

Bioinformatique structurale (approfondissement)

rappels

Bioinformatique + modélisation moléculaire + méthode spectroscopique (RX)

B. Offmann

V. Tran

M. Evain

Vinh TRAN

Localisation:

(Bureau 006 Bâtiment 25)

Coordonnées: tel: 02 51 12 57 57

mail: vinh.tran@univ-nantes.fr

Objectifs pédagogiques:

Interactivité, adaptation à la demande (CM + TP)

Inscrire un savoir dans un ensemble de connaissances (rappels ou nouveautés)
(accélération, renouvellement et mise à niveau)

Apprendre à rechercher et décrypter l'information scientifique

Aspects pratiques :

cours scindé en 2 années (M1 introduction générale avec Biostatisticiens)
(M2 approfondissement, rappels, ajouts, techniques)

Support cours (pdf disponibles)

Modélisation moléculaire

Mise en place de l'interactivité...

exercice pratique: (recherche des mots-clés)

qu'avez-vous retenu du cours M1 ?

Agencement M1-M2

Bioinformatique 3D : introduction

Relation avec la bioinformatique 1D (séquence) et la bioinformatique 2D (structure secondaire)

Informations structurales disponibles dans les bases généralistes (ex. PDB) ou spécialisées

Graphisme moléculaire :

représentation spatiale et stéréochimie

Mécanique moléculaire : champ de forces et énergies d'un système moléculaire

Bioinformatique 3D : approfondissement

Mécanique moléculaire : champ de forces et énergies d'un système moléculaire
Explorations conformationnelles et optimisations

Dynamique moléculaire

Compréhensions et prédictions fonctionnelles

Analyse des structures 3D : méthodes et outils

Notions de base (*rappels*)

Stéréochimie, généralités

Isomérisie de constitution et tautomérie

Stéréoisomérisie de configuration et de conformation

Géométrie moléculaire

Détermination de la géométrie d'une molécule

Structure chimique
(paramètres internes)

La mécanique moléculaire

énergie de liaison

énergie de flexion

énergie de torsion

énergie d'interactions non liantes

Physique

La dynamique moléculaire

Méthodologie de la dynamique moléculaire

Introduction aux différentes méthodes de dynamique moléculaire

Mathématiques

Introduction à la chimie quantique

Rappels cours année dernière ...

Concepts de la biologie

Définition: ? étude du monde vivant

⇒ caractéristique majeure: ?

extrême complexité

Facteurs de complexité ?

- * immensité des variables
- * interaction des variables
- * échelles des phénomènes
- * variabilité de comportement
- * perception dynamique

Critères antinomiques de la démarche de modélisation (pourquoi?)

Spécificité de la recherche en biologie...

Démarche scientifique ... plus difficile (nombre de paramètres et complexité...)

Accumuler : élargir la base d'information

- * explorer **systématiquement** (exhaustivement) les domaines connus
- * trouver des **nouvelles voies** d'exploration des domaines inconnus
- * **fiabiliser** l'information (vérifier, confronter, uniformiser)

Agencer, Trier : traitement de l'information

- * rejeter les **redondances**
- * **relier** les faits et variables entre eux (les échelles de perception)
- * créer des **outils de tri** (se faciliter le travail)
- * **accéder** à l'information (rentabiliser, 'ergonomiser' l'accumulation)

début de l'intelligence
l'intelligence scientifique
(nouvelle forme)

**Conceptualiser : transformer l'information en connaissances ⇒
modéliser les mécanismes de la vie**

- * **comprendre** les faits, leurs agencements,
- * **percevoir** les différentes échelles
- * **influer** le cours des faits : **prédire** les perturbations
- * **'utiliser'** la biologie

Finir d'être
intelligent...

**aller à l'essentiel,
utiliser la compréhension**

Contexte scientifique de la **biologie** : **rôle de l'informatique**

Conséquence de cette accélération de la technologie :



accélération des **disciplines liées à l'informatique**
(dont la **biologie** à travers la **bioinformatique**)

Nouvelle donne intellectuelle : **Maîtriser le flux d'informations**

- sa **gestion en temps réel** : **acquisition** (avec des outils adaptés)
tri (remise en cause du savoir)
- sa **synthèse** : raisonnement par abstraction (**modélisation**)
informations ↔ connaissances ↔ modèles

l'ère du haut débit

Quelques îlots de résistance...

- technique: **cristallogénèse** (**débit très moyen...**)
(maillon indispensable pour comprendre le repliement 3D des protéines)
- humaine : **intelligence** (conception de nouveaux modèles **à très faible débit...**)

Concepts de base de la biologie structurale: Quelles modélisations?

Définition du modèle ?

moyen **abstrait** de décrire un **système complexe** par une **représentation simplifiée**

⇒ schéma intellectuel très utile en biologie (cf. complexité)

conséquences de l'abstraction: **simplification** mais **perte d'informations**
et biais (techniques et compréhension)

Qualité d'un modèle ? (critères essentiels)

- capacité à sélectionner des **paramètres pertinents** pour décrire le phénomène
- **faible interaction** entre les paramètres (si possible...)
- capacités d'**explications** et de **prédictions** ex: drug design...

Rôle d'un modèle ?

fournir des **hypothèses**

Rôle du modélisateur ?

- vérifier la **cohérence des hypothèses** par rapports à la base expérimentale
- fournir des **modèles explicatifs et prédictifs**
- **faire tester ce modèle** par les biologistes (expériences)...

Concepts de base de la biologie: Quelles modélisations ?

Quelques exemples de modélisations :

- * modèles **mathématiques** de comportement des populations
ex. échelle macroscopique: relations des populations proies/prédateurs
ex. échelle moléculaire: analyse du contrôle métabolique
- * modèles '**symboliques**' pour un meilleur traitement des données
conversion d'une structure **3D** en représentation unidimensionnelle (**1D**)

- * modèles **géométriques** moléculaires
représentations de surface et architectures de protéines (**graphisme moléculaire**)
Conséquences du modèle énergétique

- * modèles **énergétiques** moléculaires
modélisation moléculaire (base empirique classique ou quantique)
- * modèles **statique** ou **dynamique**
concepts de **déformations** (mouvements intrinsèques) **robotique moléculaire**
et de la **chronologie** des mouvements (**dynamique moléculaire**)

Contexte du Web et Bioinformatique

Bases de données d'acides nucléiques et protéines:

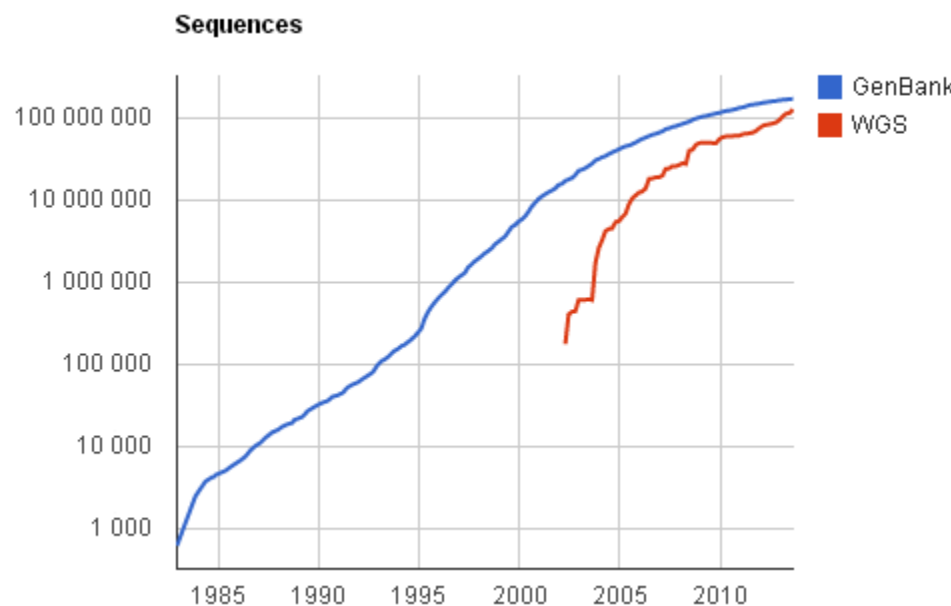
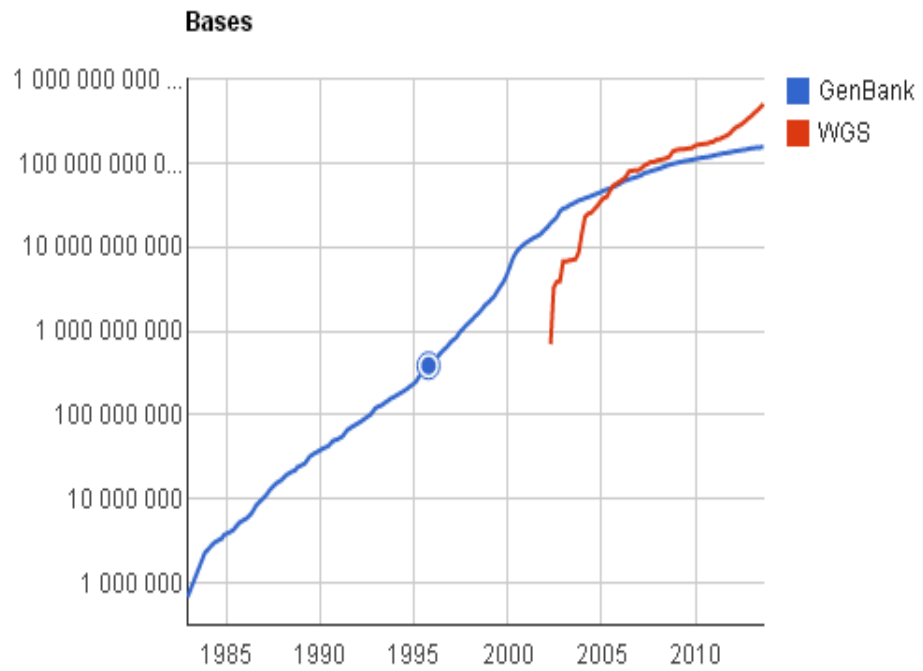
ère du haut débit, accumulation des données, gestion du flux des données

La réponse au gigantisme : Trois bases interconnectées et accessibles



Bases de données d'acides nucléiques et protéines : statistiques

Concernant le 1D...



From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months.

