Gilbert Deléage

Professeur de bioinformatique

Université Claude Bernard Lyon 1

IBCP

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: gilbert.deleage@ibcp.fr

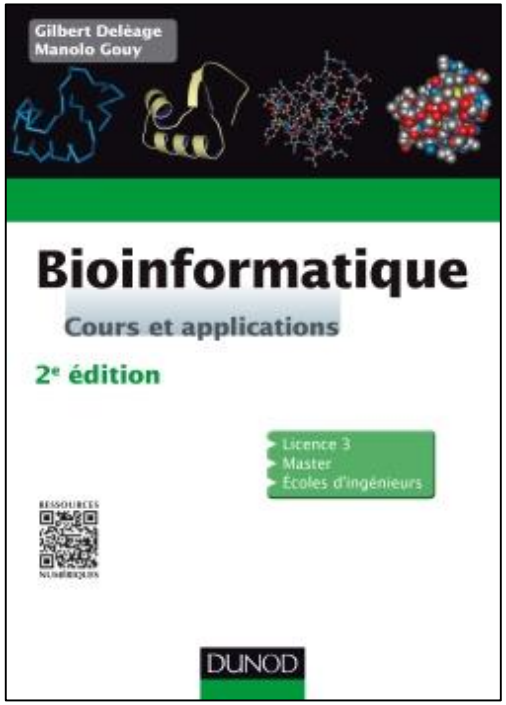
http://www.gdeleage.fr/prof/IBIS.pdf

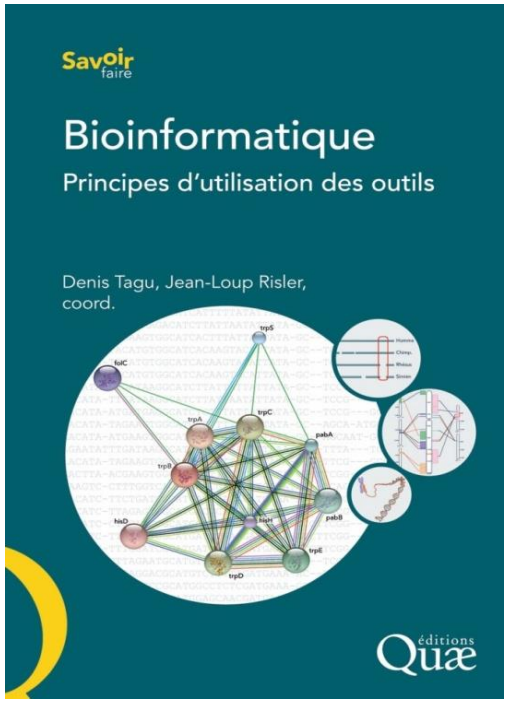
Objectif: Extraire de l'information des séquences

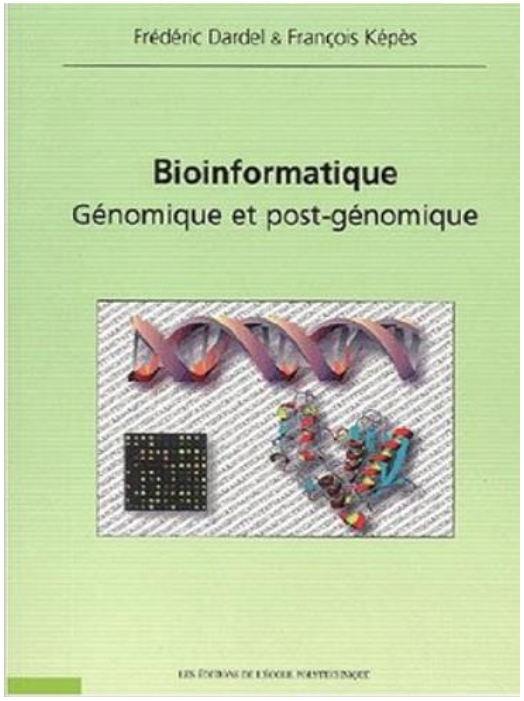
- **Introduction - contexte**
- **Banques de données**
  - Présentation - Biais
  - Interrogations
  - Formats – Pièges
- **Analyse de séquences**
- **Comparaison de séquences**
  - Matrices de points
  - Homologie-similarité
  - Recherche dans les banques
    - Fasta
    - Blast
- **Alignements binaires et multiples**
  - Clustal (Higgins), 2 versions
  - Multalin, Muscle
  - Align (Feng & Doolittle)
  - Profil d'identité
- **Fonctions, site, signature, motif**
  - PROSITE - Algo Unif / Algo non Unif
  - Profils de séquences – Blocks
  - Détection de sites (3D)
- **Profils physico-chimiques**
  - Hydrophobie
  - Amphiphilie
  - Accessibilité au solvant
  - Flexibilité
  - Antigenicité
- **Prédictions de structure secondaire**
  - Méthodes statistiques
  - Méthodes de similarité
  - Réseaux de neurones
- **Structures 3D**
  - Rasmol

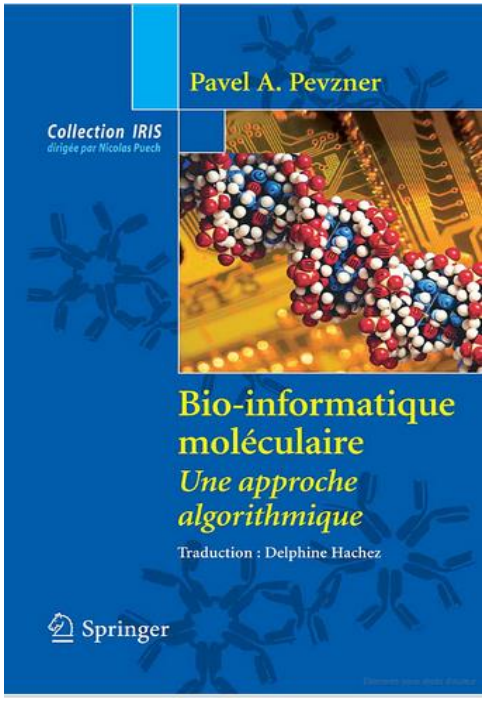
g.deleage@ibcp.fr

Quelques ouvrages de bioinformatique



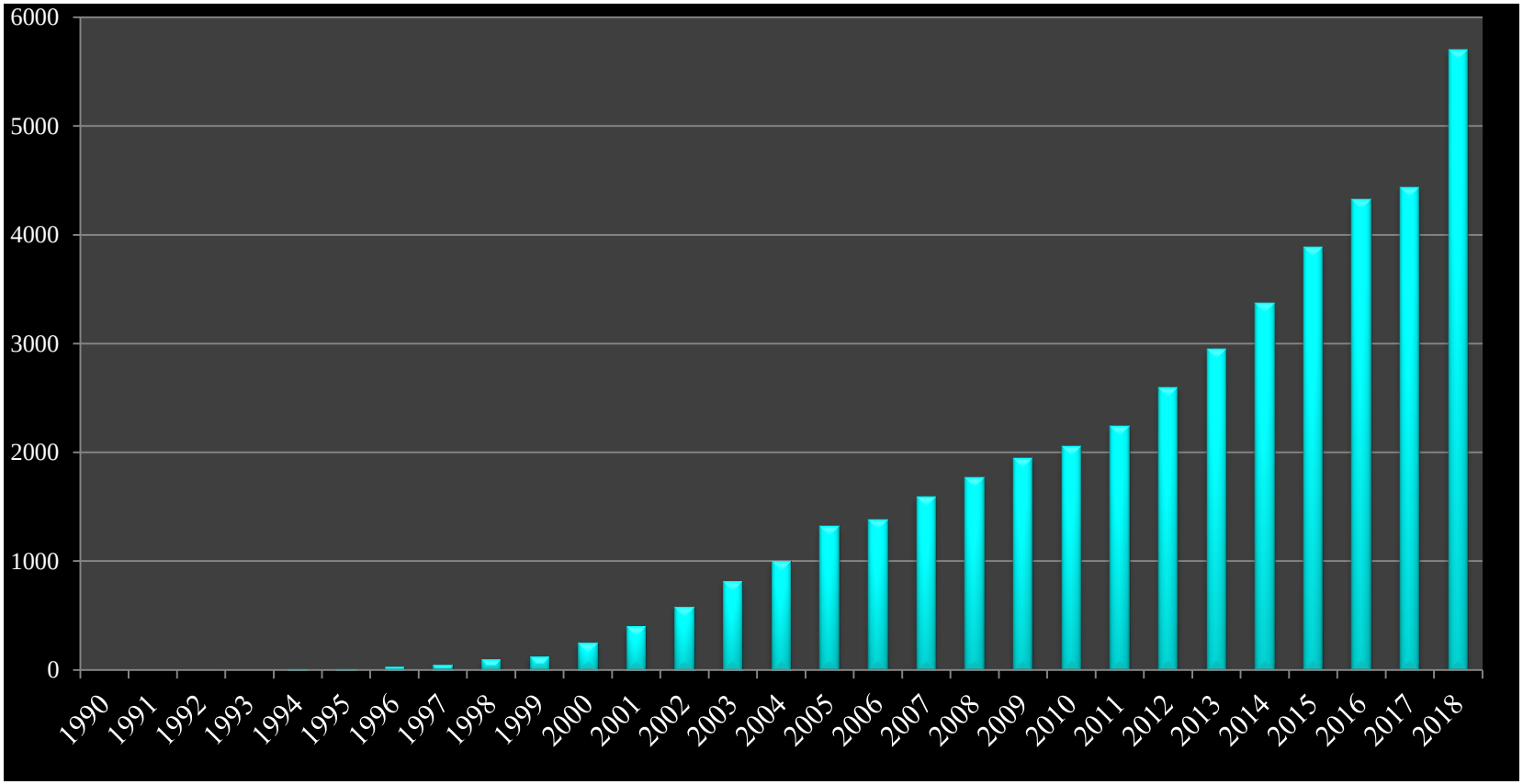






g.deleage@ibcp.fr

Nombre d'articles scientifiques référençant « bioinformatics » comme mot clef.



Articles grand public:

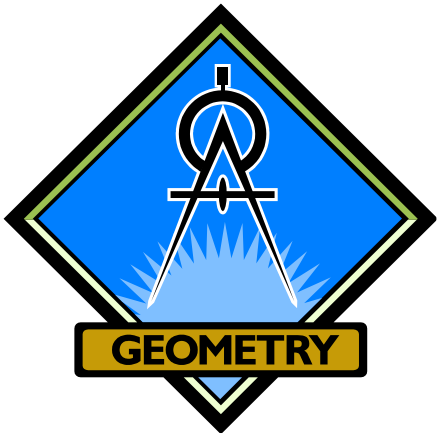
- Bioinformatique et Génome. *Biofutur* 24/251 (2005).
- Bioinformatique et Post-Génome. *Biofutur* 24/252 (2005).

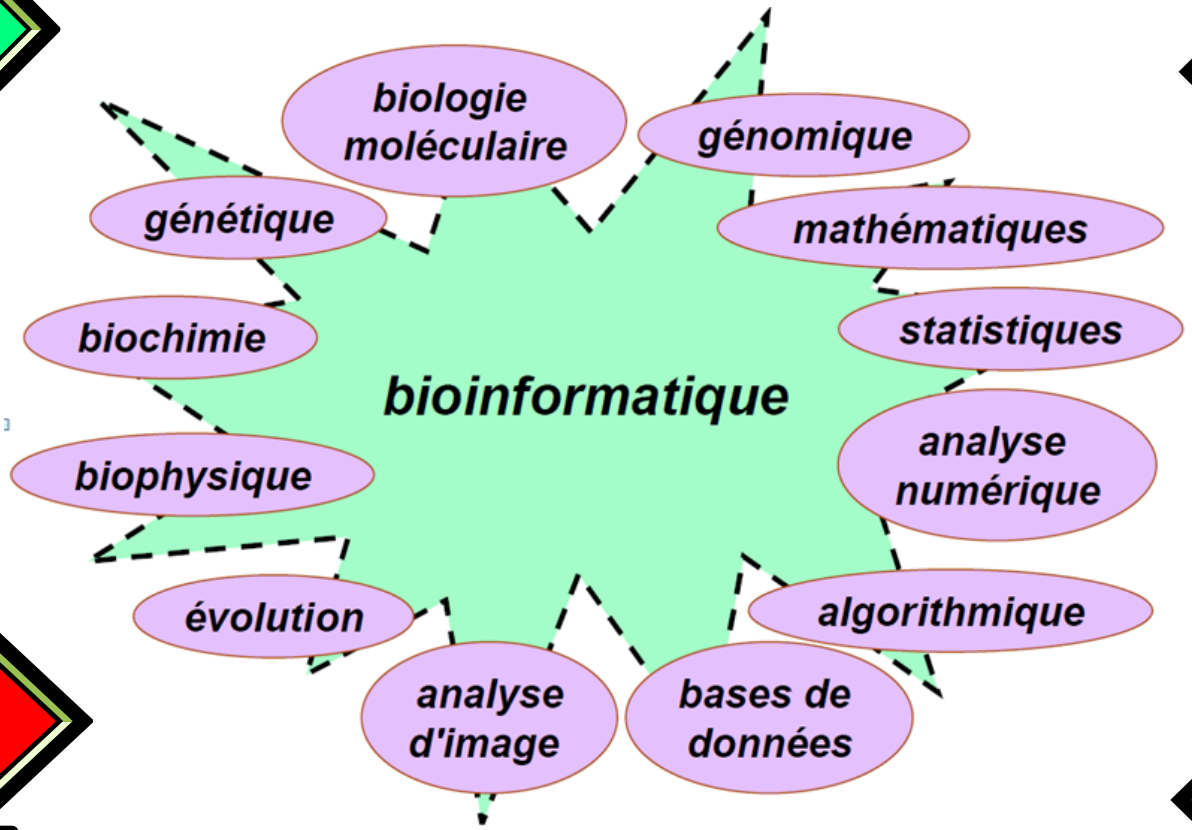
Livres de référence:

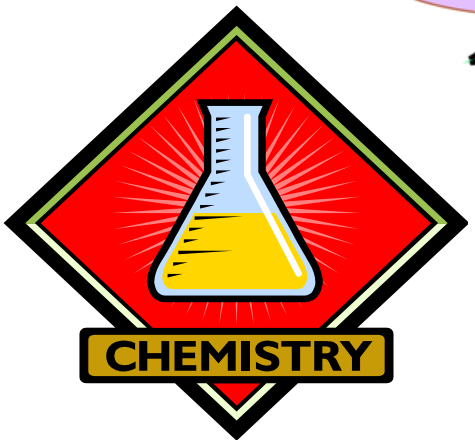
- Hancock JM & Zvelebil MJ (eds) (2004). *Dictionary of Bioinformatics and Computational Biology*. New York: John Wiley & Sons.
- Mount DW (2004). *Bioinformatics: Sequence and Genome Analysis*. NY: Cold Spring Harbor Laboratory Press.

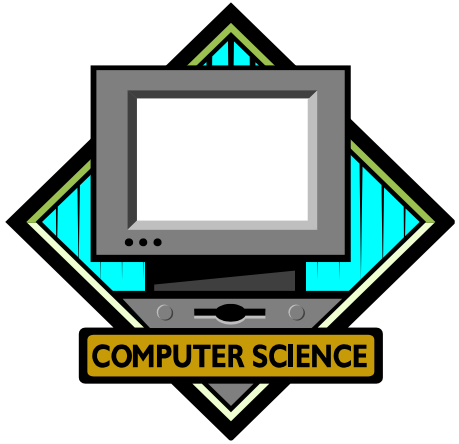
g.deleage@ibcp.fr

BIOINFORMATIQUE : Interdiscipline



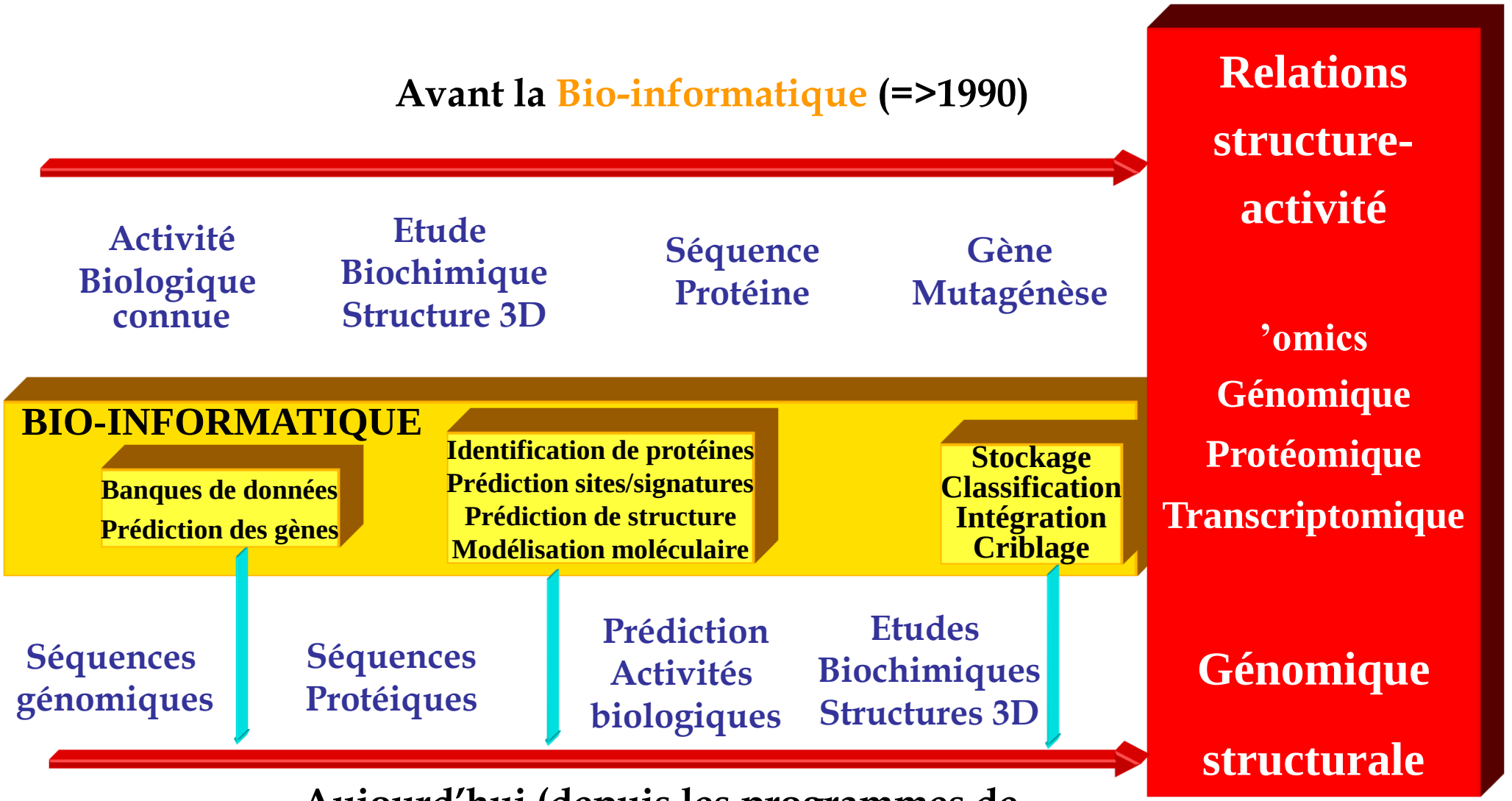






g.deleage@ibcp.fr

La révolution génomique - Une stratégie revisitée

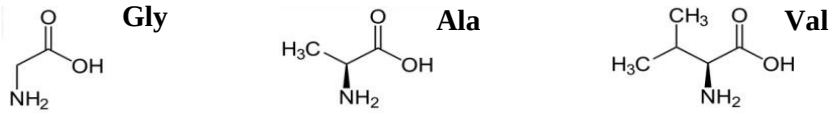


g.deleage@ibcp.fr

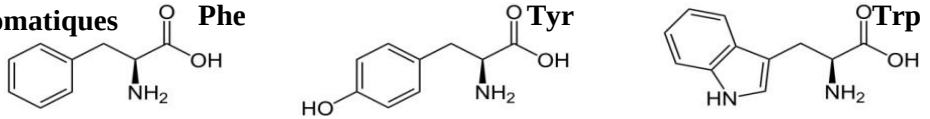
Les « briques de base » des protéines

A	Alanine	Ala
C	Cysteine	Cys
D	Aspartic Acid	Asp
E	Glutamic Acid	Glu
F	Phenylalanine	Phe
G	Glycine	Gly
H	Histidine	His
I	Isoleucine	Ile
K	Lysine	Lys
L	Leucine	Leu
M	Methionine	Met
N	Asparagine	Asn
O	**Pyrrolysine**	**Pyl**
P	Proline	Pro
Q	Glutamine	Gln
R	Arginine	Arg
S	Serine	Ser
T	Threonine	Thr
U	**Sélénocystéine**	**Sec**
V	Valine	Val
W	Tryptophane	Trp
Y	Tyrosine	Tyr
B		Asn/Asp
Z		Gln/Glu
X	Inconnu	

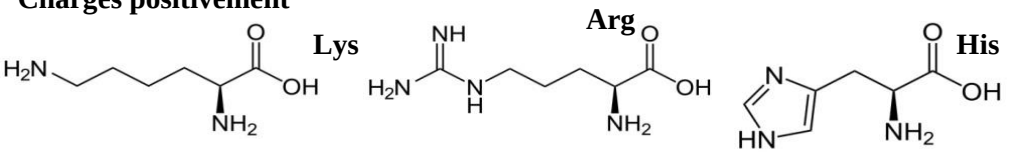
Non polaires



Aromatiques

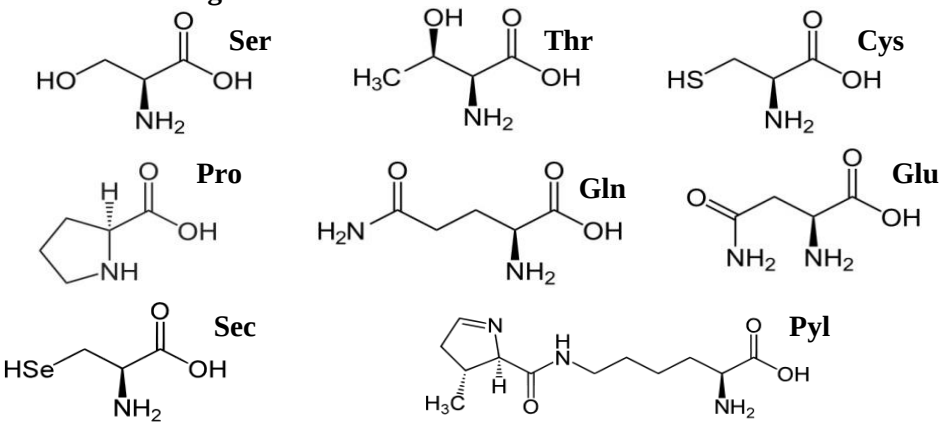


Chargés positivement



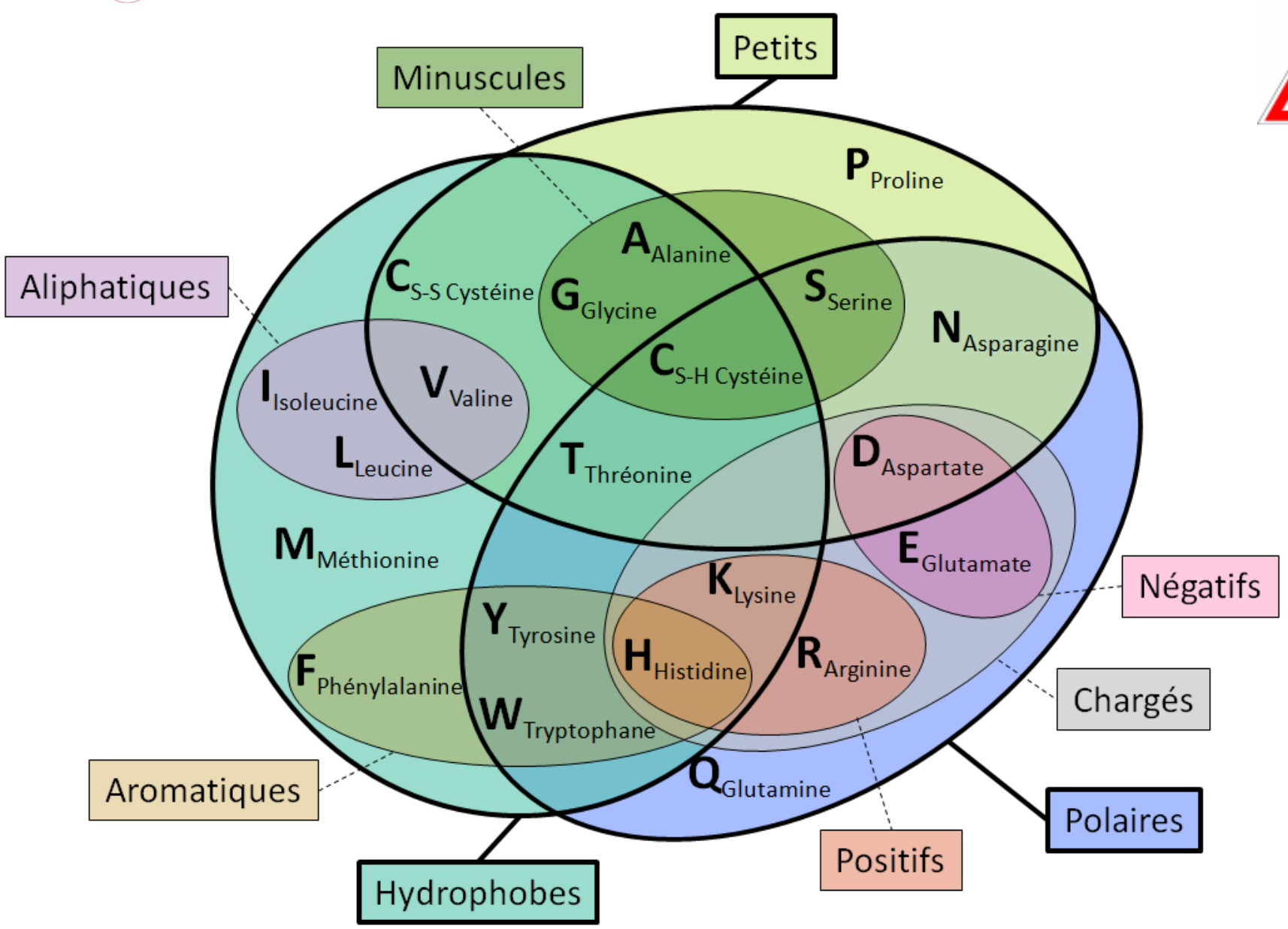
Chargés négativement

Polaires non chargés



g.deleage@ibcp.fr

Diagramme de Venn des acides aminés

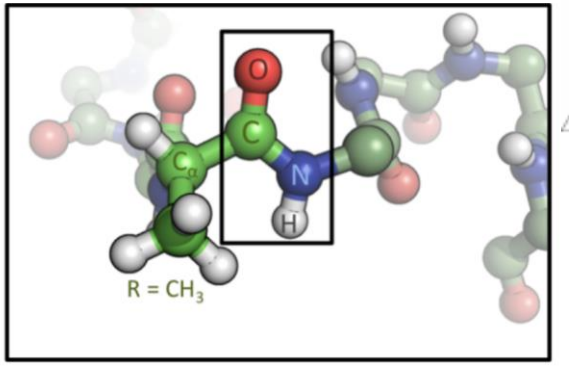


g.deleage@ibcp.fr

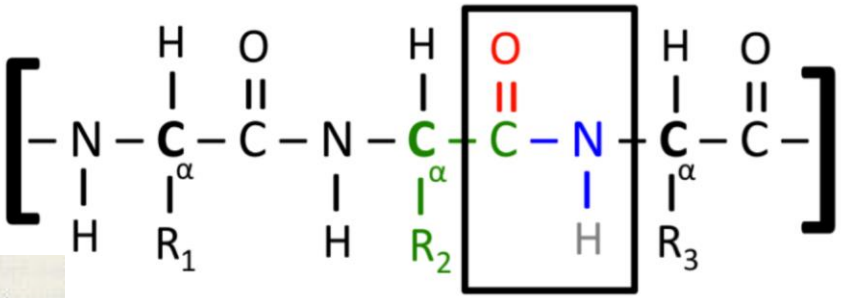
Les « briques de base » des protéines

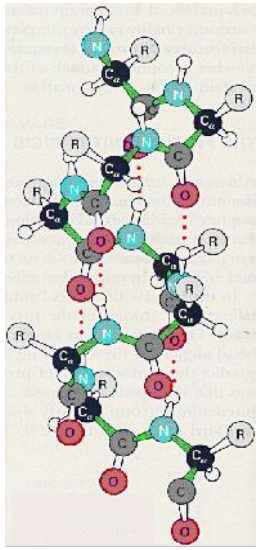
- Structure primaire ou séquence : la séquence des acides aminés dans la protéine.

```
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIVL
GFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLH
GYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGE
NHAIMGVAFTWVMALACAAPPLVGWSRYIPQGMQCSCGALYFTLKPEINN
```

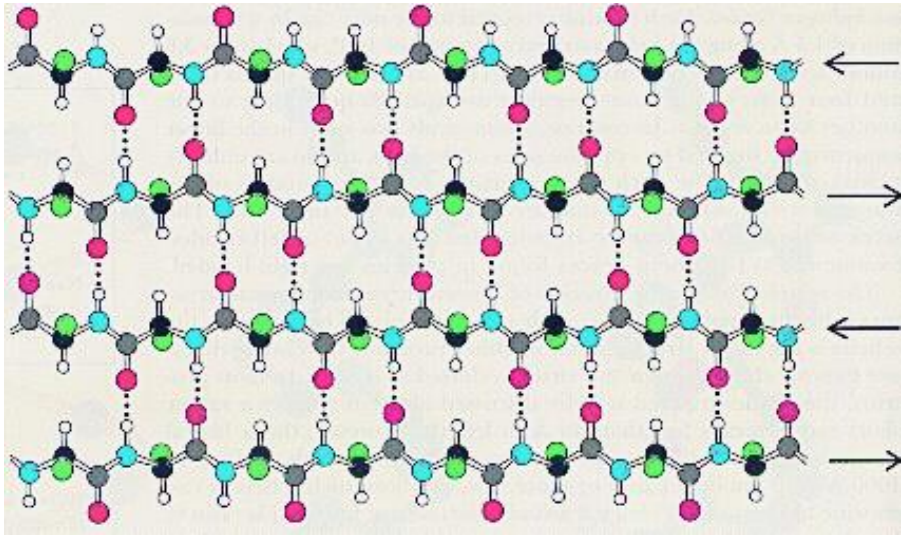


- Structure secondaire : ensemble des régions présentant un arrangement régulier (périodique) stabilisé par des liaisons hydrogènes impliquant les atomes de la chaîne principale (backbone)
  - *i.e.* hélice α, brins β, feuillets β et coudes β.

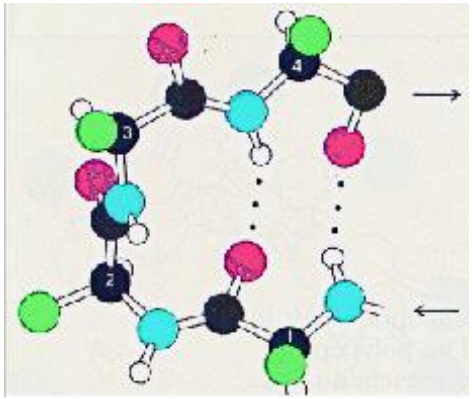




30%



20%



25%

g.deleage@ibcp.fr

La combinatoire des séquences protéiques

$$N = 20^x$$

On peut calculer :

$3{,}2\ 10^6$ séquences de 5 aa

10^{13} de protéines différentes de 10 acides aminés chacune,

$3\ 10^{19}$ de 15 acides aminés,

$3{,}3\ 10^{32}$ 25 acides aminés.

Pour 65 acides aminés > nombre d'atomes dans l'univers

et pour 130 aa (cas du lyzozyme)? 20^{130}

Et pour une protéine « moyenne »? 20^{500}

Et pour le protéome humain? Pour 20000 protéines codées $20^{10,000,000}$

Nombre d'atomes dans l'univers ~ 10^{80}

d'après Christian Magnan Collège de France, Paris
Université de Montpellier II

g.deleage@ibcp.fr

Combinaisons de ponts disulfures dans les protéines

Pour une protéine qui contient n Cys formant k ponts SS

($k \leq n/2$) le nombre N de possibilités de tous les appariements possibles est donné par :

$$N = \binom{n}{2k} \frac{(2k)!}{2^k\ k!}$$

Pour 10 Cys et 3 ponts SS :

$$N = \binom{10}{6} \frac{6!}{2^3\ 3!} = 210 \text{ x } 15 = 3150$$

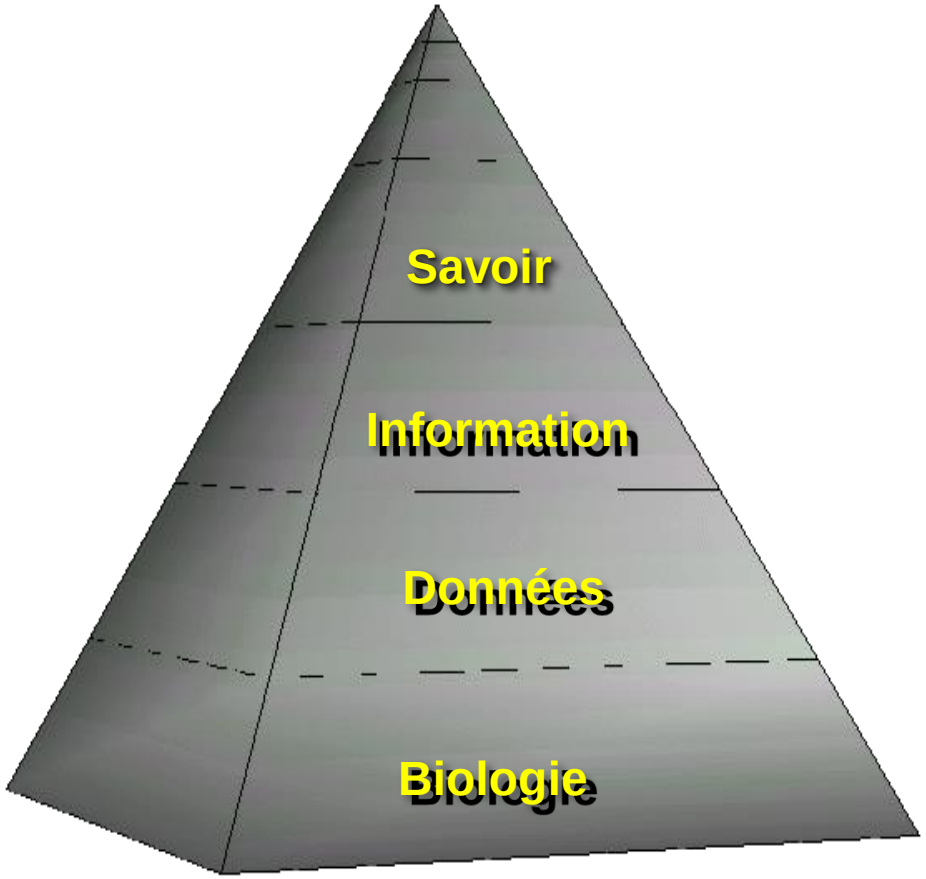
Simulateur

Pour 90 Cys il y a $1.9\ 10^{72}$ combinaisons possibles d'association des ponts SS

g.deleage@ibcp.fr

Biologie à grande échelle

Valeur ajoutée

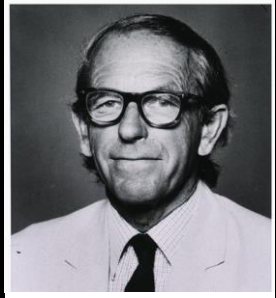


Volume

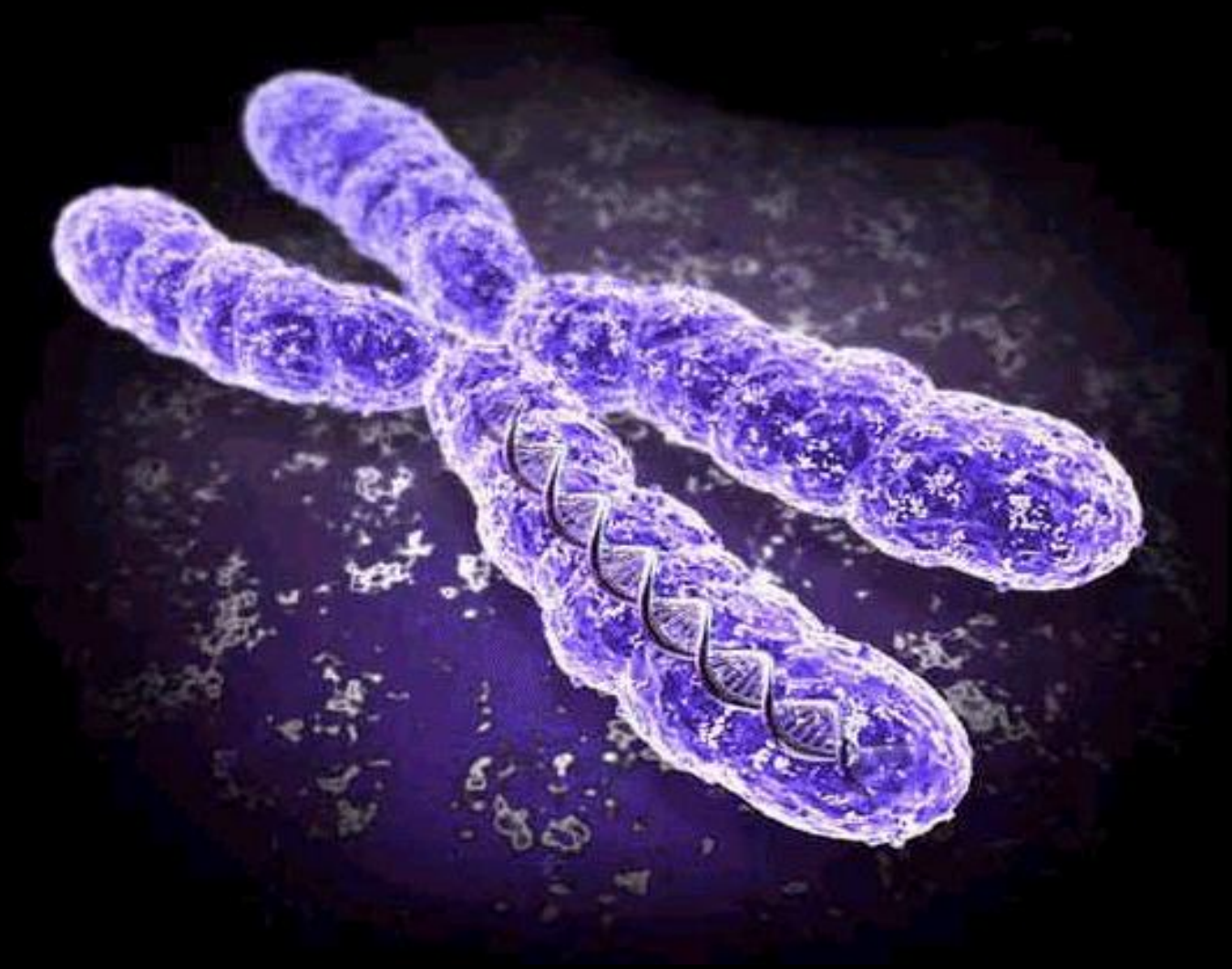
g.deleage@ibcp.fr

Défis du séquençage

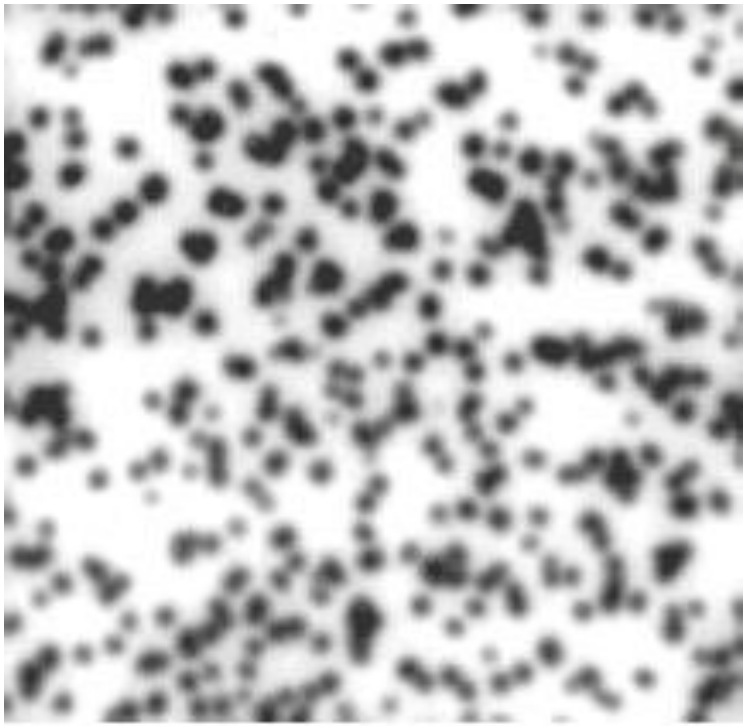
Frederick Sanger



1918-2013



SOLEXA illumina



Fluorescence *in situ*

Séquençage massif et parallèle: Chaque tache correspond à 1 ADN attaché sur un support

g.deleage@ibcp.fr

Défis du séquençage

	454	Solexa	SOLID
			
Chemistry:	**Pyrosequencing**	**Fluorescent InSitu**	**Ligation**
Parallelisation:	**400,000**	**30 million**	**50 million**
Read length:	**~ 400 nt**	**~ 50 nt**	**35 nt**
Sequence:	**~ 500 Mb**	**~ 4 Gb**	**~ 20 Gb**
Run time:	**7.5 hours**	**6 days**	**10 days**

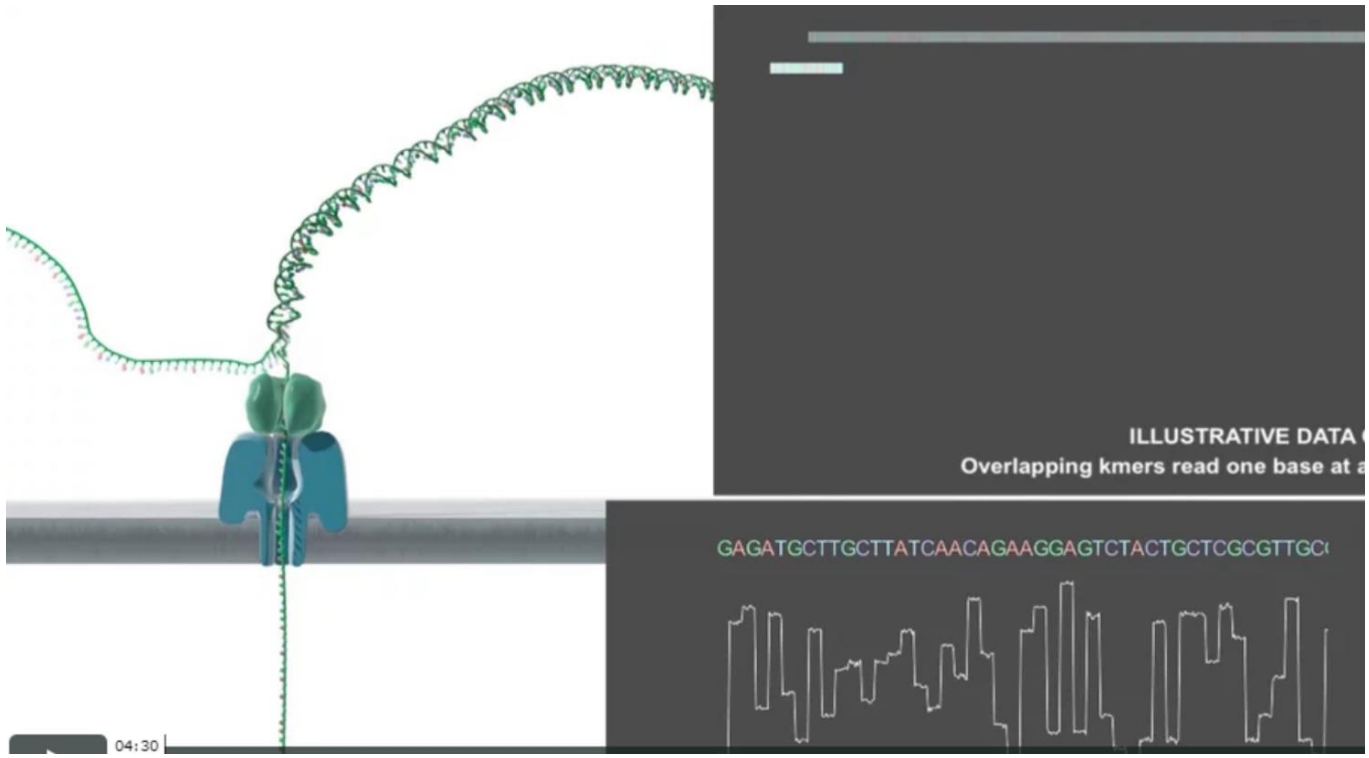
g.deleage@ibcp.fr

Système nanopore GridION (Oxford Nanopore Technologies)

https://nanoporetech.com/products/gridion

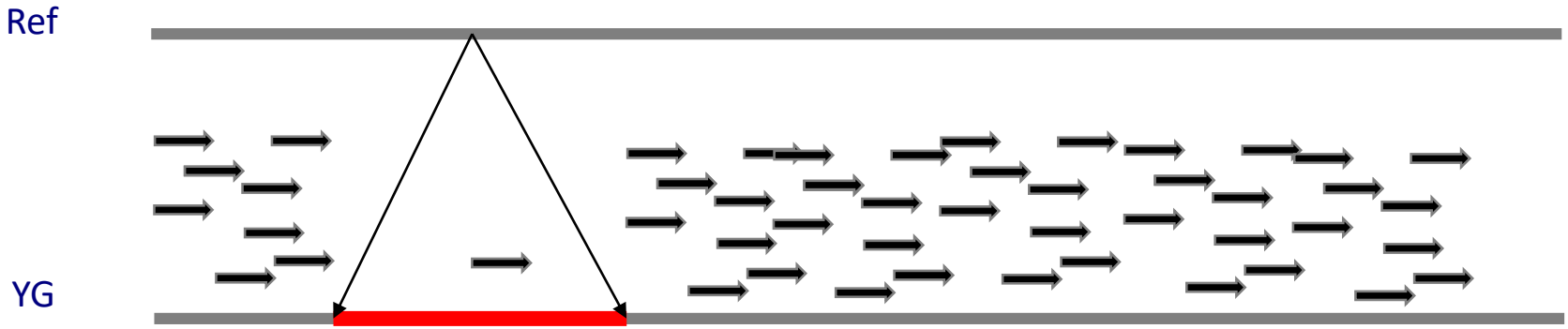
•les nanopores utilisés sont à base de protéines complexes déposées sur une membrane synthétique en polymère. Le tout est posé sur un chipset produit en **ASIC** (circuit intégré conçu pour une seule application). L'ensemble est dans une cartouche de consommable.

•un ADN double-brin est accroché à une enzyme qui va s'amarrer au nanopore et déclencher le passage de l'un des brins de l'ADN dans celui-ci puis le second brin. Les bases qui traversent le nanopore (« filer » un brin de la double hélice à travers le pore) sont détectées électriquement comme avec le GeniaChip, mais semble-t-il, par blocs de 4 ce qui réduit les erreurs. Quand le premier brin a été analysé, l'autre brin l'est également et dans l'autre sens, ce qui créé une information redondante limitant les erreurs de séquençage.



g.deleage@ibcp.fr

Single Nucleotide Polymorphism : 1/1000 positions
Détection automatique de variants structuraux (défaut de couverture) => longueur de l'insertion



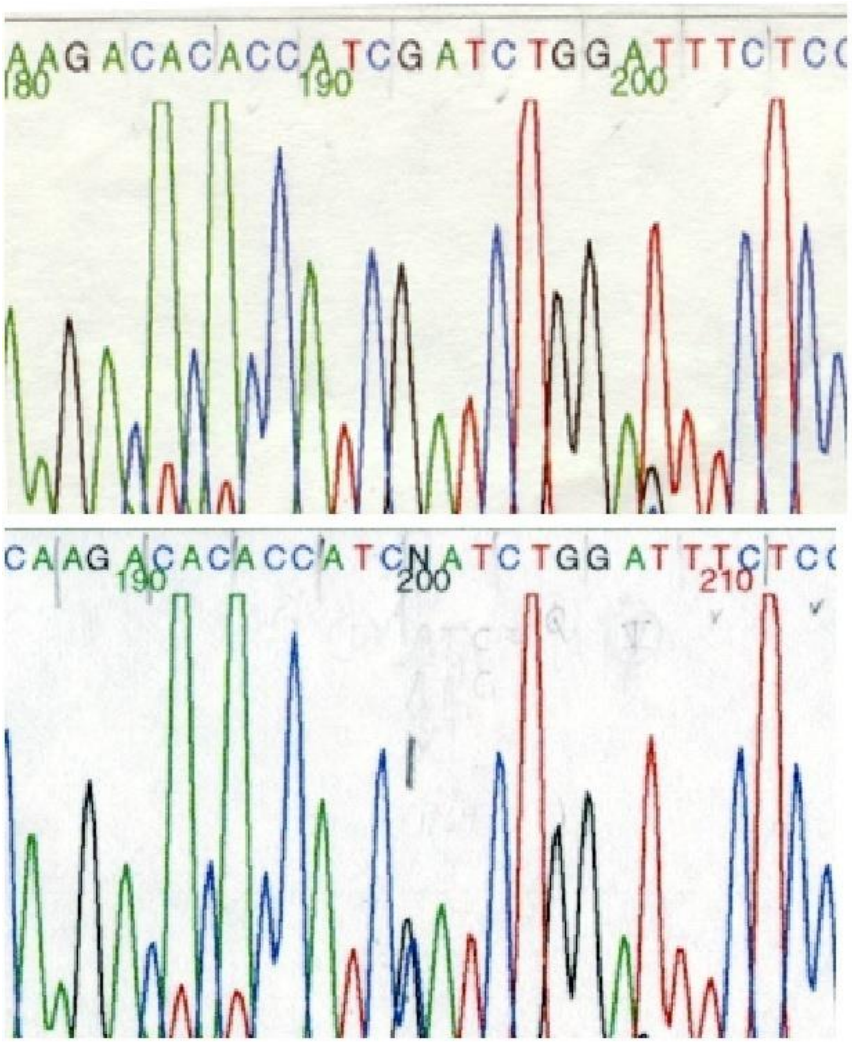
Typiquement : 40 x couverture 35 bp reads
2308 deletions, 2076 insertions

2002 : Assemblage du génome humain coût en M$ (3 années)
2010: 1000 génomes humains chacun d'un cout de 5-10k$
(capacité de séquençage de 30000x en 6 ans)

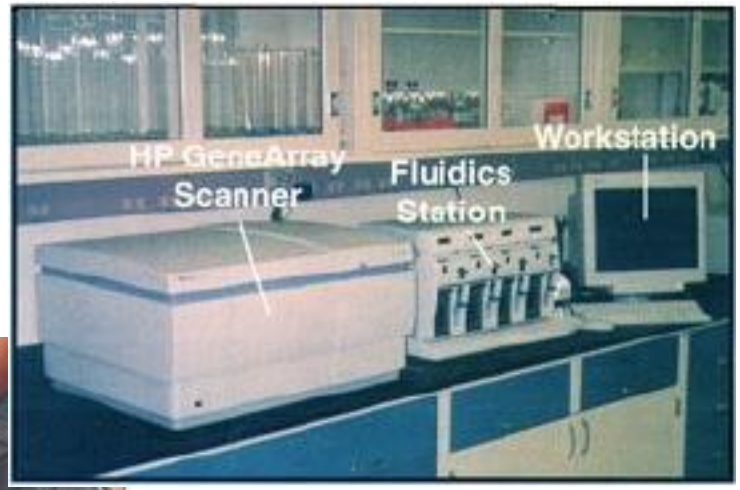
Futur: Séquençage de cohortes
P4 Médecine : Prédictive, Préventive, Personnalisée, Participative

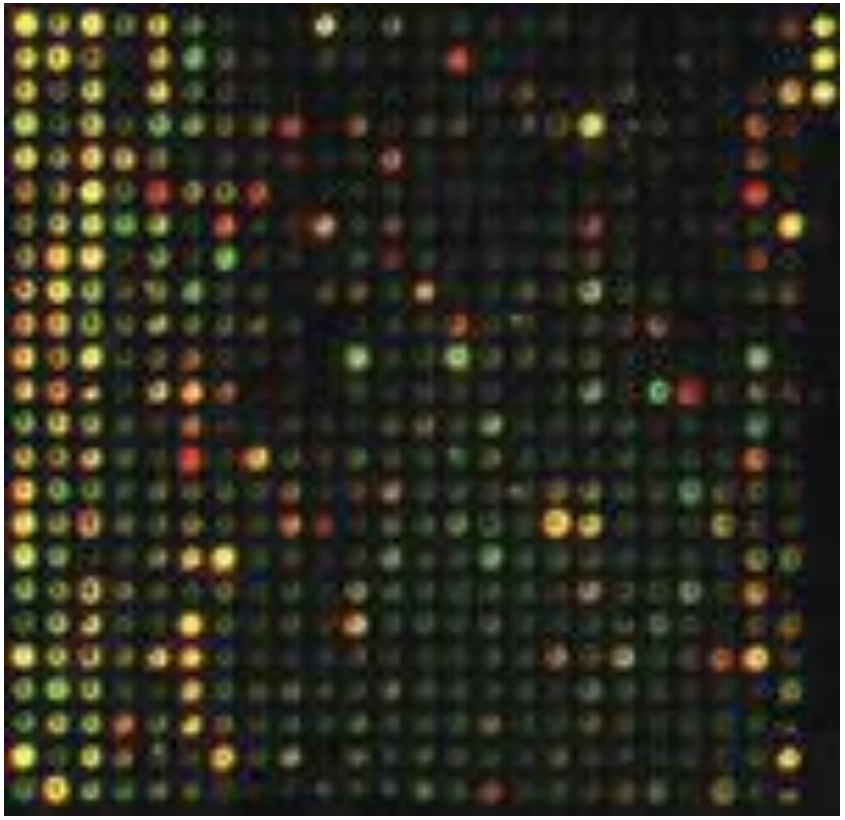
g.deleage@ibcp.fr


HITACHI
ABI PRISM
3100 Genetic Analyzer

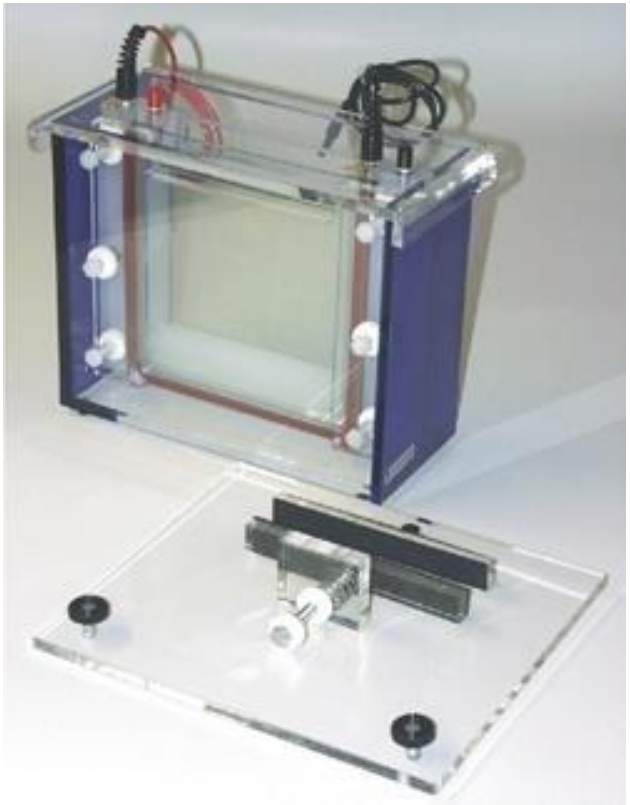

AAGACACACCATCGATCTGGATTTCTCC
180
190
200
CAAGACACACCATCNATCTGGATTTCTCC
190
200
210

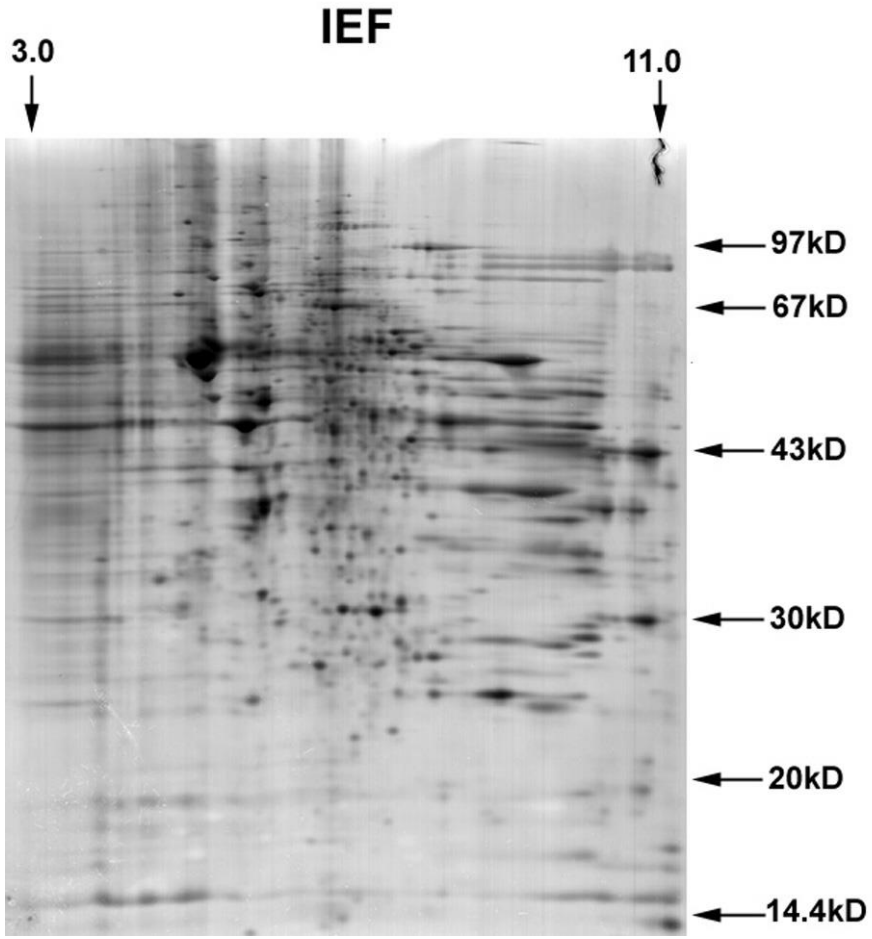
OMICS 3 / 4 Trancriptomique





g.deleage@ibcp.fr





g.deleage@ibcp.fr

Biologie à grande échelle (ordres de grandeur)

- **Hétérogénéité et nombre >2000 BD biologiques**
  - ✓ Séquences nucléiques 200 000 000
  - ✓ Séquences protéiques 600,000 annotées
  - ✓ Structures 3D 150,000
  - ✓ Structures 3D différentes (<25% Id) 20000
  - ✓ Génomes complets 3000
- **Qualité variable**
  - ✓ Erreurs
  - ✓ Propagées par l'annotation automatique
- **Biais**
  - ✓ 20 espèces = 21% des entrées de SWISS-PROT
  - ✓ Redondance
  - ✓ Génome humain **http://www.ensembl.org**
- **Volume « faible » croissance exponentielle**
  - ✓ Formats, traitements
  - ✓ Le génome du jour…

  http://www.genomesonline.org/

g.deleage@ibcp.fr

- Recherche dans les banques de données biologiques

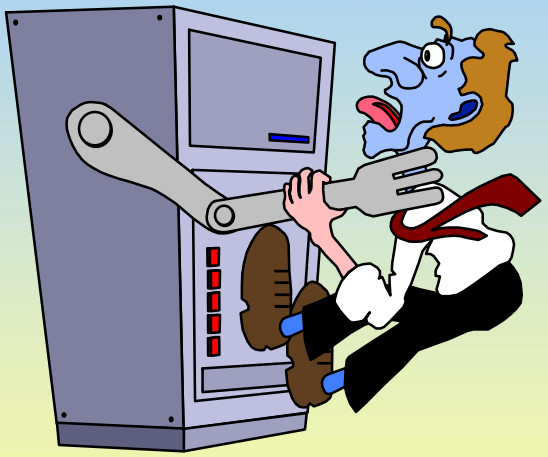
  Tache courante d'un biologiste:
  - Est-ce qu'une nouvelle séquence a déjà été complètement ou partiellement répertoriée?
  - Est-ce que cette séquence contient un gène?
  - Est-ce que ce gène appartient à une famille connue?
  - Existe-t-il d'autres gènes homologues?
- Alignement de séquences: Est-ce que deux séquences correspondent à deux gènes homologues?
- Recherche de sous-motifs communs à un ensemble de séquences (ADN, ARN). Établissement de consensus, alignement multiple
- Recherche de régions contenant des séquences répétées (en tandem ou transposées)
- Recherche d'hélices ou de brins dans les protéines
- Recherche de gènes (régions promotrices, facteurs de transcription…)
- Comment construire un modèle 3D de ma protéine.

g.deleage@ibcp.fr

Vision «immergée» de la BioInformatique

Biochimie Biologie Moléculaire

Applications et Expérimentation Biologiques



Logiciels et Serveurs
ANTHEPROT, MPSA, biolcp, NPS@ MPSAWeb, ANTHEWEB

Banques de données
Swiss-Prot
SP-Trembl
PDB, HCVDB

Méthodologies
SOPM, SOPMA, MLRC
ProScan, PattInProt,
Sumo, PROCSS, geno3D,

Informatique

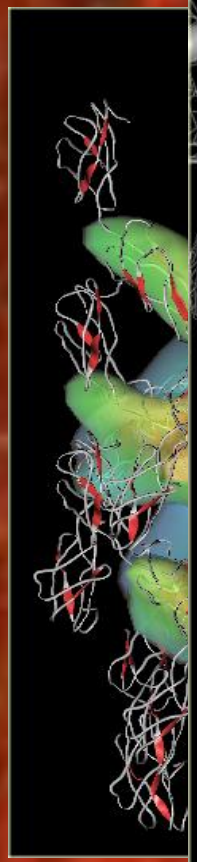
g.deleage@ibcp.fr

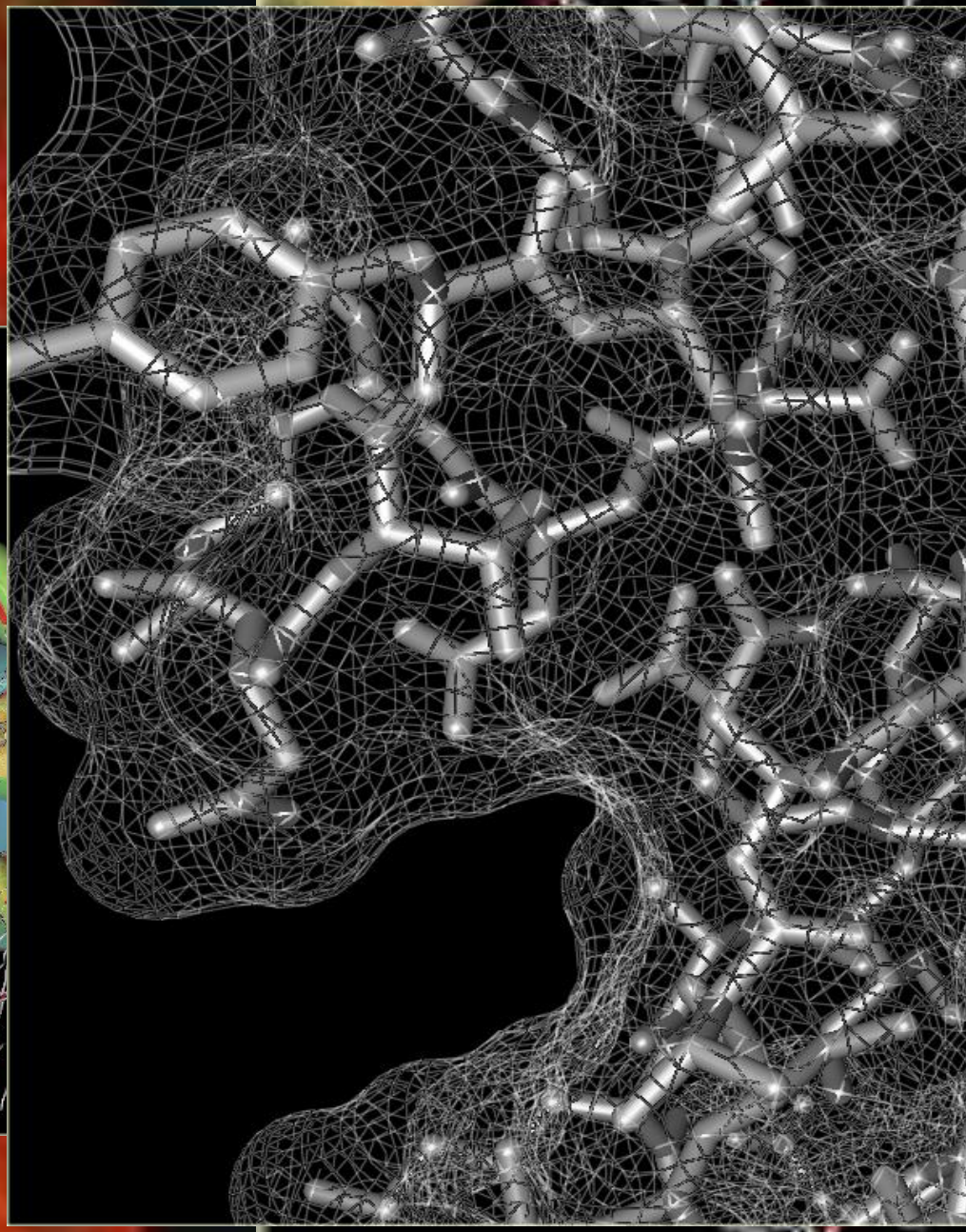
Objectifs - Enjeux de la bioinformatique structurale

- Stockage, archivage et structuration des banques de données
- Identification et classification des nouvelles protéines
  - Protéomique (Electrophorèse 2D, spectrométrie de masse)
  - Séquençage de génomes
  - Clonage « *in silico* »
- Prédiction de structures et de propriétés de protéines
  - Localisation de sites actifs
  - Prédiction, modification de fonction biologique
- Construction de structures 3D (modélisation moléculaire)
  - Mutagenèse dirigée
  - Applications biologiques
- Génomique structurale
  - Délimitation des limites de domaines fonctionnels

g.deleage@ibcp.fr

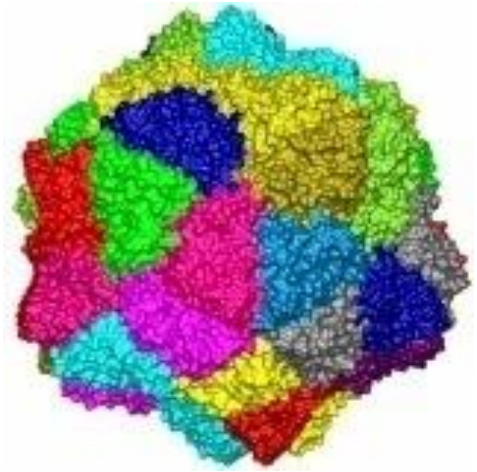
Google cell



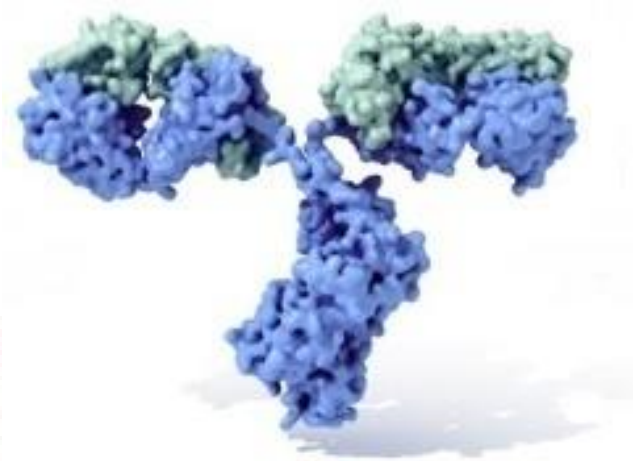


Domaines d'applications biologiques variés

- Virologie
  - Vaccins synthétiques
  - Reconnaissance moléculaire



- Immunologie
  - Synthèse de peptides antigéniques
  - Recherche d'épitopes
  - Anticorps thérapeutiques

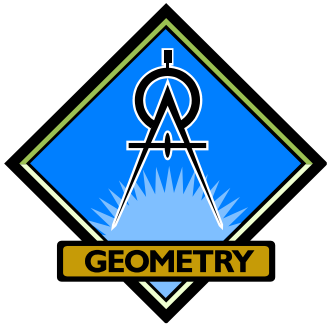


- Thérapie génique

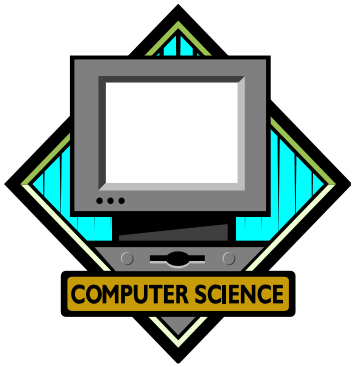
g.deleage@ibcp.fr

Pluridisciplinarité

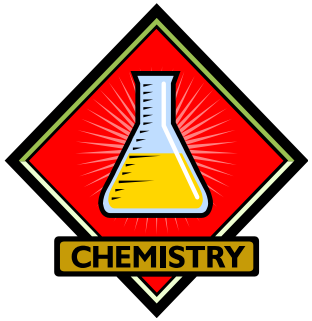
- Mathématiques
  - Programmation dynamique
  - Techniques de minimisation
  - Recuit simulé (Monte Carlo)
  - Calcul matriciel - Géométrie



- Informatique
  - Réseau et interfaces
  - Graphisme 3D
  - Algorithmique
  - Programmation dynamique

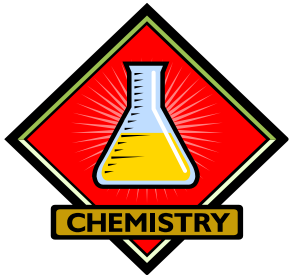


- Biologie-Chimie
  - Intégration dans le contexte biologique
  - Evolution et Phylogénie moléculaire
  - Alignements
  - Recherche de fonction
  - Modélisation moléculaire - Aide à l'expérimentation

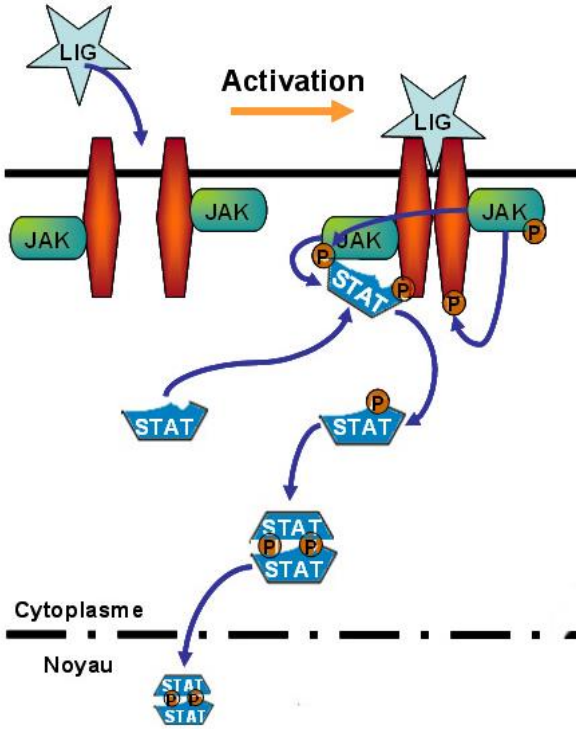


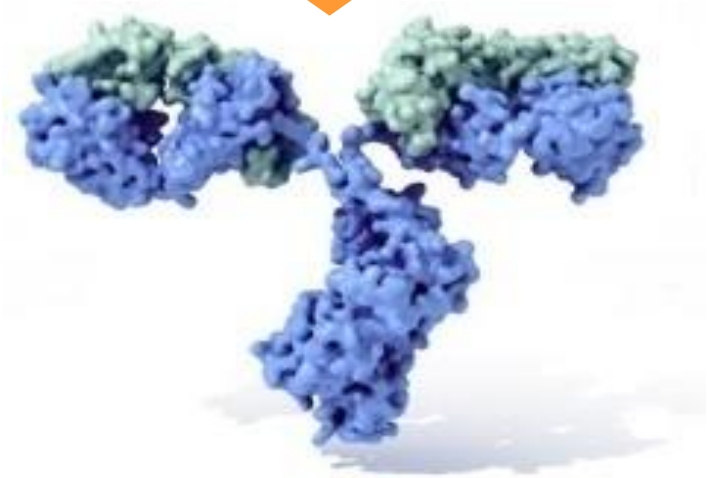
g.deleage@ibcp.fr

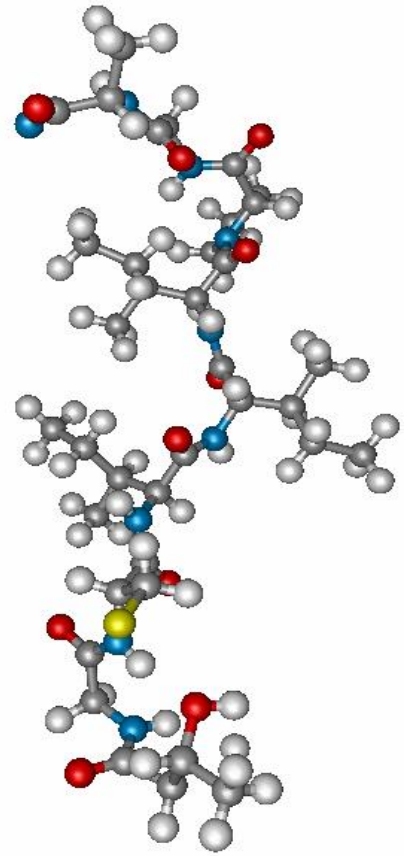
Séquences de point de vue des séquences



.. T G C I I P G A ..







g.deleage@ibcp.fr

Méthodes d'analyse de séquences biologiques

Banques de données

g.deleage@ibcp.fr

Evolution des Banques de données

- **Banques généralistes**
  - De 1980 => nos jours
  - Hétérogénéité
  - Croissance exponentielle (problème de la taille)
  - Exhaustivité
  - Erreurs
- **Banques spécialisées**
  - Thématiques
    - HIV (Los Alamos)
    - euHCVdb (IBCP-Lyon)
    - Récepteurs couplés aux protéines G (GPCRD-Njimegue)
    - Plantes (Flagdb++)
    - Anticorps (IMGT-Montpellier)
  - Espèces E.Coli (ECDB), Bacillus Subtilis,…
  - Structures
    - Secondaires de protéines non homologues (Hobohm et Sander)
    - Structures 3D (SCOP, CATH,
  - Régions codantes
  - Données biologiques (protéomique)
    - Gels 2D
    - RMN
    - Spectrométrie de masse
    - Enzymes
    - Réseaux métaboliques
  - Interactions (Mint, Smid, Bind, Interact, Mimi) http://www.imb-jena.de/jcb/ppi/

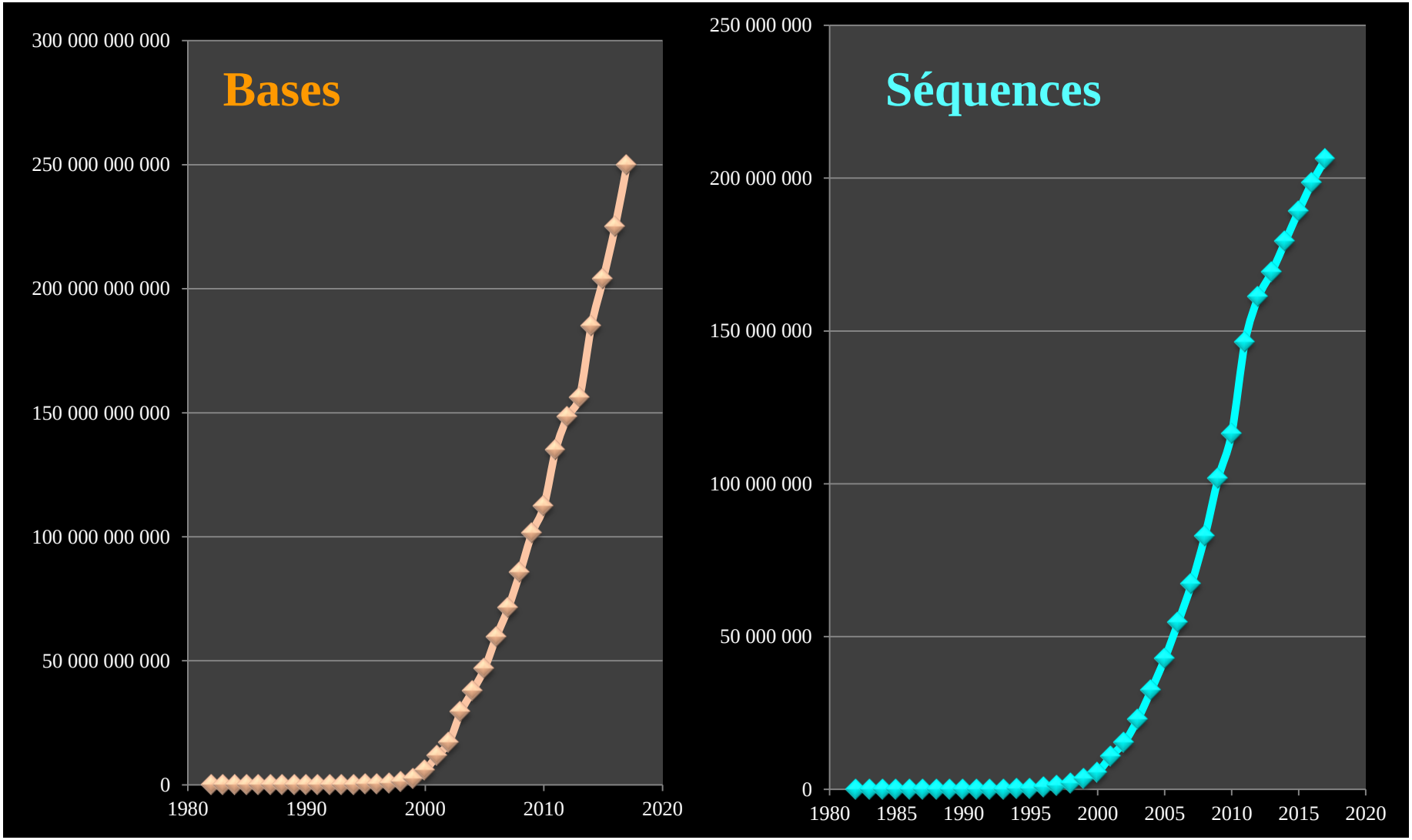
g.deleage@ibcp.fr

Caractéristiques de la biologie et des banques de données

- ✓ Les données biologiques
  - ✓ Hétérogénéité (informations, structurations, systèmes de requêtes)
  - ✓ Fortement corrélées (séquences *via* le code génétique)
  - ✓ Qualité discutable
    - ✓ Erreurs de séquences (au niveau séquençage ou saisie)
    - ✓ Propagation des erreurs par l'annotation automatique
    - ✓ Redondance (polymorphisme, gènes dupliqués, etc…)
  - ✓ Emergence constante de nouveaux types
    - ✓ Puce à ADN.
    - ✓ Spectrométrie de masse
    - ✓ RMN solide…
  - ✓ Fortement associées aux auteurs (formats)
  - ✓ Sémantique et présentation différentes selon le point de vue du biologiste
  - ✓ Volume important (tri…)
  - ✓ Tout est en anglais….
  - ✓ Changent tous les jours…(en moyenne 20-30 structures 3D déposées chaque jour)
- ✓ Les traitements informatiques
  - ✓ Comparaison de séquences (2 à 2)
  - ✓ Alignements multiples (n séquences)
  - ✓ Prédictions intro-exon sur des génomes complets,
  - ✓ Analyse de liaison pour la cartographie
  - ✓ Analyse de la structure des protéines
  - ✓ Analyse du transcriptome

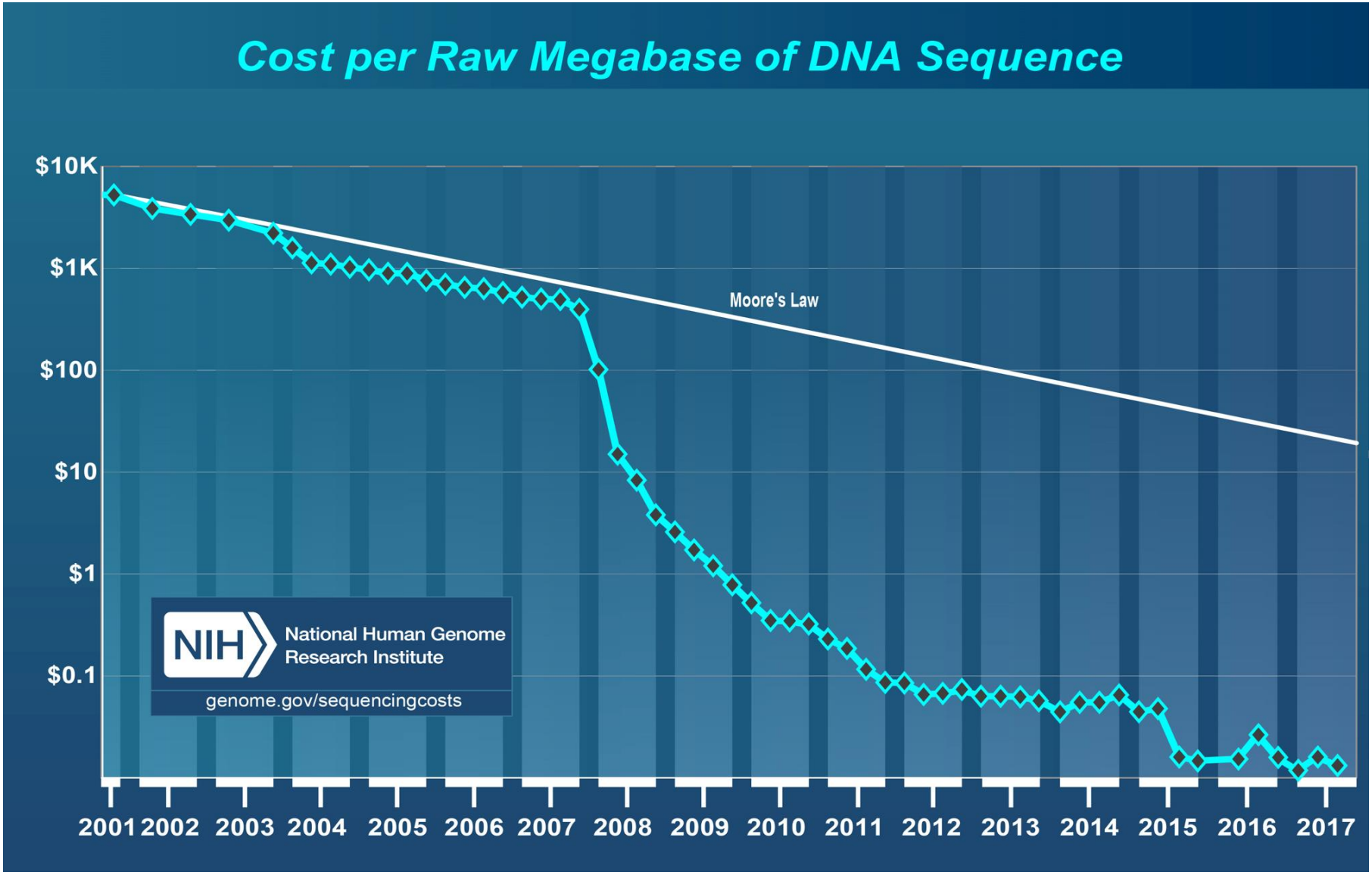
g.deleage@ibcp.fr

Croissance de genbank (1982-2017)
Banques de séquences d'acides nucléiques



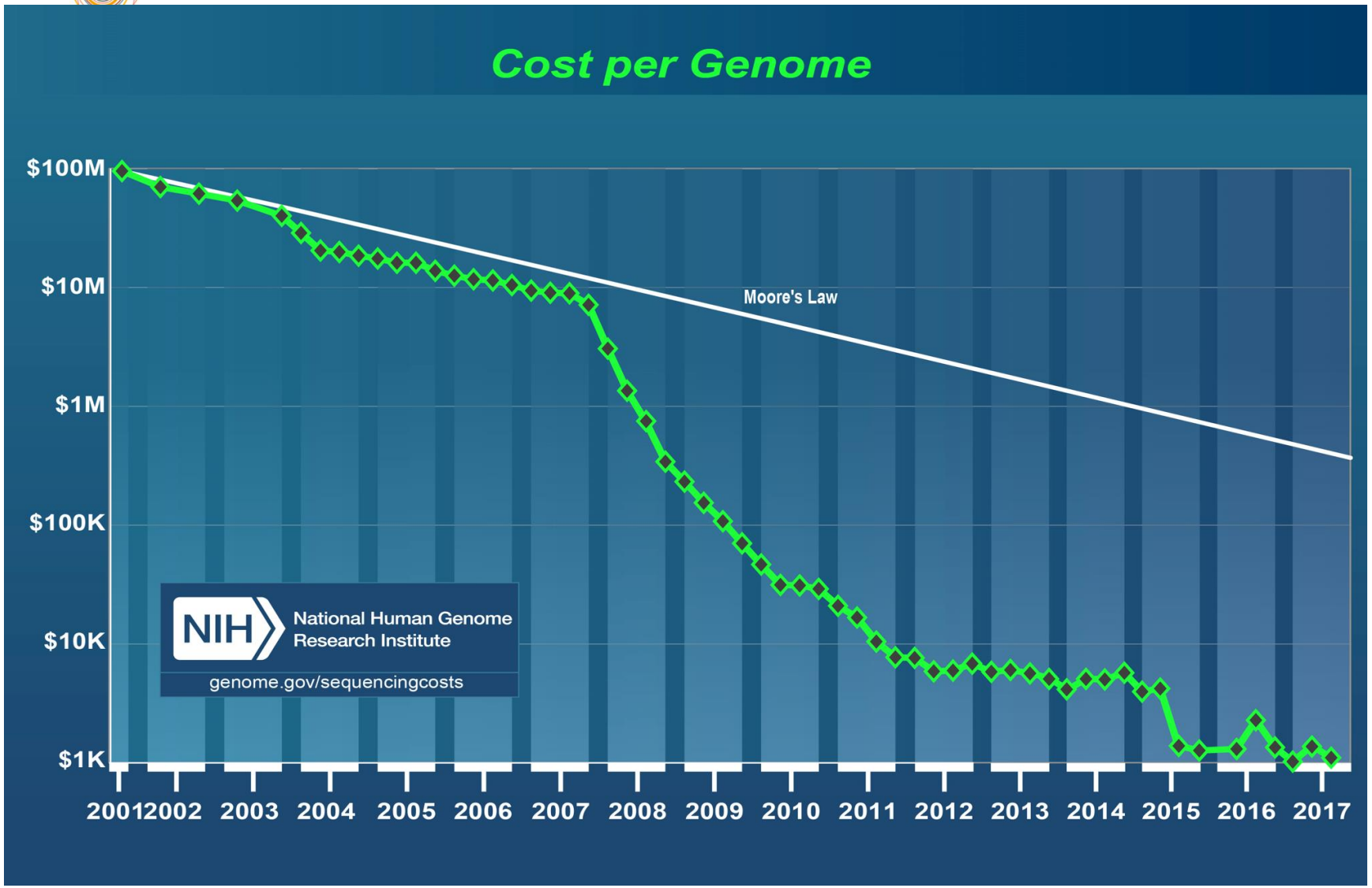
IBCP
g.deleage@ibcp.fr

Coût par mégabase



g.deleage@ibcp.fr

Coût comparé du séquençage d'un génome humain *



* La chute des coûts a fait qu'en 6 mois, les programmes de séquençages massifs de génomes ont généré autant de données que GenBank en a accumulé en 20 ans !

