliement de la microvilline sur  $1\mu\text{s}$  ..... /.... A suivre ...

# ***MODELISATION MOLECULAIRE***

## **DEFINITION**

- **MODELE**



Représentation physique  
« modèle » i.e  
simplifié des  
interactions

- **MOLECULE**



Ensemble d'atomes  
interagissant par des  
liaisons covalentes ou  
non.

# AVANT

- ***Objectif : Comprendre et éventuellement prédire les propriétés conformationnelles, énergétiques, dynamiques des molécules.***
- Simulation de petites molécules à l'aide de méthodes quantiques ab initio, semi empiriques, empiriques.
- Domaine de prédilection des chimistes et physiciens. Calculs très longs. Centres de calcul.

# *CE QUI A CHANGE ?*

- La puissance de calcul ! (bien sur !)
- La nature des machines -beaucoup de PC
- Mais surtout : les données disponibles.

=> Elargissement du domaine grâce à de nouvelles compétences ont été utilisées

- Statistique
- Algorithmie.

=> ***Naissance de la Bioinformatique***

***CE QUI DEMEURE - L'OBJECTIF***

# *LES TECHNIQUES EXPERIMENTALES*

- La plus prisée : la radiocristallographie RX
  - "Observe" la densité électronique de la molécule
  - Nécessite un cristal. Etape très longue dans certains cas (Prot.Memb.).
  - Fragilité du cristal lors de l'expérience.
  - Problème des phases:
    - Les expériences fournissent des intensités mais pas les phases. Sans les phases pas de coordonnées !
  - Rayonnement synchrotron facilite la résolution

# *Radiocristallographie RX*

## ***Avantage:***

- Fournit la position des atomes (x,y,z)
- Fournit une mesure pour estimer la qualité des données : la résolution et le facteur R.
- Les tailles des systèmes examinés ne cessent de s'accroître.
- Plus le nombre de structures augmentent, plus le nombre de structures protéiques résolubles augmentent ..... / MAIS .....

# *LA RESONANCE MAGNETIQUE NUCLEAIRE*

## *incipe:*

Résonance magnétique des noyaux dans un champ magnétique

Utilisation d'isotopes.  $^1\text{H}$ ,  $^{15}\text{N}$ ,  $^{13}\text{C}$

Déplacement chimique du proton = fonction de l'environnement de l'atome.

Avancée récente : RMN haute résolution

- Effet Overhauser Nucléaire (NOE) : couplage direct entre noyaux -> structure secondaire
- Spectre 2D de type COSY (Correlation Spectroscopy): transfert entre noyaux.
- RMN 3D

# *LA RESONANCE MAGNETIQUE NUCLEAIRE*

## **Avantage**

Pas besoin de cristal !

