Bioinformatique structurale : Un enjeu de l'après génome

IN'Tech Lyon le 23/10/2003

Gilbert DELEAGE

PBIL-IBCP-CNRS UMR 5086
Pôle Bioinformatique Lyonnais Lyon-Gerland
7, passage du Vercors
69367 Lyon cedex 07
Tél: +33 (0)4 -72-72-26-55
fax: +33 (0)4 -72-72-26 -01
mel: g.deleage@ibcp. fr

Equipe Bioinformatique et RMN structurales

Permanents

DELEAGE G. PR1 UCBL, Bioinformatique
PENIN F. IR2-CNRS, NMR, Biophysique, Biochimie
BETTLER E. MC2, Bioinformatique
BLANCHET C. IR2-CNRS, Bioinformatique
BOCKMANN A. CR2-CNRS, NMR
COMBET C. CR2-CNRS Bioinformatique
GEOURJON C. IR2-CNRS, Bioinformatique
HUET E. CR2-CNRS, Biologie-Biochimie
LAVERGNE J-P CR1-CNRS, Biochimie
MONTSERRET R. IE2-CNRS, NMR, Biophysique

Post-docs and CDD

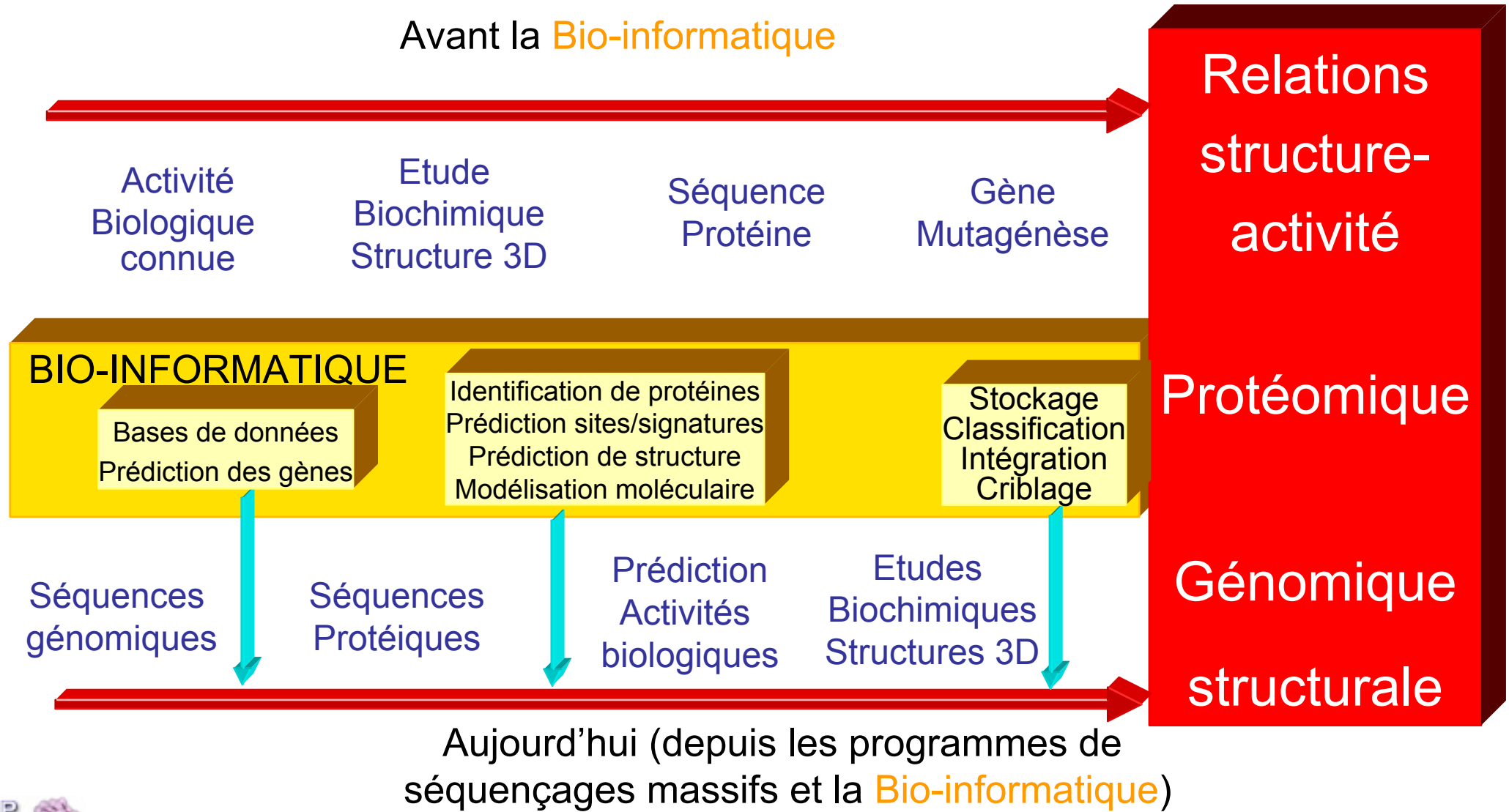
DORKELD F. (EC), Bioinformatique
LECLUSE A. (French ministry) Bioinformatique
LACORNE N. (ACI GRID) Bioinformatique
MISSEREY S. (Genopôle Rhône-Alpes) Bioinformatique

Etudiants

BOULANT S. Biochimie
GIRAUD N. NMR
RATINIER M. Biochimie
SAPAY N. Bioinformatique
VERNOIS A. GRID computing, Informatique

PBIL-IBCP

La révolution génomique - Une stratégie revisitée



g.deleage@ibcp.fr

Réalisations bioinformatiques du groupe PBIL-IBCP

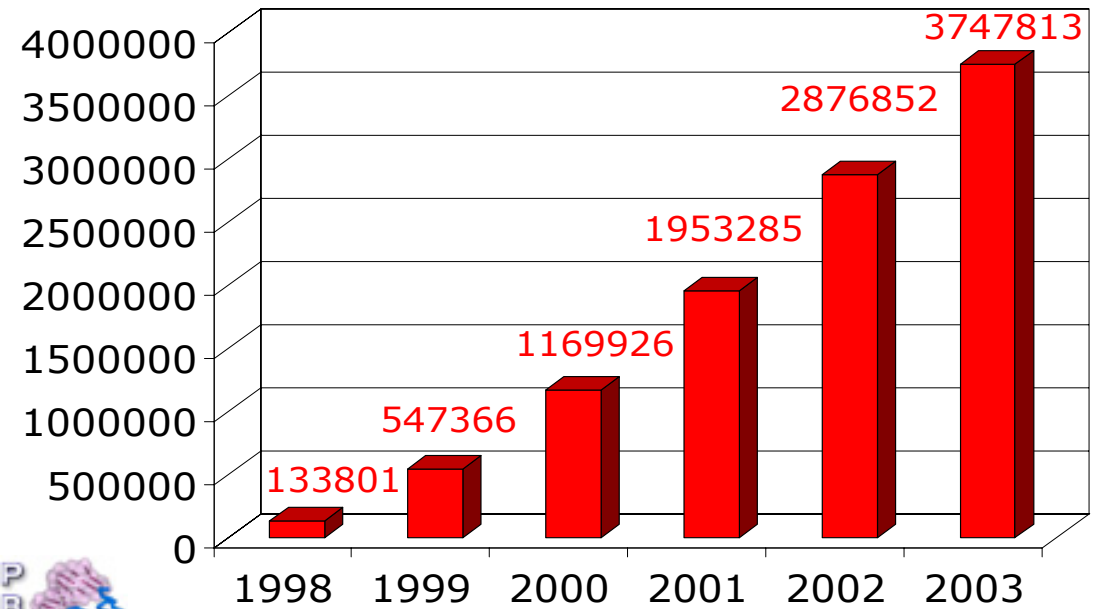
http://pbil.ibcp.fr

- **Database :**
  - **HCVDB** : Base de données de séquences annotées d'HCV; **http://hepatitis.ibcp.fr**)
- **Méthodologies :**
  - **MLRC** : Prédiction de la structure secondaire des protéines (Guermeur *et al.*, **Bioinformatics**, 1999) Coll. LIP6
  - **ProScan** : Scan PROSITE avec recherche pondérée par un système original
  - **PattInProt** : Algorithme de recherche original de signatures dans les banques. Outil de PROTEOMIQUE
  - **PROCSS** : Compatibilité des structures protéiques grâce aux structures secondaires (Geourjon *et al.*, **Protein Sci.**,2001; Errami *et al.*, **Bioinformatics**, 2003)
  - **SOPM** :Méthode auto-optimisée de prédiction de structure secondaire (Geourjon *et al.*, **Protein Eng.**, 1994)
  - **SOPMA** : Méthode auto-optimisée de prédiction de structure secondaire avec alignements (Geourjon *et al.*, **Comput Appl Biosci.**, 1995)
  - **SuMo** : Détection de sites 3D dans les protéines (Jambon *et al.*, **Proteins**, 2003, Brevet international)
- **Logiciels :**
  - **AnTheProt** : Logiciel intégré d'analyse de séquences client/serveur (Deléage *et al.*, **Comput Appl Biosci.**, 1988; Deléage *et al.*, **Comput Biol Med.**, 2001; **http://antheprot-pbil.ibcp.fr**)
  - **AnTheNuc** : Analyse de sites de restriction
  - **Bioread** : Interface graphique pour **extractblast** and **extractfasta** programmes "parser".
  - **DicroProt** : Analyse des spectres de dichroïsme circulaire des protéines (Deléage G et Geourjon C.**Comp. Appl. Biosc.,** 1993)
  - **MPSA** : integrated protein sequence analysis with client/server capabilities (Blanchet *et al.*, **Bioinformatics**, 2000; **http://mpsa-pbil.ibcp.fr**)
  - **SecTrace** : secondary structure plot
- **Services Web:**
  - **NPS@** : Analyse intégrée de séquences sur le Web (Combet *et al.*, **Trends Biochem Sci**, 2000; Perrière, Combet *et al.,* **Nucleic Acids Res.**, 2003; **http://npsa-pbil.ibcp.fr**)
  - **Geno3d** : Modélisation moléculaire de protéines à grande échelle (Combet *et al.*, **Bioinformatics**, 2002; **http://geno3d-pbil.ibcp.fr**)
  - **SuMo** : protein common 3D sites detection (Jambon *et al.*, **Proteins**, 2003; **http://sumo-pbil.ibcp.fr**)
- **Applications biologiques:**
  - Très nombreuses en collaboration avec des biologistes

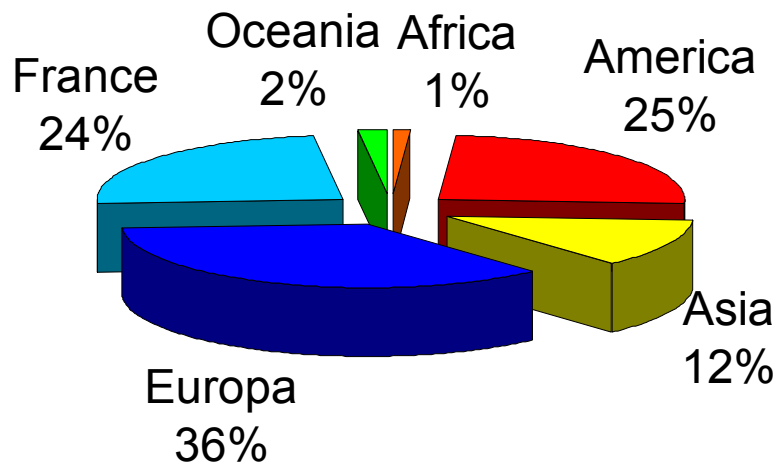
g.deleage@ibcp.fr

Portail de bioinformatique : Webiciel - NPS@

- **NPS@ http://npsa-pbil.ibcp.fr**
- **Partie «protéine et structure» du Pôle BioInformatique Lyonnais**
- Interconnexion de **46 méthodes** d'analyse de séquences de protéine
- **Récupération automatique des données** dans des logiciels clients/serveurs d'analyse biologique MPSA, AnTheProt, Clustal X, RasMol, …
- Liens hypertextes sur les données de **17 banques de données internationales** (SWISS-PROT, PROSITE, PDB, SCOP,…).
- **Mise à jour automatique** des bases de données disponibles sur le serveur (SP, SP-TrEMBL, NRL3D, Nr, PDB, PROSITE, *etc.*).
- **Références internationales:** Expasy, University of California, InfoBioGen, RSCB(PDB),….