Bases de données de structures 3D

gène: données de type séquentiel (1D) pour l'information génétique
mais sous plusieurs formes: séquence ADN \Rightarrow séquences protéines

l'expression **fonctionnelle** de cette information génétique passe par les **protéines**

relations **structure-fonction** \Rightarrow **structure 3D** (architecture)

structure 3D (protéines essentiellement, ADN moins important)

Ce qui est nouveau (par rapport au 1D):

Rôle **informatique** ? : aide précieuse mais **moins 'définitive'** que le traitement 1D

- changement d'échelles de perception (moléculaire)
- simplifications précédentes moins efficaces (ex: code lettre)

- plutôt le domaine de la **modélisation moléculaire**
- intégrer de nouveaux concepts hors domaine mathématique (**physique**)
et dans le domaine mathématique (**topologie**, **robotique**, etc...)
- **tout est à faire...**

Contexte du Web

Bases de données de structures 3D: PDB Brookhaven National Laboratory (BNL)

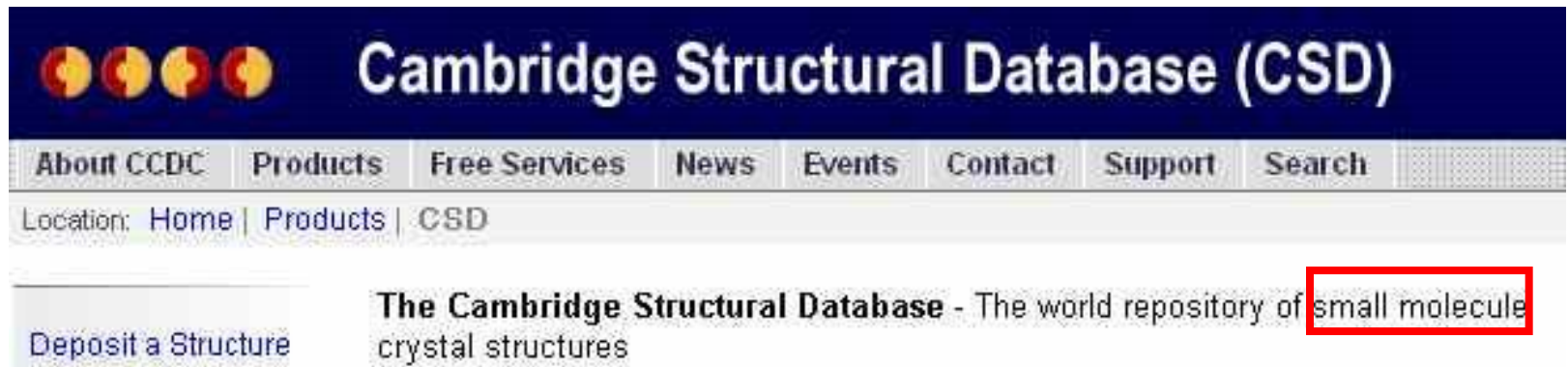
Base de données '**universelle**' pour les structures (de protéines)

- soumission obligatoire pour publier en cristallographie
- vérifications techniques lourdes (nombreuses modifications, 'releases')
- période de blocage (validations et exploitation des résultats en primeur)

Deposition Date: 19-Feb-1998

Release Date: 23-Mar-1999

Autres bases...



The screenshot shows the Cambridge Structural Database (CSD) website. At the top, there is a logo consisting of four overlapping circles in red, yellow, and blue, followed by the text "Cambridge Structural Database (CSD)". Below this is a navigation bar with links: "About CCDC", "Products", "Free Services", "News", "Events", "Contact", "Support", and "Search". A breadcrumb trail indicates the current location: "Location: Home | Products | CSD". On the left side, there is a button labeled "Deposit a Structure". On the right side, the text reads "The Cambridge Structural Database - The world repository of small molecule crystal structures", with the words "small molecule" highlighted by a red rectangular box.

Contexte du Web

Bases de données spécialisées (sous-classe de composés)

ex: CAZY <http://cazy.org>
(4 personnes à plein temps...)



Family GH11

Family GH11

CAZy Family Glycoside Hydrolase Family 11

Known Activities xylanase (EC 3.2.1.8).

Mechanism Retaining

Catalytic Nucleophile/Base Glu (experimental)

Catalytic Proton Donor Glu (experimental)

3D Structure Status Available (see PDB). Fold β -jelly roll

Clan GH-C

Note formerly known as cellulase family G.

Relevant Links HOMSTRAD; InterPro; PFAM; PROSITE

Statistics CAZY(218); GenBank/GenPept (317); Swissprot (135); PDB (35); 3D(20); cryst(2)

Protein	Organism	EC#	GenBank / GenPept	SwissProt	PDB / 3D
xylanase A (XynA)	<i>Aeromonas punctata</i> ME-1	3.2.1.8	D32065 BAA06837.1	Q43993	
xylanase	<i>Ascochyta pisi</i>	3.2.1.8	Z68891 CAA93120.1	Q00263	
β -1,4-xylanase (fragment)	<i>Ascochyta rabiei</i>	3.2.1.8	AJ245713 CAB53563.1	Q9UW04	

Toutes les informations disponibles (interconnexions) ... **dont les structures 3D**

Description détaillée Protein Data Base (PDB)

<http://www.pdb.org>

La recherche d'informations ... dont les coordonnées 3D (ou mot-clé)

et le reste ...

The screenshot shows the PDB website interface with several red annotations:

- A red box highlights the search bar at the top right, with a red arrow pointing to it from the text "coordonnées 3D (ou mot-clé)".
- A red box highlights the left sidebar menu, with a red arrow pointing to it from the text "et le reste ...".
- A red box highlights the "Molecule of the Month" section, with a red arrow pointing to it from the text "ce qui change....".

The website content includes:

- Search** | All Categories | e.g., PDB ID, molecule name, author
- Biological Macromolecular Resource**
- Full Description**
- Featured Molecules**
- Structural View of Biology**
- Molecule of the Month**
O-GlcNAc Transferase
Cells use many methods to control their proteins, to make sure that they perform their jobs when and where they are needed. Some are brutally irreversible, such as the continuous breakdown of obsolete proteins by the **ubiquitin/proteasome** system. Others, such as the modulation of enzyme function by allosteric motions, are far more subtle and respond to the second-by-second needs of the cell.
[Full Article](#)
- Protein Structure Initiative Featured System**
Bacterial Armor
Researchers at MCSG have revealed the inner workings of a surface layer protein, showing how bacteria attach their form-fitting protein coats.
[Full Article](#) | [Archive](#) | [PSI Structural Biology Knowledgebase](#)
- Explore Archive**
- Organism**
 - Homo sapiens (18472)
 - Escherichia coli (4485)

Protein Data Base (PDB)

Et le reste du monde?:



<http://www.ebi.ac.uk/pdbe/>

<http://www.pdbj.org/>

Nouveaux membres du RCSB

Europe

Asie

mais aussi des sites miroirs....

Communauté internationale:

le Net n'est pas que virtuel....

chercher la saturation moindre...

plusieurs milliers de connexions par jour...

utiliser les sites miroirs

et raisonner en fuseaux horaires

(valable pour bien d'autres bases!)

Protein Data Bank Mirror Sites

Select the server closest to you:

AR Argentina

AU Australia

BR Brazil

CN China

DE Germany

IL Israel

PL Poland

TW Taiwan

UK Cambridge

UK Hinxton

US Georgia

US New York

(Feedback to [Eric Martz](#))

Protein Data Base (PDB) : évolution des entrées

Octobre 2006

		Molecule Type				
		Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
Exp. Method	X-ray	30503	916	1406	28	32854
	NMR	4809	725	122	6	5662
	Electron Microscopy	91	10	29	0	130
	Other	75	4	3	0	83
	Total	35478	1655	1560	34	38729

septembre 2013

Exp.Method	Proteins	Nucleic Acids
X-RAY	77139	1481
NMR	8829	1044
ELECTRON MICROSCOPY	466	45
HYBRID	51	3
other	150	4
Total	86635	2577

Progression relativement très lente (en comparaison des séquences)...

Protein Data Base (PDB)

*Nécessité pour la MM de s'appuyer sur des bases **expérimentales fiables***

2 grandes catégories techniques

***cristallographie** (RX, neutrons, microscopie électronique)*

[protéines et complexes]

RMN

[fragments protéines, peptides]

+ quelques modèles théoriques [non comptabilisés]

*Base expérimentale étroite ⇒ **pour élargir cette base...***

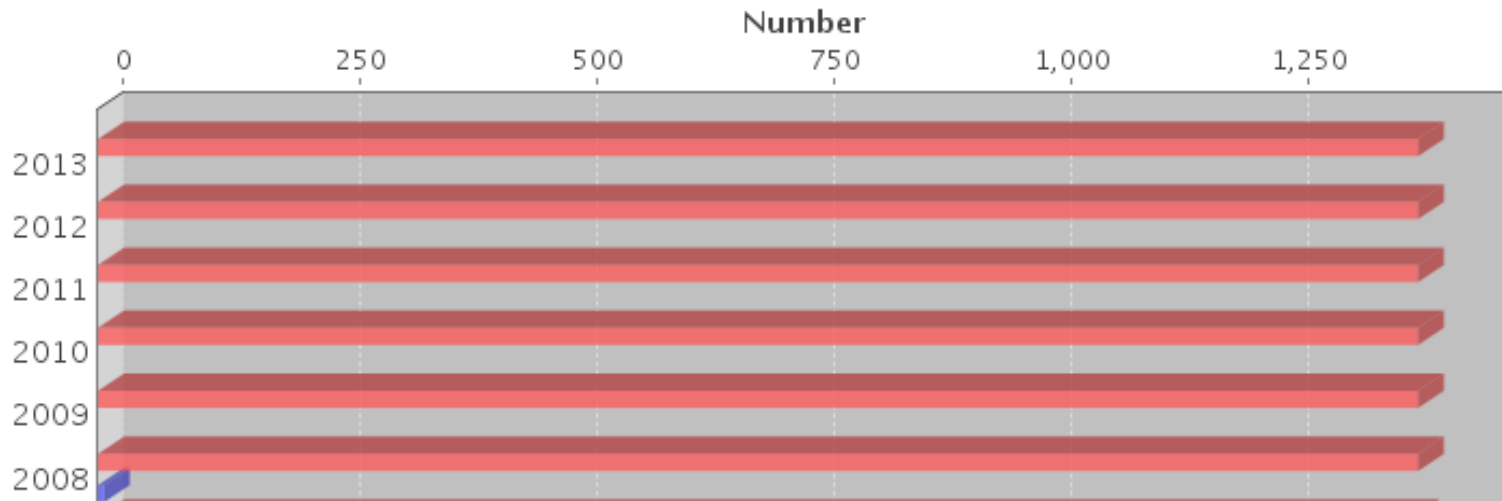
*utiliser la modélisation moléculaire pour construire des structures dérivées
par homologie (**structurale ou fonctionnelle**)*

Protein Data Base (PDB)

*Structures 3D disponibles par type de repliements (fold)
(extrait)*

Growth Of Unique Folds Per Year As Defined By SCOP (v1.75)

number of folds can be viewed by hovering mouse over the bar



depuis 2008 (~1300 folds) rien selon Scop

⇒ Y a-t-il d'autres classes?