Protéine en solution

Propriétés dynamiques

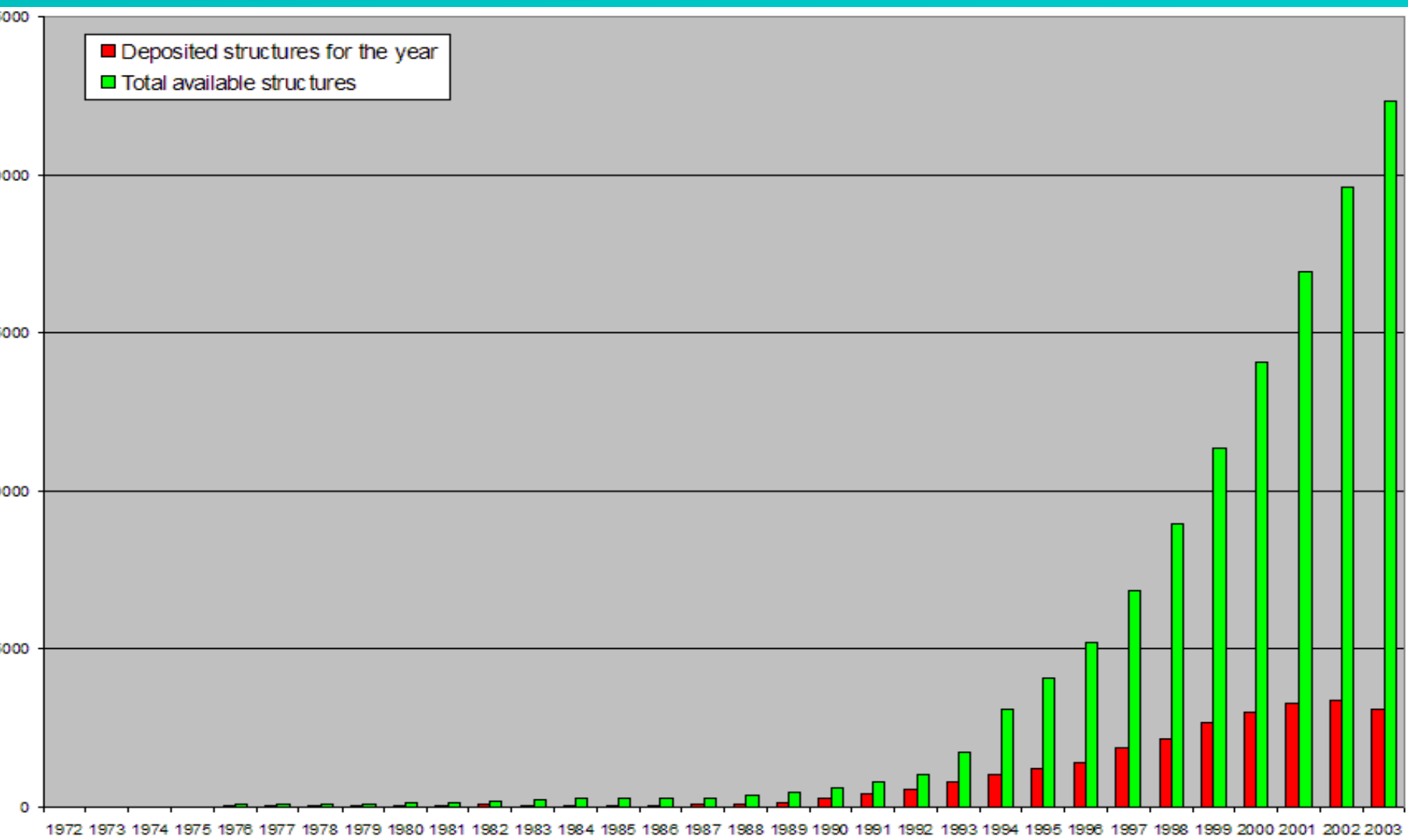
Limites :