2530 analyses / jour en 2002
3584 analyses / jour depuis janvier 2003



Combet *et al.*, **Trends Biochem Sci**, 2000; Perrière *et al.*, **Nucleic Acids Res.**, 2003

g.deleage@ibcp.fr

Projets bioinformatique en cours

- **GRID computing :**
  - **GPS@** : Portail d'analysesde séquences de protéines sur la Grille (Déploiement de NPS@ sur la grille, EU FP5 DATAGRID)
  - **GriPPS** : Services de recherche de signatures sur la grille (Ministère rehcerche ACI-GRID, CNRS-IN2P3, ENSL-LIP)
  - **RUGBI** : Prédictions de structures secondaires sur la grille GRID (Ministère rehcerche et CNRS-IN2P3 (V. BRETON), Biopôle Clermont-Limagne, CS, ECP)
  - **e-Toile** : Projet de grille expérimentale (Ministère rehcerche RNTL)
  - **GiGn** : Réseau de recherche français de **Gr**I**l**le pour la **G**e**N**omique (Ministère rehcerche Action IMPG, with V. BRETON)
  - **HealthGrid** : Organisation de conférences (HealthGrid Conference, Lyon FR, January 2003)
- **Bioinformatique structurale et clinique :**
  - **CPS@** : Version pour cliuster de NPS@ (PBS, PVM and MPI algorithms)
  - **euHCVdb** : Base européenne de séquences HCV (EU FP5 HepCVax QLK2-CT-2002-01329 ; EU FP6 viRgil)
  - Collaborarion avec D. KAHN (**PRODOM**)
    - Amélioratiomdes domaines PRODOM grâce aux prédictions de structures
    - Connexion avec Geno3D
  - **StrucAnnot** Annotation structurale de genome *Arabidopsis thaliana* : (consortium France, Belgium (P. ROUZE, Gent) and SIB (A. BAIROCH, Geneva))
  - **MSFold** : Modélisation de structure 3D à partir des contraintes de MS. Protéomique structurale, (coll. E. Forest, IBS, Grenoble)
  - **MADPROTS** : Modélisation, analyses et serveur Web pour amarrage moléculaire
  - **SYNAPSE** : Système expert d'annotation de protéines

g.deleage@ibcp.fr

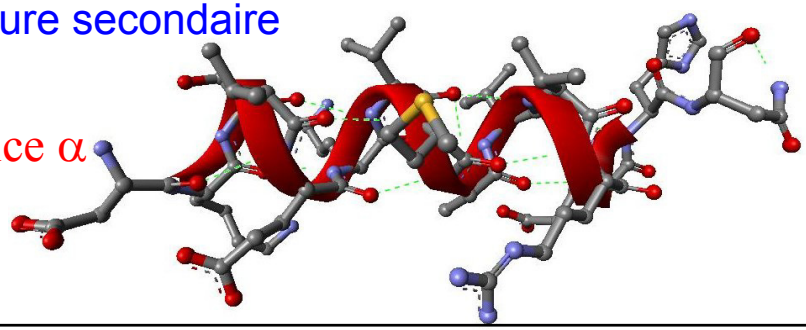
Organisation hiérarchique des protéines

Structure primaire = séquence= mot écrit avec un alphabet de 20 lettres

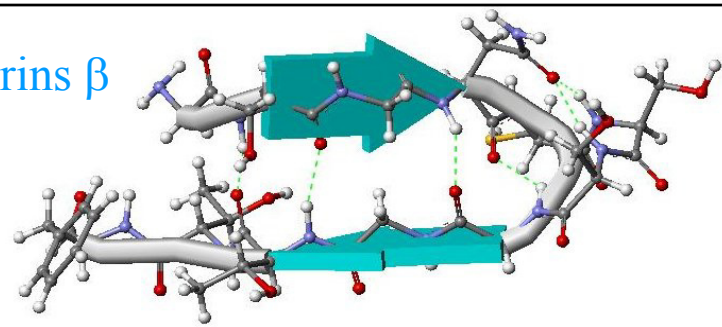
MKLD**E**IARLAGVSRTTASYVINGKAKQYRVSDKTVEKVMAVVREHNYHPNAVAAGLRAGR

Structure secondaire

Hélice α



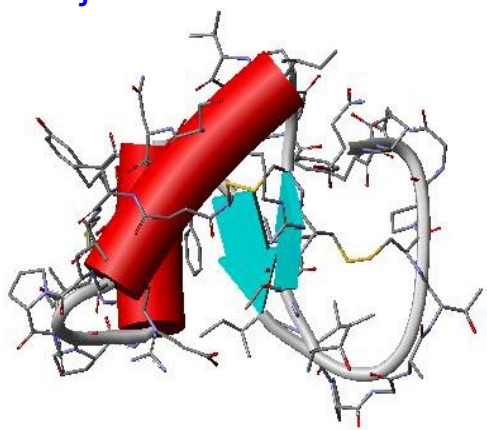
Brins β



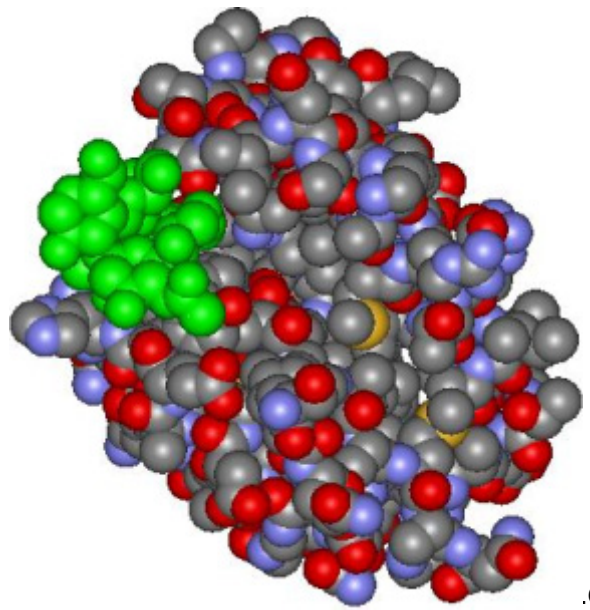
Structure secondaire = mot alphabet de 3 à 10 lettres

CCHHHHHHHHHHHCCCEEEETTTTEEEEECCCCHHHHHHHHHHHCCHHHHHHHHHGCCCC

Structure tertiaire = objet 3D



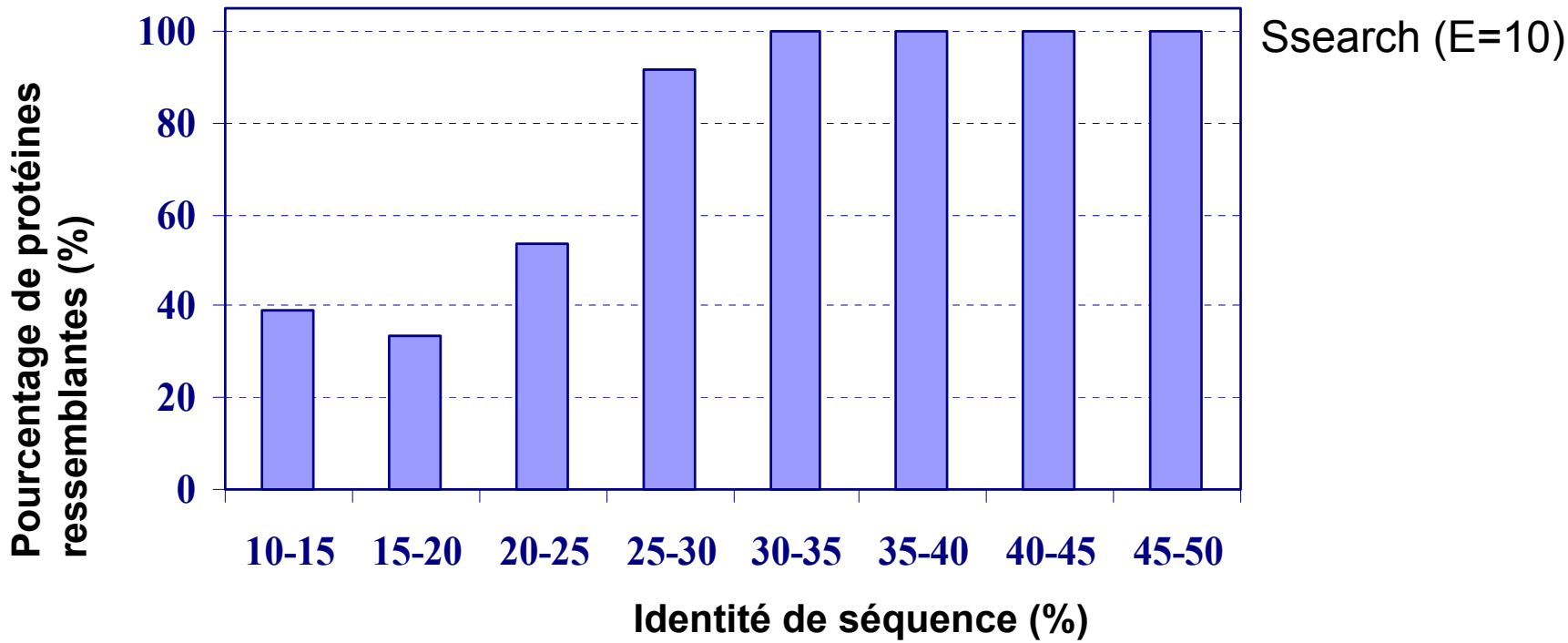
Fonction



.deleage@ibcp.fr

Relation identité de séquence ressemblance structurale

```
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
Identity         *  * : **:* *.* **   *****..** *.***:.*
```