En 2018, une séquence de génome humain coute moins de 1000$

g.deleage@ibcp.fr

http://www.insdc.org/

Consortium international

ABOUT INSDC | POLICY | ADVISORS | DOCUMENTS

ENA European Nucleotide Archive

DDBJ

NCBI

International Nucleotide Sequence Database Collaboration

- The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing foundational initiative that operates between DDBJ, EMBL-EBI and NCBI. INSDC covers the spectrum of data raw reads, though alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations.

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

- The INSDC advisory board, the International Advisory Committee, is made up of members of each of the databases' advisory bodies. The International Advisory Committee published a paper reiterating the importance of depositing data to INSDC.
- Individuals submitting data to the international sequence databases should be aware of INSDC policy.

How to submit data

- For full details of how to submit data to the databases, please select a collaborating partner.
- DDBJ, ENA, GenBank
- The INSDC Feature Table Definition Document is available here.

http://www.insdc.org/

g.deleage@ibcp.fr

Bases de données nucléiques

Genbank 239 (08/2020) 9 236 443 421 310 bases
http://www.ncbi.nlm.nih.gov/genbank 1 901 329 611 séquences

European Nucleotide Archive 140 (08/2020) 8 327 10^9 bases
http://www.ebi.ac.uk/ena/ 2 621 700 000 séquences

DDBJ 120 (06/2020) 9,253,936,453,958 bases
http://www.ddbj.nig.ac.jp/ 2,414,499,799 séquences

Bases de données protéiques

http://www.uniprot.org

UniProt/TrEMBL(08/2020) 186 961 949 séquences

UniProt/Swiss-Prot (08/2020) 563 082 sequences annotées

202 799 066 acides aminés

Structures 3D

http://www.rcsb.org/pdb/

PDB RCSB (07/09/2020) 168358 Structures dont 50% avec des séquences <>

Protéines avec <30% identité ~35000 Groupes

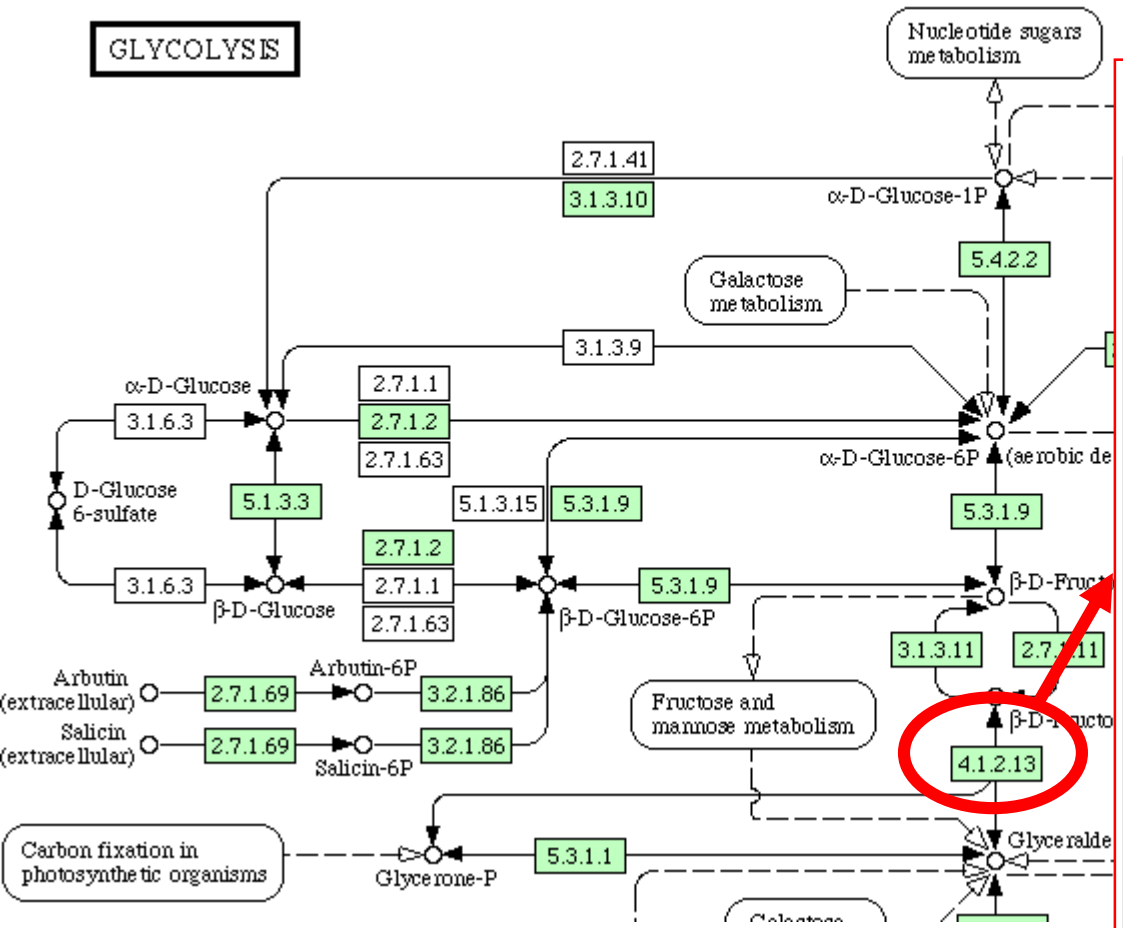
g.deleage@ibcp.fr

Données de bases pour bases de données

- **Banque des étiquettes (dbEST)**
  - Complète (09/2019) 77,827,308 Entrées
  - Homo sapiens 8,714,458 Entrées
  - Souris 4,871,688 Entrées
  - LifeSeq database privée Incyte (accès payant)
  - Human Genome Sciences exploite les données à travers des brevets http://www.hgsi.com/
- **Bases de données spécialisées**
- **Prosite 07/2018** http:www.expasy.ch/prosite
  - Banque de sites et signatures fonctionnelles **2529** entrées
- **Métabolisme**
  - Banque de données sur les enzymes et voies métaboliques
  - KEGG http://www.genome.ad.jp/kegg/
  - EcoCyc http://ecocyc.org/
- **OMIM**
  - Online Mendelian Inheritance in Man
- **HGMD**
  - Human Gene Mutation Database

g.deleage@ibcp.fr

KEGG (glycolyse)



KEGG **Escherichia coli K-12 W3110: JW2084**

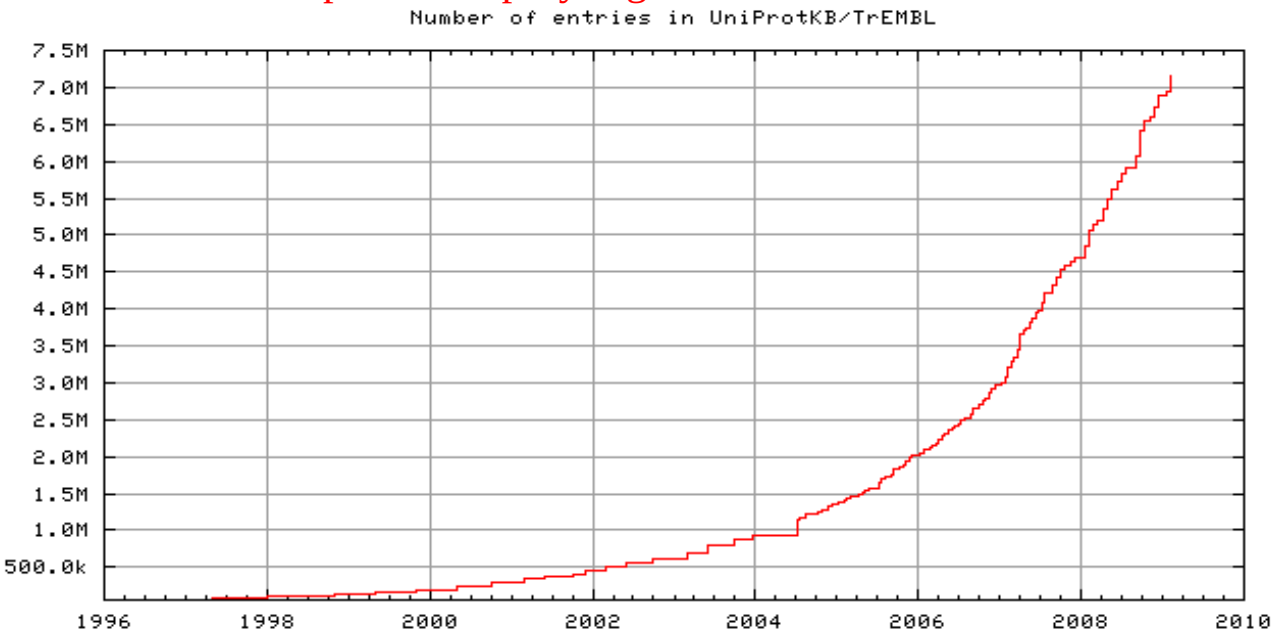
Help

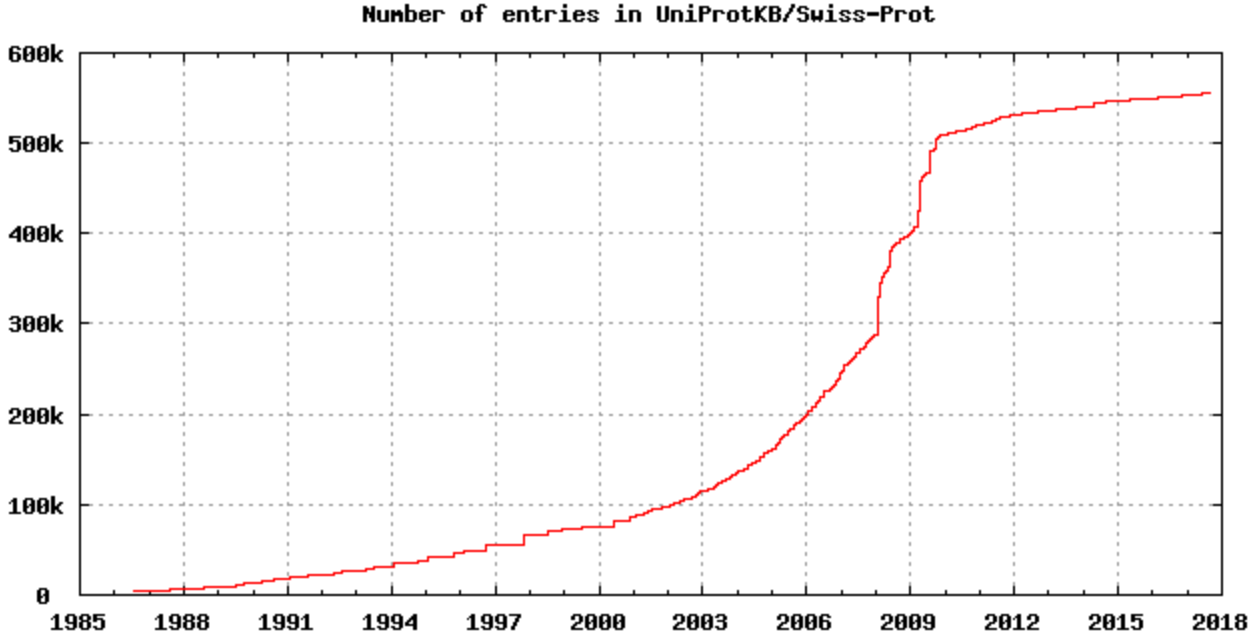
Entry	JW2084 CDS E.coli_J
Gene name	fbaB
Definition	fructose-bisphosphate aldolase, class I [EC:4.1.2.13]
KO	KO: KO1623 fructose-bisphosphate aldolase, class I OC search OC viewer
Pathway	Metabolism; Carbohydrate Metabolism; Glycolysis / Gluconeogenesis [PATH:ecj00010] Metabolism; Carbohydrate Metabolism; Pentose phosphate pathway [PATH:ecj00030] Metabolism; Carbohydrate Metabolism; Fructose and mannose metabolism [PATH:ecj00051] Metabolism; Energy Metabolism; Carbon fixation [PATH:ecj00710]
Class	Gene catalog
SSDB	Ortholog Paralog Motif Gene cluster
Other DBs	Nara: JW2084 Wisconsin: b2097 UniProt: P71295
LinkDB	PDB All DBs
Position	complement(2181707..2182831) Genome map
AA seq	374 aa AA seq FASTA-genes FASTA-sp BLAST-nr MIARKRRARTIHSRYPIGIYGSIVMTDIAQLLGKDADNLLQHRCMTIPSDQLYLPGHDYV DRVMIDNNRPPAVLRNMQTLYNTGRLAGTGYLSILPVDQGVEHSAGASFAANPLYFDPKN IVELAIEAGCNCVASTYGVLASVSRRYAHRIPFLVKLNHNETLSYPNTYDQTLYASVEQA FNMGAVAVGATIYFGSEESRRQIEEISAAFERAHELGMVTVLWAYLRNSAFKKDGVDYHV SADLTGQANHLAATIGADIVKQKMAENNGGYKAINYGYTDDRVYSKLTSENPIDLVRYQL ANCYMGRAGLINSGGAAGGETDLSDAVRTAVINKRAGGMGLILGRKAFKKSMADGVKLIN AVQDVYLDSKITIA
NT seq	1125 nt NT seq +upstream 0 nt +downstream 0 nt atgattgcccgcaaaaggcgggccaggacaatccatagccgatatccaatcggaatttac gggagcatagtaatgacagatattgcgcagttgcttggcaaagacgccgacaacctttta

g.deleage@ibcp.fr

Evolution des banques protéiques

http://web.expasy.org/docs/relnotes/relstat.html





g.deleage@ibcp.fr

Uniprot 2020 - Biais des espèces (1/3)

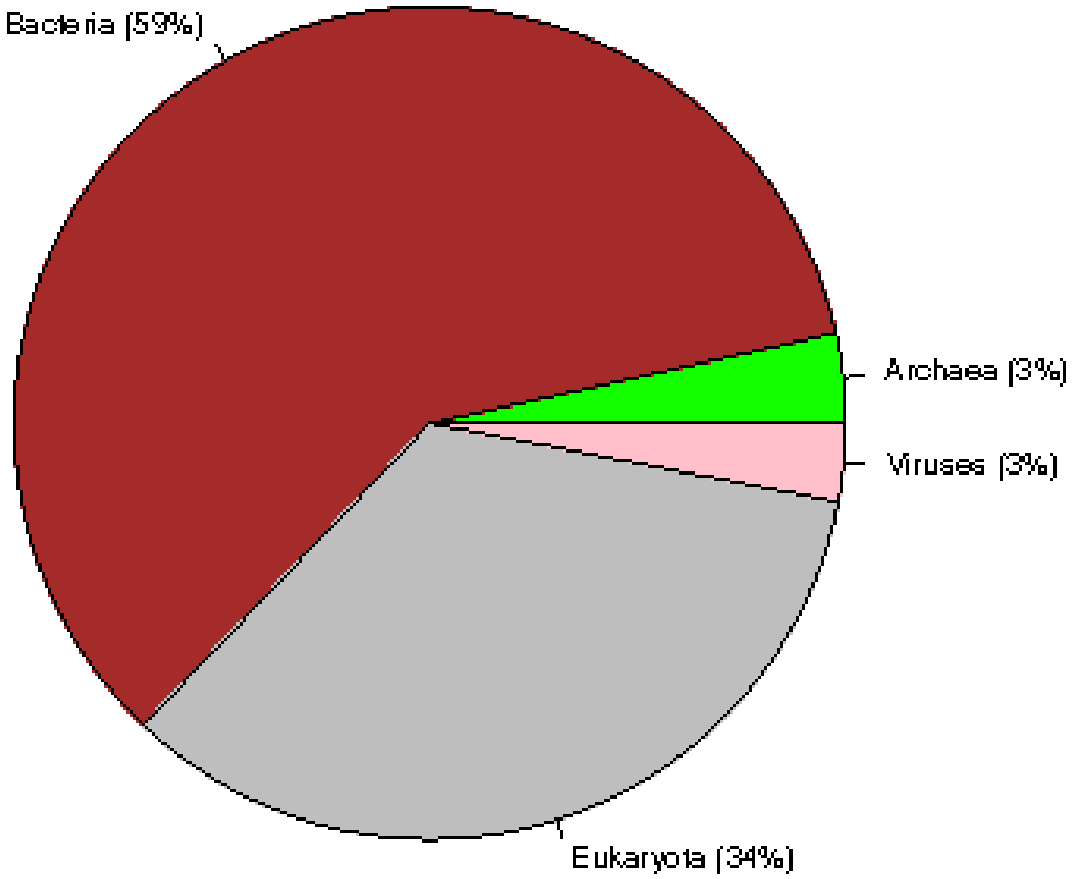
- **Nombre d'espèces représentées : 13936**
- **20 espèces représentent 121365 sequences soit 21.6% des entrées :**

1	20375	Homo sapiens (Human)
2	17046	Mus musculus (Mouse)
3	15991	Arabidopsis thaliana (Mouse-ear cress)
4	8106	Rattus norvegicus (Rat)
5	6721	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)
6	6012	Bos taurus (Bovine)
7	5141	Schizosaccharomyces pombe (strain 972 / ATCC 24843) (Fission yeast)
8	4518	Escherichia coli (strain K12)
9	4191	Bacillus subtilis (strain 168)
10	4149	Dictyostelium discoideum (Slime mold)
11	4143	Caenorhabditis elegans
12	4086	Oryza sativa subsp. japonica (Rice)
13	3610	Drosophila melanogaster (Fruit fly)
14	3451	Xenopus laevis (African clawed frog)
15	3164	Danio rerio (Zebrafish) (Brachydanio rerio)
16	2295	Gallus gallus (Chicken)
17	2218	Pongo abelii (Sumatran orangutan) (Pongo pygmaeus abelii)
18	2209	Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv)
19	2042	Escherichia coli O157:H7
20	1898	Mycobacterium tuberculosis (strain CDC 1551 / Oshkosh)

30 à 50% des nouvelles séquences sont homologues à des séquences déjà identifiées.

g.deleage@ibcp.fr

UniProt 2020 - Biais des espèces (2/3)

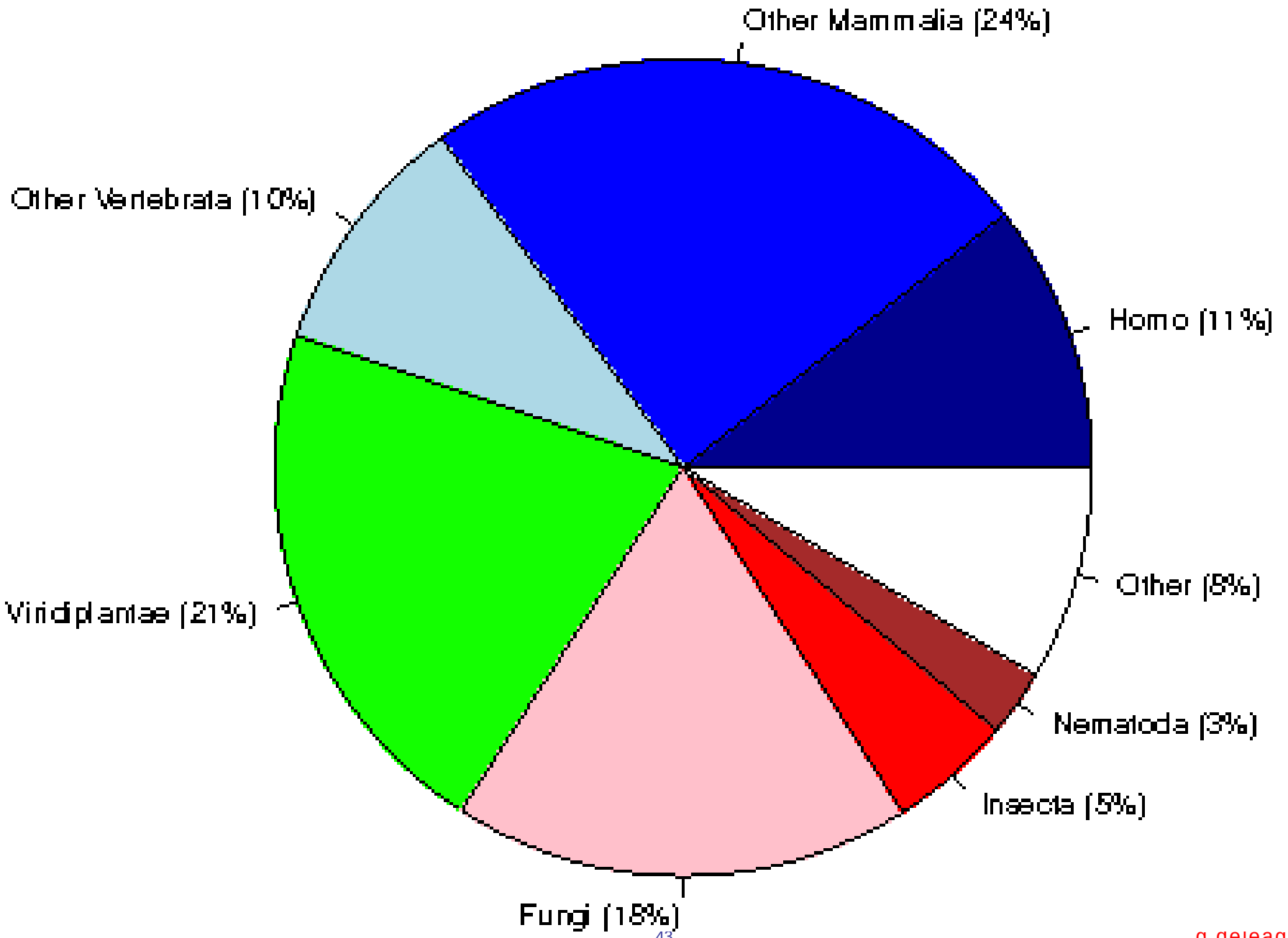


	espèce
1 seq:	5713
2 seq :	2008
3 seq :	1087
4 seq :	709
5 seq :	516
6 seq :	410
7 seq :	317
8 seq :	252
9 seq :	232
10 seq:	143
11-20:	805
21- 50:	469
51-100:	221
>100:	1054

Archaea	19634 (3%)
Bacteria	334639 (59%)
Eukaryota	191801 (34%)
Viruses	1708 (3%)

g.deleage@ibcp.fr

UniProt 2020 - Biais des espèces (3/3)

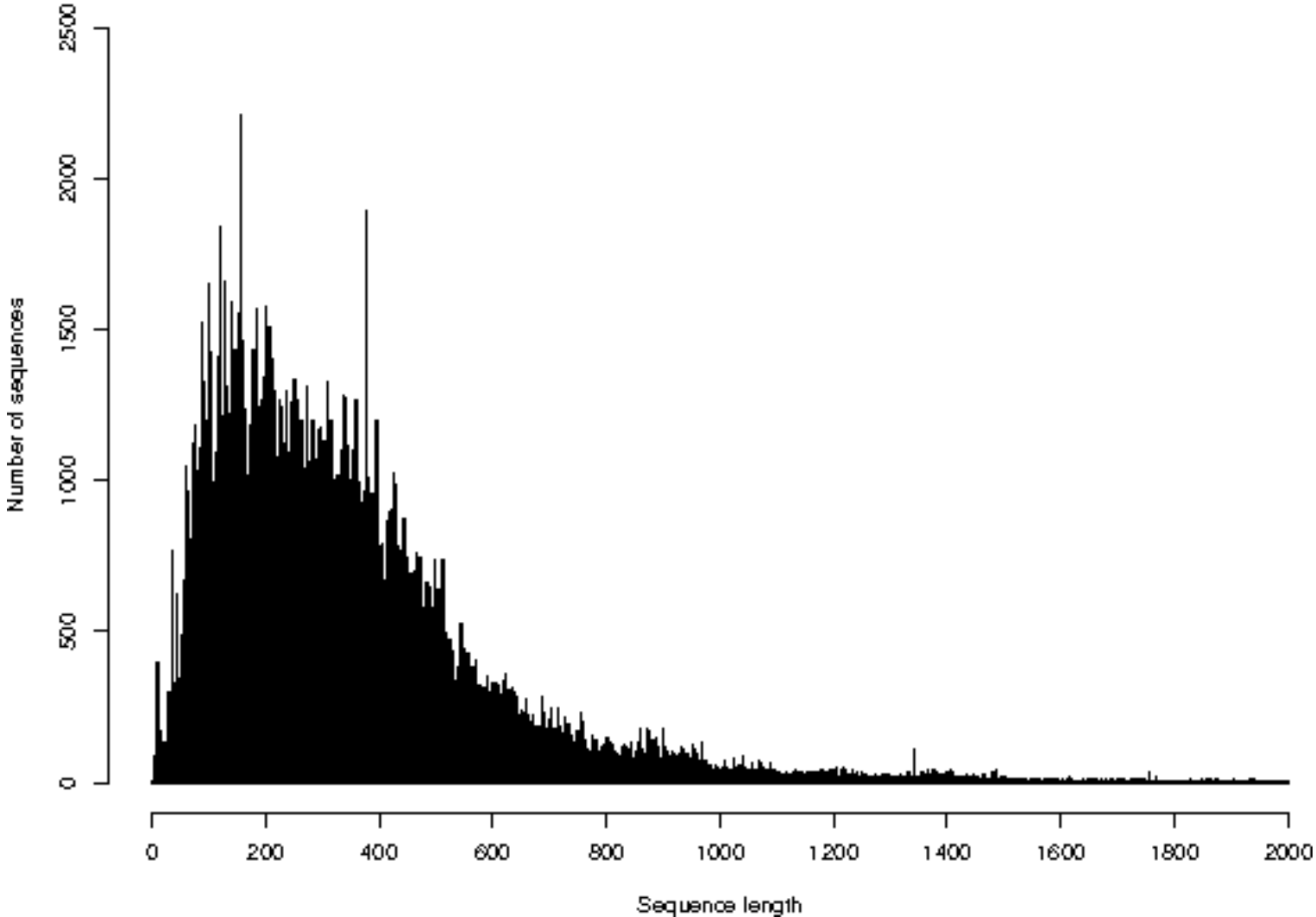


g.deleage@ibcp.fr

SWISS-PROT 2020 - Distribution

GWA_SEPOF (P83570): 2 aa
TITIN_MOUSE (A2ASS6): 35213 aa

Moyenne des protéines : 360 acides aminés



g.deleage@ibcp.fr

SWISSPROT 2020

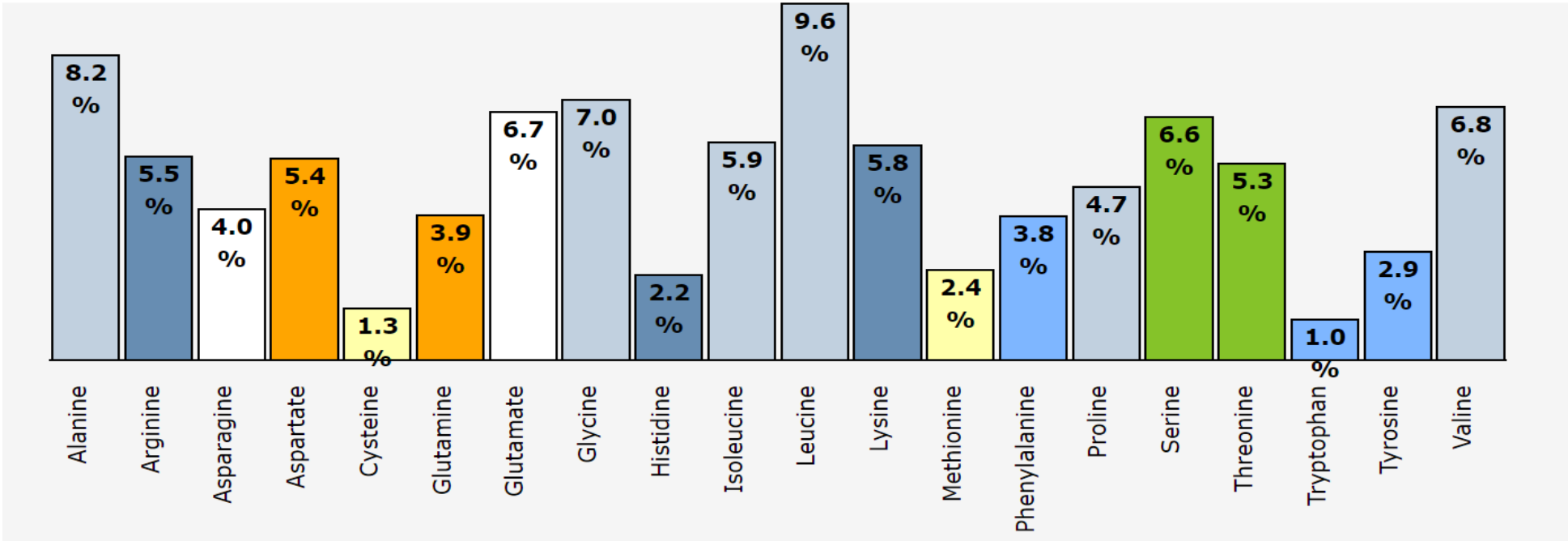
http://web.expasy.org/docs/relnotes/relstat.html

Fréquence des acides aminés (x100)

Ala (A) 8.26	Gln (Q) 3.93	Leu (L) 9.66	Ser (S) 6.57
Arg (R) 5.53	Glu (E) 6.74	Lys (K) 5.83	Thr (T) 5.34
Asn (N) 4.05	Gly (G) 7.08	Met (M) 2.41	Trp (W) 1.09
Asp (D) 5.46	His (H) 2.27	Phe (F) 3.86	Tyr (Y) 2.92
Cys (C) 1.37	Ile (I) 5.95	Pro (P) 4.71	Val (V) 6.87

Acides aminés les plus fréquents

Leu, Ala, Gly, Val, Glu, Ser, Ile, Lys, Arg, Asp,
Thr, Pro, Asn, Gln, Phe, Tyr, Met, His, Cys, Trp



g.deleage@ibcp.fr

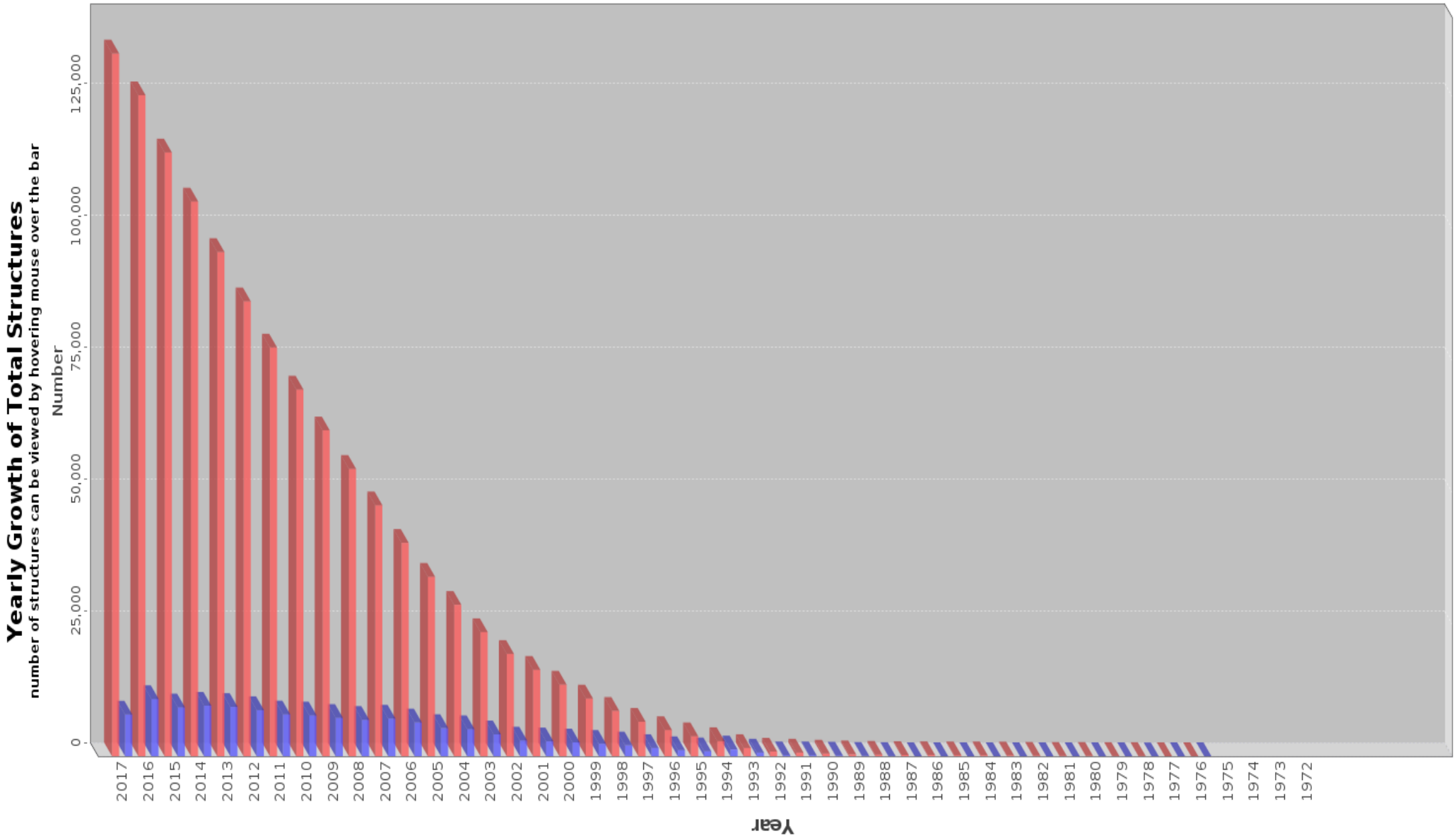
Format des séquences

- **18 formats différents**
  - Ig/Stanford
  - Genbank/GB
  - NBRF
  - EMBL
  - GCG
  - Pearson/Fasta
  - Zuker
  - Olsen
  - Fitch
  - Phylip3.2
  - Phylip
  - Plain/Raw
  - PIR/CODATA
  - MSF.1
  - PAUP
  - Pretty
  - ANTHEPROT
  - DNA strider
- **Un utilitaire de reformatage: READSEQ**

```
>sw|P02159|MYG_LYCPI Myoglobin.
GLSDGEWQIVLNIWGKVETDLAGHGQEVLIRLFKNHPETLDKFDKFKHLKTEDEMKGSED
LKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPVKYLEFISDAIIQVLQNKHS
GDFHADTEAAMKKALELFRNDIAAKYKELGFQG
>sw|Q9DEP1|MYG_PSEGE Myoglobin.
ADFDMVLKCWGLVEADYATYGSLVLTRLFTEHPETLKLFPKFAGIAHGDLAGDAGVSAHG
ATVLNKLGDLLKARGGHAALLKPLSSSHATKHKIPIINFKLIAEVIGKVMEEKAGLDAAG
QTALRNVMAVIIADMEADYKELGFTE
>sw|P02201|MYG_GRAGE Myoglobin.
GLSDDEWHHVLGIWAKVEPDLSAHGQEVIIRLFQVHPETQERFAKFKNLKTIDELRSSEE
VKKHGTTVLTALGRILKLKNNHEPELKPLAESHATKHKIPVKYLEFICEIIVKVIAEKHP
SDFGADSQAAMRKALELFRNDMASKYKEFGFQG
>sw|P02185|MYG_PHYCA Myoglobin.
VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED
LKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP
GDFGADAQGAMNKALELFRKDIAAKYKELGYQG
```


g.deleage@ibcp.fr

Evolution de la banque PDB



g.deleage@ibcp.fr

Contenu de la banque PDB (07/09/2020)

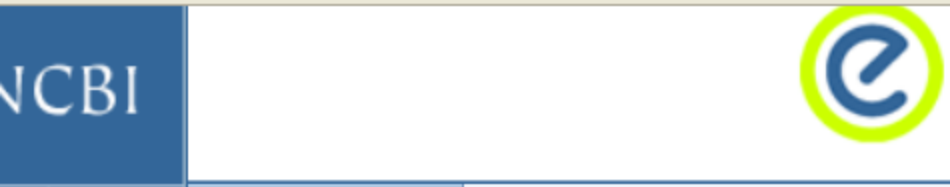
Molecular Type	X-ray	NMR	EM	Multiple methods	Neutron	Other	Total
Protein (only)	132181	11451	3880	160	67	32	147771
Other	7964	92	447	6	0	4	8513
Protein/NA	7016	265	1354	3	0	0	8638
Nucleic acid (only)	2081	1300	47	5	2	1	3436
Total	149242	13108	5728	174	69	37	168358

g.deleage@ibcp.fr

Accès aux banques - Interrogations

- Entrez développé au NCBI
  http://www.ncbi.nlm.nih.gov/Entrez/
- ACNUC (le premier…)
  - Système multi-critères très puissant
  - http://pbil.univ-lyon1.fr/search/query_fam.php
- SRS (Sequence Retrieval System) est mort!
  - Interrogations multi-banques et multi-critères
  - [http://srs.ebi.ac.uk/] [http://srs-pbil.ibcp.fr][http://srs.sanger.ac.uk]
  - SRS est un système européen relativement générique permettant d'intégrer des dizaines de bases génomiques et qui offre des outils de navigation et de recherche orientés WEB
  - C'est la référence européene en matière d'intégration de données génomiques,
  - SRS repose sur une technologie de fichiers plats ASCII et de fichiers d'index qui pointent directement vers des entrées dans les fichiers plats. SRS n'est pas basé sur un SGDB C'est un système essentiellement *read only*,
  - La technologie sur laquelle repose SRS (pointeurs directs vers des fichiers de données) n'est pas adaptée aux mises à jour incrémentales,
  - Données peu structurées,
  - Pas d'API permettant d'accéder aux données structurées.
- SGBD, SGBDO
  - Systèmes de requête (Oracle, db2, 4D, Sybase, MySQL, PostgreSQL)
  - http://www.ebi.ac.uk/biomart
  - EYEDB [http://www.sysra.com/eyedb/]
- No SQL (google…)

Banque de données	Nombre d'entrées	Taille de la base (Go)	Nombre d'objets bio	Durée d'import
PROSITE	1,5 K	0,8	108 K	6 min
SWISSPROT	100 K	2,9	2,4 M	5h30
SPTREMBL	660 K	13	8,4 M	20h33
EMBL	17 M	261	122 M	25j
PRODOM	305 K	3,1	2,5 M	3h50
PFAM	85 K	1,9	1,6 M	10h04
BLOCKS	12 K	0,6	690 K	1h40
ENZYME	4 K	0,2	42 K	5 min
RHDB	133 K	1,9	1,34 M	1h58



Search across databases [ ] GO Clear Help

Welcome to the Entrez cross-database search page

- **PubMed:** biomedical literature citations and abstracts
- **PubMed Central:** free, full text journal articles
- **Site Search:** NCBI web and FTP sites
- **Books:** online books
- **OMIM:** online Mendelian Inheritance in Man

- **Nucleotide:** Core subset of nucleotide sequence records
- **EST:** Expressed Sequence Tag records
- **GSS:** Genome Survey Sequence records
- **Protein:** sequence database
- **Genome:** whole genome sequences
- **Structure:** three-dimensional macromolecular structures
- **Taxonomy:** organisms in GenBank
- **SNP:** short genetic variations
- **dbGaP:** genotype and phenotype
- **UniGene:** gene-oriented clusters of transcript sequence
- **CDD:** conserved protein domain database
- **Clone:** integrated data for clone resources
- **UniSTS:** markers and mapping data
- **PopSet:** population study data sets
- **GEO Profiles:** expression and molecular abundance pro
- **GEO DataSets:** experimental sets of GEO data

Accès aux banques – Interrogations via ACNUC

WWW-Query

BBE contribution to PBIL in Lyon, France

Quick search | WWW-Query | Cross taxa search | History | Species | List | Modify | Retrieve | Releases | Help

Usage Example No script

Sequence search selected

(•) Search for sequences ( ) Search for families, alignments and trees

Protein sequences selected

(•) Protein databank ( ) Nucleotide databank UniprotKB/SwissProt

Selection criteria:

1. DEFAULT Author
2. AND Keyword
3. AND Keyword
4. AND Keyword

List name: query_list SUBMIT CLEAR

This form allows you to compose a query in a way to retrieve sequences or families in one of the databases available on this server. First you need to select the type of search, the type of sequence, then one of the available database, afterwards you have to write your query using the different selectors and text editors. Criteria like keywords accept the use of wildcard (character *).Use of space is allowed and search is case insensitive. For example, you may type RNA POLYMERASE, or *POLYMERASE, or RNA*. If you select a "Family Search" you will retrieve a list of families. Then a document can be generated for each family, allowing you to display its associated **alignment** and **phylogenetic tree**.

User reference :1458751951

g.deleage@ibcp.fr

EMBL European Bioinformatics Institute - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

EMBL European Bioinformatics ...

www.ebi.ac.uk

ebi

Liens Gilbert Deléage - Citati... IBCP : Institute of Biolo... Site Web du Pr. Gilbert ... Web of Knowledge [v.5....

EMBL-EBI

Services Research Training About us

The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Recherche de pseudogene avec une faute psuedogene

Explore the EBI:

psuedogene

Search

Examples: blast, keratin, bfl1...

Press release

Popular

- Services
- Research
- Training
- News
- Jobs
- Visit us
- EMBL
- Contacts

Industry Programme

EBI Search

psuedogene

Search

Examples: VAV_HUMAN, tpi1, Sulston ...

Advanced

Search results for *psuedogene*

Showing **10** results out of **108** in All results

Filter your results

Source

All results (108)
Genomes (33)
Nucleotide sequences (60)
Protein sequences (12)
Literature (3)

Literature (3 results found)

Antisense transcription of a murine FGFR-3 psuedogene during fetal developement.

Weil D, Power MA, Webb GC, Li CL
(1997 Mar 10) Gene, 187(1):115-22

Related data Views

Source: MEDLINE
ID: 9073074

View all 3 results for Literature

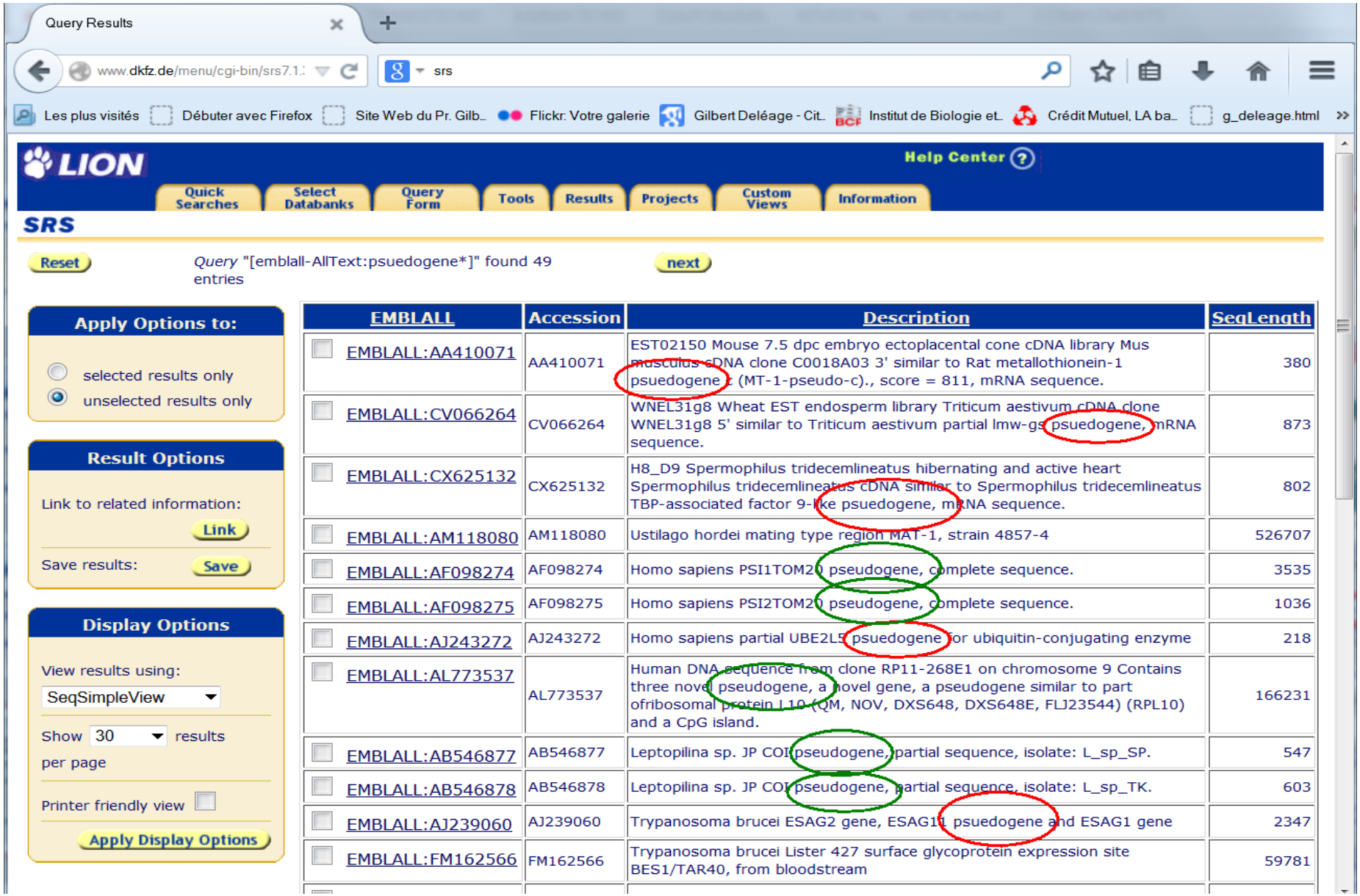
Genomes (33 results found)

OTTHUMG00000008273

Related data Views

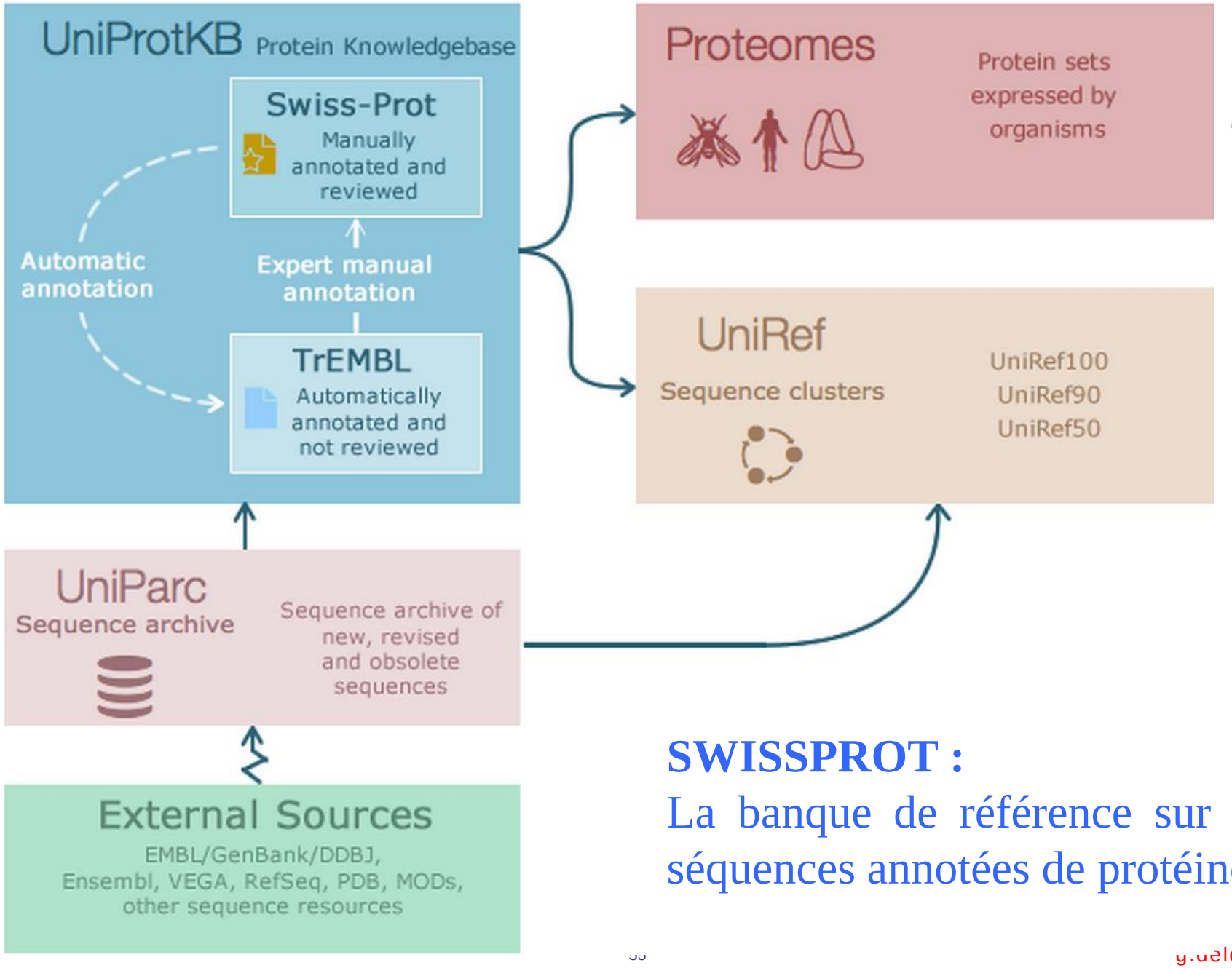
Erreur dans les banques de séquences

Recherche de pseudogene avec une faute psuedogene



g.deleage@ibcp.fr

UniProtKB http://www.uniprot.org/



SWISSPROT :

La banque de référence sur les séquences annotées de protéines

g.deleage@ibcp.fr

Results

About UniProtKB | Basket

Filter by

- Reviewed (772) Swiss-Prot
- Unreviewed (1,106) TrEMBL

Popular organisms

- Human (972)
- Mouse (150)
- Zebrafish (141)
- Rat (41)
- Bovine (7)

Other organisms

Go

Search terms

Filter "receptor" as:

- disease (4)
- domain (18)
- gene ontology (1,609)
- keyword (447)

Columns | BLAST | Align | Download | Add to basket

1 to 25 of 1,878 | Show 25

Entry	Entry name	Protein names	Gene names	Organism	Length
P08069	IGF1R_HUMAN	**Insulin-like growth factor 1 recept...**	**IGF1R**	Homo sapiens (Human)	1,367
P09208	INSR_DROME	**Insulin-like receptor**	**InR**, dinr, Dir-a, Inr-a, CG18402	Drosophila melanogaster (Fruit fly)	2,144
P06213	INSR_HUMAN	**Insulin receptor**	**INSR**	Homo sapiens (Human)	1,382
Q9UQB8	BAIP2_HUMAN	**Brain-specific angiogenesis inhibit...**	**BAIAP2**	Homo sapiens (Human)	552
P11717	MPRI_HUMAN	**Cation-independent mannose-6-phosph...**	**IGF2R**, MPRI	Homo sapiens (Human)	2,491
Q968Y9	INSR_CAEEL	**Insulin-like receptor**	**daf-2**, Y55D5A.5	Caenorhabditis elegans	1,846
P14616	INSRR_HUMAN	**Insulin receptor-related protein**	**INSRR**, IRR	Homo sapiens (Human)	1,297
P51460	INSL3_HUMAN	**Insulin-like 3**	**INSL3**, RLF, RLNL	Homo sapiens (Human)	131
Q9Y5Q6	INSL5_HUMAN	**Insulin-like peptide INSL5**	**INSL5**, UNQ156/PRO182	Homo sapiens (Human)	135
P63244	GBLP_HUMAN	**Guanine nucleotide-binding protein ...**	**GNB2L1**, HLC7, PIG21	Homo sapiens (Human)	317
Q8TDU9	RL3R2_HUMAN	**Relaxin-3 receptor 2**	**RXFP4**, GPR100, RLN3R2	Homo sapiens (Human)	374
P01308	INS_HUMAN	**Insulin**	**INS**	Homo sapiens (Human)	110

Une entrée UniProtKB (1/3) http://www.uniprot.org/uniprot/P06213.txt

```
ID   INSR_HUMAN              Reviewed;        1382 AA.
AC   P06213; Q17RW0; Q59H98; Q9UCB7; Q9UCB8; Q9UCB9;
DT   01-JAN-1988, integrated into UniProtKB/Swiss-Prot.
DT   05-OCT-2010, sequence version 4.
DT   22-JUL-2015, entry version 216.
DE   RecName: Full=Insulin receptor;
DE            Short=IR;
DE            EC=2.7.10.1;
DE   AltName: CD_antigen=CD220;
DE   Contains:
DE     RecName: Full=Insulin receptor subunit alpha;
DE   Contains:
DE     RecName: Full=Insulin receptor subunit beta;
DE   Flags: Precursor;
GN   Name=INSR;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM LONG), AND VARIANTS GLY-2;
RP   HIS-171; THR-448 AND LYS-492.
RX   PubMed=2859121; DOI=10.1016/0092-8674(85)90334-4;
RA   Ebina Y., Ellis L., Jarnagin K., Edery M., Graf L., Clauser E.,
RA   Ou J.-H., Masiarz F., Kan Y.W., Goldfine I.D., Roth R.A., Rutter W.J.;
RT   "The human insulin receptor cDNA: the structural basis for hormone-
RT   activated transmembrane signalling.";
RL   Cell 40:747-758(1985).
RN   [2]
```


g.deleage@ibcp.fr

Une entrée UniProtKB 2/3


```
..
CC   -!- FUNCTION: THIS RECEPTOR BINDS INSULIN AND HAS A TYROSINE-PROTEIN
CC       KINASE ACTIVITY.
CC   -!- CATALYTIC ACTIVITY: ATP + A PROTEIN TYROSINE = ADP +
CC       PROTEIN TYROSINE PHOSPHATE.
CC   -!- ENZYME REGULATION: AUTOPHOSPHORYLATION ACTIVATES THE KINASE
CC       ACTIVITY.
CC   -!- SUBUNIT: TETRAMER OF 2 ALPHA AND 2 BETA CHAINS LINKED BY DISULFIDE
CC       BONDS. THE ALPHA CHAINS CONTRIBUTE TO THE FORMATION OF THE LIGAND-
CC       BINDING DOMAIN, WHILE THE BETA CHAIN CARRY THE KINASE DOMAIN.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN.
CC   -!- ALTERNATIVE PRODUCTS: TWO FORMS ARE PRODUCED BY ALTERNATIVE
CC       SPLICING. THE SECOND FORM LACKS A 12 RESIDUE PEPTIDE IN THE ALPHA
CC       SUBUNIT.
CC   -!- PTM: AFTER BEING TRANSPORTED FROM THE ENDOPLASMIC RETICULUM TO THE
CC       GOLGI APPARATUS, THE SINGLE GLYCOSYLATED PRECURSOR IS FURTHER
CC       GLYCOSYLATED AND THEN CLEAVED, FOLLOWED BY ITS TRANSPORT TO THE
CC       PLASMA MEMBRANE.
CC   -!- DISEASE: MUTATIONS IN INSR CAN CAUSE VARIOUS FORMS OF INSULIN
CC       RESISTANCE AS WELL AS SOME FORMS OF DIABETES MELLITUS, NONINSULIN-
CC       DEPENDENT (NIDDM) AND OF LEPRECHAUNISM (DONOHUE SYNDROME).
CC   -!- SIMILARITY: BELONGS TO THE INSULIN RECEPTOR FAMILY OF TYROSINE-
CC       PROTEIN KINASES.
CC   -!- SIMILARITY: CONTAINS 2 FIBRONECTIN TYPE III-LIKE DOMAINS.
.
DR   EMBL; M10051; AAA59174.1; -.
DR   EMBL; X02160; CAA26096.1; -.
.
KW   Transferase; Tyrosine-protein kinase; Receptor; Transmembrane;
KW   Glycoprotein; ATP-binding; Phosphorylation; Signal; Polymorphism;
KW   Disease mutation; Diabetes; Alternative splicing; Repeat;
KW   3D-structure.
```


g.deleage@ibcp.fr

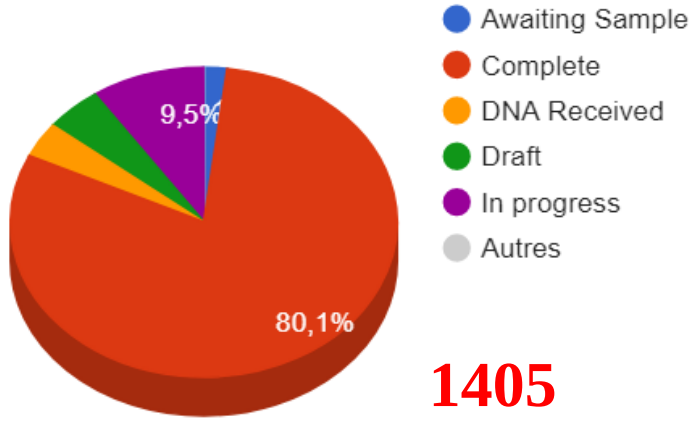
Une entrée UniProtKB 3/3


```
FT   SIGNAL        1     27
FT   CHAIN        28    758       INSULIN RECEPTOR, ALPHA-SUBUNIT.
FT   PROPEP      759    762       REMOVED IN MATURE FORM.
FT   CHAIN       763   1382       INSULIN RECEPTOR, BETA-SUBUNIT.
FT   DOMAIN      763    956       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM    957    979       POTENTIAL.
FT   DOMAIN      980   1382       CYTOPLASMIC (POTENTIAL).
FT   DOMAIN      182    339       CYS-RICH.
.
SQ   SEQUENCE   1382 AA;  156280 MW;  D681AC2E CRC32;
     MGTGGRRGAA AAPLLVAVAA LLLGAAGHLY PGEVCPGMDI RNNLTRLHEL ENCSVIEGHL
     QILLMFKTRP EDFRDLSFPK LIMITDYLLL FRVYGLESLK DLFPNLTVIR GSRLFFNYAL
     VIFEMVHLKE LGLYNLMNIT RGSVRIEKNN ELCYLATIDW SRILDSVEDN HIVLNKDDNE
     ECGDICPGTA KGKTNCPATV INGQFVERCW THSHCQKVCP TICKSHGCTA EGLCCHSECL
     GNCSQPDDPT KCVACRNFYL DGRCVETCPP PYYHFQDWRC VNFSFCQDLH HKCKNSRRQG
     CHQYVIHNNK CIPECPSGYT MNSSNLLCTP CLGPCPKVCH LLEGEKTIDS VTSAQELRGC
     TVINGSLIIN IRGGNNLAAE LEANLGLIEE ISGYLKIRRS YALVSLSFFR KLRLIRGETL
     EIGNYSFYAL DNQNLRQLWD WSKHNLTTTQ GKLFFHYNPK LCLSEIHKME EVSGTKGRQE
     RNDIALKTNG DKASCENELL KFSYIRTSFD KILLRWEPYW PPDFRDLLGF MLFYKEAPYQ
     NVTEFDGQDA CGSNSWTVVD IDPPLRSNDP KSQNHPGWLM RGLKPWTQYA IFVKTLVTFS
     DERRTYGAKS DIIYVQTDAT NPSVPLDPIS VSNSSSQIIL KWKPPSDPNG NITHYLVFWE
     RQAEDSELFE LDYCLKGLKL PSRTWSPPFE SEDSQKHNQS EYEDSAGECC SCPKTDSQIL
     KELEESSFRK TFEDYLHNVV FVPRKTSSGT GAEDPRPSRK RRSLGDVGNV TVAVPTVAAF
     PNTSSTSVPT SPEEHRPFEK VVNKESLVIS GLRHFTGYRI ELQACNQDTP EERCSVAAYV
     SARTMPEAKA DDIVGPVTHE IFENNVVHLM WQEPKEPNGL IVLYEVSYRR YGDEELHLCV
     SRKHFALERG CRLRGLSPGN YSVRIRATSL AGNGSWTEPT YFYVTDYLDV PSNIAKIIIG
     PLIFVFLFSV VIGSIYLFLR KRQPDGPLGP LYASSNPEYL SASDVFPCSV YVPDEWEVSR
     EKITLLRELG QGSFGMVYEG NARDIIKGEA ETRVAVKTVN ESASLRERIE FLNEASVMKG
     FTCHHVVRLL GVVSKGQPTL VVMELMAHGD LKSYLRSLRP EAENNPGRPP PTLQEMIQMA
     AEIADGMAYL NAKKFVHRDL AARNCMVAHD FTVKIGDFGM TRDIYETDYY RKGGKGLLPV
     RWMAPESLKD GVFTTSSDMW SFGVVLWEIT SLAEQPYQGL SNEQVLKFVM DGGYLDQPDN
     CPERVTDLMR MCWQFNPKMR PTFLEIVNLL KDDLHPSFPE VSFFHSEENK APESEELEME
     FEDMENVPLD RSSHCQREEA GGRDGGSSLG FKRSYEEHIP YTHMNGGKKN GRILTLPRSN
     PS
//
```


g.deleage@ibcp.fr

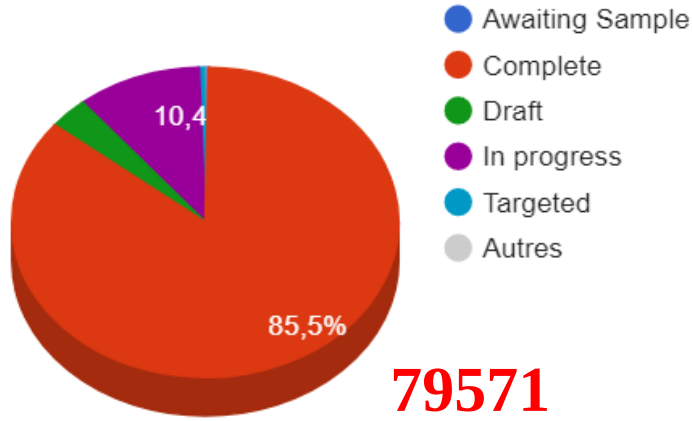
Génomes séquencés (04/09/2017) https://gold.jgi-psf.org/

Archaea Sequencing Status



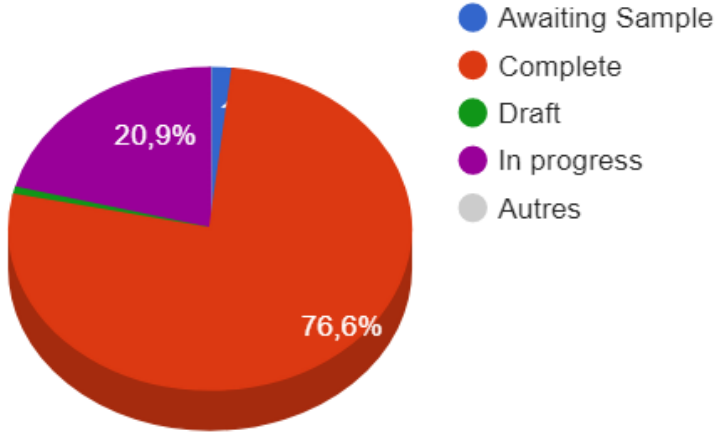
1405

Bacterial Sequencing Status



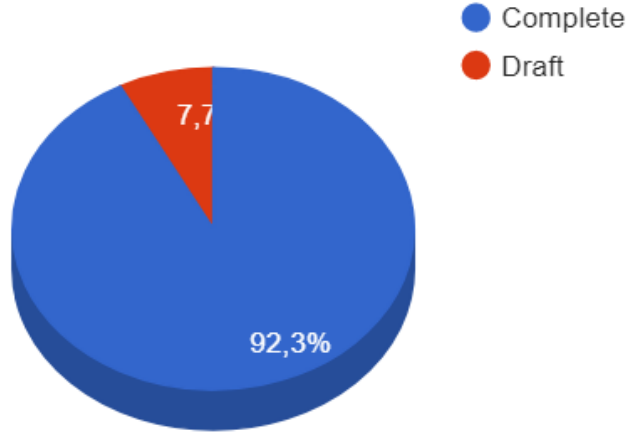
79571

Eukaryal Sequencing Status



22984

Metagenomic Sequencing Status



18126

g.deleage@ibcp.fr

Taille comparée de génomes séquencés

Organisme	Nb. chrom.	Nombre gènes	Taille Mb
Homo sapiens	23	20-30.000	3000
Mus musculus	21	30-45.000	3000
Arabidopsis thaliana	5	~20000	120
D. melanogaster	4	~ 14.000	165
C. elegans	6	~ 14.000	100
Saccharomyces cerevisiae	16	6000	13
Escherichia coli	1	4000	4,6

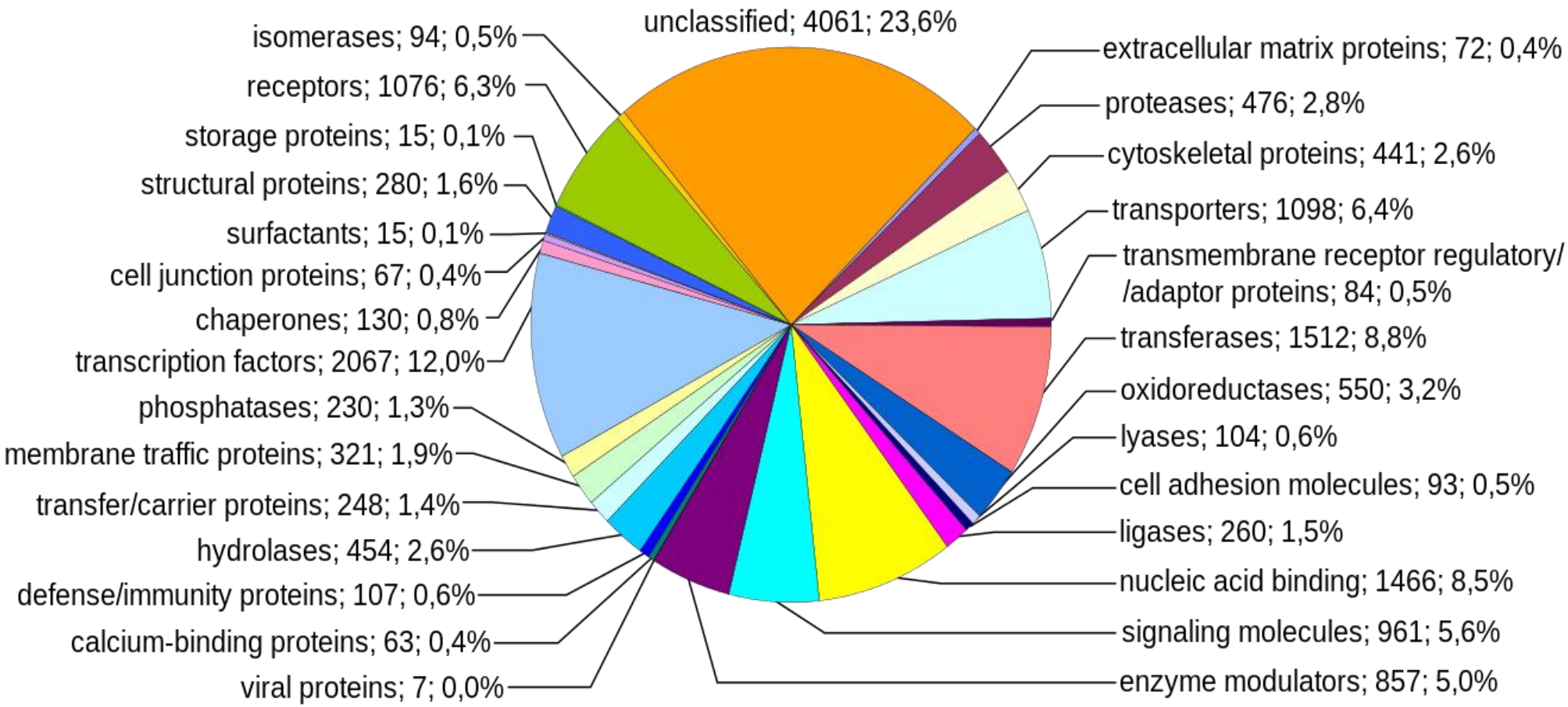
g.deleage@ibcp.fr

Répartition des ~ 20338 gènes humains (2017)

Chromosome	Base pairs	Confirmed proteins	Putative proteins	Pseudogenes
1	249,250,621	2012	31	1130
2	243,199,373	1203	50	948
3	198,022,430	104	25	719
4	191,154,276	718	39	698
5	180,915,260	849	24	676
6	171,115,067	1002	39	731
7	159,138,663	866	34	803
8	146,364,022	659	39	568
9	141,213,431	785	15	714
10	135,534,747	745	18	500
11	135,006,516	1258	48	775
12	133,851,895	1003	47	582
13	115,169,878	318	8	323
14	107,349,540	601	50	472
15	102,531,392	562	43	473
16	90,354,753	805	65	429
17	81,195,210	1158	44	300
18	78,077,248	268	20	59
19	59,128,983	1399	26	181
20	63,025,520	533	13	213
21	48,129,895	225	8	150
22	51,304,566	431	21	308
X	155,270,560	815	23	780
Y	59,373,566	45	8	327
mtDNA	16,569	13	0	0

g.deleage@ibcp.fr

Fonction des gènes humain (2014)



g.deleage@ibcp.fr

Fichier Edition Affichage Aller à Marque-pages Outils ?

http://www.ensembl.org/index.html

http://www.ensembl.org

OK

Ensembl release 40 - Aug 2006 Help

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Other links

- Home
- Sitemap
- Vega
- Pre Ensembl
- View previous release of page in Archive!
- Stable Archive! link for this page
- Archive! sites
- Trace server

2X GENOMES now in Ensembl!

What's New in Ensembl 40

- New low-coverage genomes (*L. africana*, *D. novemcinctus*, *E. telfairi*, *O. cuniculus*)
- Stickleback assembly and genebuild (*Gasterosteus aculeatus*)
- New species - Aedes aegypti (*Aedes aegypti*)
- New Macaque assembly and genebuild (*Macaca mulatta*)
- New genebuild on Rat assembly (*Rattus norvegicus*)

More news...

About Ensembl

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust.

This site provides free access to all the data and software from the Ensembl project. Click on a species name to browse the data.

Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints. Some data and software may be subject to third-party constraints.

For all enquiries, please contact the Ensembl HelpDesk (helpdesk@ensembl.org).

Other sites using the Ensembl system

- EBI Genome Reviews database - mainly archaea and bacteria.
- VEGA - Vertebrate Genome Annotation

Mammalian genomes


Homo sapiens
NCBI 36 | Vega


Pan troglodytes
PanTro 1.0 | **NEW!** *pre!*


Macaca mulatta
UPDATED! MMUL 1.0


Mus musculus
NCBI m36 | Vega


Rattus norvegicus
UPDATED! RGSC 3.4


Oryctolagus cuniculus
NEW! RABBIT


Canis familiaris
CanFam 1.0 | Vega | **UPDATED!** *pre!*


Bos taurus
Btau 2.0

Sus scrofa
NEW! (clone status map)


Dasypus novemcinctus
NEW! ARMA


Loxodonta africana
NEW! BROAD E1


Echinops telfairi
NEW! TENREC


Monodelphis domestica
MonDom 4


Ornithorhynchus anatinus

Other species


Gallus gallus
WASHUC 1


Xenopus tropicalis
JGI 4.1


Danio rerio
Zv 6 | Vega


Takifugu rubripes
FUGU 4.0


Tetraodon nigroviridis
TETRAODON 7


Gasterosteus aculeatus
NEW! BROAD S1

Oryzias latipes
MEDAKA 1


Ciona intestinalis
JGI2


Ciona savignyi
CSAV 2.0


Drosophila melanogaster
UPDATED! BDGP 4


Anopheles gambiae
AgamP3


Aedes aegypti
NEW! AaegL 1


Caenorhabditis elegans
WS 150


Saccharomyces cerevisiae

Terminé

Génome humain sur Ensembl

Human (Homo sapiens) - Mozilla Firefox

Fichier Edition Affichage Aller à Marque-pages Outils ?

http://www.ensembl.org/Homo_sapiens/index.html

e! Ensembl Human

Search e! Human: Anything Go

e.g. AL138722.15.1.44776, ENSG00000139618

Ensembl release 40 - Aug 2006 Help

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload and view data on chromosome
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Select a species

- Mammals
- Other chordates
- Other eukaryotes

Other links

- Home

Explore the *Homo sapiens* genome

Karyotype

Click on a chromosome for a closer view

Jump directly to sequence position

Chromosome: == or region

From (bp):

About the Human genome

Assembly

This release is based on the NCBI 36 assembly of the human genome [November 2005].