Protein Data Base (PDB)

*Conclusions ... sur la relation
structure (3D) - fonction (biologie)*

*On semble avoir fait le tour des repliements
(façon de construire des architectures protéiques "stables")*

a/ stratégies possibles de constructions par homologies structurales (globales)

Mais plusieurs fonctions biologiques partagent la même architecture

*b/ pas de stratégies claires pour remonter à la fonction
(et encore moins à partir de la séquence...)*

*c/ Au-delà de l'architecture générale responsable de la stabilité, il y a des
zones (a priori) très ponctuelles (quelques résidus à des positions clés)
responsables de la fonctionnalité ⇒*

*possibilités de **stratégies spécifiques** sur des **comparaisons locales** (au sein
d'une même fonctionnalité)*

Protein Data Base (PDB)

Fichiers PDB venant de la cristallographie (RX)

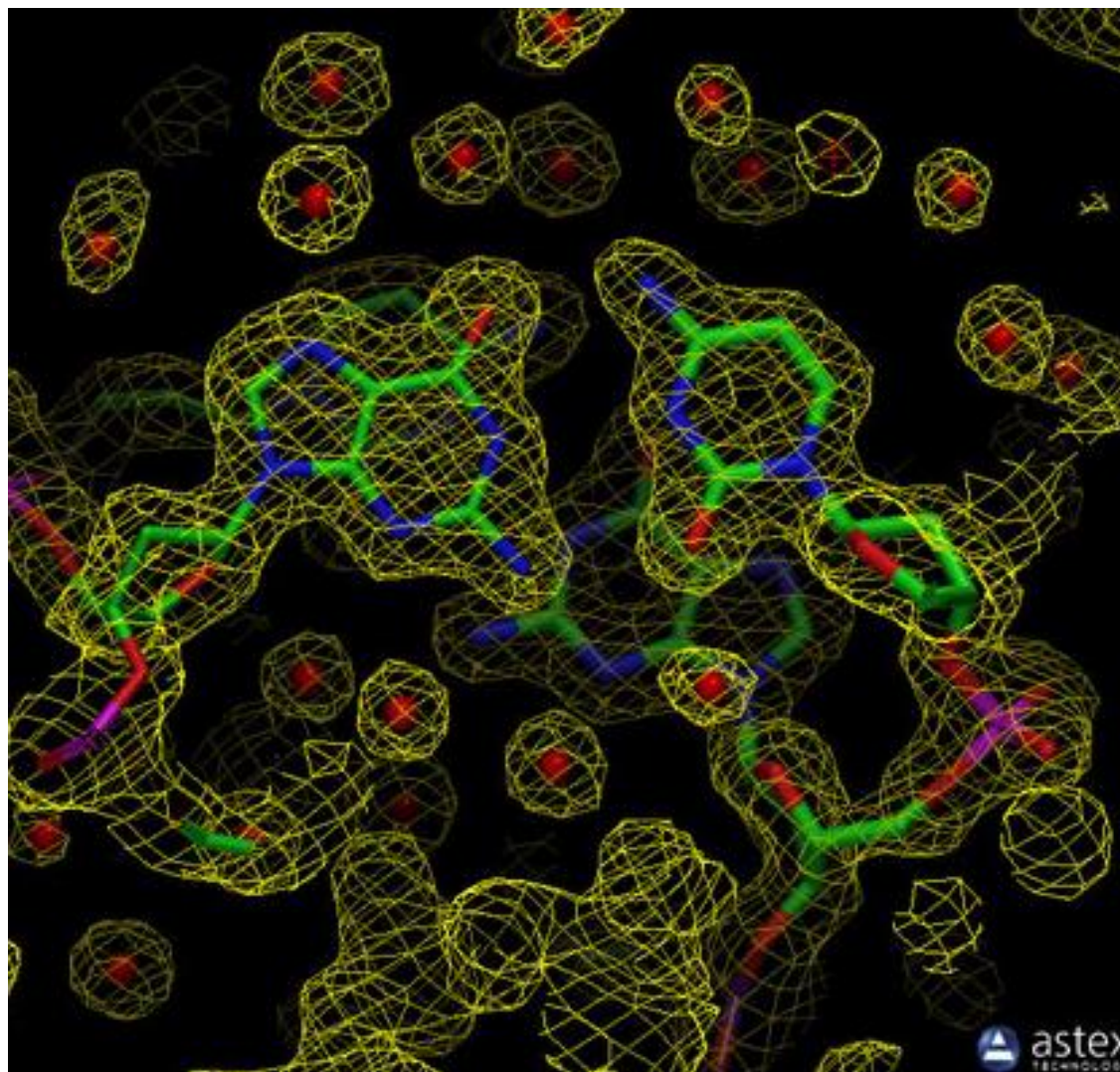
Domaine d'application : **plutôt des grosses molécules**

- un contexte 'figé' : **cristal** (même en présence de molécules d'eau)
- dans la **maille cristalline** : pratiquement toujours des **molécules d'eau** mais possibilité de trouver **plusieurs molécules** (différentes ou semblables) (les identifier notamment par leurs séquences)
- un **critère de qualité** : la **résolution** (ex: 1.5Å ou 2.8Å)
- des **contraintes techniques** : **imprécisions** sur les coordonnées déduites de nuages électroniques de diffraction
- des **contraintes biologiques** : **pas d'hydrogène**, **mobilité** de fragments \Rightarrow **résidus non localisés**



tenir compte de tous ces paramètres avant de faire de la modélisation moléculaire (notamment des calculs énergétiques)

RX



Ce qui est mesuré ... et ce qui est déduit...

Protein Data Base (PDB)

Fichiers PDB venant de la RMN

Domaine d'application : **plutôt des petites molécules (quelques centaines d'AA)**

- un contexte 'fluctuant' : **en solution** (mais pas de molécules d'eau repérées)

- **mesures structurales** : matrice de distances de **couplage proton-proton** ⇒ reconstitution de géométries **respectant plus ou moins ces critères de distances**

- **pas de critère de qualité** : la **résolution** n'a plus de signification

- des **contraintes techniques** : **hydrogènes explicites**
plusieurs conformations ou **conformation moyenne**

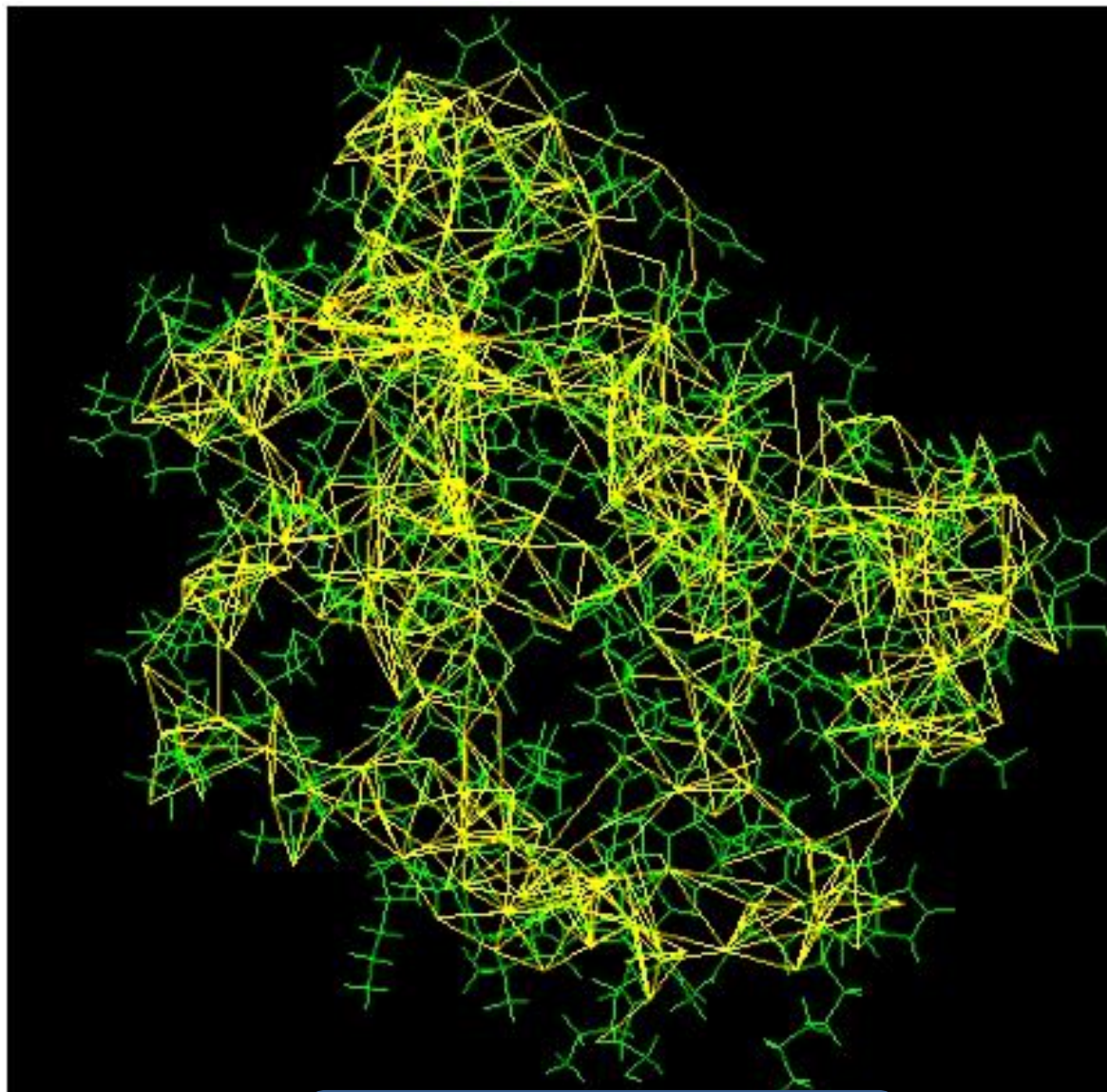
(accès la la mobilité moléculaire)

vision dynamique biaisée



tenir compte de tous ces paramètres avant de faire de la modélisation moléculaire (notamment des calculs énergétiques)

RMN



Ce qui est mesuré ... et ce qui est déduit...

Protein Data Base (PDB)

Un résumé:

STRUCTURE SUMMARY

- Protein Databank in Europe (PDBe)
- Protein Data Bank Japan (wwPDB Partner) (PDBj)
- Protein Interfaces, Surfaces and Assemblies (PISA)
- Molecular Modeling DataBase (NCBI/Entrez) (MMDB)
- PDBsum
- Jena Library
- PDBWiki
- Proteopedia
- OCA Browser (OCA)

PDBsum

Go to PDB code:

☒ Protein ☐ Ligands ☐ Clefts ☐ Links

Top page

Glycosyl hydrolase

PDB id: 1e4i

Asymmetric unit

Biological unit, dimer
- as defined in PDB file (see also PQS)

Jmol Syrup

Contents

- Description
 - [Header details](#)
 - [Header records](#)
 - [References](#)
 - [PROCHECK](#)
- Protein chain
 - [447 a.a.](#)
- Ligands
 - [G2F](#)
 - [NFG](#)
- Waters ×234

PDB id: **1e4i**

Name: **Glycosyl hydrolase**

Title: 2-deoxy-2-fluoro-beta-d-glucosyl/enzyme intermediate complex of the beta-glucosidase from bacillus polymyxa

Structure: Beta-glucosidase. Chain: a. Engineered: yes. Mutation: yes

Source: Bacillus polymyxa. Organism_taxid: 1406. Atcc: 842. Plasmid: puc derivative. Expressed in: escherichia coli. Expression_system_taxid: 562.