Taille protéine

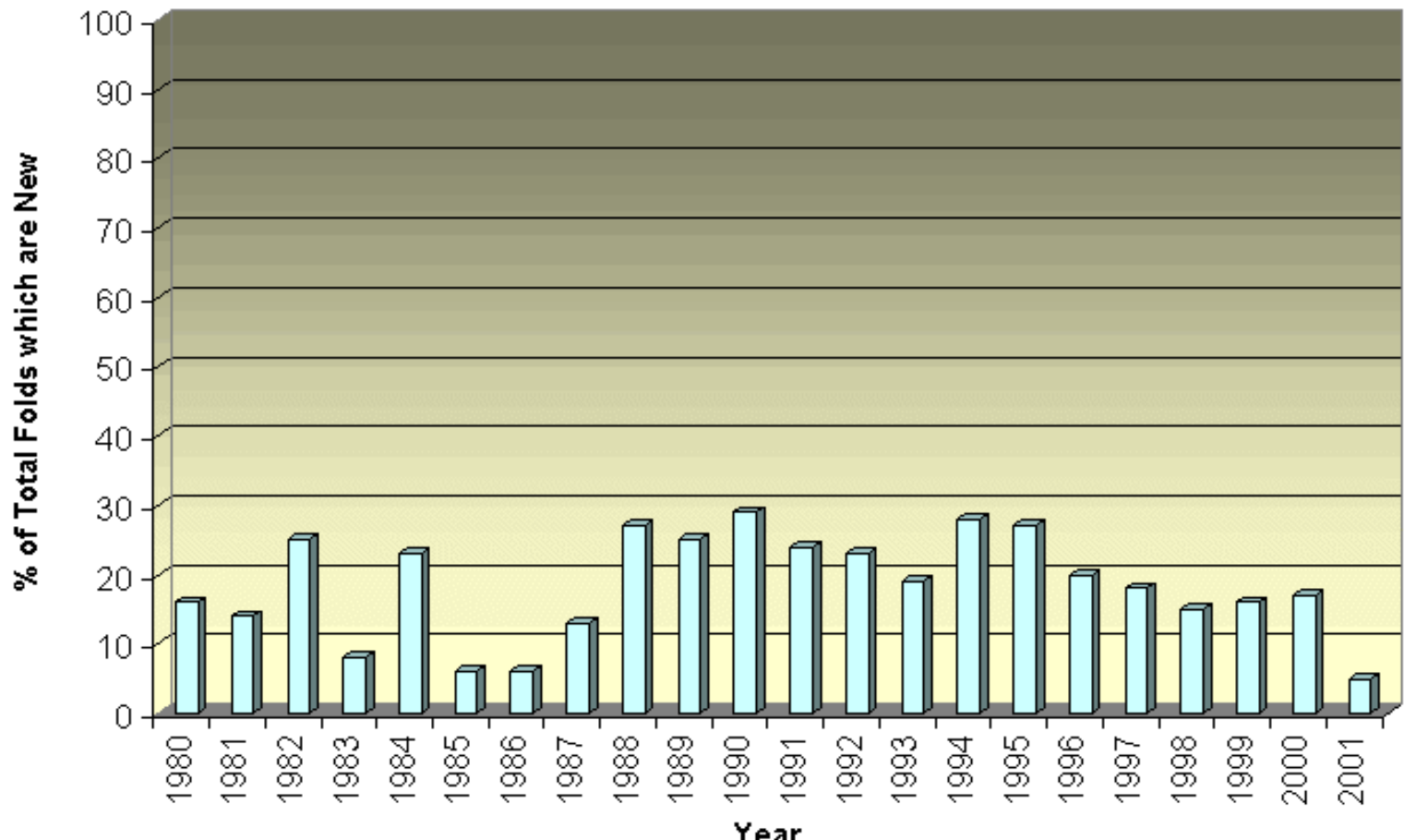
Pas de critère quantitatif de la qualité.