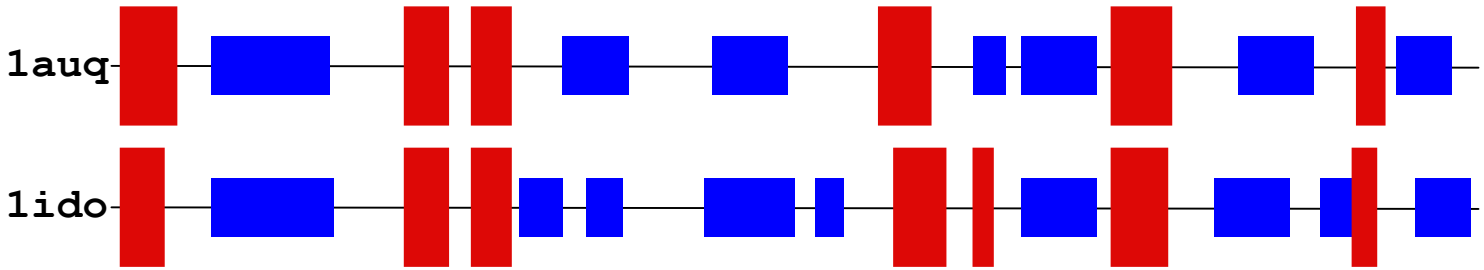
Nécessité de discriminer dans l'intervalle 10-30% d'identité

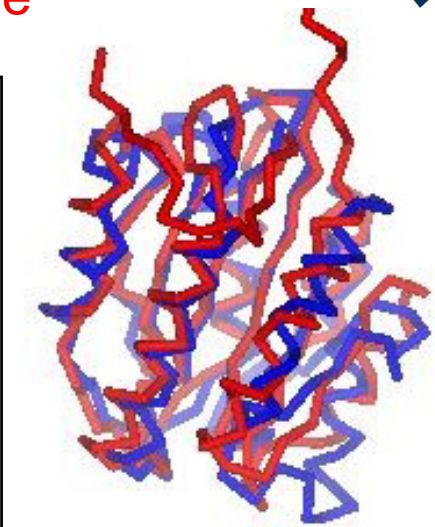
g.deleage@ibcp.fr

Utilité des prédictions de structure secondaire

Sov = 81

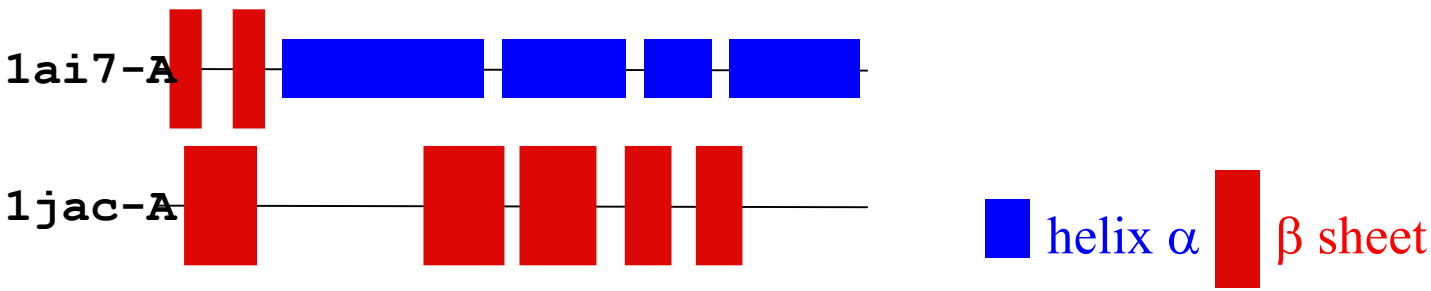
15,9% identité ; Protéines apparentées (FSSP : RMSD = 2,3A ; Z-score = 19,9)

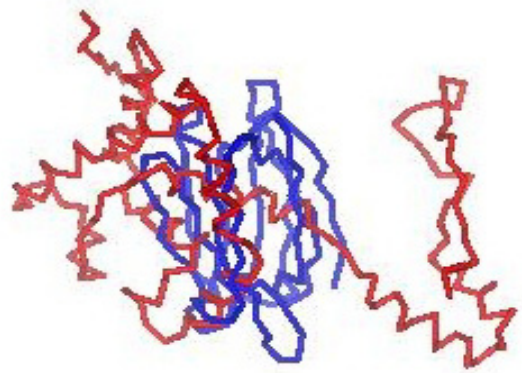




Sov = 9

16% identité ; Protéines non apparentées





Les structures secondaires permettent de discriminer les protéines ayant des structures 3D proches « zone floue » de 10-30%

(Geourjon *et al.*, **Protein Sci.**,2001; Errami *et al.*, **Bioinformatics**, 2003)

g.deleage@ibcp.fr

Prédiction de structure 3D

- Modélisation homologique « classique »
  - 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)
- Modélisation par analogie (à faible taux d'identité)
  - 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).
- Threading (en cours de développement)
  - Une séquence est testée sur une librairie de repliements pour déterminer sa compatibilité structure-séquence probable, méthode d'alignement séquence-structures tridimensionnelles.
- *Ab initio* (en progression dans CASP5, folding@home Stanford)
  - Structure directement déduite de la séquence à partir de règles empiriques.

