# Prédiction de la structure 3D


## Dr. Christophe Geourjon

*Pôle de BioInformatique Lyonnais*

*PBIL - Site de Lyon-Gerland*

**IBCP - CNRS UMR 5086**

***Bioinformatique et RMN structurales***

7, passage du Vercors

69367 Lyon cedex 07

Tél: +33 (0)4 -72-72-26-47

E-mail : c.geourjon@ibcp.fr

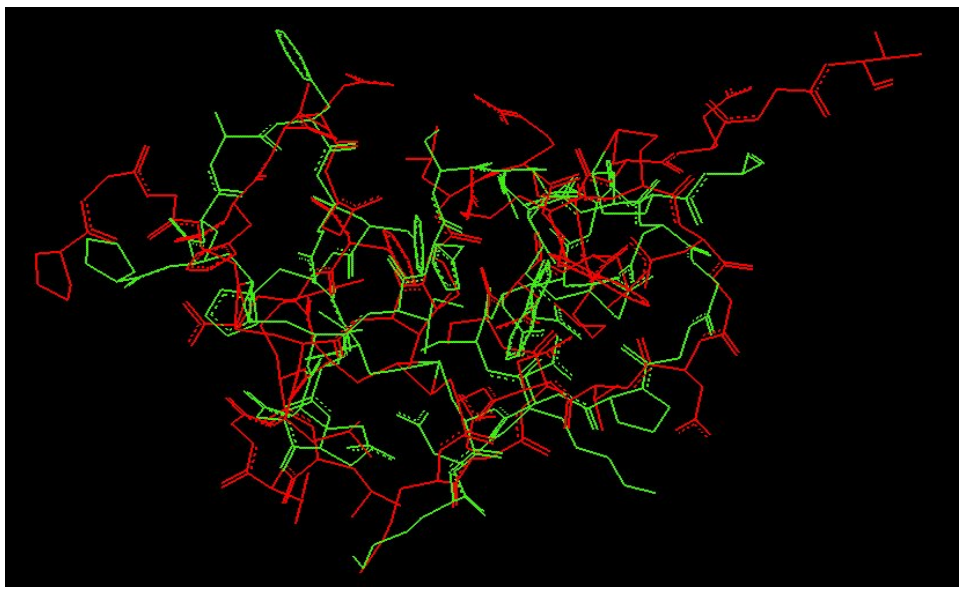
**Site Web : pbil.ibcp.fr**

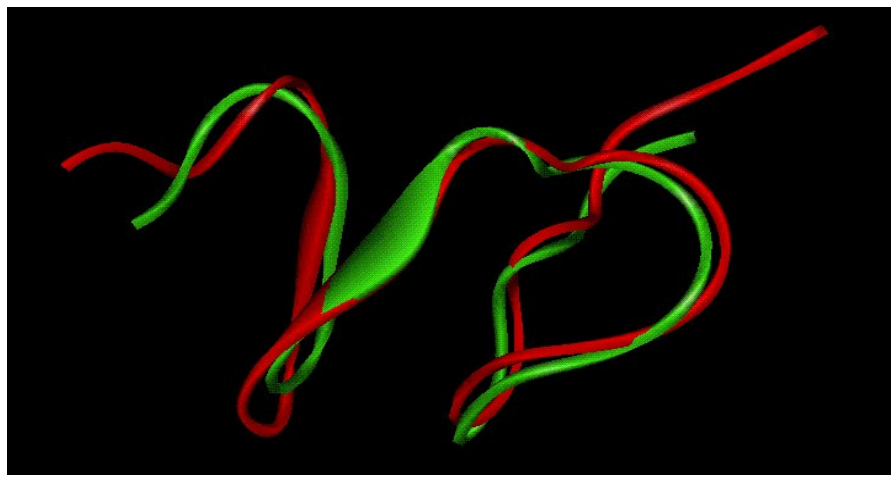

# Conservation de la structure 1D, 2D et 3D

1AJJ : LDL receptor
1CR8 : Low Density Lipoprotein Receptor Related Protein

```
                          10        20        30        40
                  .........|.........|.........|.........|..
1.pdb1ajj.ent     P--CSAFEFHC-LSGECIHSSWRCDGGPDCKDKSDEENCA--  37
2.pdb1cr8.ent     PGGCHTDEFQCRLDGLCIPLRWRCDGDTDCMDSSDEKSCEGV  42
Primary cons.     PGGC222EF2CRL2G2CI222WRCDG22DC2D2SDE22C2GV
Homology          *   * : **:* *.* **    ******..** *.***:.*
```

**50 % d'identité de séquence**





**Ecart quadratique moyen (CA des résidus conservés) : 1,6A**



c.geourjon@ibcp.fr

# Conservation de la structure 2D et 3D

**16 % d'identité de séquence**
**et**
**une topologie conservée**

En dessous de 30% d'identité, la structure est plus conservée que la séquence, ceci est vraie également des structures secondaires. Mais ceci n'est pas systématique …



c.geourjon@ibcp.fr

# *Prédiction de structure 3D*

- **Modélisation par homologie**
  - 2 protéines qui ont plus de 30% d'identité de séquences ont 80% de leurs Cα superposables avec un écart quadratique moyen de 1 Å (RMSD=1Å)

    ➔ Outils de modélisation moléculaire GENO3D

- **Modélisation par analogie**
  - 2 protéines qui ont la même fonction et une topologie « probablement » identique en dépit de l'absence de similarité importante (arguments expérimentaux ou de structures secondaires).

    ➔ Outils de modélisation moléculaire GENO3D

- **Threading**
  - Une séquence est testée sur une librairie de repliements pour déterminer sa compatibilité structure-séquence probable, méthode d'alignement séquence-structures tridimensionnelles.

- ***Ab initio***
  - Structure directement déduite de la séquence à partir de règles empiriques.