# *LES STRUCTURES RESOLUES*

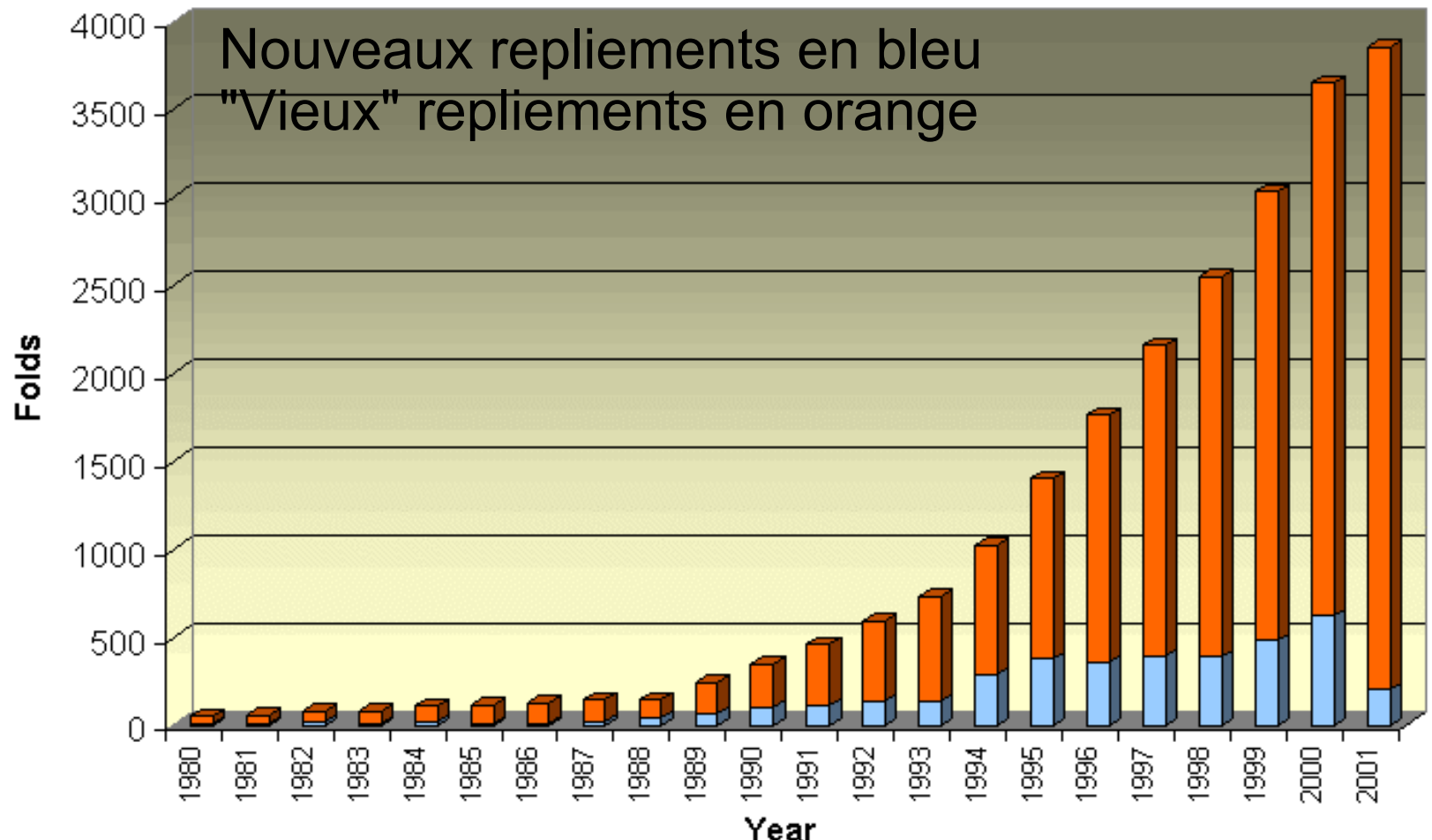
## *La Protein Data Bank*



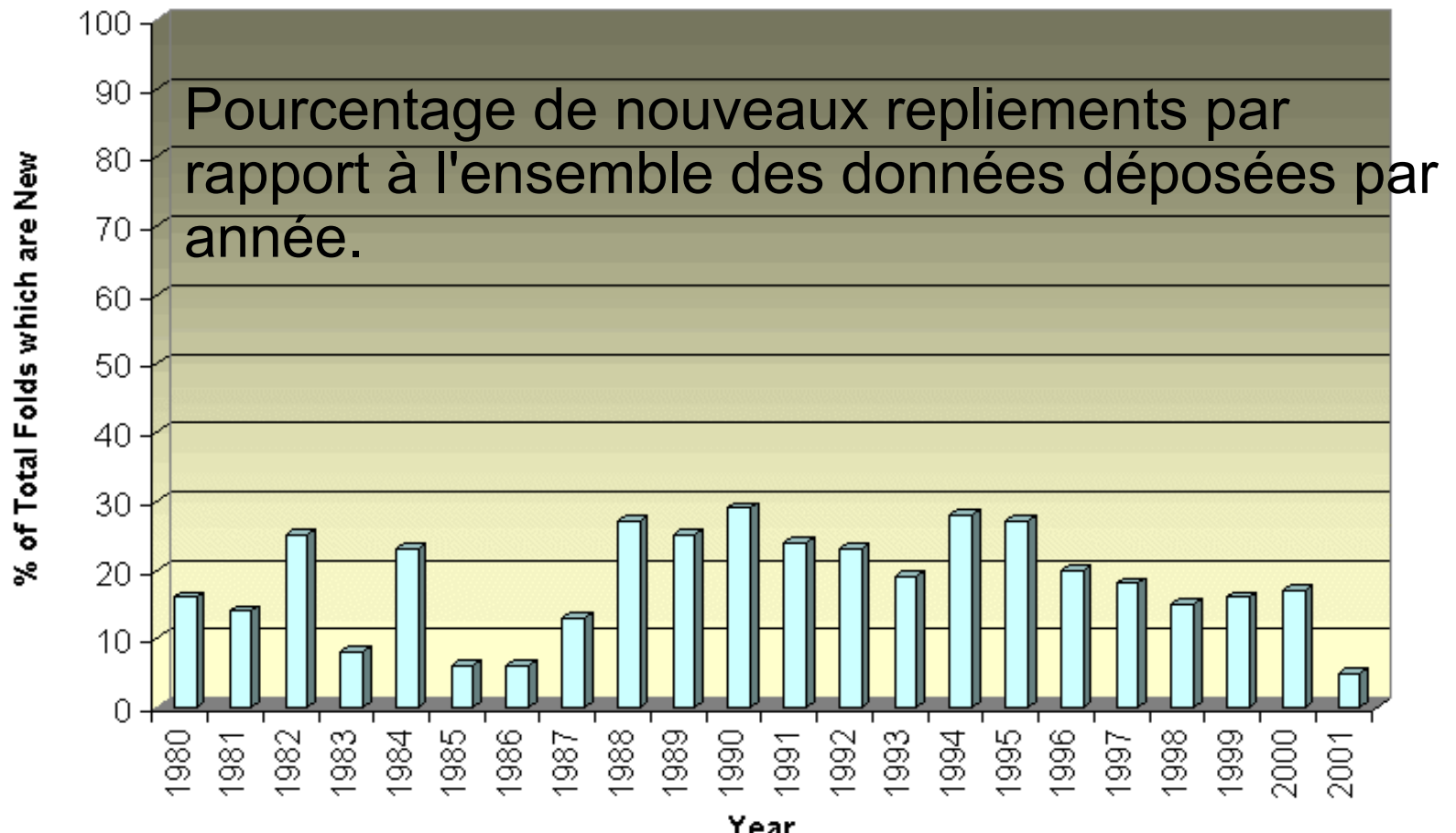
# *La Protein Data Bank*



# *La Protein Data Bank*



# *La Protein Data Bank*



# *Septembre 2003*

## **PDB Holdings List: 16-Sep-2003**

		Molecule Type				
		Proteins, Peptides, and Viruses	Protein/Nucleic Acid Complexes	Nucleic Acids	Carbohydrates	total
Exp.	X-ray Diffraction and other	17566	850	689	14	19119
Tech.	NMR	2755	95	543	4	3397
	<b>Total</b>	20321	945	1232	18	<b>22516</b>

# *REPLIEMENTS*

- Description d'un repliement ? Squelette polypeptidique. Usuellement d'après l'enchaînement des structures secondaires.
- Permet de détecter des parentés fonctionnelles alors que les séquences ont profondément divergées.
- Les séquences apparentées partagent un repliement similaire.
- Le nombre de repliements possibles pour l'ensemble des protéines est un nombre fini et "faible" à l'échelle du nombre de séquences

# ***REPLIEMENTS***

- Reconnaître de nouveaux repliements ?  
**=> Comparer des repliements.**
- Constitution de bases de données classant les structures protéiques.
- SCOP, FSSP (DALI) , CATH
- Et aussi ...../ des Modèles (MODBASE).

# *SCOP*

## *Classification hiérarchique*

<http://scop.mrc-lmb.cam.ac.uk/scop/>

- Construction manuelle par inspection visuelle et comparaison de structures
- Classification pour refléter à la fois la parenté structurale et d'évolution
- Famille : Relation d'évolution claire. En gros identité séquence > 30%. Dans certains cas parenté fonctionnelle et structurale évidente en l'absence de similitude de séquence détectable. ex les globines
- Superfamille: Probablement un ancêtre commun. Identité de séquence faible, mais parenté fonctionnelle et structurale assez marquée.

# *SCOP*

## *Classification hiérarchique*

- Famille de repliement: Similitude structurale majeure. Même type de structures secondaires organisées de manière similaire et ayant la même topologie des connexions
- Classe de repliement: Composition en structures secondaires

# SCOP

## Mars 2003

SCOP: STRUCTURAL CLASSIFICATION OF PROTEINS: 1.93 RELEASE

18946 PDB Entries (1 March 2003). 49497 Domains. 28 Literature References  
(excluding nucleic acids and theoretical models)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	171	286	457
All beta proteins	119	234	418
Alpha and beta proteins (a/b)	117	192	501
Alpha and beta proteins (a+b)	224	330	532
Multi-domain proteins	39	39	50
Membrane and cell surface proteins	34	64	71
Small proteins	61	87	135
Total	765	1029	2164