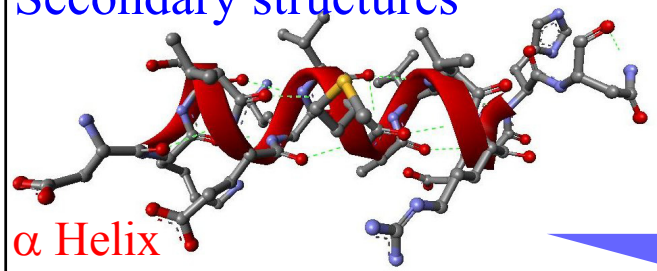
g.deleage@ibcp.fr

Modélisation moléculaire via le Serveur Web Geno3D

Protein sequence

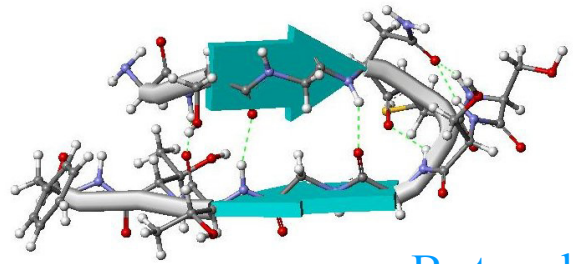
`MKLDEIARLAGVSRTTASYVINGKAKQYRVSDKTVEKVMAVVREHNYHPNAVAAGLRAGR`

Secondary structures

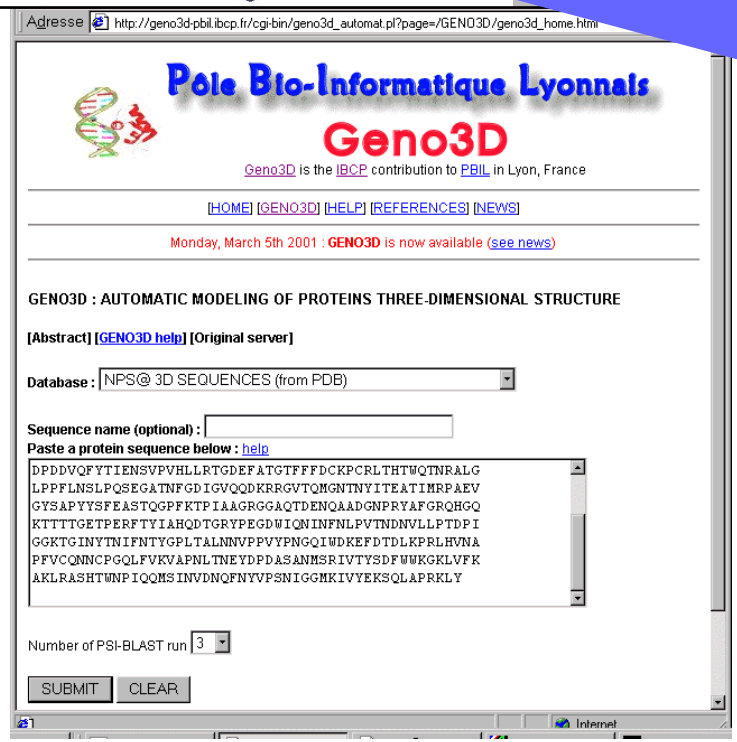


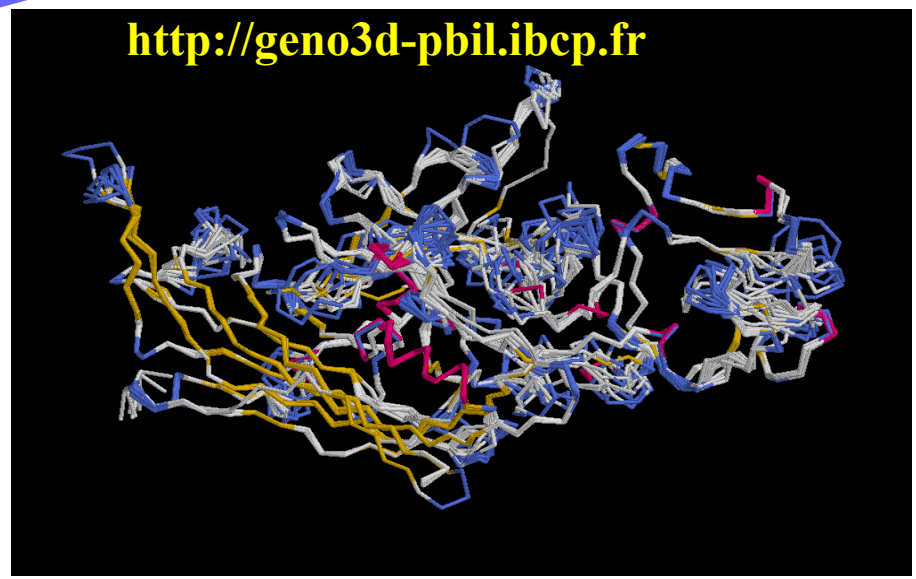
α Helix

Geno3 Tools



B strands



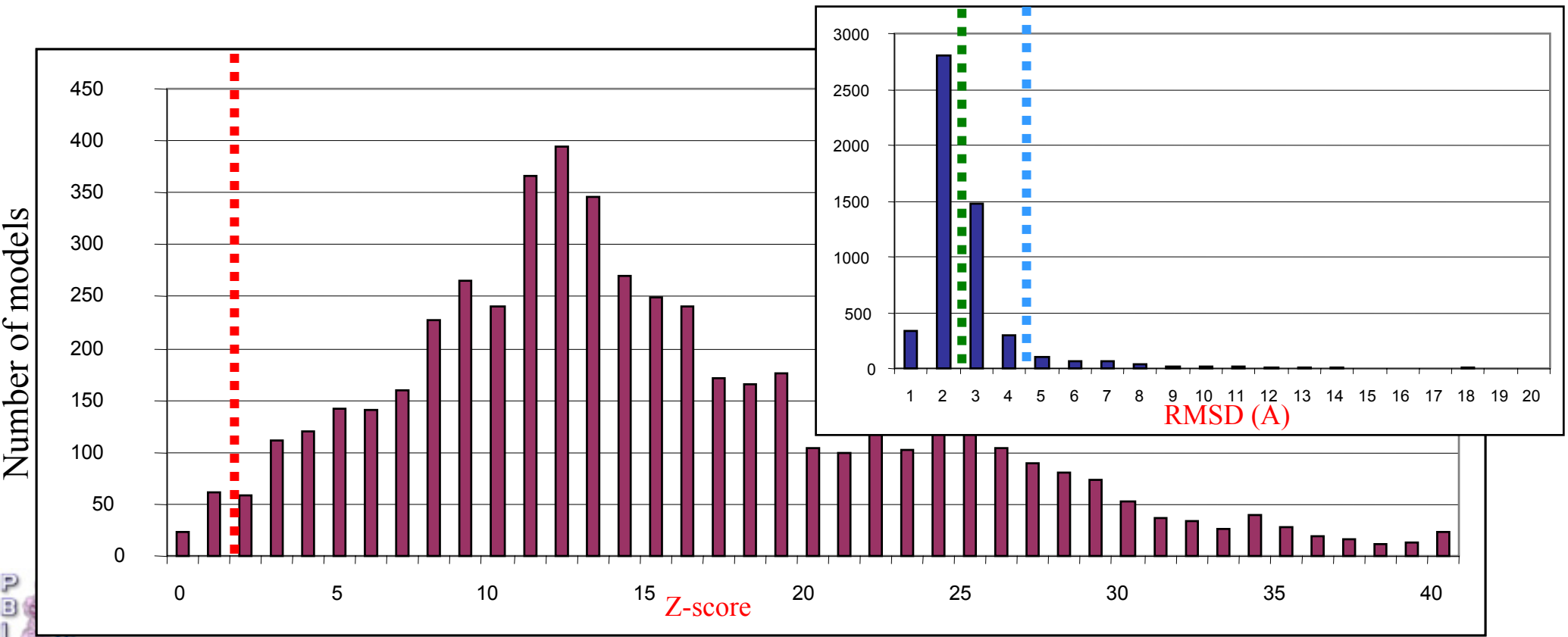


Combet C., Jambon M., Deléage G. et Geourjon, C. (2002) Bioinformatics, 18, 213-214 g.deleage@ibcp.fr

3D-Crunch : Modélisation à grande échelle

Stratégie

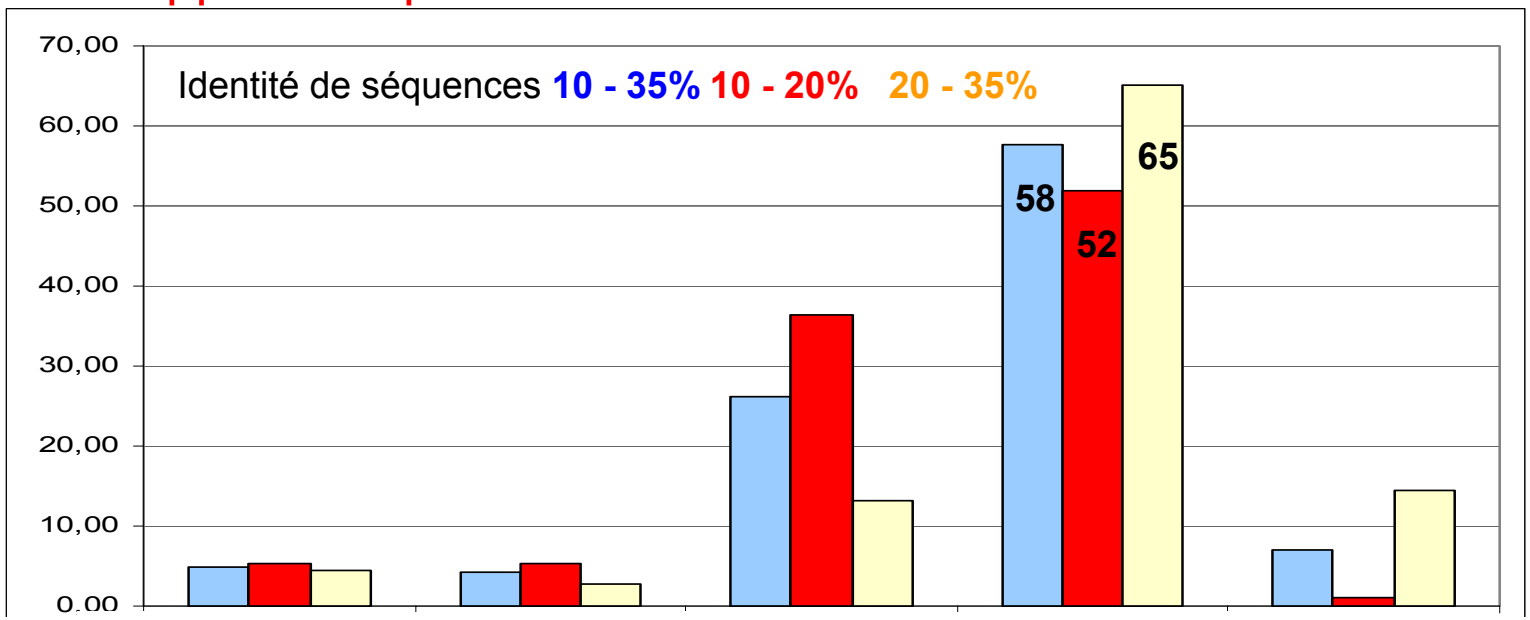
- ✓Modélisation à faible taux d'identité (entre 10-35%)
- ✓1,315 protéines de structures 3D connues (pdb 25%)
- ✓PSI-BLAST sur chaque entrée (max 5 run).
- ✓Les 5,390 protéines ayant un pourcentage d'indentité entre 10 et 35% ont été modélisées
- ✓ Les calculs ont été réalisés au CC-IN_2P_3 Villeurbanne (France) 7,881 heures de CPU



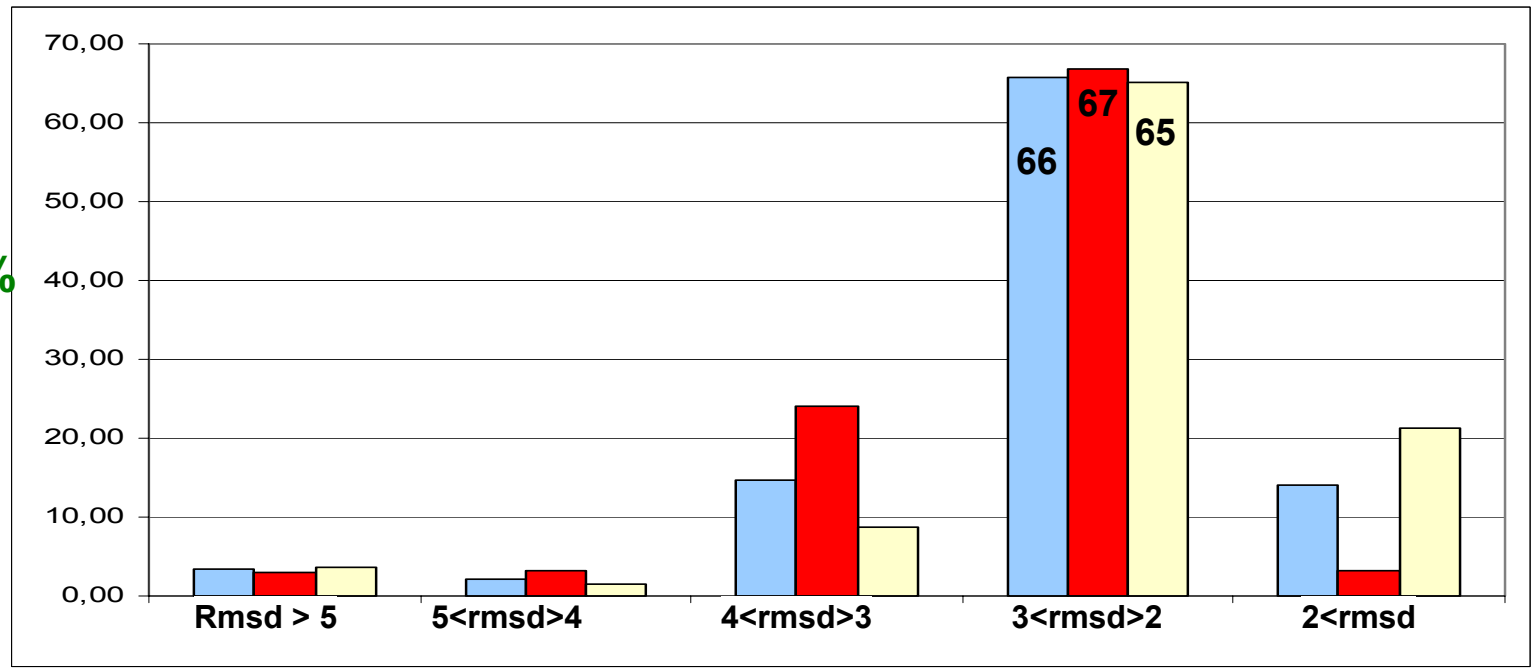
g.deleage@ibcp.fr

Apport des prédictions de structures secondaires

Sans Sov



Avec Sov >60%



g.deleage@ibcp.fr

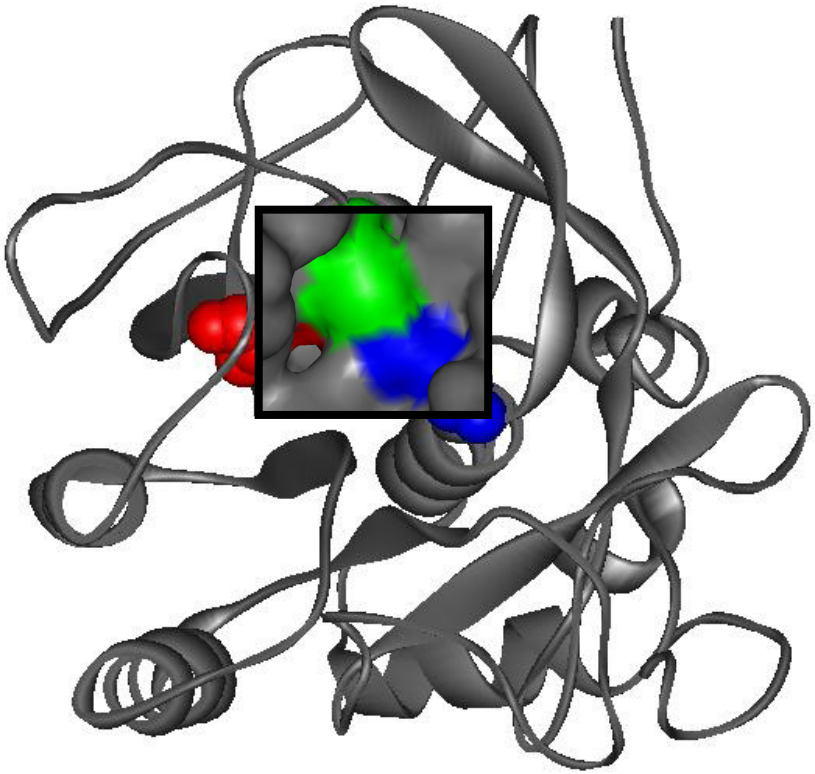
Geno3D : Bilan

- Bonne fiabilité y compris à bas taux d' identité
- Ne pas introduire d'*a priori* au niveau des «gaps» et insertions dans l'alignement.
- Possibilité d'avoir une estimation de la qualité du modèle obtenu.
- Possibilité de combiner des données provenant de protéines proches (séquences ou fonctions) mais aussi des données expérimentales.
- Possibilité de modéliser des dimères ou trimères
- Possibilité de modéliser des protéines par domaines
- Possibilité d'inclure les ligands (géométriquement puis minimisation)
- Disponible également sur serveur sécurisé Contrat de collaboration avec des industriels