The International Human Genome Sequencing Consortium have published their scientific analysis of the finished human genome.

- Nature 431, 931 - 945 (21 October 2004)
- WT Sanger Institute Press Release

Annotation

The human genome sequence is now considered sufficiently stable that the three major genome browsers have come together to produce a common set of gene IDs for their annotations. This Consensus CDS ID set has been encorporated into the Ensembl database alongside the existing identifiers.

- More information about the CCDS project.

The ENCODE (ENCyclopedia Of DNA Elements) project aims to find functional elements in the human genome.

Terminé

g.deleage@ibcp.fr

Ensembl v40: Homo sapiens Overview of Chromosome 9 - Mozilla Firefox

Fichier Edition Affichage Aller à Marque-pages Outils ?

http://www.ensembl.org/Homo_sapiens/mapview?chr=9

OK

e! Ensembl Human MapView

Search e!Human: Anything Go

e.g. 1, MT

Ensembl release 40 - Aug 2006 Help

Chromosome 9

- View Chromosome 9
- View Chr 9 Synteny
- Map your data onto this chromosome

Use Ensembl to...

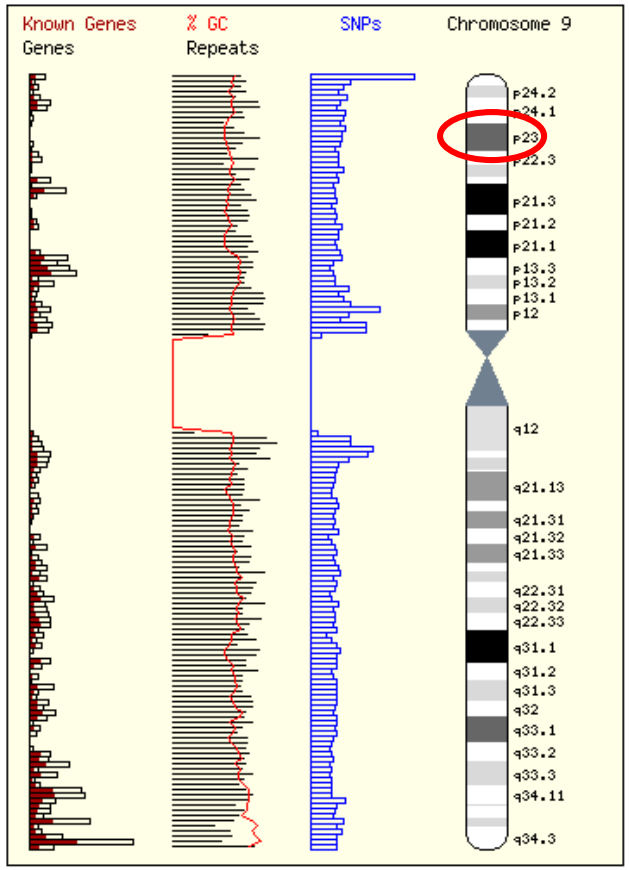
- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload and view data on chromosome
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Other links

- Home
- Sitemap
- Vega
- Pre Ensembl
- View previous release of page in Archive!
- Stable Archive! link for



Click on the image above to zoom into that point

Chromosome 9

Length:	140,273,252	bps
Known Protein-coding Genes:	926	
Novel Protein-coding Genes:	1,013	
Pseudogene Genes:	26	
miRNA Genes:	25	
rRNA Genes:	11	
snRNA Genes:	43	
snoRNA Genes:	15	
Misc RNA Genes:	47	
SNPs:	559,768	

For more information on gene statistics, see the MapView help page

Change Chromosome

Chromosome

9

Fields marked with * are required

g.deleage@ibcp.fr

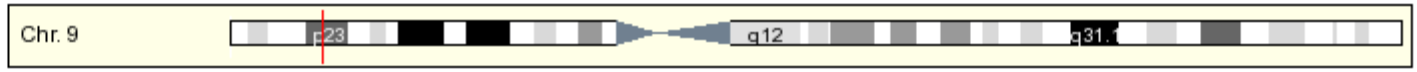
Chromosome 9
11,040,954 - 11,041,953

- View of Chromosome 9
- Graphical view
- Graphical overview
- Export information about region
- Export sequence as FASTA
- Export EMBL file
- Export Gene info in region
- Export SNP info in region
- Export Vega info in region
- View alignment with ...
- View alongside ...
- View Syntenic regions ...
- View region in NCBI browser
- View region in UCSC browser

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload and view data on chromosome

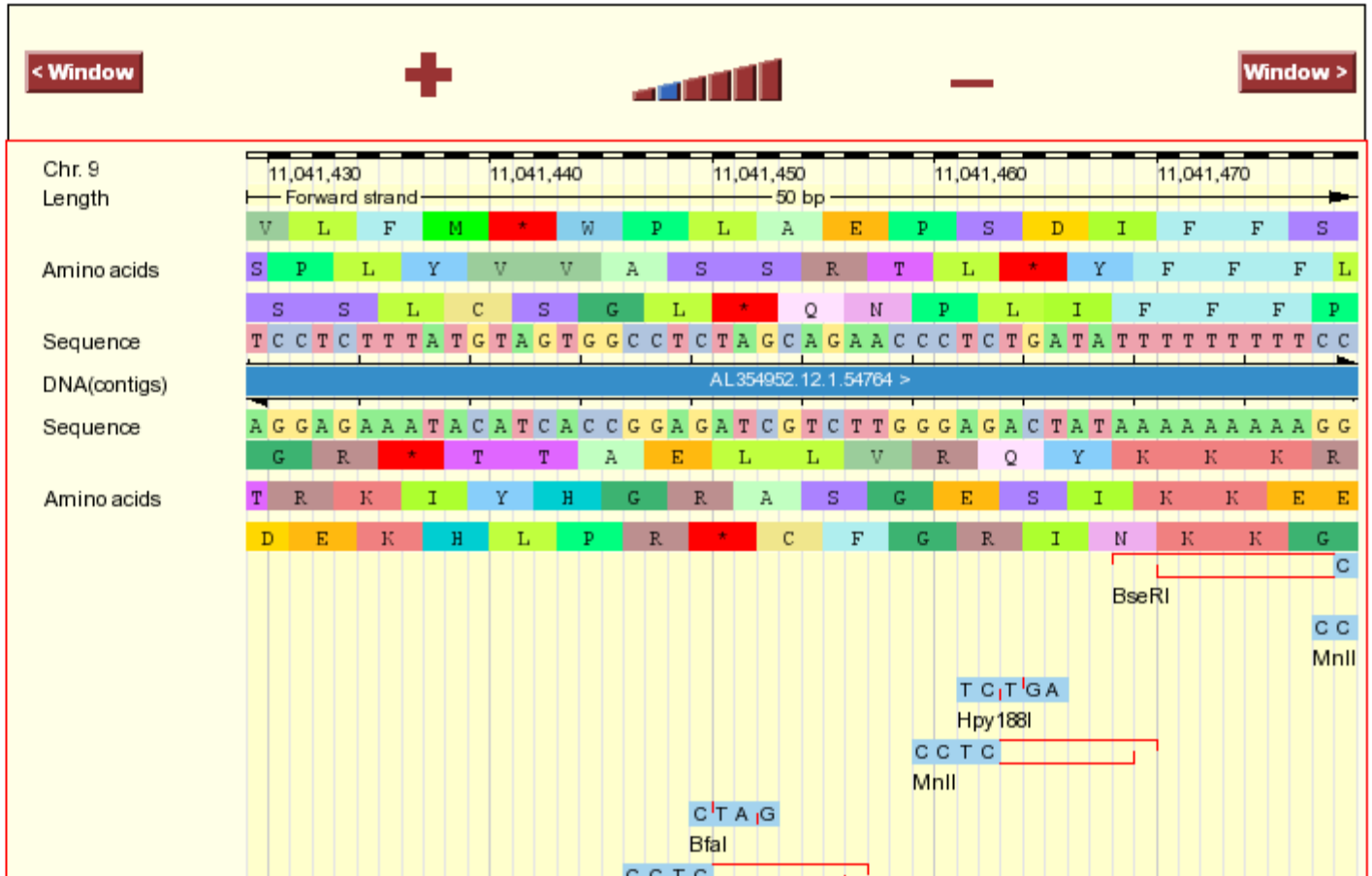
Chromosome 9



Overview

Detailed view

Basepair view



Analyse de séquences

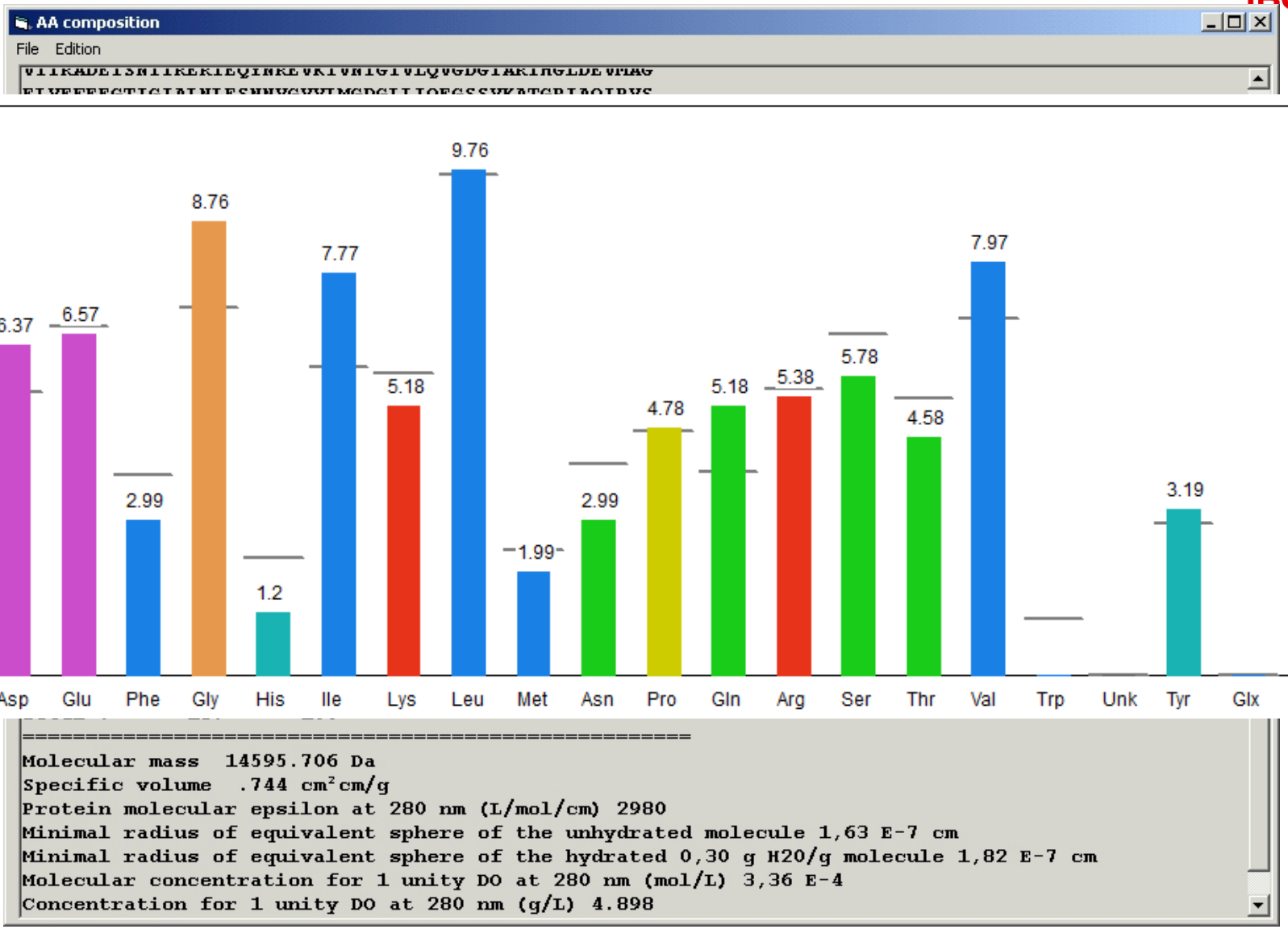
```
>P37808 (ATPA_BACSU) ATP SYNTHASE SUBUNIT ALPHA - BACILLUS SUBTILIS
MSIKAEEISTLIKQQIQNYQSDIEVQDVGTVIQVGDGIARVHGLDNCMAGELVEFSNGVLGMAQNLEESNVGIVILGPFS
EIREGDEVKRTGRIMEVPVGEELIGRIVNPLGQPVDGLGPILTSKTRPIESPAPGVMDRKSVHEPLQTGIKAIDALIPIG
RGQRELIIGDRQTGKTSVAIDAILNQKDQDMICVYVAIGQKESTVRGVVETLRKHGALDYTIVVTASASQPAPLLYLAPY
AGVTMAEEFMYNGKHVLVVYDDLSKQAAAYRELSLLLRRPPGREAFPGDVFYLHSRLLERAAKLSDAKGAGSITALPFVE
TQAGDISAYIPTNVISITDGQIFLQSDLFFSGVRPAINAGLSVSRVGGSAQIKAMKKVSGTLRLDLASYRELEAFAQFGS
DLDQATQAKLNRGARTVEVLKQDLNKPLPVEKQVAILYALTKGYLDDIPVADIRRFEEEYYMYLDQNHKDLLDGIAKTGN
LPADEDFKAAIEGFKRTFAPSN
```

PC Windows ANTHEPROT (http://antheprot-pbil.ibcp.fr)

Web NPSA: http://npsa-prabi.ibcp.fr

g.deleage@ibcp.fr

Composition en aa



- La composition est-elle « standard » ou fortement biaisée?
- Attention à la dérive isotopique (1.1% de ^{13}C pour le calcul des masses en Spectrométrie de Masse

Faible complexité (cauchemar du bioinformaticien)

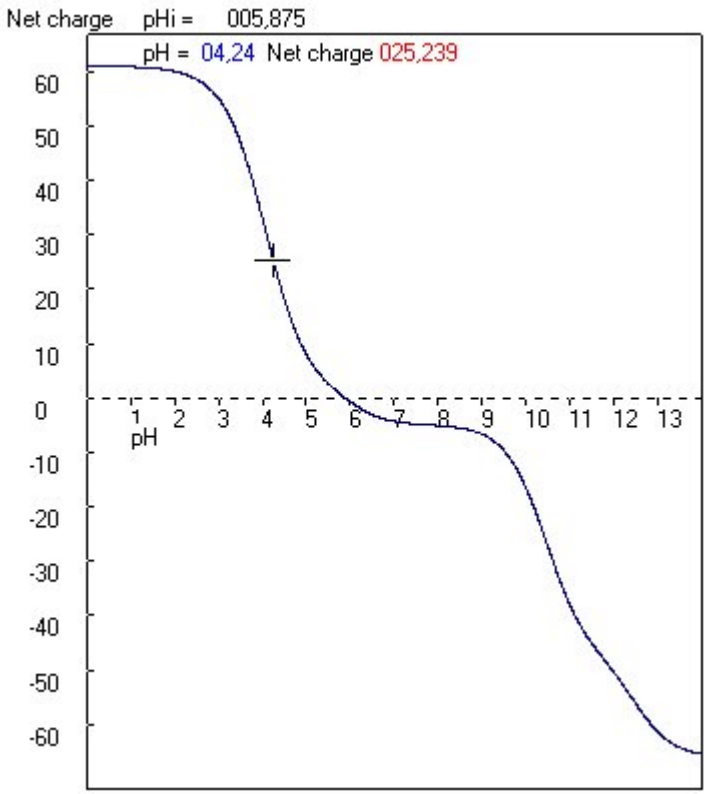
g.deleage@ibcp.fr

Courbes de titration et pI

- Définir les pKai (aa « acide ») et pKaj (aa « basique ») des chaînes latérales des acides aminés
  - pKj 1 His 6.00 pKj 2 Lys 10.53 pKj 3 Arg 12.48 pKj 4 Nter 9.80
  - pKi 5 Asp 3.86 pKi 6 Glu 4.20 pKj 7 Cys 8.33 pKj 8 Tyr 10.1
  - pKj 9 Ser 13.60 pKj 10 Thr 13.60 pKi11 Cter 2.10

$$\text{Net Charge} = \sum_i N_i \left(1 - \frac{10^{-pKai}}{10^{-pH} + 10^{-pKai}}\right) - \sum_j N_j \left(\frac{10^{-pKaj}}{10^{-pH} + 10^{-pKaj}}\right)$$

Titration



Purification

Choix des colonnes d'échanges d'ions et du pH de purification.

Solubilité minimale au pI

g.deleage@ibcp.fr

Comparaison de séquences

Comparaison de 2 séquences

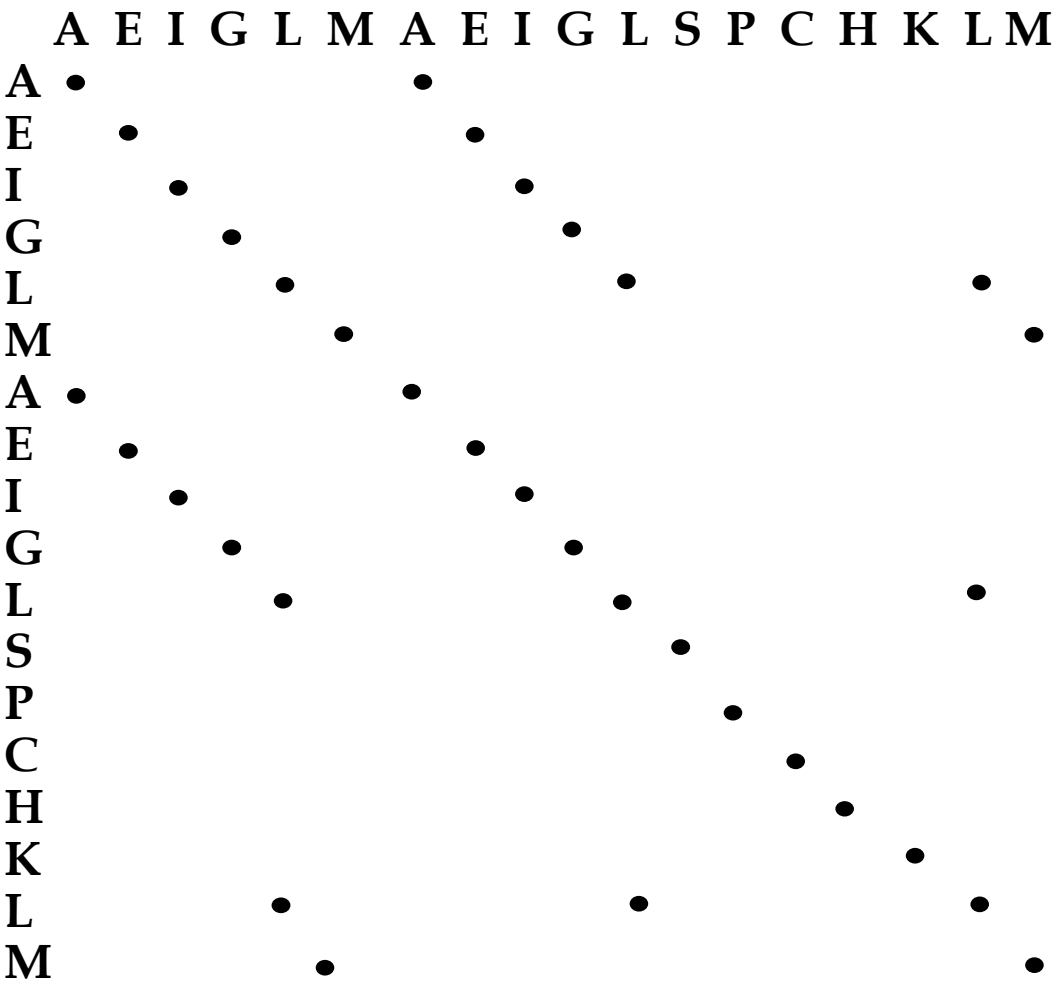
- Objectifs
  - Répétitions internes sur une séquence
  - Duplication de gènes
  - Régions de faible complexité
  - Identification de protéines homologues
  - Identification de sites similaires dans des protéines différentes
  - Zones d'insertion - délétion entre 2 séquences

- Matrice de points
  - Méthode simple - identité
    - Méthode exhaustive non ambigüe
  - Méthodes à score (filtrage pour les acides nucléiques)
    - Paramétrage du filtre (longueur du segment à choisir)
  - Ressemblance par matrice de substitution
    - Choix de la matrice
    - Paramétrage délicat
    - Comparaison des matrices difficiles

g.deleage@ibcp.fr

Comparaison de séquences

Matrices de points simple - Identité

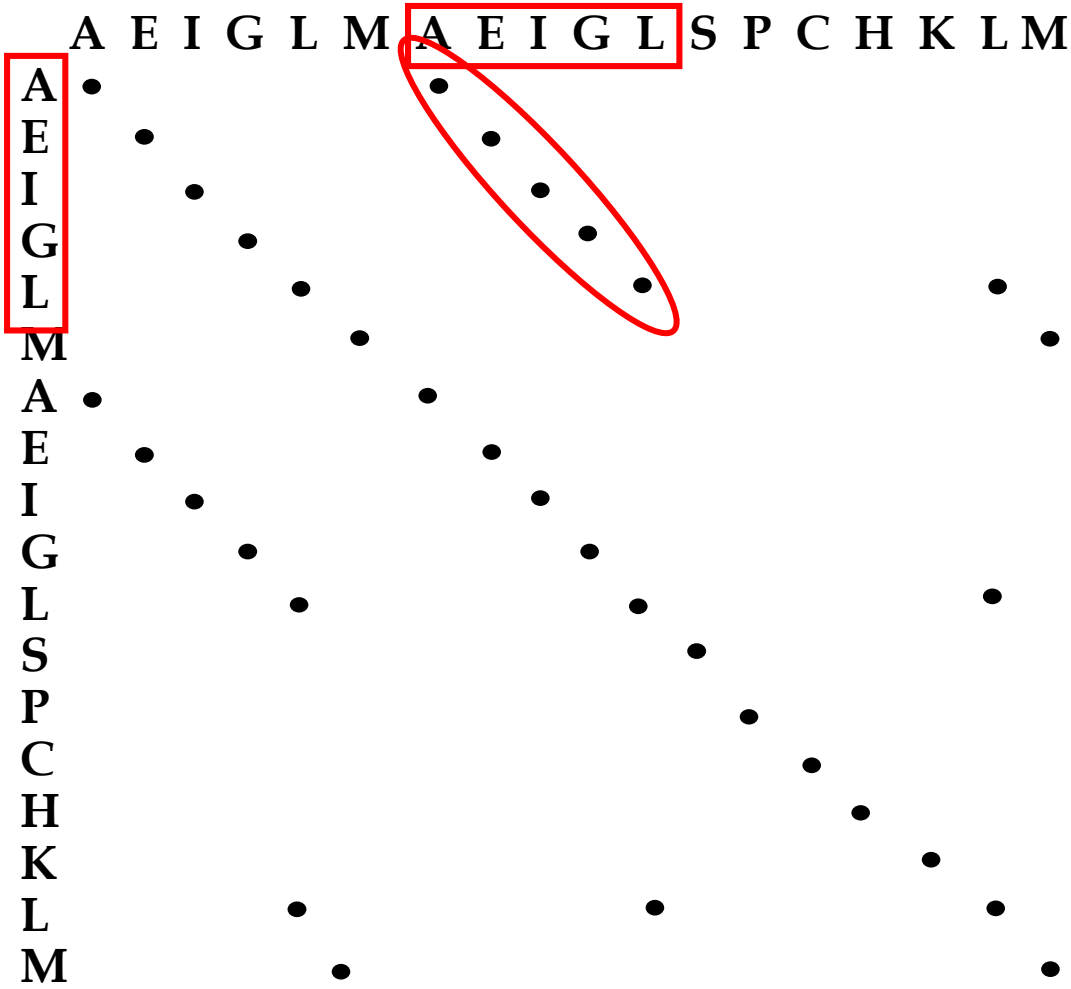


g.deleage@ibcp.fr

Comparaison de séquences

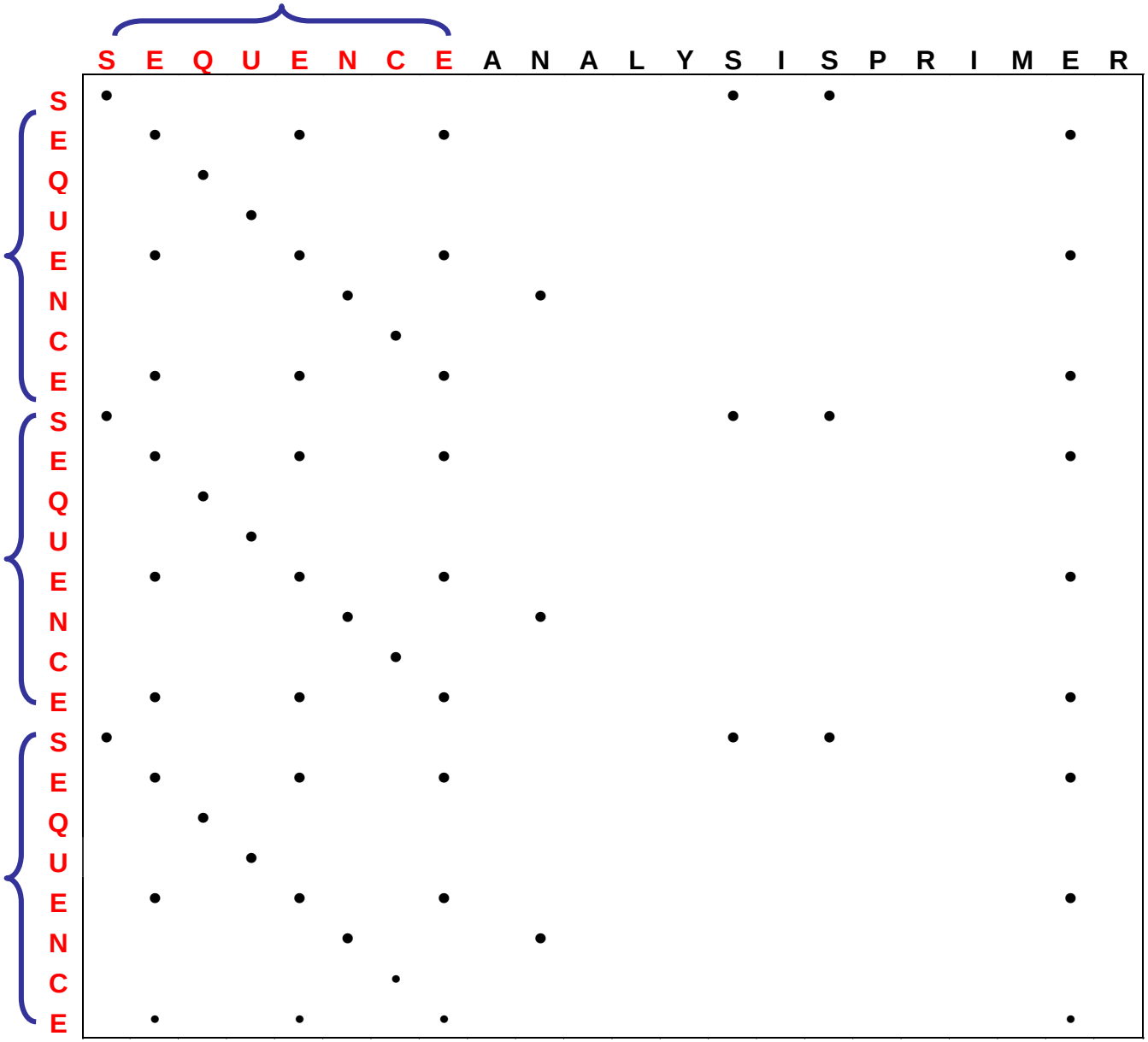
Matrices de points simple - Identité

Répétition interne AEIGL



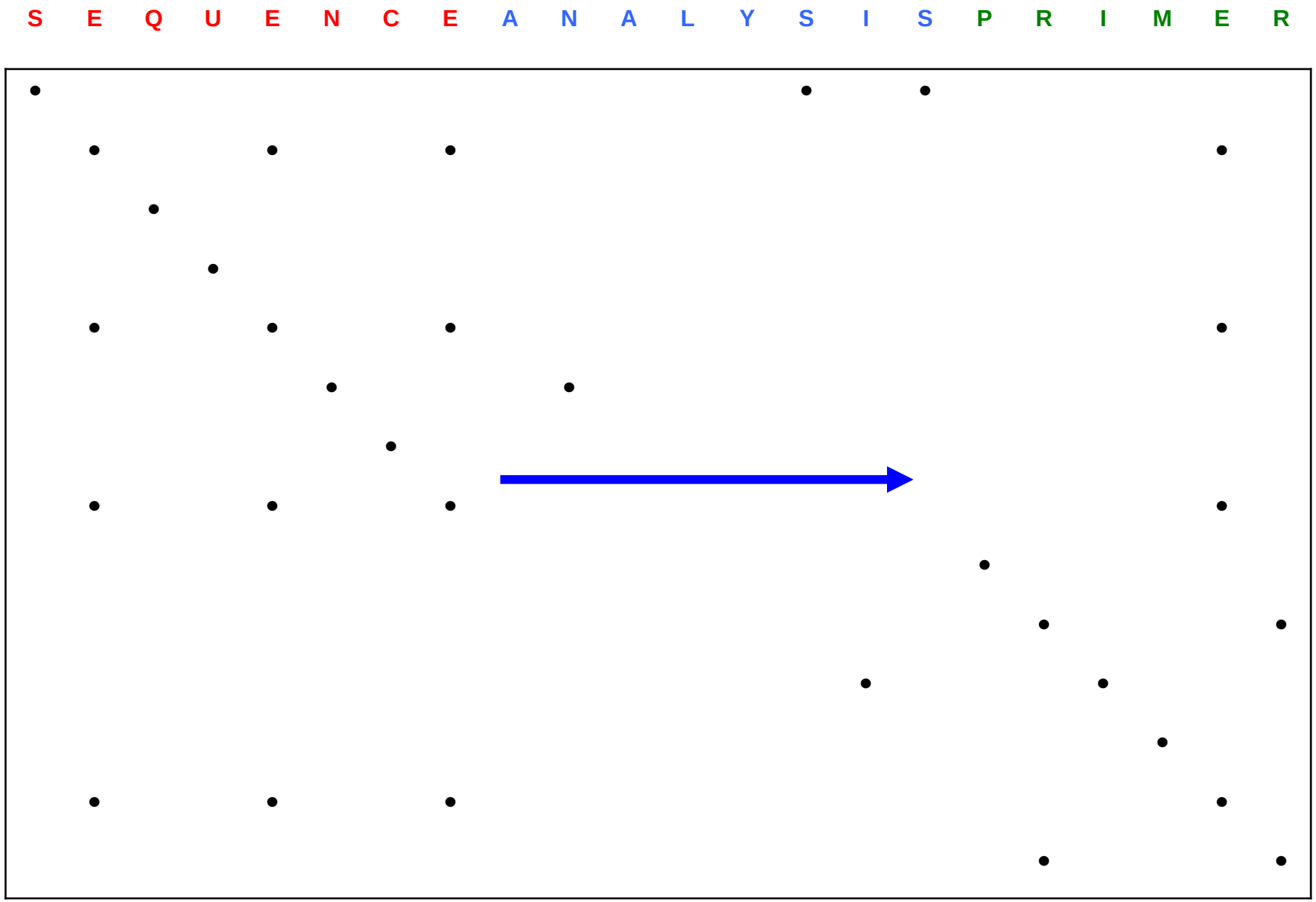
g.deleage@ibcp.fr

Répétitions internes



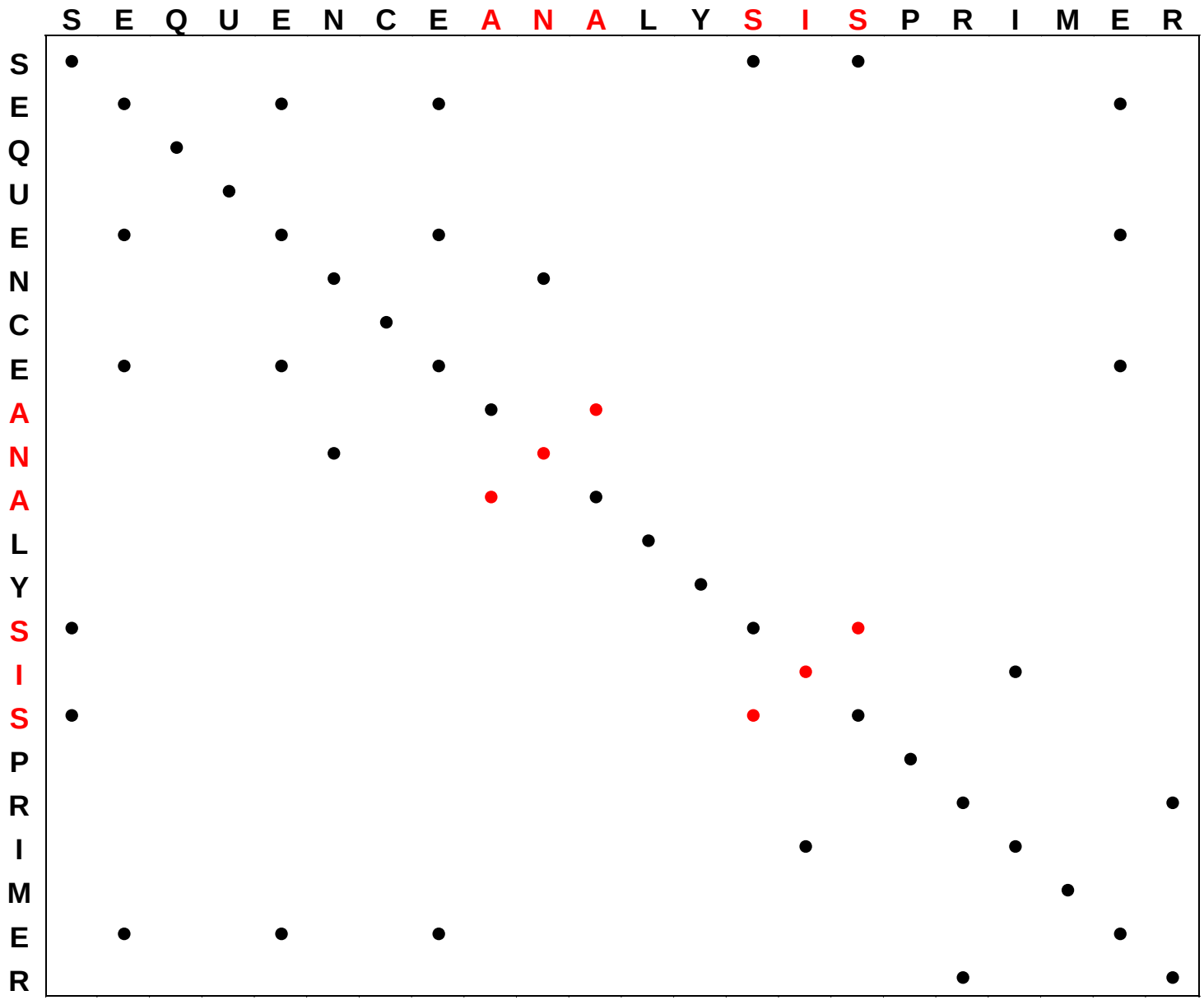
g.deleage@ibcp.fr

Insertion de « ANALYSIS » dans A



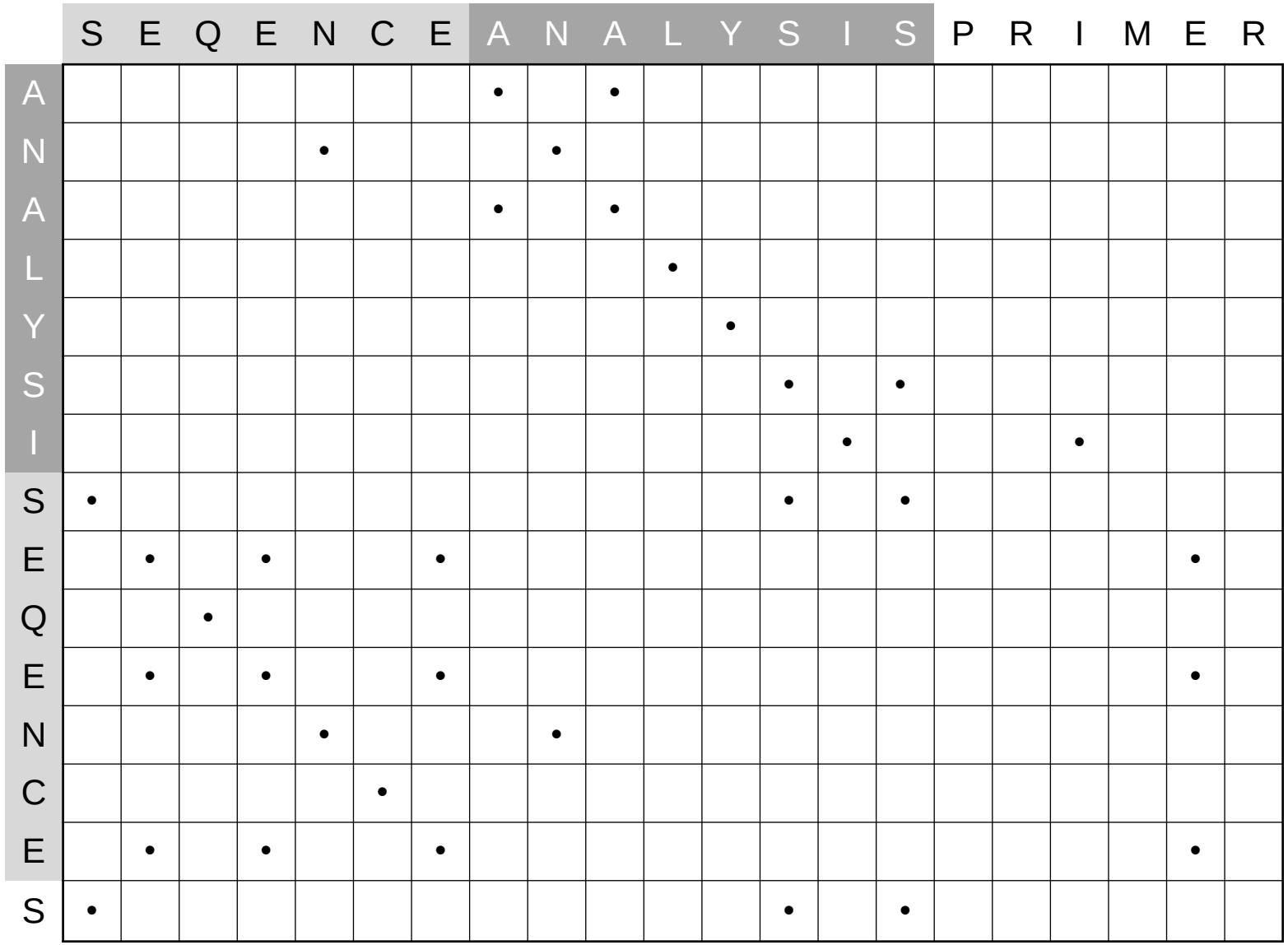
g.deleage@ibcp.fr

Palindrome= croix perpendiculaires à la diagonale (ANA et SIS)



g.deleage@ibcp.fr

Transpositions (Inversion ANALYSIS et SEQUENCE)

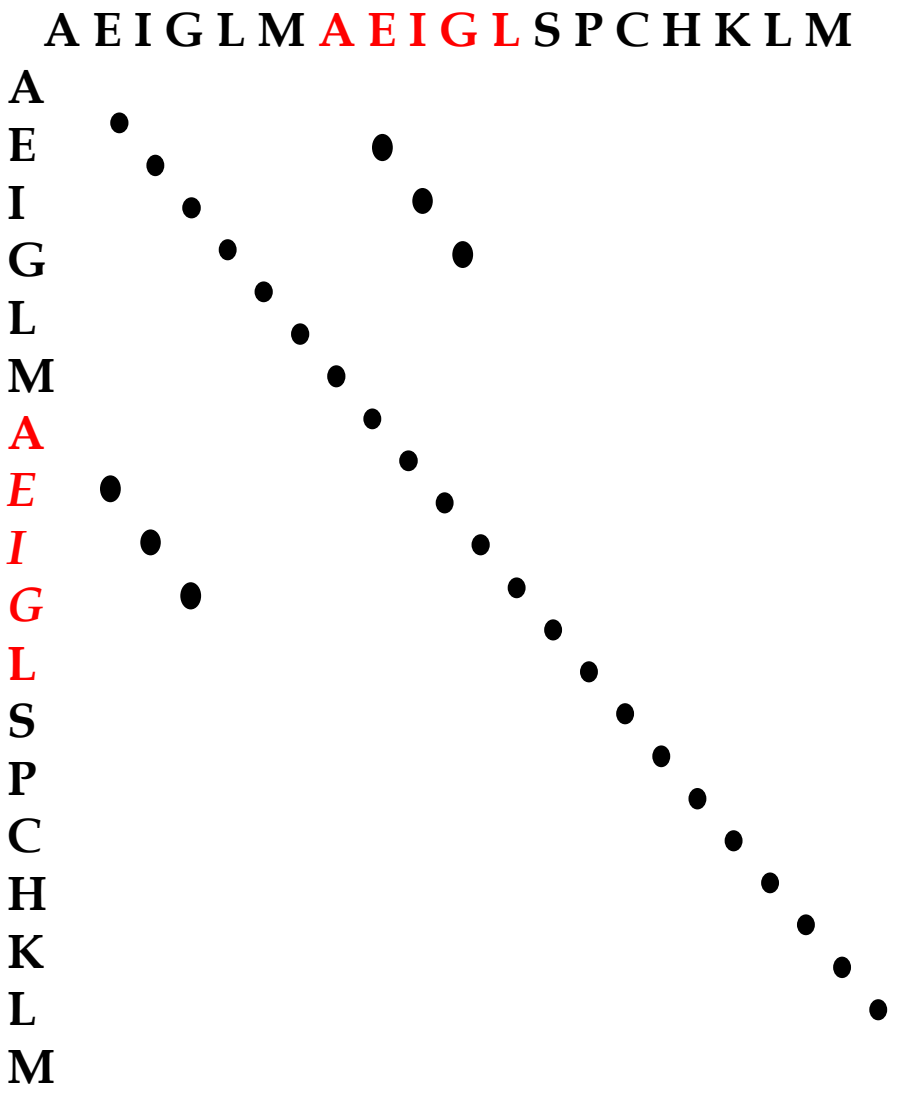


g.deleage@ibcp.fr

Matrices de points simple - Filtrage

- Plutôt que mettre un point par AA identique , le filtrage consiste à mettre un point par segment de 3AA identiques

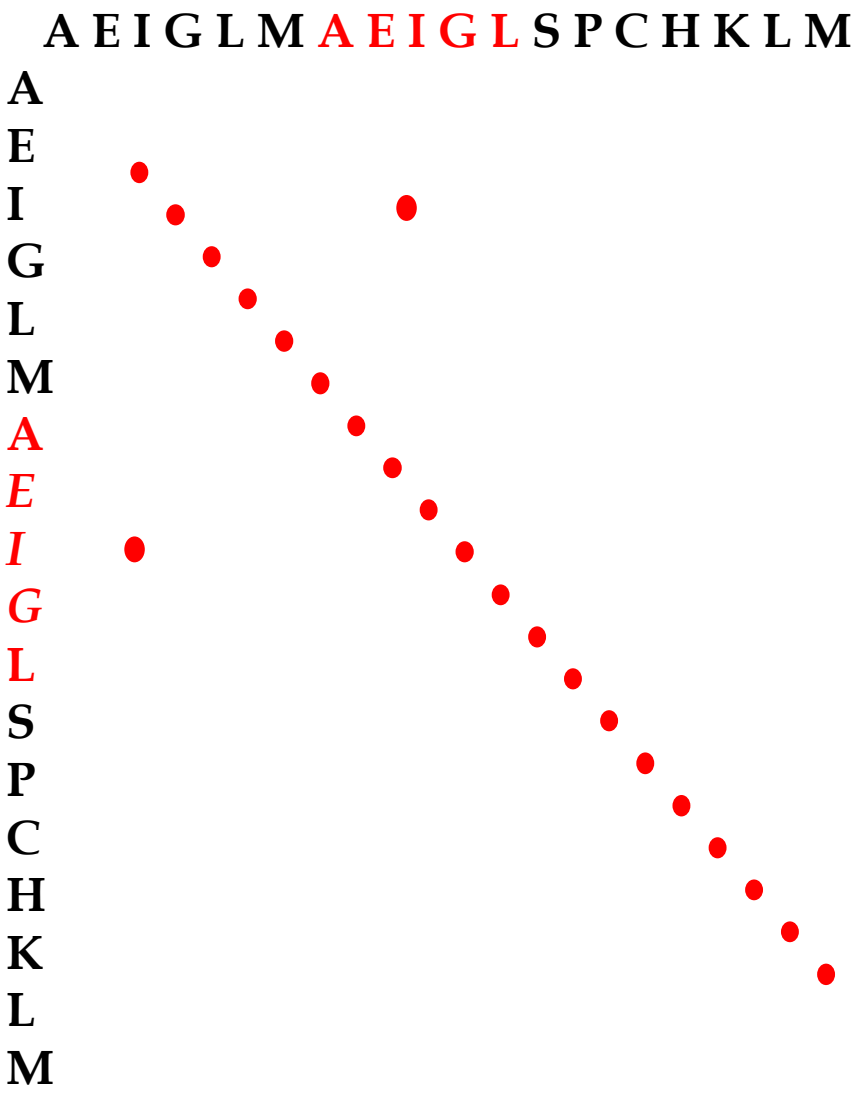
  - **1 point : 3AA**
  - **2 points: 4AA**
  - **3 points: 5 AA**



g.deleage@ibcp.fr

Matrices de points simple - Filtrage

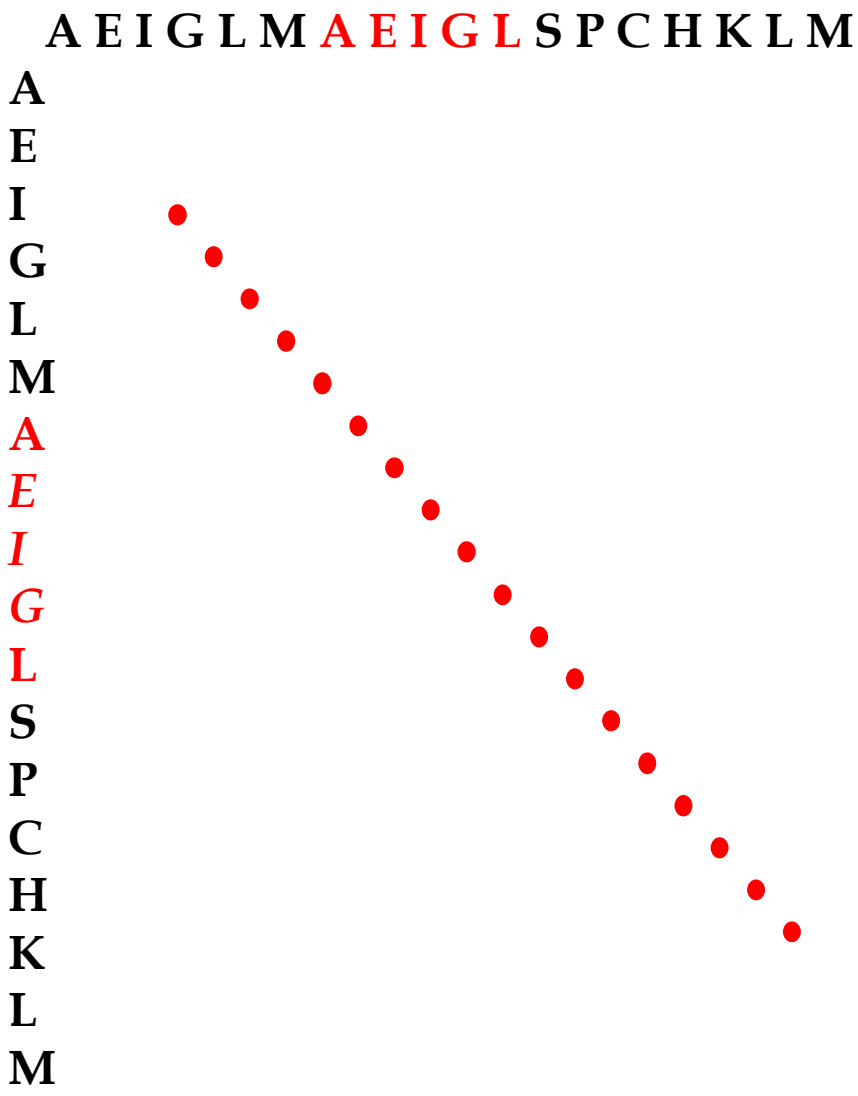
- **Un point par segment de 5 AA identiques**
  - 1 point : 5AA
  - 2 points: 6AA
  - 3 points: 7 AA



g.deleage@ibcp.fr

Matrices de points simple - Filtrage

- **Un point par segment de 7 AA identiques**
  - 1 point : 7AA
  - 2 points: 8AA
  - 3 points: 9 AA



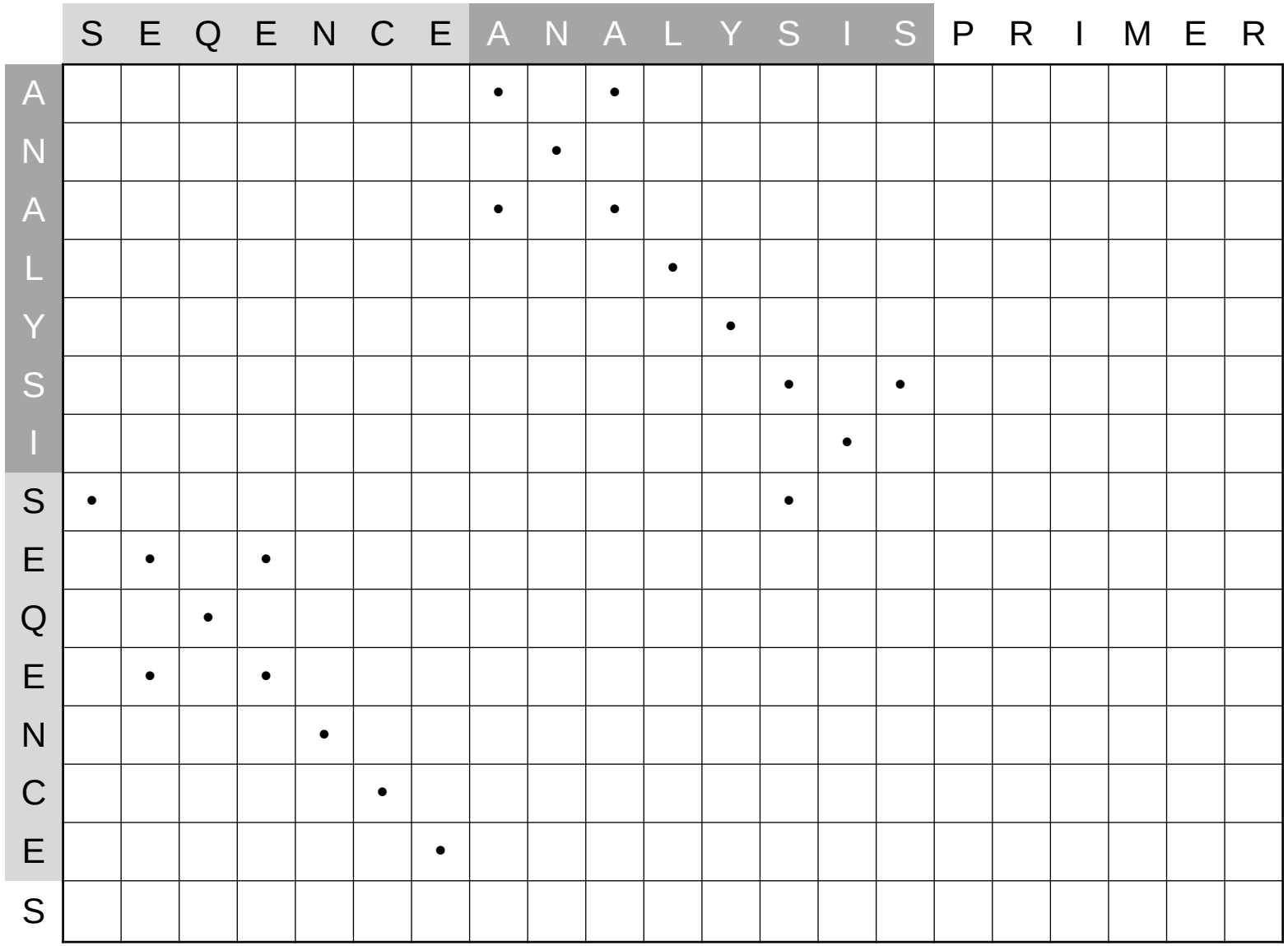
g.deleage@ibcp.fr

Avantage du filtrage Transpositions (Inversion ANALYSIS et SEQUENCE)

	S	E	Q	E	N	C	E	A	N	A	L	Y	S	I	S	P	R	I	M	E	R
A								•		•											
N					•				•												
A								•		•											
L											•										
Y												•									
S	•												•		•						
I														•				•			
S	•												•		•						
E		•		•			•													•	
Q			•																		
E		•		•			•													•	
N					•				•												
C						•															
E		•		•			•													•	
S	•												•		•						

g.deleage@ibcp.fr

Transpositions filtrées (Inversion ANALYSIS et SEQUENCE)



1 point= 2 identités sur 3

g.deleage@ibcp.fr

Passage à la similarité

Peptide 1 A E I G L M A E I G L S E K I L

Peptide 2 L D V A A I G D L A I T Q R L M

Peptide 3 W R G I Y S H H D E T W D C P C

Ces 3 peptides n'ont aucune identité et pourtant un biochimiste
saurait dire que le peptide 1 est plus proche du peptide 2 que du peptide 3

g.deleage@ibcp.fr

Utilisation de l'évolution

Les protéines évoluent via des successions de **mutations ponctuelles indépendantes** les unes des autres et **acceptées** dans la population.

- Les matrices liées à l'évolution : matrices PAM
  - représentent les échanges possibles et acceptables d'un acide aminé par un autre lors de l'évolution des protéines (M. Dayhoff, 1978).

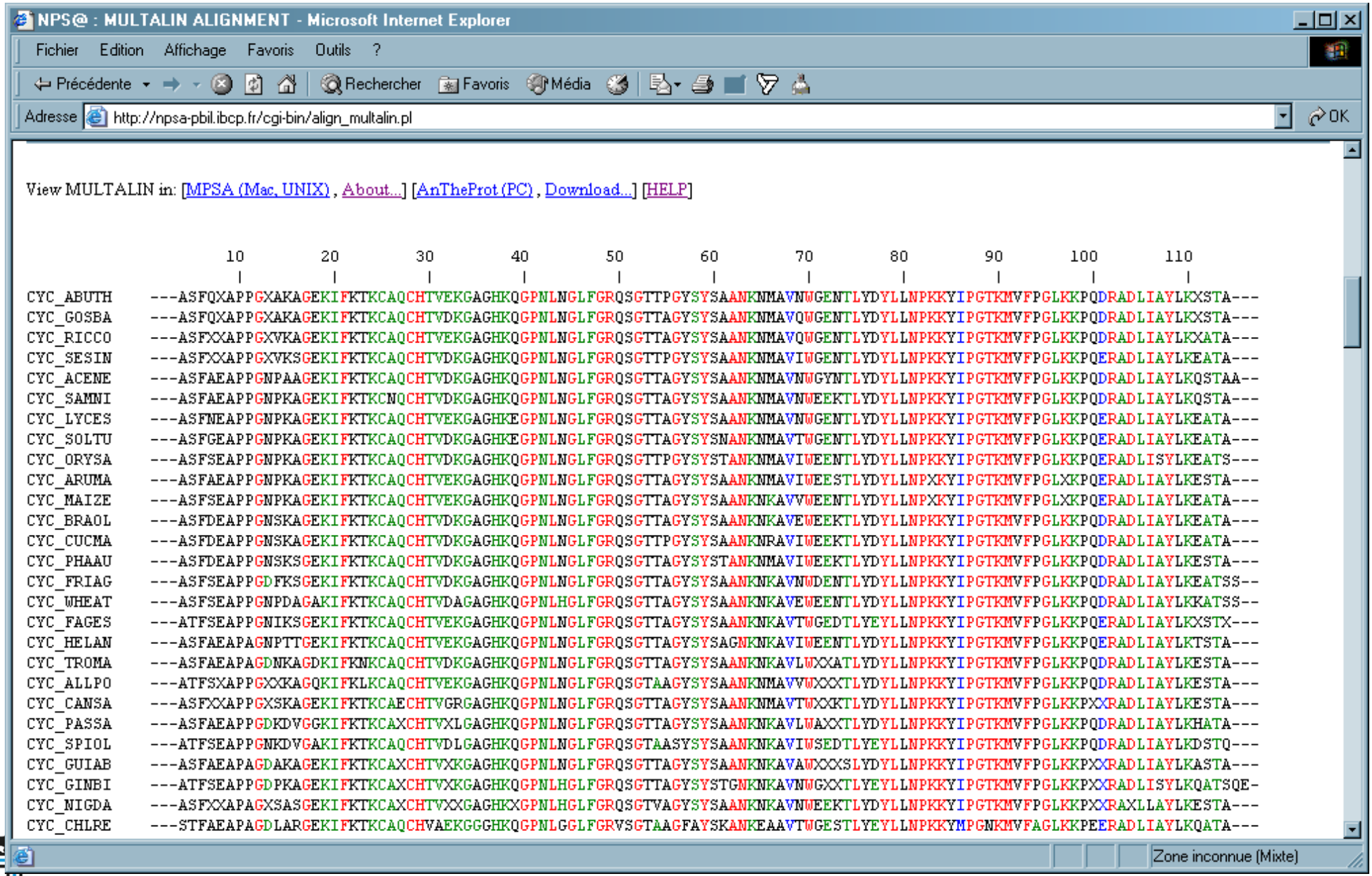
- si deux séquences appartiennent au même processus évolutif, et qu'un acide aminé de l'une a été muté pour donner l'autre, alors les deux acides aminés sont « similaires » :

  - les mutations sont dites acceptées (**P**oint **A**ccepted **M**utation)
  - elles ont été conservées au cours de l'évolution de façon à ne pas altérer la fonction de la protéine.

- Exemple les cytochromes

g.deleage@ibcp.fr

Fréquence de mutations pour une distance évolutive de 1 PAM



g.deleage@ibcp.fr

Calcul de la matrice de substitution

Alignement

Comptages

Matrice de probabilité de mutation PAM 1

Matrice « odds »

Matrice « log-odds »

```
.I D N F K N .
.I D D W K N .
```

	I	D	N	F	W	K
Nombre de changements	0	1	1	1	1	0
Nombre d'ocurences	2	3	3	1	1	2
Mutabilité m	**0**	**1/3**	**1/3**	**1**	**1**	**0**

$$M_{ij}^{1} = m_j \frac{A_{ij}}{\sum_{i=1}^{20} A_{ij}}$$

A_{ij} : nombre de mutations observées $i \longrightarrow j$

$$R_{ij}^{1} = \frac{M_{ij}^{1}}{p_i}$$

p_i : Fréquence d'occurence i

$$S_{ij}^{1} = \log R_{ij}^{1}$$

$$R_{ij}^{k} = \left(R_{ij}\right)^{k} \quad ou \quad S_{ij}^{k} = \sum_{k=1}^{n} S_{ij}^{k}$$

Simulation par k multiplications de la matrice par elle-même

g.deleage@ibcp.fr

Matrice PAM 250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0	0	0
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1	0	-1
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	1	0
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	3	3	-1
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-4	-5	-3
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	3	-1
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	3	-1
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	0	-1
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	1	2	-1
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2	-1
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	-3	-1
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	0	-1
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	-2	-1
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5	-2
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-1	0	-1
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0	0	0
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1	0
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-6	-4
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-3	-4	-2
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	-2	-1
B	0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	0	-5	-3	-2	3	2	-1
Z	0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	-1
X	0	-1	0	-1	-3	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	0	0	-4	-2	-1	-1	-1	-1

Plus la valeur est grande (>0) plus la conservation des AA est forte

Plus la valeur est petite (<O) plus la conservation des AA est faible

g.deleage@ibcp.fr

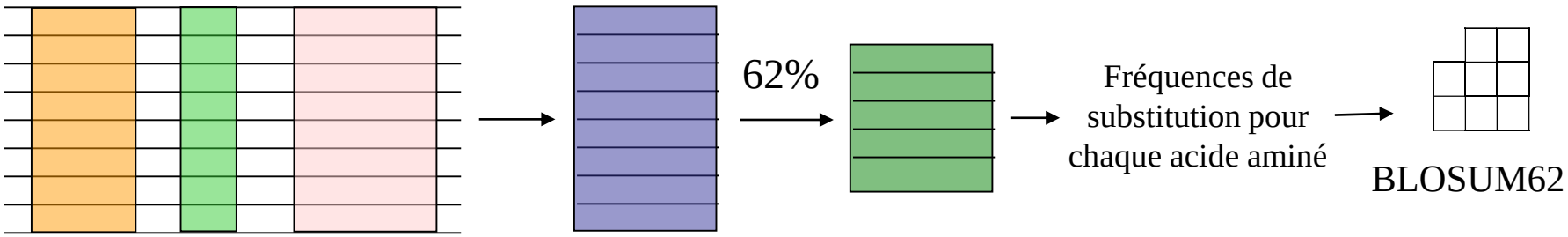
Matrices de substitution **BLOSUM** (*BLOcks Substitution Matrix*)

• observation de blocs d'acides aminés issus de protéines relativement éloignées
• chaque bloc provient d'alignements multiples sans insertions / délétions de courtes régions conservées
• les blocs sont utilisés pour regrouper tous les segments de séquences ayant un pourcentage d'identité minimum au sein de leur bloc

=> fréquences de substitution pour chaque paire d'acides aminés

=> calcul d'une matrice logarithmique de probabilité

à chaque pourcentage d'identité correspond une matrice :
BLOSUM50 avec un seuil d'identité de 50 % ;
BLOSUM62 avec un seuil d'identité de 62 %.



g.deleage@ibcp.fr

Matrice BLOSUM 62

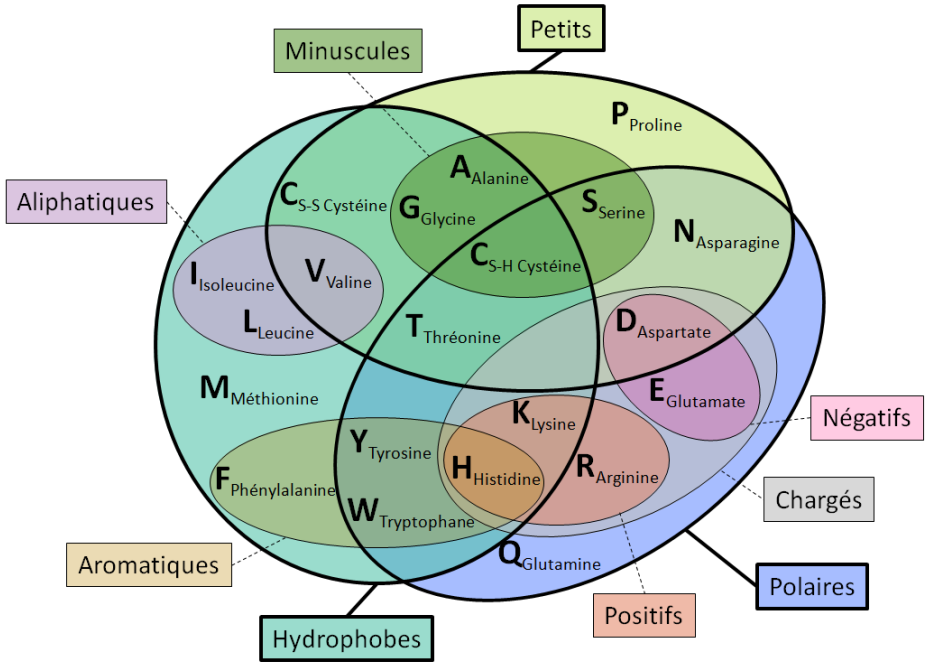
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1

g.deleage@ibcp.fr

Passage à la similarité avec une matrice PAM

Peptide 1	A	E	I	G	L	M	A	E	I	G	L	S	E	K	I	L
Peptide 2	L	D	V	A	A	I	G	D	L	A	I	T	Q	R	L	M
Peptide 3	W	R	G	I	Y	S	H	H	D	E	T	W	D	C	P	C

Ces 3 peptides n'ont aucune identité et pourtant un biochimiste saurait dire que le peptide 2 est plus proche du peptide 1 que du peptide 3



g.deleage@ibcp.fr

Passage à la similarité avec une matrice PAM 250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0	0	0
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1	0	-1
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	1	0
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	3	3	-1
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-4	-5	-3
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	3	-1
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	3	-1
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	0	-1
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	1	2	-1
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2	-1
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	-3	-1
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	0	-1
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	-2	-1
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5	-2
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-1	0	-1
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0	0	0
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1	0
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-6	-4
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-3	-4	-2
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	-2	-1
B	0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	0	-5	-3	-2	3	2	-1
Z	0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	-1
X	0	-1	0	-1	-3	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	0	0	-4	-2	-1	-1	-1	-1

Peptide 1 A E I G L M A E I G L S E K I L

-2 3 4 1 -2 2 1 3 2 1 2 1 2 3 2 4 = 27/16 **1,69**

Peptide 2 L D V A A I G D L A I T Q R L M

g.deleage@ibcp.fr

Passage à la similarité avec une matrice PAM 250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0	0	0
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1	0	-1
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	1	0
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	3	3	-1
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-4	-5	-3
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	3	-1
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	3	-1
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	0	-1
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	1	2	-1
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2	-1
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	-3	-1
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	0	-1
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	-2	-1
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5	-2
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-1	0	-1
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0	0	0
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1	0
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-6	-4
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-3	-4	-2
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	-2	-1
B	0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	0	-5	-3	-2	3	2	-1
Z	0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	-1
X	0	-1	0	-1	-3	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	0	0	-4	-2	-1	-1	-1	-1

Peptide 2 L D V A A I G D L A I T Q R L M

-2 -1 -1 -1 -3 -1 -2 1 -4 0 0 -5 2 -4 -3 -5 = -29/16 -1,82

Peptide 3 W R G I Y S H H D E T W D C P C

g.deleage@ibcp.fr

Matrices de substitution

- **Matrices nucléiques**
  - **Matrice unitaire**
  - **Matrice génétique (purine-pyrimidine)**
- **Matrices protéiques**
  - **Unitaire**
  - **Evolution**
    - **Matrice PAM (1 Point Accepted Mutation pour 100 aa), PAM 50, PAM 100, PAM 250**
      - **71 familles de protéines alignées (>85% identité) =>1300 séquences Dayhoff (1978)**
      - **Equiprobabilité des positions mutationnelles**
      - **Représentativité des séquences**
      - **Mesurée sur 6 à 15 PAM extrapolée jusqu'à 250PAM**
      - **Mise à jour en 1992 2600 familles soit 16130 séquences**
    - **Matrices BLOSUM (BLOcks Substitution Matrix, Henikoff, 1992)**
      - **2000 Blocs de séquences alignés sans insertion**
      - **Matrices à 62% d 'identité de séquences (BLOSUM 62, BLOSUM xx)**
  - **Matrices physico-chimiques**
    - **Matrices d'hydrophobie**
    - **Matrice de structures secondaires (Levin *et al*, 1986)**
    - **Matrices structurales de superposition de 32 structures (Risler *et al*,)**

g.deleage@ibcp.fr

Relation matrices de substitution et conservation des séquences

Faible divergence — **Forte divergence**

PAM50 — **PAM 100** — **PAM 120** — **PAM 250**

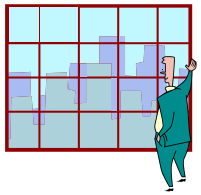
80% Id — 60% Id — 40% Id — 20% Id

BLOSUM 80 — **BLOSUM 62** — **BLOSUM 30**

g.deleage@ibcp.fr

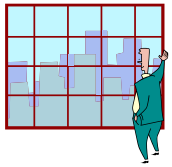
Matrice unitaire

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
R	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
N	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
D	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
C	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Q	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
E	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
G	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
H	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
I	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
L	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
K	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
M	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
F	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
B	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Z	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1



g.deleage@ibcp.fr

Matrice de structure secondaire



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	0	0	0	0	0	1	0	0	0	0	0	0	0	-1	1	0	-1	-1	0
R	0	2	0	0	0	0	0	0	0	-1	-1	1	-1	-1	0	0	0	-1	-1	-1
N	0	0	3	1	0	1	0	0	0	-1	-1	1	-1	-1	0	0	0	-1	-1	-1
D	0	0	1	2	0	0	1	0	0	-1	-1	0	-1	-1	0	0	0	-1	-1	-1
C	0	0	0	0	2	0	0	0	0	0	0	0	0	-1	0	0	0	-1	-1	0
Q	0	0	1	0	0	2	1	0	0	-1	-1	0	-1	-1	0	0	0	-1	-1	-1
E	1	0	0	1	0	1	2	0	0	-1	-1	0	-1	-1	-1	0	0	-1	-1	-1
G	0	0	0	0	0	0	0	2	0	-1	-1	0	-1	-1	1	0	0	-1	-1	-1
H	0	0	0	0	0	0	0	0	2	-1	-1	0	-1	-1	0	0	0	0	-1	-1
I	0	-1	-1	-1	0	-1	-1	-1	-1	2	0	-1	0	1	-1	-1	0	0	0	1
L	0	-1	-1	-1	0	-1	-1	-1	-1	0	2	-1	2	0	-1	-1	0	0	0	1
K	0	1	1	0	0	0	0	0	0	-1	-1	2	-1	-1	0	0	0	0	-1	-1
M	0	-1	-1	-1	0	-1	-1	-1	-1	0	2	-1	2	0	-1	-1	0	0	0	0
F	0	-1	-1	-1	-1	-1	-1	-1	-1	1	0	-1	0	2	-1	-1	0	0	1	0
P	-1	0	0	0	0	0	-1	1	0	-1	-1	0	-1	-1	3	0	0	-1	-1	-1
S	1	0	0	0	0	0	0	0	0	-1	-1	0	-1	-1	0	2	0	-1	-1	-1
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	-1	-1	0
W	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	-1	-1	-1	2	0	0
Y	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	-1	0	1	-1	-1	-1	0	2	0
V	0	-1	-1	-1	0	-1	-1	-1	-1	1	1	-1	0	0	-1	-1	0	0	0	2

g.deleage@ibcp.fr

Matrices de points utilisant la similarité avec matrice BLOSUM 62

	A	E	I	G	L	M	A	E	I	G	L	S	P	C	H	K	L	M
A																		
E																		
I			+4															
G				+6														
L					+4													
M						+5												
A							+4											
E																		
I																		
G							+0											
L								-3										
S									-2									
P										-2								
C											-1							
H																		
K																		
L																		
M																		