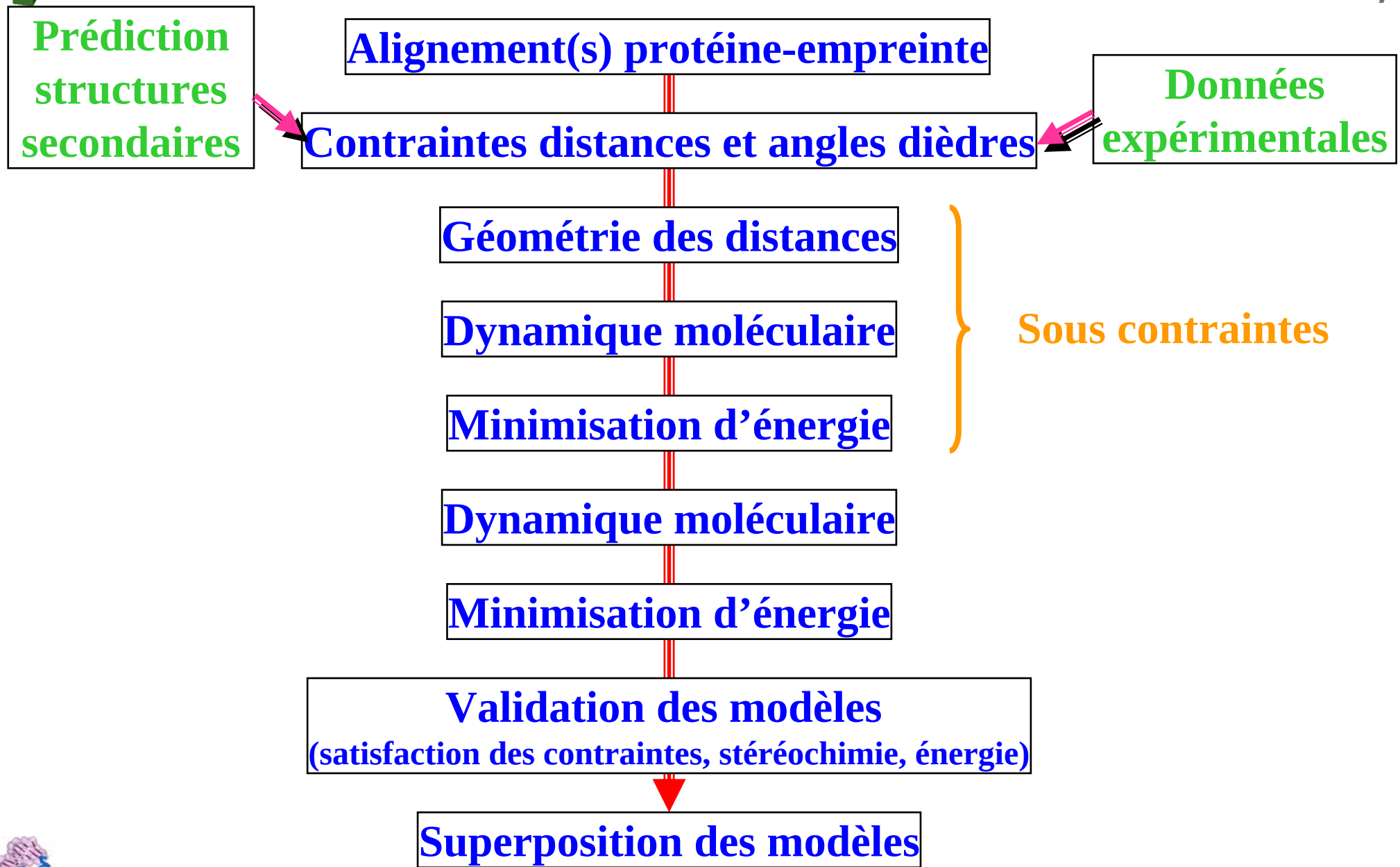

c.geourjon@ibcp.fr

# Geno3D : Serveur Web de modélisation moléculaire

- **Cahier des charges**
  - ✓ Générer des modèles 3D à partir d'une empreinte 3D y compris à bas taux d'identité de séquence.
  - ✓ Pouvoir inclure dans ce processus des informations expérimentales.
  - ✓ Pouvoir imposer à certaines régions de la protéine une structure secondaire donnée.
  - ✓ Pouvoir supporter en entrée des informations floues.
  - ✓ Générer un jeu de modèles (estimation de qualité séquentielle).
  - ✓ Notion de haute débit
  - ✓ Notion d'automatisation (critères qualitatifs)
- **Moyen**
  - ✓ Modélisation moléculaire globale sous contraintes (distances et angles dièdres) du même type que celle mise en œuvre dans le processus de modélisation moléculaire sous contraintes RMN. Dans ce cas les contraintes sont déduites à partir de la ou des structures empreintes et des alignements de séquences. Ces informations sont par la suite utilisées dans un protocole de géométrie des distances (reconstruction globale de la structure 3D)
    → Possibilité d'inclure des informations externes.
  - ✓ Moteur de mécanique moléculaire : CNS.
  - ✓ Génération d'un faisceau de structure 3D.


c.geourjon@ibcp.fr

# Geno3D : Différentes étapes

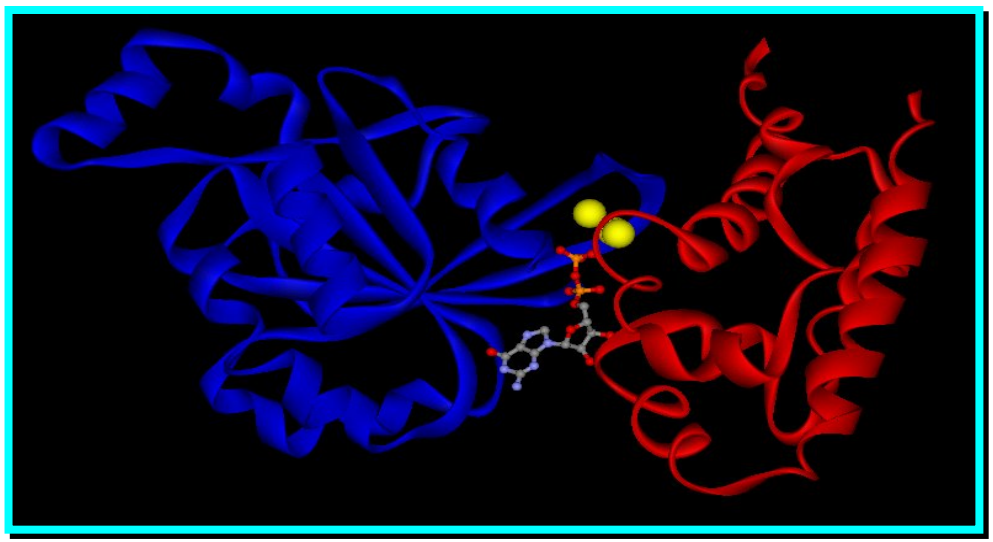





c.geourjon@ibcp.fr

# Méthode Geno3D – Version 2.0

- ✓ Système automatique de modélisation de la structure 3D des protéines par homologie et analogie.
- ✓ Disponible sur le Web : http://geno3d-pbil.ibcp.fr (111 784 soumissions)