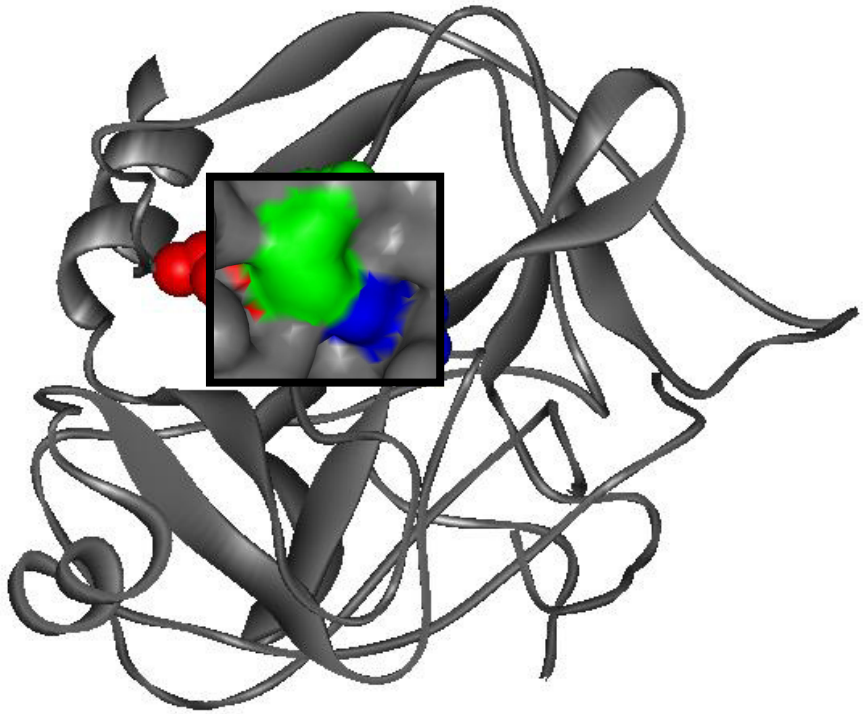
Utilisation dans le cadre de modélisation moléculaire à grande échelle
Protéome complet *Arabidopsis thaliana*. Programme GENOPLANTE

g.deleage@ibcp.fr

Problème 1 : séquences non alignables, repliement différent, même fonction, même site actif



Subtilisine



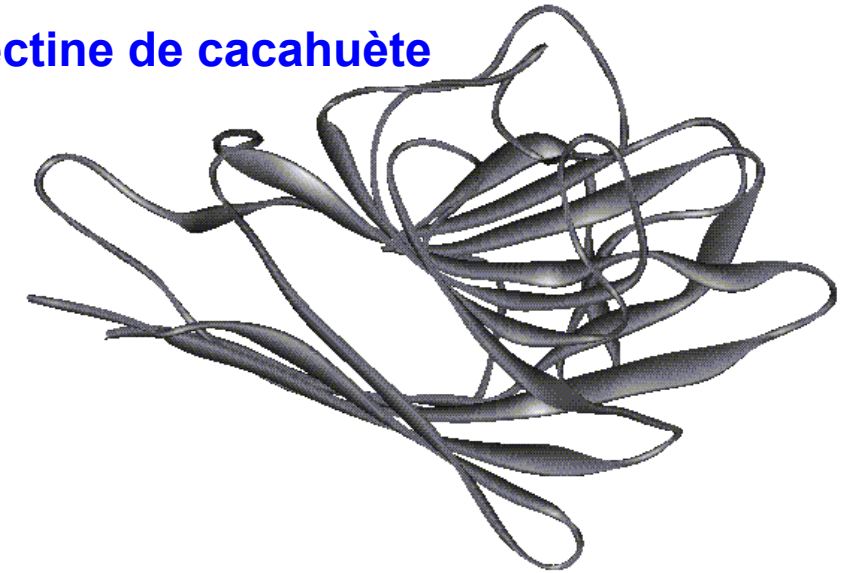
Chymotrypsine

Protéases à sérine

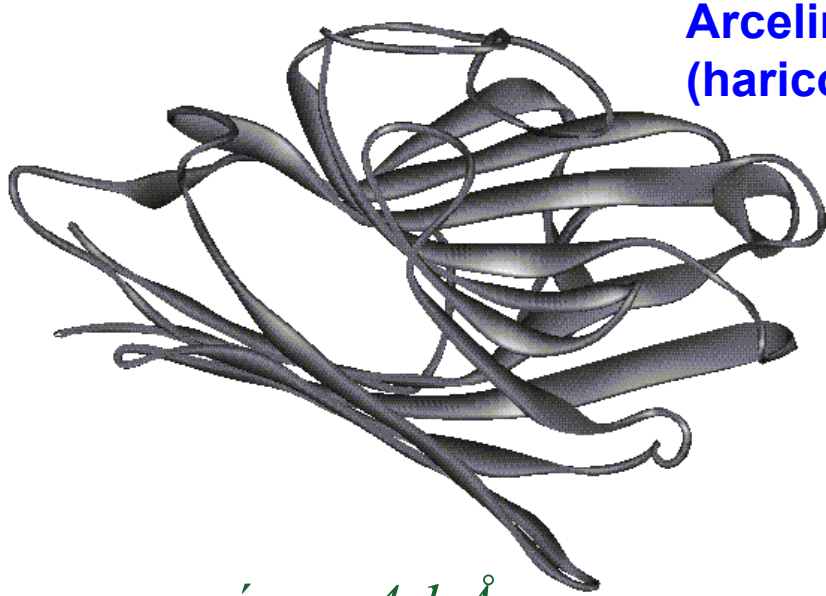
g.deleage@ibcp.fr

Problème 2 : Séquences proches, même repliement, activité différentielle

Lectine de cacahuète



Arceline (haricot)



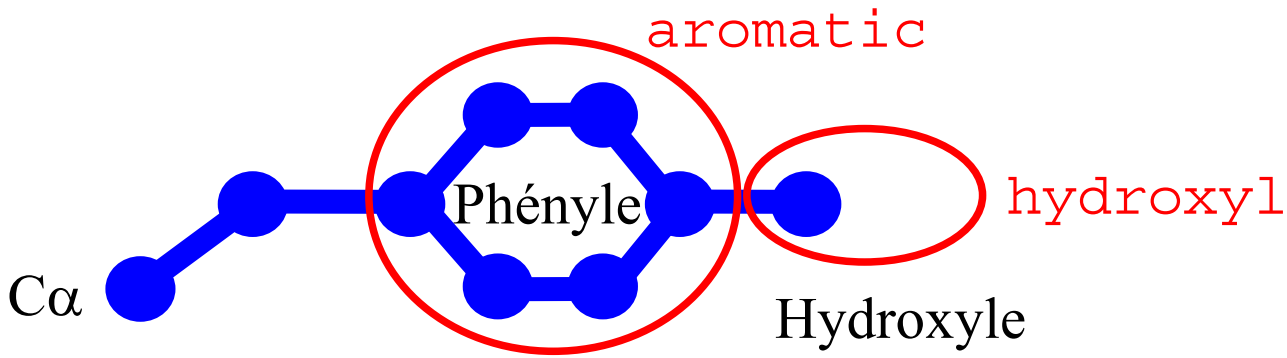
- *RMSD sur positions conservées = 4,1 Å*

```
2PELAx0     A  VSFNF SF  GN  I FQGDVTV SNGNIQLTNLN    KVNSVGRVLYA PVRI S  TGNVASF TSFSF MK  KD  PA GI F IAP DT    S   G LGVSDTK  GH VGV
1IOAAx1     A  TSFNF NF  DD  I LQGNATI SKGQLQLTGVG    RVDSLGRAFYS PIQI D  SNNVASF TNFTF IR  NQ  SA GL F LVP NS    Q   G FNTNNPE  AR VAV
            *  .**** .*  .:  * :**:.*: *:*::***.:.    :*:*:**.:*: *::* .  :.***** *.*:* ::  ::  .* *: * :.* ::    .   * :...:.:  .: *.*

2PELAx0      FDTYSN    D  TDHV   VN     KT  WNS SG  VKV V YDSSTK L V V    GDIT  IA VVDLK KL E VKFGFSASG  S G  Q H I SWSFTSTL TT  R
1IOAAx1      FNTFKN    D  KNFI   VN     EN  FHK NG  TDV I YDSSNN L V L    SQVK  VS TVHLE EV E VSVGFSPTS  T D  E H V SWSFSSKF NK  N
             *:*:.*    *  .:.:   **     :.  ::. .*  ..* : ****.: * * :    .::.  :: .*.*: :: * *..***.:.  : .  : * : ****:*.: ..  .
```

- *26 % d'identité entre les séquences*

g.deleage@ibcp.fr

Etape 1 : découpage en groupements chimiques



Exemple : tyrosine = aromatic, hydroxyl

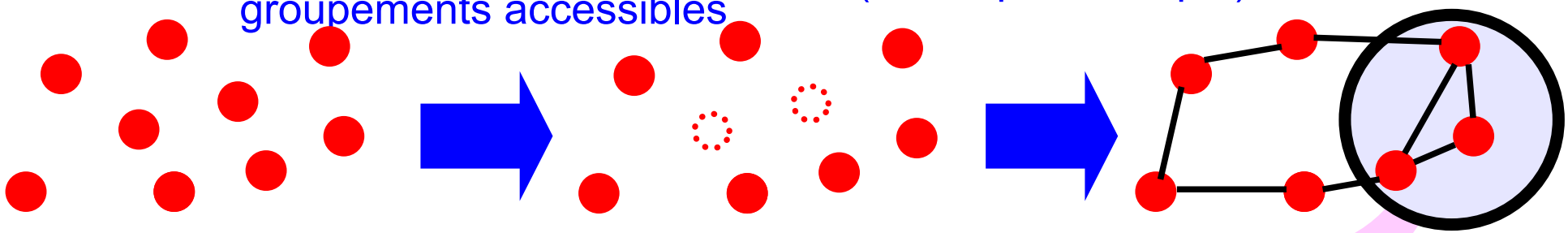
(Jambon *et al.*, **Proteins**, 2003, Brevet international)

g.deleage@ibcp.fr

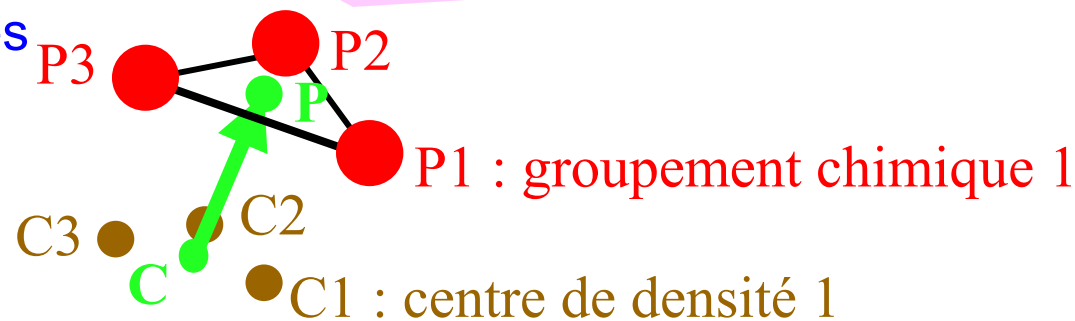
Etape 2 : génération d'un graphe de triplets de groupements chimiques