Biological unit: Homo-octamer (from PDB file)

UniProt: [P22073](#) (BGLA_PAEPO)

Seq:

Quick links

- [RCSB](#)
- [PDBe](#)
- [SRS](#)
- [MMDB](#)
- [JenaLib](#)
- [OCA](#)
- [PDBWiki](#)
- [Proteopedia](#)
- [CATH](#)
- [SCOP](#)
- [FSSP](#)
- [HSSP](#)
- [PDBSWS](#)
- [PQS](#)
- [CSA](#)
- [PROCOGNATE](#)
- [ProSAT](#)
- [Whatcheck](#)

Procheck

Clefts

un nouveau champ d'investigation...

Protein Data Base (PDB)

À partir du code PDB

Information sur chaque structure : **Download/Display** file (suite

consultation fichier PDB (texte)

Encore beaucoup d'informations pratiques:
séquences,
résidus manquants,
ponts disulfures,
éléments de structures secondaires
(selon les auteurs)

...

mais surtout:

coordonnées cartésiennes {X,Y,Z}

dans le format PDB [≠ coordonnées internes]

```
SHEET 1 K 5 LYS B 347 ALA B 355 0
SHEET 2 K 5 GLN B 336 GLU B 344 -1 N TRP B 337 0 TYR B 354
SHEET 3 K 5 ILE B 326 GLY B 333 -1 N ILE B 326 0 TYR B 342
SHEET 4 K 5 LYS B 388 LEU B 391 1 0 LYS B 388 N ALA B 327
SHEET 5 K 5 GLU B 413 PHE B 416 1 0 GLU B 413 N PHE B 389
CISPEP 1 PRO A 225 PRO A 226 0 -0.07
CISPEP 2 PRO A 420 PRO A 421 0 0.83
CRYST1 135.100 112.700 75.100 90.00 90.00 90.00 P 21 21 21 4
ORIGX1 1.000000 0.000000 0.000000 0.000000
ORIGX2 0.000000 1.000000 0.000000 0.000000
ORIGX3 0.000000 0.000000 1.000000 0.000000
SCALE1 0.007402 0.000000 0.000000 0.000000
SCALE2 0.000000 0.008873 0.000000 0.000000
SCALE3 0.000000 0.000000 0.013316 0.000000
ATOM 1 N PRO A 4 14.340 -55.053 42.032 1.00101.59 N
ATOM 2 CA PRO A 4 15.674 -55.453 41.548 1.00 98.73 C
ATOM 3 C PRO A 4 16.009 -54.938 40.158 1.00 96.22 C
ATOM 4 O PRO A 4 15.939 -55.663 39.163 1.00 96.99 O
ATOM 5 CB PRO A 4 15.717 -56.967 41.605 1.00 98.61 C
ATOM 6 CG PRO A 4 14.850 -57.213 42.828 1.00102.59 C
ATOM 7 CD PRO A 4 13.686 -56.206 42.678 1.00103.68 C
```

→ "traducteur" d'informations moléculaires (logiciel de graphisme)

Conversion coordonnées cartésiennes en coordonnées internes

Protein Data Base (PDB)

Informations structurales : classification

Summary Sequence Annotations Seq. Similarity 3D Similarity Literature Biol. & Chem. Methods Geometry **Links**

STRUCTURE CLASSIFICATION AND COMPARISON

- Structural Classification of Proteins (SCOP)
- Protein Structure Classification (CATH)
- Vector Alignment Search Tool (VAST)
- Flexible structure Alignment by Chaining Aligned fragment pairs allowing Twists (FATCAT)
- DALI
- SUPERFAMILY

* Dans quelle famille ou sous-famille de repliements appartient la structure étudiée ?

* Y a-t-il des homologues intéressants ?

Protein Data Base (PDB)

classification

SCOP

Structural Classification Of Proteins

CATH

Classe (C) : selon la composition des structures secondaires et l'empaquetage au sein de la structure (automatisé à 90% selon Michie et al. 1996)

Architecture (A) disposition relative des domaines (orientation structures secondaires sans connectivité) (manuel en cours d'automatisation)

Topology (T) (Fold family),
repliement selon la forme et la connectivité des structures secondaires

Homologous Superfamily (H)
regroupement selon en **ancêtre supposé commun** (concept d'évolution)
homologie de séquences (de fonction?)

Sequence families (S)
selon le pourcentage d'identité des séquences (>35%).
grande similarité de structures et de fonctions.

Protein Data Base (PDB)

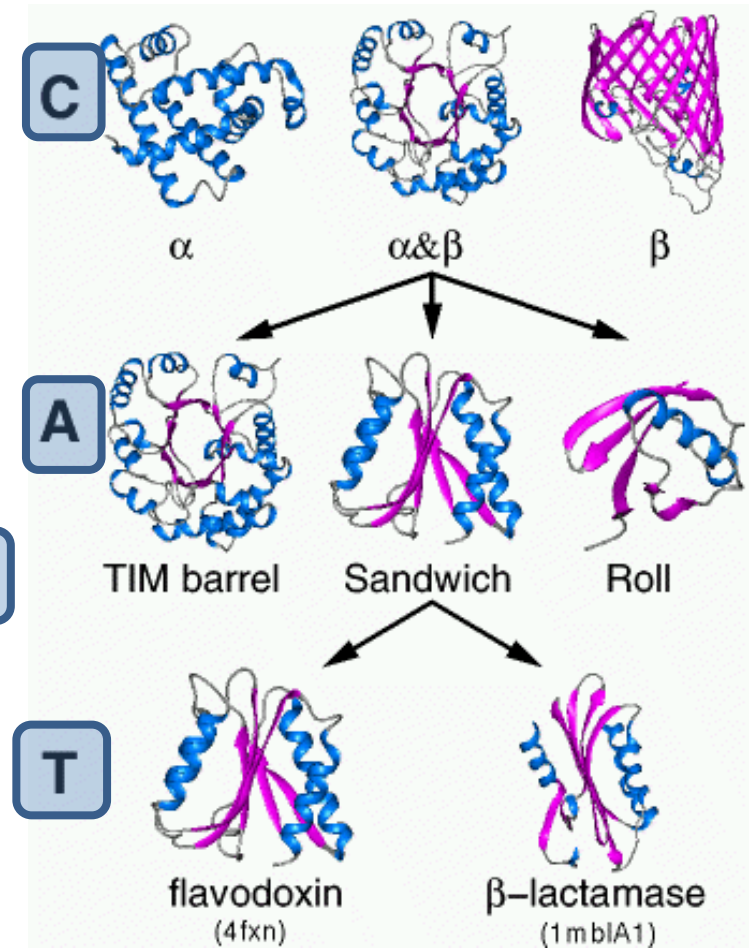
Information sur chaque structure : **CATH** (suite)

Philosophie:

CATH is a novel hierarchical classification of protein domain structures, which clusters proteins at four major levels, Class(C), Architecture(A), Topology(T) and Homologous superfamily (H).

Exemple d'arborescence
à partir de la classe 2
(**Mixed Alpha Beta**)
3 architectures
et 2 topologies de sandwich

trois premiers niveaux liés à la forme



Protein Data Base (PDB)






























Information sur chaque structure : **Structural Neighbors** (suite) **SCOP**

Philosophie: similaire à CATH

(sauf que le premier niveau descriptif est plus détaillé)

Root: scop

Classes:

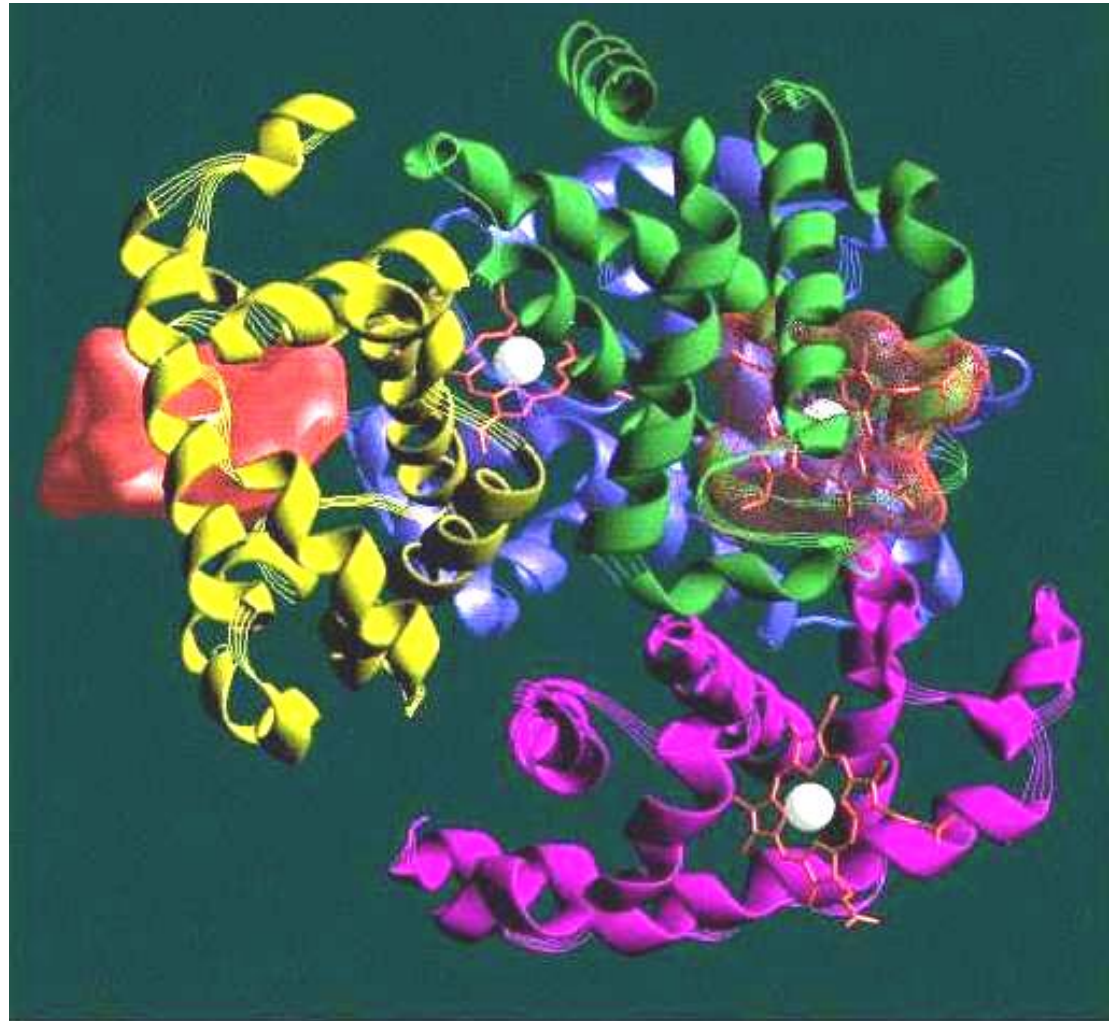
1. [All alpha proteins](#) [46456] (179)   
2. [All beta proteins](#) [48724] (126)   
3. [Alpha and beta proteins \(a/b\)](#) [51349] (121)   
Mainly parallel beta sheets (beta-alpha-beta units)
4. [Alpha and beta proteins \(a+b\)](#) [53931] (234)   
Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. [Multi-domain proteins \(alpha and beta\)](#) [56572] (38)   
Folds consisting of two or more domains belonging to different classes
6. [Membrane and cell surface proteins and peptides](#) [56835] (36)   
Does not include proteins in the immune system
7. [Small proteins](#) [56992] (66)   
Usually dominated by metal ligand, heme, and/or disulfide bridges
8. [Coiled coil proteins](#) [57942] (6)   
Not a true class
9. [Low resolution protein structures](#) [58117] (18)  
Not a true class
10. [Peptides](#) [58231] (105)   
Peptides and fragments. Not a true class
11. [Designed proteins](#) [58788] (39)   
Experimental structures of proteins with essentially non-natural sequences. Not a true class