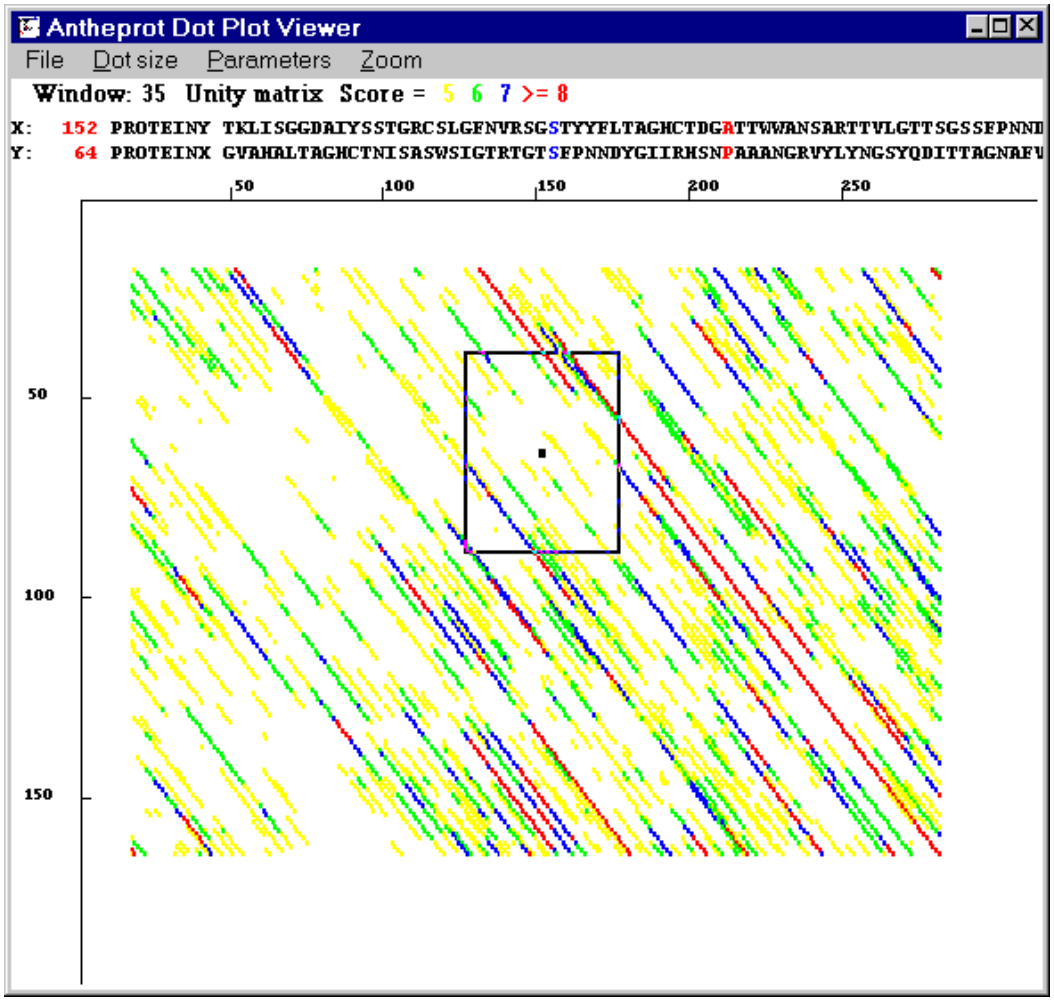
Score 23 donc >0

Score -8 donc < 0

g.deleage@ibcp.fr

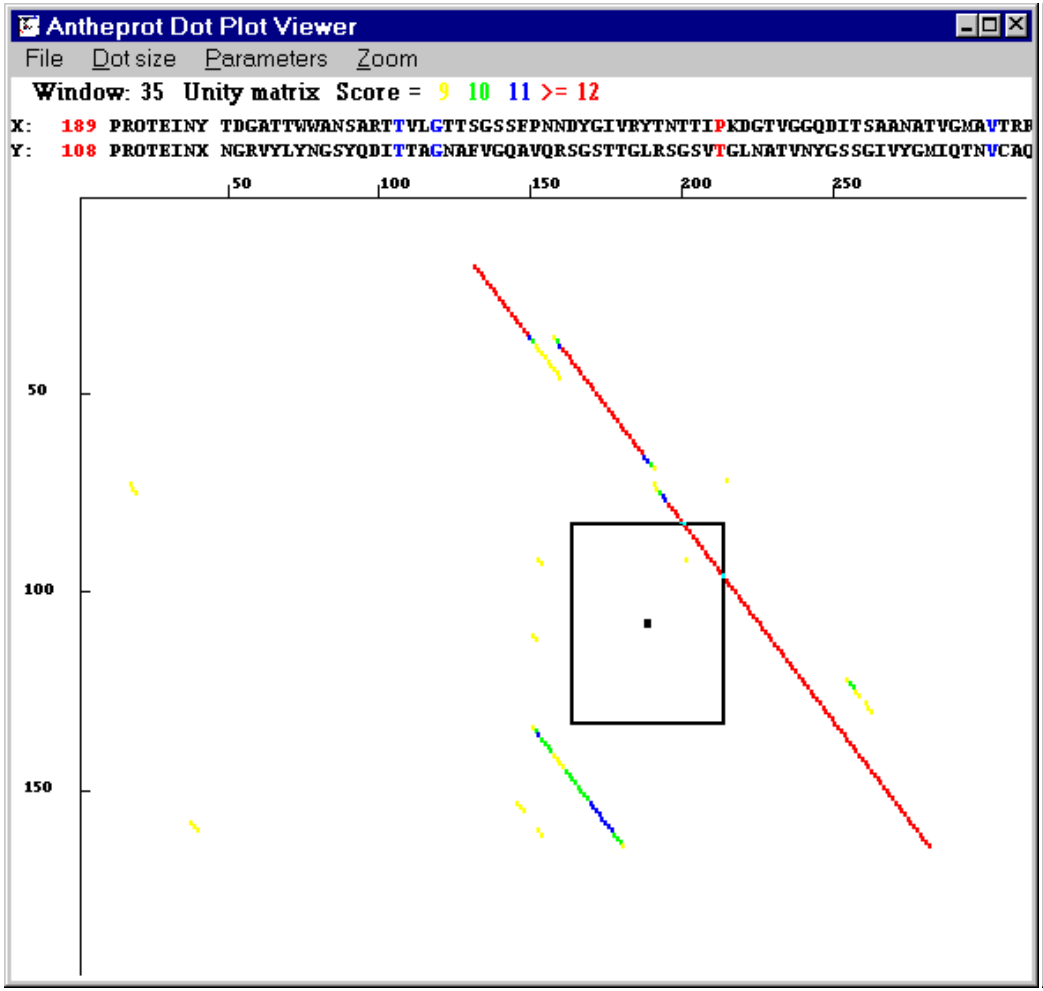
Matrice de points

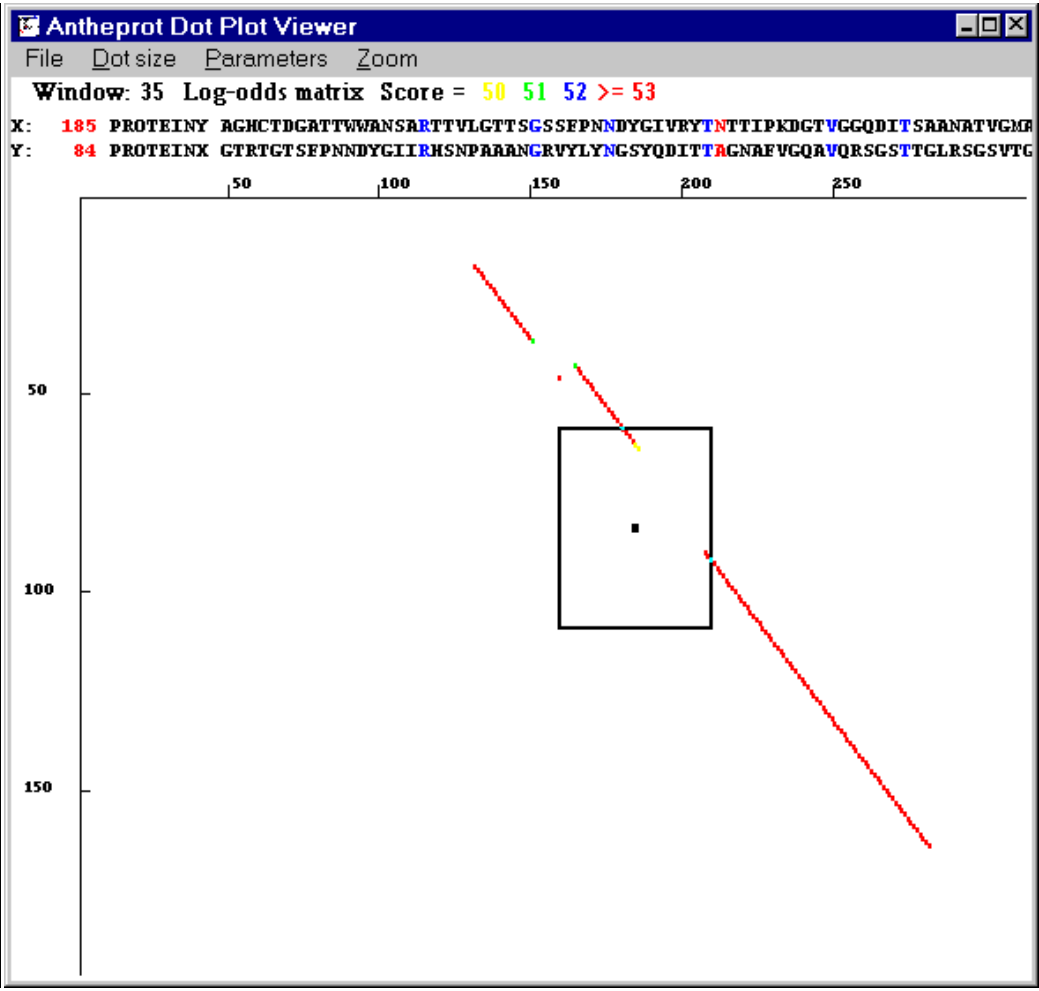
- **Avantage de la couleur**



g.deleage@ibcp.fr

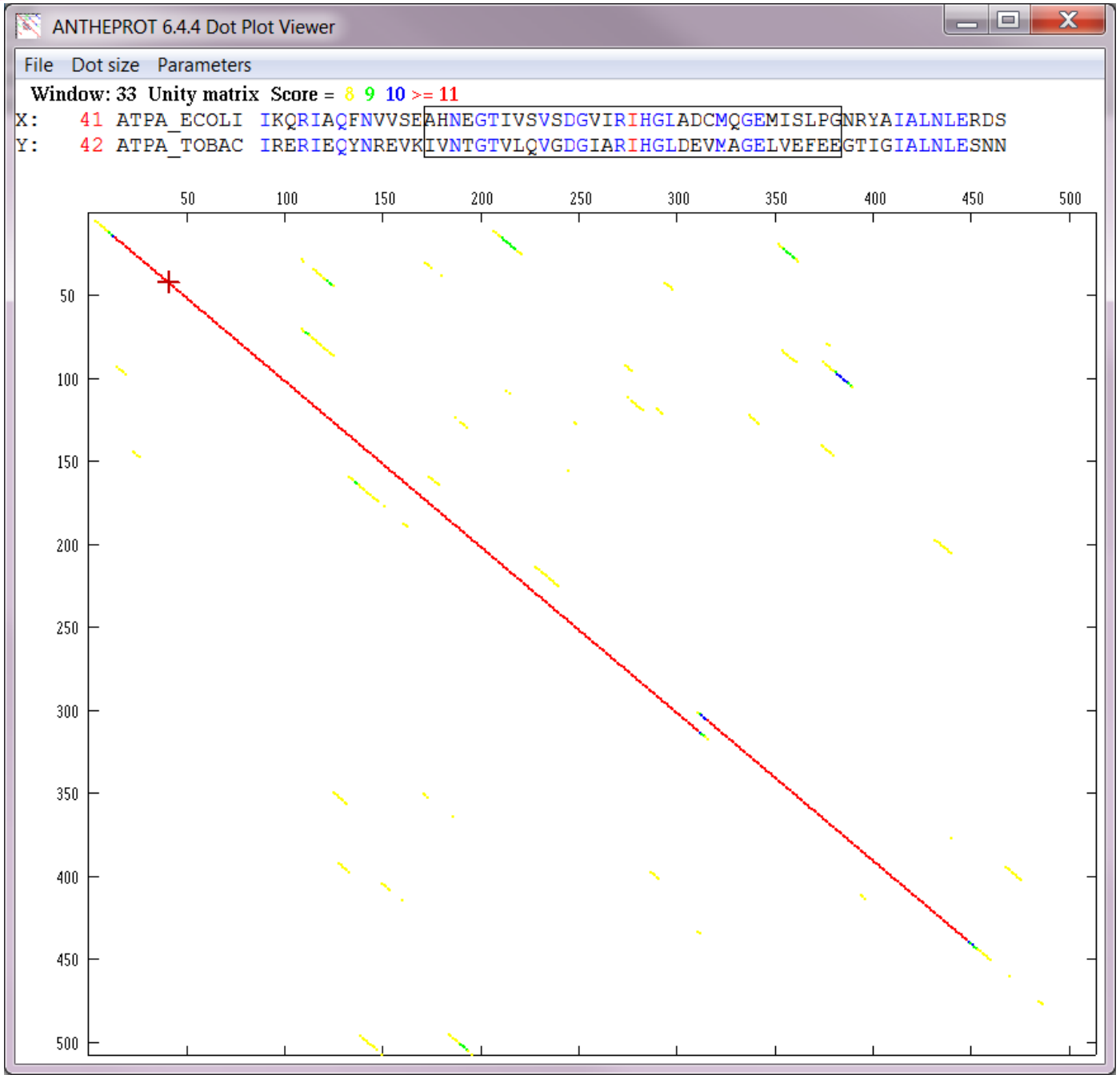
Matrice de points effet du paramétrage





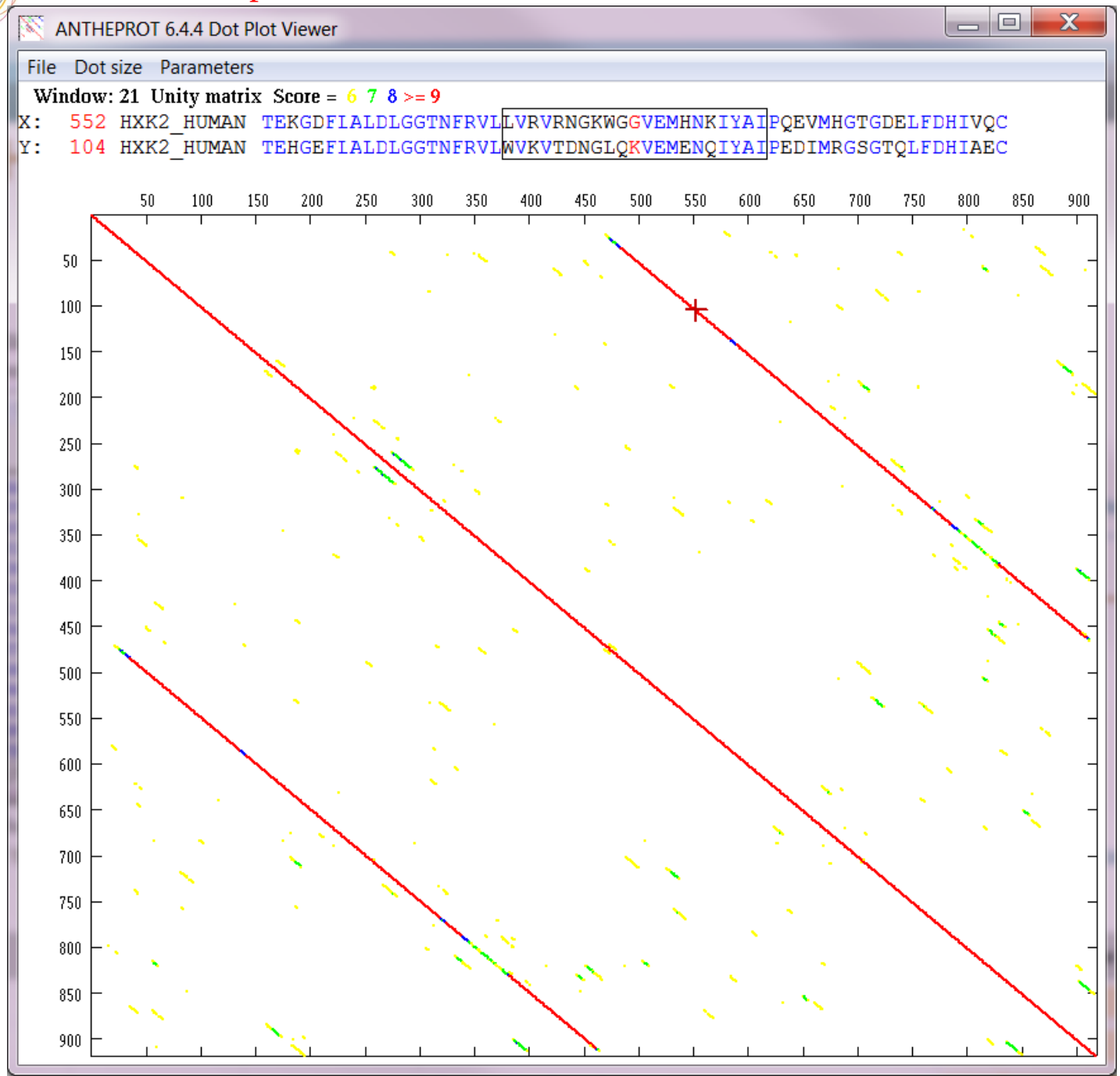
g.deleage@ibcp.fr

Exemples - Protéines homologues



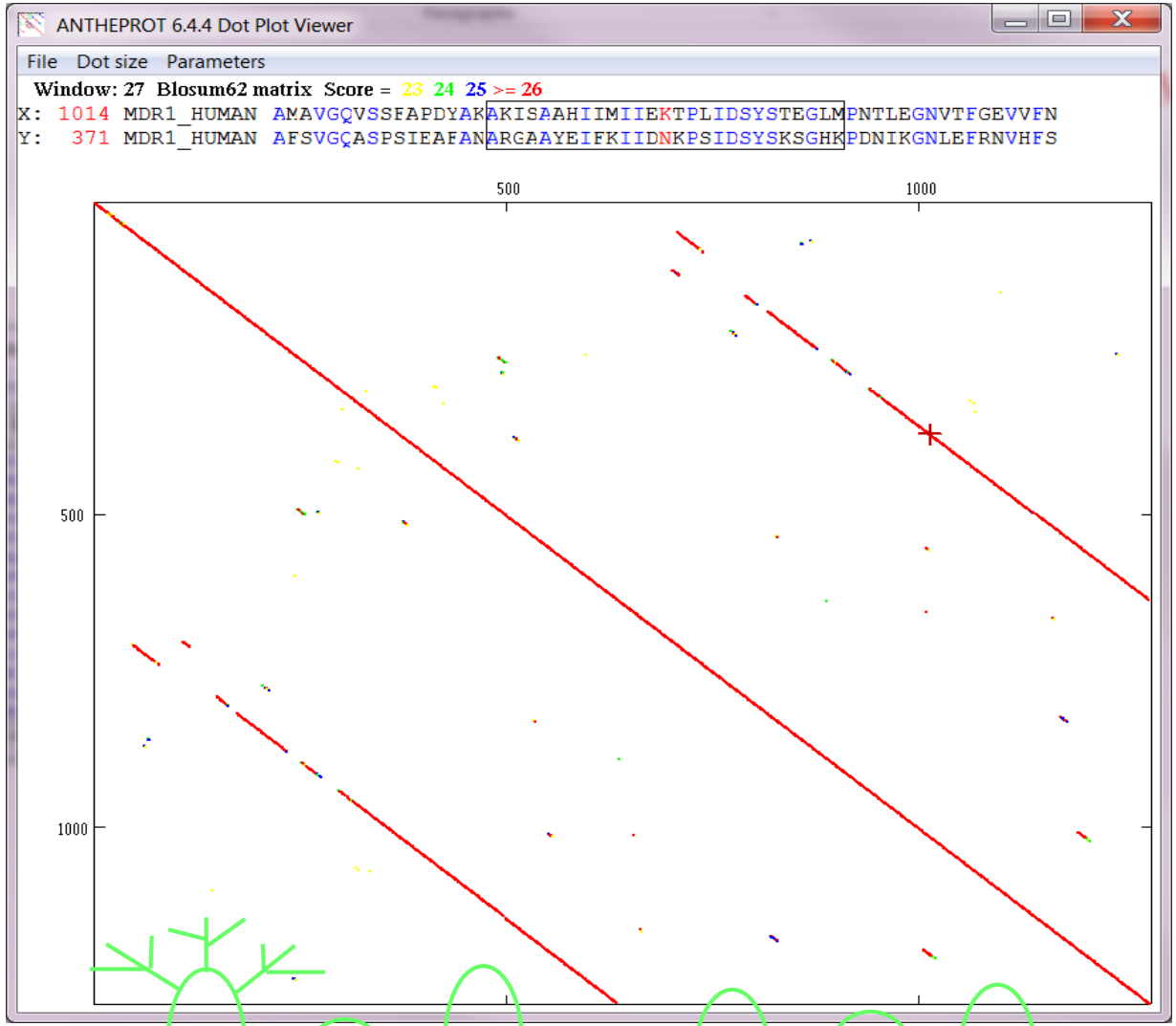
g.deleage@ibcp.fr

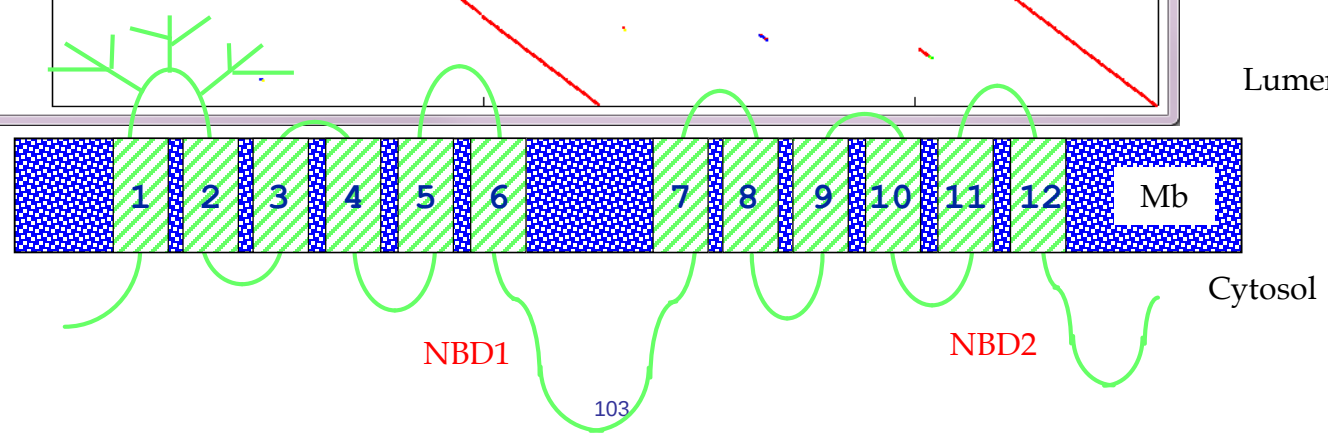
Duplication interne : hexokinase HXK2_HUMAN



g.deleage@ibcp.fr

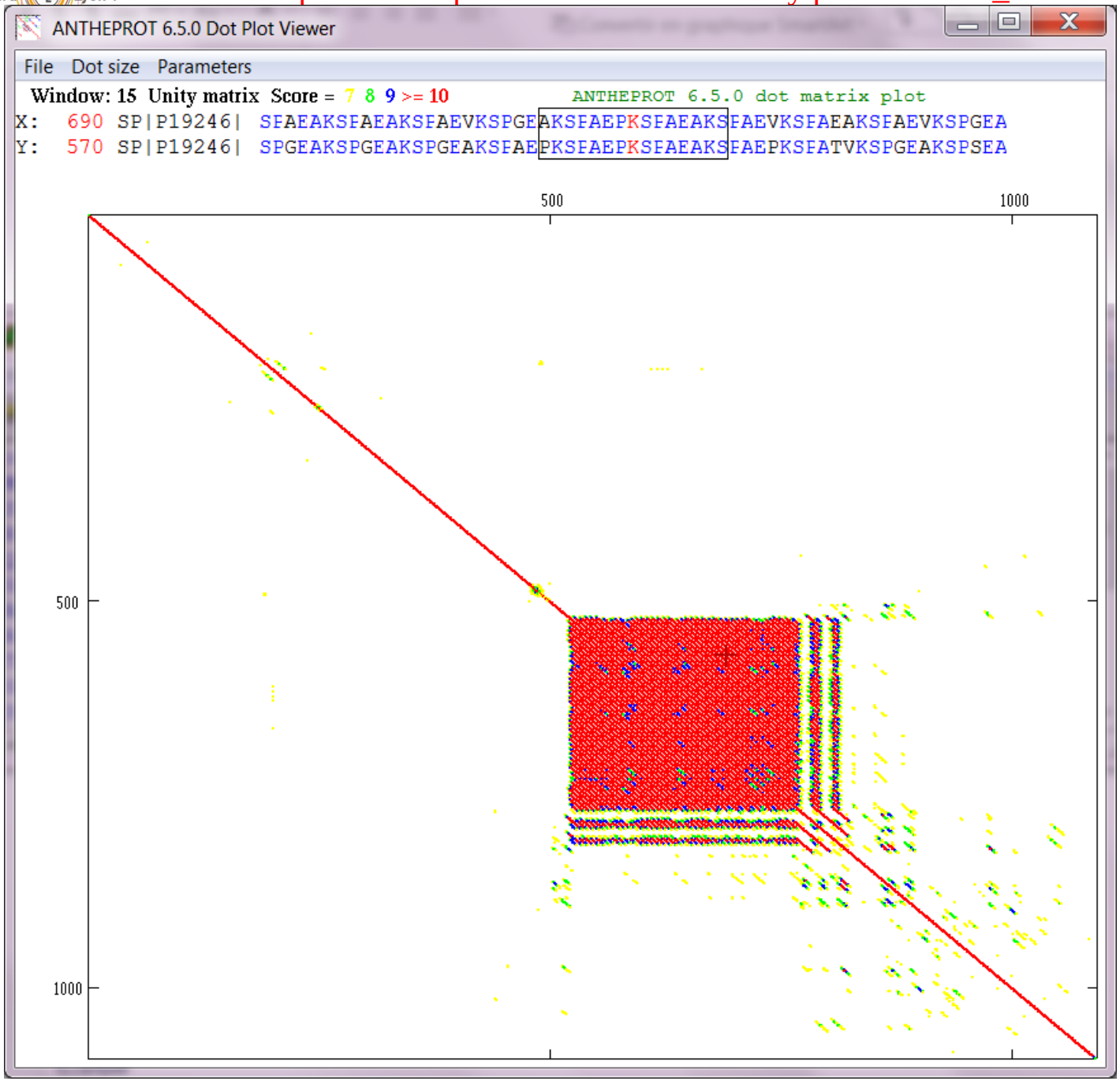
Duplication de gène : répétition interne MDR1_HUMAN





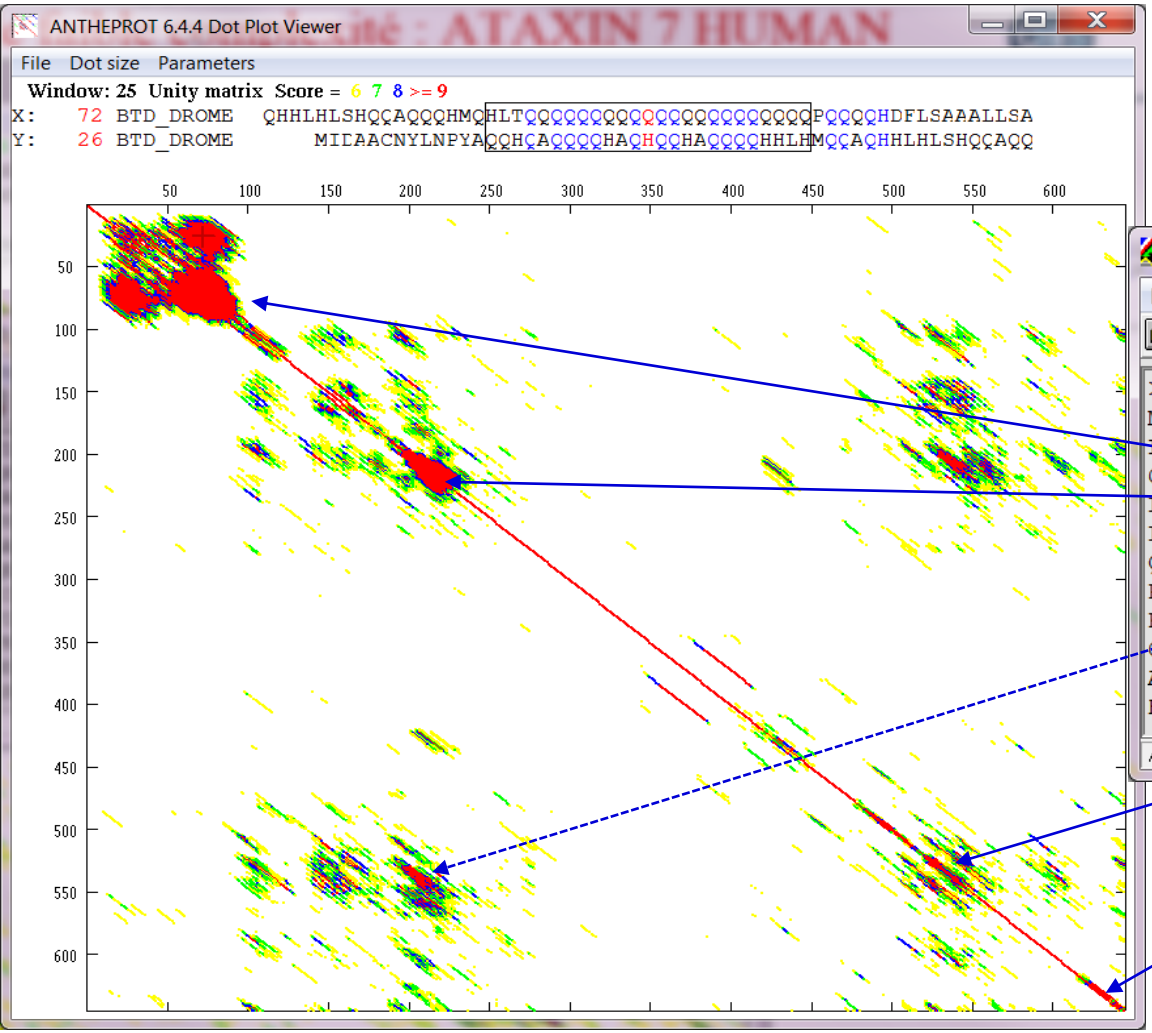
g.deleage@ibcp.fr

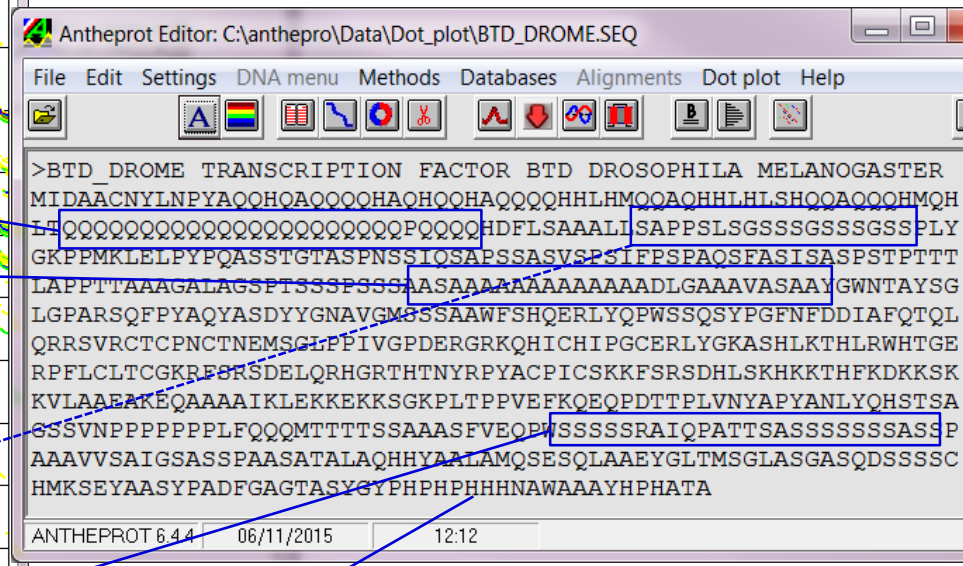
Faible complexité unique : Neurofilament heavy protein NFH_MOUSE



g.deleage@ibcp.fr

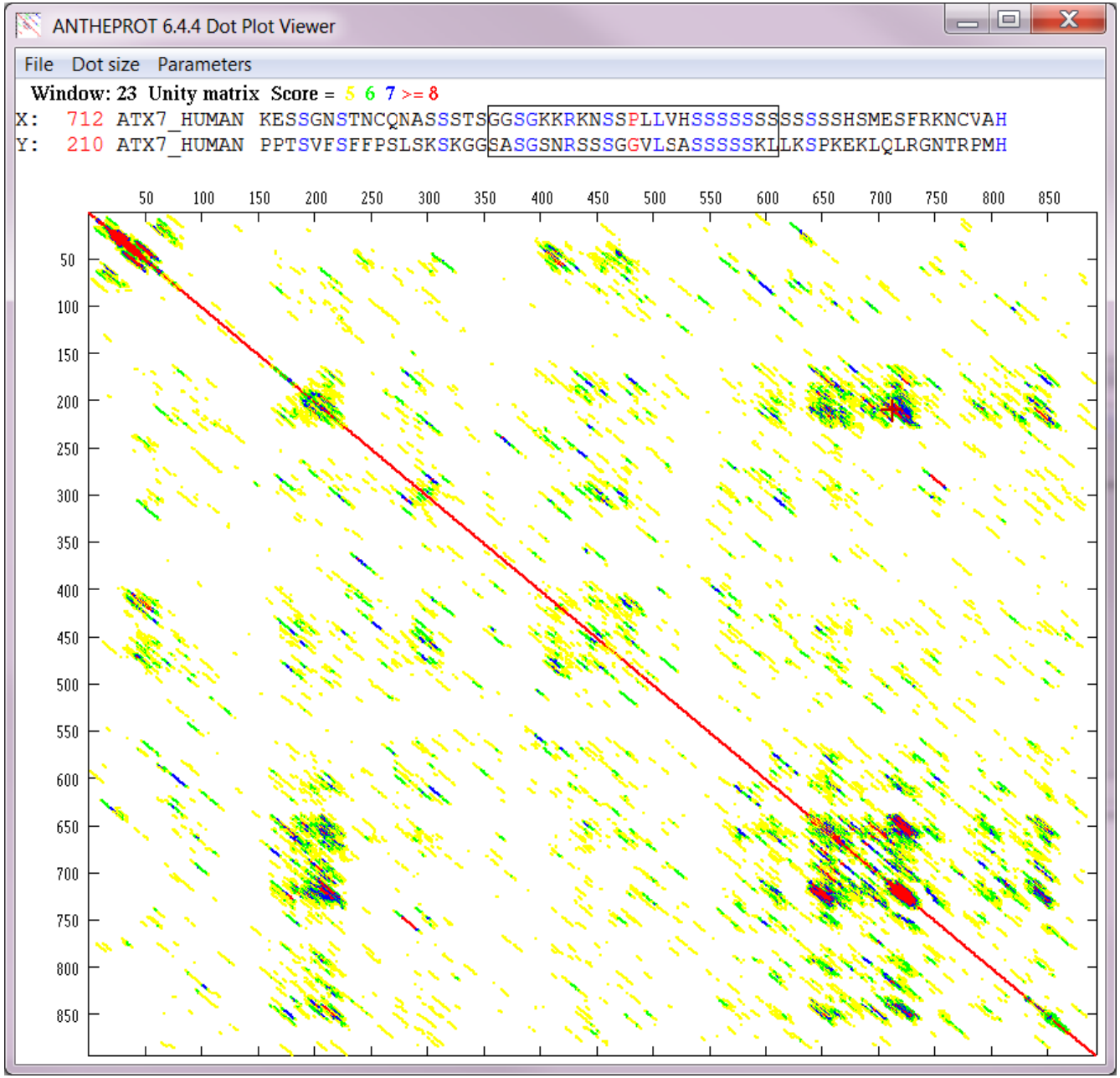
Faible complexité localisées multiples: Facteur de transcription BTD





g.deleage@ibcp.fr

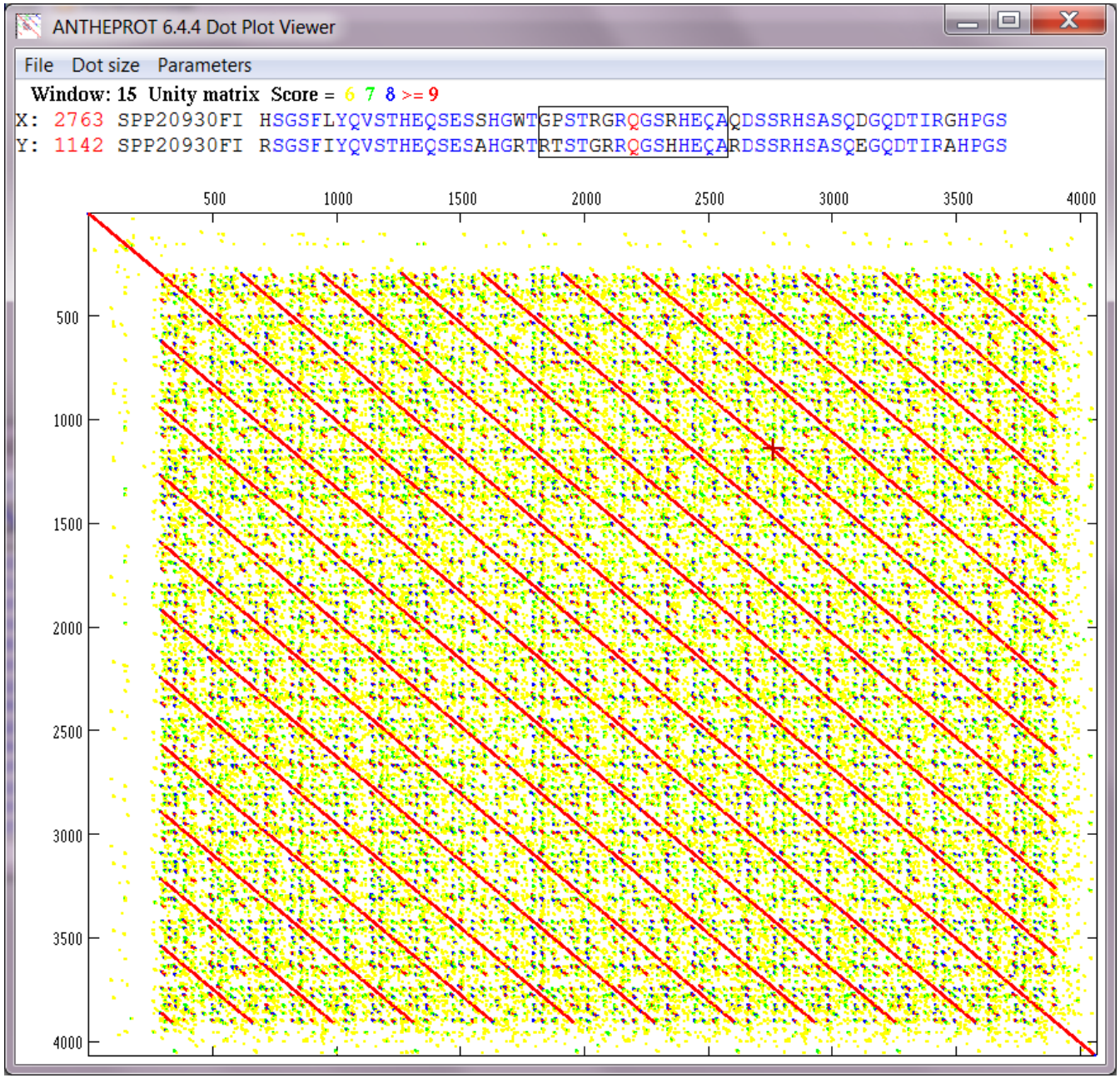
Régions de faible complexité diffuse: ATAXIN 7 HUMAN



Amino acid	Number	%
Ala(A):	56	6,27
Asx(B):	0	0
Cys(C):	20	2,24
Asp(D):	27	3,02
Glu(E):	36	4,03
Phe(F):	19	2,13
Gly(G):	57	6,39
His(H):	35	3,92
Ile(I):	22	2,46
Lys(K):	59	6,61
Leu(L):	58	6,5
Met(M):	17	1,9
Asn(N):	37	4,14
Pro(P):	108	12,1
Gln(Q):	44	4,93
Arg(R):	52	5,82
Ser(S):	141	15,8
Thr(T):	44	4,93
Val(V):	49	5,49
Trp(W):	4	0,44
Unk(X):	0	0
Tyr(Y):	7	0,78
Glx(Z):	0	0
Total :	892	100

g.deleage@ibcp.fr

Répétitions internes multiples et faible complexité : Fillagrin FILA_HUMAN

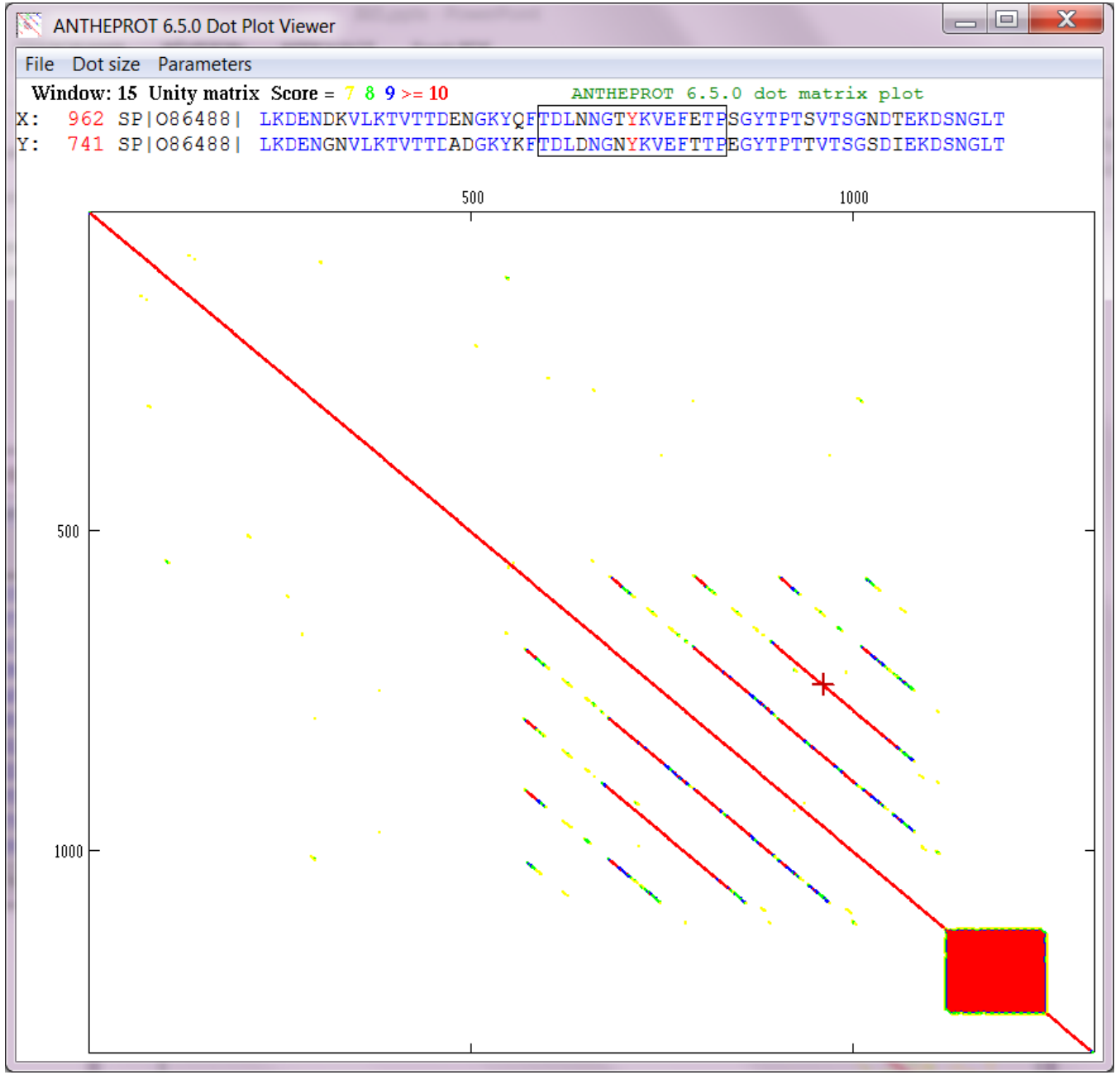


Amino acid	Number	%
Ala(A):	246	6,05
Asx(B):	0	0
Cys(C):	2	0,04
Asp(D):	222	5,46
Glu(E):	256	6,3
Phe(F):	24	0,59
Gly(G):	518	12,75
His(H):	413	10,16
Ile(I):	38	0,93
Lys(K):	57	1,4
Leu(L):	43	1,05
Met(M):	5	0,12
Asn(N):	62	1,52
Pro(P):	69	1,69
Gln(Q):	367	9,03
Arg(R):	440	10,83
Ser(S):	977	24,05
Thr(T):	166	4,08
Val(V):	83	2,04
Trp(W):	21	0,51
Unk(X):	0	0
Tyr(Y):	52	1,28
Glx(Z):	0	0
Total :	4061	100

G, H, R, S= 57,8 %

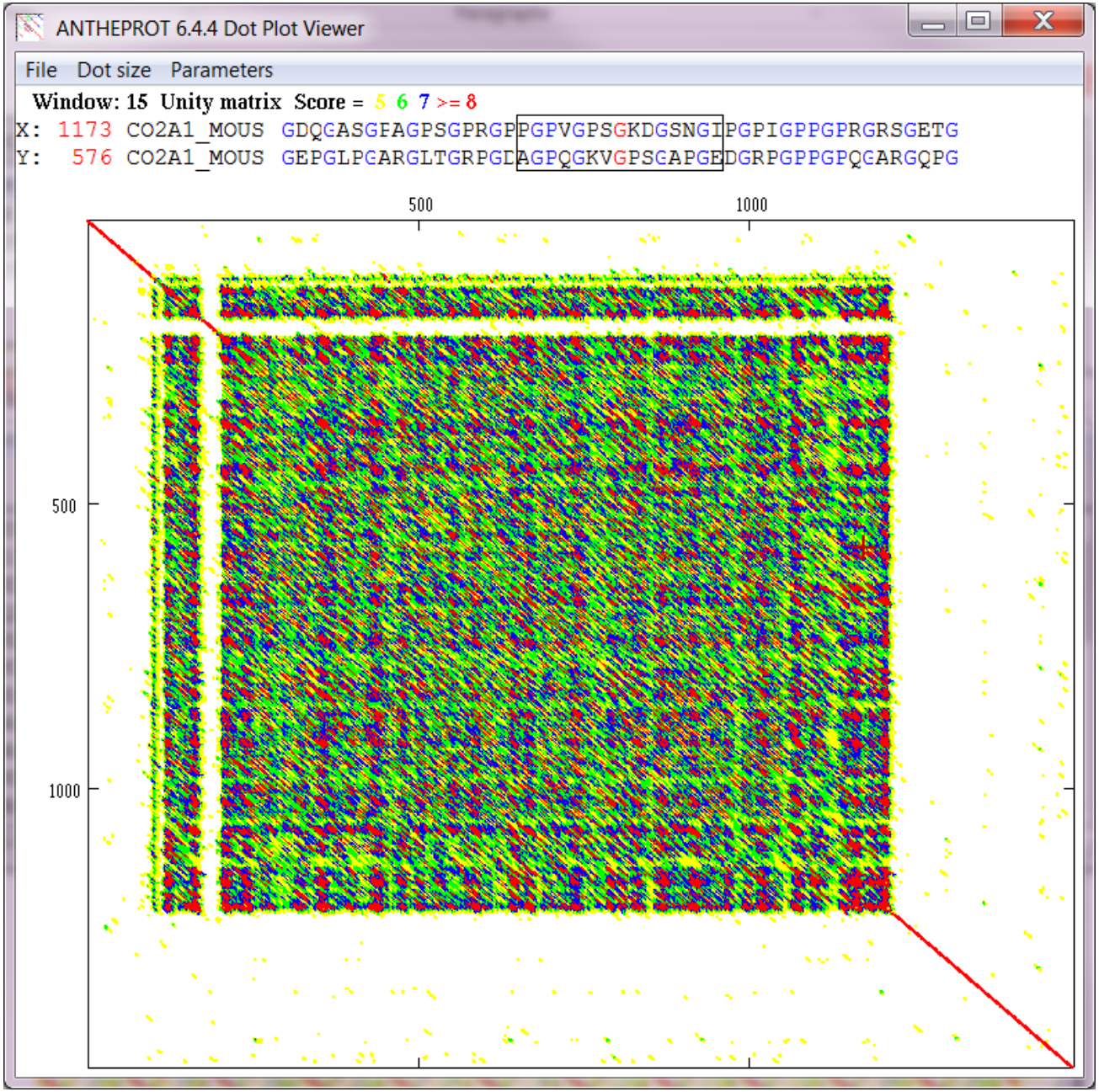
g.deleage@ibcp.fr

Répétitions internes multiples et faible complexité : SDRD_STAAE



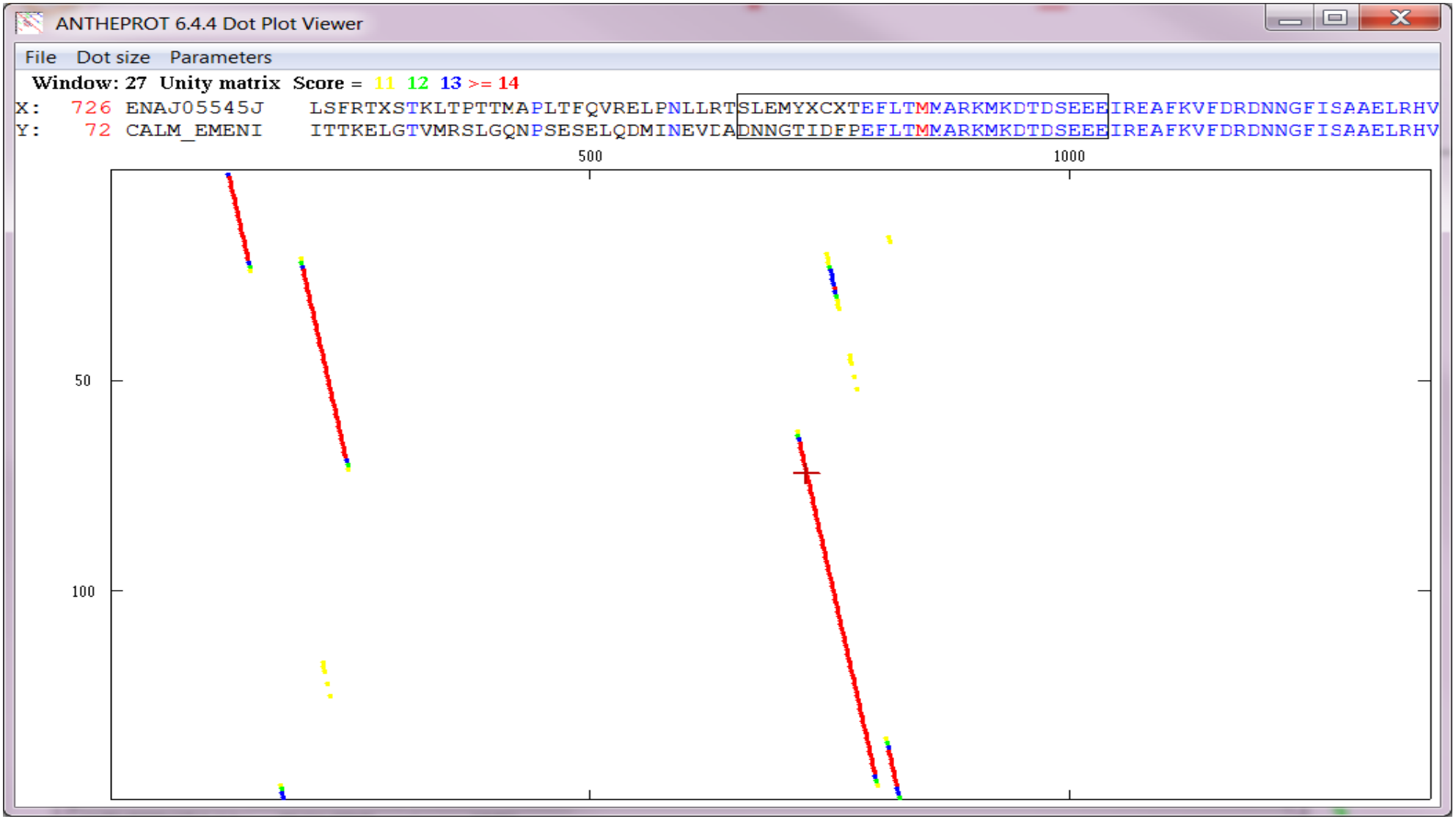
g.deleage@ibcp.fr

Faible complexité globale : Collagen alpha1(II) de souris



g.deleage@ibcp.fr

Exemples – Exons de la calmodulin

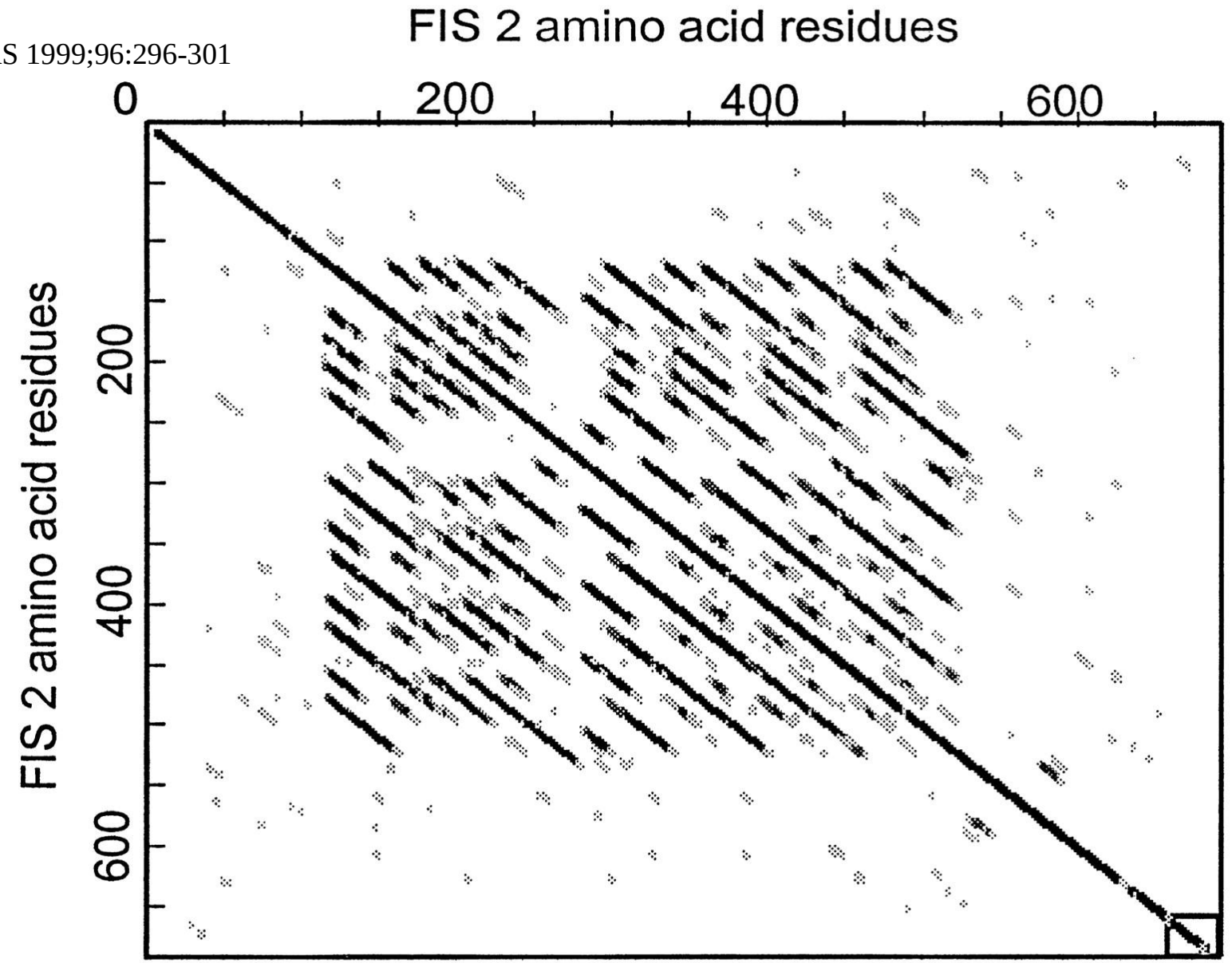


Mise bout à bout de la traduction en 3 phases du gene calmodulin

g.deleage@ibcp.fr

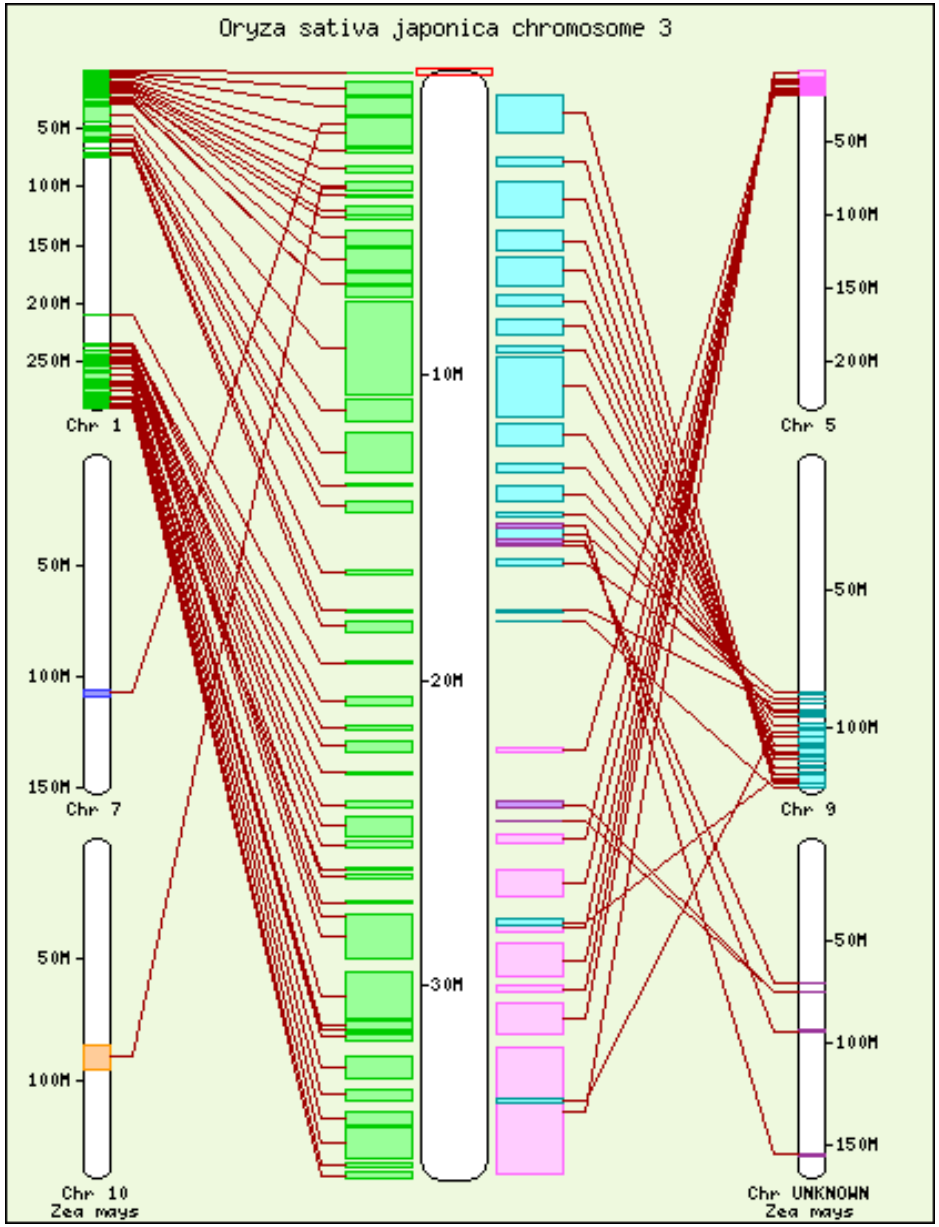
Bi-dimensional plot of the FIS2 predicted protein sequence showing the tandem repeats between residues 120 and 520.

Luo M et al. PNAS 1999;96:296-301



PNAS

Exemple des synténies



g.deleage@ibcp.fr

Recherche de segments similaires

- **Soient 2 séquences A et B à comparer**
- **Définition de la longueur L de comparison**
- **Comparaison des fenêtres de longueur L déplacées le long des 2 séquences**
  - Une première fenêtre de longueur L (1 à L) est définie sur la séquence A et les identités sont comptées sur chaque fenêtre de même longueur sur la séquence B
  - Un incrément de 1 est appliqué pour définir une 2° fenêtre sur A (2 à L+1)
  - re-examen de chaque fenêtre sur B
- **Score d'identité**
  - Pour chaque comparaison entre 2 segments de longueur L, le nombre d'identité sur le segment est calculé.
- **Sauvegarde des meilleurs scores (supérieurs à un pourcentage d'identité)**
- **Re-calcul des scores sur les fenêtres adjacentes**
- **Affichage des meilleurs scores avec les segments trouvés**

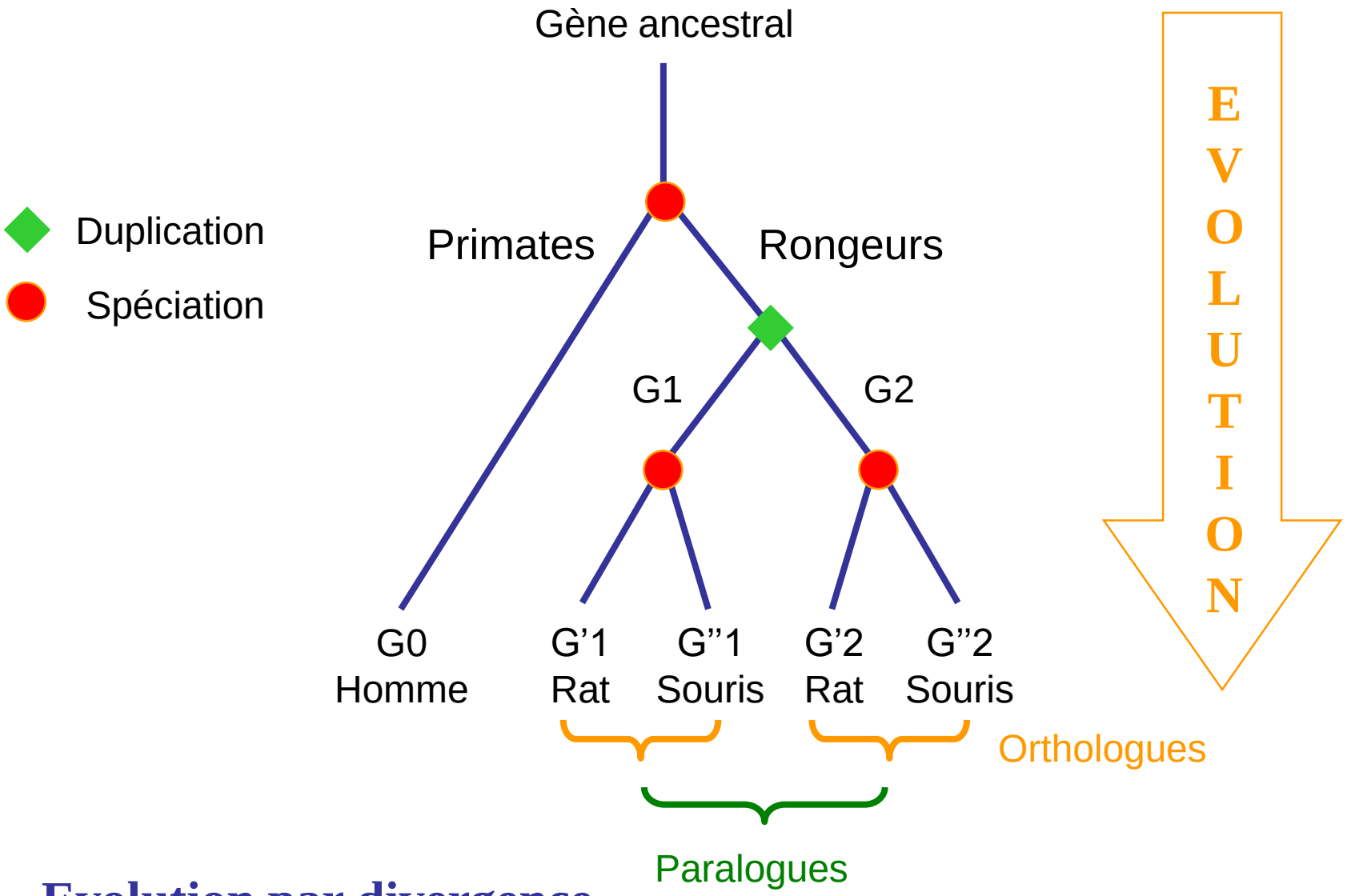
g.deleage@ibcp.fr

%Identité - %Similarité - Homologie

- **% Identité**
  - Quantité qui se mesure en % d'acides aminés identiques entre 2 séquences (après alignement des séquences)
- **%Similarité**
  - Quantité qui se mesure en % d'acides aminés ressemblants entre 2 séquences
- **Homologie**
  - 2 protéines sont homologues ssi elles ont un ancêtre commun
    - Paralogues: Séquences homologues qui ont évoluées par duplication
    - Orthologues: Gènes homologues qui ont divergé suite à la spéciation (à la séparation d'une espèce en deux espèces différentes)
  - Il est possible d'observer la ressemblance résiduelle entre les séquences originelles après l'évolution, ce qui permet d'inférer l'homologie.
  - En général, pour des séquences de longueur standard, on peut inférer l'homologie entre 2 protéines si leurs séquences présentent 30% ou plus d'identités résiduelles mais…
  - Il existe des séquences homologues avec moins de 30% d'identité dans ce cas là:
    - Homologie est transitive
      - si A homologue à B
      - et B homologues à C
      - alors A homologue à C même si A et C ont peu de similarités
    - Utilisation des structures

g.deleage@ibcp.fr

Homologie - Paralogues – Orthologues

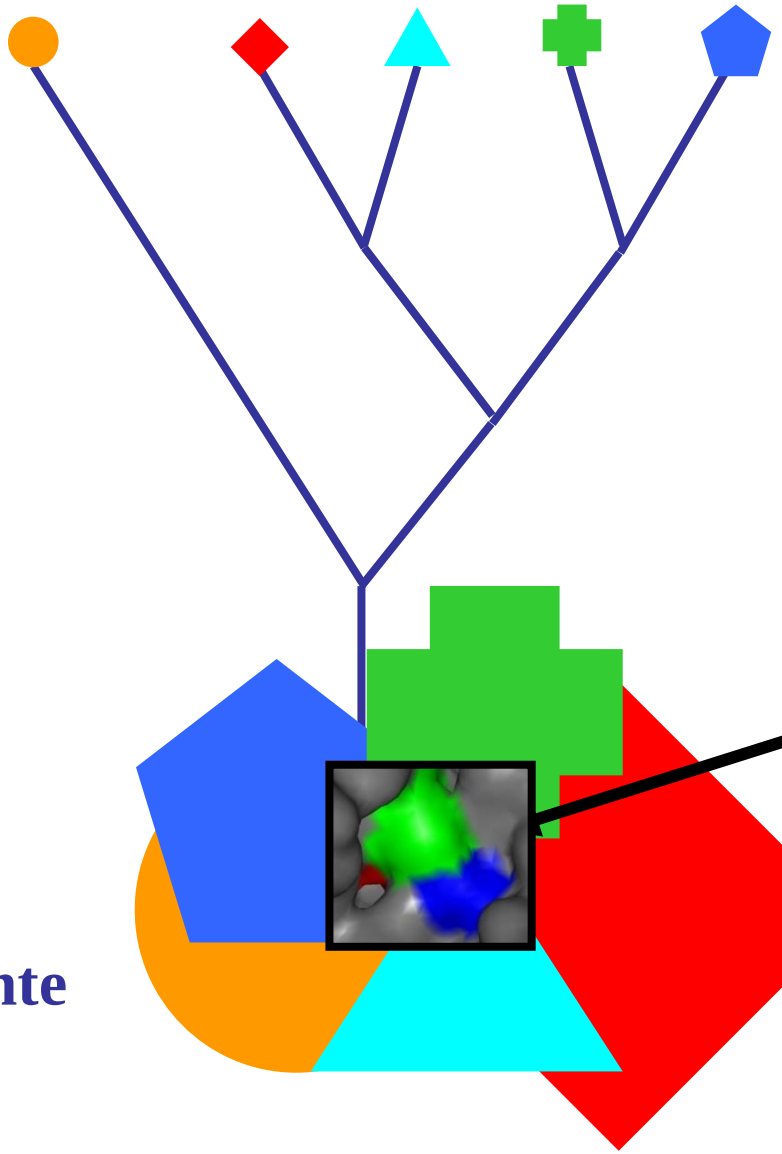


Evolution par divergence

g.deleage@ibcp.fr

Analogie

EVOLUTION



Ancêtres différents
et structures différents

Fonction identique
Site actif commun

Evolution convergente

g.deleage@ibcp.fr

Homologie-Analogie

HOMOLOGIE

Ancêtre 3D commun
Myoglobine de cachalot
Leghemoglobine de lupin
15% identité de séquences
Fonctions différentes

ANALOGIE

Ancêtres différents
Structures 3D parfois différentes
Fonctions identiques
Protéases Ser, His et Asp

Les structures 3D évoluent moins vite que les séquences

Les structures 3D sont plus préservées par l'évolution que les séquences

La pression de l'évolution est plus forte sur les structures que sur les séquences

Les structures 3D s'accomodent des séquences (plasticité)

Evolution divergence
Structures 3D proches

Evolution convergente
Structures 3D différentes
Site actif identiques

g.deleage@ibcp.fr

Homologie-Analogie

HOMOLOGIE

Ancêtre 3D commun
Myoglobine de cachalot
Leghemoglobine de lupin
15% identité de séquences
Fonctions différentes

ANALOGIE

Ancêtres différents
Structures 3D parfois différentes
Fonctions identiques
Protéases Ser, His et Asp

Les structures 3D évoluent moins vite que les séquences

Les structures 3D sont plus préservées par l'évolution que les séquences

La pression de l'évolution est plus forte sur les structures que sur les séquences

Les structures 3D s'accomodent des séquences (plasticité)

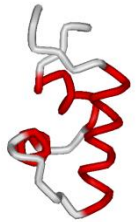
Evolution divergence
Structures 3D proches

Evolution convergente
Structures 3D différentes
Site actif identiques

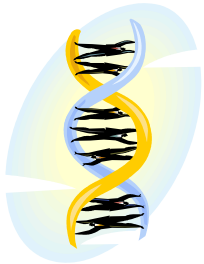
g.deleage@ibcp.fr

Choix des séquences
Acides nucléiques ou Protéines?

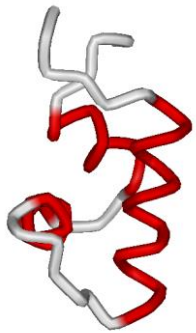
- **Protéines**
  - Identification de séquences homologues après 1 milliard d'années d'évolution
  - Identification de similarités significatives après 2,5 milliards d'années



- **Acides nucléiques**
  - DNA non codant =>200 millions d'années
  - DNA codant => 600 millions d'années

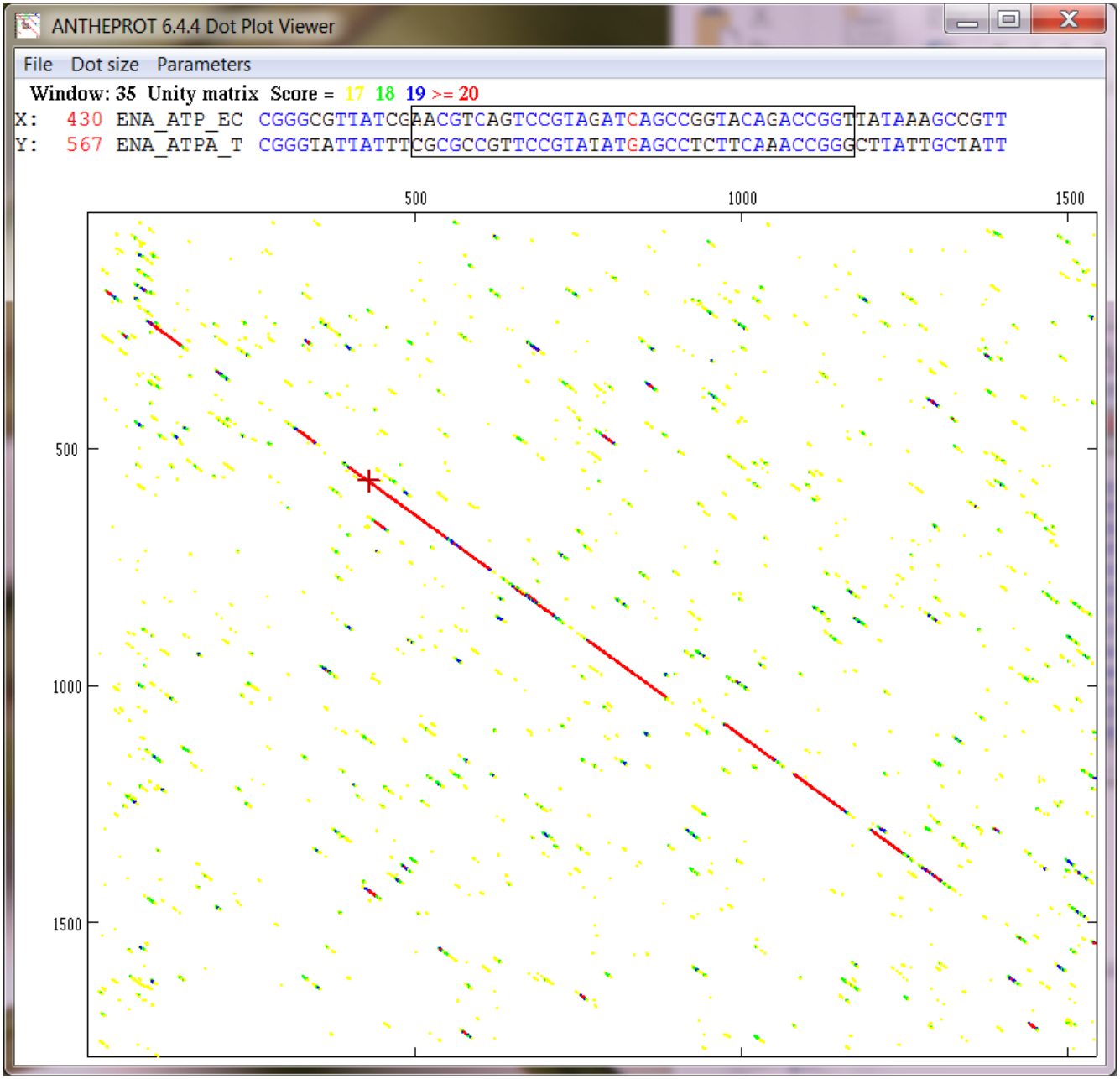


- **Conclusion =>*PROTEINES***



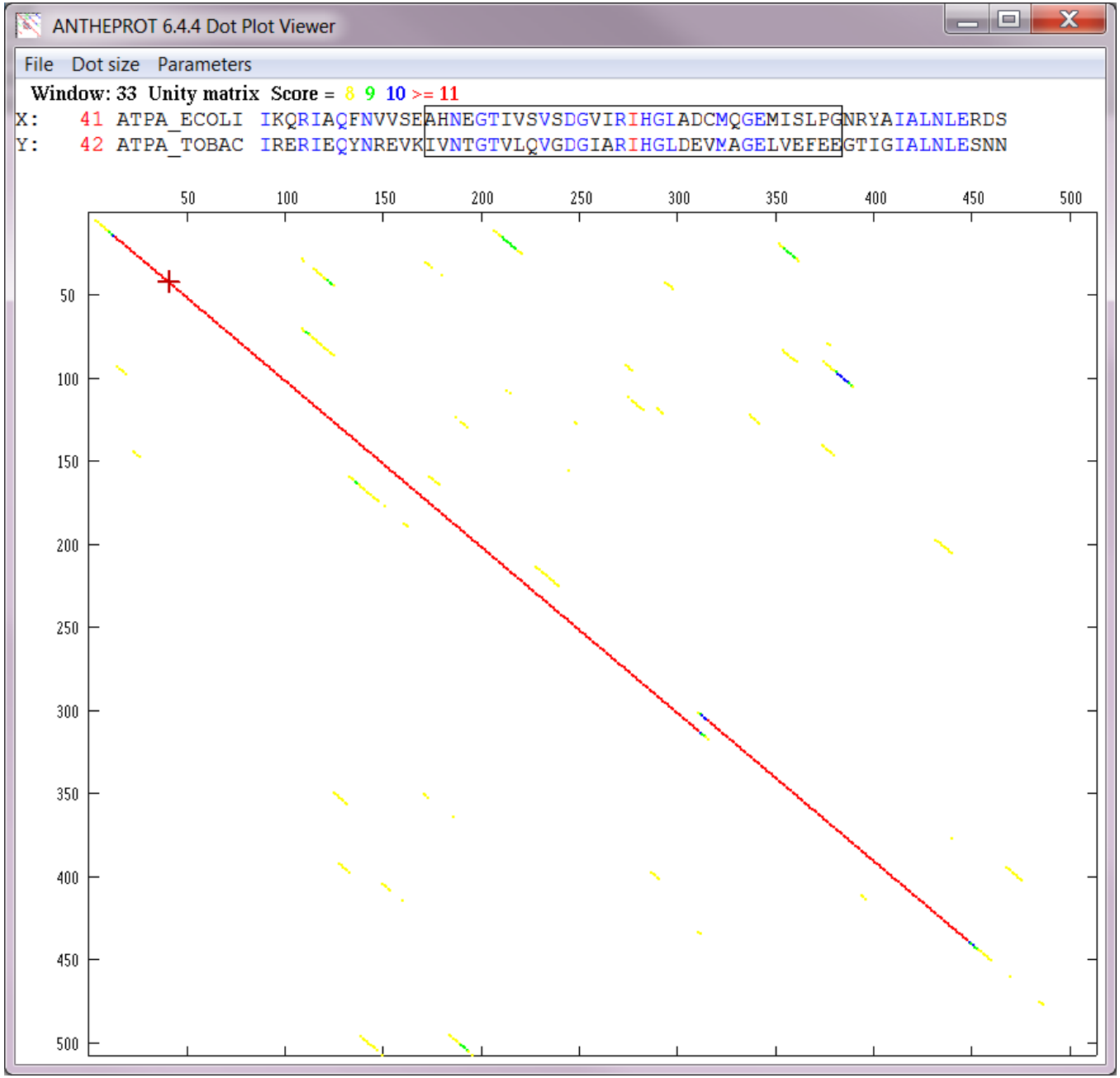
g.deleage@ibcp.fr

Exemples – Gènes homologues



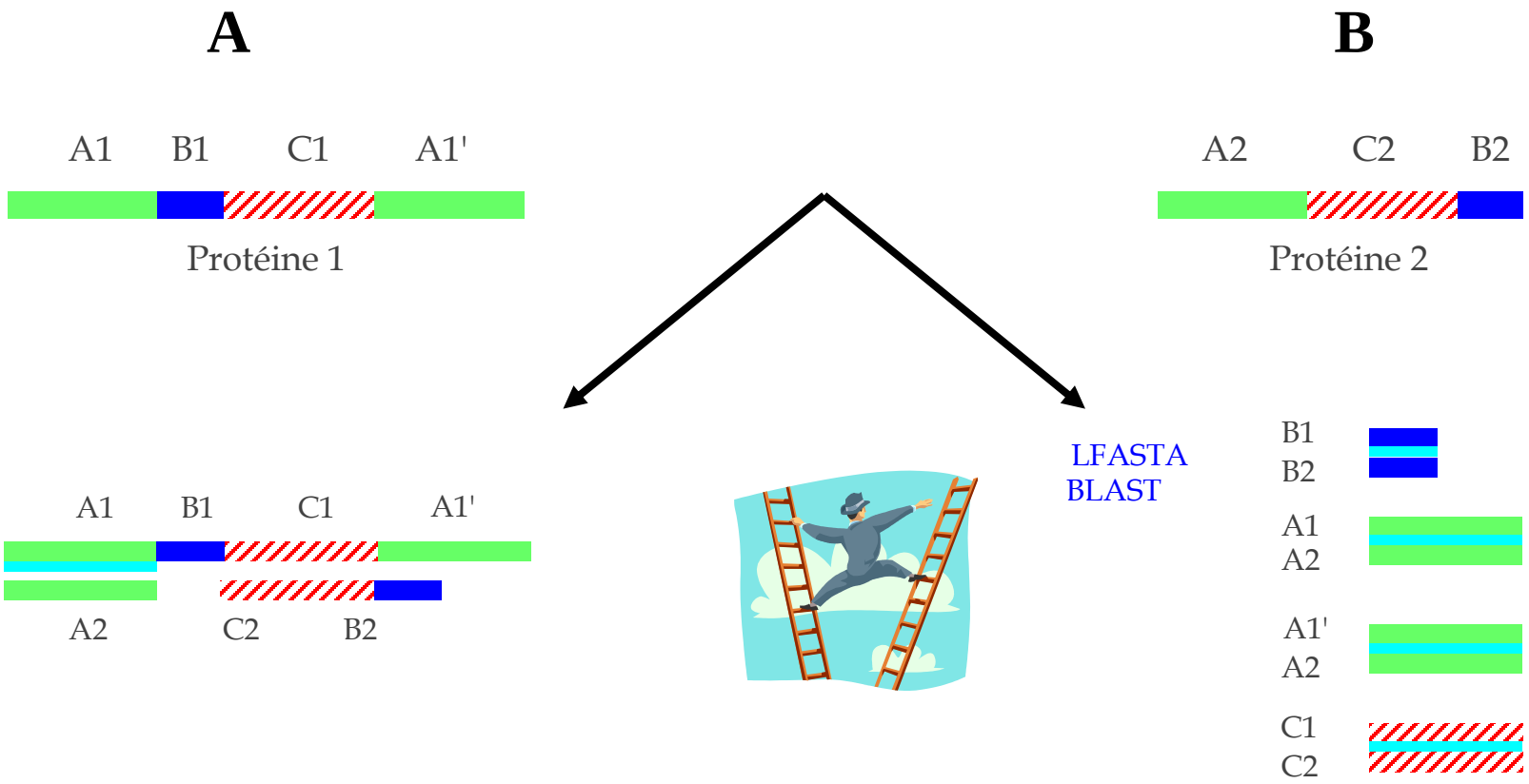
g.deleage@ibcp.fr

Exemples - Protéines homologues



g.deleage@ibcp.fr

Similarités locales ou globales



Recherche de similarité globale

FASTA

Recherche de similarité locale

BLAST

g.deleage@ibcp.fr

Fasta – 2 Heuristiques - 4 étapes

Heuristiques (approximations) de Fasta

En examinant les séquences par mot (uplets) de longueur k avec k > 1

2 séquences homologues auront :

Heur. 1) des segments de longueurs > k identiques (H1)

(heuristique si un aa/2 est identique => 50% Id inférence d'homologie)

On pourrait à la limite avoir 2 séquences ayant Id = 50% sans avoir de mot de 2 aa (k=2) identiques

Heur. 2) Les séquences homologues auront surtout des uplets proches de la diagonale de la matrice de points.

Cas des séquences homologues avec longues insertions

g.deleage@ibcp.fr

Fasta - Etapes

Principe de Fasta

- Découpage de la séquence en uplet de longueur k (k>1) (découpage en k-mots)
- Recherche de k-mots identiques (2 à 4 protéines, 4 à 11 acides nucléiques) => score init1 (score de la plus forte densité) (Heur. 1)
- Chaînage des k-mots sur la même diagonale =>score initn (extension de la densité)
- Filtrage par une matrice PAM
- Construction des zones fortement identiques à une distance d de la diagonale en cours (Heur.2)
- recalcul des scores après avoir rabouté les k-mots distant de d =>score opt
- Statistique associée => Calcul de Z-score, de E()

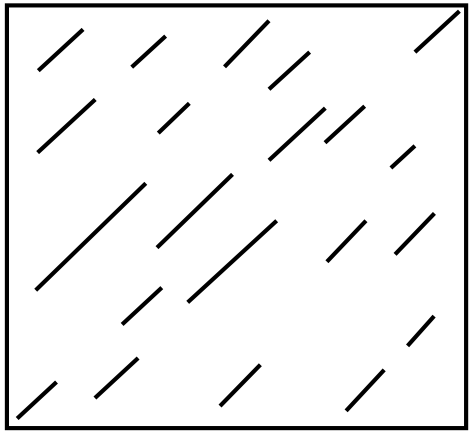
g.deleage@ibcp.fr

Recherche de séquences homologues par Fasta

- 4 étapes successives

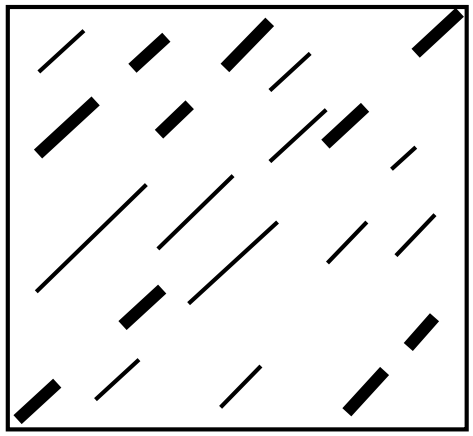
1) Recherche de k mots identiques

Score init1

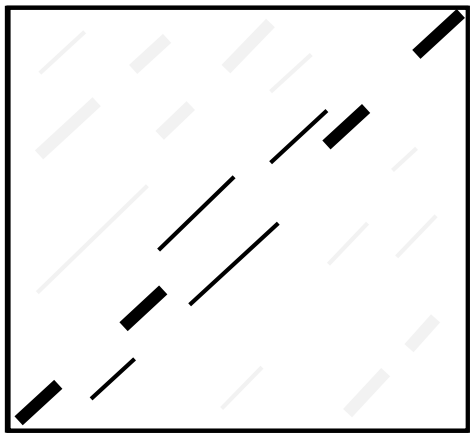


2) Chaînage des k mots sur la même diagonale+filtrage

Score initn

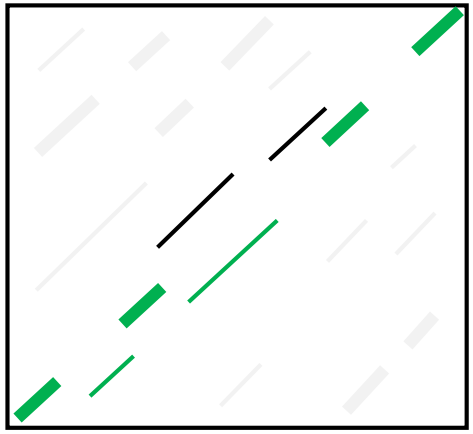


3) Elimination des régions éloignées de la diagonale



4) Alignement optimal

Score opt



125

g.deleage@ibcp.fr

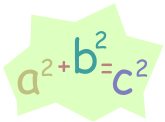
Score de similarité pour 2 séquences

1. Nécessité d'aligner les séquences
2. Besoin de définir le mode de calcul du score
3. Besoin de savoir si l'obtention d'un score Sr est fortuit:

Sr score (*i.e.* nombre d'identité)

- Génération de séquences au hasard de même composition en AA
- Calcul du score moyen pour les séquences au hasard = score de base= Sb
- Comparaison de Sr et Sb par calcul d'un Z-score.