Sélection des groupements accessibles

Connexion des groupements voisins (< 6 Å, par exemple)



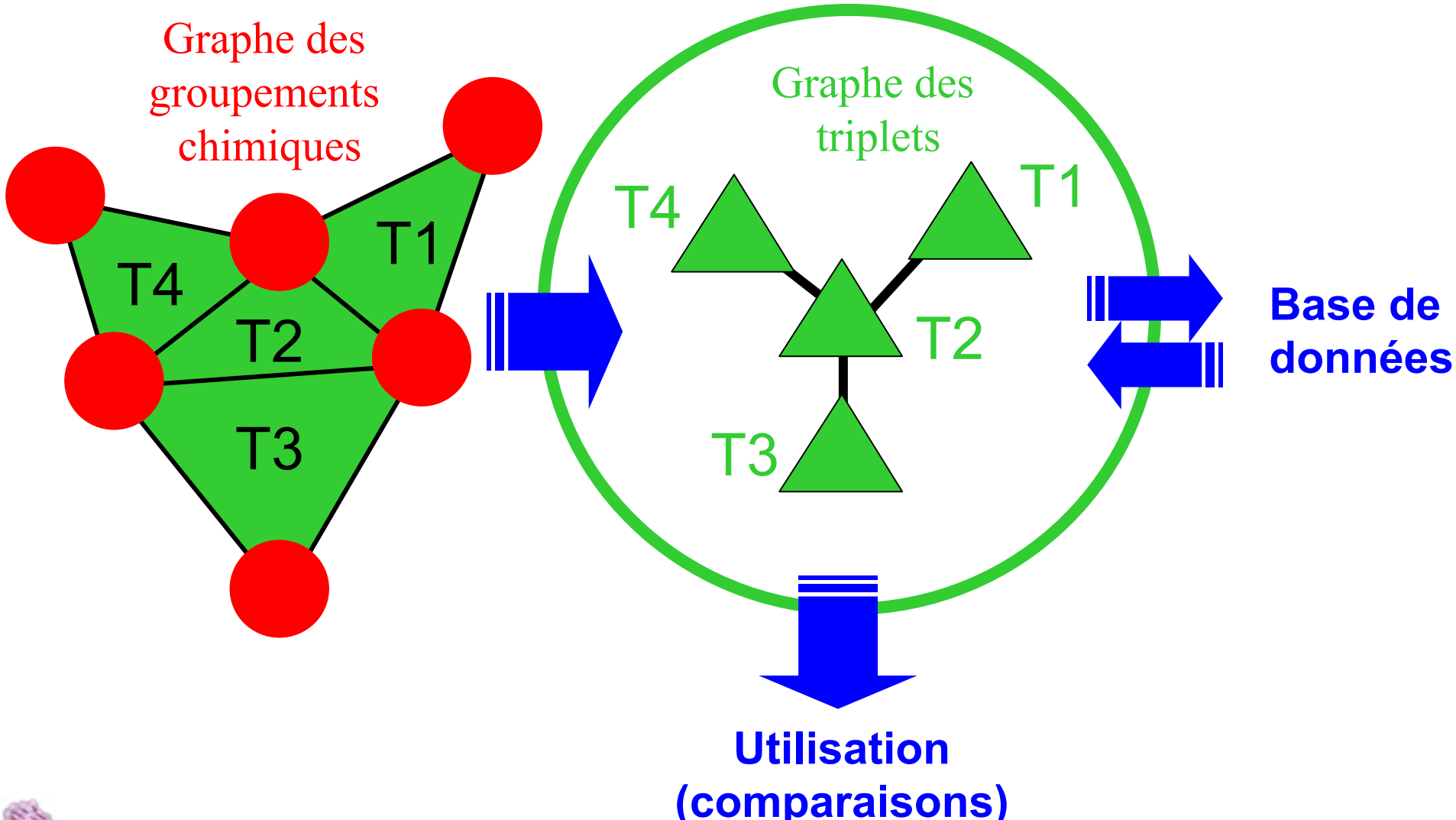
Définition des triplets de groupements chimiques