Protein Data Base (PDB)

Classification **SCOP**

Premier niveau: **Classes** 1. All alpha proteins

essentiellement
hélices alpha

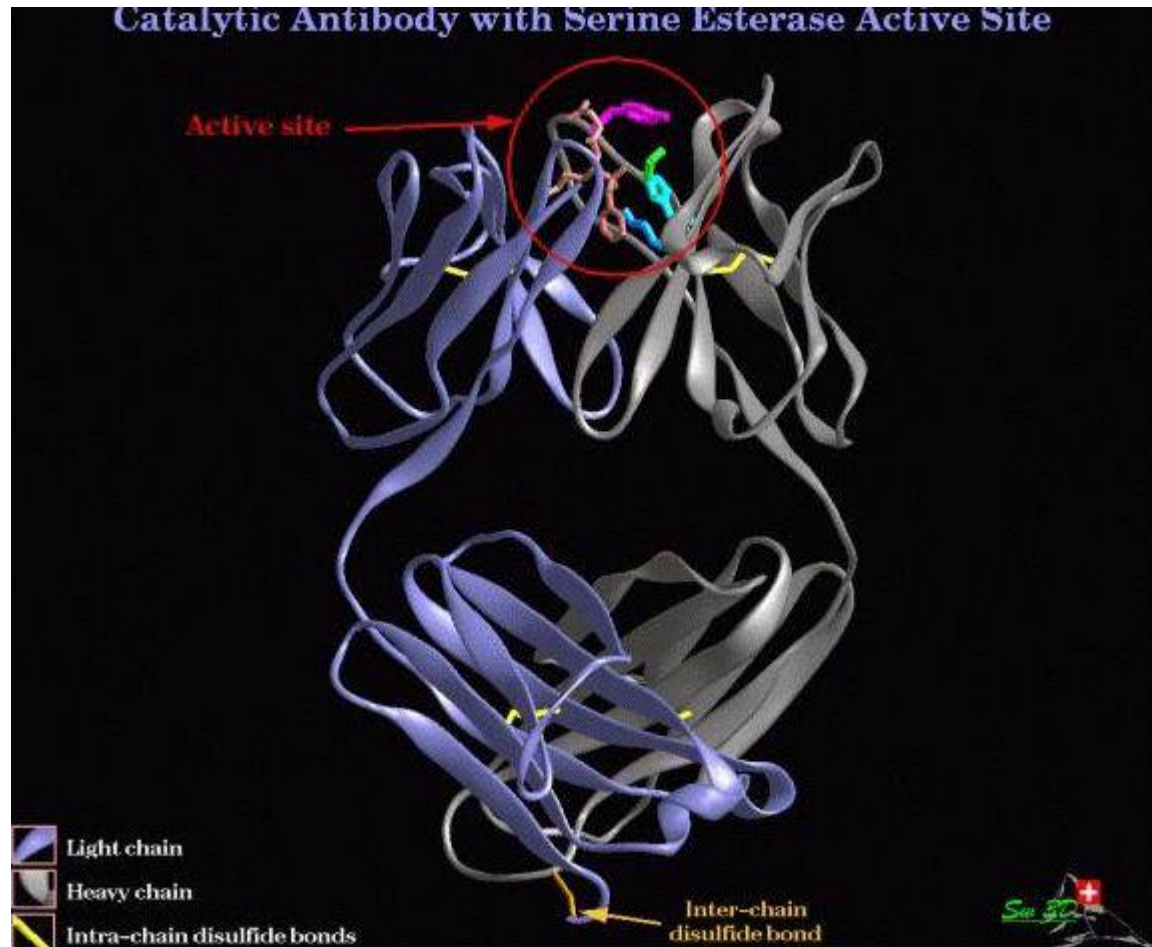


Protein Data Base (PDB)

Classification **SCOP**

Premier niveau: **Classes** 2. All beta proteins

essentiellement
feuilletés beta



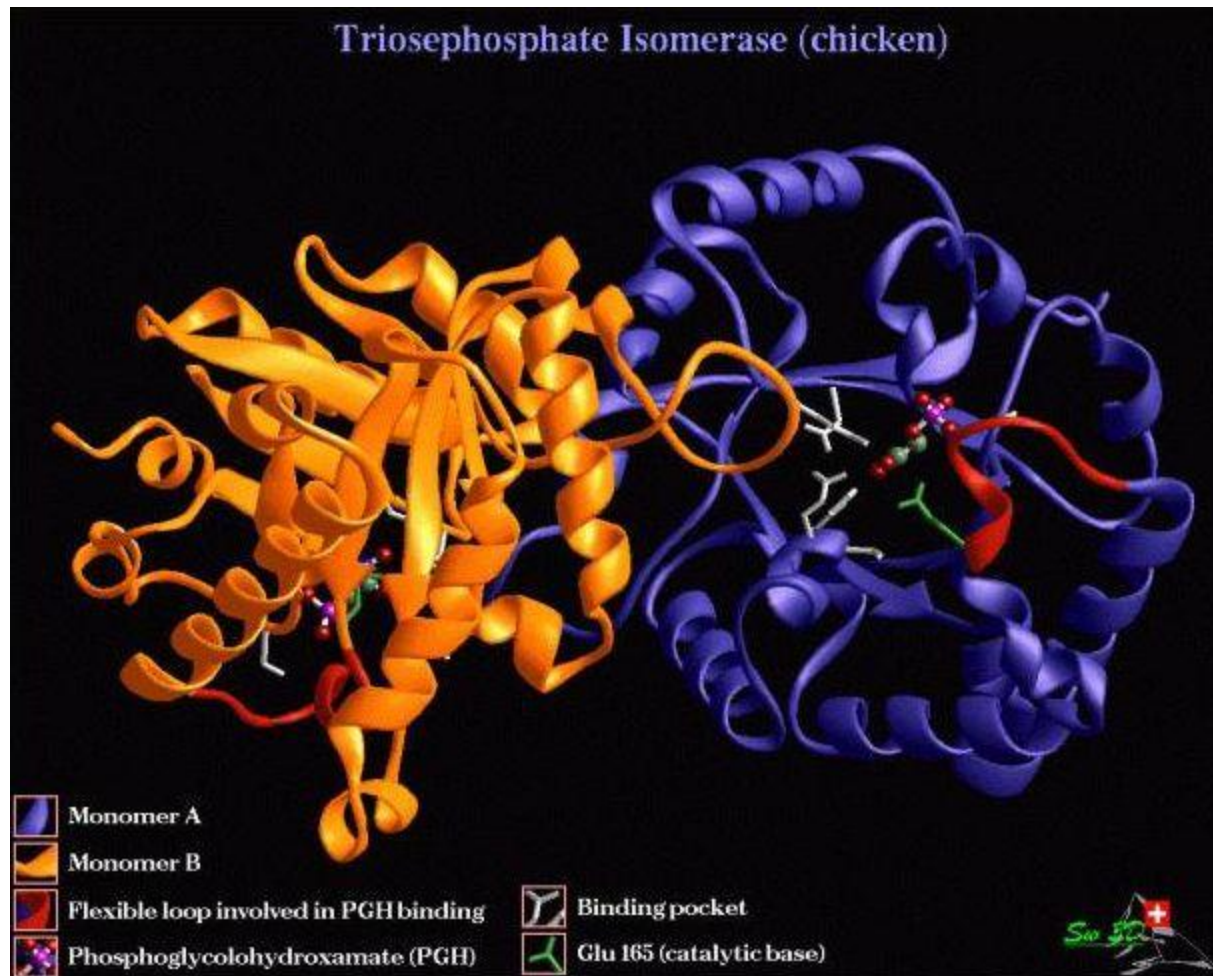
Protein Data Base (PDB)

Classification **SCOP**

Premier niveau: **Classes**

3. Alpha and beta proteins (a/b)

mainly parallel beta sheets (beta-alpha-beta units)

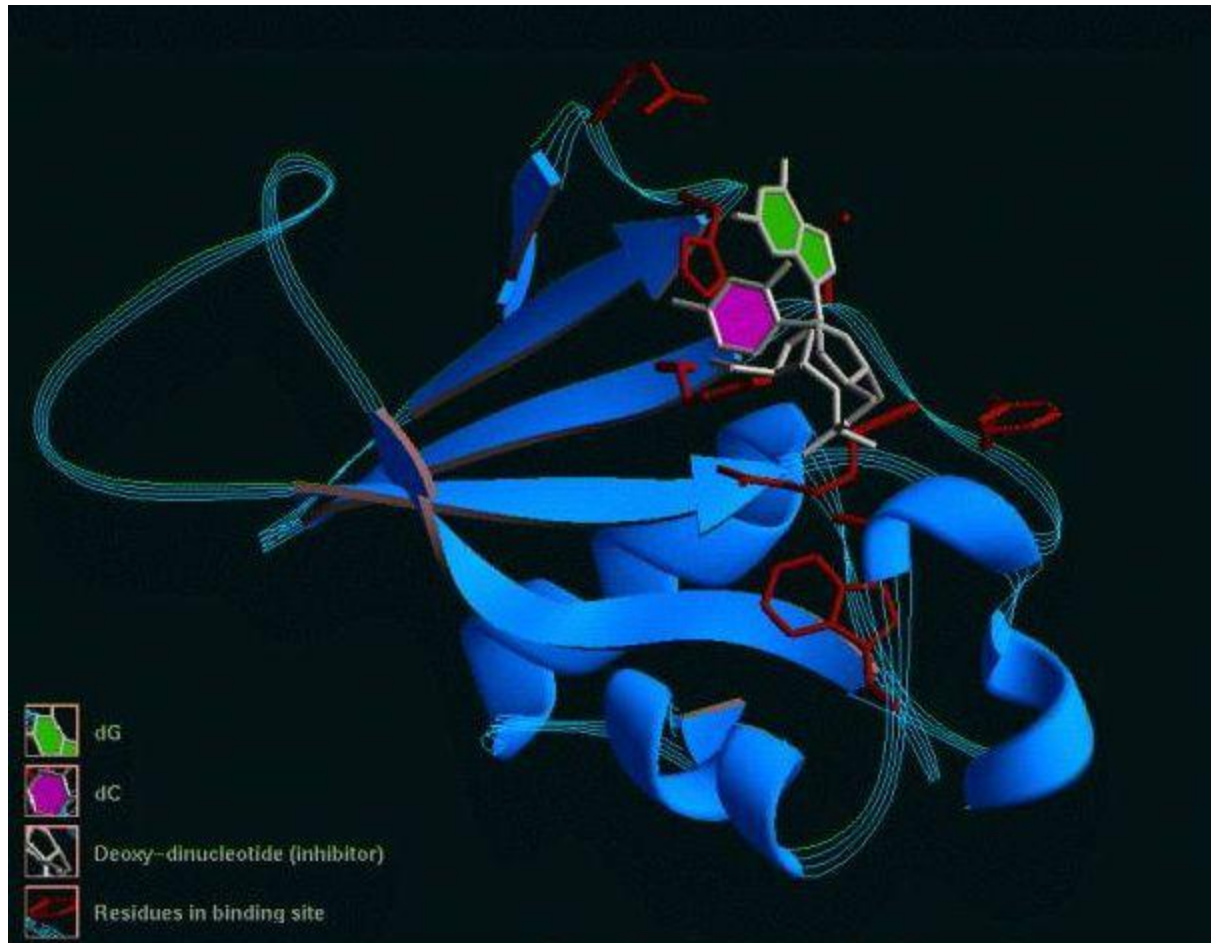


Protein Data Base (PDB)