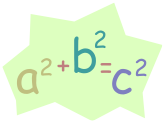
$$\text{Z-score} = [Sr - Sb] / \sigma$$



σ est la déviation standard de la distribution des scores

- E() value= nombre de fois que 2 séquences de longueur n et m auront par chance un score = Z

$$E() = K\, n\, m\, e^{-\lambda Z}$$



n: longueur de séquence 1
m : longueur de séquence 2
λ et K sont des constantes

g.deleage@ibcp.fr

Score de similarité pour 2 séquences


```
S     DDLSKQAVAYRQMSLLL-RG
S1    -D-GKTAVA-TDTILLLQG-
       *   *  ***      ***
```

8 identitiés

```
Salea 1  -GRLLLSMQRYAVA---QKSLDD
S1       DG--------KTAVATDTILLLQG
          *           ***       *
```

5 identitiés

```
Salea 2  DG-LYV---QSLLMDRRAAKLSQ
S1       DGKTAVATDTILL-------LQG
         **    *      **        *
```

6 identitiés

```
Salea 3   QSLK-A-ARRD--MLLQSYV
S1        -DGKTAVA-TDTILLLQ---
             * * *   *    ***
```

7 identitiés

```
Salea 4  ----A-ARRQSLGLYVSQLLMDKD
S1       DGKTAVA--TDTIL----LL--QG
            * *        *     **
```

5 identitiés

Z-score= (8-5,75)/(0,95)
Z-score = 2,4

g.deleage@ibcp.fr

Fasta - Paramètres

FASTA parameters:

Number of best scores to display (-b, int) : 500

Number of alignments to show (-d, int) : 500

Expectation value upper limit (-E, real) : 10.0

Expectation value lower limit (-F, real) : 0

Turn off histogram display (-H) : ☐

Report long sequence description in alignments (-L) : ☐

Matrix (-s) : BLOSUM50

Threshold for band optimization (-c) : 1

Penalty for the first residue in a gap (-f, real) : -12

Penalty for adfditional residues in a gap (-g, real) : -2

ktup : 2

Width for band optimization (-y, int) : 16 (16 for protein and ktup=2, 32 for protein ktup=1)

Number of processors to use (-T) : 2

-D, -m, -N, -q/-Q, -r, -S, -w, -x, -z, -Z, and -1 options are not supported.

g.deleage@ibcp.fr

FASTA- Distribution


```
         opt      E()
<  20    216      0:==
   22      1      0:=            one = represents 162 library sequences
   24      3      0:=
   26      9      2:*
   28     45     23:*
   30    189    142:*=
   32    594    547:===*
   34   1465   1484:==========*
   36   2954   3048:====================*
   38   5066   5038:==================================*
   40   6876   7027:================================================*
   42   8534   8590:============================================================*
   44   9720   9475:=====================================================================*=
   46   9693   9651:=====================================================================*
   48   9359   9239:=================================================================*
   50   8825   8431:=========================================================*==
   52   7364   7412:====================================================*
   54   6173   6331:============================================*
   56   5120   5289:==================================*
   58   4280   4342:=========================*
   60   3568   3517:========================*=
   62   2671   2820:==================*
   64   2256   2243:==============*
   66   1853   1772:==========*=
   68   1303   1394:=========*
   70    987   1093:=======*
   72    885    854:======*
   74    596    666:=====*
   76    474    518:====*
   78    439    403:==*
   80    311    313:=*
   82    244    239:=*
   84    178    189:=*
   86    122    147:*
   88    122    113:*
   90    103     88:*
   92     75     68:*
   94     47     53:*
   96     52     41:*
   98     29     31:*
  100     21     24:*
  102     16     19:*
  104     13     15:*
  106      7     11:*
  108      8      9:*
  110      9      7:*
  112      7      5:*
  114      6      4:*
  116      5      3:*
  118      3      2:*
>120    362      2:*==
38031984 residues in 103258 sequences
```

Distribution des séquences
= réelle
* attendue

```
  inset = represents 8 library sequences

:========*=
:=======*
:======*=
:====*
:==*
:==*
:=*
:=*
:*=
:*
:*
:*
:*=========================================
```

Zoom

Séquences sortant du bruit

g.deleage@ibcp.fr

FASTA - Classement des « hits » par score décroissant


```
FASTA (3.43 Nov 2001) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 37, opt: 25, gap-pen: -12/-2, width:  16
 Scan time:  2.520
The best scores are:                                      opt bits E(103258)
sw||ATPA_TOBAC (P00823) ATP synthase alpha chain    ( 507) 3142  652 1.1e-186
sw||ATPA_SPIOL (P06450) ATP synthase alpha chain    ( 507) 2995  622 1.2e-177
sw||ATPA_ARATH (P56757) ATP synthase alpha chain    ( 507) 2980  619  1e-176
sw||ATPA_PEA (P08215) ATP synthase alpha chain (E   ( 501) 2909  605 2.4e-172
sw||ATPA_MARPO (P06283) ATP synthase alpha chain    ( 507) 2814  585 1.7e-166
sw||ATPA_MAIZE (P05022) ATP synthase alpha chain    ( 507) 2776  577 3.8e-164
sw||ATPA_ORYSA (P12084) ATP synthase alpha chain    ( 507) 2770  576 8.9e-164
sw||ATPA_WHEAT (P12112) ATP synthase alpha chain    ( 504) 2728  568 3.4e-161
sw||ATPA_PINTH (P41602) ATP synthase alpha chain    ( 494) 2697  561 2.7e-159
sw||ATPA_CHLVU (P56294) ATP synthase alpha chain    ( 506) 2570  535 1.8e-151
sw||ATPA_MESVI (Q9MUT2) ATP synthase alpha chain    ( 505) 2552  532 2.3e-150
sw||ATPA_NEPOL (Q9TL16) ATP synthase alpha chain    ( 501) 2540  529 1.3e-149
sw||ATPA_EUGGR (P30392) ATP synthase alpha chain    ( 506) 2512  524 6.8e-148
sw||ATPA_PORPU (P51242) ATP synthase alpha chain    ( 504) 2457  512 1.7e-144
sw||ATPA_CYAPA (P48080) ATP synthase alpha chain    ( 505) 2456  512 1.9e-144
sw||ATPA_GUITH (O78475) ATP synthase alpha chain    ( 502) 2450  511 4.4e-144
sw||ATPA_SYNP1 (Q05372) ATP synthase alpha chain    ( 503) 2447  510 6.8e-144
sw||ATPA_ANTSP (Q02848) ATP synthase alpha chain    ( 505) 2442  509 1.4e-143
sw||ATPA_CHLRE (P26526) ATP synthase alpha chain    ( 507) 2426  506 1.3e-142
sw||ATPA_SYNP6 (P08449) ATP synthase alpha chain    ( 505) 2407  502  2e-141
sw||ATPA_GALSU (P35009) ATP synthase alpha chain    ( 505) 2357  492 2.4e-138
sw||ATPA_ODOSI (Q00820) ATP synthase alpha chain    ( 503) 2352  491 4.8e-138
```

-
-
-

E() value faible

g.deleage@ibcp.fr

FASTA - Classement des « hits » par score décroissant

sw||VATA_PHAAU	(P13548) Vacuolar ATP synthase cat	(622)	230	57	2.5e-07
sw||ATP2_HEVBR	(P29685) ATP synthase beta chain,	(562)	228	56	3.1e-07
sw||VATA_GOSHI	(P31405) Vacuolar ATP synthase cat	(623)	228	56	3.4e-07
sw||VATA_BETVU	(Q39442) Vacuolar ATP synthase cat	(623)	227	56	3.9e-07
sw||VATA_BORBU	(O51121) V-type ATP synthase alpha	(575)	225	56	4.9e-07
sw||VAA2_HUMAN	(P38607) Vacuolar ATP synthase cat	(615)	225	56	5.1e-07
sw||VAA1_DROME	(P48602) Vacuolar ATP synthase cat	(614)	224	56	5.9e-07
sw||VATA_PLAFA	(Q03498) Vacuolar ATP synthase cat	(611)	223	55	6.8e-07
sw||VATA_MANSE	(P31400) Vacuolar ATP synthase cat	(617)	222	55	7.8e-07
sw||VATA_CYACA	(P48414) Vacuolar ATP synthase cat	(587)	220	55	1e-06
sw||VATA_PYRAB	(Q9UXU7) V-type ATP synthase alpha	(1017)	223	56	1e-06
sw||VAA2_DROME	(Q27331) Vacuolar ATP synthase cat	(614)	219	55	1.2e-06
sw||VATA_AEDAE	(O16109) Vacuolar ATP synthase cat	(615)	219	55	1.2e-06
sw||VAA1_PIG	(Q29048) Vacuolar ATP synthase catal	(617)	219	55	1.2e-06
sw||VATA_NEUCR	(P11592) Vacuolar ATP synthase cat	(607)	218	54	1.4e-06
sw||ATPB_DROVI	(Q24751) ATP synthase beta chain,	(228)	211	53	1.7e-06
sw||ATPB_CYTLY	(P13357) ATP synthase beta chain (	(502)	215	54	1.8e-06
sw||VATA_CANTR	(P38078) Vacuolar ATP synthase cat	(1088)	209	53	7.7e-06
sw||VATA_YEAST	(P17255) Vacuolar ATP synthase cat	(1071)	204	52	1.5e-05
sw||ATPB_BACFR	(P13356) ATP synthase beta chain (	(505)	189	48	7.2e-05
sw||VATB_DROME	(P31409) Vacuolar ATP synthase sub	(490)	178	46	0.00034
sw||VATB_CAEEL	(Q19626) Probable vacuolar ATP syn	(491)	178	46	0.00034
sw||VATB_YEAST	(P16140) Vacuolar ATP synthase sub	(517)	178	46	0.00035
sw||VAB1_ACEAT	(Q38681) Vacuolar ATP synthase sub	(492)	176	46	0.00045
sw||VATB_MANSE	(P31401) Vacuolar ATP synthase sub	(494)	172	45	0.00079
sw||VATB_HELVI	(P31410) Vacuolar ATP synthase sub	(494)	172	45	0.00079
sw||VAB2_ACEAT	(Q38680) Vacuolar ATP synthase sub	(492)	171	45	0.00091
sw||ATPB_ASPND	(O03063) ATP synthase beta chain (	(284)	167	44	0.001
sw||RHO_DEIRA	(P52153) Transcription termination	(426)	168	44	0.0012
sw||VAB2_BOVIN	(P31408) Vacuolar ATP synthase sub	(511)	168	44	0.0014
sw||VAB2_HUMAN	(P21281) Vacuolar ATP synthase sub	(511)	168	44	0.0014

E() value intermédiaire

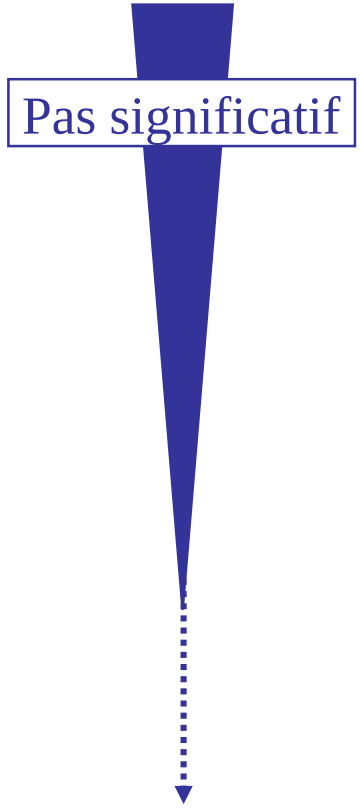
Significatif ?

g.deleage@ibcp.fr

FASTA - Classement des « hits » par score décroissant

sw||ATPB_LONHI (O03073) ATP synthase beta chain (	(208)	148	40	0.012
sw||ATPB_MICPL (O03074) ATP synthase beta chain (	(208)	147	40	0.014
sw||VATB_THETH (Q56404) V-type ATP synthase beta	(478)	151	41	0.015
sw||RHO_CHRVI (P52152) Transcription termination	(433)	150	40	0.016
sw||RHO_AQUAE (O67031) Transcription termination	(436)	144	39	0.038
sw||RHO_MICLU (P52154) Transcription termination	(690)	146	40	0.041
sw||ATPB_OSMCI (O03077) ATP synthase beta chain (	(220)	137	37	0.06
sw||RHO_THEMA (P38527) Transcription termination	(427)	140	38	0.066
sw||RHO_PSEFL (P52155) Transcription termination	(419)	137	38	0.099
sw||RHO_RHOSH (P52156) Transcription termination	(422)	137	38	0.1
sw||ATPB_HYPHO (O03070) ATP synthase beta chain (	(208)	130	36	0.16
sw||RHO_BORBU (P33561) Transcription termination	(419)	132	37	0.2
sw||RHO_BUCAI (P57652) Transcription termination	(419)	131	36	0.23
sw||RHO_NEIGO (Q06447) Transcription termination	(419)	128	36	0.36
sw||RHO_STRLI (P52157) Transcription termination	(707)	130	36	0.4
sw||RHO_HELPJ (Q9ZLS9) Transcription termination	(438)	127	36	0.42
sw||RHO_HELPY (P56466) Transcription termination	(438)	127	36	0.42
sw||RHO_RICPR (Q9ZD24) Transcription termination	(457)	126	35	0.51
sw||AG43_ECOLI (P39180) Antigen 43 precursor (AG4	(1039)	130	37	0.54
sw||RHO_ECOLI (P03002) Transcription termination	(419)	125	35	0.54
sw||RHO_SALTY (P26980) Transcription termination	(419)	125	35	0.54
sw||RHO_HAEIN (P44619) Transcription termination	(420)	124	35	0.63
sw||RHO_TREPA (O83281) Transcription termination	(519)	125	35	0.64
sw||ATPB_DENPU (O03068) ATP synthase beta chain (	(215)	119	34	0.76
sw||YLC7_YEREN (P21212) Hypothetical protein in L	(58)	109	31	1.1
sw||DCDA_ZYMMO (Q9Z661) Diaminopimelate decarboxy	(421)	119	34	1.3
sw||RHO_BACSU (Q03222) Transcription termination	(427)	119	34	1.3
sw||ATPA_BRYMA (P26965) ATP synthase alpha chain	(29)	102	30	1.8
sw||SR54_METJA (Q57565) Signal recognition 54 kDa	(451)	117	34	1.8
sw||TPO_MOUSE (P40226) Thrombopoietin precursor (	(356)	115	33	2
sw||FMT_PASMU (P57949) Methionyl-tRNA formyltrans	(317)	112	32	2.8
sw||RECA_CHLAU (O52394) RecA protein (Recombinase	(351)	111	32	3.4
sw||Z198_HUMAN (Q9UBW7) Zinc finger protein 198 (	(1377)	117	34	4.3
sw||TRBB_RHISN (P55395) Probable conjugal transfe	(325)	105	31	7.6
sw||YI73_AQUAE (O67720) Hypothetical protein AQ_1	(407)	105	31	9.1
sw||SYA_HELPY (P56452) Alanyl-tRNA synthetase (EC	(847)	109	32	9.1
sw||FMT_HAEIN (P44787) Methionyl-tRNA formyltrans	(318)	103	31	9.9

E() value « forte »



g.deleage@ibcp.fr

Sensibilité - Sélectivité /Spécificité

g.deleage@ibcp.fr

Sensibilité - Sélectivité /Spécificité

- **Sensibilité**
  - Capacité de détecter des séquences homologues de façon lointaine (faible score de similarité et faible taux d'identité) Trouve le maximum de « bons » => Peu de faux négatifs

$$Sn=VP/(VP+FN)$$

- **Sélectivité ou spécificité**
  - Capacité d'éviter de trouver des séquences non homologues avec de forts scores de similarités. Minimise le nombre de mauvais => peu de faux positifs

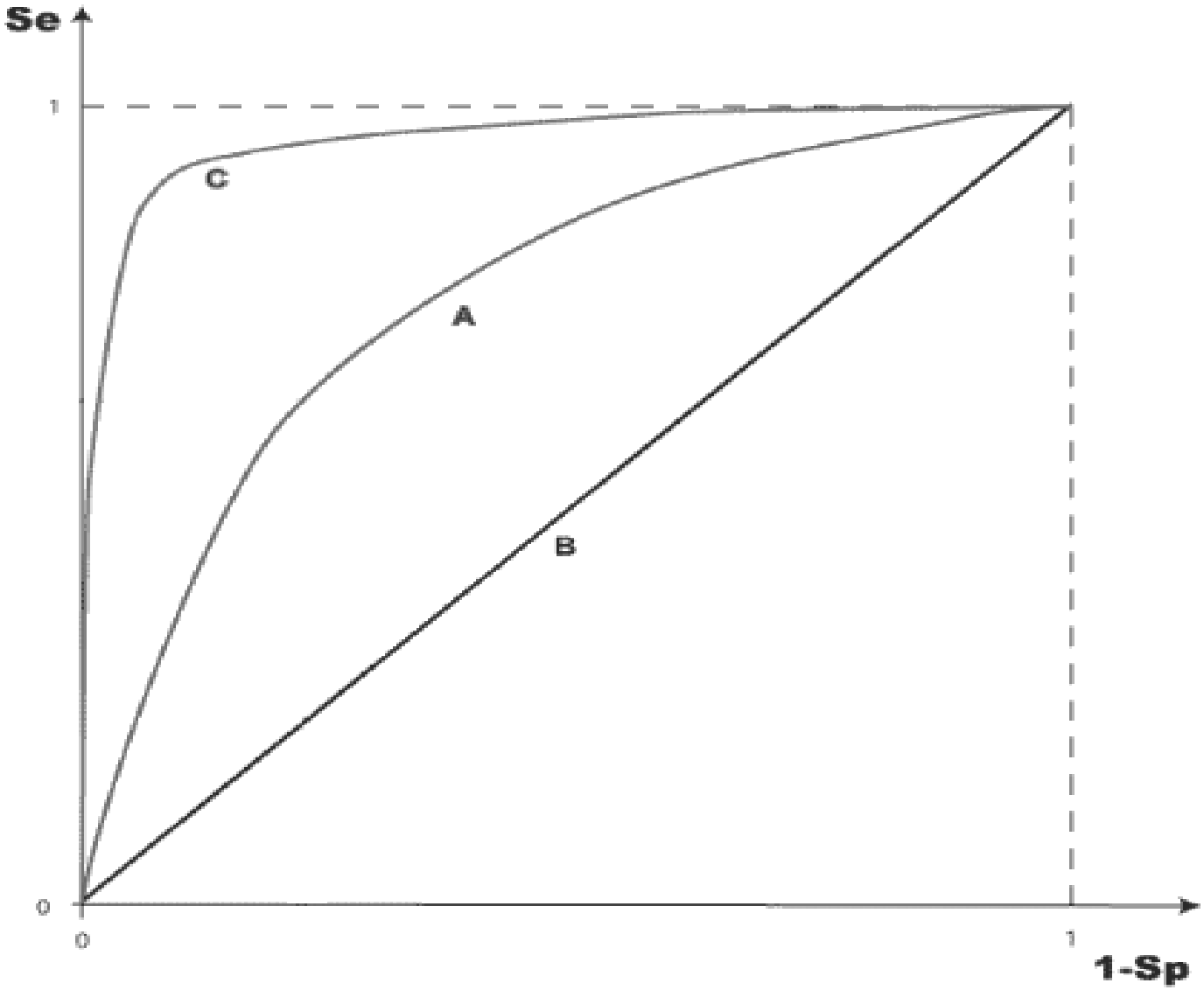
$$Sp=VN/(VN+FP)$$

Sensibilité Selectivité

- **Comparaison des algorithmes de recherche de similarité**
  - Algorithme exhaustif (Smith & Waterman,1981)
  - Algorithmes heuristiques
    - FASTA (Lipman &Pearson)
    - BLAST (Altschul et al.)

g.deleage@ibcp.fr

Sensibilité – Spécificité courbe de ROC



La courbe B montre une méthode aléatoire, la courbe C (superposition faible des 2 courbes) une meilleure méthode que la courbe A (superposition importante des 2 courbes).

g.deleage@ibcp.fr

FASTA – Un alignement


```
>>sw||ATPA_ECOLI (P00822) ATP synthase alpha chain (EC 3  (513 aa)
 initn: 1765 init1: 1143 opt: 1143  Z-score: 1264.3  bits: 243.5 E(): 1.3e-63
Smith-Waterman score: 1738;  55.000% identity (56.468% ungapped) in 500 aa overlap (4-491:3-501)

               10         20         30         40         50         60
QUERY  MVTIRADEISNIIRERIEQYNREVKIVNTGTVLQVGDGIARIHGLDEVMAGELVEFEEGT
          . . ::::.:.::: :.:    .  : ::...:.::. ::::: . : ::.. .   .
sw||AT  MQLNSTEISELIKQRIAQFNVVSEAHNEGTIVSVSDGVIRIHGLADCMQGEMISLPGNR
                10         20         30         40         50

               70         80         90        100        110        120
QUERY  IGIALNLESNNVGVVLMGDGLLIQEGSSVKATGRIAQIPVSEAYLGRVINALAKPIDGRG
        .::::::: ..::::.::     . :: .:: :::: ..:.... :::.:.:. :::::.:
sw||AT YAIALNLERDSVGAVVMGPYADLAEGMKVKCTGRILEVPVGRGLLGRVVNTLGAPIDGKG
       60         70         80         90        100        110

              130        140        150        160        170        180
QUERY  EISASEFRLIESAAPGIISRRSVYEPLQTGLIAIDSMIPIGRGQRELIIGDRQTGKTAVA
        .. . :  .:. :::.: :.:: .:.:::  :.:::::::::::::::::::::::::.:
sw||AT PLDHDGFSAVEAIAPGVIERQSVDQPVQTGYKAVDSMIPIGRGQRELIIGDRQTGKTALA
      120        130        140        150        160        170

              190        200        210        220        230        240
QUERY  TDTILNQQGQNVICVYVAIGQKASSVAQVVTTLQERGAMEYTIVVAETADSPATLQYLAP
        :.:::. ... :::::::::::::::  :.:::.  :::::. ::.  :.:::::::
sw||AT IDAIINQRDSGIKCIYVAIGQKASTISNVVRKLEEHGALANTIVVVATASESAALQYLAP
      180        190        200        210        220        230

              250        260        270        280        290        300
QUERY  YTGAALAEYFMYRERHTLIIYDDPSKQAQAYRQMSLLLRRPPGREAYLGDVFYLHSRLLE
       :.: :.:::::  : . .:::::: :::: ::::::::::::::::. ::::::::::::
sw||AT YAGCAMGEYFRDRGEDALIIYDDLSKQAVAYRQISLLLRRPPGREAFPGDVFYLHSRLLE
      240        250        260        270        280        290

                     310              320        330        340
QUERY  RAAKLSS------SLGE-----GSMTALPIVETQSGDVSAYIPTNVISITDGQIFLSADL
       :::....       . ::     ::.::::::.::::::::.:::::::::::::::  ..:
sw||AT RAARVNAEYVEAFTKGEVKGKTGSLTALPIIETQAGDVSAFVPTNVISITDGQIFLETNL
      300        310        320        330        340        350

      350        360        370        380        390        400
QUERY  FNSGIRPAINVGISVSRVGSAAQIKAMKQVAGKLKLELAQFAELEAFAQFASDLDKATQN
       ::.::::::.: ::::::::::::: : ::...: ..  :::. :: ::.:::::::: ::..
sw||AT FNAGIRPAVNPGISVSRVGGAAQTKIMKKLSGGIRTALAQYRELAAFSQFASDLDDATRK
      360        370        380        390        400        410

      410        420        430        440        450        460
QUERY  QLARGQRLRELLKQSQSAPLTVEEQIMTIYTGTNGYLDSLEVGQVRKFLVELRTYL-KTN
       :: .:::. ::::::.: ::.:: .: ..... ::: ..:.... .: . : .:. . .
sw||AT QLDHGQKVTELLKQKQYAPMSVAQQSLVLFAAERGYLADVELSKIGSFEAALLAYVDRDH
      420        430        440        450        460        470
```

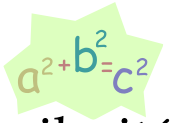

g.deleage@ibcp.fr

Statistiques des scores de similarité

- **Analyse du score de similarité**
  - Calcul de la probabilité que le score obtenu soit fortuit
  - Génération de séquences aléatoires de même longueur et de même composition
  - Recalcul des scores de similarité avec des séquences aléatoires
  - Comparaison de la distribution des scores Attendus <=> Obtenus
- **Similarité fortuite - Bruit de fond - Statistiques**
  - Relation logarithmique entre score (Sb) de similarité basale et la longueur

n la banque)

$$Sb = a + b \ln n$$



$$\text{Z-score} = [Sr - (a + b \ln n] / \sigma$$

S : Similarité réelle

a + b Ln n : Sb

σ = Ecart-type de la distribution de similarité aléatoire

a et b sont calculés à partir des 10000 à 20000 premières séquences en ôtant les séquences ayant les plus forts scores

g.deleage@ibcp.fr

Recherche de séquences homologues par BLAST (Altschul *et al*, 1990)

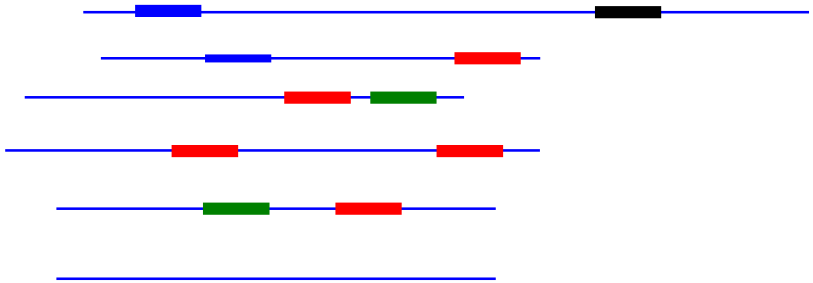
Séquence de longueur L

Liste des (L - W + 1) mots de longueur W= 3 protéines ou longueur W=11 nucléiques

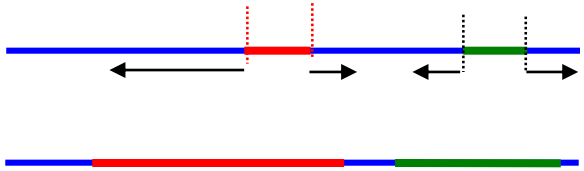
Séquences de la banque

Liste de mots

Identification des mots exacts dans la banque



Extension des mots tant que le Z-score est supérieur à une valeur seuil



Itération éventuelle (PSI-BLAST)

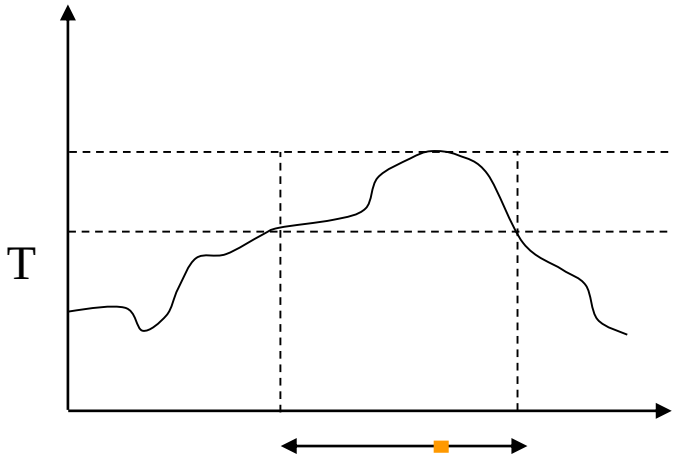
g.deleage@ibcp.fr

Exemple BLAST (Altschul *et al*, 1990)

Liste de mots proches T = Somme de S(i)

S(L,L) = 7
S(Q,R) = 1
S(G,G) = 6
S(Q,Q) = 5

S L A A L L N K C K T L Q G Q R L V N Q W



L Q G	18
L E G	15
L R G	14
L K G	14
L N G	13
L D G	13
L H G	13
L M G	13
L S G	13
L Q A	12
L Q N	12

Score seuil *T* = 13

```
Query : 325  S L A A L L N K C K T L Q G Q R L V N Q W  345
             + L A + + L +       T L   G   R + +   + W
Sbjct : 290  T L A S V L D C T V T L M G S R M L K R W  310
```


g.deleage@ibcp.fr

Scores de recherche dans les banques

1. Sb dépend de la taille s de la banque **Sb = a + b ln s**
2. Sb est déduit de la pente b et ordonnée à l'origine a estimés sur un jeu de séquences réelles de la banque (sur 10k or 20 k seqs avec des scores faibles)
3. La banque est considéré comme une séquence virtuelle de longueur N

$$E() = K\ m\ N\ e^{-\lambda Sr}$$

N taille de la banque, m longueur de la séquence,
K et λ sont des constantes,
m longueur de la séquence requête,
Sr score

- E() est utilisée pour l'extension des mots tant que

  E (expected) < E (threshold) => HSP High Scoring Pair

- Avec une distribution de Poisson, la P-value est la probabilité que 2 séquences aient un score>= Sr

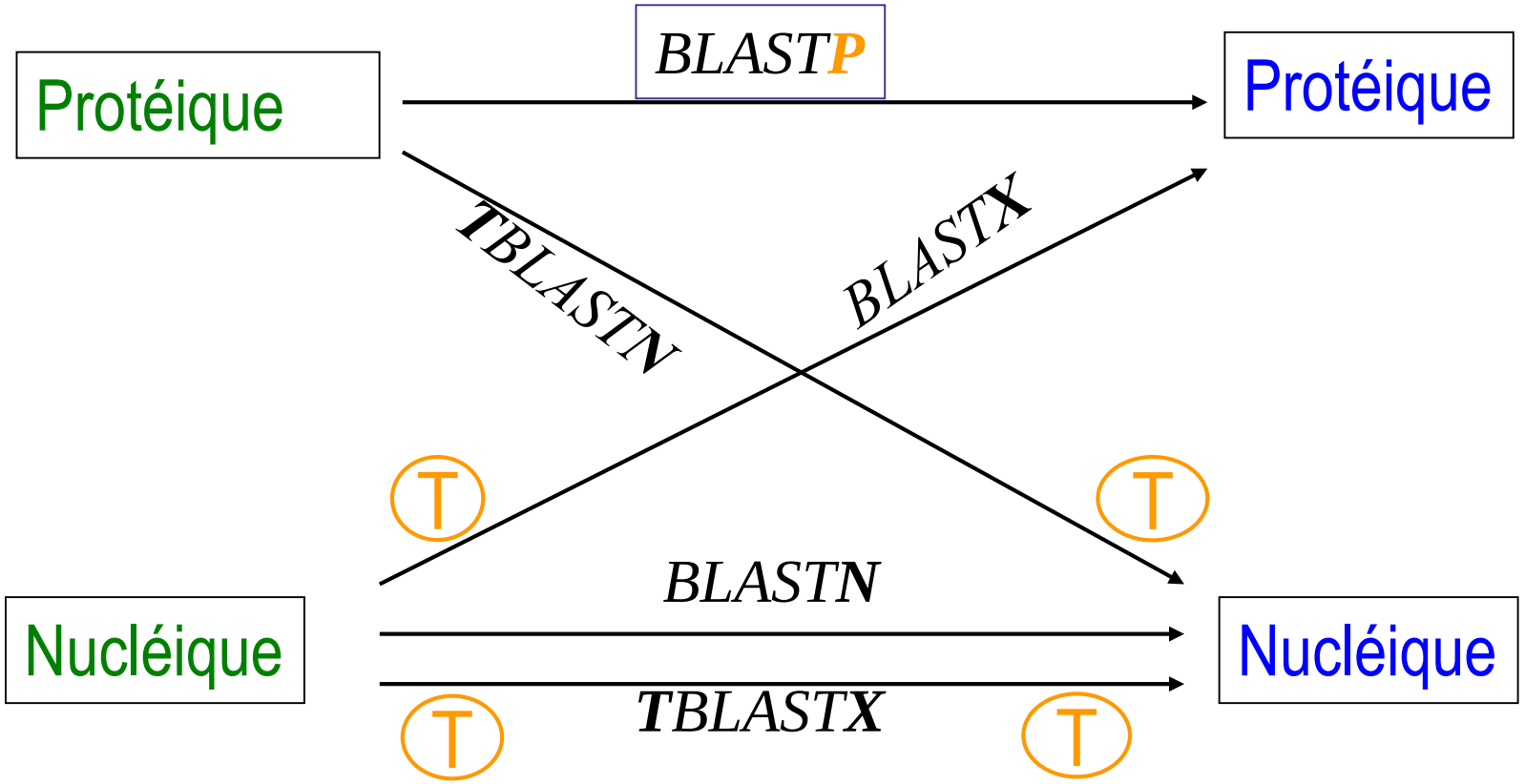
$$p\text{-value} = 1 - e^{-E()}$$

g.deleage@ibcp.fr

Différents versions de programmes de recherche de similarité

http://npsa-prabi.ibcp.fr

SEQUENCE **BANQUE**



T Traduction en 6 phases

http://www.ncbi.nlm.nih.gov/blast/

g.deleage@ibcp.fr

Utilisation de BLAST

Rechercher la localisation chromosomique d'un gène d'une protéine humaine

>xxxxx_HUMAN

MGNRGMEDLIPLVNRLQDAFSAIGQNADLDLPQIAVVGGQSAGKSSVLENFVGRDFLPRGSG
IVTRRPLVLQLVNATTEYAEFLHCKGKKFTDFEEVRLEIEAETDRVTGTNKGISPVPI
NLRVYSPHVLNLTLVDLPGMTKVPVGDQPPDIEFQIRDMLMQFVTKENCLILAVSPANSDLA
NSDALKVAKEVDPQGQRTIGVITKLDLMDEGTDARDVLENKLLPLRRGYIGVVNRSQK
DIDGKKDITAALAAERKFFLSHPSYRHLADRMGTPYLQKVLNQQLTNHIRDTLPGLRNKLQS
QLLSIEKEVEEYKNFRPDDPARKTKALLQMVQQFAVDFEKRIEGSGDQIDTYELSGGA
RINRIFHERFPFELVKMEFDEKELRREISYAIKNIHGIRTGLFTPDMAFETIVKKQVKKIRE
PCLKCVDMVISELISTVRQCTKKLQQYPRLREEMERIVTTHIREREGRTKEQVMLLID
IELAYMNTNHEDFIGFANAQQRSNQMNKKKTSGNQDEILVIRKGWLTINNIGIMKGGSKEYW
FVLTAENLSWYKDDEEKEKKYMLSVDNLKLRDVEKGFMSSKHIFALFNTEQRNVYKDY
RQLELACETQEEVDSWKASFLRAGVYPERVGDKEKASETEENGSDSFMHSMDPQLERQVETI
RNLVDSYMAIVNKTVRDLMPKTIMHLMINNTKEFIFSELLANLYSCGDQNTLMEESAE
QAQRRDEMLRMYHALKEALSIIGNINTTTVSTPMPPPVDDSWLQVQSVPAGRRSPTSSPTPQ
RRAPAVPPARPGSRGPAPGPPPAGSALGGAPPVPSRPGASPDPFGPPPQVPSRPNRAP
PGVPSRSGQASPSRPESPRPPFDL

g.deleage@ibcp.fr

Rechercher la localisation chromosomique d'une protéine
http://www.ensembl.org

Ensembl Human

The Wellcome Trust Sanger Institute

EBI

Human Genome Browser

Ensembl Entry Points

Search for: Anything — with: Lookup

Display Chr: 1 — From: 1 — To: 100000 — Lookup

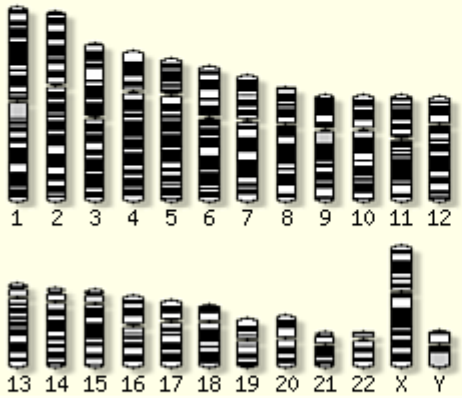
Retrieve a sequence: Export

BLAST your sequence: Blast

Advanced data retrieval tool: EnsMart

For fast identity search try: SSAHA

Browse a Chromosome



Current Release 12.31.1

This release is based on the NCBI 31 assembly of the human genome.

View the status history of the human assemblies.

Last Update: 27-02-2003

Ensembl gene predictions : 24847

Genscan gene predictions : 58770

Ensembl gene exons : 248449

Pre! Finished Human Assembly

The human Ensembl pre-build site provides a preliminary data set based on the Human NCBI33 assembly. **This is the first essentially complete assembly of the human genome.**

Please note that the site shows only the DNA sequence, initial gene placement, repeatmasking and raw BLAST hits on this genome. The annotated assembly will be released on the main ensembl site at the begining of July 2003.

View pre-build Human NCBI33 at pre.ensembl.org.

Documentation & Help

About Ensembl — e!Home

g.deleage@ibcp.fr

Ensembl Blast server

Ensembl BLAST Server

Alternative sequence search: SSAHA

RETRIEVE BLAST RESULTS Help

Enter the blast retrieval ID: [ ] Retrieve

SUBMIT A BLAST QUERY Help

Paste your sequence here in FASTA or plain text format.

```
VMAGELVEFEEGT
IGIALNLESNNVGVVLMGDGLLIQEGSSVKATGRI
AQIPVSEAYLGRVINALAKPIDGRG
EISASEFRLIESAAPGIISRRSVYEPLQTGLIAID
SMIPIGRGQRELIIGDRQTGKTAVA
TDTILNQQGQNVICVYVAIGQKASSVAQVVTT
```

Search Rétablir

OR select the sequence file you wish to search [ ] Parcourir...

BLAST OPTIONS Help

Database: Homo sapiens, peptides (genscan)

Executable: BLASTP (protein vs. protein)

Report 100 alignments.

- [x] Mask repetitive sequences using Repeatmasker.
- [x] Filter low complexity regions.
- [ ] Display histogram of score statistics.

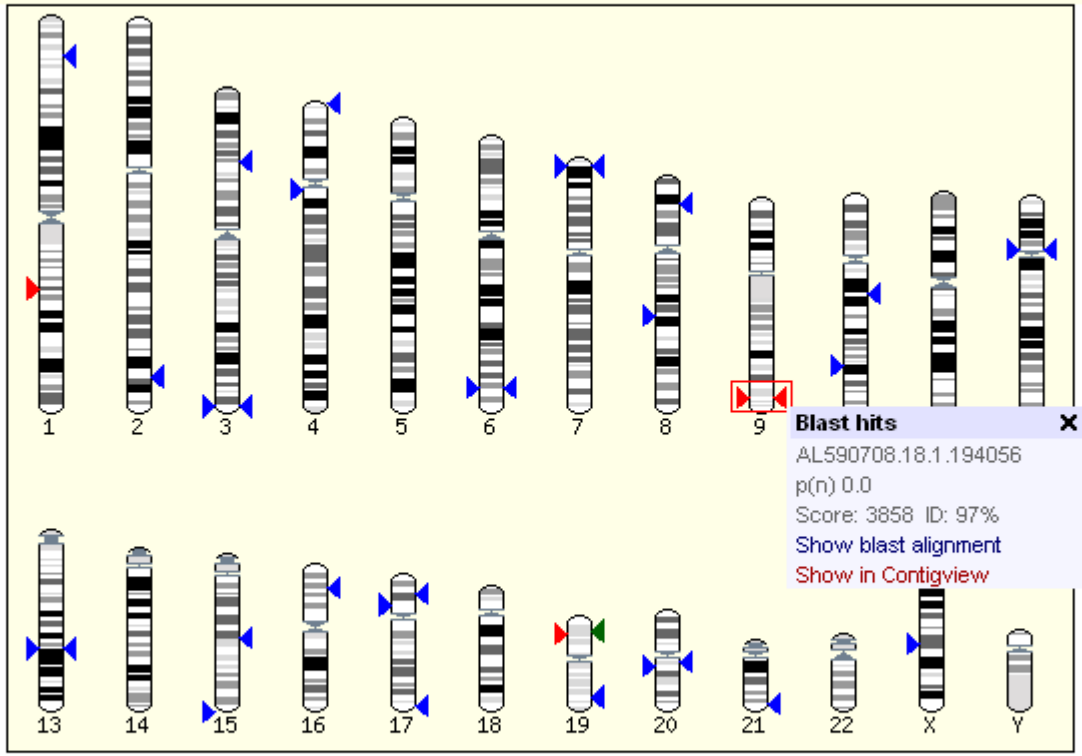
ADVANCED BLAST OPTIONS Help

Matrix: blosum62 Expect (E): 10

g.deleage@ibcp.fr

Ensembl Blast server

Ensembl Human BLASTView



Blast score ranges for this search: [The highest scoring hit(s) are boxed]

37 - 1310 | 1311 - 2584 | 2585 - 3858

Location of Blast hits

Pourquoi la séquence ne présente que 97% d'identité après alignement?

g.deleage@ibcp.fr

Ensembl Blast server

>xxxxx_HUMAN

MGNRGMEDLIPLVNRLQDAFSAIGQNADLDLPQIAVVGGQSAGKSSVLENFVGRDFLPRGSG
IVTRRPLVLQLVNATTEYAEFLHCKGKKFTDFEEVRLEIEAETDRVTGTNKGISPVPI
NLRVYSPHVLNLTLVDLPGMTKVPVGDQPPDIEFQIRDMLMQFVTKENCLILAVSPANSDLA
NSDALKVAKEVDPQGQRTIGVITKLDLMDEGTDARDVLENKLLPLRRGYIGVVNRSQK
DIDGKKDITAALAAERKFFLSHPSYRHLADRMGTPYLQKVLNQQLTNHIRDTLPGLRNKLQS
QLLSIEKEVEEYKNFRPDDPARKTKALLQMVQQFAVDFEKRIEGSGDQIDTYELSGGA
RINRIFHERFPFELVKMEFDEKELRREISYAIKNIHGIRTGLFTPDMAFETIVKKQVKKIRE
PCLKCVDMVISELISTVRQCTKKLQQYPRLREEMERIVTTHIREREGRTKEQVMLLID
IELAYMNTNHEDFIGFANAQQRSNQMNKKKTSGNQDEILVIRKGWLTINNIGIMKGGSKEYW
FVLTAENLSWYKDDEEKEKKYMLSVDNLKLRDVEKGFMSSKHIFALFNTEQRNVYKDY
RQLELACETQEEVDSWKASFLRAGVYPERVGDKEKASETEENGSDSFMHSMDPQLERQVETI
RNLVDSYMAIVNKTVRDLMPKTIMHLMINNTKEFIFSELLANLYSCGDQNTLMEESAE
QAQRRDEMLRMYHALKEALSIIGNINTTTVSTPMPPPVDDSWLQVQSVPAGRRSPTSSPTPQ
RRAPAVPPARPGSRGPAPGPPPAGSALGGAPPVPSRPGASPDPFGPPPQVPSRPNRAP
PGVPSRSGQASPSRPESPRPPFDL

g.deleage@ibcp.fr

Principe de l'alignement de 2 séquences

- **2 Séquences à comparer: mot1 de longueur n1=9; mot2 de longueur n2=8 (n1>n2)**

```
mot1   A G V S I L N Y A   Identité = 0 longueur = 9 %id=0
mot2   V S I L Y A K R
```

- **Alignement optimum par glissement g1=2**

```
A G V S I L N Y A     Identités = 4 longueur = 10 %id=40
    * * * *
    V S I L Y A K R
```

- **Alignement optimum par glissement g2=3**

```
A G V S I L N Y A     Identités = 2 longueur = 11 %id=18
          * *
    V S I L Y A K R
```

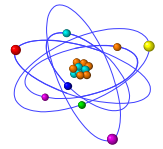
- **Alignement optimum avec insertion**

```
A G V S I L N Y A     Identités = 6 longueur = 11 %id=55
    * * * *   * *
    V S I L - Y A K R       Score 5
```

Score d'identité : identité=+1
Pénalité d'insertion : gap -1

$$N(l) = \binom{2l}{l}$$

N(25) ~ 10^{14}
N(50) = 10^{28}
N(100) ~ 10^{58}
N(200) ~10^{120}
N(350) ~1.6^{209}



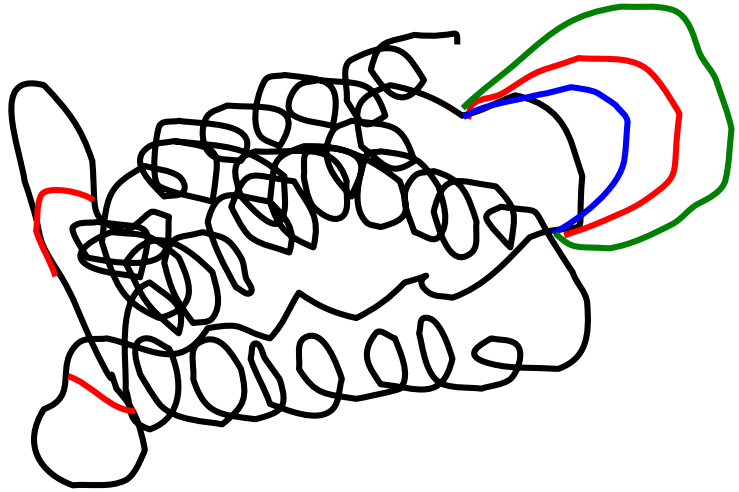
g.deleage@ibcp.fr

Insertions-délétions et évolution

- **Un événement évolutif**

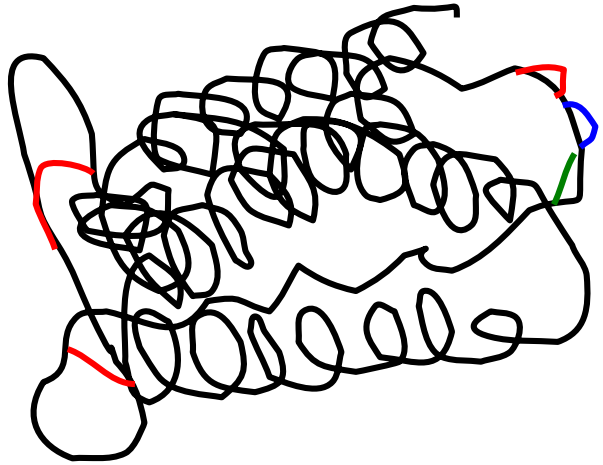
  Insertion d'un acide aminé (une lettre dans le mot)

  Insertion ou suppression de plusieurs acides aminés (ajout d 'un sous mot)



Total 3 événements évolutifs
quelle que soit la longueur du sous-mot

- **Plusieurs insertions délétions = Plusieurs événements évolutifs**



Total 5 événements évolutifs

g.deleage@ibcp.fr

Pénalité pour les Indels (Insertion-Délétion)

- **Pénalité fixe par indel (exemple -1)**
- **Pénalité fixe pour un indel x + pénalité variable y pour extension d'indel avec x>y.**
  - Cette dernière pénalité y est moins lourde, mais permet de prendre en compte la longueur de l'indel. L'expression de cette pénalité à deux paramètres peut être décrite par la fonction affine suivante:

$$P = x + y\,L$$

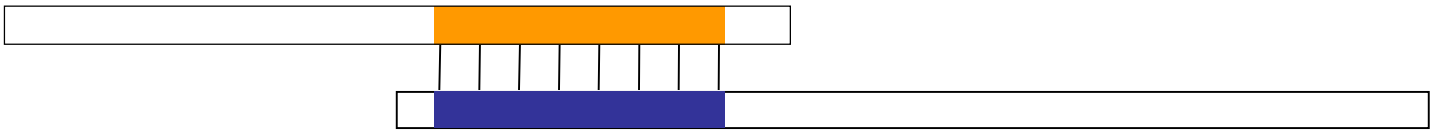
  - où P est la pénalité pour une insertion de longueur L
    - x la pénalité fixe d'insertion indépendante de la longueur
    - y la pénalité d'extension pour un élément (souvent x = 10 y)
  - Une longue insertion est légèrement plus pénalisante qu'une courte, ce qui revient en fait à minimiser le poids de la longueur des insertion par rapport à l'introduction même d'une insertion.
- **Pénalité variable en fonction de la localisation (structure secondaire)**

g.deleage@ibcp.fr

Alignement global (Needlman & Wunsch, 1970)


```
G G C T G A C C A C C - T T
|   |   | |     | |   |
G A - T C A C T T C C A T G
```

Alignement local (Smith & Waterman, 1981)




```
G G C T G A C C A C C T T
    | |   | | |   | |
    G A T C A C - T T C C A T G
```

Algorithme de Needleman et Wunsch 1

- Algorithme général de comparaison globale de séquences
- Maximise un score de similarité => Accord maximum

ou

- Différence minimale (ou minimise les différences)
- Accord maximum = le plus grand nombre de résidus d'une séquence qui peut correspondre à une autre séquence en autorisant des gaps
- Trouve l'alignement optimal entre 2 séquences
- Calcul itératif d'une méthode matricielle qui calcule:
  - Toutes les paires possibles (base ou aa) sont présentés sous forme d'un tableau 2D
  - Tous les alignements sont représentés par des chemins dans le tableau

g.deleage@ibcp.fr

Algorithme de Needleman et Wunsch 2

- 3 étapes

1. Assignation des scores de similarité et de pénalisation (substitution, insertion-délétion)

Cas simple :

- identité = +2 (favorable)
- substitution = -1 (défavorable)
- insertion-délétion = -1 (défavorable)

2. Pour chaque case, calculer les scores (en autorisant insertion et délétion)

3. Construit un chemin en retour depuis la case de plus fort score pour donner le gain maximal de l'alignement

g.deleage@ibcp.fr

Algorithme de Needleman et Wunsch
Programmation dynamique

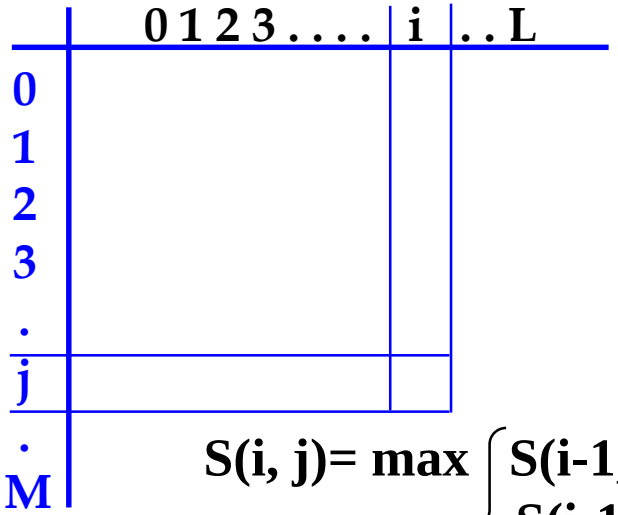
```
tableaux : S1[N], S2[M], Matrice[N][M]

S1 <-- séquence 1                                  /* N caractères */
S2 <-- séquence 2                                  /* M caractères */
PENALITE=-1                                        /* PENALITE pour indel*/
Matrice[i][0] <-- - (i x -GAP)                     /* init matrice */
Matrice[0][j] <-- - (j x -GAP)
SUBS[S1[i],S2[j]]=+2 si S1[i]=S2[j]
SUBS[S1[i],S2[j]]=-1 S1[i]<>S2[j]

 pour j=1 jqa M faire
  {
    pour i=1 jqa N faire
     {
                                  | Matrice[i][j-1] + PENALITE
       Matrice[i][j] <-- MAX      | Matrice[i-1][j-1] + SUBS[S1[i]],[S2[j]]
                                  | Matrice[i-1][j] + PENALITE
     }
  }
```

Complexité : NxM

g.deleage@ibcp.fr

Principe de l'alignement de 2 séquences

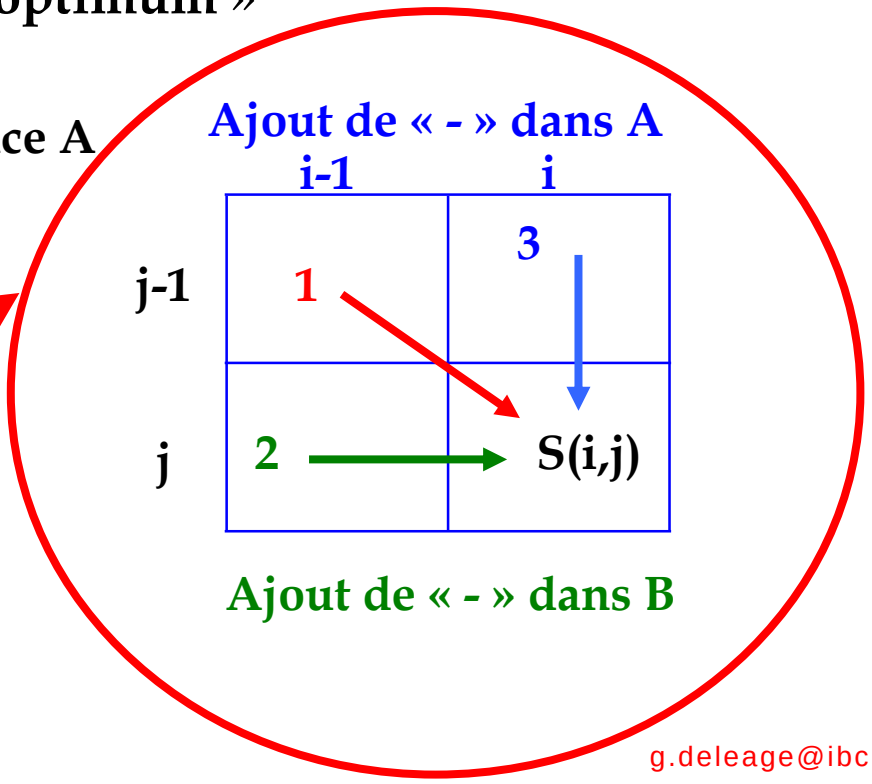
- **Programmation dynamique (solutions optimales)**
- **Objectif**
  - Maximiser les identités (en minimisant les insertions -délétions indel) entre les 2 séquences. Dans une séquence, une délétion d'aa est réprésenté par « – ».
- **Principe**
  - Mode de représentation de séquences au moyen de substitutions, insertions ou délétions.
  - Remplissage pas à pas d'une « **matrice de score** » LxM, de la case (1,1) à la case (L,M) et d 'une matrice du « **chemin optimum** »
  - Ligne par ligne de gauche à droite



Séquence A

Case i, j

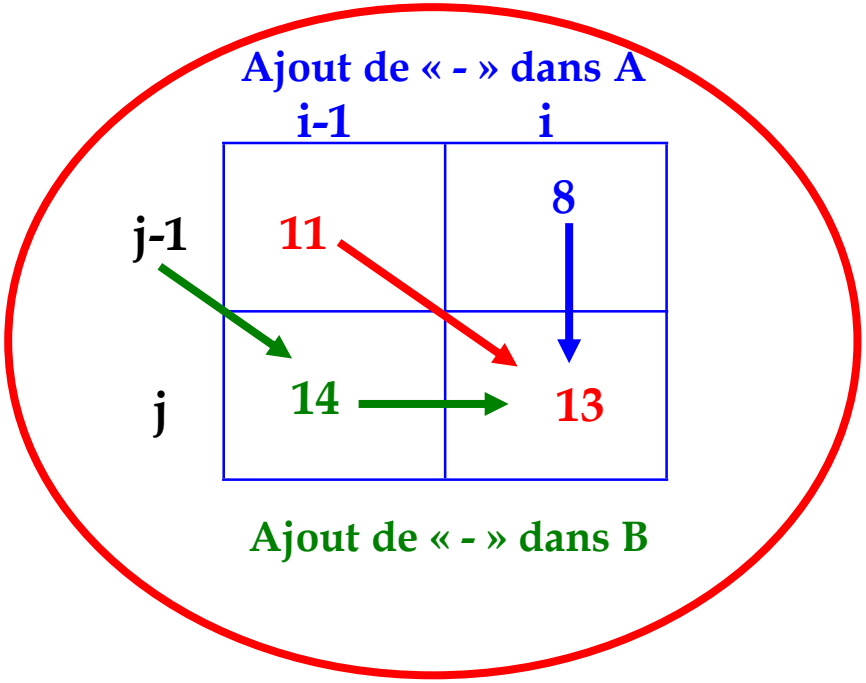
$$S(i,j) = \max \begin{cases} S(i-1, j-1) + s(a_i, b_j) \\ S(i-1, j) + s(a_i, -) \\ S(i, j-1) + s(-, b_j) \end{cases}$$

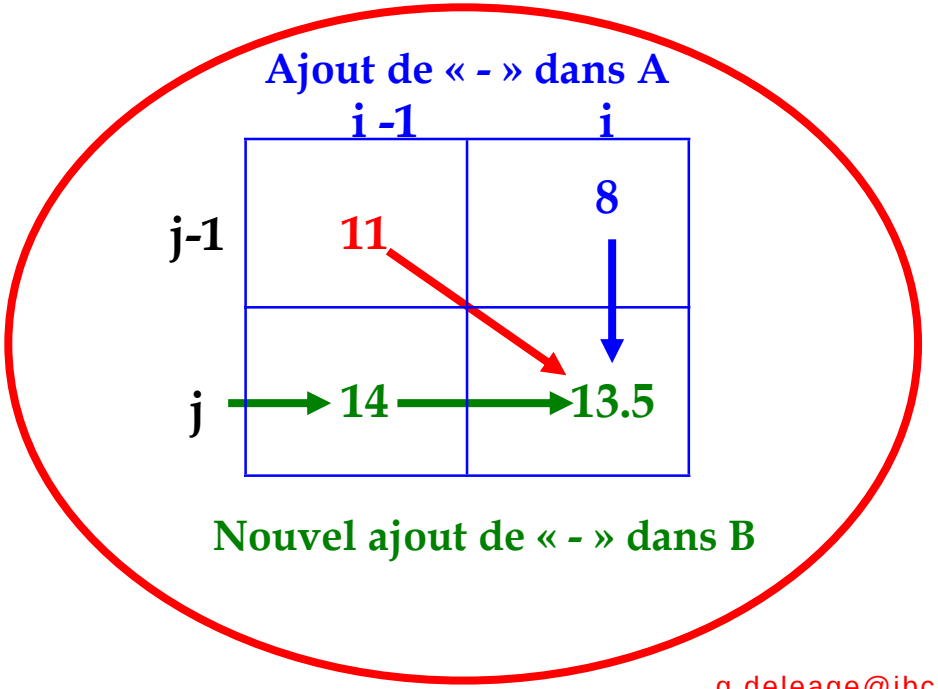


g.deleage@ibcp.fr

Principe de l'alignement de 2 séquences

- **Calcul des 3 scores S1, S2 et S3 correspondant aux 3 façons d'obtenir le score S(i,j)**
  - **S1 S(i,j) =S(i-1, j-1) + Bonus si i = j (+2)**
    Pénalité< 0 si i ≠ j (-1)
  - **S2 S(i,j)=S(i-1, j) + Pénalité1 = -1 lors de la création d'un gap**
    Pénalité2 = - 0.5 extension d'un gap
  - **S3 S(i,j)=S(i,j-1) + Pénalité (idem 2)**
- **S(i, j) = max (S1, S2, S3)**
  - En cas d'égalité, la diagonale est favorisée





g.deleage@ibcp.fr

Algorithme de Needleman et Wunsch

identité = +3

substitution = -1

indel = -2

Initialisation

		M	P	R	C	L	C	Q	R	I	N	C	Y	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22	-24	-26
P	-2	?												
Y	-4													
R	-6													
C	-8													
K	-10													
C	-12													
R	-14													
N	-16													
I	-18													
C	-20													
I	-22													
A	-24													

g.deleage@ibcp.fr

Algorithme de Needleman et Wunsch

identité = +3

substitution = -1

indel = -2

Initialisation

		M	P	R	C	L	C	Q	R	I	N	C	Y	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22	-24	-26
P	-2	-1	?											
Y	-4													
R	-6													
C	-8													
K	-10													
C	-12													
R	-14													
N	-16													
I	-18													
C	-20													
I	-22													
A	-24													

g.deleage@ibcp.fr

Algorithme de Needleman et Wunsch

identité = +3

substitution = -1

indel = -2

Initialisation

		M	P	R	C	L	C	Q	R	I	N	C	Y	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22	-24	-26
P	-2	-1	1											
Y	-4													
R	-6													
C	-8													
K	-10													
C	-12													
R	-14													
N	-16													
I	-18													
C	-20													
I	-22													
A	-24													

g.deleage@ibcp.fr

Algorithme de Needleman et Wunsch

identité = +3

substitution = -1

indel = -2

Scores

		M	P	R	C	L	C	Q	R	I	N	C	Y	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22	-24	-26
P	-2	-1	1	-1	-3	-5	-7	-9	-11	-13	-15	-17	-19	-21
Y	-4	-3	-1	0	-2	-4	-6	-8	-10	-12	-14	-16	-14	-16
R	-6	-5	-3	2	0	-2	-4	-6	-5	-7	-9	-11	-16	-15
C	-8	-7	-5	0	5	3	1	-1	-3	-5	-7	-6	-8	-10
K	-10	-9	-7	-2	3	4	2	0	-2	-4	-6	-8	-7	-9
C	-12	-11	-9	-4	1	2	7	5	3	1	-1	-3	-5	-7
R	-14	-13	-11	-6	-1	0	5	6	8	6	4	2	0	-2
N	-16	-15	-13	-8	-3	-2	3	4	6	7	9	7	5	3
I	-18	-17	-15	-10	-5	-4	1	2	4	9	7	8	6	4
C	-20	-19	-17	-12	-7	-6	-1	0	2	7	8	10	8	6
I	-22	-21	-19	-14	-9	-8	-3	-2	0	5	6	8	9	7
A	-24	-23	-21	-16	-11	-10	-5	-4	-2	3	4	6	7	12

g.deleage@ibcp.fr

Algorithme de Needleman et Wunsch

identité = +3

substitution = -1

indel = -2

Chemins

		M	P	R	C	L	C	Q	R	I	N	C	Y	A
		←	→	→	→	→	→	→	→	→	→	→	→	→
P	↓	↘	↘	→	→	→	→	→	→	→	→	→	→	→
Y	↓	↘	↓	↘	↘	↘	↘	↘	↘	↘	↘	↘	↘	→
R	↓	↘	↓	↘	→	→	→	→	↘	→	→	→	↓	↘
C	↓	↘	↓	↓	↘	→	↘	→	→	→	→	↘	→	→
K	↓	↘	↓	↓	↓	↘	↘	↘	↘	↘	↘	↘	↘	↘
C	↓	↘	↓	↓	↘	↘	↘	←	→	→	→	↘	→	→
R	↓	↘	↓	↘	↓	↘	↓	↘	↘	→	→	→	→	→
N	↓	↘	↓	↓	↓	↘	↓	↘	↓	↘	↘	→	→	→
I	↓	↘	↓	↓	↓	↘	↓	↘	↓	←	→	↘	↘	↘
C	↓	↘	↓	↓	↘	↘	↘	↘	↓	↓	↘	↘	→	→
I	↓	↘	↓	↓	↓	↘	↓	↘	↓	↘	↘	↓	↘	↘
A	↓	↘	↓	↓	↓	↘	↓	↘	↓	↓	↘	↓	↘	↘

Alignement

```
MP-RCLCQR-INCYA
 | || | | | | |
-PYRCKC-RNI-CIA
```


g.deleage@ibcp.fr

Algorithme de Needleman et Wunsch

identité = +3

substitution = -1

indel = -2

Chemins

		M	P	R	C	L	C	Q	R	I	N	C	Y	A
		←	→	→	→	→	→	→	→	→	→	→	→	→
P	↓	↘	↘	→	→	→	→	→	→	→	→	→	→	→
Y	↓	↘	↓	↘	↘	↘	↘	↘	↘	↘	↘	↘	↘	→
R	↓	↘	↓	↘	→	→	→	→	↘	→	→	→	↓	↘
C	↓	↘	↓	↓	↘	→	↘	→	→	→	→	↘	→	→
K	↓	↘	↓	↓	↓	↘	↘	↘	↘	↘	↘	↘	↘	↘
C	↓	↘	↓	↓	↘	↘	↘	←	→	→	→	↘	→	→
R	↓	↘	↓	↘	↓	↘	↓	↘	↘	→	→	→	→	→
N	↓	↘	↓	↓	↓	↘	↓	↘	↓	↘	↘	→	→	→
I	↓	↘	↓	↓	↓	↘	↓	↘	↓	↘	→	↘	↘	↘
C	↓	↘	↓	↓	↘	↘	↘	↘	↓	↓	↘	↘	→	→
I	↓	↘	↓	↓	↓	↘	↓	↘	↓	↘	↘	↓	↘	↘
A	↓	↘	↓	↓	↓	↘	↓	↘	↓	↓	↘	↓	↘	↘

Alignement

```
MP-RCLCQRIN-CYA
 |  ||  |  |  |  |
-PYRCKC-R-NICIA
```


g.deleage@ibcp.fr

Algorithme de Needleman et Wunsch

identité = +3

substitution = -1

indel = -2

Scores

		M	P	R	C	L	C	Q	R	I	N	C	Y	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22	-24	-26
P	-2	-1	1	-1	-3	-5	-7	-9	-11	-13	-15	-17	-19	-21
Y	-4	-3	-1	0	-2	-4	-6	-8	-10	-12	-14	-16	-14	-16
R	-6	-5	-3	2	0	-2	-4	-6	-5	-7	-9	-11	-16	-15
C	-8	-7	-5	0	5	3	1	-1	-3	-5	-7	-6	-8	-10
K	-10	-9	-7	-2	3	4	2	0	-2	-4	-6	-8	-7	-9
C	-12	-11	-9	-4	1	2	7	5	3	1	-1	-3	-5	-7
R	-14	-13	-11	-6	-1	0	5	6	8	6	4	2	0	-2
N	-16	-15	-13	-8	-3	-2	3	4	6	7	9	7	5	3
I	-18	-17	-15	-10	-5	-4	1	2	4	9	7	8	6	4
C	-20	-19	-17	-12	-7	-6	-1	0	2	7	8	10	8	6
I	-22	-21	-19	-14	-9	-8	-3	-2	0	5	6	8	9	7
A	-24	-23	-21	-16	-11	-10	-5	-4	-2	3	4	6	7	12

g.deleage@ibcp.fr

Algorithme de Needleman et Wunsch

identité = +3

substitution = -1

indel = -2

Chemins alternatifs

		M	P	R	C	L	C	Q	R	I	N	C	Y	A
		→	→	→	→	→	→	→	→	→	→	→	→	→
P	↓	↘	↘	→	→	→	→	→	→	→	→	→	→	→
Y	↓	↘	↓	↘	↘	↘	↘	↘	↘	↘	↘	↘	↘	→
R	↓	↘	↓	↘	→	→	→	→	↘	→	→	→	↓	↘
C	↓	↘	↓	↓	↘	→	↘	→	→	→	→	↘	→	→
K	↓	↘	↓	↓	↓	↘	↘	↘	↘	↘	↘	↘	↘	↘
C	↓	↘	↓	↓	↘	↘	↘	→	→	→	→	↘	→	→
R	↓	↘	↓	↘	↓	↘	↓	↘	↘	→	→	→	→	→
N	↓	↘	↓	↓	↓	↘	↓	↘	↓	↘	↘	→	→	→
I	↓	↘	↓	↓	↓	↘	↓	↘	↓	↘	→↓	↘	↘	↘
C	↓	↘	↓	↓	↘	↘	↘	↘	↓	↓	↘	↘	→	→
I	↓	↘	↓	↓	↓	↘	↓	↘	↓	↘	↘	↓	↘	↘
A	↓	↘	↓	↓	↓	↘	↓	↘	↓	↓	↘	↓	↘	↘

Alignements équivalents

```
MP-RCLCQR-INCYA
 |  ||  |  |  |  |  |
-PYRCKC-RNI-CIA
```

```
MP-RCLCQRIN-CYA
 |  ||  |  |  |  |  |
-PYRCKC-R-NICIA
```

g.deleage@ibcp.fr

Initialisation des premières lignes et colonnes

- **Alignement global optimal (NW)**
  - Initialisation avec la valeur de la délétion à la position dans la séquence
  - Revenir de (n,m) à (0,0) en favorisant la diagonale
- **Alignements globaux sub optimaux**
  - Initialisation des lignes et colonnes avec des 0.
  - Déplacements différents en fonction de l'objectif de l'alignement.
  - Exemples:
    - Recherche de la plus grande sous-séquence ressemblante.
    - Trouver le meilleur chevauchement entre 2 séquences (reconstruction de séquences à partir de fragments cartes de séquençage)
- **Alignements locaux**
  - Recherche de la meilleure similitude locale (Smith & Waterman, 1981)
  - Paramétrage
    - Score >0 caractère identique (+10)
    - Score <0 si mutation (-9) ou insertion-délétion (-20)
  - La zone la plus similaire est celle du plus fort score qui finit la zone (le début a la valeur 0 sur la même diagonale)

g.deleage@ibcp.fr

Algorithme de Smith & Watermann

identité = +2

substitution = 0

indel = -1

		L	I	B	R	E	S	E	Q	E	N	C	E
	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	2	1	0	0	0	0	0
E	0	0	0	0	0	2	1	4	3	2	1	0	2
Q	0	0	0	0	0	1	2	3	6	5	4	3	1
A	0	0	0	0	0	0	1	2	5	6	5	4	3
N	0	0	0	0	0	0	0	1	4	5	8	7	6
C	0	0	0	0	0	0	0	0	3	4	7	10	9
E	0	0	0	0	0	2	1	2	2	5	6	9	12
L	0	2	1	0	0	1	2	1	2	4	5	8	11
I	0	1	4	3	2	1	1	2	1	3	4	7	10
B	0	0	3	6	5	4	3	2	2	2	3	6	9
R	0	0	2	5	8	7	6	5	4	3	2	5	8
E	0	0	1	4	7	10	9	8	7	6	5	4	7

```
Seq1 -------LIBRESEQENCE
Seq2 SEQANCELIBRE-------
      *****
```

```
Seq1 LIBRESEQENCE-----
Seq2 -----SEQANCELIBRE
      *** ***
```

g.deleage@ibcp.fr

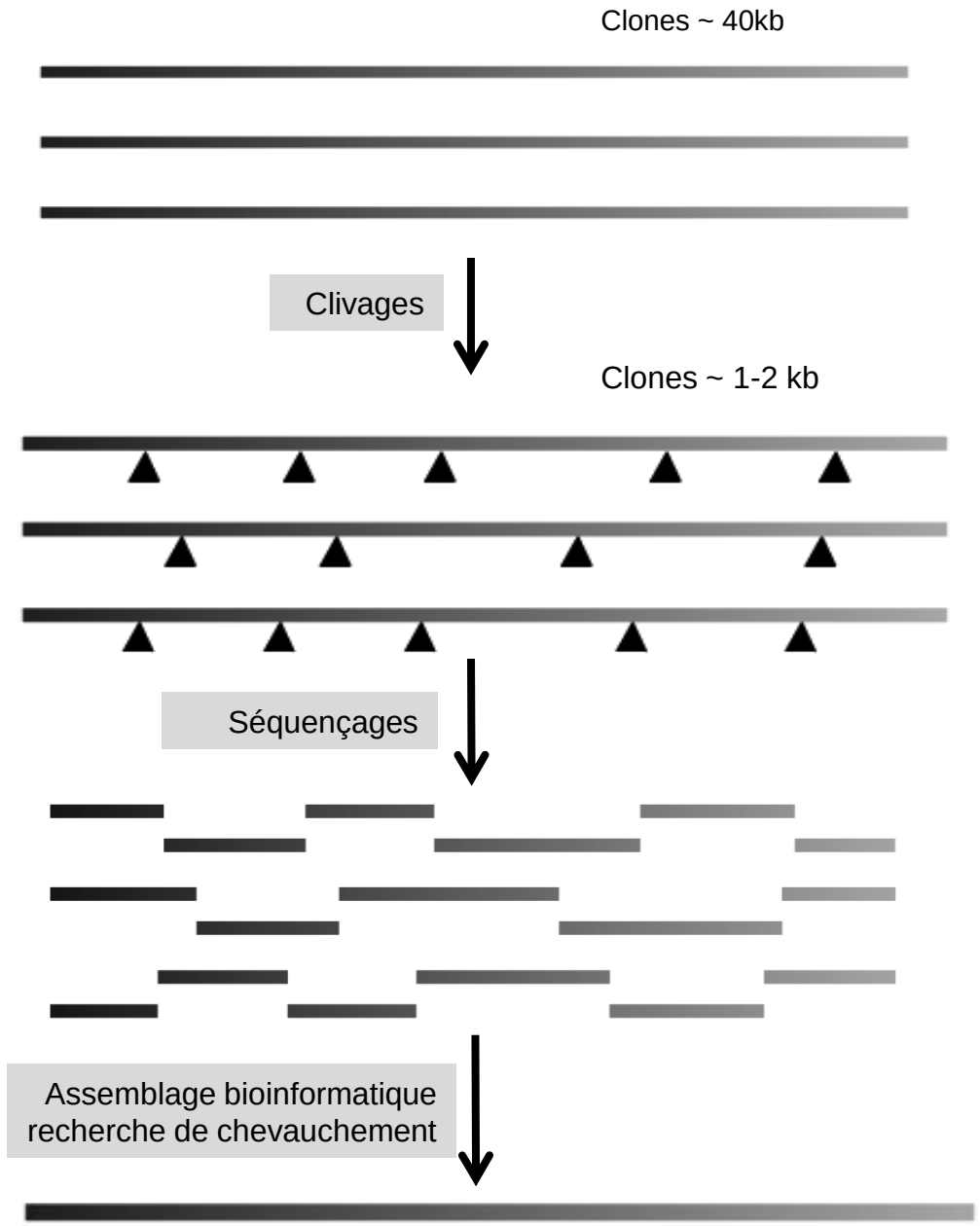
Algorithme de Smith & Watermann

		L	I	B	R	E	S	E	Q	E	N	C	E
		-	-	-	-	-	-	-	-	-	-	-	-
S	-	↘	↘	↘	↘	↘	↘	→	↘	↘	↘	↘	↘
E	-	↘	↘	↘	↘	↘	↓	↘	→	↘	→	↘	↘
Q	-	↘	↘	↘	↘	↓	↘	↓	↘	→	→	→	↓
A	-	↘	↘	↘	↘	↘	↘	↘	↓	↘	↘	↘	↘
N	-	↘	↘	↘	↘	↘	↘	↘	↓	↘	↘	→	→
C	-	↘	↘	↘	↘	↘	↘	↘	↓	↘	↓	↘	→
E	-	↘	↘	↘	↘	↘	→	↘	↓	↘	↓	↓	↘
L	-	↘	→	↘	↘	↓	↘	↘	↘	↓	↘	↓	↓
I	-	↓	↘	→	→	→	↘	↘	↘	↓	↘	↓	↓
B	-	↘	↓	↘	→	→	→	→	↘	↓	↘	↓	↓
R	-	↘	↓	↓	↘	→	→	→	→	→	↘	↓	↓
E	-	↘	↓	↓	↓	↘	→	↘	→	↘	→	↓	↘

```
Seq1 -------LIBRESEQENCE      Seq1 LIBRESEQENCE-----
Seq2 SEQANCELIBRE-------      Seq2 -----SEQANCELIBRE
          *****                         *** ***
```


g.deleage@ibcp.fr

Principe du séquençage global aléatoire : shotgun



g.deleage@ibcp.fr

Alignement multiple de séquences pourquoi faire?