# *FSSP*

## *Fold classification based on Structure-Structure alignment of Proteins*

<http://www2.ebi.ac.uk/dali//fssp/>

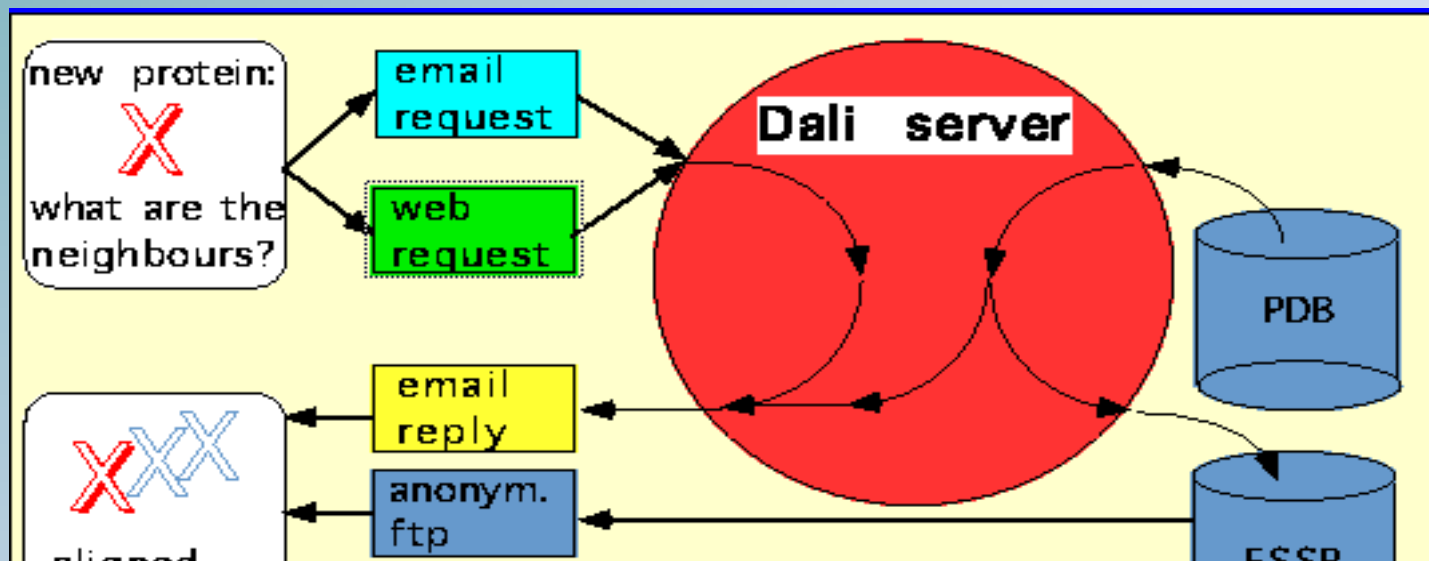
- Basée sur une comparaison exhaustive toutes contre toutes des différentes structures de la PDB. La classification et les alignements sont actualisés régulièrement grâce au moteur DALI
- Définition des domaines

# FSSP

## *Fold classification based on Structure-Structure alignment of Proteins*

<http://www2.ebi.ac.uk/dali//fssp/>

- Basée sur une comparaison exhaustive toutes contre toutes des différentes structures de la PDB. La classification et les alignements sont actualisés régulièrement grâce au moteur DALI

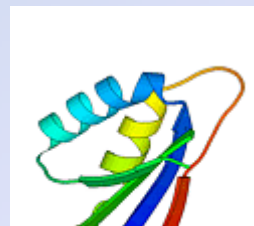
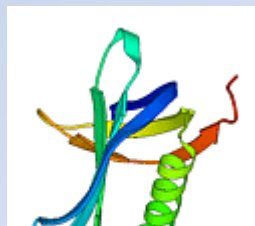
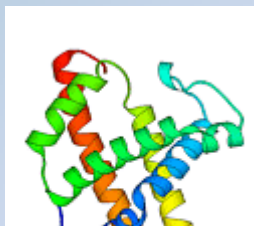


# FSSP

## *Fold classification based on Structure-Structure alignment of Proteins*

<http://www2.ebi.ac.uk/dali//fssp/>

- Définition des domaines: Automatiquement selon des critères de récurrence et de compacité
  - 1er niveau: Composition en str. sec. et en motifs supersecondaires
  - 5 régions "attractrices" ont été définies et sont caractérisées par les repliements suivants.