Classification **SCOP**

Premier niveau: **Classes** 4. Alpha and beta proteins (a+b)

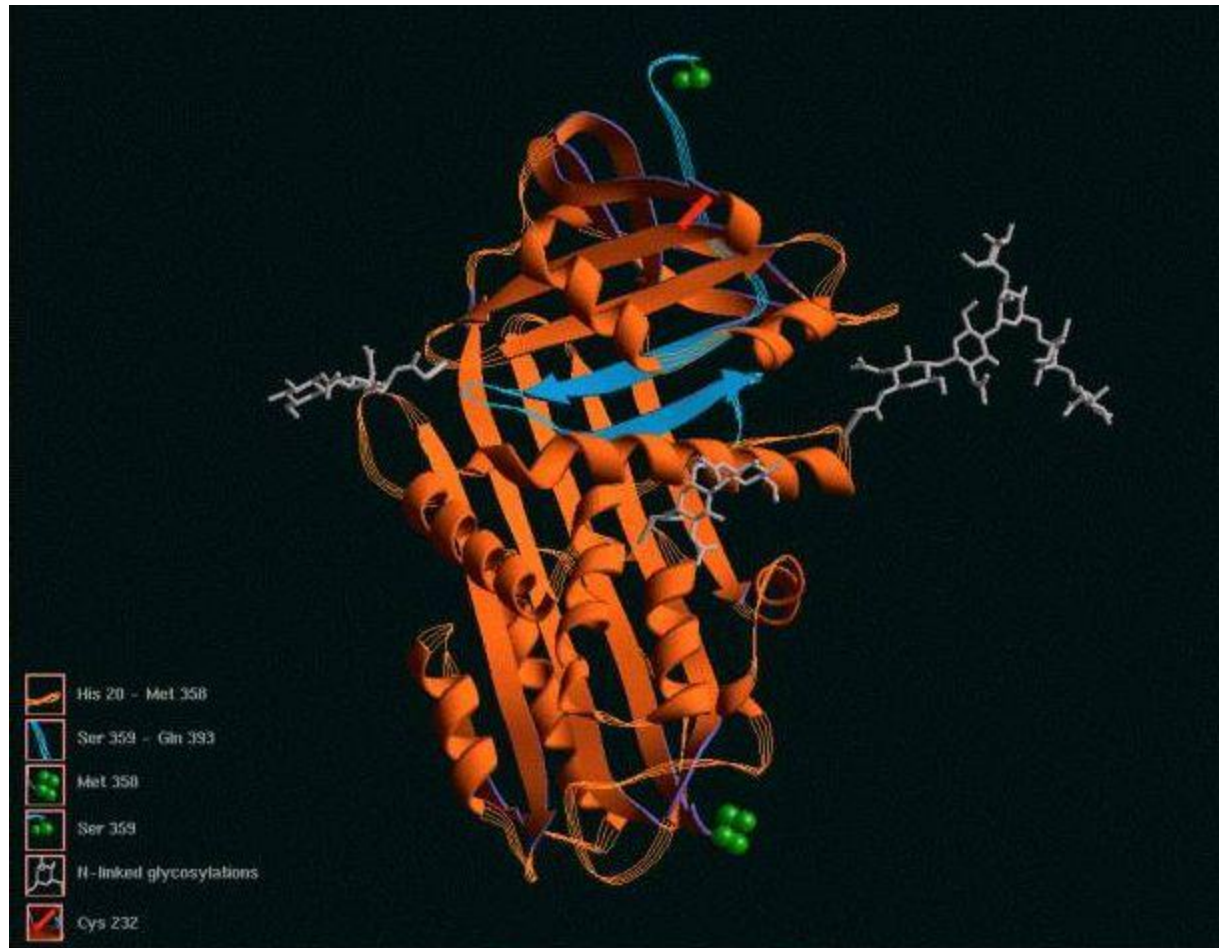
mainly antiparallel beta sheets (segregated alpha and beta regions)



Protein Data Base (PDB)

Classification **SCOP**

Premier niveau: **Classes** 5 **Multi-domain proteins** (alpha and beta)
folds consisting of two or more domains belonging to different classes



2 modes de pensées

Bio-informatique (plutôt le 1D)

traitement de la séquence (exemple: outils d'alignement)

Modélisation moléculaire (plutôt le 2D, 3D)

éléments de structures secondaires, traitement de la disposition dans l'espace (exemple: outils de superposition)



la complexité d'une molécule ne peut pas se résumer à une séquence
(notion de modèles...)

Cette **complexité** nécessite d'autres apports :

- sur la matière (**chimie, biochimie**)
- sur le comportement de la matière (**physique**)
- d'autres branches des **mathématiques**

structure

énergie

mvt.

Bioinformatique et modélisation moléculaire

Cours de Modélisation moléculaire en 2 blocs

Bloc: **modélisation moléculaire** (incontournable)

graphisme moléculaire

mécanique moléculaire

dynamique moléculaire

exemples et applications

Nécessite des connaissances préalables... **donc**

Bloc **pré- et post-modélisation** (modulable)

la chimie (stéréochimie)

la biochimie (groupements fonctionnels, structures biologiques)

la physique (spectroscopies, calculs énergétiques)

les mathématiques (exemple de robotique moléculaire)

I. Introduction

Concept de modèle

hypothèses (plus ou moins simplistes) => explications et prédictions
modélisation moléculaire à base de chimie, physique et mathématique

Attention: **modèle de référence quantique** car il part de la composition des molécules (atomes, électrons \Rightarrow disposition moléculaire...)

II.1. Les éléments constitutifs des molécules organiques

- C, H, O, N
- des non-métaux : Cl, Br, I, S, P, As, ...
- des métaux : Na, Li, Mg, Zn, Fe, Co, Cu, Cd, Pb, Sn, ..

{C, H et N} : 31% du corps humain

et 1,78% des éléments présents dans l'écorce terrestre

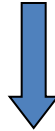
Et beaucoup plus avec l'eau!

importance en masse de C, H, **O**, N dans la biologie
et **structuration basée sur le carbone**

Modélisation bien développée pour C, H, O, N
et ... des lacunes pour la paramétrisation du reste...

II.2. Structure et propriétés des molécules organiques

La position **médiane** du carbone dans la classification périodique et dans l'échelle des **électronégativités** :



chimie organique \approx chimie composés **covalents** avec liaisons **peu ou non polarisées**

II.3. Caractéristiques intrinsèques du carbone

configuration électronique du carbone : $1s^2 2s^2 p^2$

avec **4** électrons de valence sur sa couche externe

donc **jusqu'à 4 liaisons covalentes** pour compléter sa couche externe à 8 électrons (**règle de l'octet**).

II. 4. Formation des liaisons chimiques

L'explication de la formation des liaisons chimiques passe nécessairement par la mécanique quantique

II.4.1. Forces de Coulomb

Une des caractéristiques essentielles de la formation des liaisons.

Deux atomes entrent en **interaction** si le nouvel état est globalement plus favorable que la somme des états indépendants antérieurs.

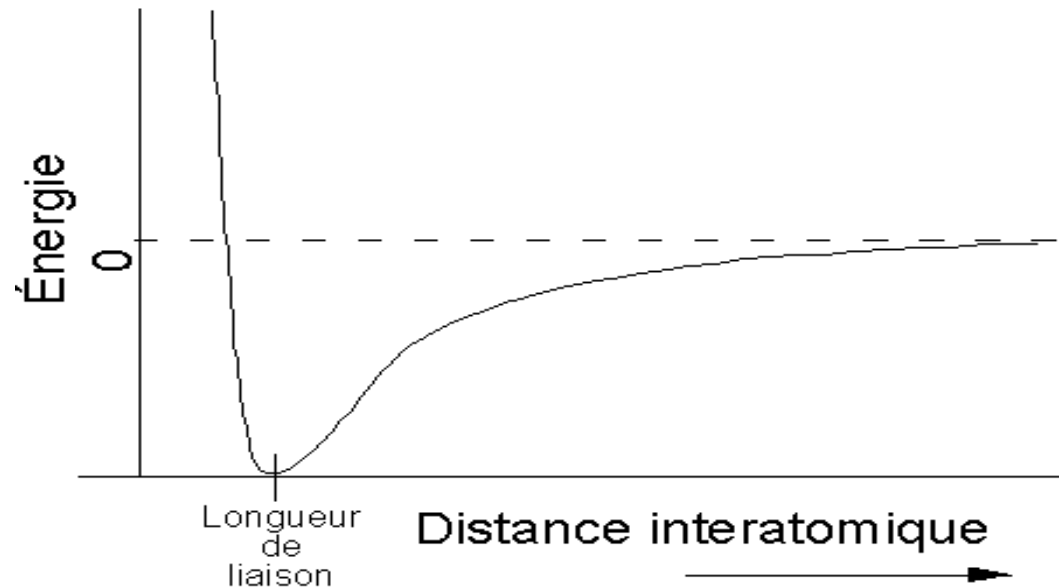
Les forces de Coulomb : des charges électriques éloignées de signes opposés s'attirent mutuellement :

$$F = \text{cte} * \frac{\text{charge}(+) \times \text{charge}(-)}{\text{distance}^2}$$

II.4.1. Forces de Coulomb (suite)

Les électrons d'un atome attirent le proton de l'autre et les électrons et protons portés par chacun des atomes se repoussent.

⇒ distance optimale (d'équilibre) = longueur de la **liaison chimique** formée



La position de ce minimum dépend de nombreux facteurs :

électronégativité, rayon atomique, orbitales impliquées et charges en présence...