- ✓ **Choix de « primers » consensus pour la PCR**
- ✓ **Caractériser une nouvelle famille de protéines**
- ✓ **Détecter une homologie entre différentes protéines**
- ✓ **Etablir une phylogénie**
- ✓ **Détecter des résidus identiques ou similaires ayant un rôle fonctionnel ou structural**
- ✓ **Prédictions de structures secondaires**

g.deleage@ibcp.fr

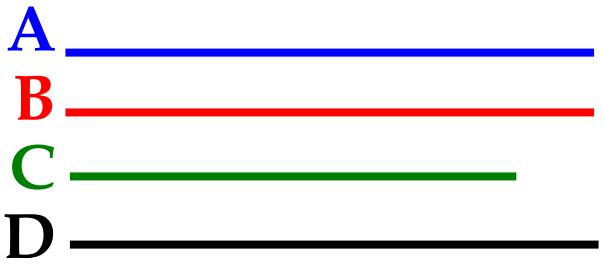
Alignement multiple et programmation dynamique

	Taille moyenne des séquences (1 octet/élément)					
N	100 AA		500 AA		1000 AA	
	Elements	Mémoire	Elements	Mémoire	Elements	Mémoire
2	100^2	10 Ko	500^2	250ko	1000^2	1Mo
3	100^3	1 Mo	500^3	125 Mo	1000^3	1 Go
5	100^5	10 Go	500^5	30 Po	1000^5	1000 Po
10	100^{10}	100, 000 Po	500^{10}	10^{11} Po	1000^{10}	10^{15} Po

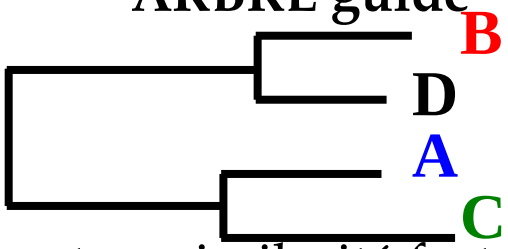
g.deleage@ibcp.fr

Principe de l'alignement progressif de n séquences

- **Alignements de n séquences => n alignements de 2 séquences**
- **Alignements par paires (n*(n-1)/2) comparaisons**
  - Calcul d'un dendogramme ou arbre de similarité => phylogénie
  - 6 comparisons pour 4 séquences A, B, C, D

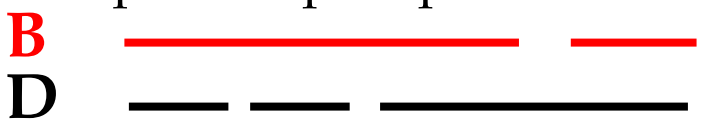


ARBRE guide



Distance courte = similarité forte

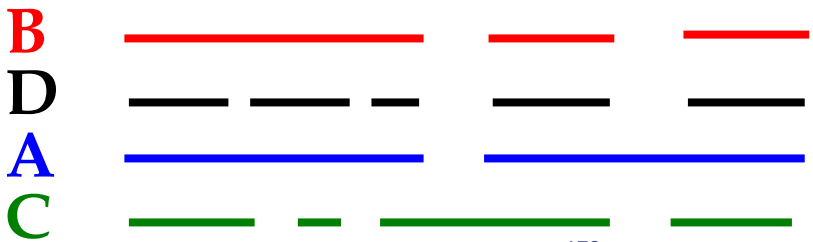
- **Alignements par paires en suivant l'arbre**
  - Alignement de la paire la plus proche



Insertions créées

  - Puis de la suivante

- **Alignements progressifs des paires alignées entre elles**



Nouvelles insertions

g.deleage@ibcp.fr

Coloration f(Pourcentage de conservation des acides aminés)

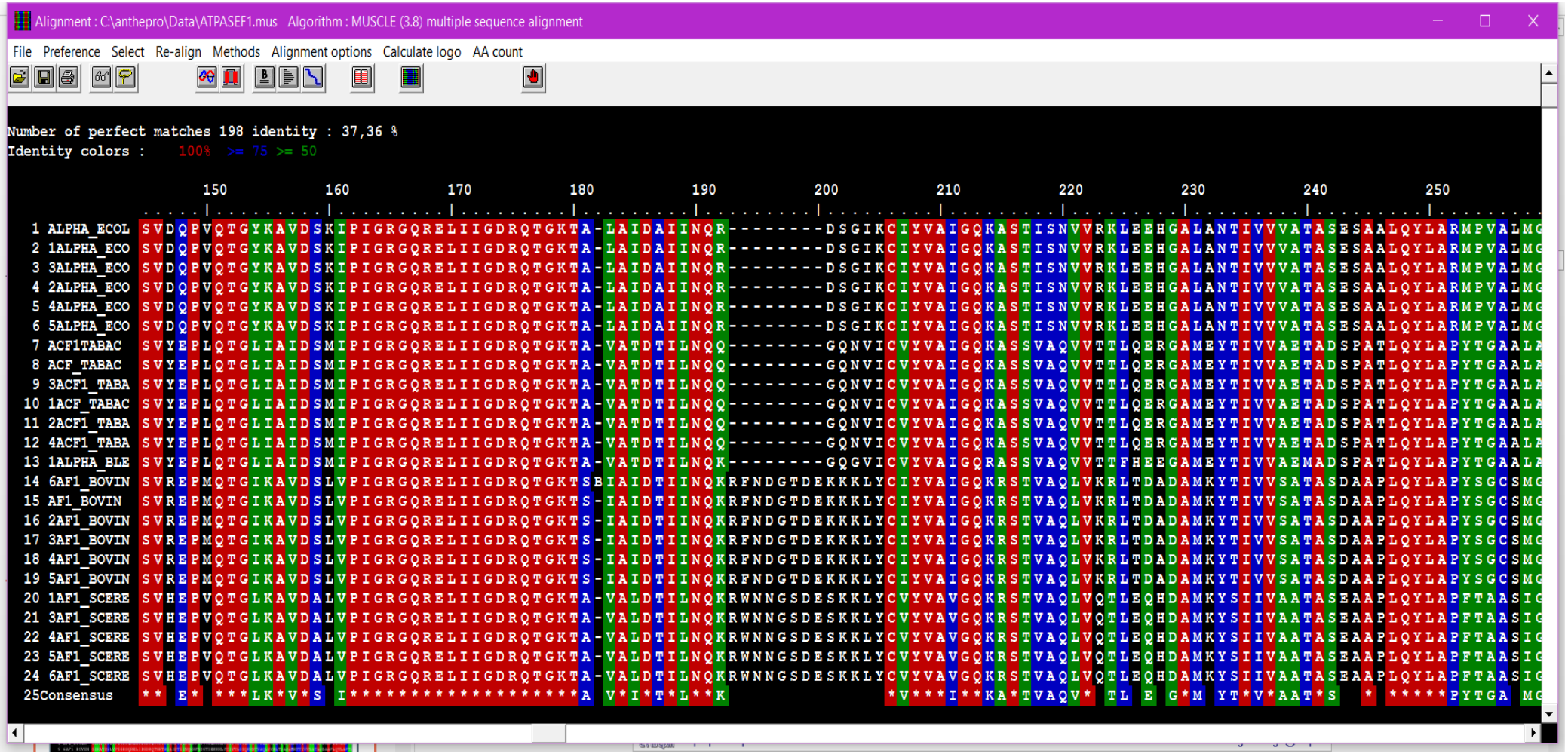
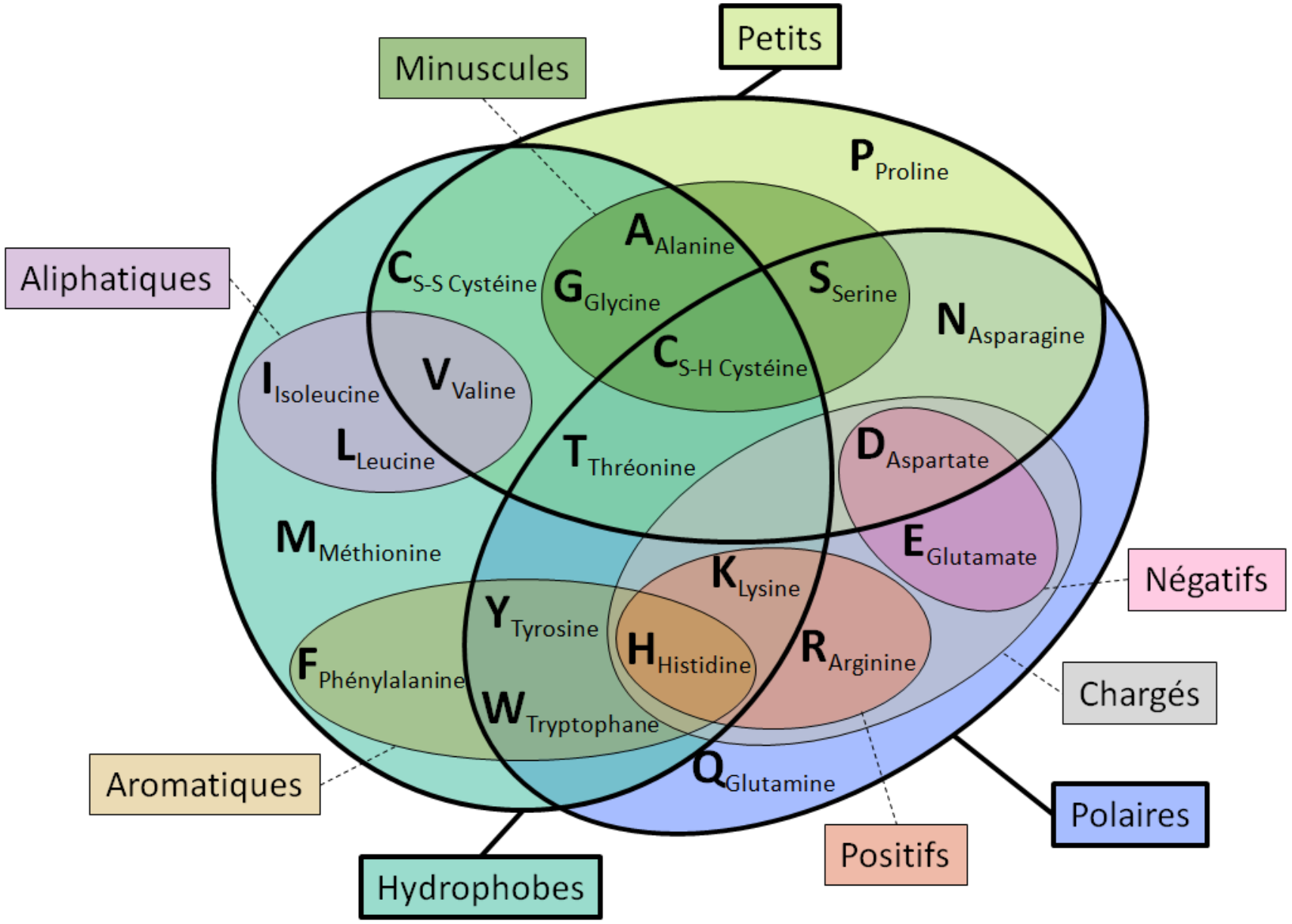
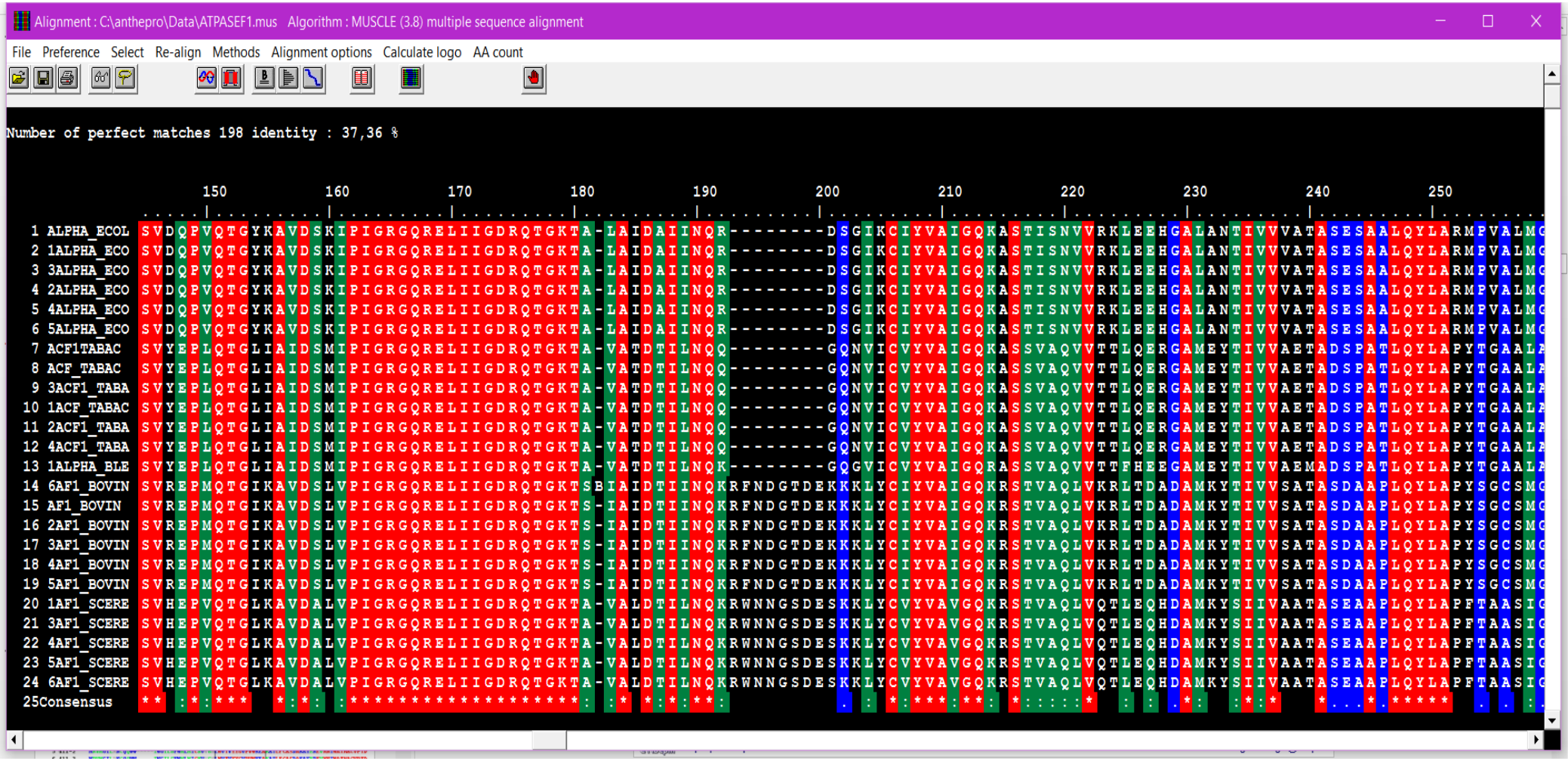


Diagramme de Venn des acides aminés

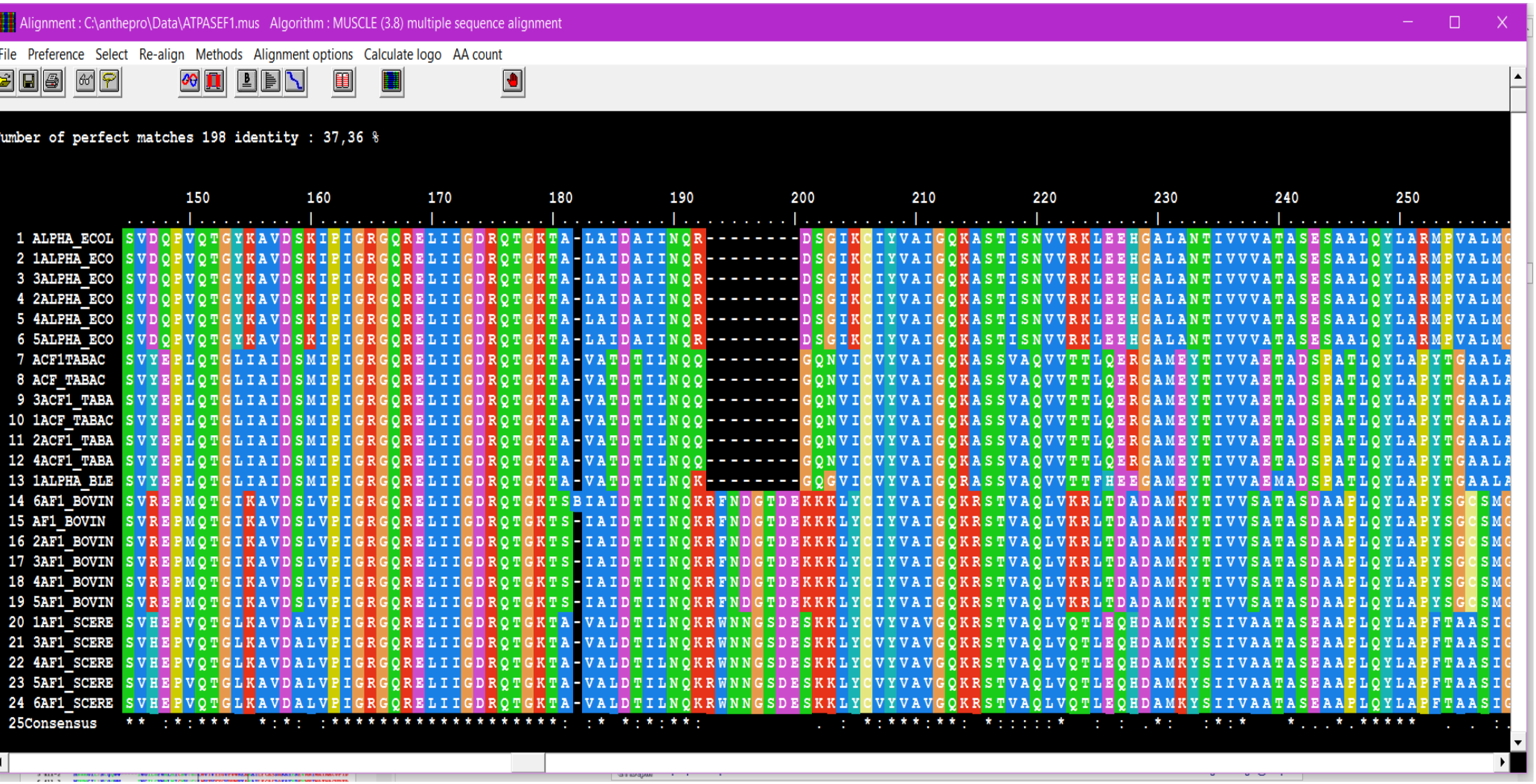


g.deleage@ibcp.fr

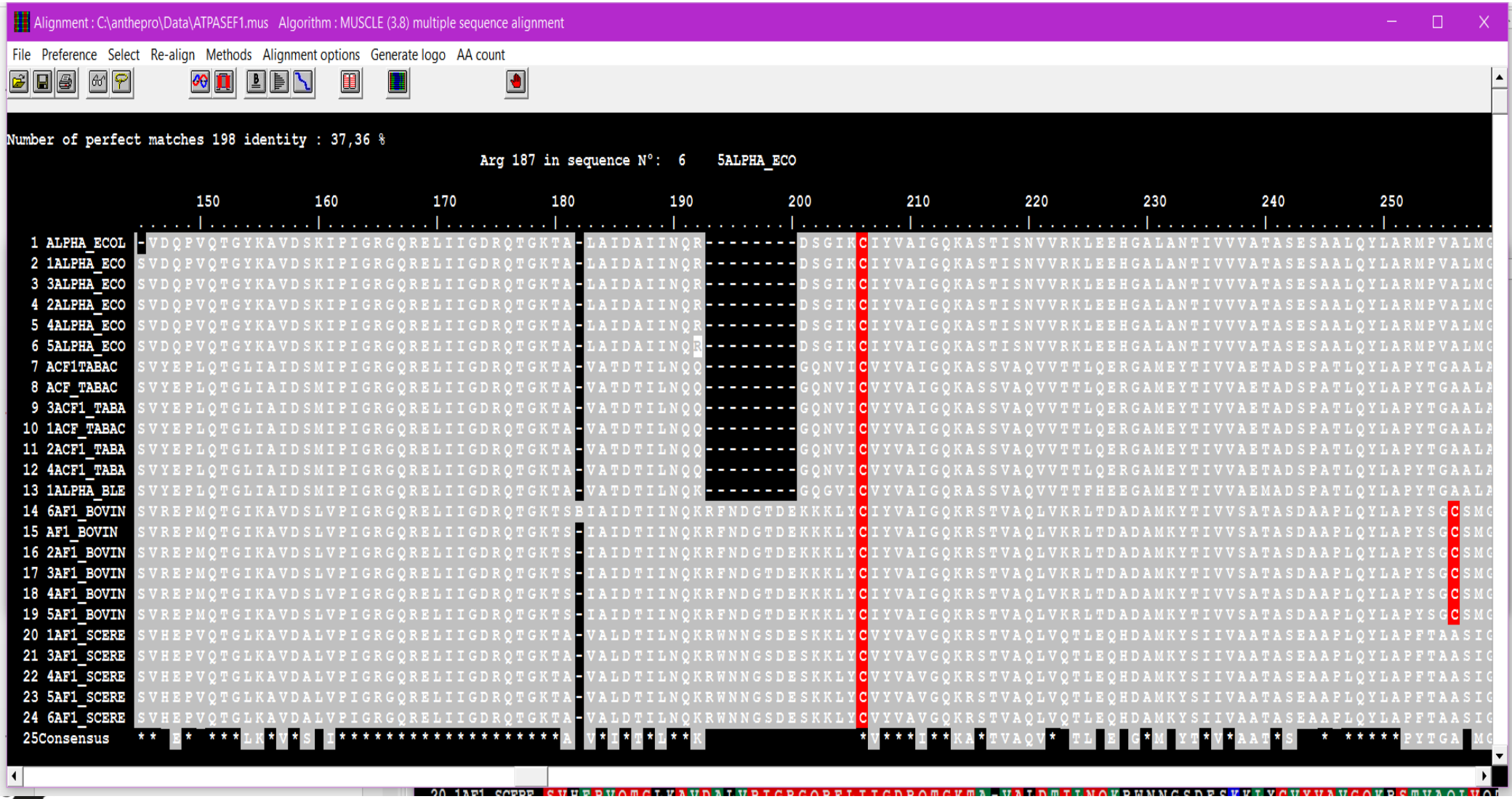
Coloration f(groupe d'acides aminés)



Coloration f(type d'acides aminés)



Coloration f(choix de l'utilisateur)



Programmes d'alignement multiple

- ClustalW (Bien mais lent 20 sec)

  Alignment data :

  Alignment length : 529
  Identity (*) : 198 is 37.43 %
  Strongly similar (:) : 100 is 18.90 %
  Weakly similar (.) : 38 is 7.18 %
  Different : 193 is 36.48 %

  - http://npsa-pbil.ibcp.fr/NPSA/npsa_clustalw.html
- ClustalO (Le plus récent de la série) (Alignement itératif)
- Multalin (2 à 10 fois + rapide, pour des séquences proches 7sec)
  - http://www.toulouse.inra.fr/multalin.html
  - http://npsa-pbil.ibcp.fr/NPSA/npsa_multalin.html

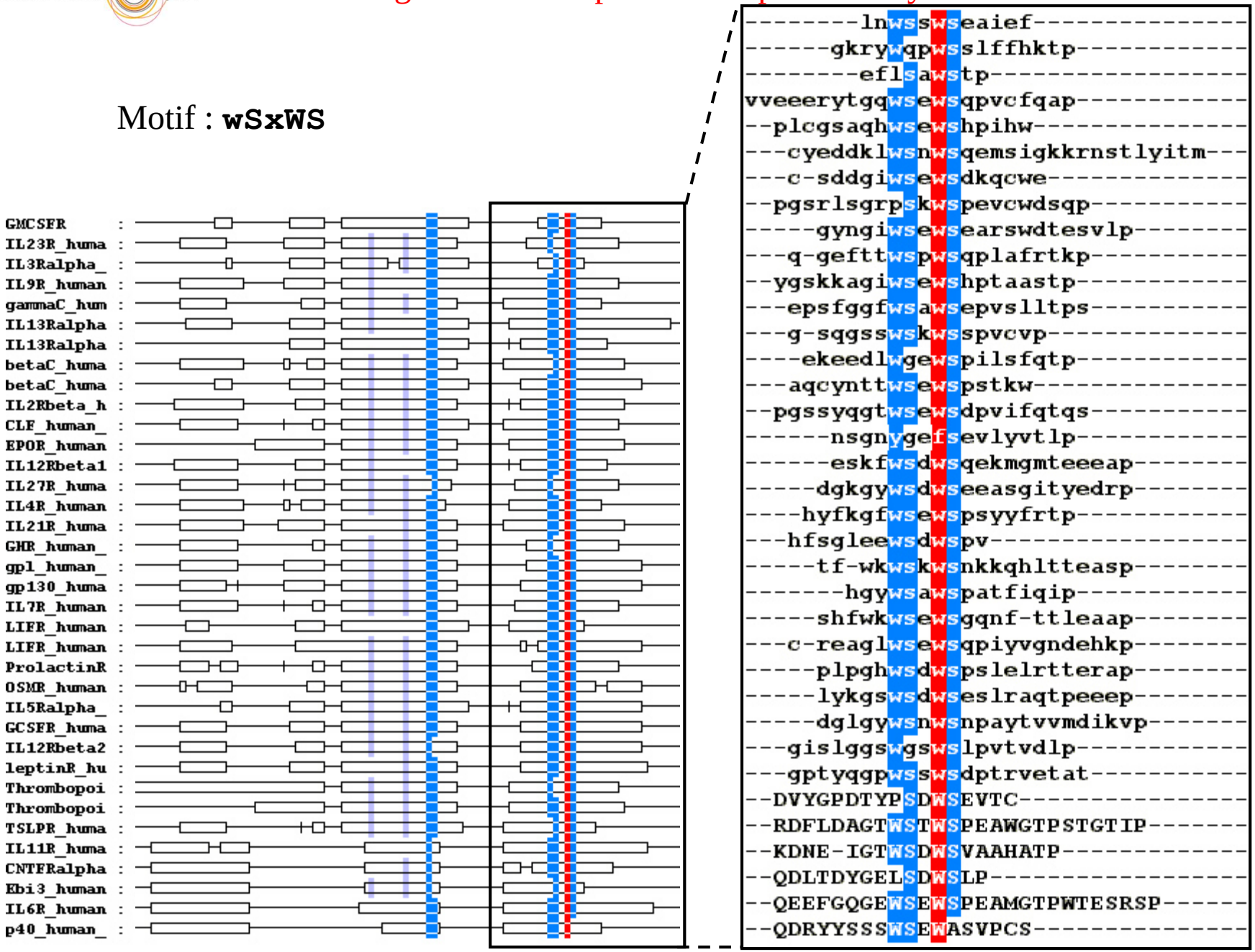
  Alignment data :

  Alignment length : 529
  Residues conserved for 90 % or more (upper-case letters) : 228 is 43.10 %
  Residues conserved for 50 % and less than 90 % (lower-case letters) : 241 is 45.56 %
  Residues conserved less than 50 % (white space) : 16 is 3.02 %
  IV conserved positions (!) : 19 is 3.59 %
  LM conserved positions ($) : 5 is 0.95 %
  FY conserved positions (%) : 4 is 0.76 %
  NDQEBZ conserved positions (#): 16 is 3.02 %
- Muscle 3.8 http://www.drive5.com/muscle/ (Alignement itératif)
  - Permet de réaliser des très gros alignements (500 séquences ou plus)
  - Un des plus rapides
  - Très bon rapport performance/qualité/temps CPU
- T-Coffee
  - Permet l'intégration de données structurales comme contraintes de gap
- MAFFT 7.023
  - Très performant sur des cas difficiles

g.deleage@ibcp.fr

Alignement multiple des récepteurs de cytokines

Motif : **wSxWS**

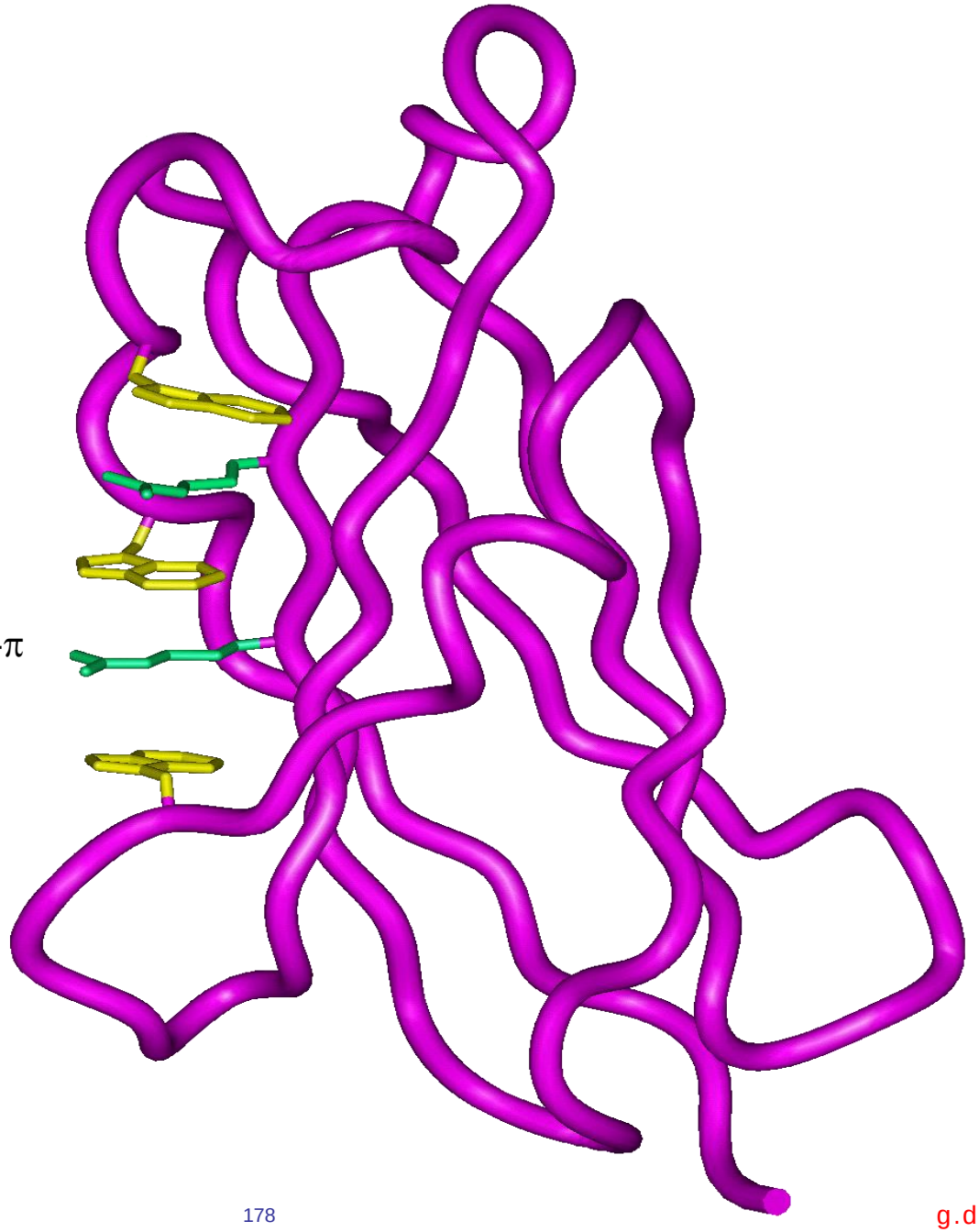


g.deleage@ibcp.fr

Motif structural de cytokines

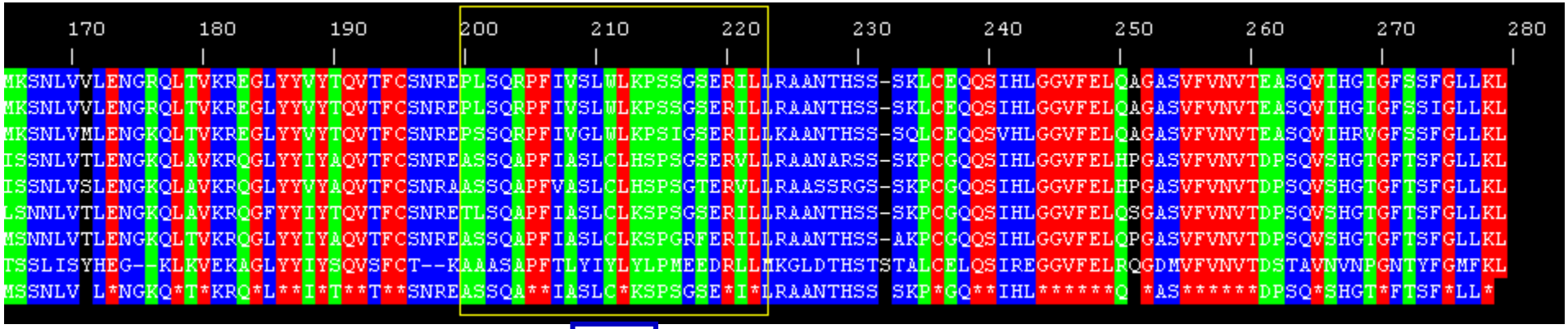
Motif : **wSxWS**

<u>Motif structural</u> : *interactions cation-π entre les résidus **W** et **R***



g.deleage@ibcp.fr

Web logo http://weblogo.berkeley.edu



La hauteur de la colonne indique le degré de conservation à la position considérée

La hauteur d'une lettre est fonction de la fréquence relative d'une lettre

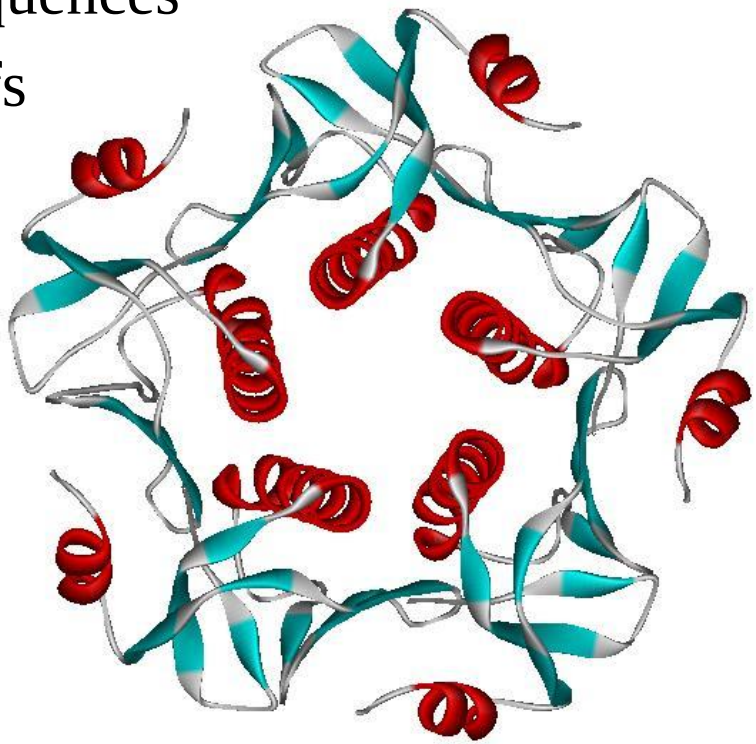
g.deleage@ibcp.fr

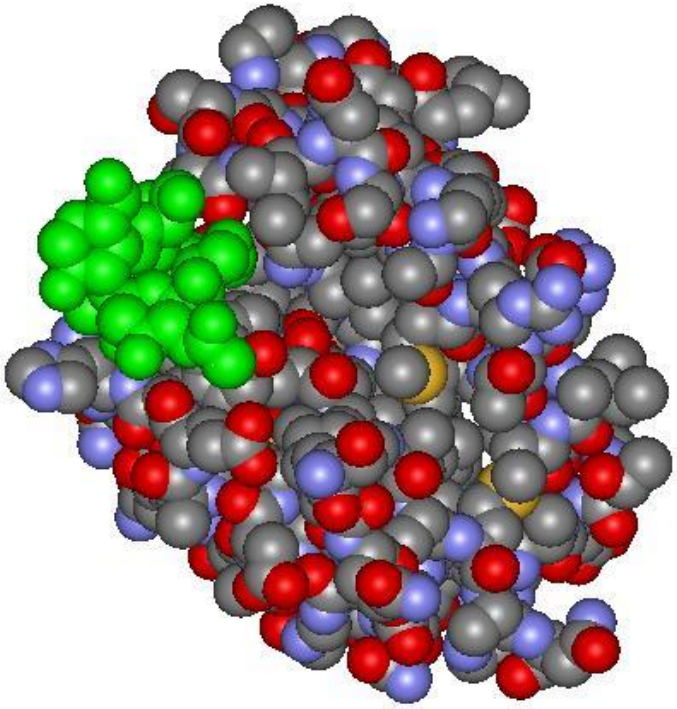
Introduction : Signatures, motifs et sites

ACGYVRLAKCDD

Y-x(2)-[LV]-x-KC

- Alignements de séquences
- Recherche de motifs séquentiels





- Reconnaissance de repliement
- Classification structurale

Analyse de sites fonctionnels

g.deleage@ibcp.fr

Consensus

- ✓ **PROSITE** Signatures et Matrices par programmation dynamique **1309 patterns, 1220 profiles**
- ✓ **PRINTS** **1950 patterns**

Alignements

- ✓ **BLOCKS** profils issus de PROSITE **8909 blocs**
- ✓ **PFAM** Base d'alignments multiple et Hidden Markov Model **16712 entrées (familles)**

Clustering method (BLAST)

- ✓ **PRODOM**

INTERPRO Base intégrée comprenant PROSITE, PRINTS, Pfam and ProDom

http://www.ebi.ac.uk/interpro/

g.deleage@ibcp.fr

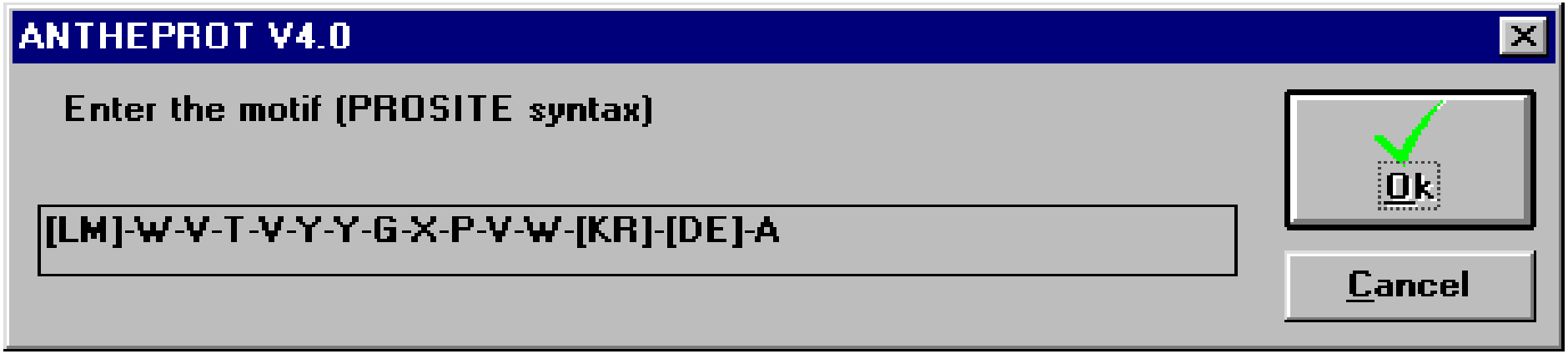
Alignement multiple de séquences


```
Matrix  % of identity for delay 0
Gap opening penalty  0    Hydrophilic gaps    Gap extension penalty  0
Number of perfect matches 175
 100%   >= 75 >= 50 < 50                    -9 in sequence N° 13
                                  10        20        30        40        50        60        70        80
                                  |         |         |         |         |         |         |         |
   1 413-1     MRVKGILRNWQQWW-----IWGILGFWMIMLCNVAGNLWVTVYYGVPVWRDAKATLFCASDAKAYEREVHNIWATHACVPTD
   2 413-2     MRVKGILRNWQQWW-----IWGILGFWMIMLCNVAGNLWVTVYYGVPVWRDAKATLFCASDAKAYEREVHNIWATHACVPTD
   3 413-3     MRVKGILRNWQQWW-----IWGILGFWMIMLCNVAGNLWVTVYYGXPVWRDAKATLFCASDAKAHEREVHNIWATHACVPTD
   4 411-1     MRVMGTLRNCQQWW-----IWGILGFWMLMICNVVGNLWVTVYYGVPVWKEAKATLFCASDAKAYEREVHNIWATHACVPTD
   5 411-2     MRVMGILRNCQQWW-----IWGILGFWMLMICNVVGNLWVTVYYGVPVWKEAKATLFCASDAKAYEREVHNIWATHACVPTD
   6 411-3     MRVMGILRNCQQWW-----IWGILGFWMLMICNVVGNLWVTVYYGVPVWKEAKATLFCASDAKAYEREVHNIWATHACVPTD
   7 426       MRVKGIQMNWPNLW-----KWGTLILGLGIICSASDNLWVTVYYGVPVWRDADTTLFCASDAKAHETEAHNVWATHACVPTD
   8 420       MRVRGMQRNWQPLG-----KWGLLFLGMLIICNAADNLWVTVYYGVPVWKEATTTLFCASDAKAYDREVHNVWATHACVPTD
   9 127       MRVMGIQRNYPLLW-----RWGMIIFWIMIIC-SAEKLWVTVYYGVPVWRDAETTLFCASDAKAYDTEVHNVWATHACVPTD
  10 324       ---MGIQKKYPLLW-----GWGTIIFWIMLICCNAEKMWVTVYYGVPVWKEAKTTLFCASDAKAYDTEVHNVWATHACVPTD
  11 133-1     MRVKEIRRNYQHLW-----RWGTMLLGMFMICSAEEQLWVTVYYGVPVWKEATTTLFCASDAKAYDTEVHNVWATHACVPTD
  12 133-2     MRVKEIRRNYQHLW-----RWGTMLLGTFMICSAEEQLWVTVYYGVPVWKEATTTLFCASDAKAYDTEVHNVWATHACVPTD
  13 153-1     MRVKGIRKNYQHLLW----RWGIMLLGMLRICNAAEKLWVTVYYGVPVWKEATTTLFCASDAKAYDTEVHNVWATHACVPTD
  14 153-2     MRVKGIRKNYQHLLW----RWGIMPLGMLRICNAAEKLWVTVYYGVPVWKEATTTLFCASDAKAYNTEVHNVWATHACVPTD
  15 112       MRVKGIRKSCQHLWW----RWGMMLLGILMICSATNNLWVTVYYGVPVWKEATTTLFCASDAKAYDTEVHNVWATHACVPTD
  16 113       MRVRGIRKSYQHLW-----RWGIMLLGILMICSAANNLWVTVYYGVPVWKEATTTLFCASDAKAYDTEVHNVWATHACVPTD
  17 120       MRVKGIRRNYQHLW-----RWGIMLLGILMICSAAENLWVTVYYGVPVWKEATTTLFCASDAKAYDTEVHNVWATHACVPTD
  18 146-1     MKAKGIRKNYQHLWRGGIWRWGIMLLGMLMICSATERLWVTVYYGVPVWKEASTTLFCASDAKAYDTEVHNVWATHACVPTD
  19 146-2     MKAKGIRKNYQHLWRGGIWRWGIMLLGMLMICSATERLWVTVYYGVPVWKEASTTLFCASDAKAYDTEVHNVWATHACVPTD
  20 146-3     MKAKGIRKNYQHLWRGGIWRWGIMLLGMLMICSATEQLWVTVYYGVPVWKEASTTLFCASDAKAYDTEVHNVWATHACVPTD
  21 149-1     MKAKGIRKNYQHLWRGGIWRWGIMLLGMLLICSATEQLWVTVYYGVPVWKEASTTLFCASDAKAYDTEVHNVWATHACVPTD
  22 149-3     MKAKGIRKNYQHLWRGGIWRWGIMLLGMLLICSATEQLWVTVYYGVPVWKEASTTLFCASDAKAYDTEVHNVWATHACVPTD
  23 149-4     MKAKGIRKNYQHLWRGGIWRWGIMLLGMLLICSATEQLWVTVYYGVPVWKEASTTLFCASDAKAYDTEVHNVWATHACVPTD
  24 149-2     MKAKGIRKNYQHLWRGGIWRWGIMLLGMLLICSATEQLWVTVYYGVPVWKEASTTLFCASDAKAYDTEVHNVWATHACVPTD
  25 149-5     MKAKGIRKNYQHLWRGGIWRWGIMLLGMLLICSATEQLWVTVYYGVPVWKEASTTLFCASDAKAYDTEVHNVWATHACVPTD
  26 374       MRAKGMRKNYQHLWR-----WGMMLLGILMICNAKEQLWVTVYYGVPVWKEATTTLFCASDAKAYDTEVHNVWATHACVPTD
```


g.deleage@ibcp.fr

Génération de séquences consensus

- Alignements de protéines homologues
- Génération de séquences consensus
- Identification de résidus importants (conservés)
- Création d'une signature fonctionnelle
- Constitution d'un dictionnaire de signatures



g.deleage@ibcp.fr

Recherche de motifs

- **Utilisation du dictionnaire PROSITE**
  - Identifications des signatures possibles dans une séquence recherche de fonctions
  - Recherche d'une signature dans une banque de séquence recherche d'homologues distants
- **Méthodes de profils**
  - Principe
  - Recherche des "Blocks" (Patmat)
  - HMMER, PFTools
- **Autres...**

g.deleage@ibcp.fr

PROSITE: Dictionnaire de signatures

- **Auteur :** Amos Bairoch
  Département de Biochimie Médicale
  Université de Genève
  1, Rue Michel Servet, 1 211 Geneva 4, Suisse
  bairoch@cmu.unige.ch

- **Version :** 2016, 1400 sites et signatures
- **Exemple :**

```
ID      HTH_LACI_FAMILY; PATTERN.
DE      Bacterial regulatory proteins, LacI family signature.
PA      [LIVM]-x-[DE]-[LIVM]-A-x(2)-[STAG]-[LIVMA]-x(2)-[FLIVMAN]-
PA      [LIVMC].
3D      1LCC; 1LCD; 1LTP
```

- **Utilisation du dictionnaire PROSITE**
  - **Recherche de fonctions => Identifications des signatures possibles dans une séquence**
  - **Recherche d'homologues distants => Recherche d'une signature dans une banque de séquence**

Index du dictionnaire PROSITE

- **Post-translational modifications**
- **Domains**
- **DNA or RNA associated proteins**
- **Enzymes**
  - Oxidoreductases signatures
  - Transferases
  - Hydrolases
  - Lyases
  - Isomerases
  - Ligases
  - Others
- **Electron transport proteins**
- **Other transport proteins**
- **Structural proteins**
- **Receptors**
- **Cytokines and growth factors**
- **Hormones and active peptides**
- **Toxins signatures**
- **Inhibitors signature**
- **Protein secretion and chaperones**
- **Others**

g.deleage@ibcp.fr

Syntaxe des motifs = syntaxe de PROSITE

- **Les positions sont séparées par des "-".**
- **A une même position:**
  - Les accolades "{ }" excluent un ou plusieurs acides aminés.
  - les crochets "[ ]" autorisent plusieurs acides aminés.
  - "X" autorise tous les acides aminés.
- **Une position peut être répétée un nombre fixe de fois "( )" ou un nombre variable de fois "(,)".**
- **"<" et ">" signifient respectivement motif en position N-terminale ou C-terminale.**
- **Exemple:**

<M-**[AGVS](2)-Y(2,8)-X-A-L-{AGVS}**

g.deleage@ibcp.fr

Algorithme de recherche

- **Analyse syntaxique du motif de longueur L**
- **Génération de matrices de position (une par position) avec**
  - 1 si AA autorisé
  - 0 si AA non autorisé
- **Découpage de la séquence en segments de longueur L**
  - 1 - L , 2 - L+ 1, 3 - L + 2 , . . . , (n - L) - n
  - Application des matrices correspondant aux positions du motif sur chaque segment et somme S des scores obtenus sur un segment
  - Si le score S est égal à L alors le motif est trouvé (toutes les positions sont bonnes)
  - Si le score S est égal à L - k alors le motif est trouvé avec k erreurs
  - Si le score S est supérieur à L x τ et τ est % de similarité
- **Tri et affichage des résultats**

g.deleage@ibcp.fr

Etablissement des matrices de substitutions locales

Longueur=6 positions

[LIVA]-[LIVAMY]-[VAT]-H-N-[STC]

	L	I	V	A
L	1	1	1	1
I	1	1	1	1
V	1	1	1	1
A	1	1	1	1

	L	I	V	A	M	Y
L	1	1	1	1	1	1
I	1	1	1	1	1	1
V	1	1	1	1	1	1
A	1	1	1	1	1	1
M	1	1	1	1	1	1
Y	1	1	1	1	1	1

	V	A	T
V	1	1	1
A	1	1	1
T	1	1	1

	H		N
H	1	N	1

	S	T	C
S	1	1	1
T	1	1	1
C	1	1	1

g.deleage@ibcp.fr

Algorithme de recherche

[LIVA]-[LIVAMY]-[VAT]-H-N-[STC]

	L	I	V	A
L	1	1	1	1
I	1	1	1	1
V	1	1	1	1
A	1	1	1	1

	L	I	V	A	M	Y
L	1	1	1	1	1	1
I	1	1	1	1	1	1
V	1	1	1	1	1	1
A	1	1	1	1	1	1
M	1	1	1	1	1	1
Y	1	1	1	1	1	1

	V	A	T
V	1	1	1
A	1	1	1
T	1	1	1

	H		N
H	1	N	1

	S	T	C
S	1	1	1
T	1	1	1
C	1	1	1

```
AYITGFRPLVTHNSLCVHNS...
110000
 011000
  100000
   000000
    000000
     000001
      000000
       011000
        111111
         100000
          000001
           000000
            000000
             010000
              101111
AYITGFRPLVTHNSLCVHNS...
```

Trouvé exactement

Trouvé avec une erreur

g.deleage@ibcp.fr

Exemple de calcul de sites PROSITE sur une séquence

ANTHEPROT (http://antheprot-pbil.ibcp.fr)

PROSCAN http://npsa-pbil.ibcp.fr/NPSA/npsa_proscan.html

Interpro : http://npsa-pbil.ibcp.fr/iprscan/iprscan

g.deleage@ibcp.fr

Exemples


```
Protein PROTEINY using C:\PROSITE.DAT as reference site file
Similarity percentage 100 NUmber of mismatch allowed 0
-----------------------------------------------------------------------
N-glycosylation site PS00001
N-{P}-[ST]-{P}
-----------------------------------------------------------------------
Site :       8   to     11   NRSN    Observed frequency: 6.08E-006
Site :     185   to    188   NTTI    Observed frequency: 7.71E-006
-----------------------------------------------------------------------
Casein kinase II phosphorylation site PS00006
[ST]-x(2)-[DE]
-----------------------------------------------------------------------
Site :      84   to     87   SKAE    Observed frequency: 4.19E-003
Site :     116   to    119   SGGD    Observed frequency: 3.75E-003
-----------------------------------------------------------------------
Serine proteases, trypsin family, histidine active site PS00134
[LIVM]-[ST]-A-[STAG]-H-C
-----------------------------------------------------------------------
Site :     143   to    148   LTAGHC    Observed frequency: 1.14E-008
-----------------------------------------------------------------------
Serine proteases, trypsin family, serine active site PS00135
[DNSTAGC]-[GSTAPIMVQH]-x(2)-G-[DE]-S-G-[GS]-[SAPHV]-[LIVMFYWH]
-----------------------------------------------------------------------
Site :     249   to    260   CAEPGDSGGPL    Observed frequency: 3.11E-013

Number of motifs 1358 Elasped searching time: 9.61 Sec.
```


g.deleage@ibcp.fr

Exemples

Serine proteases, trypsin family, active sites
****

The catalytic activity of the serine proteases from the trypsin family is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which itself is hydrogen-bonded to a serine. The sequences in the vicinity of the active site serine and histidine residues are well conserved in this family of proteases [1]. A partial list of proteases known to belong to the trypsin family is shown below.

- Acrosin.
- Blood coagulation factors VII, IX, X, XI and XII, thrombin, plasminogen, and protein C.
- Apolipoprotein(a).

All the above proteins belong to family S1 in the classification of peptidases [2,E1] and originate from eukaryotic species. It should be noted that bacterial proteases that belong to family S2A are similar enough in the regions of the active site residues that they can be picked up by the same patterns. These proteases are listed below.

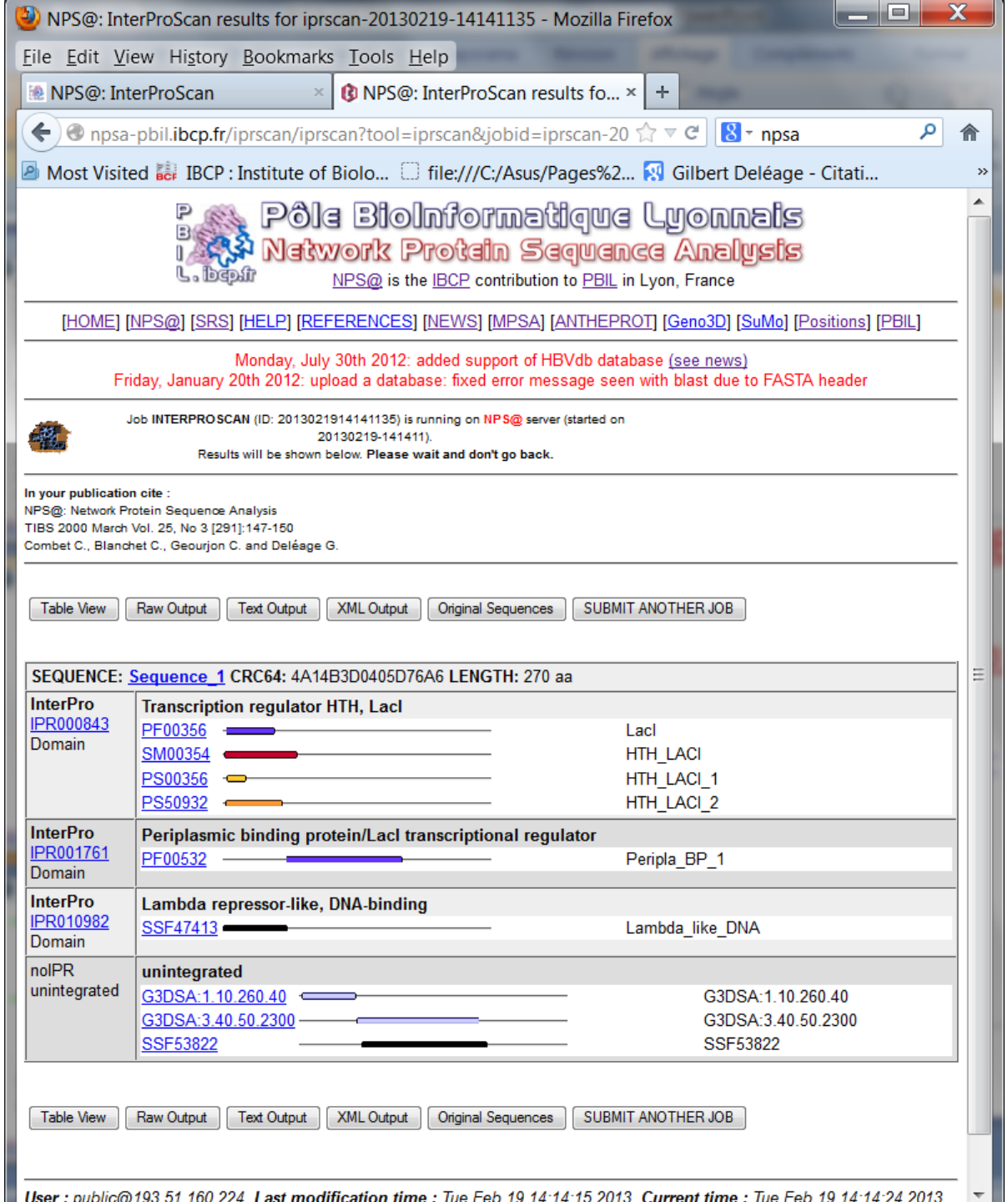
- Achromobacter lyticus protease I.
- Lysobacter alpha-lytic protease.
- Streptomyces fradiae proteases 1 and 2.

-Consensus pattern: [LIVM]-[ST]-A-[STAG]-H-C [H is the active site residue]

-Last update: November 1997 / Text revised.

[1] Brenner S.
Nature 334:528-530(1988).
[2] Rawlings N.D., Barrett A.J.
Meth. Enzymol. 244:19-61(1994).

g.deleage@ibcp.fr



NPS@: InterProScan results for iprscan-20130219-14141135 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

NPS@: InterProScan | NPS@: InterProScan results fo...

npsa-pbil.ibcp.fr/iprscan/iprscan?tool=iprscan&jobid=iprscan-20

Most Visited | IBCP : Institute of Biolo... | file:///C:/Asus/Pages%2... | Gilbert Deléage - Citati...

Pôle BioInformatique Lyonnais
Network Protein Sequence Analysis

NPS@ is the IBCP contribution to PBIL in Lyon, France

[HOME] [NPS@] [SRS] [HELP] [REFERENCES] [NEWS] [MPSA] [ANTHEPROT] [Geno3D] [SuMo] [Positions] [PBIL]

Monday, July 30th 2012: added support of HBVdb database (see news)
Friday, January 20th 2012: upload a database: fixed error message seen with blast due to FASTA header

Job **INTERPROSCAN** (ID: 2013021914141135) is running on NPS@ server (started on 20130219-141411).
Results will be shown below. **Please wait and don't go back.**

In your publication cite :
NPS@: Network Protein Sequence Analysis
TIBS 2000 March Vol. 25, No 3 [291]:147-150
Combet C., Blanchet C., Geourjon C. and Deléage G.

Table View | Raw Output | Text Output | XML Output | Original Sequences | SUBMIT ANOTHER JOB

SEQUENCE: Sequence_1 **CRC64:** 4A14B3D0405D76A6 **LENGTH:** 270 aa

InterPro IPR000843 Domain	**Transcription regulator HTH, LacI**	
	PF00356	LacI
	SM00354	HTH_LACI
	PS00356	HTH_LACI_1
	PS50932	HTH_LACI_2
InterPro IPR001761 Domain	**Periplasmic binding protein/LacI transcriptional regulator**	
	PF00532	Peripla_BP_1
InterPro IPR010982 Domain	**Lambda repressor-like, DNA-binding**	
	SSF47413	Lambda_like_DNA
noIPR unintegrated	**unintegrated**	
	G3DSA:1.10.260.40	G3DSA:1.10.260.40
	G3DSA:3.40.50.2300	G3DSA:3.40.50.2300
	SSF53822	SSF53822

Table View | Raw Output | Text Output | XML Output | Original Sequences | SUBMIT ANOTHER JOB

User : public@193.51.160.224 Last modification time : Tue Feb 19 14:14:15 2013 Current time : Tue Feb 19 14:14:24 2013

g.deleage@ibcp.fr

Recherche avec erreur

- **Motif : [LIVA]-[LIVMY]-[VAT]-H-N-[STC]**
- **Considérons une séquence de protéine :**

  AYITGFRPLVTINS**LCVHNS…**

  $Sc_0 = 100\%$, $Sc_{seuil} = \tau = 80\%$
- **La pénalité pour une position non trouvée est uniforme, et vaut 100/6 = 16,7%**
- **Pour ce motif potentiel, seule la position 4 ne correspond pas. Le score du motif est donc égal à :**

  100 - 16,7 = 83,3 %
- **Le motif LVTINS est donc trouvé bien qu'une position stricte [H] ne soit pas conservée.**

g.deleage@ibcp.fr

Recherche avec pénalité variable

- Cet algorithme (pondération variable) de recherche de motif(s) dans les protéines doit privilégier la conservation des positions strictes dans les motifs trouvés qui sont des positions biologiquement importantes.
- L'algorithme permet d'augmenter de façon notable le rapport signal/bruit dans les résultats de la recherche.
- Ce gain est réalisé lors d'une recherche de motif(s) suivant un taux de similarité minimal.
- L'algorithme utilise la valeur de la fréquence du motif et les valeurs des fréquences des acides aminés pour pondérer les erreurs.

g.deleage@ibcp.fr

Recherche avec pénalité variable - Résumé

- La fréquence du motif complet est calculée d'après les fréquences des acides aminés dans SwissProt.
- La valeur de cette fréquence est assignée comme score initial (Sc_0) du motif.
- Lors de l'introspection de la séquence, pour chaque position ne correspondant pas, le score du motif est divisé par une pondération.
- Cette pondération est égale à la valeur de la fréquence de la position ne correspondant pas.
- Un score seuil (Sc_{seuil}) est défini avec: $Sc_{seuil} = 10^{\tau.\log Sc_0}$ où τ est le taux de similarité désiré pour la recherche.
- Le score à la position doit être inférieur au Sc_{seuil}

g.deleage@ibcp.fr

Recherche avec pondération variable - Calcul de Sc0

- **Motif : [LIVA]-[LIVMY]-[VAT]-H-N-[STC]**
- **Protéine : AYITGFRPLVTINSLCVHNS…**
- **Taux de similarité : 80%**

$$P(j) = \sum_{i=1}^{N} F(i)$$

Où: **i est le type d'acide aminé autorisé à la position j**
F(i) est la fréquence de l'acide aminé i en moyenne dans SWISS-PROT
N est le nombre d'acides aminés différents autorisés à la position j.

- **Valeur de pondération pour chaque position:**
  - **P_1 = 0,289**
  - **P_2 = 0,268**
  - **P_3 = 0,200**
  - **P_4 = 0,022**
  - **P_5 = 0,044**
  - **P_6 = 0,147**

$$Y = \prod_{j=1}^{L} P(j)$$

où : j est la position dans le motif
L la longueur du motif (L=6)

$$\left.\right\} \quad Y = P_1 \times P_2 \times \ldots \times P_6$$

- **Valeur initiale du score du motif : Sco = Y= $2{,}27.10^{-6}$**

g.deleage@ibcp.fr

Recherche avec pondération variable – Calcul de Sc(j)

- **Lors de la recherche évaluation du score Sc(j)**

Si la position J est satisfaite alors:

$$Sc(j) = Sc(j\text{-}1)$$

Sinon:

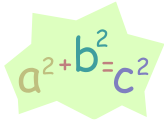
$$Sc(j) = \frac{Sc(j-1)}{P(j)}$$

g.deleage@ibcp.fr

Recherche avec pénalité variable - 1 erreur sur position fixe

- **Motif : [LIVA]-[LIVMY]-[VAT]-H-N-[STC]**
- **Motif potentiel dans la protéine :**

  AYITGFRP<u>LVTINS</u>LCVHNS…



  $Sc_0 = 2{,}27.10^{-6}$, $\tau = 0{,}8$

  $Sc_{seuil} = 10^{\tau.\log Sc_0} = 3{,}05.10^{-5}$

- **Position 1 : L concorde ⇒ $Sc_1 = Sc_0$**
- **Position 2 : V concorde ⇒ $Sc_2 = Sc_1$**
- **Position 3 : T concorde ⇒ $Sc_3 = Sc_2$**
- **Position 4 : I ne concorde pas ⇒ $Sc_4 = Sc_3 / P_4$**

  $Sc_4 = 2.27\ 10^{-6} / 0.022$

  $Sc_4 = 1{,}03.10^{-4} > Sc_{seuil}$ ⇒ motif non correct

g.deleage@ibcp.fr

Recherche avec pénalité variable - 1 erreur position dégénérée

- **Motif : [LIVA]-[LIVMY]-[VAT]-H-N-[STC]**
- **Motif potentiel dans la protéine :**

  AYITGFRPLVTINS<u>LCVHNS</u>…

  $Sc_0 = 2{,}27.10^{-6}$, $\tau = 0{,}8$

  $Sc_{seuil} = 10^{\tau.\log Sc_0} = 3{,}05.10^{-5}$
- **Position 1 : L concorde ⇒ $Sc_1 = Sc_0$**
- **Position 2 : C ne concorde pas ⇒ $Sc_2 = Sc_1 / P_2$**

  $Sc_2 = 2{,}27 .10^{-6} / 0{,}268$

  $Sc_2 = 8{,}47.10^{-6} < Sc_{seuil}$ ⇒ **on continue**
- **Position 3 : V concorde ⇒ $Sc_3 = Sc_2$**
- **Position 4 : H concorde ⇒ $Sc_4 = Sc_3$**
- **Position 5 : N concorde ⇒ $Sc_5 = Sc_4$**
- **Position 6 : S concorde ⇒ <u>motif correct</u>**

g.deleage@ibcp.fr

Recherche avec pénalité variable - 2 erreurs 1 dégénérée et une fixe

- **Motif : [LIVA]-[LIVMY]-[VAT]-H-N-[STC]**
- **Motif potentiel dans la protéine :**

  AYITGFRPLVTINSLCVHNS…

  $Sc_0 = 2{,}27.10^{-6}$, $\tau = 0{,}8$

  $Sc_{seuil} = 10^{\tau.\log Sc_0} = 3{,}05.10^{-5}$
- **Position 1 : A concorde $\Rightarrow Sc_1 = Sc_0$**
- **Position 2 : Y concorde $\Rightarrow Sc_2 = Sc_1$**
- **Position 3 : I ne concorde pas $\Rightarrow Sc_3 = Sc_2 / P_3$**

  $Sc_3 = 2{,}27.10^{-6} / 0.2$

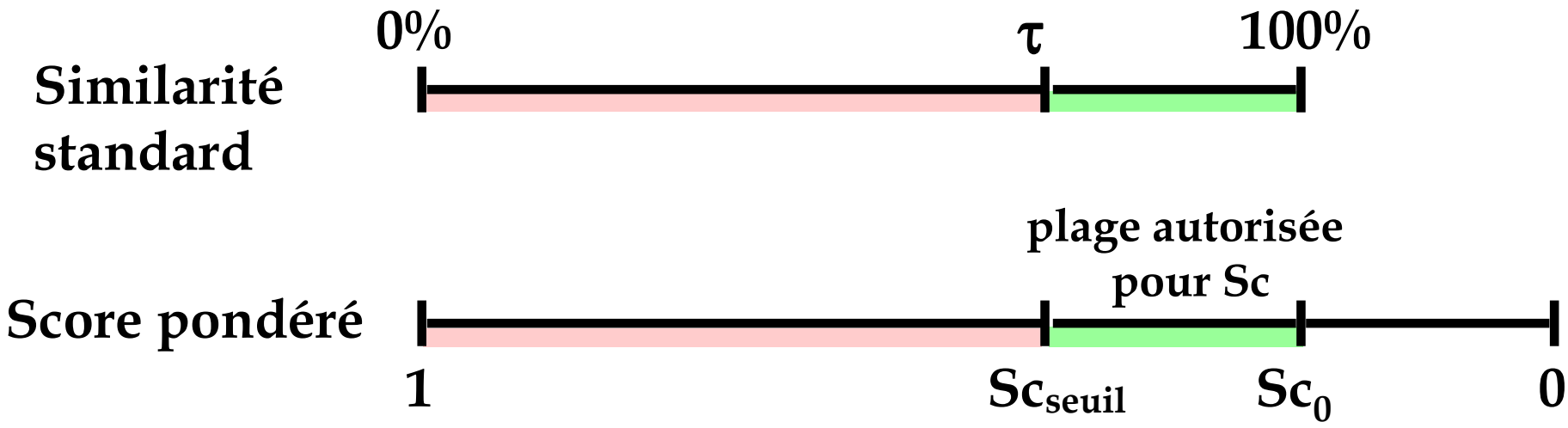
  $Sc_3 = 1{,}13.10^{-5} < Sc_{seuil} \Rightarrow$ on continue
- **Position 4 : T ne concorde pas $\Rightarrow Sc_4 = Sc_3 / P_4$**

  $Sc_4 = 1{,}13.10^{-5} / 0.022$

  $Sc_4 = 5{,}16.10^{-4} > Sc_{seuil} \Rightarrow$ motif non correct

g.deleage@ibcp.fr

Calcul du score seuil Sc_{seuil}



Recherche avec pénalité variable - Exemple

Site d'épissage des protéines (PS00881 de Prosite) recherché avec 80% de similarité dans SwissProt

[LIVA]-[LIVMY]-[VAT]-H-N-[STC]

- **La recherche traditionnelle (sans pondération) trouve plus de 100 000 occurences. Elle autorise jusqu'à 2 erreurs, mais les positions les plus substituées sont les deux positions strictes : H et N.**

- **La méthode pondérée autorise aussi jusqu'à 2 erreurs, favorise la conservation des positions les plus strictes et trouve 7401 occurrences avec les 2 positions strictes toujours conservées**

Lancer la recherche

g.deleage@ibcp.fr

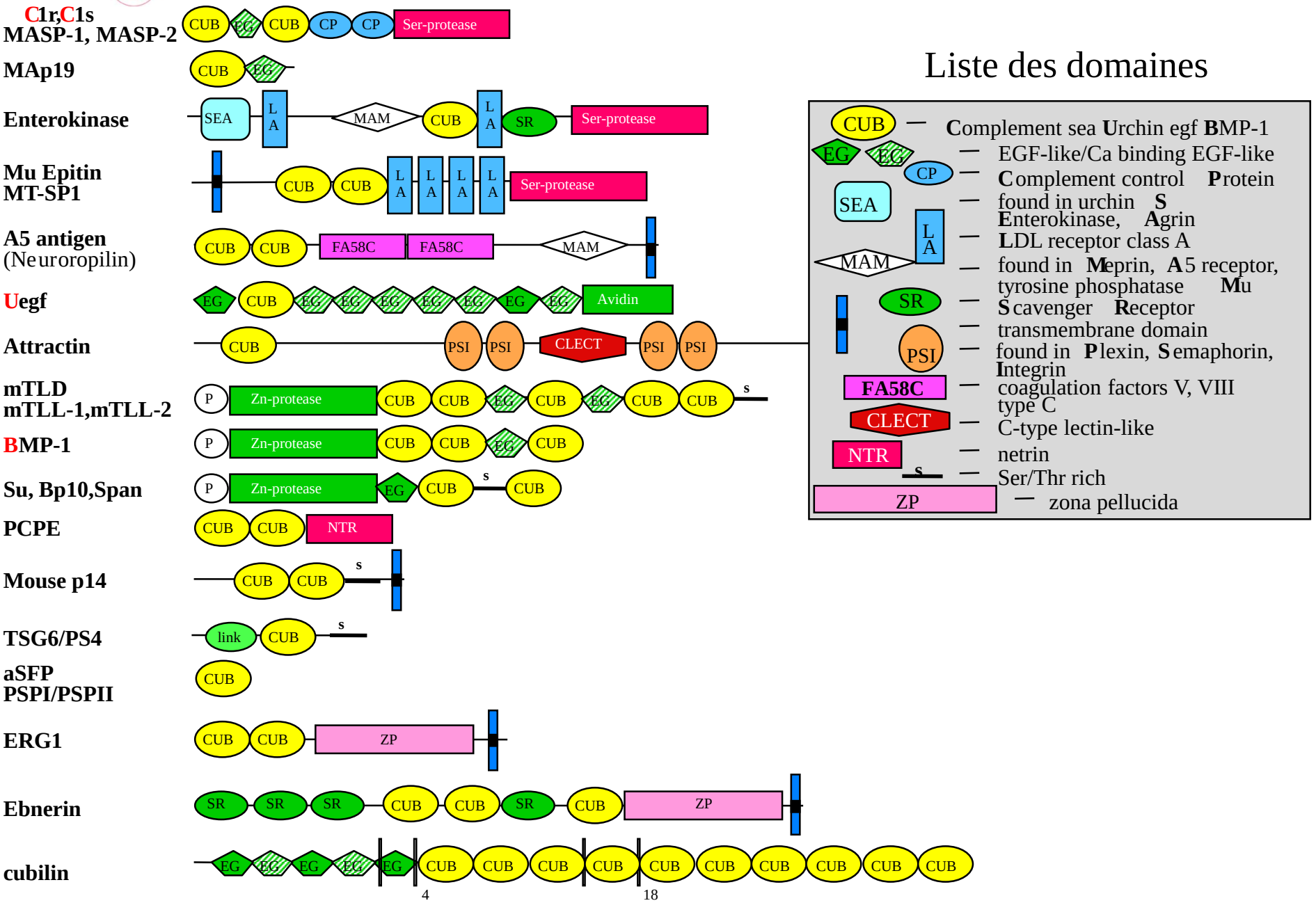
Stratégie d'utilisation PattInProt sur le Web

[EQR]-C-[LIVMFYAH]-x-C-x(5,8)-C-x(3,8)-[EDNQSTV]-C-{C}-x(5)-C-x(12,24)-C	24 min	103184
[EQR]-C-[LIVMFYAH]-x-C	0.40 min	1877
[EQR]-C-[LIVMFYAH]-x-C-x(5,8)-C-x(3,8)-[EDNQSTV]-C-{C}-x(5)-C-x(12,24)-C	1.10 min	56

En cas de recherche de motif très dégénéré avec beaucoup de positions flottantes
Découper la signature en morceaux
Rechercher en premier le motif plus rare (sans ,) recherches successives

g.deleage@ibcp.fr

Organisation modulaire des protéines



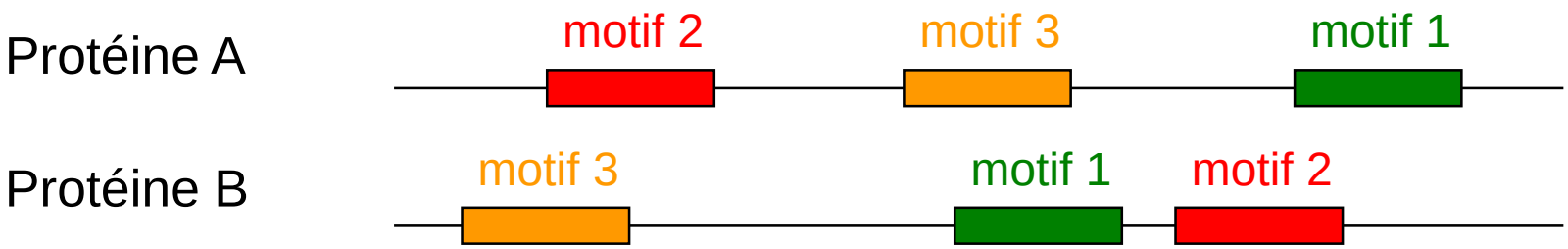
Liste des domaines

g.deleage@ibcp.fr

Recherche de motifs dans les banques de séquences

- **Recherche d'un seul motif:**
  - Identité stricte
  - Positions fausses autorisées (mismatch)
  - Taux de similarité
  - Pénalité variable
- **Recherche de plusieurs motifs**

  Les mêmes critères sont appliqués à la recherche de plusieurs motifs sensés appartenir à la même protéine. L'ordre de soumission des motifs n'est pas corrélé à leur ordre d'apparition dans la protéine.



https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_pattinprot.html

g.deleage@ibcp.fr

Définition

Un **profil** ou **matrice de poids** est une table de fréquences (ou %) par position d'acides aminés dans un alignement. La table est utilisée pour calculer un score de similarité entre tous les alignements d'un profil avec une séquence. Un alignement avec un score de similarité supérieur ou égal à un seuil représente une occurrence.