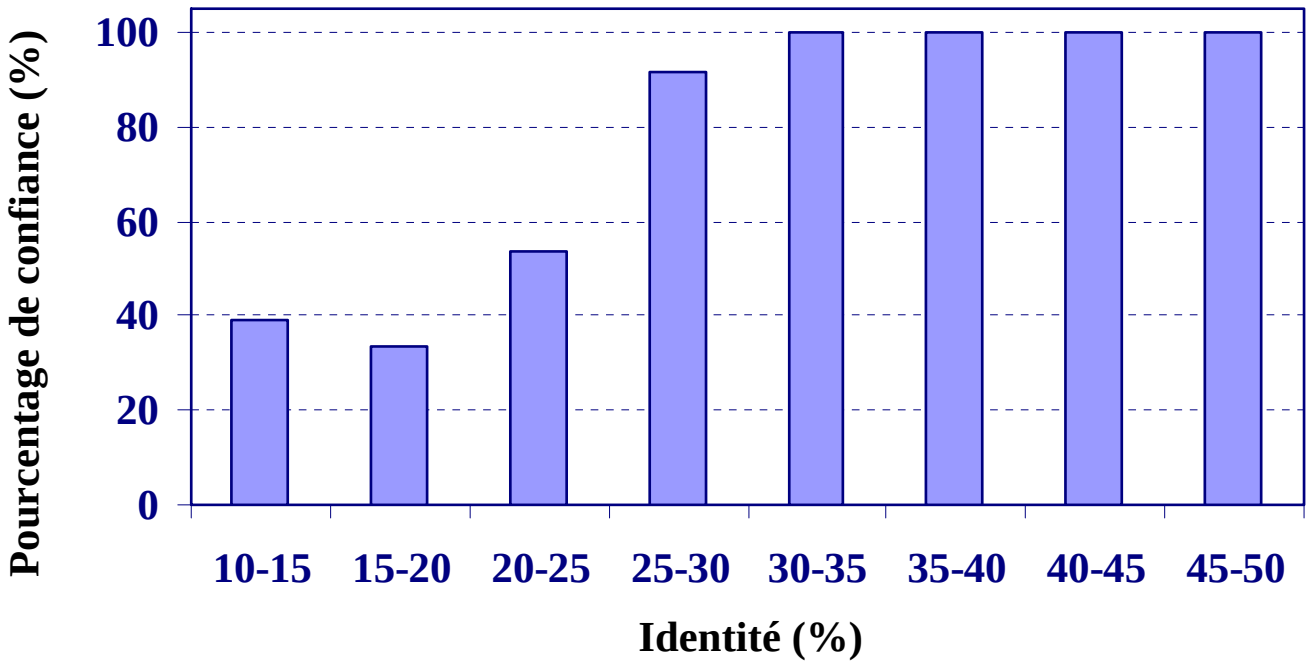
- ✓ Utilisé à grande échelle (modélisation de l'ensemble des protéines d'un protéome).

En moyenne 15% des requêtes aboutissent à la prédiction d'une structure 3D (seuil pour la construction du modèle fixé à 25% d'identité entre la protéine à modéliser et la ou les empreintes).


c.geourjon@ibcp.fr

# Recherche d'homologues distants

- Recherche de similarité (pdb_select_25 versus pdb_select_95)
- Sélection des fragments
  - 10%<identité<50%
  - gap <10%
  - longueur > 100AA
- Classification selon FSSP en « apparenté » ou « non apparenté »



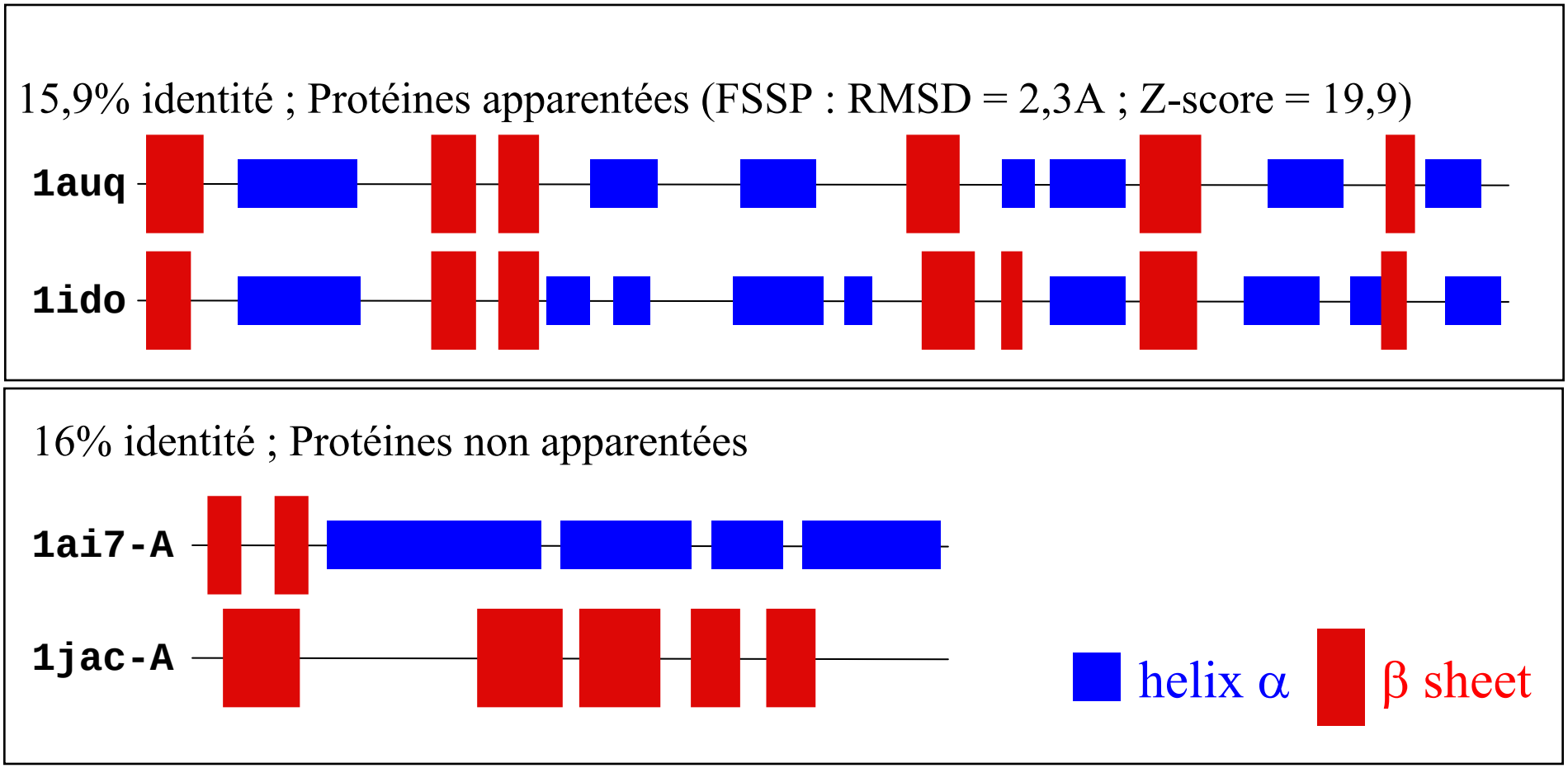

Ssearch (E=10)