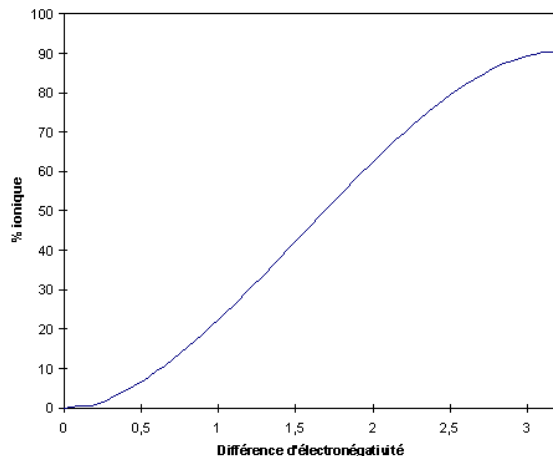
En modélisation, résultat brut: **longueur de la liaison** (1er paramètre interne)

II.4.2. Liaisons covalentes et ioniques

La **liaison covalente** résulte du **partage** des électrons entre les deux atomes

La **liaison ionique** résulte du **transfert** d'un électron d'un atome vers l'autre pour créer 2 atomes chargés (cation + et anion -)

Le pourcentage de liaison ionique dans n'importe quelle liaison dépend de la **différence d'électronégativité** entre les deux atomes.



$$\% \text{ ionique} = -4.935 \Delta^3 + 22.996 \Delta^2 + 5.748 \Delta - 1.337$$

la charge est un facteur correctif à la vision géométrique de la liaison covalente

Chimie et structures

La **mécanique quantique** donne l'explication de la disposition spatiale des atomes et des électrons qui constituent une molécule.

Orbitales atomiques

mouvement des électrons décrit par une **fonction d'onde**.

∃ plusieurs solutions toutes d'énergies **quantifiées**

⇒ **mouvements** des électrons autour du noyau (**orbitales**) d'énergies et de symétries spécifiques : **1s, 2s, 2px, 2py, 2pz....**

Orbitales moléculaires et hybrides

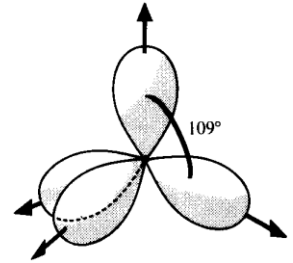
La MQ formalise le recouvrement des orbitales atomiques pour former des **orbitales moléculaires**.

- * autant d'orbitales moléculaires que d'orbitales atomiques de départ
- * plusieurs possibilités de combinaisons avec réorganisation (**hybridation**) des symétries des orbitales atomiques de départ avant la formation de la liaison.

Chimie et structures

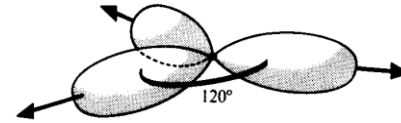
Première solution d'orbitales hybrides:

avec 3 orbitales 2p (+ 1 orbitale 2s) \Rightarrow 4 orbitales hybrides sp^3



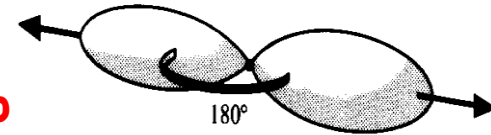
Deuxième solution d'orbitales hybrides:

avec 2 orbitales 2p (+ 1 orbitale 2s) \Rightarrow 3 orbitales hybrides sp^2



Troisième solution d'orbitales hybrides:

avec 1 orbitale 2p (+ 1 orbitale 2s) \Rightarrow 2 orbitales hybrides sp



Ces trois orbitales hybrides sont les fondements de la disposition dans l'espace des molécules (**stéréochimie**) basée sur la répulsion des électrons dans un modèle de **combinaisons hybrides des orbitales atomiques** pour la formation des molécules.

En modélisation, l'angle de valence est le 2eme paramètre interne

III. Stéréochimie

stéréochimie : représentation 3D des molécules, leur **arrangement spatial**.
principes de stéréochimie (par l'exemple de la **projection de Newman** mais bien d'autres types de représentations **pseudo-3D**).

Stéréochimie et informatique

deux grands domaines de développements (**visuels**):

a/ **support plan** (2D) + **illusion de la troisième dimension**
(techniques d'ombrage par exemple)

b/ directement **vision stéréoscopique** (décalage de perception de nos 2 yeux :
(lunettes stéréo devant un écran équipé, simulations de déplacement physique dans une molécule avec des lunettes d'environnement)

III.2. Passage représentations 2D vers 3D

Techniques actuelles d'infographie

algorithmes de représentation moléculaire: **artifices efficaces**
pour donner la disposition dans l'espace des atomes et des liaisons.

- * **masquage** (ce qui est devant masque ce qui est derrière),
- * **éclairage** (ce qui est devant est plus lumineux que ce qui est derrière)
- * **ombrage**, combinaison (masquage et éclairage) :

un angle d'éclairage bien choisi permet de représenter l'ombre portée de ce qui est devant sur ce qui est derrière.

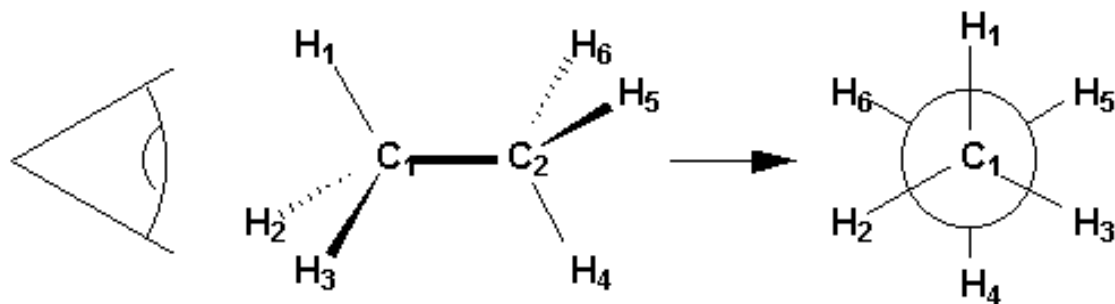
- * **transparence** (pour voir au travers dans un environnement compliqué)
- * **'fenêtre'** de vision (**clipping**)

III.2.3. Projection de Newman et angles dièdres

Représenter **locale** (disposition des atomes **autour une liaison centrale**).

Elle décrit toutes les liaisons attachées aux atomes formant la liaison centrale et plus particulièrement elle indique l'orientation de ces liaisons telles qu'on les verrait si on regardait la molécule dans **l'axe de la liaison qui unit les deux atomes centraux**.

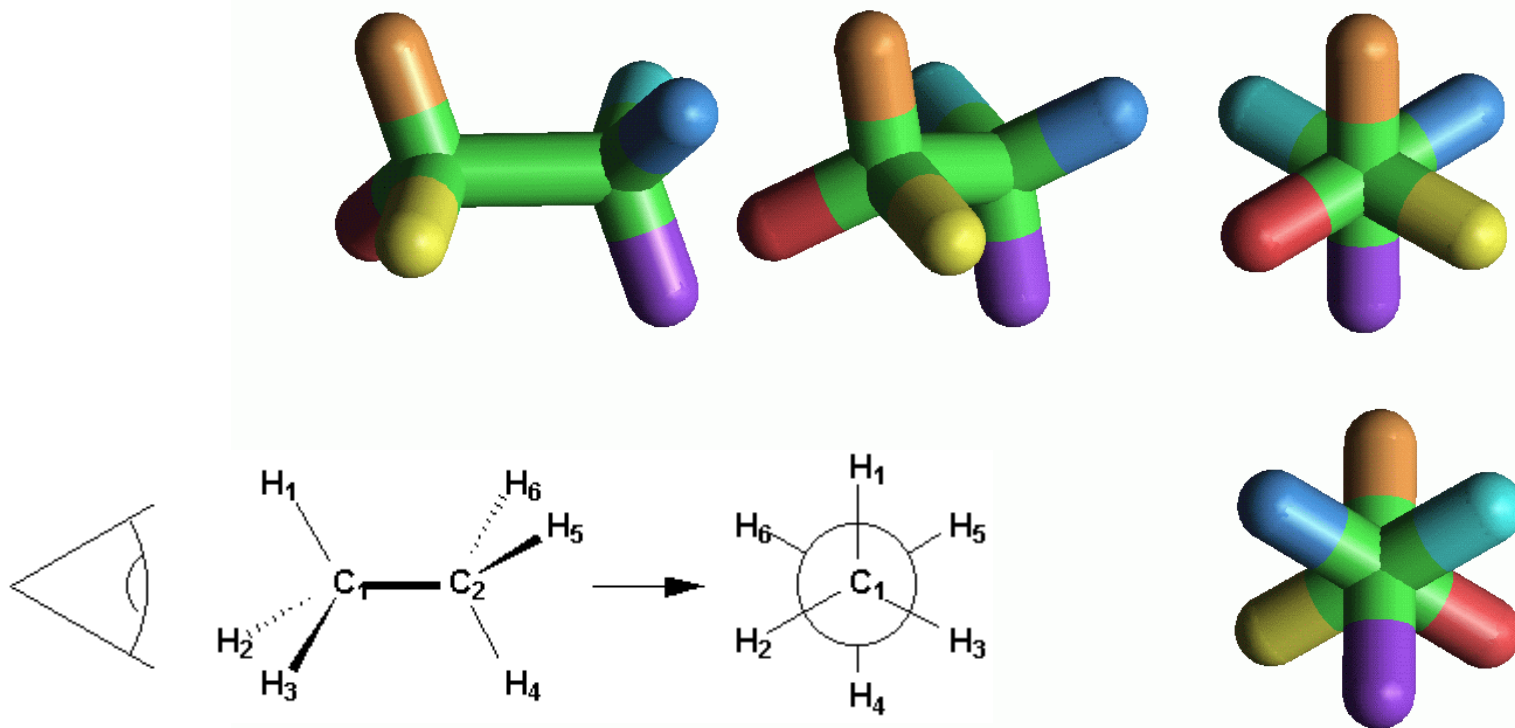
Cas de l'éthane :



Le cercle permet de visualiser les liaisons "**devant**" (C1-H1, C1-H2, C1-H3) et de les différencier de celles qui sont "**derrière**" (C2-H4, C2-H5, C1-H6) par convention C2 (masqué) est situé dans l'exact prolongement de C1.

III.2.3. Projection de Newman et angles dièdres (suite)

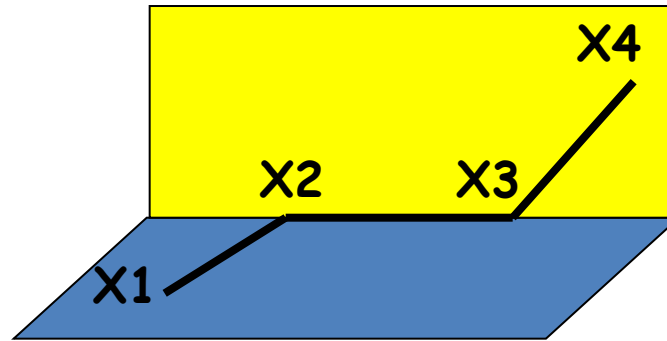
rotation 'virtuelle' de l'image autour de la liaison centrale.