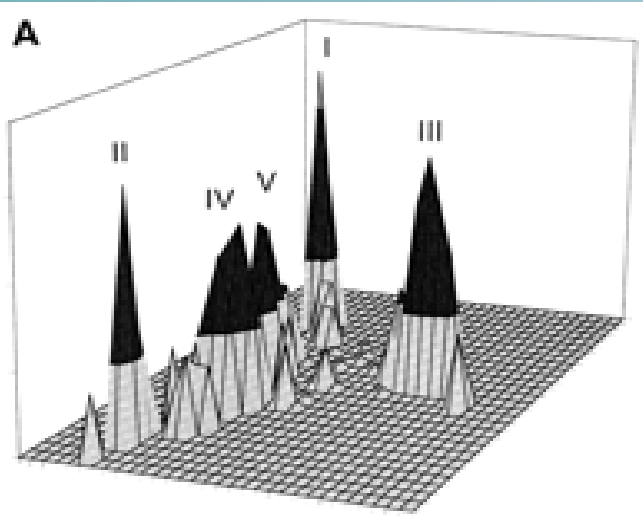


# FSSP

## *Fold classification based on Structure-Structure alignment of Proteins*

<http://www2.ebi.ac.uk/dali//fssp/>

- A chaque domaine est attribué un nombre DC\_l\_m\_n\_p avec l, région attractrice du repliement, m, topologie de repliement, n famille fonctionnelle et n famille de séquences.



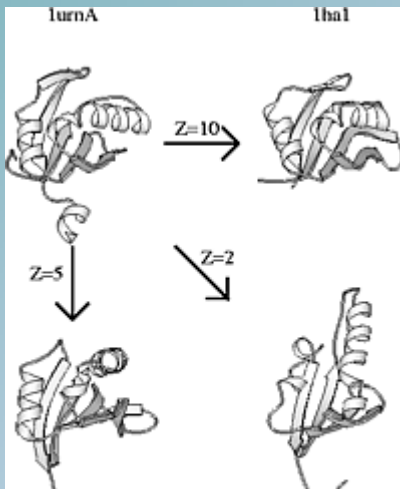
- Si pas de préférence pour une des régions attractrices , 6ème classe.

nsité de distribution des domaines

## *Fold classification based on Structure-Structure alignment of Proteins*

- 2ème niveau : type de repliement

Défini comme des clusters de voisins structuraux dans l'espace des repliements avec une moyenne de Zscore par paire  $> 2$



Type de repliement  
Voisins structuraux de 1urnA  
et 1mli

## *Fold classification based on Structure-Structure alignment of Proteins*

- 3ème niveau : famille fonctionnelle  
Branches du dendogramme de repliement où toutes les paires ont une moyenne élevée d'être homologue, prédite par réseau de neurone
- 4ème niveau: famille de séquence basée sur un alignement des séquences avec un seuil de 25% d'identité

# CATH

<http://www.biochem.ucl.ac.uk/bsm/cath/>

- Classification en Classe (C), Architecture (A), Topology (T) et superfamille homologue (H)

*Classe*: basée sur contenu en structure secondaire (attribuée automatiquement pour 90% des protéines)

*Architecture*: décrit l'orientation grossière des str.sec, indépendamment des connexions (manuel)

*Topology*: regroupement des structures selon la topologie des connexions et le nombre de str. sec.

*Superfamille Homologue*: regroupement selon des similitudes très marquées de fonction et de structures.

H et T sont basées à la fois sur des comparaisons de structures et de séquences

# CATH

11/08/2003: CATH2.5.0

	A	T	H	S	N	I	D
mainly Alpha	5	228	433	957	1640	3611	9013
mainly Beta	19	139	286	961	2240	4605	12962
Alpha Beta	12	361	659	2008	3444	7873	20411
new Secondary Structures	1	85	89	110	208	345	843

# CATH

C	A	T	H	S	N	I	D
Mainly Alpha	5	228	433	957	1640	3611	9013
Mainly Beta	19	139	286	961	2240	4605	12962
Alpha Beta	12	361	659	2008	3444	7873	20411
Few Secondary Structures	1	86	90	111	209	346	844
Preliminary single domain assignments	1	363	368	432	486	787	1540
Multi-domain domains	2	28	28	31	40	73	132
CATH-35 Sequence families	1	688	688	688	774	1485	2422
Fragments from multi-chain	1	25	25	27	33	54	122

# *VAST*

## *Vector Alignment Search Tool*

<http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html>

- Serveur de recherche de comparaison structurale
- Les alignements de toutes les structures protéiques de la pdb ont été réalisés.
- Le coeur de l'approche est basée sur la définition d'unité de similitude 3D comme les paires d'éléments de structure secondaire qui ont le même type, la même orientation et la même connectivité.