Constitution d'alignements et d'un jeu de référence

BLOCKS d'Henikoff (dérive de PROSITE)	8909 blocs
PFAM Base d'alignments multiple et HMM	3071 familles
PRINTS	9800 motifs uniques=> 1600 empreintes
PRODOM	283 772 familles de domaine de séquences

g.deleage@ibcp.fr

Recherches de motifs - Blocks de Henikoff


```
ID   ATPASE_ALPHA_BETA; BLOCK
AC   BL00152B; distance from previous block=(26,29)
DE   ATP synthase alpha and beta subunits proteins.
BL   DGT motif; width=42; seqs=78; 99.5%=873; strength=2546
FLII_BACSU ( 142) EKMGVGVRSIDSLLTVGKGQRIGIFAGSGVGKSTLMGMIAKQ

FLII_SALTY ( 157) HVLDTGVRAINALLTVGRGQRMGLFAGSGVGKSVLLGMMARY

ATP0_BETVU ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN
ATP0_BOVIN ( 187) EPMQTGIKAVDSLVPIGRGQRELIIGDRQTGKTSIAIDTIIN
ATP0_BRANA ( 146) EPMQTGLKAVDSLVPIGRGQRELLIGDRQTGKTTIAIDTILN
ATP0_HELAN ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN
ATP0_MAIZE ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN
ATP0_MARPO ( 145) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN
ATP0_NICPL ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN
ATP0_OENBI ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN
ATP0_ORYSA ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN
  ATP0_PEA ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN
ATP0_PHAVU ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN
ATP0_RAPSA ( 146) EPMQTGLKAVDSLVPIGRGQRELLIGDRQTGKTTIAIDTILN
ATP0_SOYBN ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN
ATP0_WHEAT ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN
ATPA_ANTSP ( 147) EPLQTGITAIDSMIPIGRGQRELIIGDRQTGKTTVALDTIIN
ATPA_CHLRE ( 145) EPLATGLVAVDAMIPVGRGQRELIIGDRQTGKTAIAVDTILN
ATPA_ECOLI ( 144) QPVQTGYKAVDSMIPIGRGQRELIIGDRQTGKTALAIDAIIN
ATPA_EUGGR ( 145) EPLQTGLIAIDAMIPIGRGQRELIIGDRQTGKTAVATDTILN
ATPA_GALSU ( 145) EPLQTGITAIDSMIPIGRGQRELIIGDRQTGKTSIALDTIIN
ATPA_HUMAN ( 187) EPMQTGIKAVDSLVPIGRGQRELIIGDRQTGKTSIAIDTIIN
ATPA_MAIZE ( 145) EPLQTGLIAIDSMIPIGRGQRELIIGDRQTGKTAVATDTILN
ATPA_MARPO ( 145) EPMQTGLIAIDSMIPIGRGQRELIIGDRQTGKTAVAIDTILN
ATPA_MOUSE ( 187) EPMQTGIKAVDSLVPIGRGQRELIIGDRQTGKTSIAIDTIIN
ATPA_ODOSI ( 145) EPLQTGITSIDAMIPIGRGQRELIIGDRQTGKTAIAVDTIIN
ATPA_ORYSA ( 145) EPLQTGLIAIDSMIPIGRGQRELIIGDRQTGKTAVATDTILN
  ATPA_PEA ( 145) EPLQTGLIAIDSMIPIGRGQRELIIGDRQTGKTAVATDTILN
ATPA_PROMO ( 144) EPLQTGIKSIDGMVPIGRGQRELIIGDRQTGKTAVALDAIIN
ATPA_RHOBL ( 144) EPMATGLKAVDAMIPIGRGQRELIIGDRQTGKTAVALDTILN
ATPA_SCHPO ( 172) EPMQTGLKAIDSMVPIGRGQRELIIGDRQTGKTAIALDTILN
ATPA_SPIOL ( 145) EPLQTGLIAIDAMIPVGRGQRELIIGDRQTGKTAVATDTILN
ATPA_SYNP6 ( 145) EPMQTGITAIDAMIPIGRGQRELIIGDRQTGKTAIAIDTILN
ATPA_THEP3 ( 144) EPLQTGIKAIDALVPIGRGQRELIIGDRQTGKTSVAIDTIIN
ATPA_TOBAC ( 145) EPLQTGLIAIDSMIPIGRGQRELIIGDRQTGKTAVATDTILN
ATPA_VIBAL ( 144) QPVQTGYKSVDSMIPIGRGQRELIIGDRQIGKTALAIDAIIN
ATPA_WHEAT ( 145) EPLQTGLIAIDSMIPIGRGQRELIIGDRQTGKTAVATDTILN
ATPA_XENLA ( 188) EPMQTGIKAVDSLVPIGRGQRELIIGDRQTGKTSIAIDTIIN
ATPA_YEAST ( 181) EPVQTGLKAVDALVPIGRGQRELIIGDRQTGKTAVALDTILN
ATP2_HEVBR ( 212) QILVTGIKVVDLLAPYQRGGKIGLFGGAGVGKTVLIMELINN
ATP2_MAIZE ( 203) QILVTGIKVVDLLAPYQRGGKIGLFGGAGVGKTVLIMELINN
ATP2_NICPL ( 210) QILVTGIKVVDLLAPYQRGGKIGLFGGAGVGKTVLIMELINN
ATP2_ORYSA ( 201) QILVTGIKVVDLVAPYQRGGKIGLFGGAGVGKTVLIMELINN
ATPB_AEGCO ( 147) SIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
ATPB_ANASP ( 137) SVFETGIKVVDLLTPYRRGGKIGLFGGAGVGKTVIMMELINN
ATPB_ANGLY ( 145) SIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
ATPB_BACFR ( 132) EVLFTGIKVIDLLEPYSKGGKIGLFGGAGVGKTVLIMELINN
ATPB_BOVIN ( 181) EILVTGIKVVDLLAPYAKGGKIGLFGGAGVGKTVLIMELINN
ATPB_CHLRE ( 147) SIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
ATPB_CUSRE ( 145) SIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
ATPB_CYTLY ( 131) EVLFTGIKVIDLIEPYAKGGKIGLFGGAGVGKTVLIQELINN
ATPB_DICDH ( 136) AIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
ATPB_ECOLI ( 125) ELLETGIKVIDLMCPFAKGGKVGLFGGAGVGKTVNMMELIRN
ATPB_EUGGR ( 136) SIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
ATPB_HORVU ( 147) SIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
ATPB_HUMAN ( 181) EILVTGIKVVDLLAPYAKGGKIGLFGGAGVGKTVLIMELINN
ATPB_IPOBA ( 145) SIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
ATPB_LACCA ( 130) EILETGIKVIDLLEPYLRGGKVGLFGGAGVGKTVLIQELIHN
ATPB_MAIZE ( 147) SIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
ATPB_MARPO ( 145) SIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
ATPB_MYCGA ( 131) EIFETGIKVIDLLIPYAKGGKIGLFGGAGVGKTVLVQELIHN
ATPB_NEUCR ( 170) EILVTGIKVVDLLAPYARGGKIGLFGGAGVGKTVFIQELINN
ATPB_NICPL ( 147) SIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
ATPB_ORYSA ( 147) SIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
  ATPB_PEA ( 147) SIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
ATPB_PECFR ( 130) QILETGIKVVDLIAPYSRGGKIGLFGGAGVGKTVLIMELIHN
ATPB_PYLLI ( 136) AIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
  ATPB_RAT ( 181) EILVTGIKVVDLLAPYAKGGKIGLFGGAGVGKTVLIMELINN
ATPB_RHOBL ( 130) QILVTGIKVIDLLAPYSKGGKIGLFGGAGVGKTVLIQELINN
ATPB_SCHPO ( 178) EILETGIKVVDLLAPYARGGKIGLFGGAGVGKTVFIQELINN
ATPB_SPIOL ( 147) SIFETGIKVVNLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
ATPB_SYNP6 ( 137) KVFETGIKVIDLLAPYRQGGKIGLFGGAGVGKTVLIQELINN
ATPB_THEP3 ( 133) EILETGIKVVDLLAPYIKGGKIGLFGGAGVGKTVLIQELIHN
ATPB_TOBAC ( 147) SIFETGIEVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
ATPB_VIBAL ( 124) ALLETGVKVIDLICPFAKGGKIGLFGGAGVGKTVNMMELINN
ATPB_WHEAT ( 147) SIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINN
ATPB_YEAST ( 165) EILETGIKVVDLLAPYARGGKIGLFGGAGVGKTVFIQELINN
ATPX_BACFI ( 131) EILETGIKVVDLLAPYIIGGKIGLFGGAGVGKTVLIQELINN
//
```


g.deleage@ibcp.fr

Recherches de motifs – Construction du profil

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
A	3,8	0	0	2,6	0	0	0	0	44,9	0	0	11,5	0	41	0	0	12,8	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	2,6	0	0	0	0	0	0	0
D	0	0	0	1,3	0	0	0	0	0	0	97,4	0	0	0	0	0	0	0	0	0	0
E	62,8	0	0	35,9	0	0	0	1,3	0	0	0	0	0	3,8	0	0	0	0	0	0	0
F	0	0	25,6	2,6	0	0	0	0	0	0	0	0	0	0	0	2,6	0	0	0	0	0
G	0	0	0	1,3	0	100	0	0	0	0	0	1,3	0	0	0	0	50	0	100	50	0
H	1,3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	42,3	0	0	0	0	61,5	10,3	0	32,1	0	0	3,8	21,8	0	44,9	2,6	1,3	0	0	0
K	1,3	1,3	0	0	0	0	0	79,5	0	0	0	0	0	0	0	0	0	14,1	0	0	50
L	0	2,6	42,3	0	0	0	32,1	0	0	0	0	50	70,5	2,6	0	0	1,3	0	0	0	0
M	0	0	28,2	0	0	0	0	0	0	0	0	0	24,4	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	2,6	0	0	0	0	0	0	0	0	0	0
P	0	47,4	0	0	0	0	0	0	0	0	0	0	0	0	97,4	0	0	0	0	0	0
Q	10,3	0	0	44,9	0	0	0	0	0	0	0	0	0	0	0	0	5,1	1,3	0	50	0
R	0	0	0	0	0	0	0	2,6	0	0	0	0	0	0	0	0	24,4	83,3	0	0	50
S	20,5	0	0	0	0	0	0	0	5,1	0	0	37,2	0	0	0	0	3,8	0	0	0	0
T	0	0	0	0	98,7	0	0	5,1	0	0	0	0	0	1,3	2,6	0	0	0	0	0	0
V	0	6,4	3,8	11,5	1,3	0	3,8	1,3	50	67,9	0	0	1,3	26,9	0	5,1	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	2,6	0	0	0	0	0	0	0	0	47,4	0	0	0	0	0

g.deleage@ibcp.fr

Détection de motifs –Application du profil 1/2

- **Découpage de la séquence en j segments de longueur L**

  - 1 - L , 2 - L+ 1, 3 - L + 2 , . . . , (n - L) - n
  - Application du profil à chaque segment et calcul des scores obtenus sur un segment

$$\text{Score}(j)=\sum_{i=1}^{L} P(i)$$

- **Tri et affichage des x meilleurs scores**
  - Tous les segments
  - Tous les blocs
  - Normalisation du score d'un bloc par la valeur contenue dans le champ 99.5%=

g.deleage@ibcp.fr

Recherches de motifs – Application sur un segment 1/2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
	E	P	M	Q	T	G	I	K	A	V	D	S	L	V	P	I	G	R	G	Q	R	E
A	3,8	0	0	2,6	0	0	0	0	45	0	0	12	0	41	0	0	13	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	2,6	0	0	0	0	0	0	0	0
D	0	0	0	1,3	0	0	0	0	0	0	97	0	0	0	0	0	0	0	0	0	0	0
E	63	0	0	36	0	0	0	1,3	0	0	0	0	0	3,8	0	0	0	0	0	0	0	47
F	0	0	26	2,6	0	0	0	0	0	0	0	0	0	0	0	2,6	0	0	0	0	0	0
G	0	0	0	1,3	0	100	0	0	0	0	0	1,3	0	0	0	0	50	0	100	50	0	0
H	1,3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	42	0	0	0	0	62	10	0	32	0	0	3,8	22	0	45	2,6	1,3	0	0	0	49
K	1,3	1,3	0	0	0	0	0	80	0	0	0	0	0	0	0	0	0	14	0	0	50	0
L	0	2,6	42	0	0	0	32	0	0	0	0	50	71	2,6	0	0	1,3	0	0	0	0	0
M	0	0	28	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	1,3
N	0	0	0	0	0	0	0	0	0	0	2,6	0	0	0	0	0	0	0	0	0	0	0
P	0	47	0	0	0	0	0	0	0	0	0	0	0	0	97	0	0	0	0	0	0	0
Q	10	0	0	45	0	0	0	0	0	0	0	0	0	0	0	0	5,1	1,3	0	50	0	0
R	0	0	0	0	0	0	0	2,6	0	0	0	0	0	0	0	0	24	83	0	0	50	0
S	21	0	0	0	0	0	0	0	5,1	0	0	37	0	0	0	0	3,8	0	0	0	0	0
T	0	0	0	0	99	0	0	5,1	0	0	0	0	0	1,3	2,6	0	0	0	0	0	0	0
V	0	6,4	3,8	12	1,3	0	3,8	1,3	50	68	0	0	1,3	27	0	5,1	0	0	0	0	0	2,6
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	2,6	0	0	0	0	0	0	0	0	47	0	0	0	0	0	0

Somme sur tout le segment =2527 SCORE=2527/873*1000=2895

Résultats avec ANTHEPROT

g.deleage@ibcp.fr

Recherche avec ANTHEPROT et l'ATPsynthase de S. cerevisae

Site file viewer:C:\anthepro\Data\1AF1_SCEREV.BLO

File Settings

```
Protein C:1AF1_SCEREV
Reference BLOCKS :C:.DAT Content : 2884 profiles.
Uppercase blue characters have already been observed at the current position in the given
block
Lowercase green characters have never been observed at the current position in the given
block
Elapsed searching time: 3 s
==============================   Best 100 BLOCKS   ==============================
 BL00152E;  ATP synthase alpha and beta subunits proteins.
 Normed score :  2535  Threshold :  2940  99.5% value :  981 from  323 to  376 in
sequence : GSLTALPVIETQGGDVSAYIPTNVISITDGQIFLEAEeFYKGIRPAINVGLSVS
 BL00152A;  ATP synthase alpha and beta subunits proteins.
 Normed score :  2229  Threshold :  2044  99.5% value :  635 from  95 to  120 in sequence
: TGNIVDVPVGPGLLGRVVDALGNPID
 BL00152B;  ATP synthase alpha and beta subunits proteins.
 Normed score :  2105  Threshold :  2546  99.5% value :  873 from  148 to  189 in
sequence : EPVQTGLKAVDALVPIGRGQRELIIGDRQTGKTAVALDTILN
 BL00152C;  ATP synthase alpha and beta subunits proteins.
 Normed score :  1754  Threshold :  1664  99.5% value :  405 from  252 to  263 in
sequence : FTAASIGEWFRD
 BL00152D;  ATP synthase alpha and beta subunits proteins.
 Normed score :  1306  Threshold :  1603  99.5% value :  449 from  286 to  299 in
sequence : SLMLRRPPGREAYP
 BL00170B;  Cyclophilin-type peptidyl-prolyl cis-trans isomerase signatur.
 Normed score :  1231  Threshold :  3163  99.5% value :  776 from  319 to  359 in
sequence : kEGSgsltalPvietQGGDVSAyipTnvISItdGqiflEAe
```

Détection de motifs –Améliorations

- **Méthode la plus sensible pour rechercher des motifs très flous**
  - Détection de motifs encore non décrits dans les banques
  - Problème du calcul statistique d'un « score seuil de significativité »
  - Temps de calcul non négligeable pour « balayer les banques »
- **Amélioration du système de construction de profils**
  - Applications de groupes d'acides aminés
  - Prise en compte des sous blocs (représentativité des séquences d'un sous bloc)
  - Utilisations de matrices de substitution
  - Gestion des gaps
- **HMMER** (Chaine de Markov)
- **PFTools** (Profils généralisés)

g.deleage@ibcp.fr

Amélioration des profils - Groupes d'acides aminés

Alignement

F	K	L	L	S	H	C	L	L	V
F	K	A	F	G	Q	T	M	F	Q
Y	P	I	V	G	Q	E	L	L	G
F	P	V	V	K	E	A	I	L	K
F	K	V	L	A	A	V	I	A	D
L	E	F	I	S	E	C	I	I	Q
F	K	L	L	G	N	V	L	V	C

Exemple de profil

A	-18	-10	-1	-8	8	-3	3	-10	-2	-8
C	-22	-33	-18	-18	-22	-26	22	-24	-19	-7
D	-35	0	-32	-33	-7	6	-17	-34	-31	0
E	-27	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	14	-26	-29	-15	4	12	-29
G	-30	-20	-28	-32	28	-14	-23	-33	-27	-5
H	-13	-12	-25	-25	-16	14	-22	-22	-23	-10
I	3	-27	21	25	-29	-23	-8	33	19	-23
K	-26	25	-25	-27	-6	4	-15	-27	-26	0
L	14	-28	19	27	-27	-20	-9	33	26	-21
M	3	-15	10	14	-17	-10	-9	25	12	-11
N	-22	-6	-24	-27	1	8	-15	-24	-24	-4
P	-30	24	-26	-28	-14	-10	-22	-24	-26	-18
Q	-32	5	-25	-26	-9	24	-16	-17	-23	7
R	-18	9	-22	-22	-10	0	-18	-23	-22	-4
S	-22	-8	-16	-21	11	2	-1	-24	-19	-4
T	-10	-10	-6	-7	-5	-8	2	-10	-7	-11
V	0	-25	22	25	-19	-26	6	19	16	-16
W	9	-25	-18	-19	-25	-27	-34	-20	-17	-28
Y	34	-18	-1	1	-23	-12	-19	0	0	-18

g.deleage@ibcp.fr

Amélioration des profils - Prise en compte des sous blocs

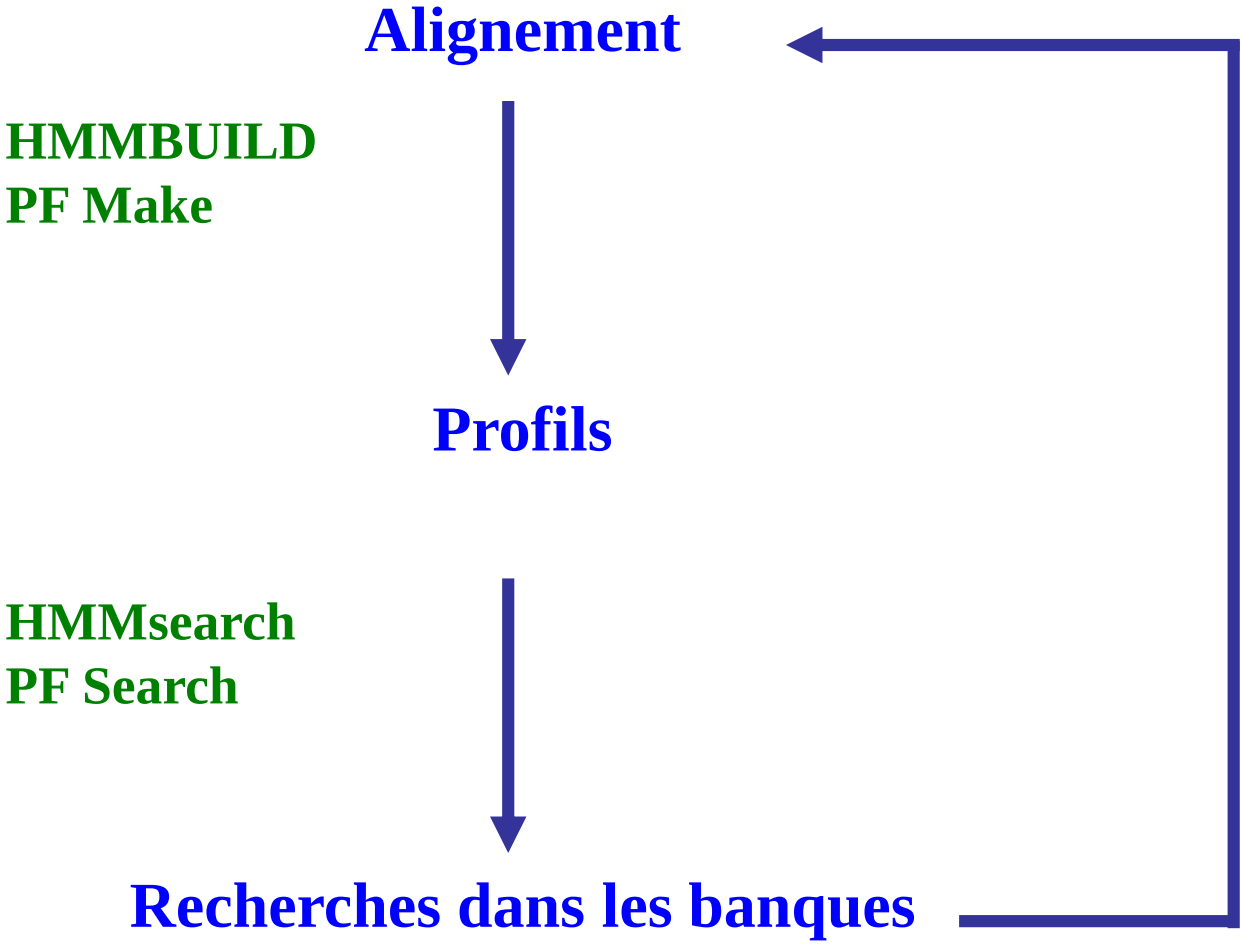
1. **Comptage du nombre de sous-blocs = NBLOCKS [3]**
2. **Pour chaque sous-bloc IBL{1,NBLOCKS} comptage du nombre de séquences ISEQ(IBL) [1,1,15]**
3. **Pour chaque position de chaque séquence d'un sous-bloc IBL**
4. **Comptage de chaque type aa {1,20} divisé par NBLOCKS et par ISEQ(IBL)*100**
5. **Calcul du profil de conservation des acides aminés**

```
FLII_BACSU ( 142) EKMGVGVRSIDSLLTVGKGQRIGIFAGSGVGKSTLMGMIAKQ 100

FLII_SALTY ( 157) HVLDTGVRAINALLTVGRGQRMGLFAGSGVGKSVLLGMMARY  98

ATP0_BETVU ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN  10
ATP0_BOVIN ( 187) EPMQTGIKAVDSLVPIGRGQRELIIGDRQTGKTSIAIDTIIN  11
ATP0_BRANA ( 146) EPMQTGLKAVDSLVPIGRGQRELLIGDRQTGKTTIAIDTILN  11
ATP0_HELAN ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN  10
ATP0_MAIZE ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN  10
ATP0_MARPO ( 145) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN  10
ATP0_NICPL ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN  10
ATP0_OENBI ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN  10
ATP0_ORYSA ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN  10
  ATP0_PEA ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN  10
ATP0_PHAVU ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN  10
ATP0_RAPSA ( 146) EPMQTGLKAVDSLVPIGRGQRELLIGDRQTGKTTIAIDTILN  11
ATP0_SOYBN ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN  10
ATP0_WHEAT ( 146) EPMQTGLKAVDSLVPIGRGQRELIIGDRQTGKTAIAIDTILN  10
ATPA_ANTSP ( 147) EPLQTGITAIDSMIPIGRGQRELIIGDRQTGKTTVALDTIIN  13
```


g.deleage@ibcp.fr

Profils Hmmer et PF tools

http://hmmer.org/



g.deleage@ibcp.fr

PF profiles


```
ID   SEQUENCE_RPOFILE; MATRIX.
AC   ZZ99999;
DT   Mon Feb  3 16:44:20 2003
DE   Generated from MSF file: '/pbil/servers/npsa/www/tmp/20116.wmsf'.
MA   /GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=20;
MA   /DISJOINT: DEFINITION=PROTECT; N1=3; N2=18;
MA   /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=-0.7424185; R2=0.0218336; TEXT='-LogE';
MA   /CUT_OFF: LEVEL=0; SCORE=424; N_SCORE=8.5000; MODE=1; TEXT='!';
MA   /CUT_OFF: LEVEL=-1; SCORE=332; N_SCORE=6.5000; MODE=1; TEXT='?';
MA   /DEFAULT: M0=-7; D=-20; I=-20; MI=-105; MD=-105; IM=-105; DM=-105;
MA   /I: E1=*; BI=-105; BD=-105;
MA   /M: SY='M'; M=-10,-20,-20,-30,-20,0,-20,0,20,-10,20,60,-20,-20,0,-10,-20,-10,10,-20,0,-10;
MA   /M: SY='N'; M=-10,40,-20,20,0,-20,0,10,-20,0,-30,-20,60,-20,0,0,10,0,-30,-40,-20,0;
MA   /M: SY='T'; M=0,0,-10,-10,-10,-10,-20,-20,-10,-10,-10,-10,0,-10,-10,-10,20,50,0,-30,-10,-10;
MA   /M: SY='M'; M=-13,2,-23,2,-7,-13,-17,0,1,-7,4,31,-7,-17,0,-10,-14,-10,-3,-26,-6,-4;
MA   /M: SY='M'; M=-9,3,-23,3,-1,-19,-15,-4,-12,7,-12,10,-2,-14,3,0,-3,-5,-9,-27,-10,1;
MA   /M: SY='K'; M=-8,13,-23,4,3,-23,-14,-6,-23,22,-26,-13,19,-13,3,13,2,5,-19,-28,-13,3;
MA   /M: SY='I'; M=-7,-22,-22,-31,-24,-2,-30,-20,30,-20,14,25,-18,-19,-14,-20,-12,1,22,-23,-3,-21;
MA   /M: SY='T'; M=-3,0,-15,-7,-5,-15,-20,-17,-15,6,-15,-10,0,-10,-5,1,12,34,-5,-27,-10,-5;
MA   /M: SY='I'; M=-10,-30,-27,-37,-27,3,-37,-27,42,-30,28,20,-23,-23,-20,-27,-23,-10,25,-20,0,-27;
MA   /M: SY='Y'; M=-20,-1,-30,5,-9,11,-25,15,-11,-7,-8,-8,-9,-25,-7,-10,-15,-10,-15,11,53,-12;
MA   /M: SY='D'; M=-17,39,-30,56,31,-37,-13,0,-37,3,-27,-27,15,-7,5,-7,0,-10,-30,-37,-20,18;
MA   /M: SY='V'; M=-5,-30,-17,-33,-28,2,-33,-28,34,-25,20,15,-27,-27,-25,-23,-16,-5,37,-25,-5,-28;
MA   /M: SY='A'; M=50,-10,-10,-20,-10,-20,0,-20,-10,-10,-10,-10,-10,-10,-10,-20,10,0,0,-20,-20,-10;
MA   /M: SY='R'; M=-16,-5,-30,-3,13,-24,-20,-2,-30,30,-22,-12,0,-15,12,50,-8,-10,-22,-22,-12,9;
MA   /M: SY='L'; M=0,-12,-22,-10,12,-10,-21,-13,-4,-12,14,0,-15,-15,-3,-13,-12,-8,-7,-24,-11,4;
MA   /M: SY='A'; M=43,-8,-10,-16,-8,-20,0,-18,-12,-10,-14,-12,-6,-10,-8,-18,16,4,-2,-24,-20,-8;
MA   /M: SY='G'; M=0,-3,-27,-5,-15,-28,54,-15,-35,-16,-30,-20,9,-19,-15,-16,5,-13,-28,-25,-28,-15;
MA   /M: SY='V'; M=0,-30,-10,-30,-30,0,-30,-30,30,-20,10,10,-30,-30,-30,-20,-10,0,50,-30,-10,-30;
MA   /M: SY='S'; M=10,0,-10,0,0,-20,0,-10,-20,-10,-30,-20,10,-10,0,-10,40,20,-10,-40,-20,0;
MA   /M: SY='M'; M=-2,-17,-23,-24,-14,-9,-20,-9,6,-2,3,21,-13,-18,-2,7,-12,-8,3,-20,-6,-11;
MA   /I: B1=*; IE=-105; DE=-105;
CC   /RESCALED_BY=" /pbil/bin/pftools/pfscale /pbil/servers/npsa/www/tmp/20116.finalsortlist
/pbil/servers/npsa/www/tmp/20116.pr..";
CC   /GENERATED_BY="/pbil/bin/pftools/pfmake -3s /pbil/servers/npsa/www/tmp/20116.wmsf
/pbil/bin/pftools/blosum45.cmp E=0.2 S=0...";
//
```


g.deleage@ibcp.fr

PSORT (Norton & Nakai): Intégration de plusieurs recherches de motifs

Méthode de plus proche voisins (k-NN)

- ✓Recognition of Signal Sequence (Matrice de poids-Von Heijne)
- ✓Recognition of Transmembrane Segments
- ✓Prediction of Membrane Topology
- ✓Recognition of Mitochondrial Proteins (Analyse discriminante+patterns)
- ✓Recognition of Nuclear Proteins (Signatures)
- ✓Recognition of Peroxisomal Proteins (Signatures)
- ✓Recognition of Chloroplast Proteins (Amphiphilie forte)
- ✓Recognition of ER (endoplasmic reticulum) Proteins
- ✓Analysis of Proteins in Vesicular Pathway (YQRL signature)
- ✓Lysosomal and Vacuolar Proteins
- ✓Lipid Anchors (Signature)
- ✓Miscellaneous Motifs
- ✓Coiled-coil Structure (Lupas)

g.deleage@ibcp.fr

Modifications post-traductionnelles

- Sulfinator Prediction of tyrosine sulfation sites
- PSORT - Prediction of protein sorting signals and localization sites
- SignalP - Prediction of signal peptide cleavage sites
- ChloroP - Prediction of chloroplast transit peptides
- MITOPROT - Prediction of mitochondrial targeting sequences
- Predotar - Prediction of mitochondrial and plastid targeting sequences
- NetOGlyc - Prediction of type O-glycosylation sites in mammalian proteins
- DictyOGlyc - Prediction of GlcNAc O-glycosylation sites in Dictyostelium
- YinOYang - O-beta-GlcNAc attachment sites in eukaryotic protein sequences
- big-PI Predictor - GPI Modification Site Prediction
- DGPI - Prediction of GPI-anchor and cleavage sites
- NetPhos - Prediction of Ser, Thr and Tyr phosphorylation sites in eukaryotic proteins
- NetPicoRNA - Prediction of protease cleavage sites in picornaviral proteins

g.deleage@ibcp.fr

Exemple de calcul de sites PROSITE sur une séquence

ANTHEPROT (http://antheprot-pbil.ibcp.fr)

PROSCAN http://npsa-pbil.ibcp.fr/NPSA/npsa_proscan.html

Interpro : http://npsa-pbil.ibcp.fr/iprscan/iprscan

g.deleage@ibcp.fr

Profils physico-chimiques

Profil d'hydrophobie

Prédire les régions transmembranaires
Le degré de structuration d'une protéine (PINS)
Les régions enfouies (stratégie d'obtention d'anticorps)

g.deleage@ibcp.fr

Obtention d'AC anti-protéine

Séquence

KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNSQATNRNTDGSTDYGVLQINSRWWCNDG
KTPGSRNLCNIPCSALLSSDITATVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL

Prédiction des zones antigéniques

KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNSQATNRNTDGSTDYGVLQINSRWWCNDG
KTPGSRNLCNIPCSALLSSDITATVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL

Synthèse chimique

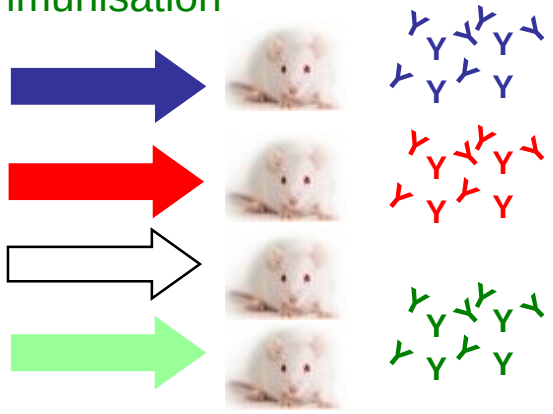
Immunisation

DNYRGYS

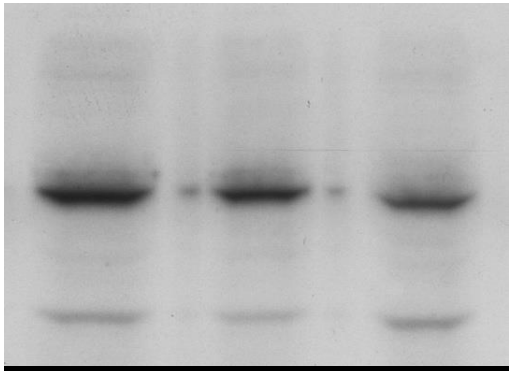
ATNRNTDGSTD

NDGKTPGSRN

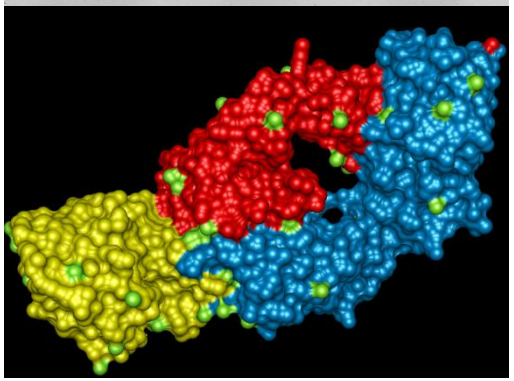
RNRCKGTD



Test en Western blot



Structure 3D du complexe Ag-Ac

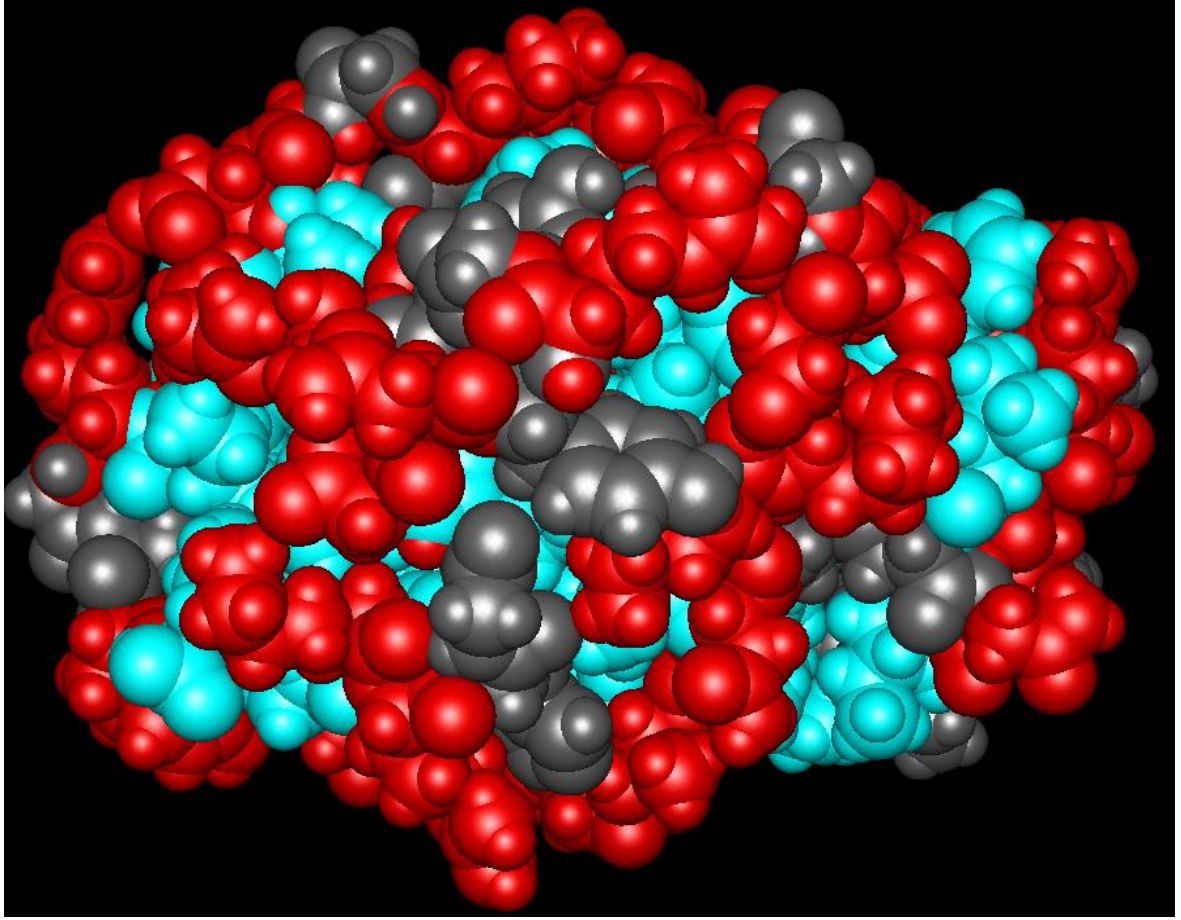


g.deleage@ibcp.fr

Prédiction des régions antigéniques

Les résidus hydrophobes sont majoritairement enfouis et ne peuvent pas être reconnus par les Ac

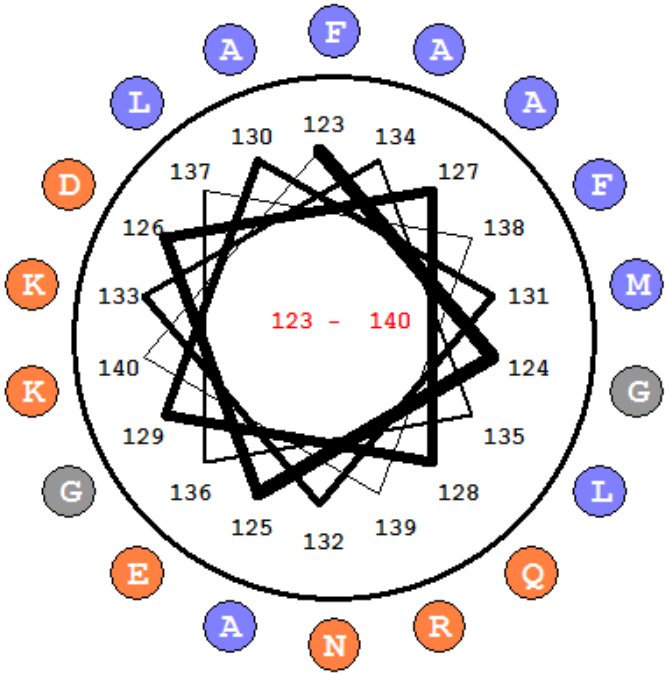
Les résidus hydrophiles sont exposés au solvant et peuvent être reconnus par les Ac



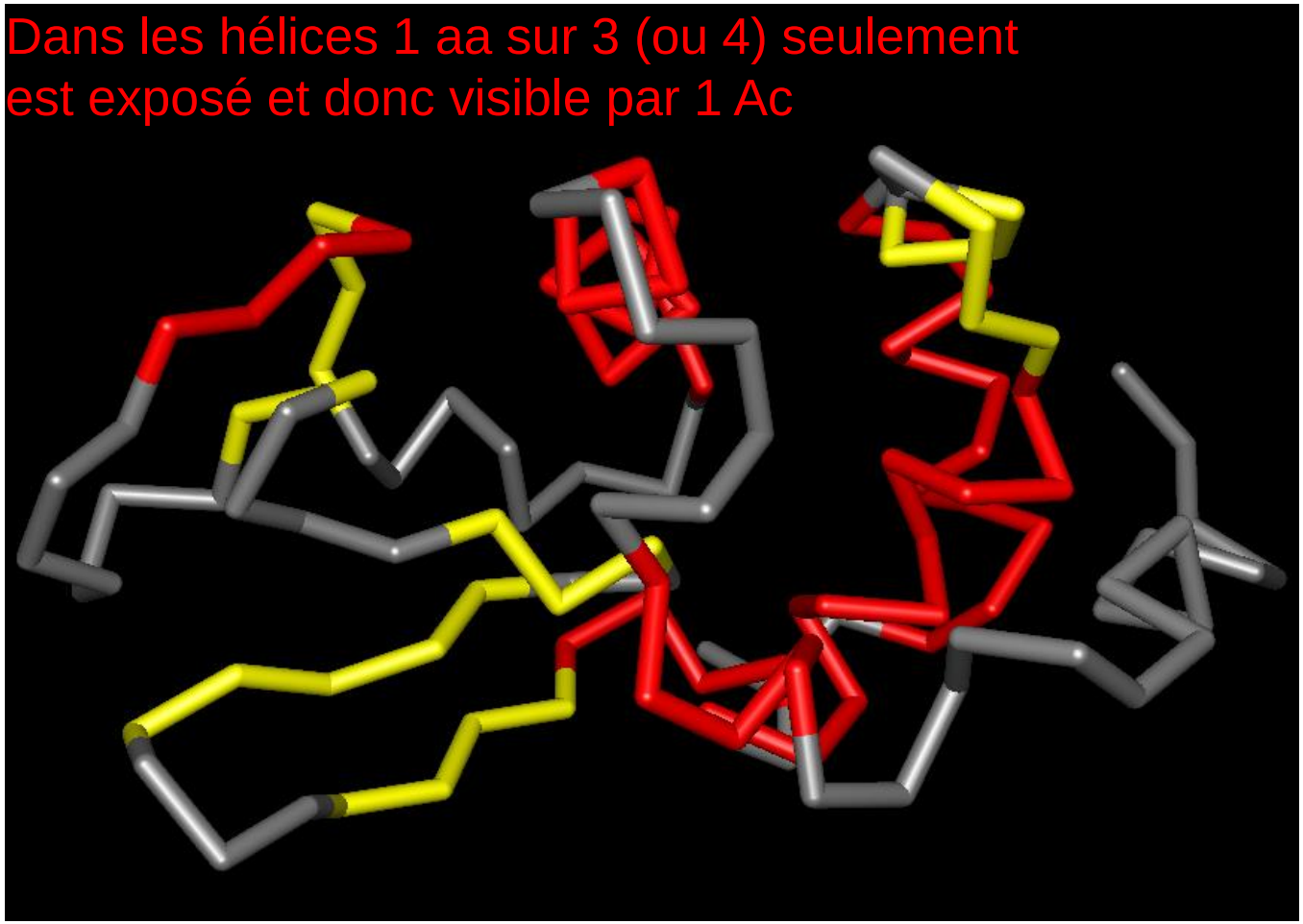
g.deleage@ibcp.fr

Prédiction des régions antigéniques

C:\anthepro\mbn.seq
FGADAQGAMNKALELFRK

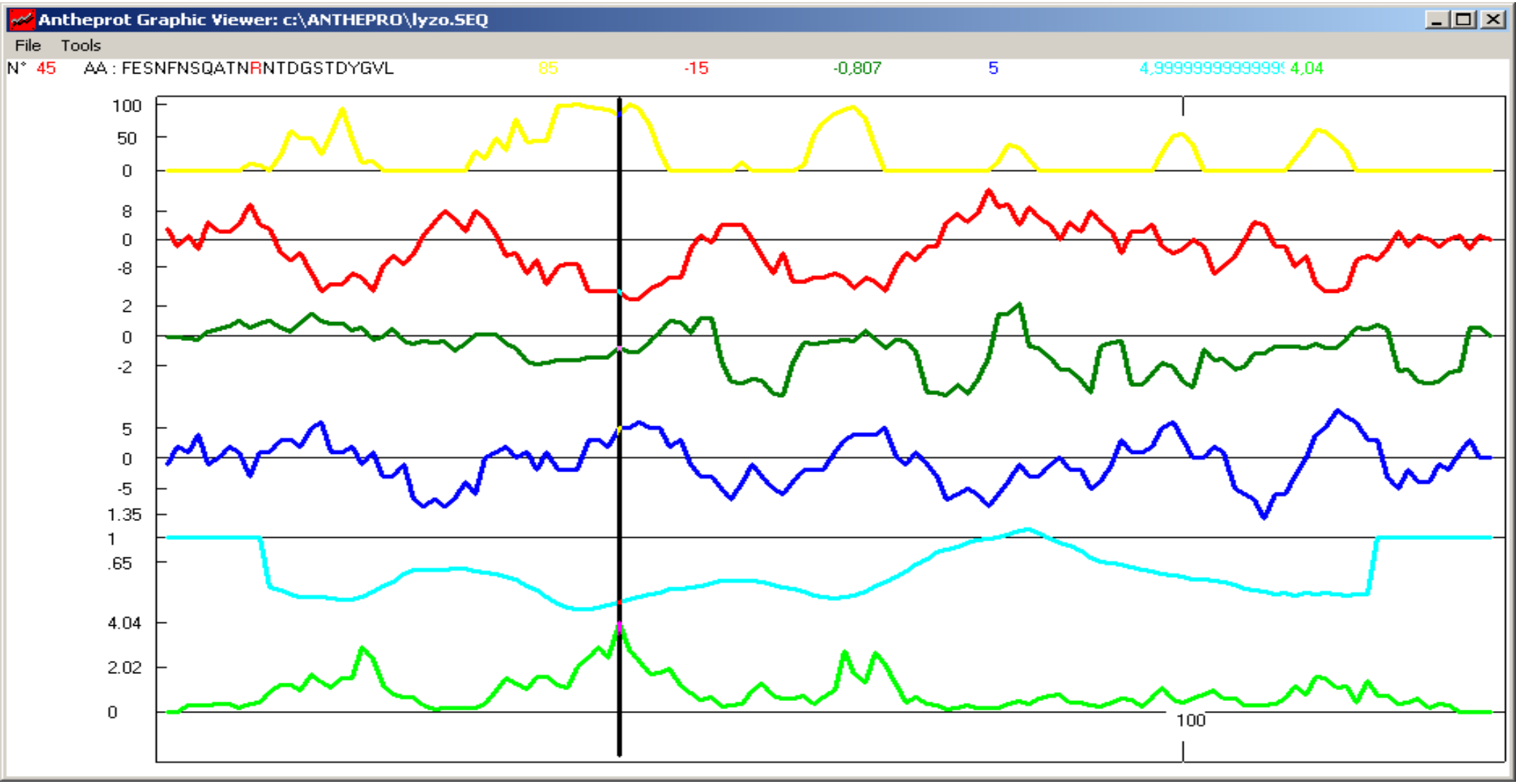


Dans les hélices 1 aa sur 3 (ou 4) seulement est exposé et donc visible par 1 Ac



g.deleage@ibcp.fr

Prédiction des régions antigéniques

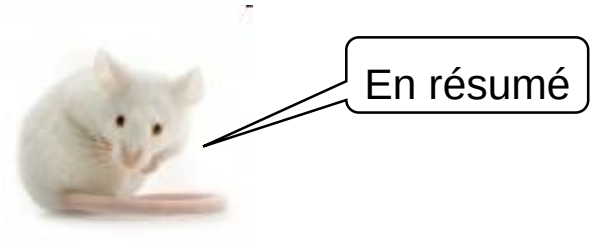


KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNSQATNRNTDGSTDYGVLQINSRWWCNDG
KTPGSRNLCNIPCSALLSSDITATVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL

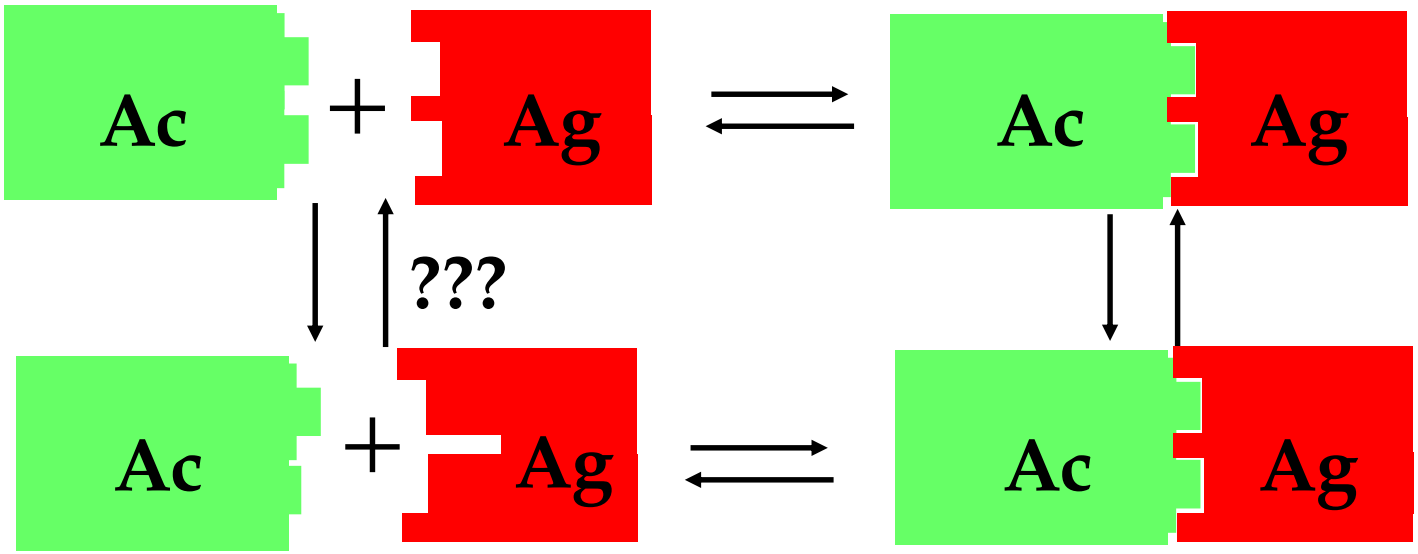
g.deleage@ibcp.fr

Prédictions d'antigénicité

- ***Hydrophilie*** : Les protéines globulaires solubles enfouissent les résidus hydrophobes et exposent les zones hydrophiles
  - Intérêt pour les protéines qui interagissent
  - Discrimination des zones externes
  - Protéines membranaires



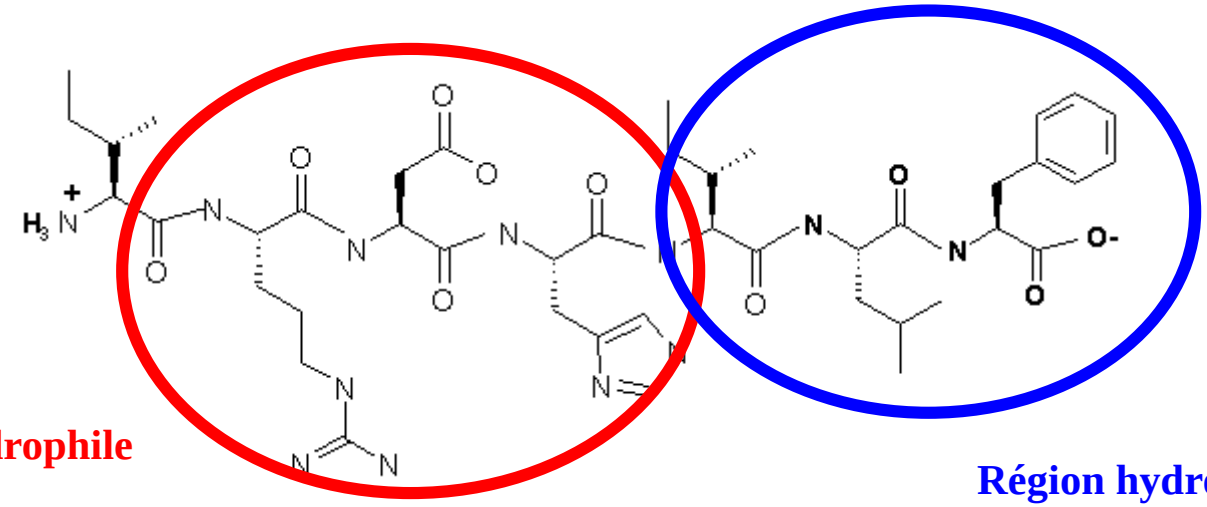
- ***Accessibilité*** : ce qui est accessible peut être antigénique (nécessaire pas suffisant)
- ***Flexibilité*** : ce qui est flexible peut être prédit en prenant en compte les données des structures 3D.



g.deleage@ibcp.fr

Profils d'hydrophobie

- **Hydrophobie** : Enchaînement de *résidus hydrophobes* (R1,R2,R3) (tous les aa sont hydrophiles)



Région hydrophile

Région hydrophobe

- Δ G de transfert de chaîne latérale de H20 dans ethanol
  - Δ G < 0 Ile, Val, Phe... Δ G = -1 à -3 kcal/mole
  - Δ G > 0 Asp, Glu, Ser... Δ G = 1 à 3 kcal/mole
- Δ G de transfert de chaîne latérale de H20 en vapeur
  - Δ G < 0 Ile, Val, Phe... Δ G = -3 kcal/mole
  - Δ G > 0 Asp, Glu, Ser... Δ G = 10 kcal/mole

g.deleage@ibcp.fr

Echelle et profils d'hydrophobie

Paramètres :

Ile	4,5	Val	4,2	Leu	3,8
Phe	2,8	Cys	2,5	Met	1,9
Ala	1,8	Gly	-0,4	Thr	-0,7
Ser	-0,8	Trp	-0,9	Tyr	-1,3
Pro	-1,6	His	-3,2	Asx	-3,5
Glx	-3,5	Lys	-3,9	Arg	-4,5

Algorithme:

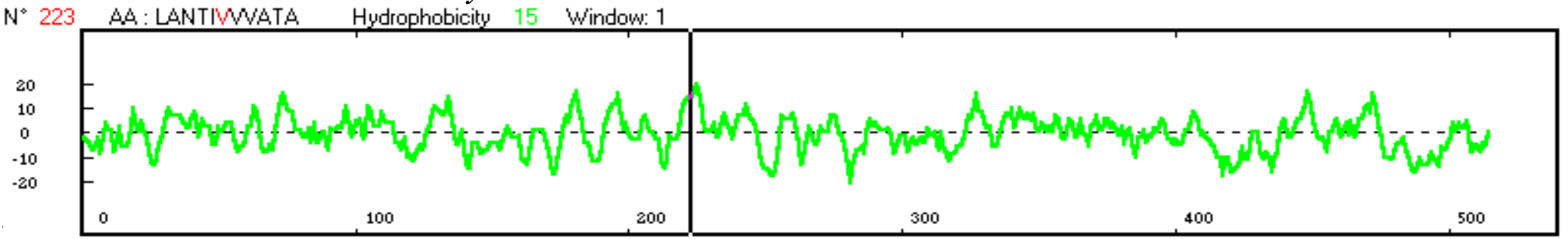
- *Moyenne sur 7 , 9, 11, 13 ou toute valeur impaire:*

..Q - N - V - E - D - S - G - I....

H(E)= (-3,5 - 3,5 +4,2 -3,5 -3,5 - 0,8 -0,4) / 7 =-1,57

H(D)= (- 3,5 +4,2 -3,5 -3,5 - 0,8 -0,4+4,5) / 7 =-0,42

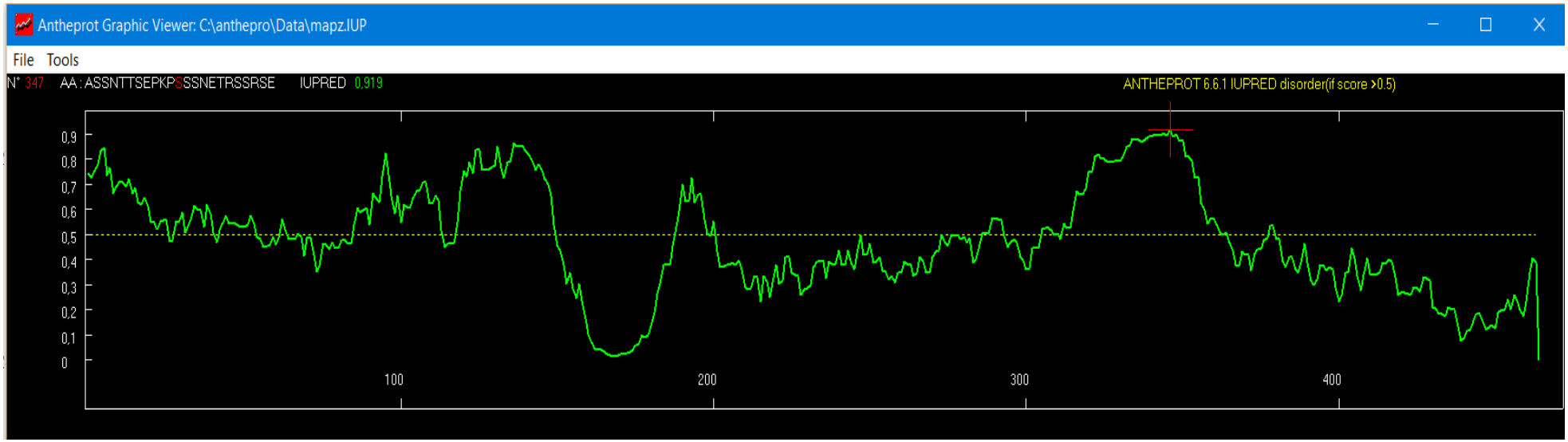
- *Courbes des moyennes obtenues*



g.deleage@ibcp.fr

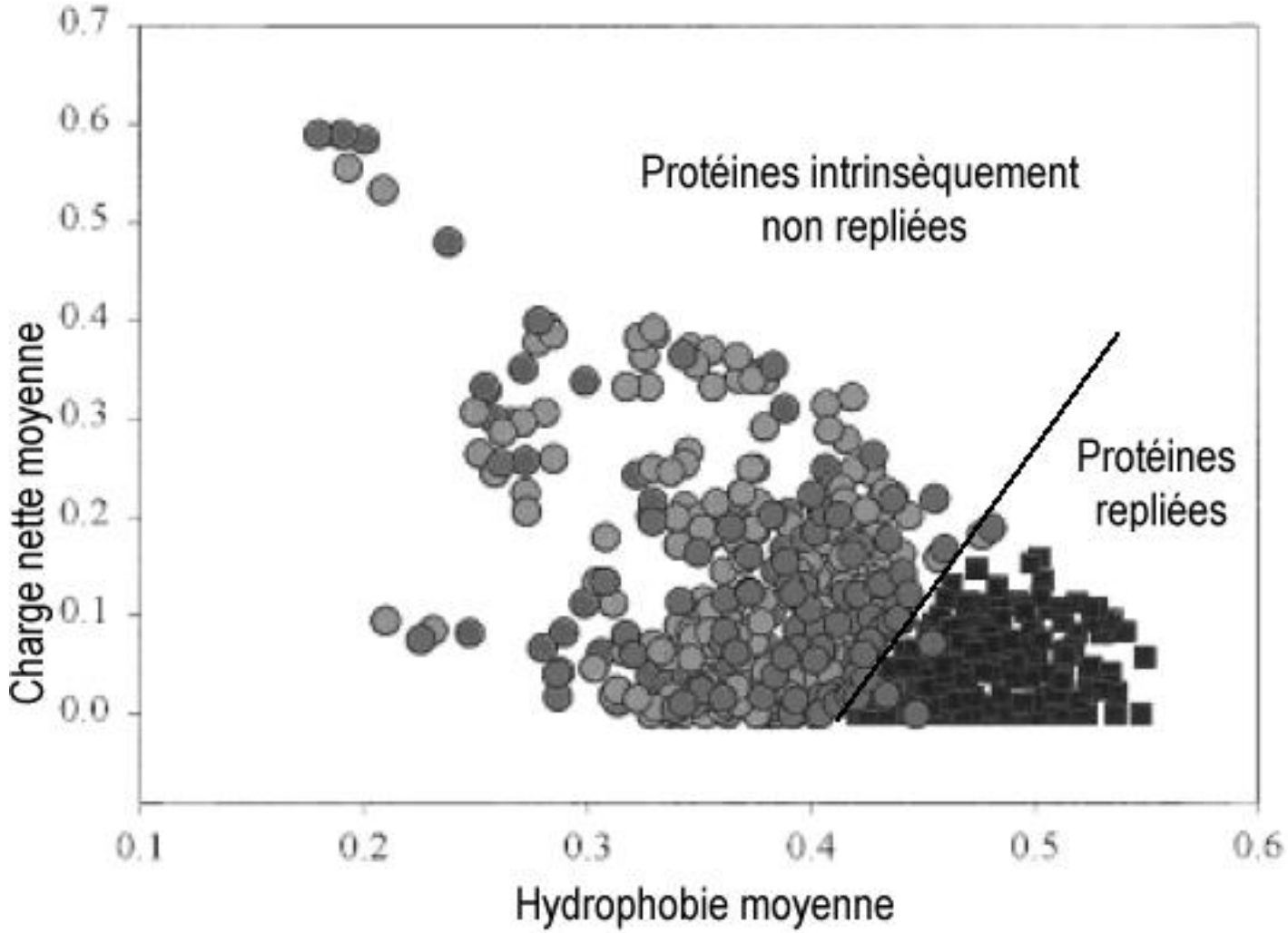
Prédire les régions « PINS »

PONDR	*http://www.pondr.com*
Charge/hydropathy method	*http://www.pondr.com*
DisEMBL	*http://dis.embl.de*
GLOBPLOT	*http://globplot.embl.de*
FOLDINDEX	*http://bip.weizmann.ac.il/fldbin/findex*
Hydrophobic cluster analysis (HCA)	*http://smi.snv.jussieu.fr/hca/hca-seq.html*
RONN	*http://www.strubi.ox.ac.uk/RONN*
NORSp	*http://cubic.bioc.columbia.edu/services/NORSp*
DISOPRED	*http://bioinfo.cs.ucl.ac.uk/disopred*
IUPRED	*http://iupred.enzim.hu*



g.deleage@ibcp.fr

Protéines Intrinsèquement non structurées (PINS)

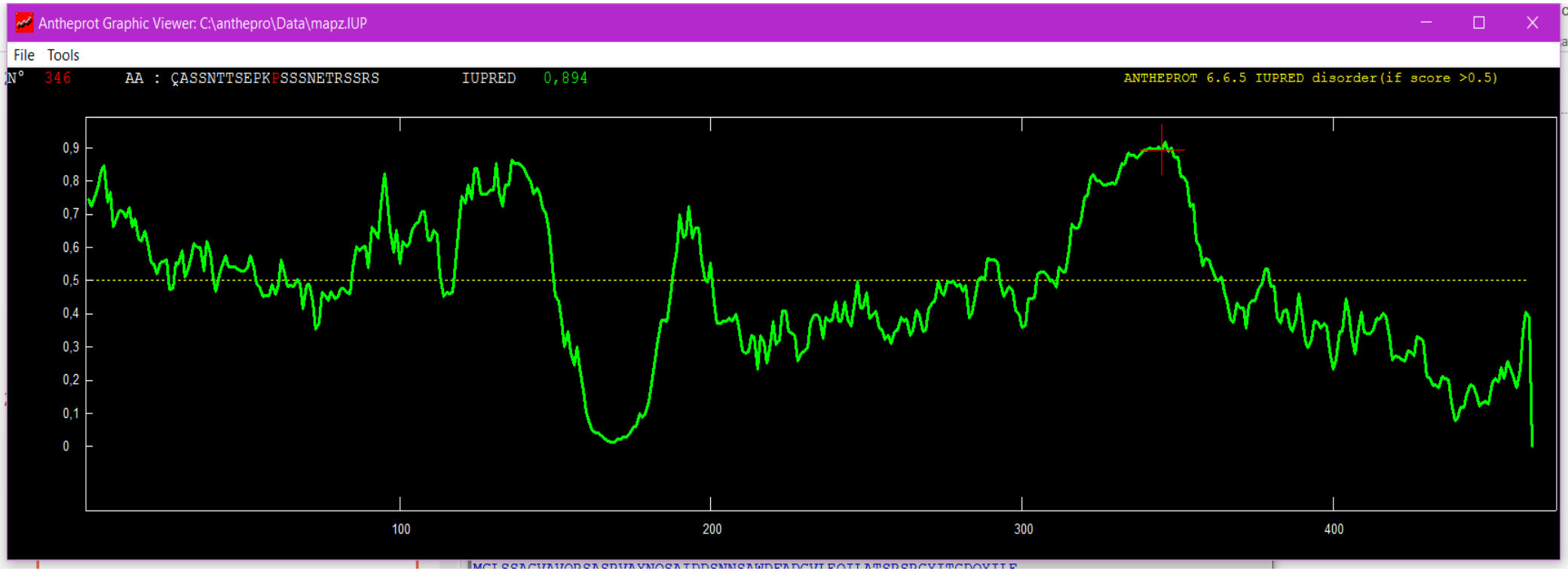


g.deleage@ibcp.fr

Protéines Intrinsèquement non structurées (IUPRED)

http://iupred.enzim.hu/

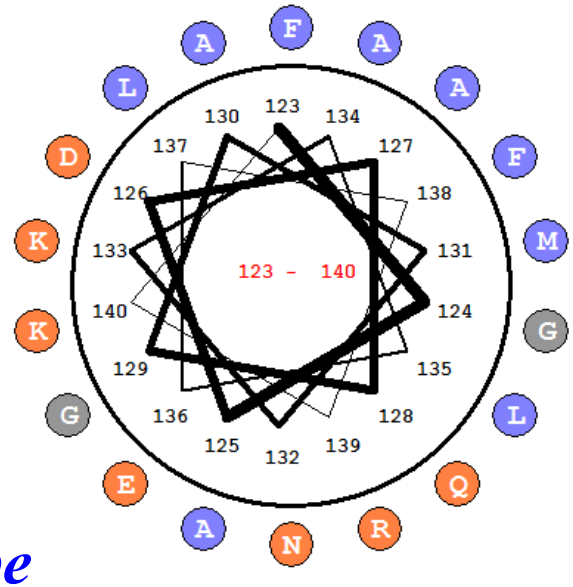
IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content
Zsuzsanna Dosztányi, Veronika Csizmók, Péter Tompa and István Simon
Bioinformatics (2005) 21, 3433-3434.



Les régions au dessus de 0.5 (trait pointillé) sont considérées comme non structurées.

g.deleage@ibcp.fr

Amphiphilie des structures



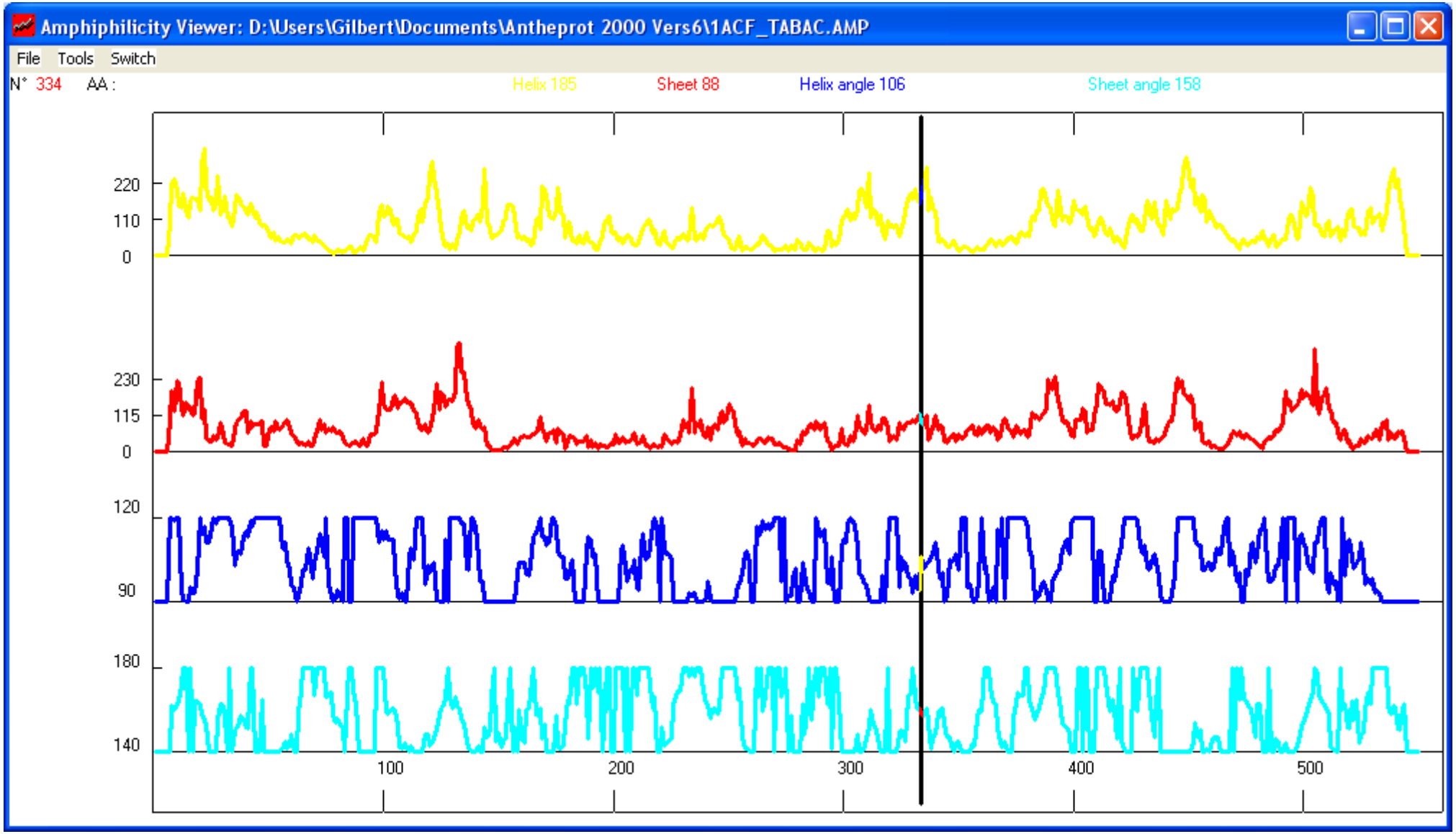
1MBN

- ***Principe***
  - Inégalité de répartition des résidus hydrophobes et hydrophiles
- ***Paramètres***
  - Une échelle d'hydrophobicité
- ***Algorithme:***

$$A(i) = \sqrt{\left[\sum_{i=1}^{7}\left(H(i)\sin(\delta i)\right)\right]^2 + \left[\sum_{i=1}^{7}\left(H(i)\cos(\delta i)\right)\right]^2}$$

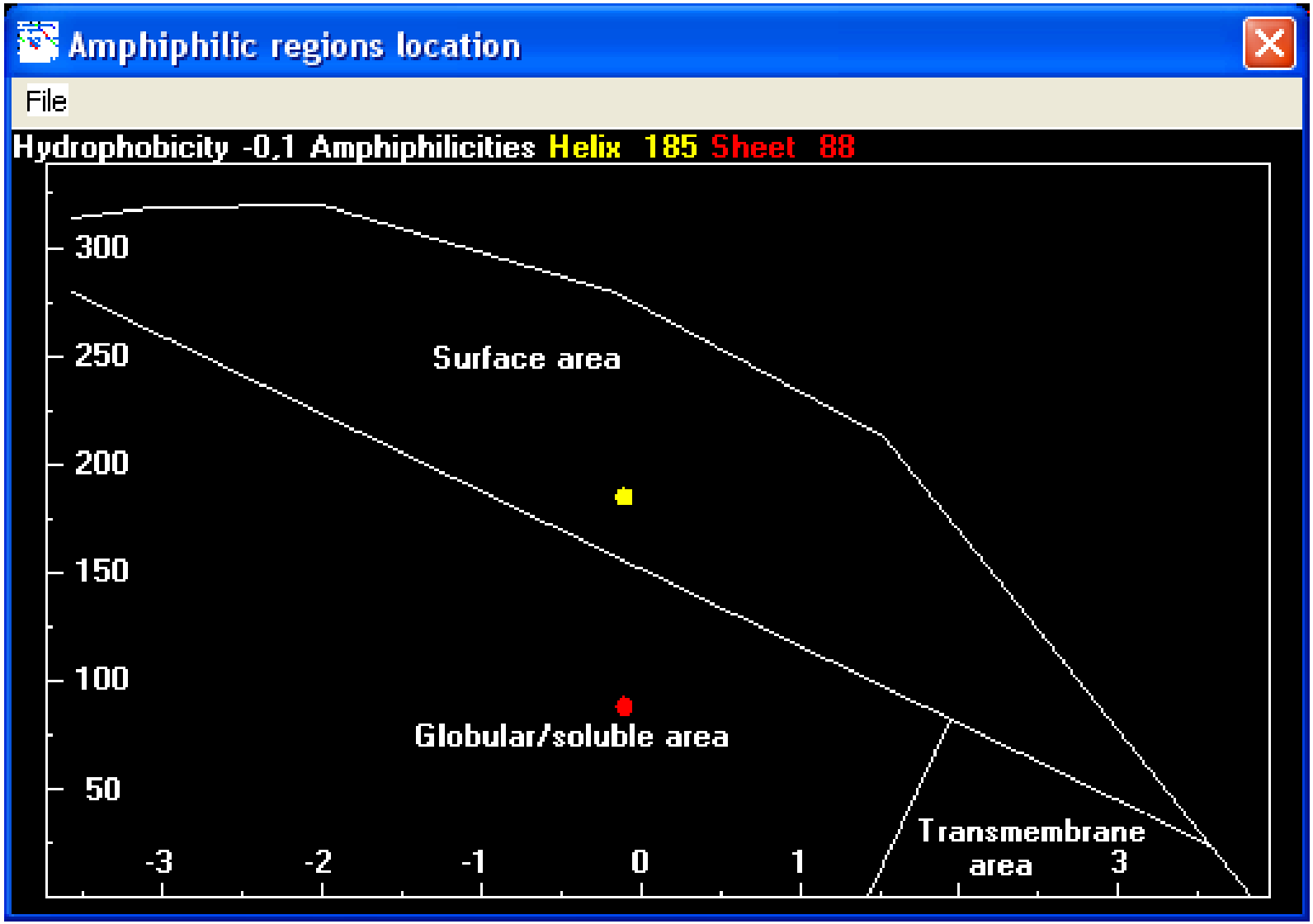
g.deleage@ibcp.fr

Profils d'amphiphilicité



g.deleage@ibcp.fr

Profils d'amphiphilicité



Amphiphilie

Hydrophobie

g.deleage@ibcp.fr

Profils de surface accessible au solvant

- ***Principe***
  - Approche heuristique pour la sélection d'antigènes peptidiques synthétiques
- ***Paramètres***
  - Une échelle d 'accessibilité mesurée comme la fraction δ d'acides aminés exposés à la surface des protéines (>20 A²) à partir des structures 3D.
  - $0,26 < \delta < 0,97$

Ile	0,34	Val	0,36	Leu	0,40
Phe	0,42	Cys	0,26	Met	0,48
Ala	0,49	Gly	0,48	Thr	0,70
Ser	0,65	Trp	0,51	Tyr	0,76
Pro	0,75	His	0,66	Lys	0,97
Arg	0,95	Glx	0,84	Asx	0,80

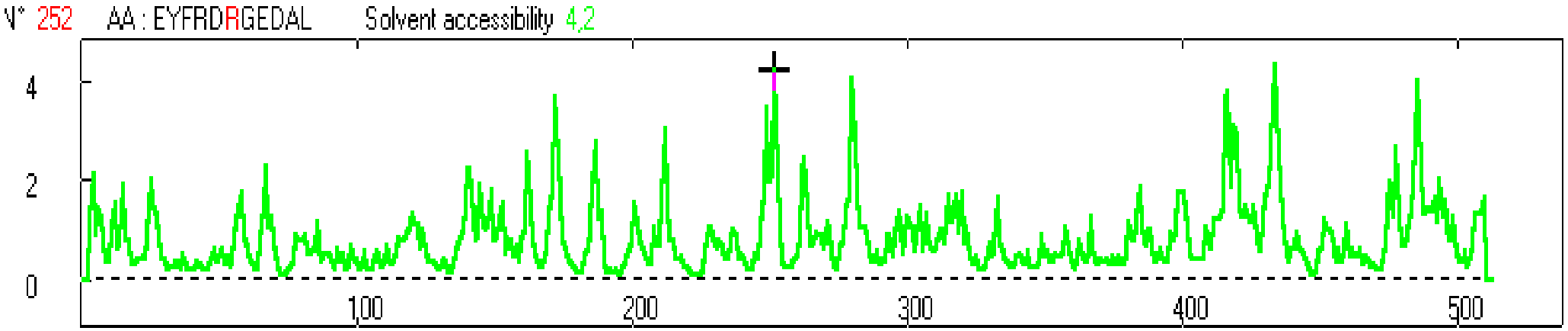
- ***Algorithme:***
  - *Produit normé sur 6 acides aminés*

$$Sn = \left[\prod_{i=1}^{6} \delta_{n+4-i} \right] .(0,62)^{-6}$$

g.deleage@ibcp.fr

Profils de surface accessible au solvant

- ***Caractéristiques***
  - Max de Sn = 14,7
  - Min de Sn =0.005
  - Séquence moyenne = 1



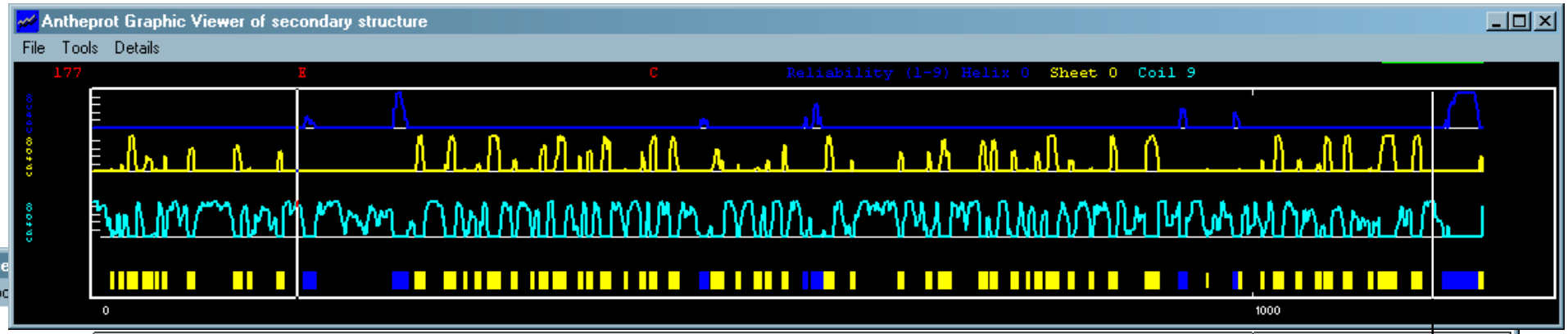
- Boger et al., (1986) 6th international congress of Immunology, Toronto

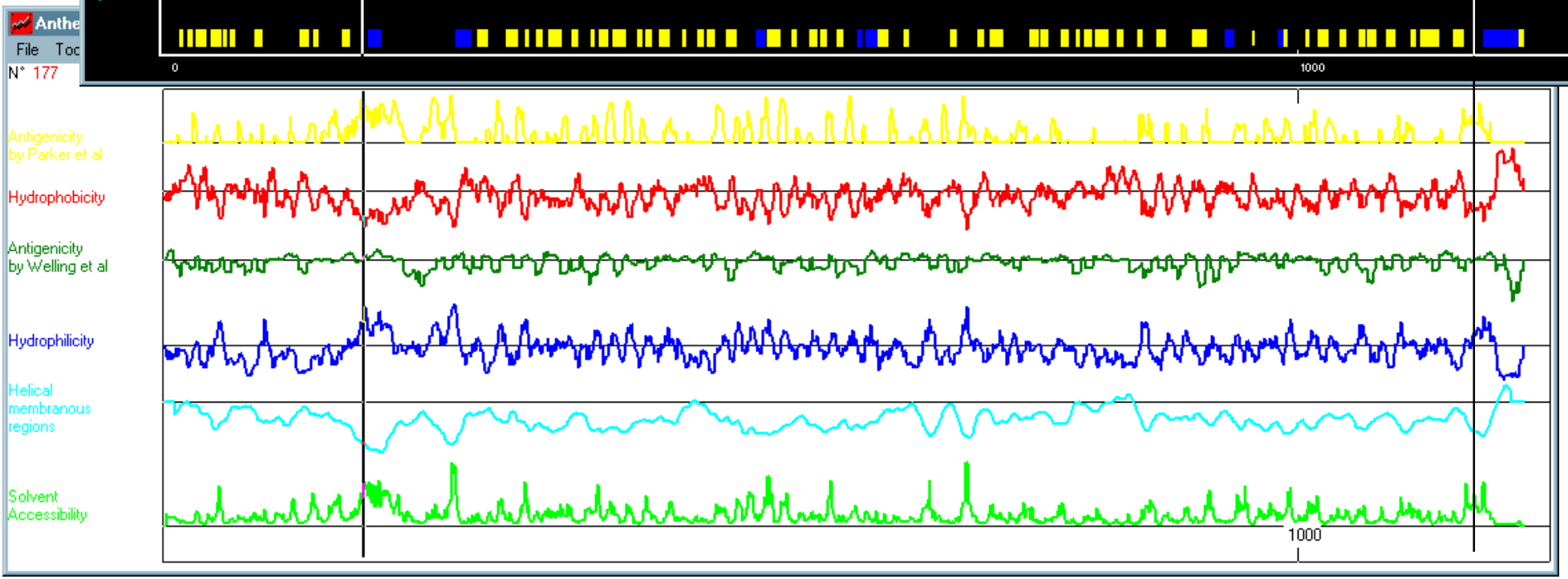
g.deleage@ibcp.fr

Anticorps obtenus par synthèse peptidique et immunisation

R 12 R

S 15 S





g.deleage@ibcp.fr

Profils de flexibilité 1/2

- ***Principe***
  - Mesure des facteurs de température B (agitation atomique)
- ***Paramètres***
  - 3 échelles de mobilité
    - Intermédiaire 20 valeurs de F
    - Flexibilité =>AA avec F>1 sur échelle I
    - Rigidité =>AA avec F<1 sur échelle I
- ***Algorithme:***
  - **1)** CHOIX DE L'ECHELLE E(i)
    - Examen des 2 voisins (i -1 et i+1) sur échelle intermédiaire :

i-1	i	i+1		i-1	i	i+1	sinon
M		M		R		R	
	M(i)				R(i)		I(i)

    - Exemple

	A	-	V	-	L	-	I	-	S	-	T	-	D	-	E	-	L
	R		R		R		R		M		M		M		M		R
E(i)	-		R		R		I		I		M		M		I		-

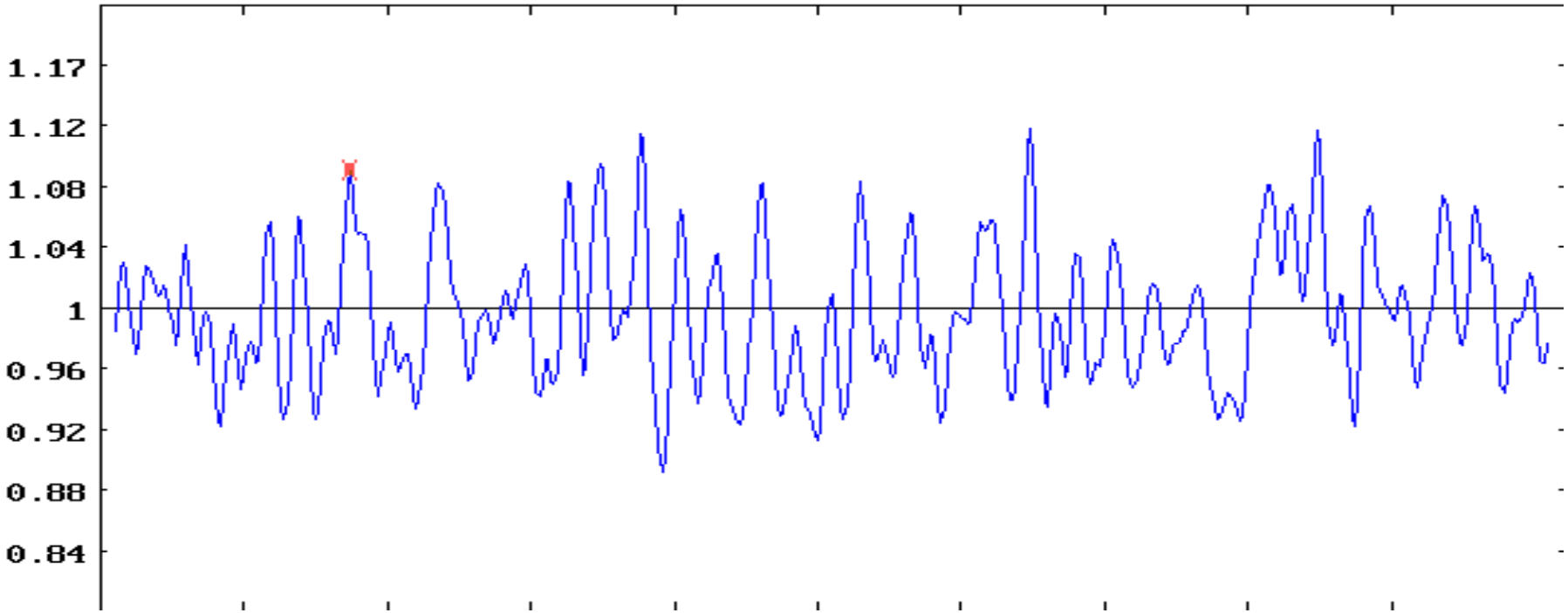
g.deleage@ibcp.fr

Profils de flexibilité 2/2

- 2) MOYENNE PONDEREE SUR 7 AA DES E(i)

$$F(s) = \frac{1}{4}\left[\frac{R(V)}{4} + \frac{R(L)}{2} + 3\frac{I(I)}{4} + I(S) + 3\frac{M(T)}{4} + \frac{M(D)}{2} + \frac{I(E)}{4}\right]$$

- Karplus et Schulz (1985) Naturewissenschaften 72, 212



g.deleage@ibcp.fr

Applications au clonage in silico

- **Déterminer la séquence nucléotidique des KARAP**
  - KARAP : Protéine impliquée dans le contrôle de l'activation des lymphocytes NK après intervention du CMH de classe 1
- **Portrait robot**
  - Protéine membranaire
  - Protéine d'environ 12 kDa
  - Un acide aminé chargé dans la région trans-membranaire (R,K,D,E)
  - Une cystéine dans la partie intra-cytoplasmique
  - Un motif ITAM dans la partie intra-cytoplasmique

Y-x(2)-[LI]-x(7,8)-Y-x(2)-[LI]

g.deleage@ibcp.fr

Clonage in silico

- **Utilisation de la base de données dbEST** (1 266 608)
- **Traduction de toutes les EST en 6 phases de lecture** (16 277 716)

Sélection des protéines comportant de 70 à 120 AA (1 500 007)

Sélection des protéines possédant un site ITAM (4 772)

- **Protéines comportant une zone trans-membranaire** (1 837)
- **Sélection des protéines possédant :** (95)
  - un acide aminé chargé dans la région trans-membranaire
  - une cystéine dans la région extra-cytoplasmique
- **Analyse qualitative, peptide signal**