III.2.3. Projection de Newman et angles dièdres

Quantification de la disposition des atomes en 3D : **angle dièdre**.

* soit 4 atomes liés entre eux dans un ordre défini : $X1 - X2 - X3 - X4$.



* **Définition mathématique**: l'angle de **projection de 2 plans** (définis par les atomes $X1-X2-X3$ et $X2-X3-X4$ respectivement).

* Sur la projection de Newman, on voit que l'angle dièdre ne peut être défini que dans une **plage de variations de 360** (modulo 360).

* le **module** de l'angle dièdre est égal à la rotation nécessaire pour amener X1 sur X4 selon l'axe défini par les points centraux X2 et X3.

(**ramener virtuellement l'atome de devant sur celui de derrière**)

* le **signe** est donné par la règle du **tire-bouchon** (rotation dextrogyre) (ou celle du **sens des aiguilles d'une montre**)

* la convention la plus utilisée : **angle dièdre entre -180 et +180**

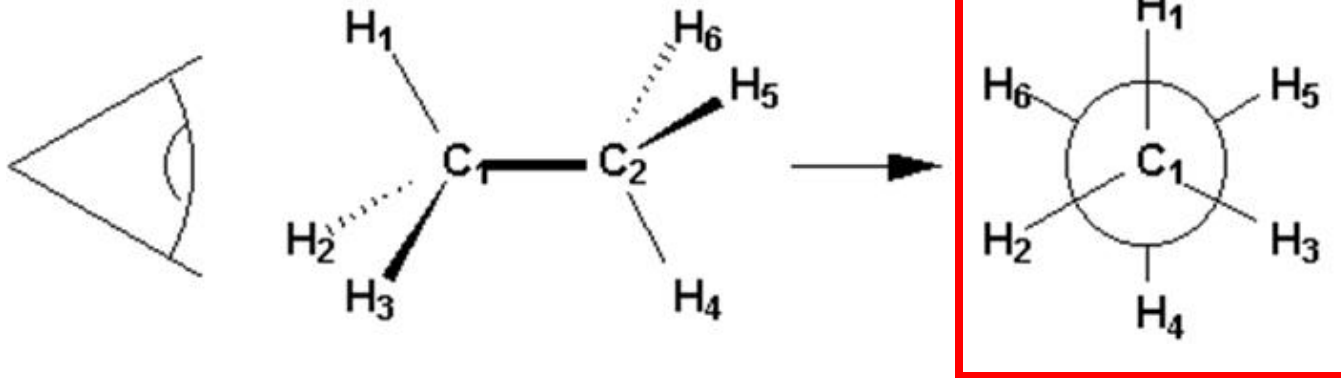
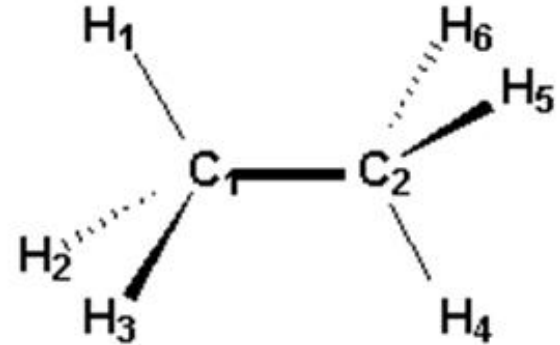
III.2.3. Projection de Newman et angles dièdres en pratique....toujours sur l'éthane

Valeurs des dièdres suivants: ?

[H1-C1-C2-H5]

[H3-C1-C2-H5]

Est-ce plus simple comme ça?



Réponse:

respectivement +60 et -60 .

Dans ces 2 cas, les atomes extrêmes sont dits "**décalés**".

Si l'angle dièdre vaut **0** , les atomes extrêmes sont dits "**éclipsés**"

Si l'angle dièdre vaut **180** , on parlera d'une position **trans**.

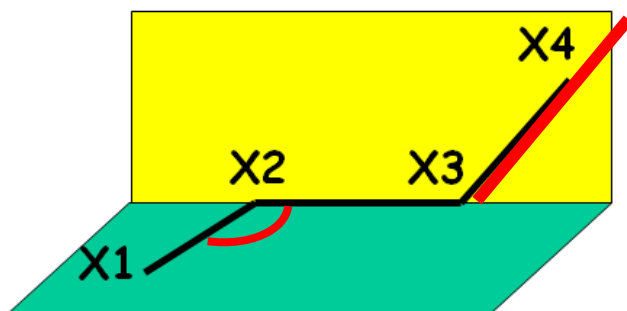
III.2.3. Projection de Newman et angles dièdres (suite)

Propriétés remarquables :

* la valeur de l'angle dièdre est **indépendante de l'angle de vision** dans la projection de Newman. Par exemple, $[H1-C1-C2-H5]$ vaut toujours $+60^\circ$, que l'on regarde de $C1$ vers $C2$ ou l'inverse.

* l'angle dièdre est une **valeur de projection**

- **indépendante** des angles de valence $X1-X2-X3$ ou $X2-X3-X4$.
- **indépendante** des distances $X1-X2$, $X2-X3$, $X3-X4$



* les angles dièdres modulo 360 identiques ($+180^\circ = -180^\circ$).

En pratique, **le dièdre est le dernier des trois paramètres internes** permettant de définir dans l'espace la position des atomes d'un système moléculaire les uns par rapport aux autres.

Rappel 2 autres paramètres: **distances de liaison et angles de valence.**

III.3.5. Configuration absolue, règles de Cahn-Ingold-Prelog

la configuration relative de 2 énantiomères est l'opposée l'une de l'autre.

Mais qu'en est-il de leurs configurations absolues ?

La configuration absolue est sans rapport avec le signe de la rotation optique

A/ Elle ne peut être établie qu'avec des spectroscopies accédant à la disposition relative des positions atomiques de la molécule (ex diffraction des rayons X).

B/ On peut la déduire par corrélation chimique avec une structure de configuration connue.

La nomenclature des molécules chirales utilise le système de Cahn-Ingold-Prelog (C.I.P) .

Ce système permet de désigner sans ambiguïtés tous les stéréo-isomères puisque chaque carbone asymétrique ou chaque élément de chiralité aura sa propre distinction (*R* ou *S*).

III.3.5. Configuration absolue, règles de Cahn-Ingold-Prelog (suite)

Règles de classement selon un ordre de priorité défini par convention:

Règle 1: La priorité est établie d'après les nombres atomiques des atomes qui sont attachés. En cas de substitution isotopique, le noyau dont la masse atomique est la plus élevée est prioritaire vis-à-vis de celui dont la masse atomique est inférieure.

Règle 2: Si 2 substituants ou plus ont le même grade lorsque l'on considère les atomes attachés au stéréocentre, continuer le long des chaînes substitutives jusqu'à ce que l'on atteigne un atome de nature différente qui permette une distinction de priorité.

Règle 3: Les liaisons multiples sont considérées comme si elles étaient saturées

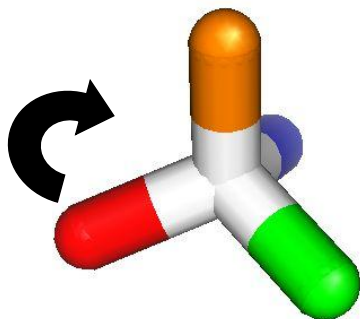
III.3.5. Configuration absolue, règles de Cahn-Ingold-Prelog (exemple)

Soit un ordre de priorité des substituants ($a > b > c > d$) autour d'un carbone asymétrique:

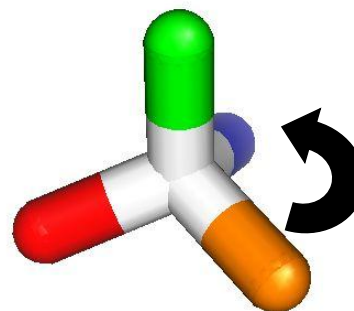


On positionne la molécule pour que le substituant de plus basse priorité soit le plus éloigné de l'observateur [projection de Newman du carbone central asymétrique selon l'axe $C \rightarrow d$].

Si l'ordre de priorité $[a, b, c]$ suit la rotation des aiguilles d'une montre, la configuration est **R** (*rectus*, latin, droit). Si l'ordre de priorité suit la rotation inverse des aiguilles d'une montre, la configuration est **S** (*sinister*, latin, gauche). Le symbole est ajouté sous forme de préfixe au nom de la molécule.



R et **S**



Il faut pratiquer les orientations des énantiomères à l'aide de modèles moléculaires (type Dreiding) ou de manipulations devant un écran graphique disposant de facilités interactives.

III.3.5. Configuration absolue, règles de Cahn-Ingold-Prelog (suite)

La chiralité est extrêmement importante puisqu'elle conditionne toute la **chimie des systèmes vivants**.

Par exemple:

- tous les acides aminés naturels intégrés dans les systèmes vivants possèdent la configuration **S** dans la classification **CIP**.
- les composés suivants ne possèdent qu'un seul stéréoisomère à l'état naturel: camphre, menthol, glucose, cholestérol (sur les 256 stéréoisomères possibles)
- la forme lévogyre de l'adrénaline est 12 fois plus active que sa forme dextrogyre.
- la forme dextrogyre de l'asparagine a un goût sucré, la forme lévogyre a un goût amer.

III.3.5. Configuration absolue, règles de Cahn-Ingold-Prelog (suite)

Importance pour l'industrie pharmaceutique:

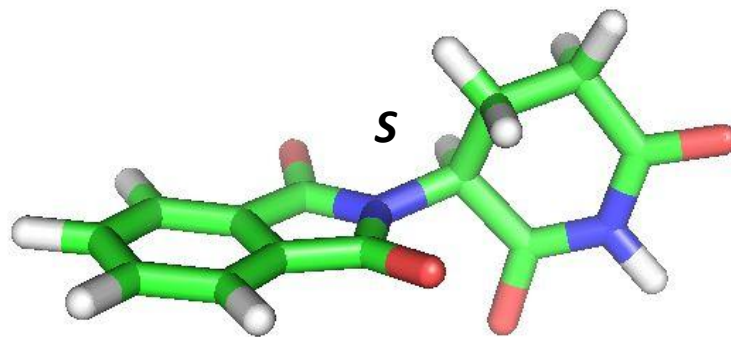
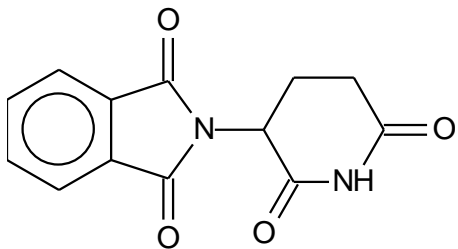
exemple dramatique de la **Thalidomide**

(médicament donné aux femmes enceintes dans les années 60).

R, principe actif anti-nauséeux

S, tératogène (*malformation du fœtus*)

Formule brute: $C_{13}H_{10}N_2O_4$



Conséquences : plus aucune délivrance de mise sur la marché de médicaments ne se fait sans études d'énantiométrie très poussées.

Commercialisation très stricte de racémiques (qu'il y aient des différences d'activités ou non)