$\overrightarrow{CP}$: orientation vers l'extérieur de la protéine

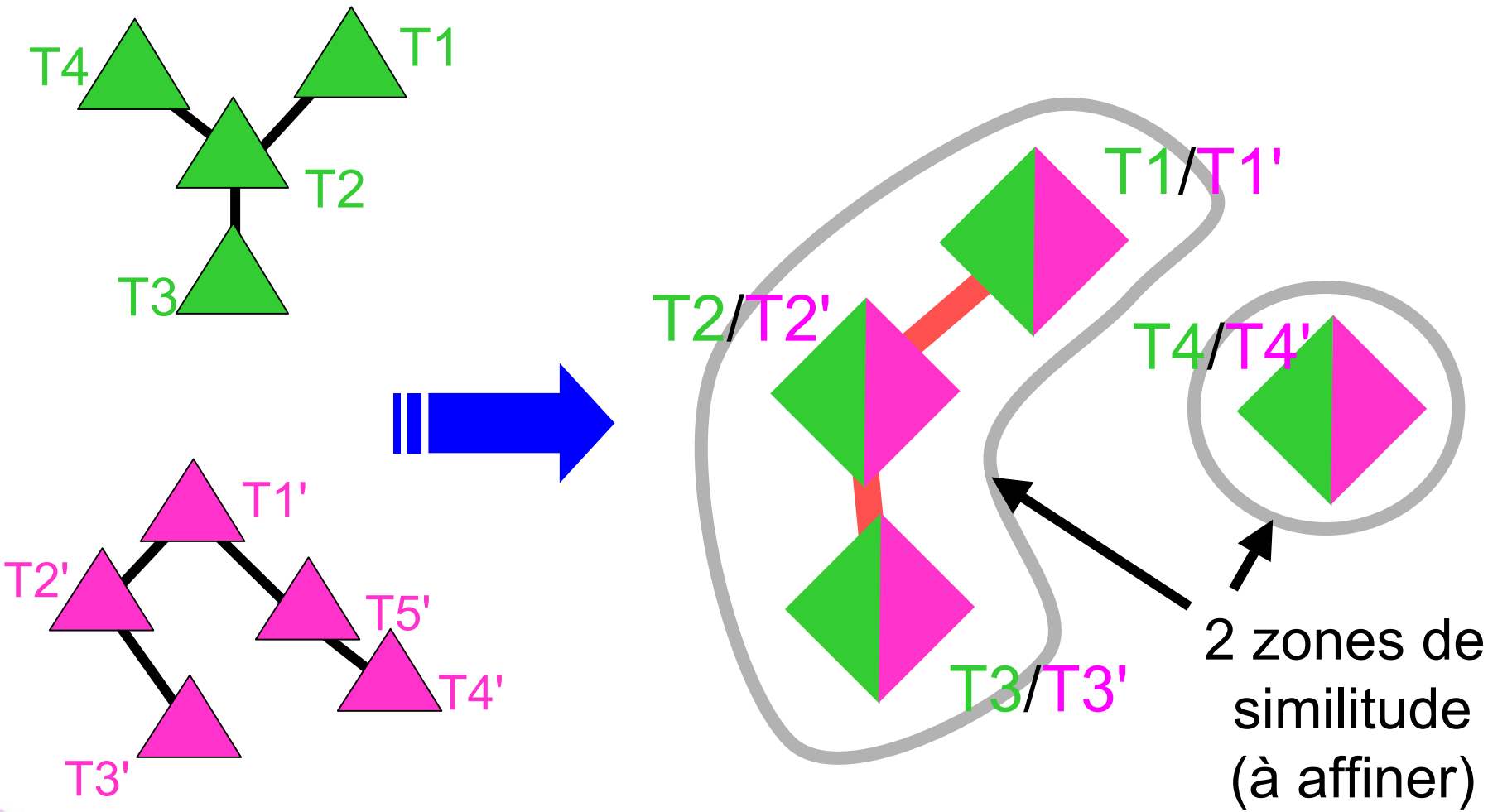
g.deleage@ibcp.fr

Représentation finale des molécules et stockage



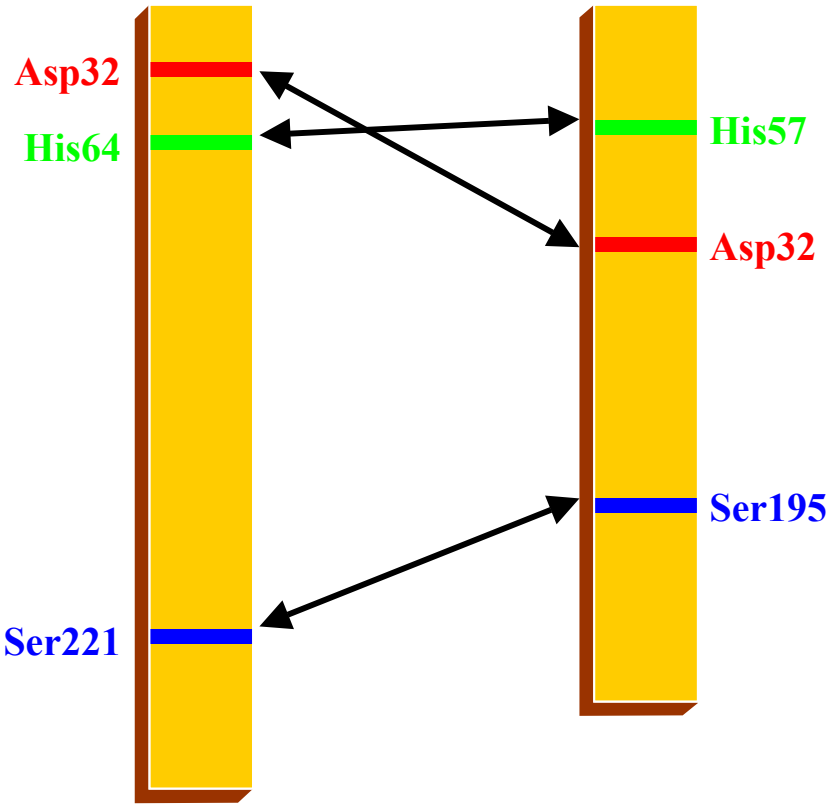
g.deleage@ibcp.fr

Principe de la comparaison entre 2 molécules



g.deleage@ibcp.fr

Exemple 1 : protéases à sérine



Données biochimiques

Comparaison
subtilisine 1SBC / chymotrypsine 1AFQ

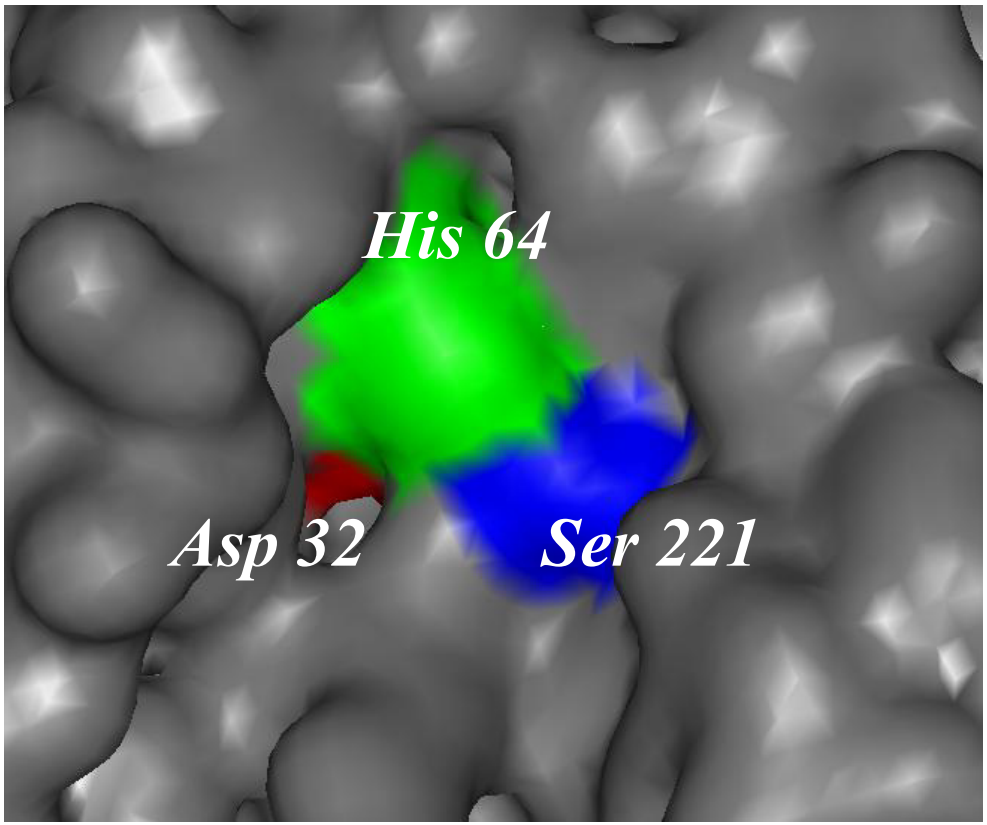
Résultat :

```
[Comparison result of (1AFQ,1SBC):
---------------------------- Patch number 1 --...
Number of groups: 4
Score = 1.116 [RMSD = 0.708] [penalty = 0.408]

Selected pairs of groups:
  ammonium B HIS   57 | 0.55 |   30.245    2.96...
  ammonium   HIS   64 | 0.54 |   30.407    2.70...

  aromatic B HIS   57 | 0.56 |   31.077    3.61...
  aromatic   HIS   64 | 0.56 |   31.309    3.27...

      acyl B ASP  102 | 0.67 |   32.459    7.49...
      acyl   ASP   32 | 0.66 |   31.730    7.17...

  hydroxyl C SER  195 | 0.59 |   27.628    2.51...
  hydroxyl   SER  221 | 0.57 |   27.964    3.42...

]
```

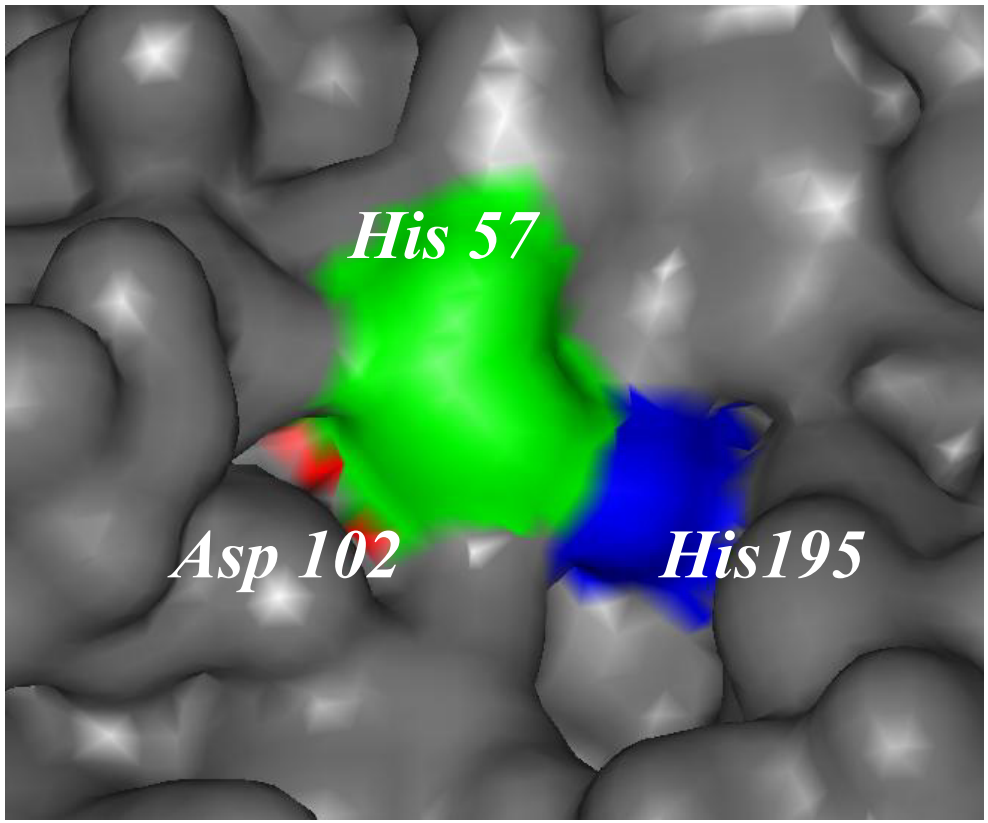
g.deleage@ibcp.fr

Vue de surface des 2 sites catalytiques de protéases

Subtilisine (1SBC)



Chymotrypsine (1AFQ)



g.deleage@ibcp.fr

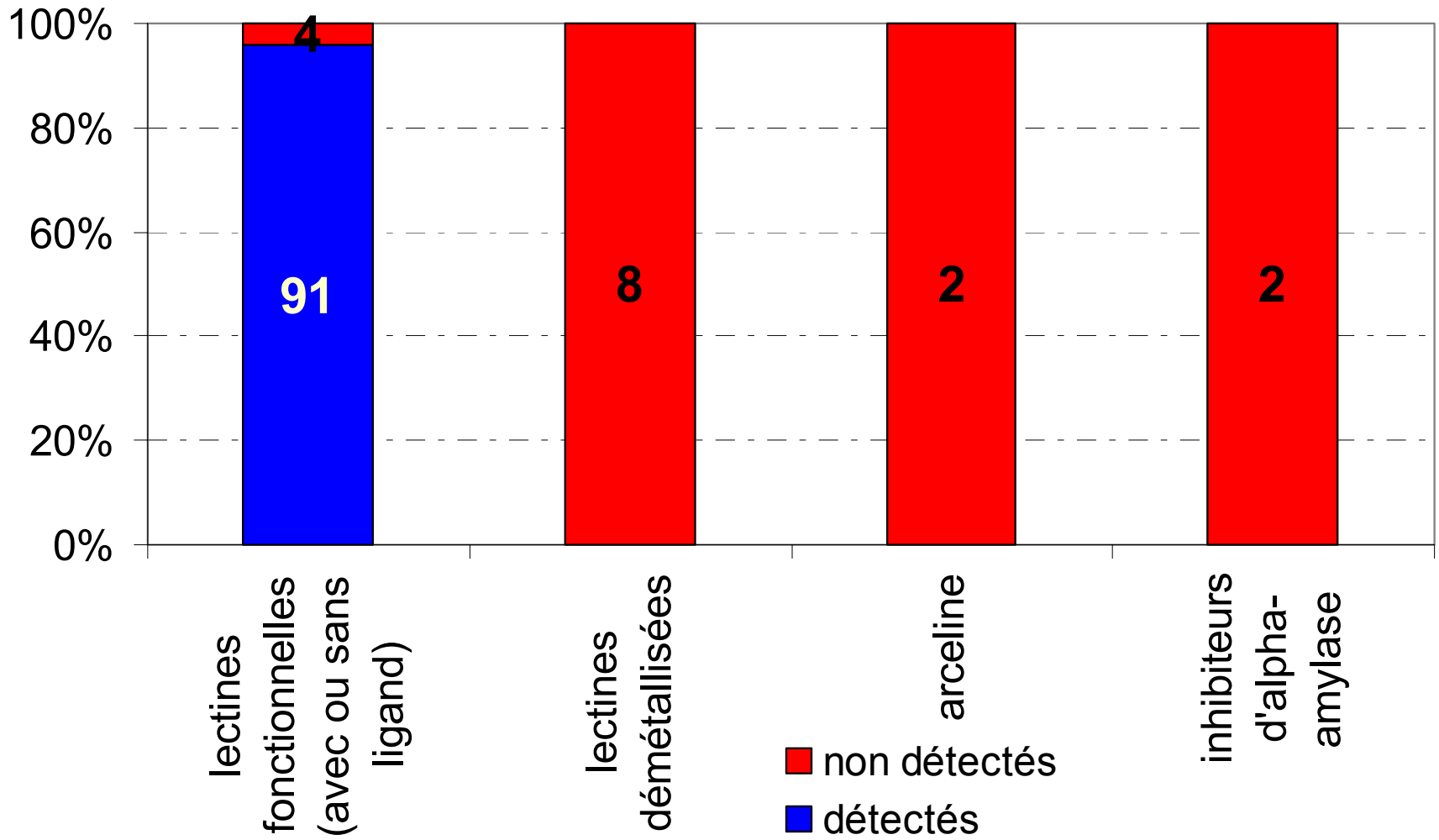
Exemple 2 : lectines de légumineuses

Collaboration avec Anne Imberty, CERMAV (Grenoble)

- Travail sur la famille des lectines de légumineuses
- Séquences voisines
- Même repliement