**Nécessité d'utiliser une autre source d'informations …**



c.geourjon@ibcp.fr

# Utilisation des prédictions de structures secondaires





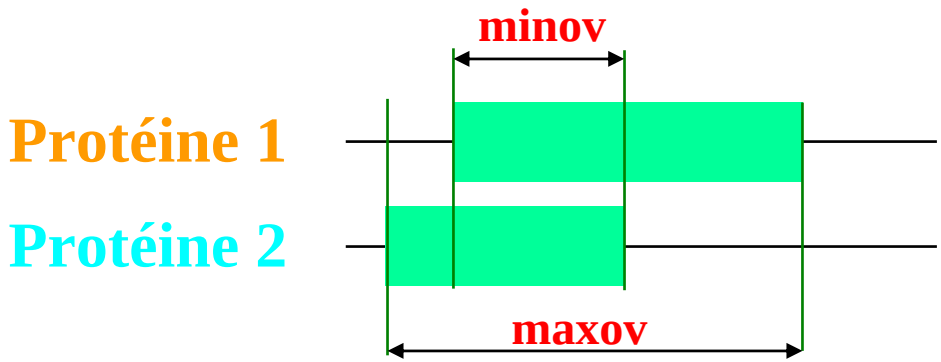
c.geourjon@ibcp.fr

# Mesure de l'accord entre les structures secondaires Sov

- Sov coefficient (Structural Overlap) (Rost *et al.*, 1994 ; Zemla *et al.*, 1999)

$$Sov = 100 \times \left[ \frac{1}{N} \sum_{i \in [H,E,C]} \sum_{S(i)} \frac{minov(s_q, s_t) + \delta(s_q, s_t)}{maxov(s_q, s_t)} \times len(s_q) \right]$$

- *minov* : longueur de la structure secondaire chevauchante entre la source $S_q$ et la cible $S_t$
- *maxov* : longueur maximale des structures secondaires chevauchantes entre la source $S_q$ et la cible $S_t$




- $\delta$ est défini par :

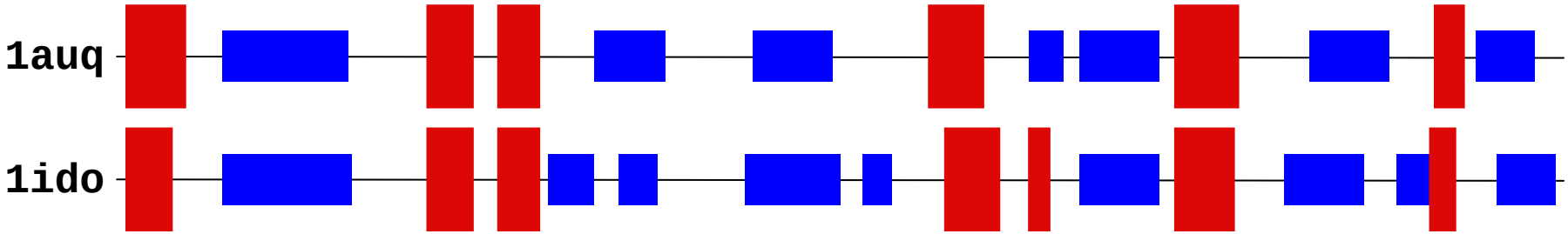
$$\delta(s_q, s_t) = \min \left\{ \begin{array}{l} (maxov(s_q, s_t) - minov(s_q, s_t)); minov(s_q, s_t); \\ int(len(s_q/2)); int(len(s_t/2) \end{array} \right\}$$


c.geourjon@ibcp.fr

# Utilisation des prédictions de structures secondaires

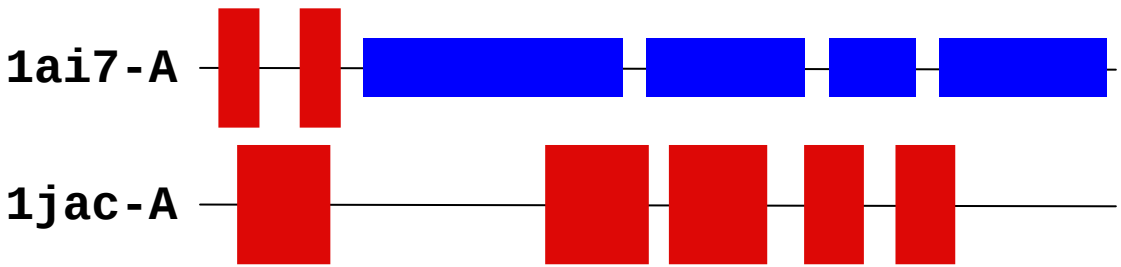
**Sov = 81**

15,9% identité ; Protéines apparentées (FSSP : RMSD = 2,3A ; Z-score = 19,9)




**Sov = 9**

16% identité ; Protéines non apparentées




helix α  β sheet

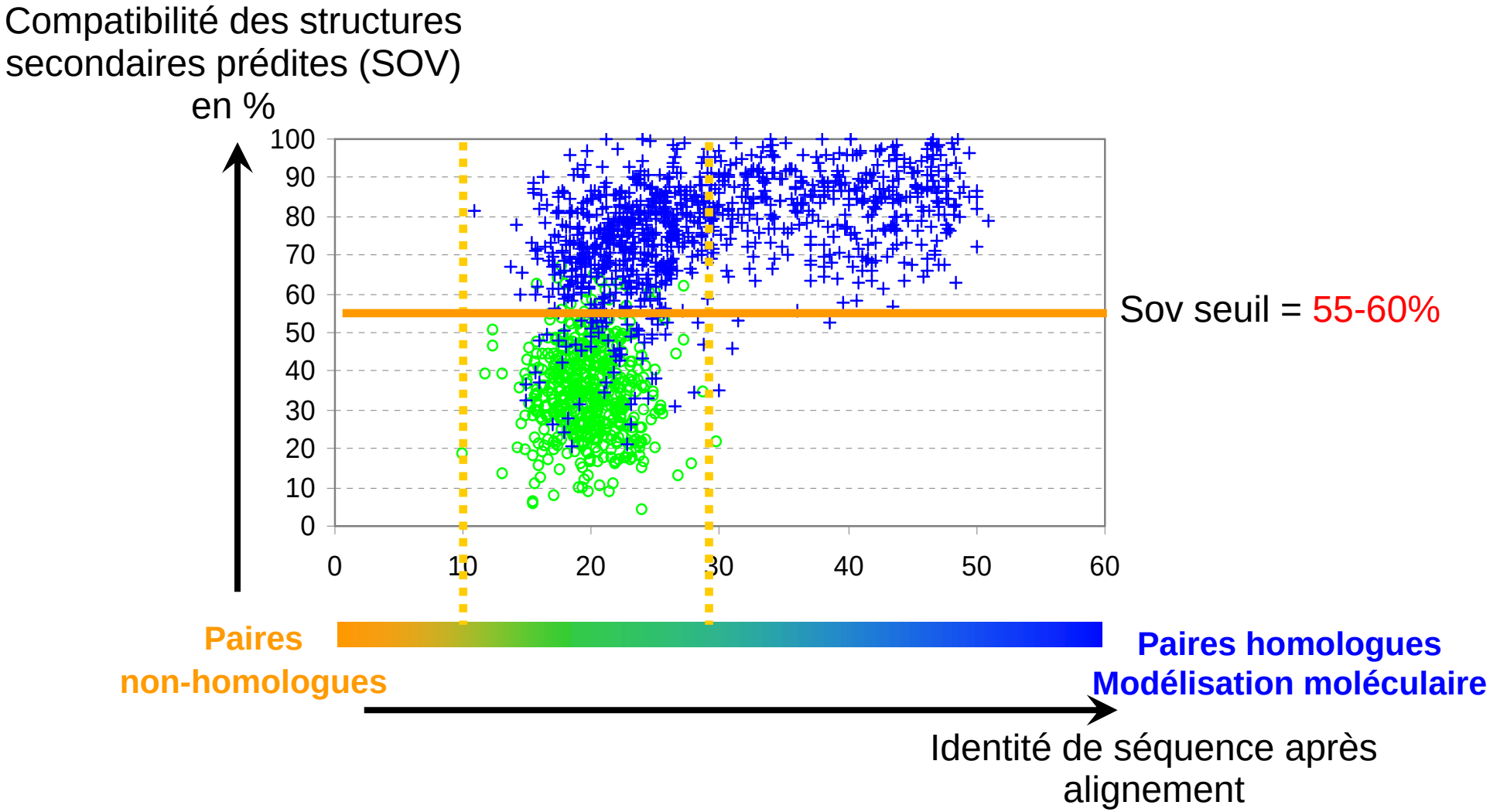
Les prédictions de structures secondaires permettent donc de valider des similarités de séquence à bas taux d'identité.

**Annotation fonctionnelle**



c.geourjon@ibcp.fr

# Détection d'empreintes pour la modélisation à faible taux d'identité




Identification of related proteins with weak sequence identity using secondary structure information
Geourjon, C. Combet, C, Blanchet, C & Deléage G
**Protein Science** (2001) 10, 788-797



c.geourjon@ibcp.fr

# Exemple : modélisation à faible taux d'identité

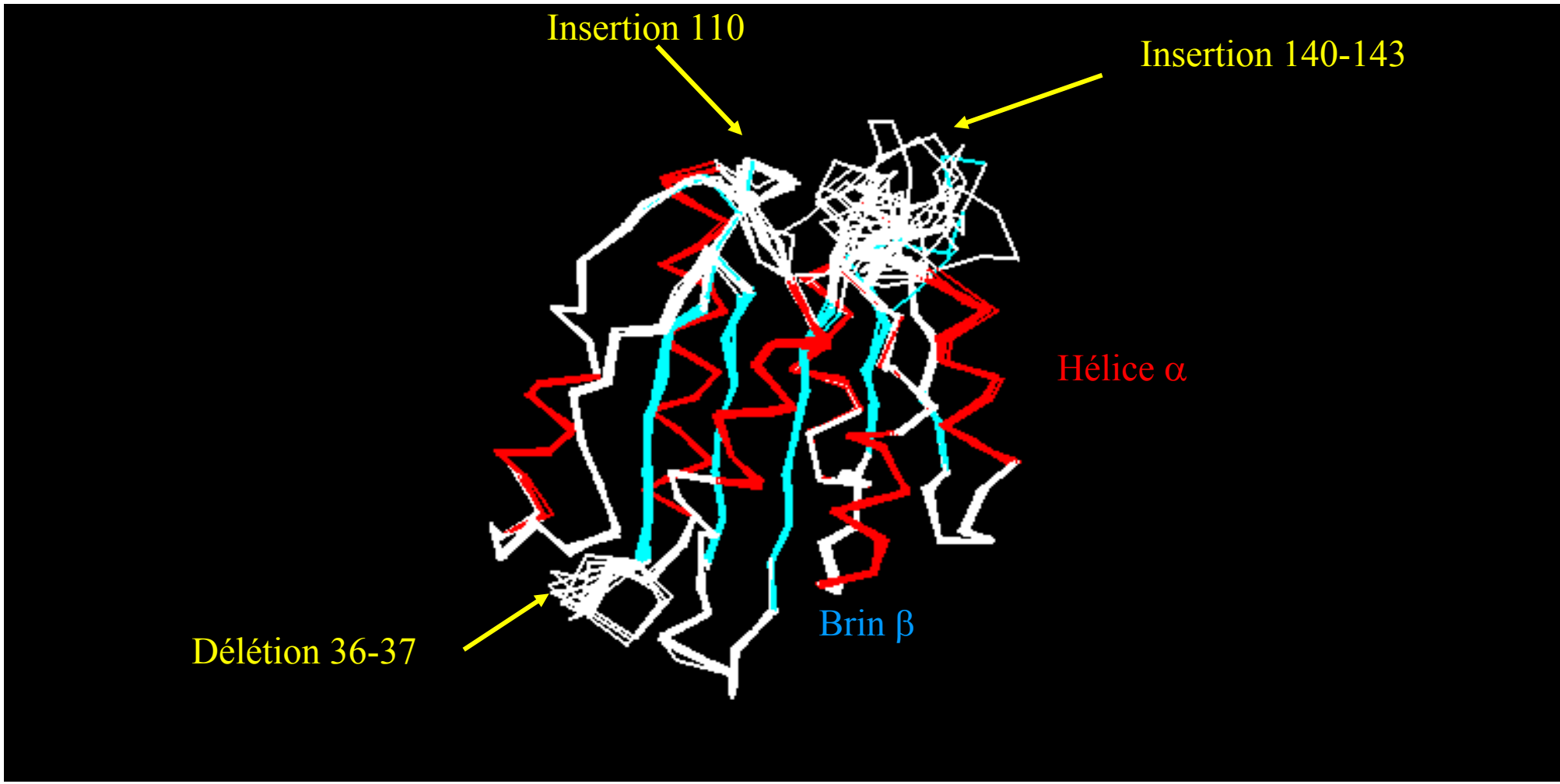
1auq : A1 domain of Von Willebrand factor
1ido : Integrin CR3

```
                    10        20        30        40        50        60
                    .........|.........|.........|.........|.........|.........|
1.1auq              DLVFLLDGSSRLSEAEFEVLKAFVVDMMERLRISQKWVRVAVVEYHDGSHAYIGLKDRKR
 .predicted         ceeeeeeeeccccchhhhhhhhhhhhhhhhhcccccccccceeeeeeccceeeeeccccccc
2.1ido              DIAFLIDGSGSIIPHDFRRMKEFVSTVMEQLKKSK--TLFSLMQYSEEFRIHFTFKEFQN
 .predicted         ceeeeeeccccccchhhhhhhhhhhhhhhhhhccccc??ccceeeeeeccceeeeeechhhhhh
Homology            *:.**:***. :   :*. :* **  :**:*: *:  . .::::* :  : :: :*: :.

                           70        80        90       100       110       120
                    .........|.........|.........|.........|.........|.........|
1.1auq              PSELRRIASQVKYAGSQVASTSEVLKYTLFQIFSKIDRPEASRIALLLMASQ-EPQRMSR
 .predicted         hhhhhhhhhccccccccccchhhhhhhhhhcccccccccccceeeeeeeccc?cchhhhc
2.1ido              NPNPRSLVKPITQLLGRTHTATGIRK-VVRELFNITNGARKNAFKILVVITDGEKFGDPL
 .predicted         ccchhhhhccccccccccchhhhhhh?hhhhccchhhccccccceeeeeeeccccceeccc
Homology             .: * :.. :.   .:. ::: : * .: ::*.  : .. . : :*:: :: *    .

                         130       140       150       160       170       180
                    .........|.........|.........|.........|.........|.........|
1.1auq              NFVRYVQGLKKKKVIVIPVGIG----PHANLKQIRLIEKQAPENKAFVLSSVDELEQQRD
 .predicted         chhhhhhhhhhcceeeeeeeeec????ccccchhhhhhhhhhcccccceeeecchhhhhhhcc
2.1ido              GYEDVIPEADREGVIRYVIGVGDAFRSEKSRQELNTIASKPPRDHVFQVNNFEALKTIQN
 .predicted         chhhhhhhhhhccceeeeeeeecccccccchhhhhhhhhhccccchhhheeecccccchhhhhh
Homology            .:   :   .:: **   :*:*    .. . :::. * .:.*.::.* :...: *:  ::
```