g.deleage@ibcp.fr

Recherche de site à sucre dans la famille des lectines de légumineuses



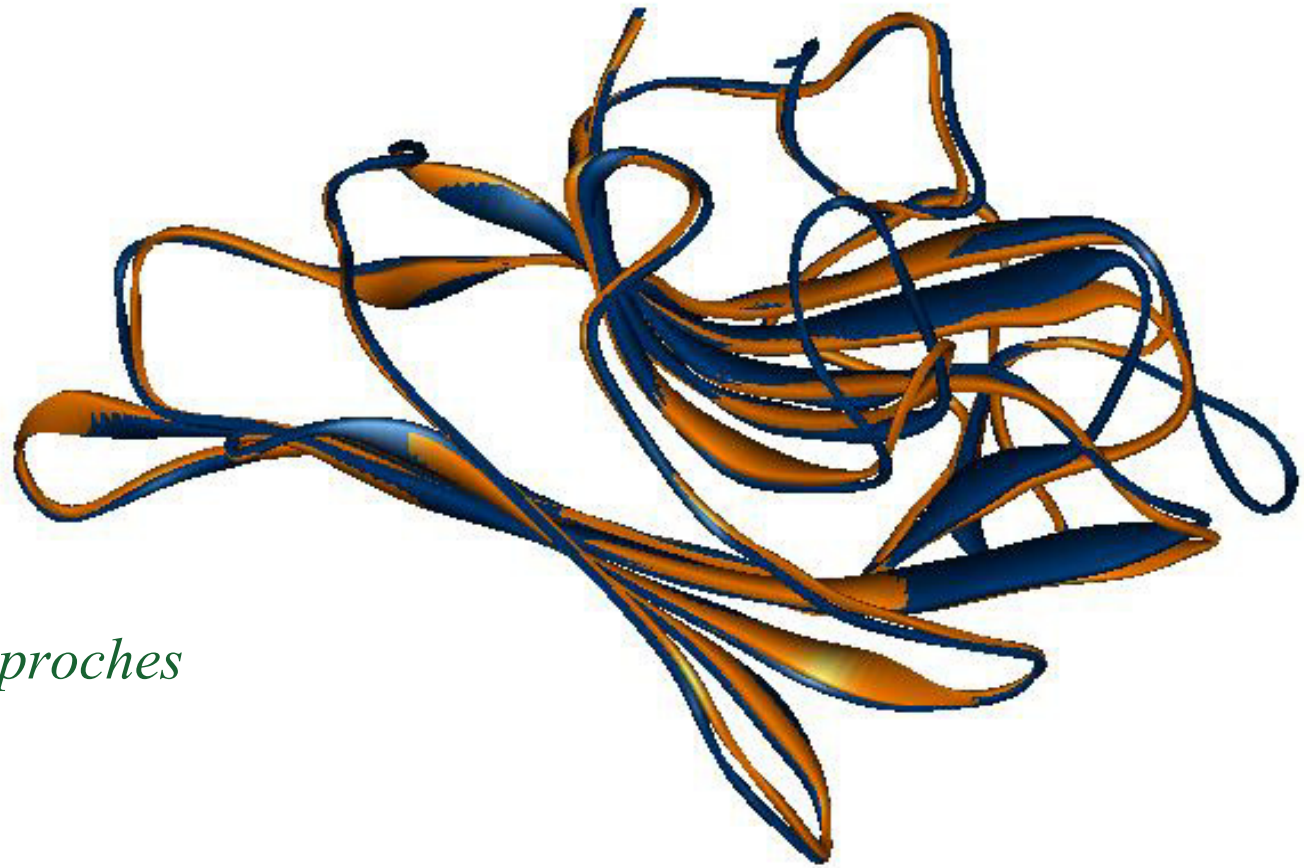
Site recherché : lectine de cacahuète (2PEL)
Nombre de protéines analysées = 107
Pourcentage de réussite global = 96 %

g.deleage@ibcp.fr

Illustration : 1DQ1 et 1DQ2 : Concanavaline A

Orange : concanavaline A fonctionnelle

Bleu : concanavaline A non fonctionnelle (démétallisée)

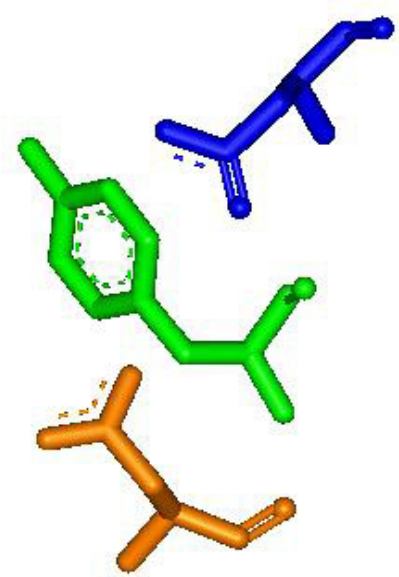


- *RMSD = 0,9 Å*
- *séquences très proches*

g.deleage@ibcp.fr

Vue des sites 1DQ1 et 1DQ2 : concanavaline A native et démétallisée

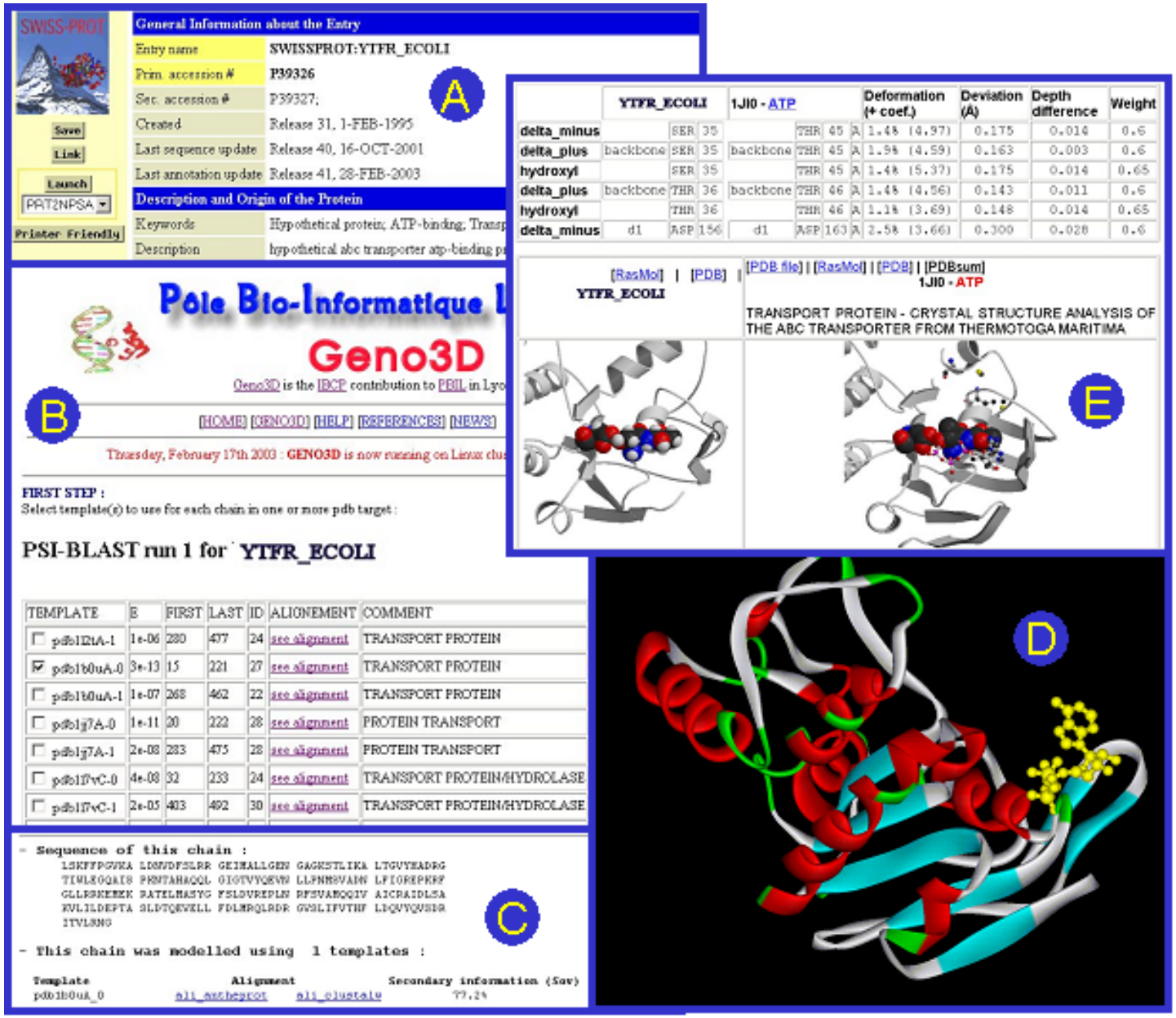
Site de fixation du sucre (mannose)



Superposition : Site fonctionnel et site détruit en absence de calcium et de zinc

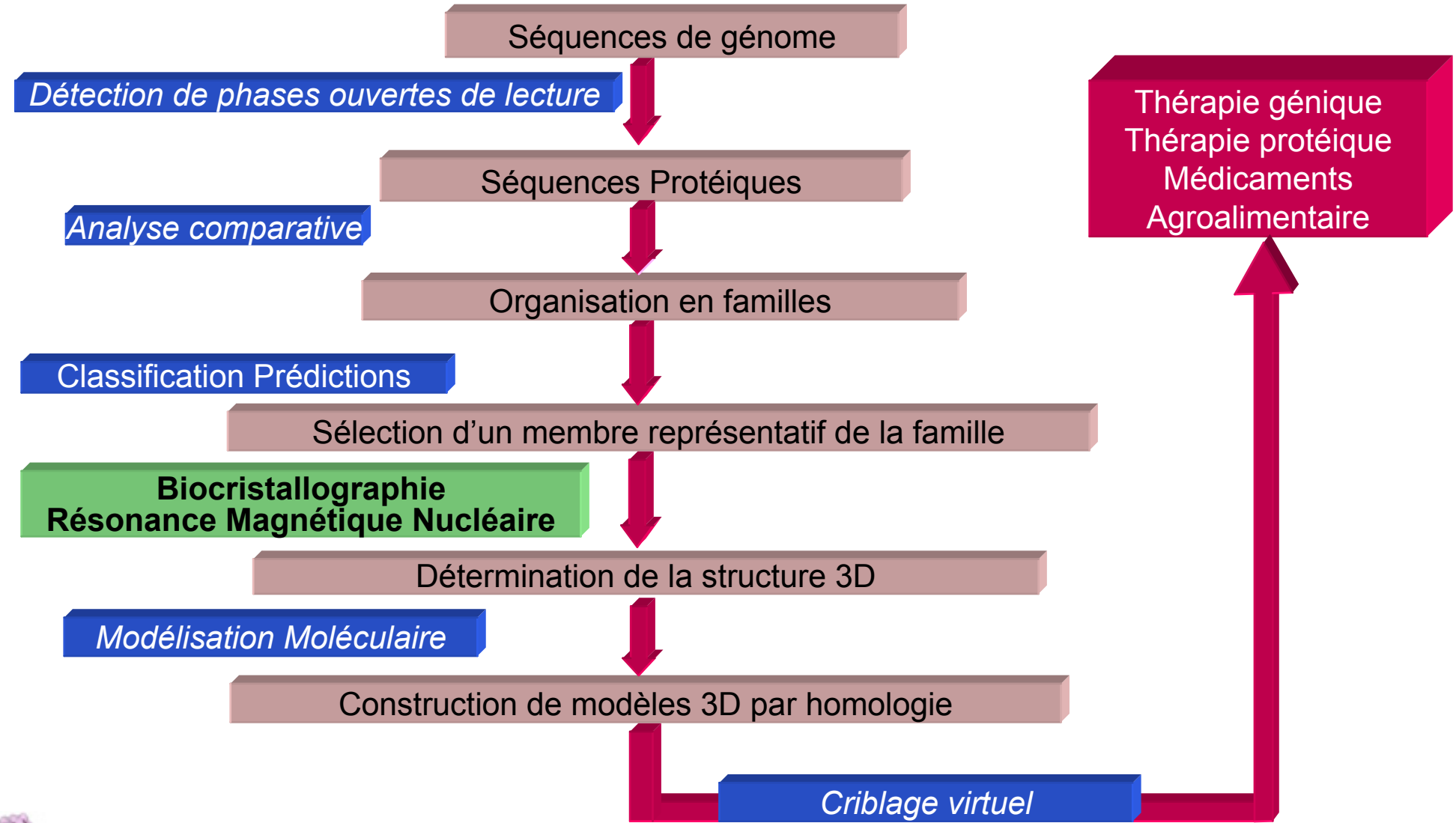
g.deleage@ibcp.fr

Annotation structurale



g.deleage@ibcp.fr

Etude structure fonction des protéines.
Génomique structurale



g.deleage@ibcp.fr