15,9 % d'identité ; Blast E-value = 0,05 ; Sov = 81,7


c.geourjon@ibcp.fr

# Geno3D : Résultat de la modélisation




Les structures géométriquement correctes, sans violation des contraintes introduites.
« Energie » = -547 kcal $mol^{-1}$ ; « Rmsd » = 0,36Å
Rmsd par rapport à la structure expérimentale (rayons X) : 3,3 Å

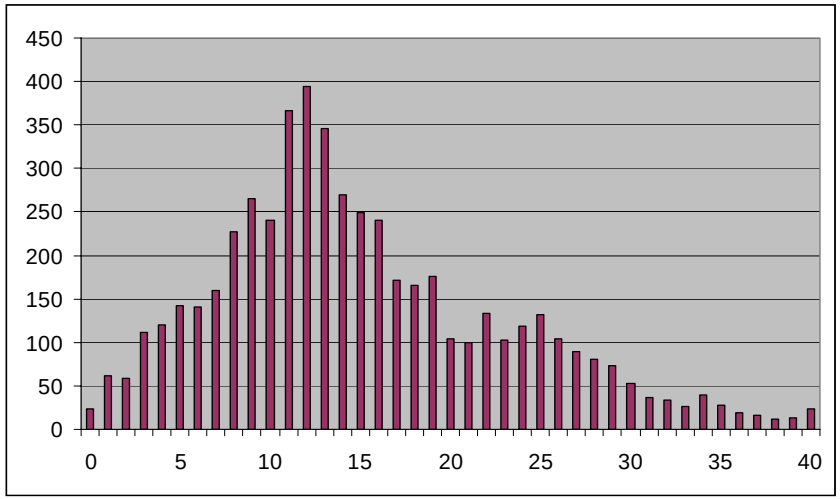

c.geourjon@ibcp.fr

## Stratégie mise en oeuvre

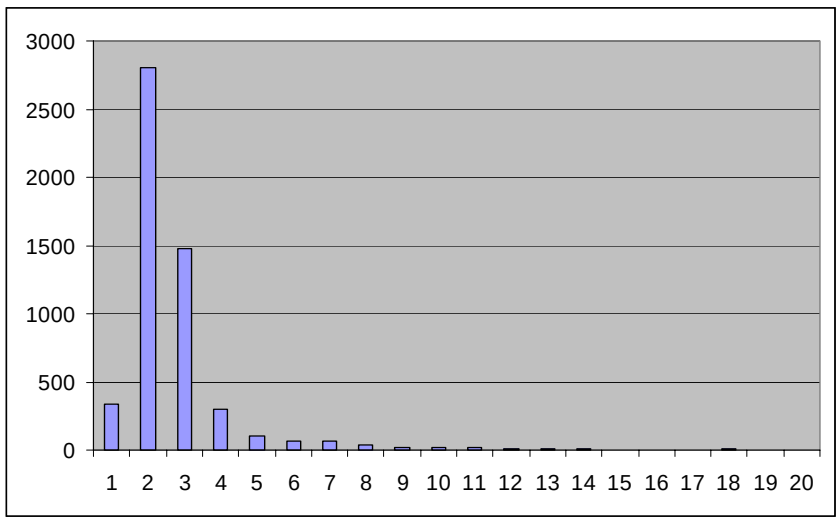
- ✓Modélisation moléculaire à bas taux d'identité (entre 10 et 35% d'identité) des 1 315 protéines représentatives de l'ensemble des protéines de structure 3D connues. (pdb 25%)
- ✓Pour chacune de ces entrées nous avons réalisé une recherche de similarité sur la banque PDB avec le programme PSI-BLAST (maximum 5 itérations). Nous avons sélectionné les empreintes possibles présentant seulement entre 10 et 35% d'identité de séquence (région délicate pour la modélisation moléculaire) soit un total de 5 390.
- ✓La modélisation de l'ensemble de ces modèles a été effectuée sur la ferme de PC sous Linux (382 cpu) du centre de calcul de l'IN2P3 sur le campus de La Doua à Villeurbanne. Ceci a nécessité 7 881 heures de calcul.


c.geourjon@ibcp.fr

# Résultats (1)




Répartition Z-score




Répartition rmsd

Analyse qualitative des modèles générés réalisée en utilisant le logiciel DALI (Holm & Sander, 1998). Chaque modèle a été comparé à la structure expérimentale.

Dans 5110 cas (96%) les modèles possèdent une similarité structurale avec la structure expérimentale, 94% ont un rmsd (au niveau de la chaîne principale) inférieur à 5Å et peuvent donc être considéré comme pertinents.

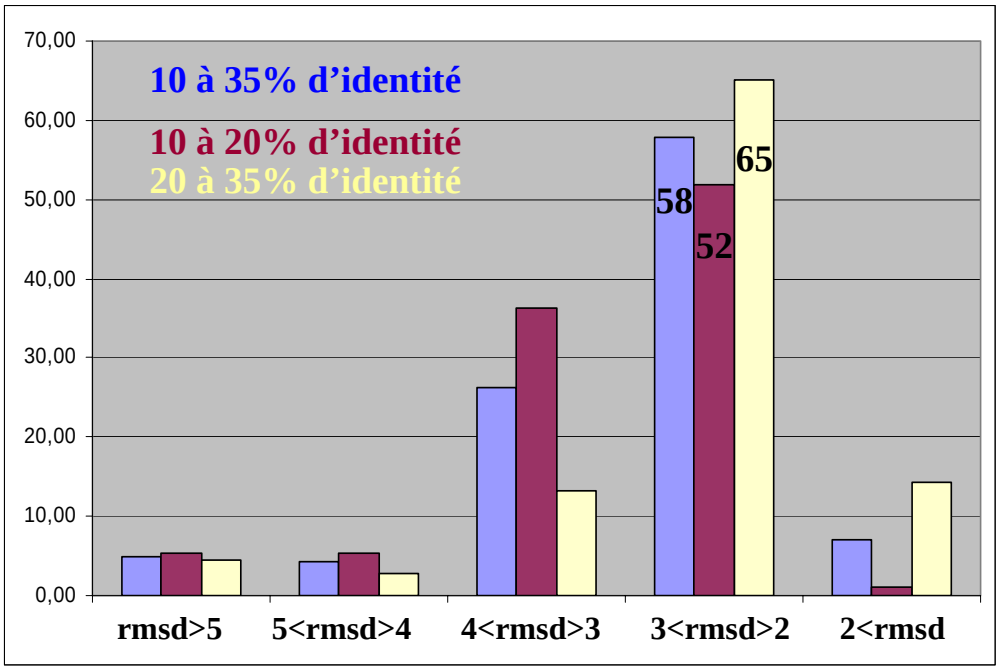
**Globalement, le taux de succès de ce test à grande échelle et à bas taux d'identité est donc de 90% (66% à 3Å).**

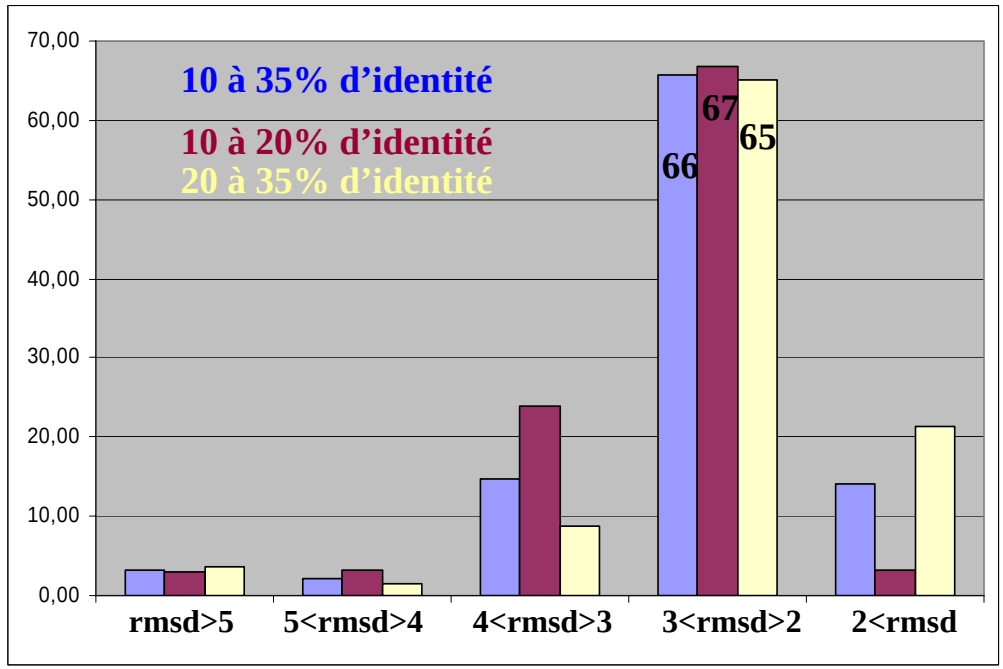
**Utilisable pour des applications en biologie et chimie.**



c.geourjon@ibcp.fr

# Résultats (2)




**Sans utilisation de l'information Sov**

**Avec utilisation de l'information Sov (seuil fixé à 60%)**

Utilisable pour des applications en biologie (mutagénèse dirigée, immunologie), en biologie structurale (remplacement moléculaire) et dans les cas les plus favorables pour l'étude d'interaction avec des ligands.

**En utilisant l'information des prédictions de la structure secondaire, le taux de prédiction sur le serveur Geno3D atteint presque 40% des requêtes.**



c.geourjon@ibcp.fr

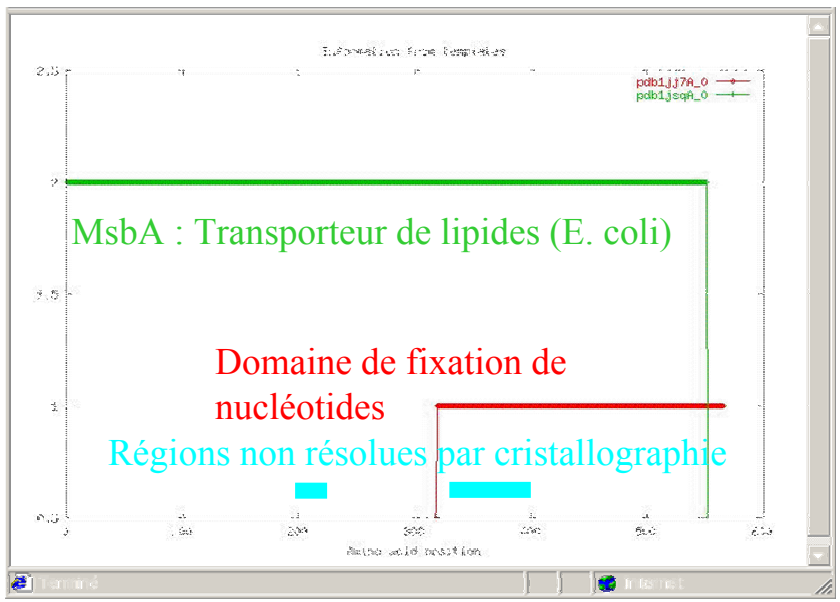
# Serveur Geno3D : exemple d'application

✓Possibilité d'utiliser plusieurs empreintes plus ou moins chevauchantes. Modélisation par domaines ou sous domaines.

Protéine a modéliser

Empreintes possibles (Cas 1)

➢Exemple : Modélisation d'un membre de la famille des transporteurs ABC (collaboration avec Dr. Jean Michel Jault, CEA Grenoble)






c.geourjon@ibcp.fr

# Geno3D : Bilan

- Bonne fiabilité y compris à bas taux d' identité
- Possibilité d'avoir une estimation de la qualité du modèle obtenu.

- Ne pas introduire d'*a priori* au niveau des «gaps» et insertions dans l'alignement.

- Possibilité de modéliser des dimères ou trimères
- Possibilité de modéliser des protéines par domaines (notion de logo moléculaire)

- Possibilité d'inclure les ligands (géométriquement puis minimisation)

- Disponible sur le Web pour les académique : **http://geno3d-pbil.ibcp.fr**

- Disponible également sur serveur sécurisé

- Automatisation en cours (fonction de scoring)

- Utilisation dans le cadre de modélisation moléculaire à grande échelle (protéome complet). Programme GENOPLANTE – *Arabidopsis thaliana*.

- Possibilité de combiner des données provenant de protéines proches (séquences ou fonctions) mais aussi des données expérimentales ou théorique **du type : Ponts di-sulfure, Ponts salins, Interaction entre les brins β**

- Dans la suite du programme nous souhaitons augmenter de manière très significative le taux de génération de modèles 3D grâce à l'utilisation de modèle d'apprentissage coopératif pour la classification, l'extraction de la structure prototype et la prédiction de la structure 3D

Geno3D: Automatic comparative molecular modelling of protein
C. Combet, M. Jambon, G. Deléage & C. Geourjon
*Bioinformatics*, 2002, 18, 213-214



c.geourjon@ibcp.fr

## Utilisation d'un jeu de distance pour rechercher dans la banque de structures 3D les protéines de même repliement

## Travaux menés sur des données expérimentales (spectrométrie de masse), mais utilisable aussi avec des données théoriques du type :

-Ponts di-sulfure
-Ponts salins
-Interaction entre les brins β


c.geourjon@ibcp.fr

**Séquence de la protéine d'intérêt**

**Analyse de séquence**

- Prédiction des structures secondaires

**Données expérimentales**

- Réticulation chimique
- Protéolyse
- Identification des peptides par SM

**Contraintes de distances**

**Distances ambiguës**

**Banque de données structurale (banque PDB)**

**Familles structurales**

- Génération d'une base de données des familles structurales selon FSSP (DALI)
- Recherche des segments conservés
- Construction des matrices de distances pour chaque famille sur le cœur structural

**Recherche des hits pour lesquels les distances correspondent**

- Génération de toutes les combinaisons de contraintes en fonction du cœur structural
- Criblage de la base de donnée des matrices de distances

**Filtrage grâce aux contraintes ambiguës**

**Filtrage des résultats**

- Pourcentage de structure secondaire moyen
- Hydrophobie, Accessibilité au solvant, Amphiphilie, Flexibilité

**Sélection des hits les plus représentés pour chaque empreinte**

- Construction des histogrammes pondérés des positions trouvées selon leur occurrence

**Regroupement des hits par famille de repliement semblables**

- Phylogénie structurale par alignement des structures 3D (calcul d'une déviation standard moyen au niveau des carbones alpha et d'un Z-score, algorithme CE)

**Repliements potentiels de la protéine d'intérêt**

**Retour à l'expérience**

- Spécificité différente des agents réticulants
- Longueur du bras espaceur différente
- Utilisation d'une protéine homologue (au niveau expérimental)

**Génération des modèles 3D possibles**



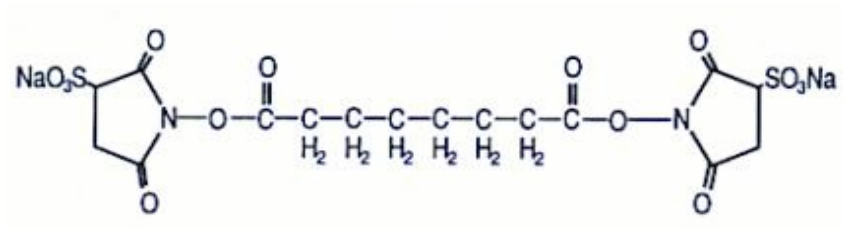
c.geourjon@ibcp.fr

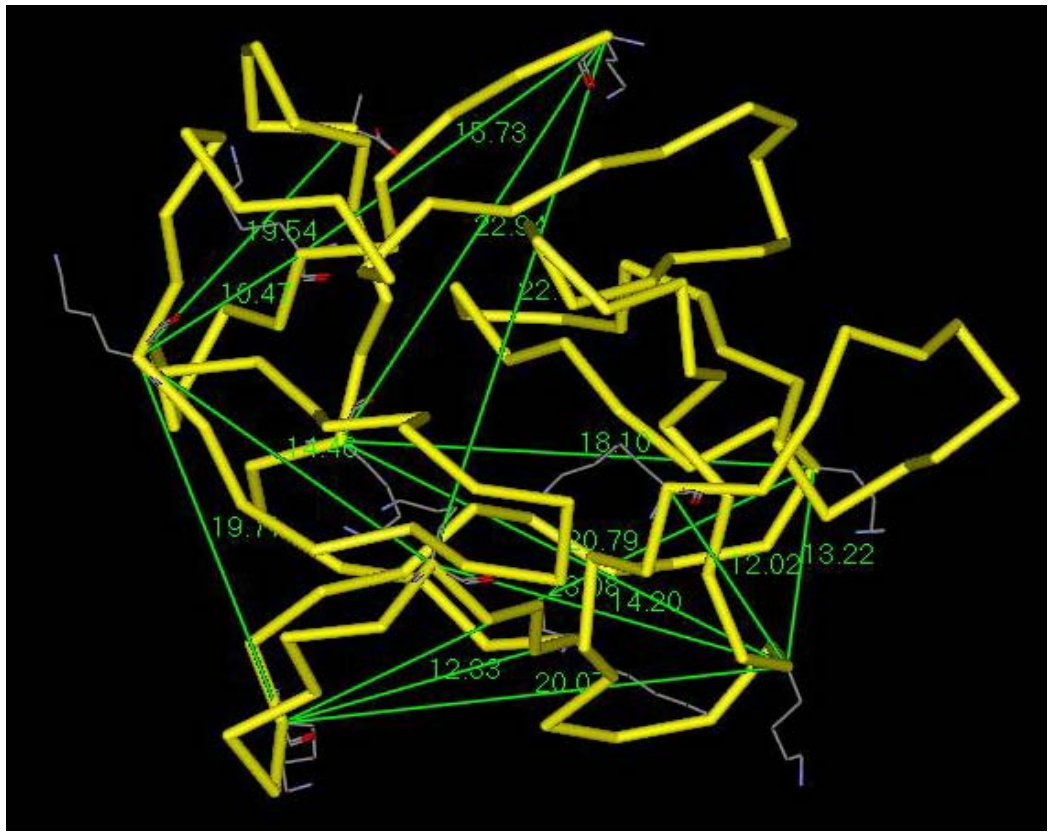
# 1er jeu de données test

Travail sur les données de l'article de Young *et al* (2000)

- **Protéine étudiée :**
  - basic fibroblast growth factor : FGF-2
    (code pdb : 1BLA)

- **Linker** :
  - $BS^3$ : bis(sulfosuccinimidyl) suberate
  - cross-link **LYS-LYS (NH$_2$)**
  - longueur du bras espaceur : 12 Å




- **Résultats expérimentaux** :
  - 15 contraintes donnant des informations


c.geourjon@ibcp.fr

# Résultats : FGF-2

| Filtres appliqués<br>Tolérance | Distances | Pourcentage de structures secondaires<br>20 % | Hydrophobie<br>μ+2σ | Accessibilité au solvant<br>μ+2σ | Flexibilité<br>μ+2σ | Amphiphilie<br>μ+2σ | Histogramme |
|---|---|---|---|---|---|---|---|
| Total des hits | 7 486 913 | 1 116 582 | 338 031 | 288 646 | 269 001 | 261 654 | 95 |
| Hits parmi les bons repliements | | 6 611 | 5 226 | 4334 | 4274 | 3886 | 19 |
| | | **19 / 20** | **18 / 20** | **18 / 20** | **18 / 20** | **18 / 20** | **17 / 20** |
| **Enrichissement** | | **0.59** | **1.55** | **1.50** | **1.59** | **1.49** | **20.00** |

⇒ Près de 7,5 millions de hits

⇒ 95 hits conservés au final



c.geourjon@ibcp.fr