11 candidats … dont le bon gène

g.deleage@ibcp.fr

Prédiction de Structures secondaires

g.deleage@ibcp.fr

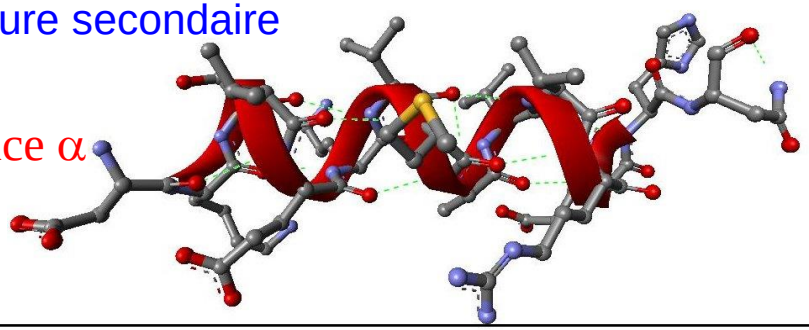
Organisation hiérarchique des protéines

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

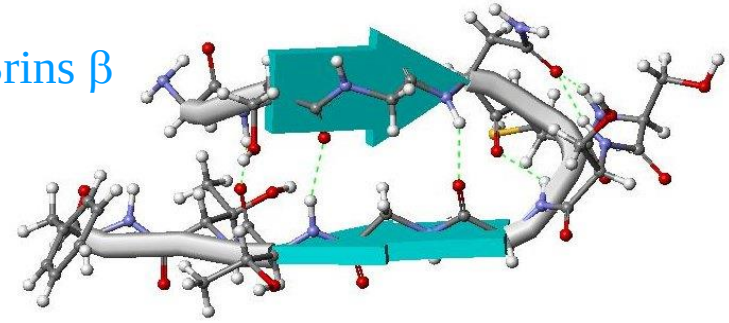
MKLD**E**IARLAGVSRTTASYVINGKAKQYRVSDKTVEKVMAVVREHNYHPNAVAAGLRAGR

Structure secondaire

Hélice α



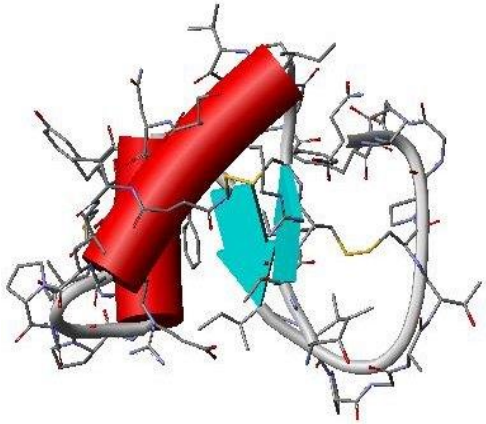
Brins β



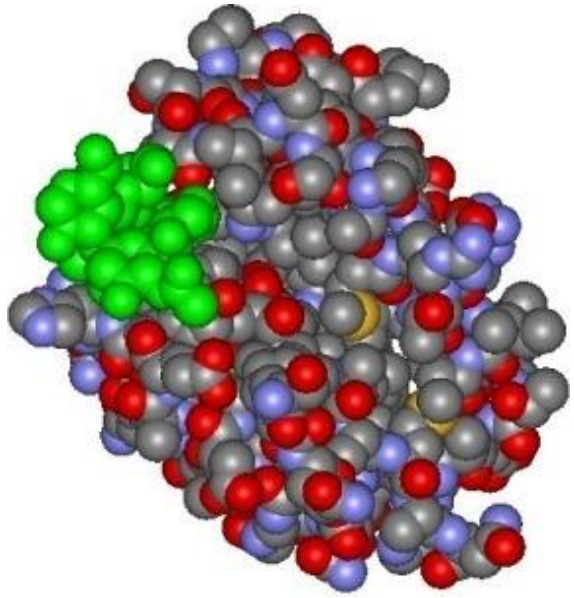
Structure secondaire = mot alphabet de 3 à 10 lettres

CCHHHHHHHHHHHCCCEEEETTTTEEEEECCCCHHHHHHHHHHHCCHHHHHHHHHGCCCC

Structure tertiaire = objet 3D



Fonction

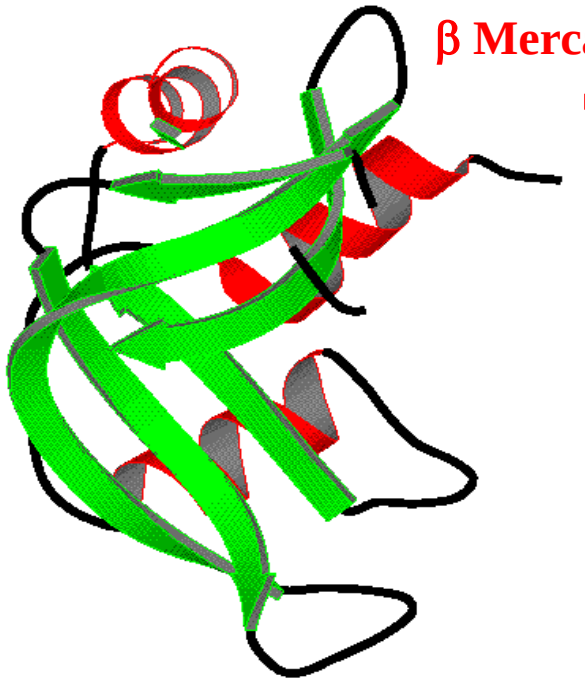


g.deleage@ibcp.fr

Expérience de dénaturation-renaturation (Anfinsen)
Paradoxe de Levinthal

Ribonucléase native, active, thermodynamiquement stable

+ UREE et
β Mercaptoethanol



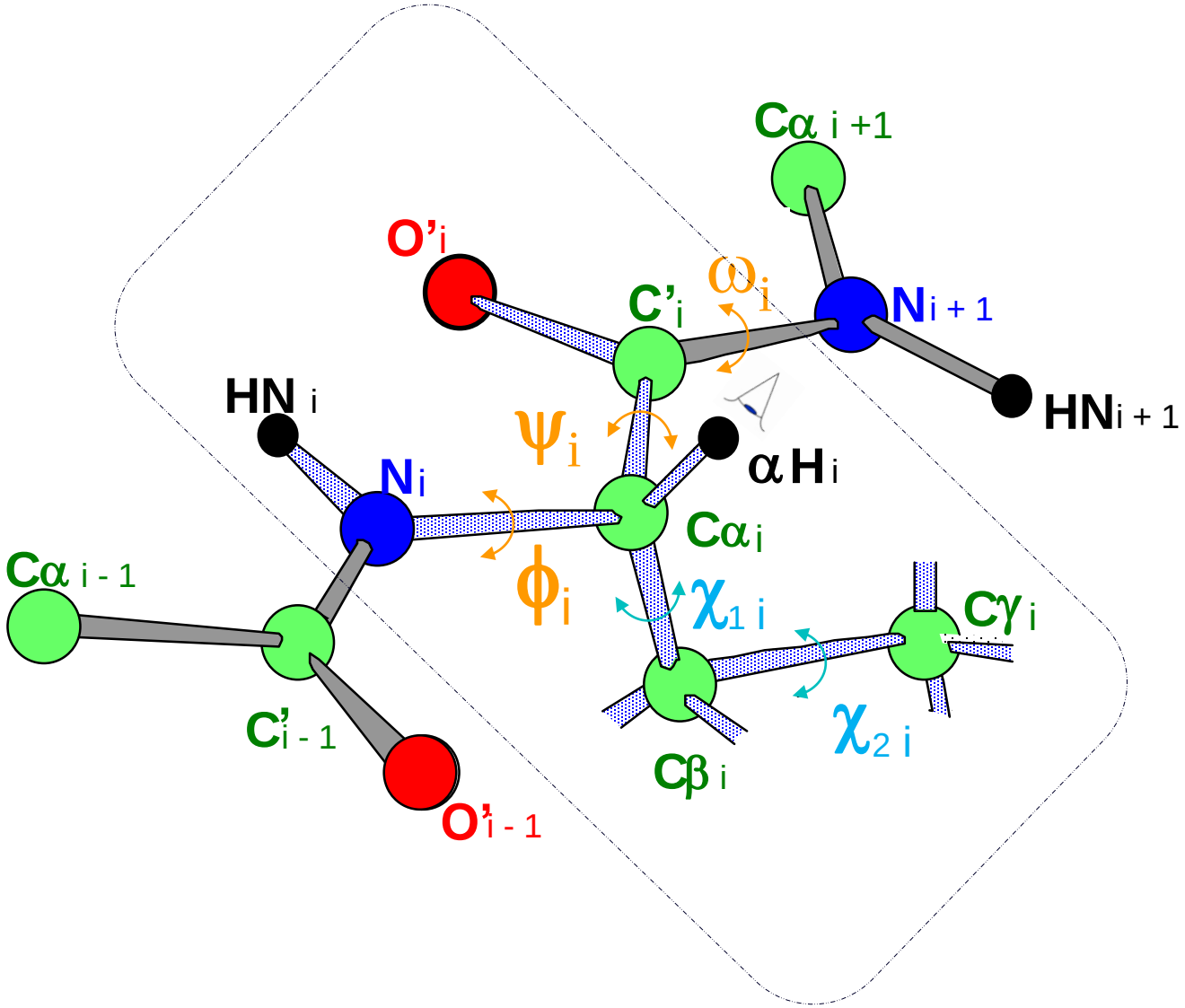
Quelques ms à 48h
- UREE et
β Mercaptoethanol

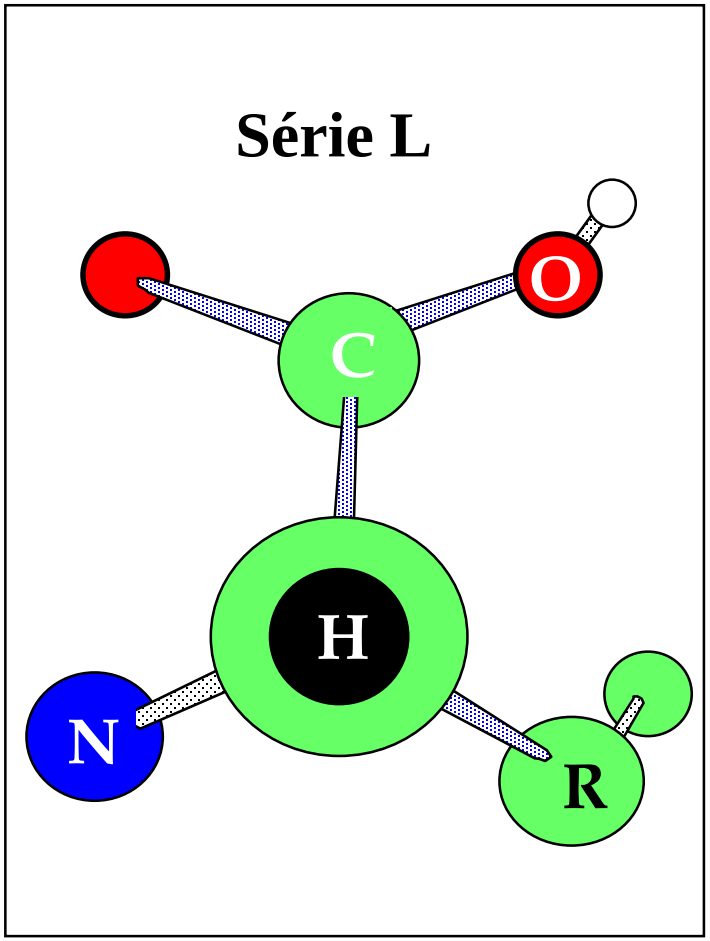
Ribonucléase, renaturée active, thermodynamiquement stable

Structure instable
Plus d'activité
Information de séquence
Attention exceptions

g.deleage@ibcp.fr

Géométrie du squelette protéique
Degrés de liberté dans la chaîne protéique

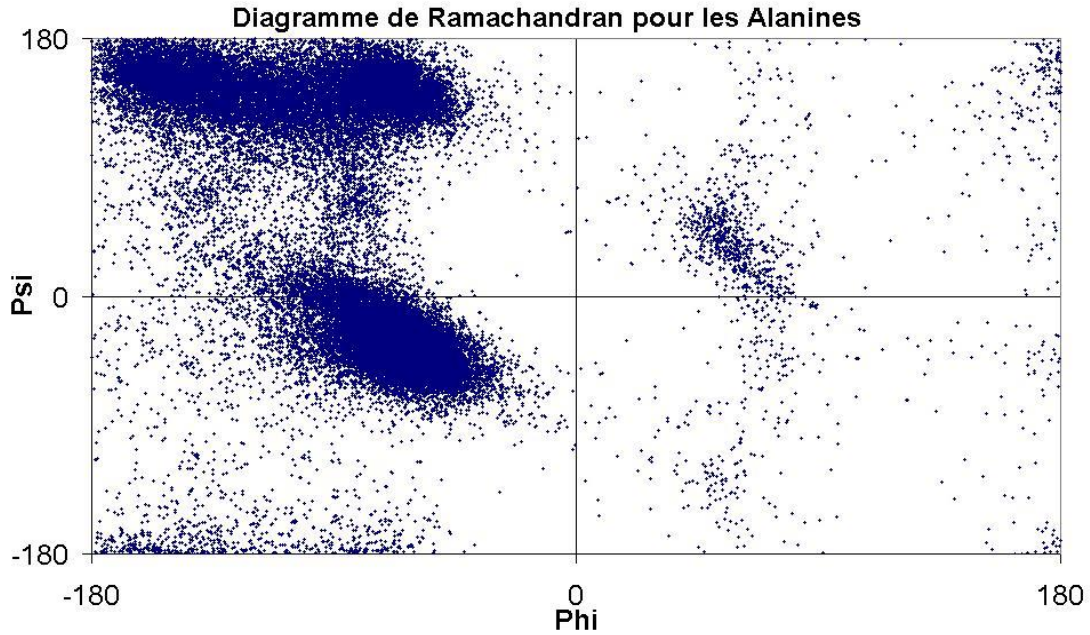


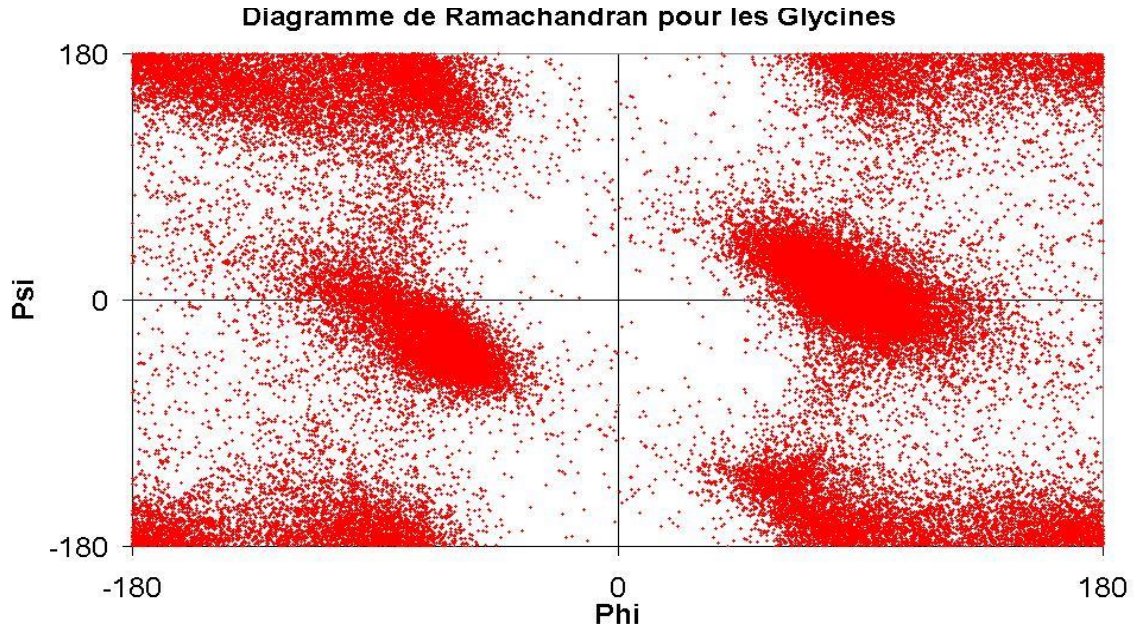


Lire « **C O R N** »

g.deleage@ibcp.fr

Diagramme de Ramachandran

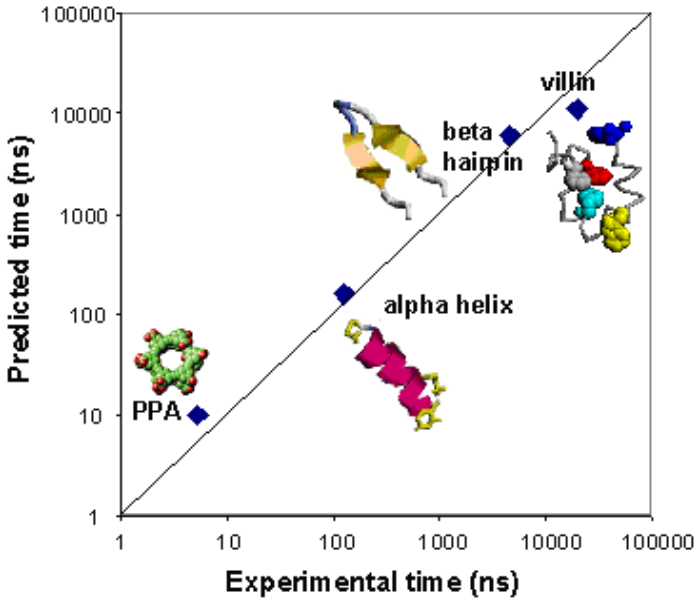


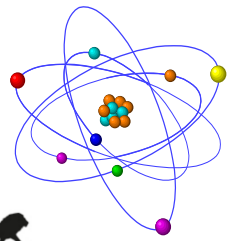


g.deleage@ibcp.fr

Loi des nombres

- Soit une petite protéine (100 aa)
- 20^{100} séquences théoriquement possibles
- Supposons 10 conformations par aa (10^{100} conformations)
- même si 2 conformations par aa
  - hélice ou non hélice
  - 2^{100} 10^{30} conformations
- Durée de vie d'une conformation 0,1 ps
  - ceci donne 10^{17}s pour que la protéine se replie soit 3 milliards d'années
  - et en admettant le calcul de l'énergie de 10^{10} molécules/seconde il faudrait 10^{20} s pour simuler par minimisation d'énergie soit 30 milliards de siècles!

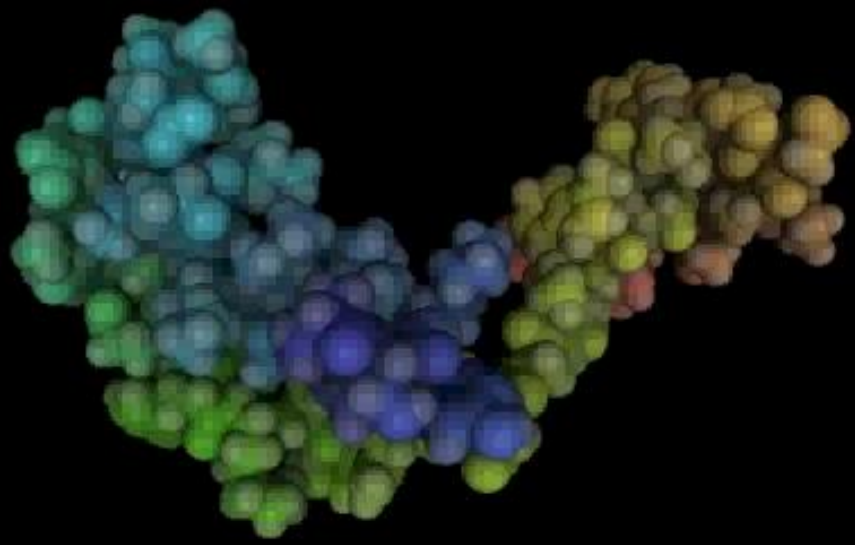




Nombre d'atomes dans l'univers $\approx 10^{80}$...

g.deleage@ibcp.fr

Repliement d’une protéine

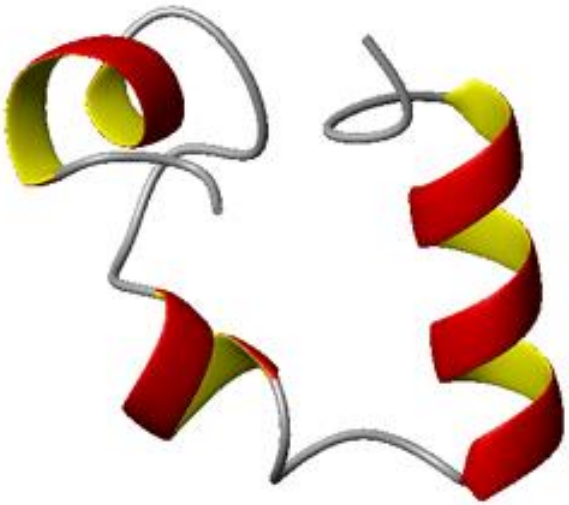


SimbiosMovies

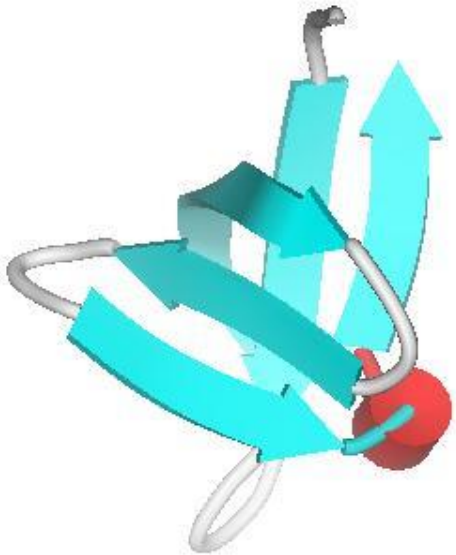
Simulation de dénaturation-renaturation

Projet folding@home de V. Pande

http://folding.stanford.edu/



Simulation du repliement de La villine



Formation d'une hélice

HIV intégrase

Simulation de la dénaturation de HIV intégrase

g.deleage@ibcp.fr

Prédiction de structures de protéines

- Méthodes statistiques
  - Chou/Fasman
  - GOR I, II, III Information directionnelle
  - Double Prediction Method
  - Combinaison statistiques (Discrimination linéaire DSC)
- Méthodes utilisant la similarité et optimisées
  - Plus proches voisins (simple)
  - Self Optimised Prediction Method (auto-optimisée)
  - Self Optimised Prediction Method from Alignments (+alignements)
- Méthodes neuronales (2 réseaux en tandem)
  - Méthode HNN Hierarchical Neural Network
  - Réseaux de neurones (PHD) avec alignements
- Méthode combinée
  - Multivariate Linear Regression Combination
- Disponibilité sur le Web
  - NPS@ (Network Protein Sequence @nalysis http://npsa-pbil.ibcp.fr)
- Utilisation en modélisation moléculaire
  - A faible taux d'identité pour l'identification d'empreinte potentielle.

g.deleage@ibcp.fr

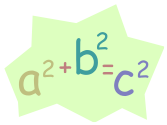
Méthode de Chou & Fasman (1978)

- Prédictions des hélices α
- Prédictions des feuillets β
- Elimination des zones chevauchantes
- Prédiction des coudes β

g.deleage@ibcp.fr

Nombre d'occurences (29 protéines,1974)

	N	Alpha	Beta	Apério.	Coudes
Ala	434	234	71	129	85
Arg+	142	53	26	63	40
Asn	230	40	58	132	106
Asp-	273	105	29	139	118
Cys	94	25	22	47	33
Gln	162	68	35	59	47
Glu-	234	134	17	83	51
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
Trp	78	32	21	25	22
Tyr	184	48	53	83	62
Val	357	144	119	94	53
Total	4741	1798	930	2013	1400
Freq		0.38	0.20	0.42	0.30



$$P\alpha_{Ala} = \frac{\frac{234}{434}}{\frac{1798}{4741}} = 1.42$$

g.deleage@ibcp.fr

Paramètres conformationnels (29 protéines,1974)

	P α	Hélice α		P β	Feuillet β
Glu	1.51	Hα	Val	1.70	Hβ
Met	1.45	Hα	Ile	1.60	Hβ
Ala	1.42	Hα	Tyr	1.47	Hβ
Leu	1.21	Hα	Phe	1.38	hβ
Lys+	1.16	hα	Trp	1.37	hβ
Phe	1.13	hα	Leu	1.30	hβ
Gln	1.11	hα	Cys	1.19	hβ
Trp	1.08	hα	Thr	1.19	hβ
Ile	1.08	hα	Gln	1.10	hβ
Val	1.06	hα	Met	1.05	hβ
Asp-	1.01	Iα	Arg+	0.93	iβ
His+	1.00	Iα	Asn	0.89	iβ
Arg+	0.98	iα	His+	0.87	iβ
Thr	0.83	iα	Ala	0.83	iβ
Ser	0.77	iα	Ser	0.75	bβ
Cys	0.70	iα	Gly	0.75	bβ
Tyr	0.69	bα	Lys+	0.74	bβ
Asn	0.67	bα	Pro	0.55	Bβ
Pro	0.57	Bα	Asp-	0.54	Bβ
Gly	0.57	Bα	Glu-	0.37	Bβ

g.deleage@ibcp.fr

Méthode de Chou & Fasman (1978)
Prédictions des hélices α

- **4 conditions et 1 règle**
  - a) Formation de l'hélice α 4aa/6hα ou Hα => noyau d'hélice α
  - b) Elongation des 2 côtés => tétrapeptides ayant P(α) < 1

    b4, b3i, b3h, b2i2, b2ih, b2h2, bi3, bi2h, bih2
  - c) L'hélice α complète >=1/2 de H, h

    < 1/3 de B,b

- **Tout segment de 6 aa (ou plus) avec un P(α) >=1.03 et P(α)>P(β) satisfaisant les conditions est prédit comme une hélice α**

g.deleage@ibcp.fr

Méthode de Chou & Fasman (1978)
Prédictions des zones en brins β

- **4 conditions et 1 règle**
  - a) Formation du brin β 3aa/5hβ ou Hβ => noyau de brin β
  - b) Elongation des 2 côtés => tétrapeptides ayant P(β) < 1

    b4, b3i, b3h, b2i2, b2ih, b2h2, bi3, bi2h, bih2
  - c) Le brin β complet >=1/2 de H, h

    < 1/3 de B,b
  - d) Glu et Pro rares dans les brins

- **Tout segment de 5 aa (ou plus) avec un P(β) >=1,05 et P(β)>P(α) satisfaisant les conditions est prédit comme un brin β**

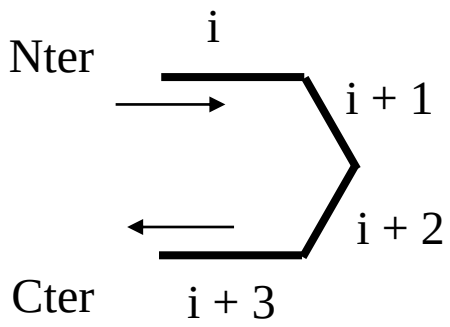
g.deleage@ibcp.fr

Méthode de Chou & Fasman (1978)
Elimination des zones chevauchantes

- **4 étapes**
  - a) Calculer les P(β) et les P(α) sur la zone chevauchante :
    - si P(β) < P(α) alors Hélice α
    - si P(β) > P(α) alors Feuillet β
  - b) Comparer les préférences conformationelles (H, h, I, i, b, B)
  - c) Faire une analyse des limites (hélices et feuillets)
  - d) Autres critères subjectifs

g.deleage@ibcp.fr

Méthode de Chou & Fasman (1978)
Prédictions des coudes β



- **3 conditions:**
  - a) Examiner la zone par tétrapeptide
    - P(β) <P(t)> P(α)
  - b) P(t)> 1.00
  - c) Fi x F(i+1) x F(i+2) x F(i+3)>= 0.55 10^{-4}

g.deleage@ibcp.fr

Bilan de la Méthode de Chou & Fasman (1978)

- Avantages
  - La première méthode utilisable par les biologistes
  - Méthode manuelle
- Inconvénients
  - Non reproductible
  - Difficile à implémenter
  - Qualité faible (52%)

g.deleage@ibcp.fr

Méthode de l'information directionnelle (Garnier et al., 1978)

- **Principe de base**

  Chaque aa possède une influence sur la conformation de tous les autres aa

- **Mesure statistique et Théorie de l'Information**

$$I(k, i) = Ln \frac{p(k / i)}{p(k)}$$

i : numéro de l'aa
k : numéro de la conformation
p(k/i) : probabilité d'avoir l'état k sachant que l'aa est i
p(k) : probabilité d'avoir l'état k

- **Influence mesurable => I(k,i)≠ 0 de i - 8 à i + 8**
- **20 x 17 x 4 = 1360 paramètres**

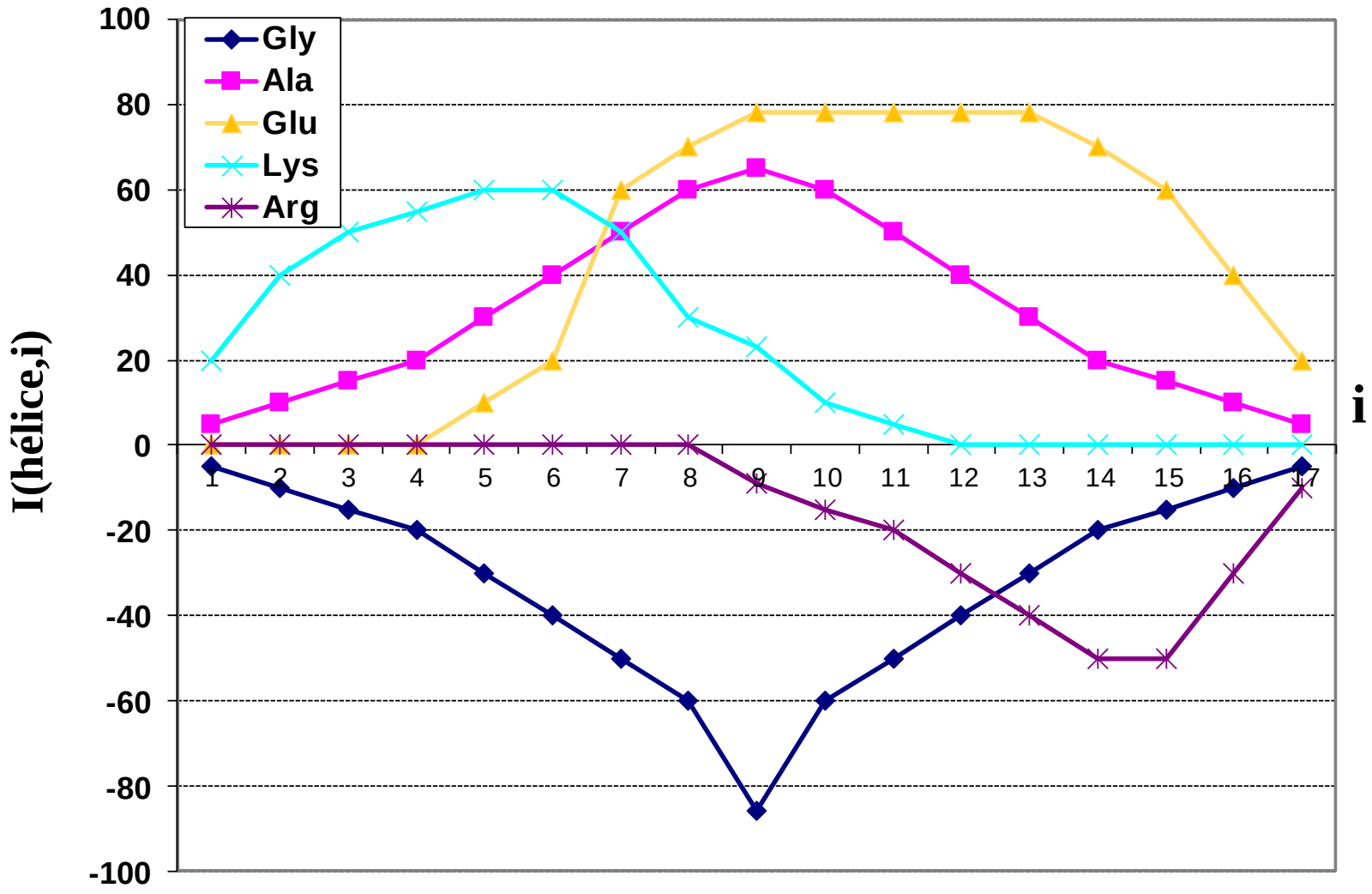
g.deleage@ibcp.fr

Paramètres conformationnels de l'hélice (GOR)

Gly	-5	-10	-15	-20	-30	-40	-50	-60	-86	-60	-50	-40	-30	-20	-15	-10	-5
Ala	5	10	15	20	30	40	50	60	65	60	50	40	30	20	15	10	5
Val	0	0	0	0	0	0	5	10	14	10	5	0	0	0	0	0	0
Leu	0	5	10	15	20	25	28	30	32	30	28	25	20	15	10	5	0
Ile	5	10	15	20	25	20	15	10	6	0	-10	-15	-20	-25	-20	-10	-5
Ser	0	-5	-10	-15	-20	-25	-30	-35	-39	-35	-30	-25	-20	-15	-10	-5	0
Thr	0	0	0	-5	-10	-15	-20	-25	-26	-25	-20	-15	-10	-5	0	0	0
Asp	0	-5	-10	-15	-20	-15	-10	0	5	10	15	20	20	20	15	10	5
Glu	0	0	0	0	10	20	60	70	78	78	78	78	78	70	60	40	20
Asn	0	0	0	0	-10	-20	-30	-40	-51	-40	-30	-20	-10	0	0	0	0
Gln	0	0	0	0	5	10	20	20	10	-10	-20	-20	-10	-5	0	0	0
Lys	20	40	50	55	60	60	50	30	23	10	5	0	0	0	0	0	0
His	10	20	30	40	50	50	50	30	12	-20	-10	0	0	0	0	0	0
Arg	0	0	0	0	0	0	0	0	-9	-15	-20	-30	-40	-50	-50	-30	-10
Phe	0	0	0	0	0	5	10	15	16	15	10	5	0	0	0	0	0
Tyr	-5	-10	-15	-20	-25	-30	-35	-40	-45	-40	-35	-30	-25	-20	-15	-10	-5
Trp	-10	-20	-40	-50	-50	-10	0	10	12	10	0	-10	-30	-50	-40	-20	-10
Cys	0	0	0	0	0	0	-5	-10	-13	-10	-5	0	0	0	0	0	0
Met	10	20	25	30	35	40	45	50	53	50	45	40	35	30	25	20	10
Pro	-10	-20	-40	-60	-80	-100	-120	-140	-77	-60	-30	-20	-10	0	0	0	0

g.deleage@ibcp.fr

Profils conformationnels de l'hélice (GOR)



g.deleage@ibcp.fr

Méthode de l'information directionnelle (Garnier et al., 1978)

- Algorithme prédictif en 2 tours

$$\mathbf{Info\ (k,\ \ i)\ =\ \sum_{j=-8}^{j=+8} I\ (k,\ \ i+\ \ j)}$$

- Info(k,i) = Information directionnelle que possède l 'acide aminé i pour la conformation k

g.deleage@ibcp.fr

Prédiction de l'Arg 9

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	L	I	D	G	E	M	L	A	R	Y	Q	N	F	K	R	L	I
	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
	0	10	-10	-20	10	40	28	60	-9	-40	-20	-20	0	0	-50	5	-5
Gly	-5	-10	-15	-20	-30	-40	-50	-60	-86	-60	-50	-40	-30	-20	-15	-10	-5
Ala	5	10	15	20	30	40	50	60	65	60	50	40	30	20	15	10	5
Val	0	0	0	0	0	0	5	10	14	10	5	0	0	0	0	0	0
Leu	0	5	10	15	20	25	28	30	32	30	28	25	20	15	10	5	0
Ile	5	10	15	20	25	20	15	10	6	0	-10	-15	-20	-25	-20	-10	-5
Ser	0	-5	-10	-15	-20	-25	-30	-35	-39	-35	-30	-25	-20	-15	-10	-5	0
Thr	0	0	0	-5	-10	-15	-20	-25	-26	-25	-20	-15	-10	-5	0	0	0
Asp	0	-5	-10	-15	-20	-15	-10	0	5	10	15	20	20	20	15	10	5
Glu	0	0	0	0	10	20	60	70	78	78	78	78	78	70	60	40	20
Asn	0	0	0	0	-10	-20	-30	-40	-51	-40	-30	-20	-10	0	0	0	0
Gln	0	0	0	0	5	10	20	20	10	-10	-20	-20	-10	-5	0	0	0
Lys	20	40	50	55	60	60	50	30	23	10	5	0	0	0	0	0	0
His	10	20	30	40	50	50	50	30	12	-20	-10	0	0	0	0	0	0
Arg	0	0	0	0	0	0	0	0	-9	-15	-20	-30	-40	-50	-50	-30	-10
Phe	0	0	0	0	0	5	10	15	16	15	10	5	0	0	0	0	0
Tyr	-5	-10	-15	-20	-25	-30	-35	-40	-45	-40	-35	-30	-25	-20	-15	-10	-5
Trp	-10	-20	-40	-50	-50	-10	0	10	12	10	0	-10	-30	-50	-40	-20	-10
Cys	0	0	0	0	0	0	-5	-10	-13	-10	-5	0	0	0	0	0	0
Met	10	20	25	30	35	40	45	50	53	50	45	40	35	30	25	20	10
Pro	-10	-20	-40	-60	-80	-100	-120	-140	-77	-60	-30	-20	-10	0	0	0	0

g.deleage@ibcp.fr

Décision pour Arg 9 (Garnier et al., 1978)

- **Résultat avec toutes les constantes nulles (1er tour):**
  - Si Info (Hélice, R9) = - 21
  - si Info (Brin, R9) = - 256
  - si Info (Coude, R9) = -115
  - et si Info (Apériod., R9) = - 85

L'état prédit pour Arg 9 est Hélice

g.deleage@ibcp.fr

Constantes de décisions pourquoi?

Protéines avec un contenu en SS>50% au premier tour
Sous prédiction des SS=>
Il faut augmenter le % en SS => DC(hélice) = -100, DC(feuillet) = - 88

Les méthodes statistiques ont tendance à moyenner

si % 20 < (hélice+feuillet) < 50 alors
DC(hélice) = - 75, DC(Feuillet) = - 88

Protéines avec un contenu en SS<20% au premier tour
Sur prédiction des SS=>
Il faut diminuer le % en SS => DC(hélice) =158, DC(Feuillet) = 50

g.deleage@ibcp.fr

Méthode de l'information directionnelle (Garnier et al., 1978)

- Algorithme prédictif en 2 tours

$$\mathbf{Info\ (k,\ \ i)\ =\ \sum_{j=-8}^{j=+8} I\ (k,\ \ i+\ \ j)\ -\ DC\ (k)}$$

- Info(k,i) = Information directionnelle que possède l 'acide aminé i pour la conformation k
- DC(k) = Contante de décision et DC (Coude) et DC (Apériodique) =0
  - **Premier tour**
    - **DC (hélice) = 0 et DC (brin) = 0**
  - **Deuxième tour**
    - **si % (hélice+feuillet) <20 alors DC(hélice) =158, DC(Feuillet) = 50**
    - **si % 20 < (hélice+feuillet) < 50 alors DC(hélice) = - 75, DC(Feuillet) = - 88**
    - **si % (hélice+feuillet) >50 alors DC(hélice) = -100, DC(feuillet) = - 88**

g.deleage@ibcp.fr

Constante décision pour Arg 9 (Garnier et al., 1978)

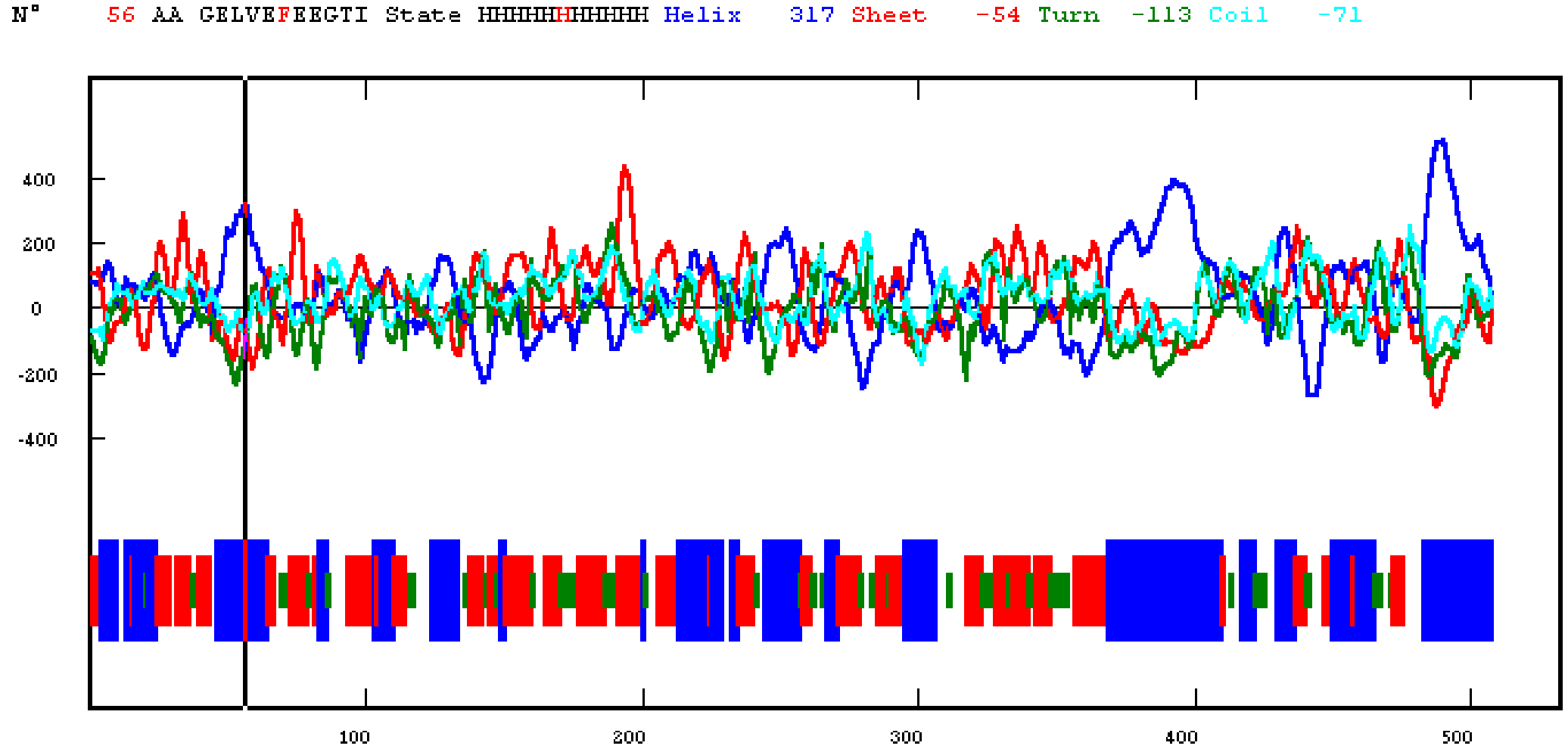
- **Résultat avec toutes les constantes nulles (1er tour):**
  - Si Info (Hélice, R9) = - 21
  - si Info (Brin, R9) = - 256
  - si Info (Coude, R9) = -115
  - et si Info (Apériod., R9) = - 85

  L'état prédit pour Arg 9 est Hélice
- **Résultat si DCH= 158, DCB=50 , DCT et DCC=0**
  - Si Info (Hélice, R9) = - 179
  - si Info (Brin, R9) = - 306
  - si Info (Coude, R9) = - 115
  - et si Info (Apériod., R9) = - 85

  L'état prédit pour Arg 9 est Coil

g.deleage@ibcp.fr

Profils conformationnels (GOR)



g.deleage@ibcp.fr

Bilan de la Méthode de Garnier et al., (1978)

- Avantages
  - Méthode automatique non ambiguë
  - Théorie de l'information
  - Rapide (instantanée)
  - Méthode insensible à l'homologie
  - Prise en compte de l'environnement séquentiel
- Inconvénients
  - Qualité moyenne (56%)
  - Faible évolutivité

g.deleage@ibcp.fr

Bilan de la Méthode de Garnier et al., (1978)

- Avantages
  - Méthode automatique non ambiguë
  - Théorie de l'information
  - Rapide (instantanée)
  - Méthode insensible à l'homologie
  - Prise en compte de l'environnement séquentiel
- Inconvénients
  - Qualité moyenne (56%)
  - Faible évolutivité

g.deleage@ibcp.fr

Méthode des plus proches voisins (Levin et al., 1986)

- **Principe**

  De courtes séquences similaires ont tendance à adopter des structures secondaires identiques

- **Algorithme** 

Chaque heptapeptide (1-7, 2-8, . . ., n - 7 à n de la protéine "à prédire" est comparé avec tous les heptapeptides de chaque protéine d'une base de données de référence :

Prot 1 (1-7 , 2-8, . . . , n_1 -7 à n_1)

Prot 2 (1-7 , 2-8, . . . , n_2 -7 à n_2)

.

Prot i (1-7 , 2-8, . . . , n_i -7 à n_i)

.

Prot n (1-7 , 2-8, . . . , n -7 à n)

g.deleage@ibcp.fr

Base de référence (126 protéines Rost & Sander, 1993)

Fichier Edition Format Affichage ?

```
>1ACX-1ANTIBACTERIALPROTEIN17-DE
APAFSVSPASGASDGQSVSVSVAAAGETYYIAQCAPVGGQDACNPATATSFTTDASGAASFSTVRKSYAGQTPSGTPVGS
VDCATDACNLGAGNSGLNLGHVALTF*
CCEEEEECCCCCCCCCCEEEEEEECCCCEEEEEEECEETTEECCCTTTCCEEECCCCCCCEEEECCCEEEEECTTCCEEEE
EETTTCCCEEEEECCCCCCCCCCCCC*
>1AK3-ATRANSFERASE(PHOSPHOTRANSFERASE)17-JA
RLLRAIMGAPGSGKGTVSSRITKHFELKHLSSGDLLRDNMLRGTEIGVLAKTFIDQGKLIPDDVMTRLVLHELKNLTQYN
WLLDGFPRTLPQAEALDRAYQIDTVINLNVPFEVIKQRLTARWIHPGSGRVYNIEFNPPKTMGIDDLTGEPLVQREDDRP
ETVVKRLKAYEAQTEPVLEYYRKKGVLETFSGTETNKIWPHVYAFLQTKLPQRS*
CCEEEEECCTTCCHHHHHHHHHHCCCEEEEHHHHHHHHHTTTCHHHHHHHHHHHTTCCCCHHHHHHHHHHHHHTTCCCC
EEEECCCCCHHHHHHHTTCCCCEEEEEECCHHHHHHHHTCEEEETTTTEEEETTTCCCCCTTCCTTTCCCCCCCTTCCH
HHHHHHHHHHHHHHHHHHHHHHHHTCEEEEECCCHHHHHHHHHHHHHTTCCCCC*
>1AZU-1ELECTRONTRANSPORT(COPPERBINDING)04-AU
SVDIQGNDQMQFNTNAITVDKSCKQFTVNLSHPGNLPKNVMGHNWVLSTAADMQGVVTDGMASGLDKDYLKPDDSRVIAH
TKLIGSGEKDSVTFDVSKLKEGEQYMFFCTFPGHSALMKGTLTL*
CEEEECCCCCCCCCCCEECCCCCCEEEEEEECCCCCCTTTCCCCCEEEETTTTHHHHHHHHHHCHHHCCCCCCTTCCCC
CCCCCTTCCEEEEEECCCCCCCCCCEEEECCCTTTTTTCEEEEEC*
>1BBP-ABILINBINDING19-SE
NVYHDGACPEVKPVDNFDWSNYHGKWWEVAKYPNSVEKYGKCGWAEYTPEGKSVKVSNYHVIHGKEYFIEGTAYPVGDSX
XXKIGKIYHKLTYGGVTKENVFNVLSTDNKNYIIGYYCKYDEDKKGHQDFVWVLSRSKVLTGEAKTAVENYLIGSPVVXD
SQKLVYSDFSEAACKVN*
CEEEECCCCCCCCCCCCCHHHCCEEEEEEEECCCCTTTTCEEEEEEEEECCCCEEEEEEEETTEEEEEEEEEECCCTC
CCTCCEEEEEEEETTEEEEEEEEEECCCCEEEEEEEEETTTTEEEEEEEEECCCCCCHHHHHHHHHHHHCCCCCC
HHHCEECCCCHHHHCCC*
>1BDS-1ANTI-HYPERTENSIVEANTI-VIRALPROTEIN14-NO
AAPCFCSGKPGRGDLWILRGTCPGGYGYTSNCYKWPNICCYPH*
CCCCCCCCCCCCCCEEECCCCCCCTTTCCCCEEEETTEEEECCC*
>1BMV-1VIRUS09-OC
```


g.deleage@ibcp.fr

Matrice de substitution de Levin et al., 1986

G	2																			
P	1	3																		
D	0	0	2																	
E	0	-1	1	2																
A	0	-1	0	1	2															
N	0	0	1	0	0	3														
Q	0	0	0	1	0	1	2													
S	0	0	0	0	1	0	0	2												
T	0	0	0	0	0	0	0	0	2											
K	0	0	0	0	0	1	0	0	0	2										
R	0	0	0	0	0	0	0	0	0	1	2									
H	0	0	0	0	0	0	0	0	0	0	0	2								
V	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	2							
I	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	1	2						
M	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	0	0	2					
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2				
L	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	1	0	2	0	2			
F	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	0	1	0	-1	0	2		
Y	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	-1	0	1	2	
W	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	0	0	0	0	-1	0	0	0	2
	G	P	D	E	A	N	Q	S	T	K	R	H	V	I	M	C	L	F	Y	W

A - V - K - L - M - S - T

0 + 1 + 1 + 1 + 2 + 0 + 0 = 5

I - L - R - V - M - N - S

g.deleage@ibcp.fr

Méthode des plus proches voisins (Levin et al., 1986)

Exemple seuil de similarité =7

```
A - V - K - L - M - S - T
0 + 1 + 1 + 1 + 2 + 0 + 0  = 5 Score < 7
I - L - R - V - M - N - S

A - V - K - L - M - S - T
1 + 2 + 2 + 2 + 2 + 2 + 2 = 13 Score >7
E - V - K - L - M - S - T
H - H - H - H - C - C - E

A - V - K - L - M - S - T
1 + 1 + 1 + 2 + 2 + 0 + 2 = 9  Score >7
S - L - R - L - L - T - T
C - H - H - H - H - H - E
```

AA	Hélice	Etendu	Coil	
A	13	0	9	= H 13
V	13 + 9	0	0	= H 22
K	13 + 9	0	0	= H 22
L	13 + 9	0	0	= H 22
M	0 + 9	0	13	= C 13
S	0 + 9	0	13	= C 13
T	0	13 + 9	0	= E 22

g.deleage@ibcp.fr

Algorithme des plus proches voisins

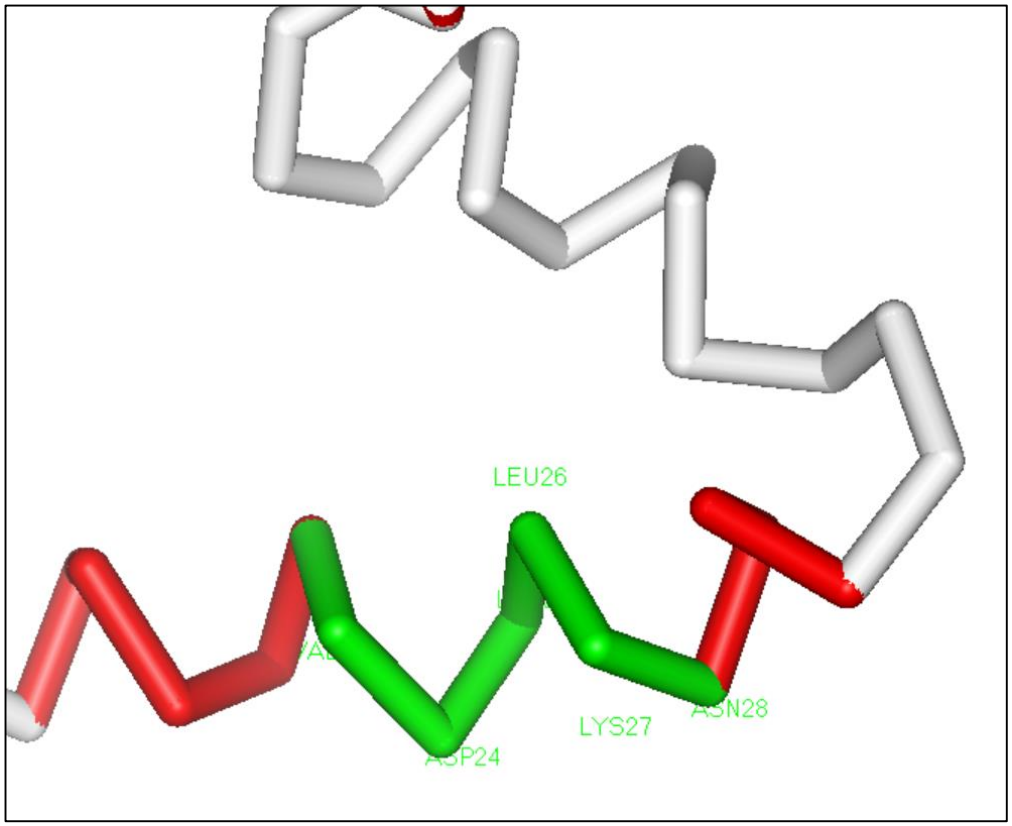

```
pour i=1 jqa M faire (500)              /* parcourir la sequence a predire*/
{
   pour l=1 jqa nombre_prot (2000)      /* pour toutes les protéines */
   {
        pour j=1 jqa N(l) faire (~500) /* parcourir la sequence de la banque*/
        {
                score=0
                pour k=1 jqa fenetre (17) faire /* pour chaque peptide*/
                {
                        Score = score + SUBS[Seq[i+k]][Seq[j+k,l]]
                }
                Si score >= seuil alors
                        pour k=1 jqa longueur_fenetre (17) faire
                        {
                                confo(i+k) = confo(i+k)+score
                        }
                }
        }
   }
}
```

~8,5 10^9 comparaisons pour prédire une séquence de 500 acides aminés

g.deleage@ibcp.fr

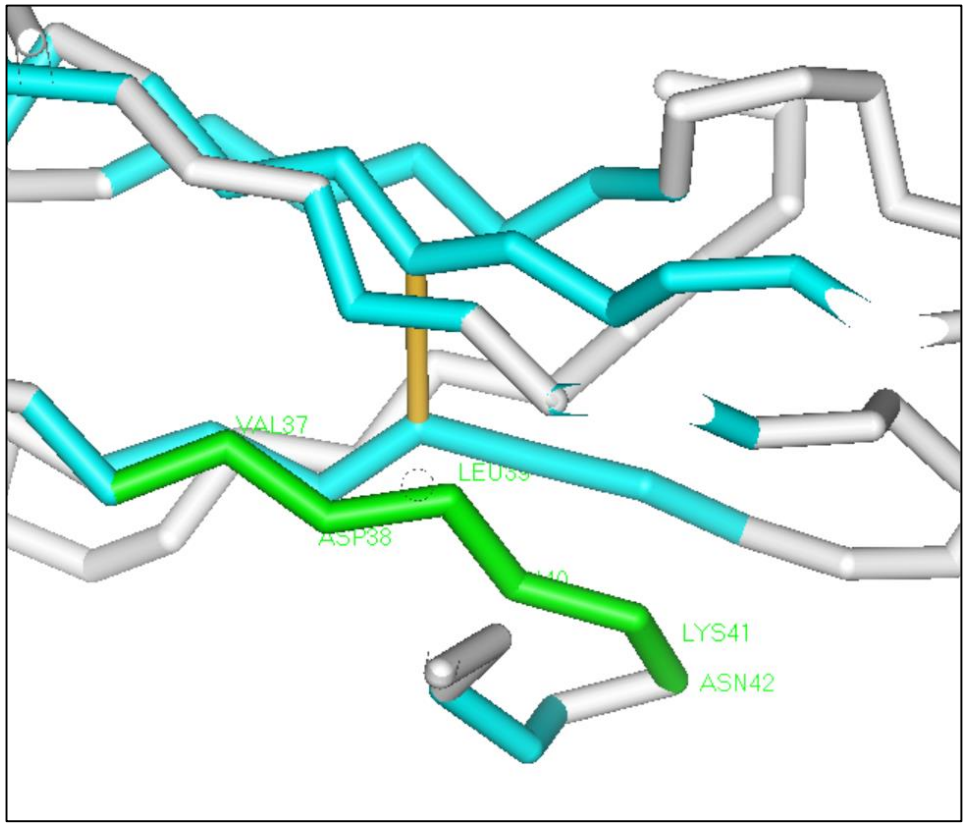
Séquences courtes avec structures 3D observées différentes (RCSB)

VDLLKN
23 28



pdb : 1CO0 A

VDLLKN
37 42



pdb : 1B0G A

g.deleage@

Bilan de la Méthode de Levin et al., (1986)

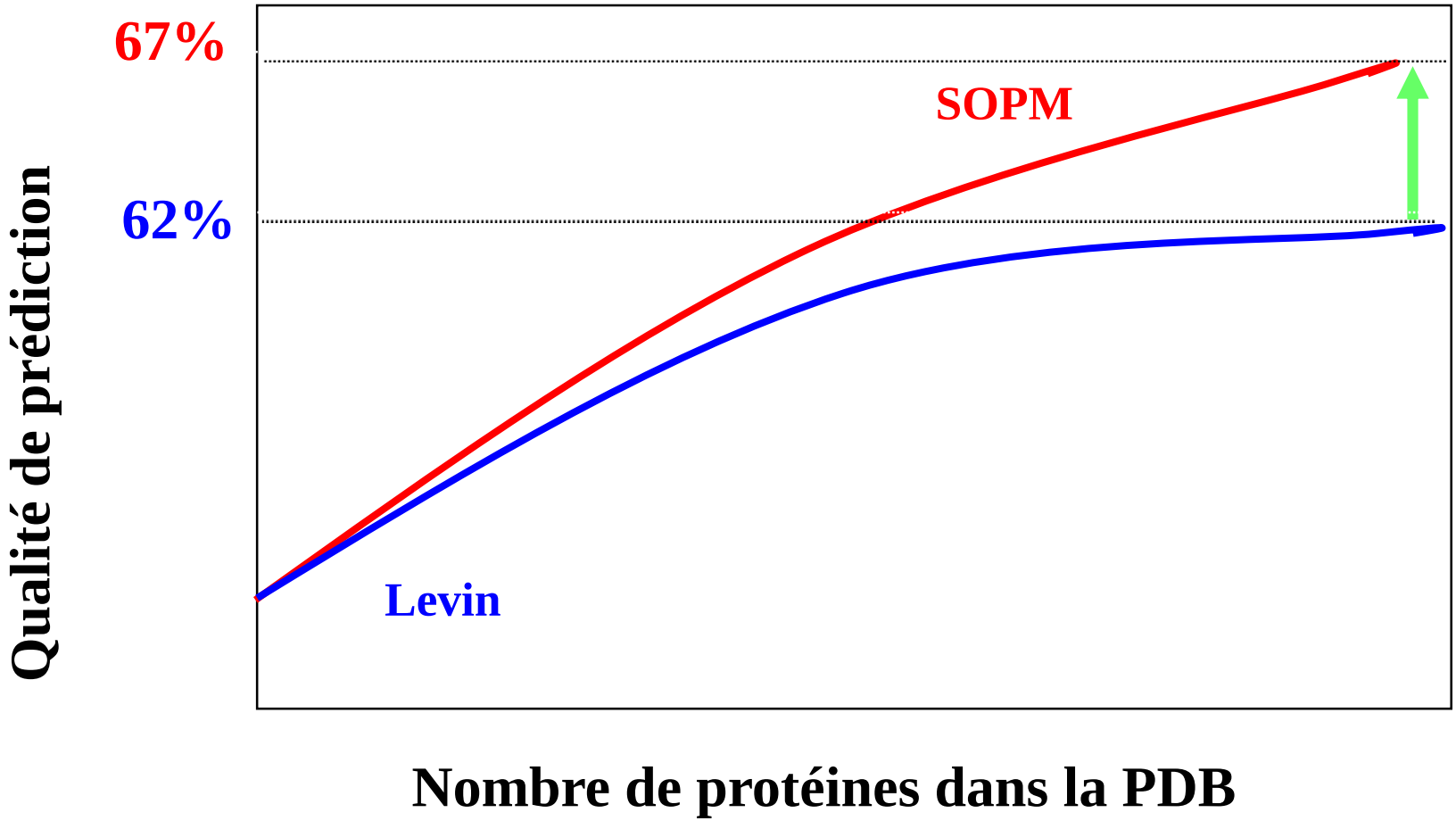
- Avantages
  - Méthode automatique non ambiguë
  - Bonne qualité de prédiction (62%)
  - La qualité augmente avec la taille de la base de données (tant qu'on apporte plus de signal que de bruit)
- Inconvénients
  - Méthode sensible à l'homologie
  - Temps de calcul assez long (mur des combinaisons)

g.deleage@ibcp.fr

Limitation de la qualité de prédiction => SOPM

- **Objectifs**

Améliorer de façon constante la qualité de prédiction en fonction de la taille de la base de données =>Meilleur rapport signal/bruit



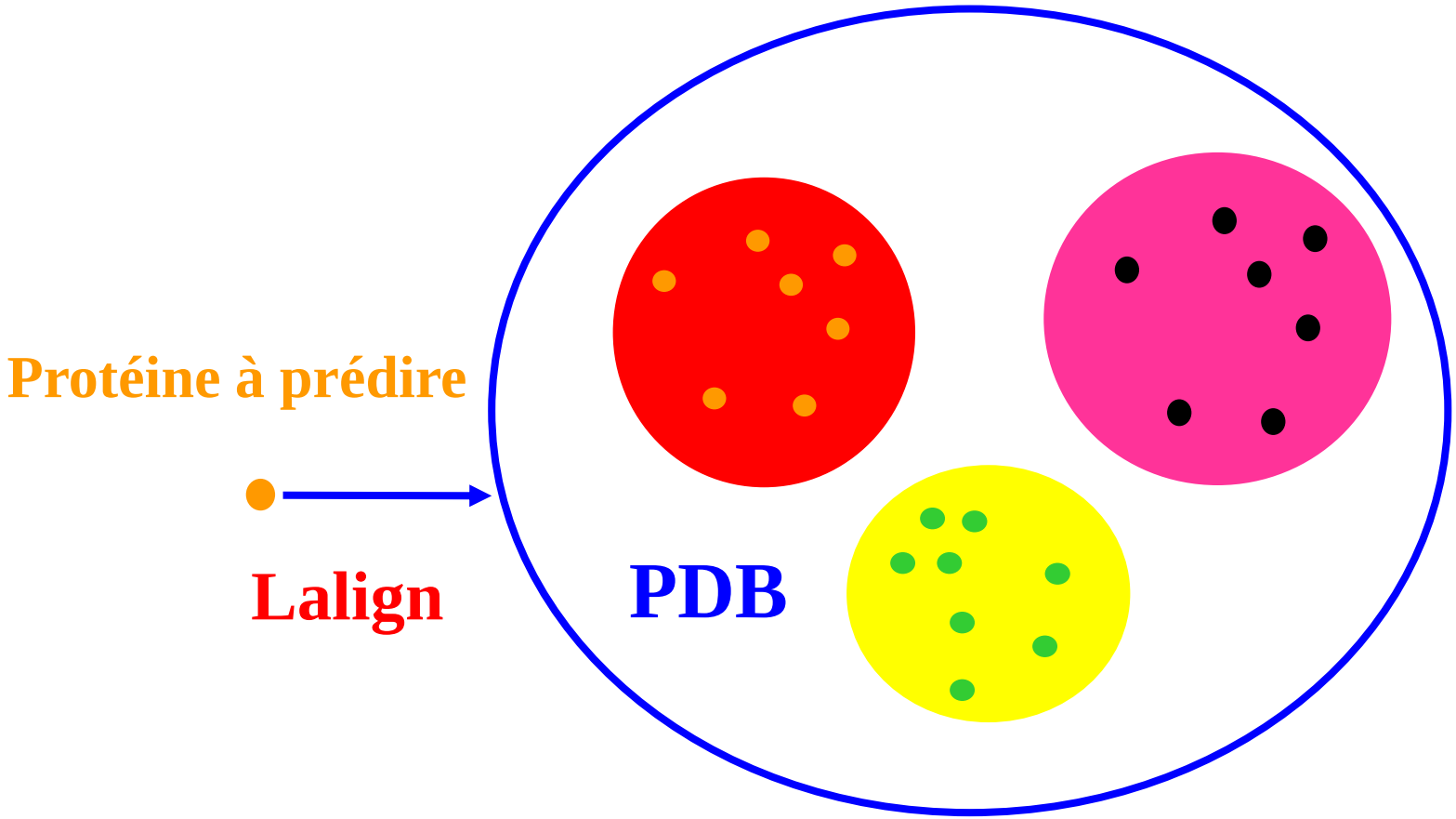
g.deleage@ibcp.fr

SOPM => Principe

- **Principe**

  Constituer à la volée un sous ensemble des protéines de la PDB les plus similaires et de même classe structurale que la protéine à prédire

- **Optimiser les paramètres prédictifs sur le sous-ensemble**



g.deleage@ibcp.fr

Méthode SOPM (1994)

Principe

Etablir une base de données appropriée pour optimiser les paramètres prédictifs.

g.deleage@ibcp.fr

Bilan de la Méthode SOPM (1994)

- Avantages
  - Méthode automatique non ambiguë
  - Bonne qualité de prédiction (69%)
  - La qualité augmente avec la taille de la base de données
- Inconvénients
  - Temps de calcul assez long
  - Méthode basée sur une seule séquence

g.deleage@ibcp.fr

Méthode SOPMA (avec Alignements) (1995-2000)

- Principe
  - Utiliser des familles fonctionnelles de protéines

Séquence à prédire

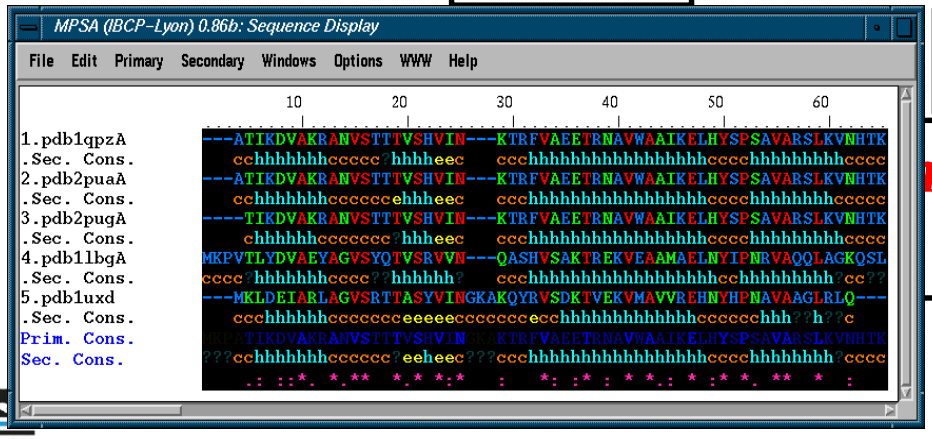
FASTA — **SWISSPROT**

Protéines homologues

CLUSTALW

Séquences alignées

SOPM — **Séquences PDB**



Structures alignées

Prédiction finale

g.deleage@ibcp.fr

Bilan de la Méthode SOPMA (1995)

- Avantages
  - Méthode optimisée pour chaque séquence
  - Bonne qualité de prédiction (69%)
  - Prise en compte des familles de protéines homologues
  - La qualité augmente avec la taille de la base de données
- Inconvénients
  - Temps de calcul très long (20 à 30 minutes par séquence)
  - Mise à jour des bases de données régulière
  - Comparaison de séquences difficiles

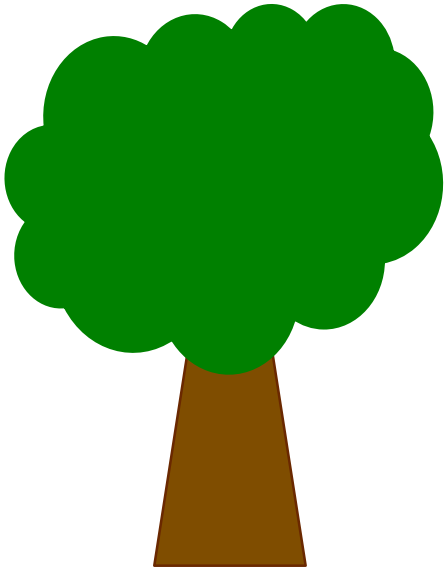
g.deleage@ibcp.fr

Mémoire associative

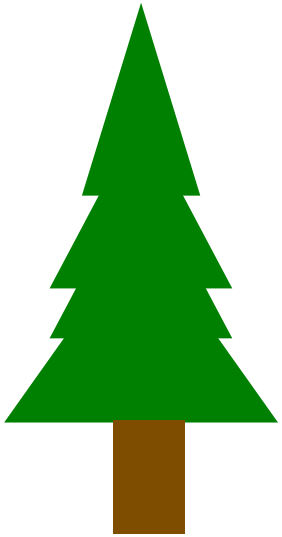
Les réseaux de neurones

g.deleage@ibcp.fr

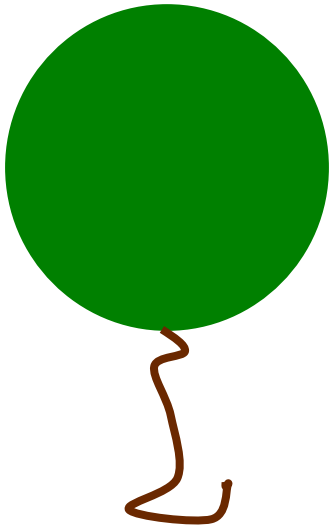
Mémoire associative



?



?



??

g.deleage@ibcp.fr

Vous avez dit apprentissage?

si vuos pvueoz lrie ccei, vuos aevz assui nu dôrle de cvreeau. Puveoz-vuos lrie ccei? Seleuemnt 56 porsnenes sur cnet en snot cpabales. Je n'en cyoaris pas mes yuex que je sios cabaple de cdrpormene ce que je liasis. Le povuoir phoémanénl du crveeau hmauin. Soeln une rcheerche fiate à l'Unievristé de Cmabridge, il n'y a pas d'iromtpance sur l'odrre dnas luqeel les ltertes snot, la suele cohse imotrpnate est que la priremère et la derènire lterte du mot siot à la bnone pcale. La raoisn est que le cveerau hmauin ne lit pas les mtos ltrete par letrte mias ptuôlt cmome un tuot. Étaonnnt n'est-ce pas? Et moi qui ai tujoours psneé que svaoir élpeer éatit ipomratnt! Si vuos poevuz le lrie, arols bavro !!!

g.deleage@ibcp.fr

Vous avez dit tolérance d'erreur?

UN B34U JOUR D'373,

J'37415 5UR L4 PL4G3 37 J3 R3G4RD415 D3UX J3UN35 F1LL35 JOU4N7 D4N5 L3 54BL3. 3LL35 CON57RU15413N7 UN CHÂ734U D3 54BL3, 4V3C 7OUR5, P4554G35 C4CH35 37 PON7-L3V15. 4LOR5 QU'3LL35 73RM1N413N7, UN3 V4GU3 357 4RR1V33 37 4 7OU7 D37RU17, R3DU154N7 L3 CH4734U 3N UN 745 D3 54BL3 37 D'3CUM3.J'41 CRU QU'4PR35 74N7 D'3FFOR7, L35 F1LL37735 COM3NÇ3R413N7 4 PL3UR3R, M415 4U CON7R41R3 3LL35 COURRUR3N7 5UR L4 PL4G3, R14N7 37 JOU4N7 37 COMM3NÇ3R3N7 4 CON57RU1R3 UN 4U7R3 CHÂ734U. J'41 COMPR15 QU3 J3 V3N415 D'4PPR3NDR3 UN3 GR4ND3 L3ÇON. NOU5 P455ON5 UN3 GR4ND3 P4R713 D3 NO7R3 V13 4 CON57RU1R3 D35 CHO535 M415 LOR5QU3 PLU5 74RD UN3 V4GU3 L35 D3MOL17, L35 53UL35 CHO535 QU1 R3573N7 5ON7 L'4M1713, L'4MOUR 37 L '4FF3C71ON 37 L35 M41N5 D35 G3N5 QU1 5ON7 C4P4BL35 D3 NOU5 F41R3 5OUR1R3.

g.deleage@ibcp.fr

Mémoire associative

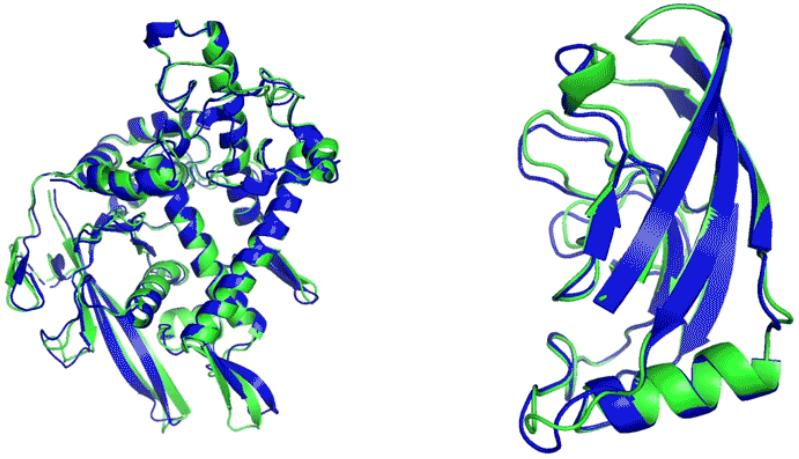
LA CDECANE DE CET OTRIRAOO SEBLME
ECVSSEIXE
PUOR UN CEITVALICNSE NOTEPYHE.

Le réseau de neurone →

Réseaux de neurones
Machine learning
Méthode d'apprentissage
Intelligence articielle
Deep learning

↓

https://deepmind.com/ ←



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

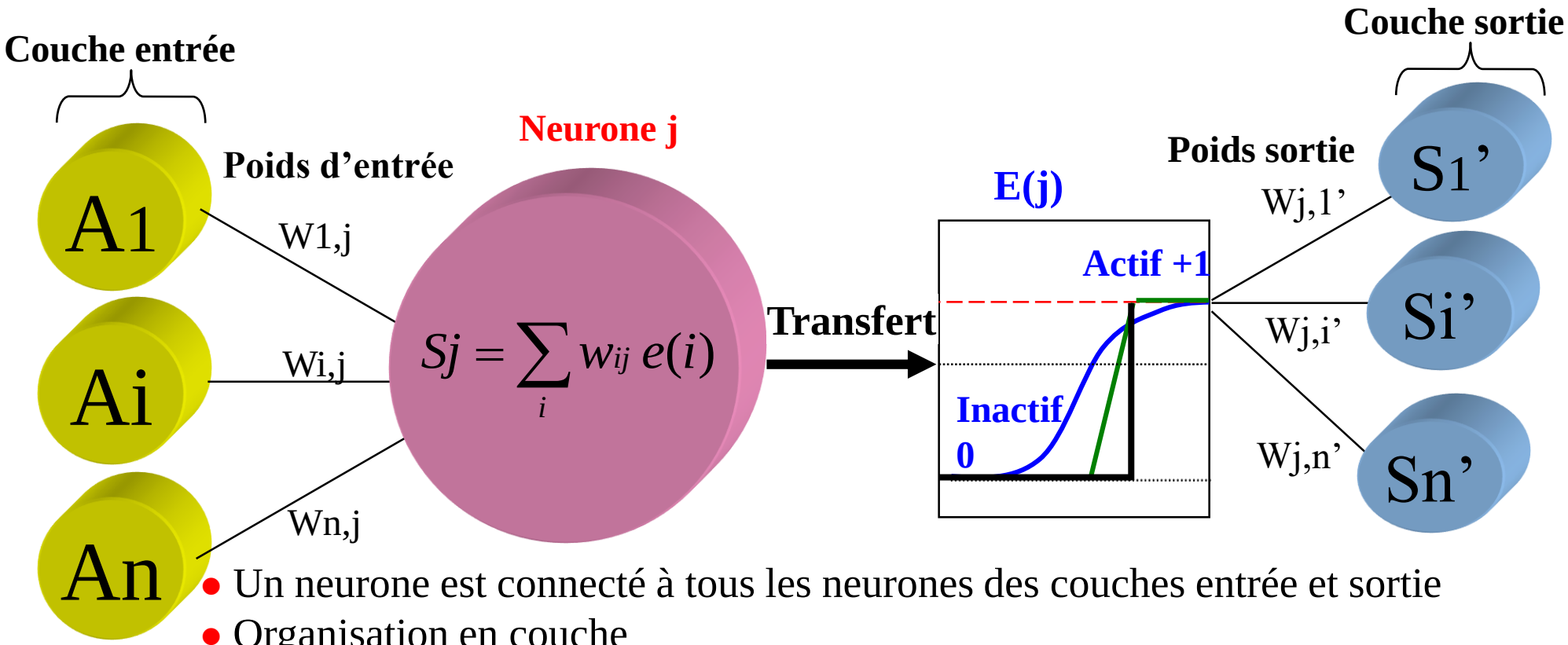
T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

30/11/2020

g.deleage@ibcp.fr

Le neurone artificiel binaire j



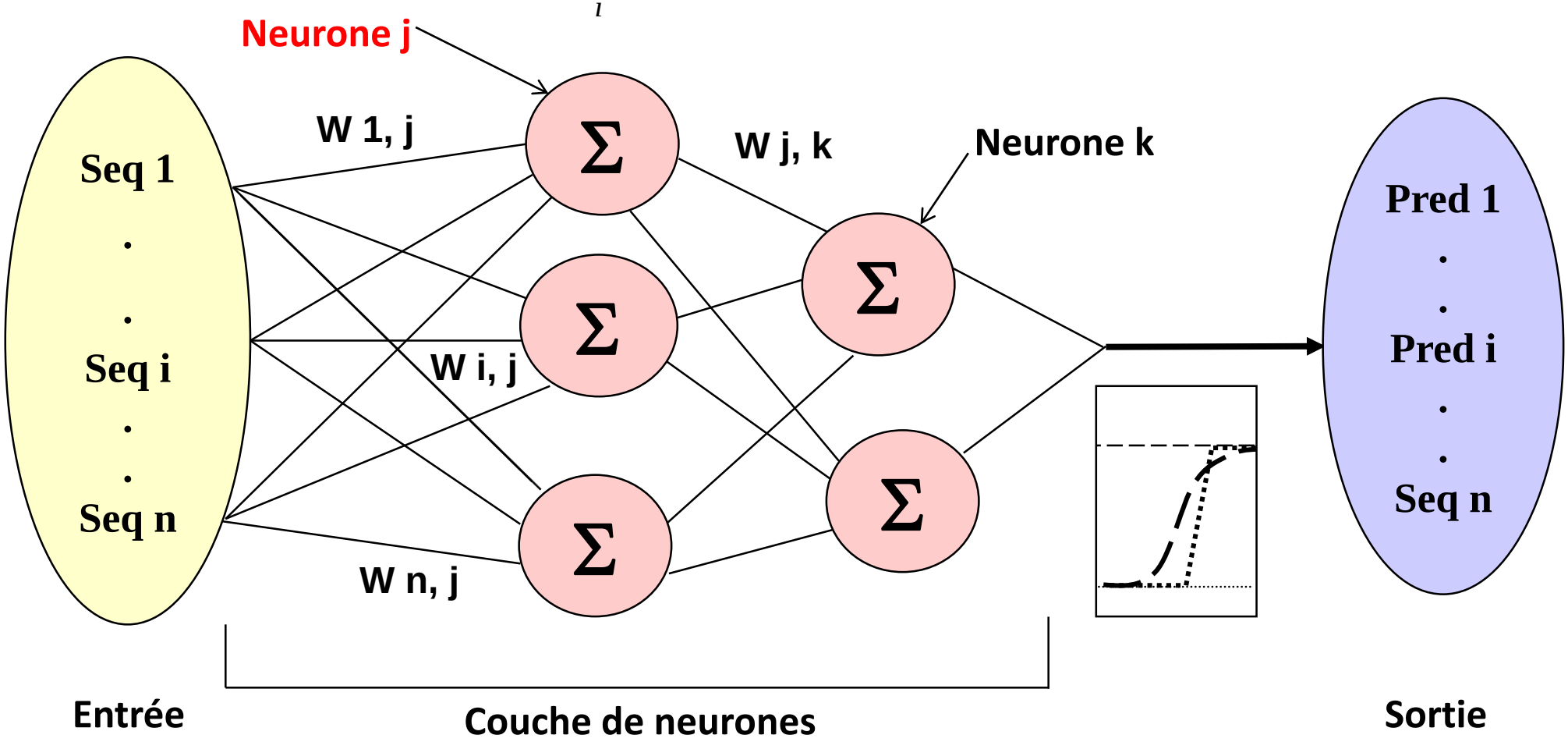
- Un neurone est connecté à tous les neurones des couches entrée et sortie
- Organisation en couche
- 3 phases:
  - Apprentissage => Trouver les poids w(i, j) qui optimisent les sorties avec les entrées fournies (long)
  - Fonctionnement : Utilisation des poids en production (rapide)
  - Généralisation du réseau

Non utilisable sur des problèmes non linéairement séparables

g.deleage@ibcp.fr

Réseau de neurones Multilayer Perceptron

$$Sj = \sum_i w_{ij}\, e(j)$$

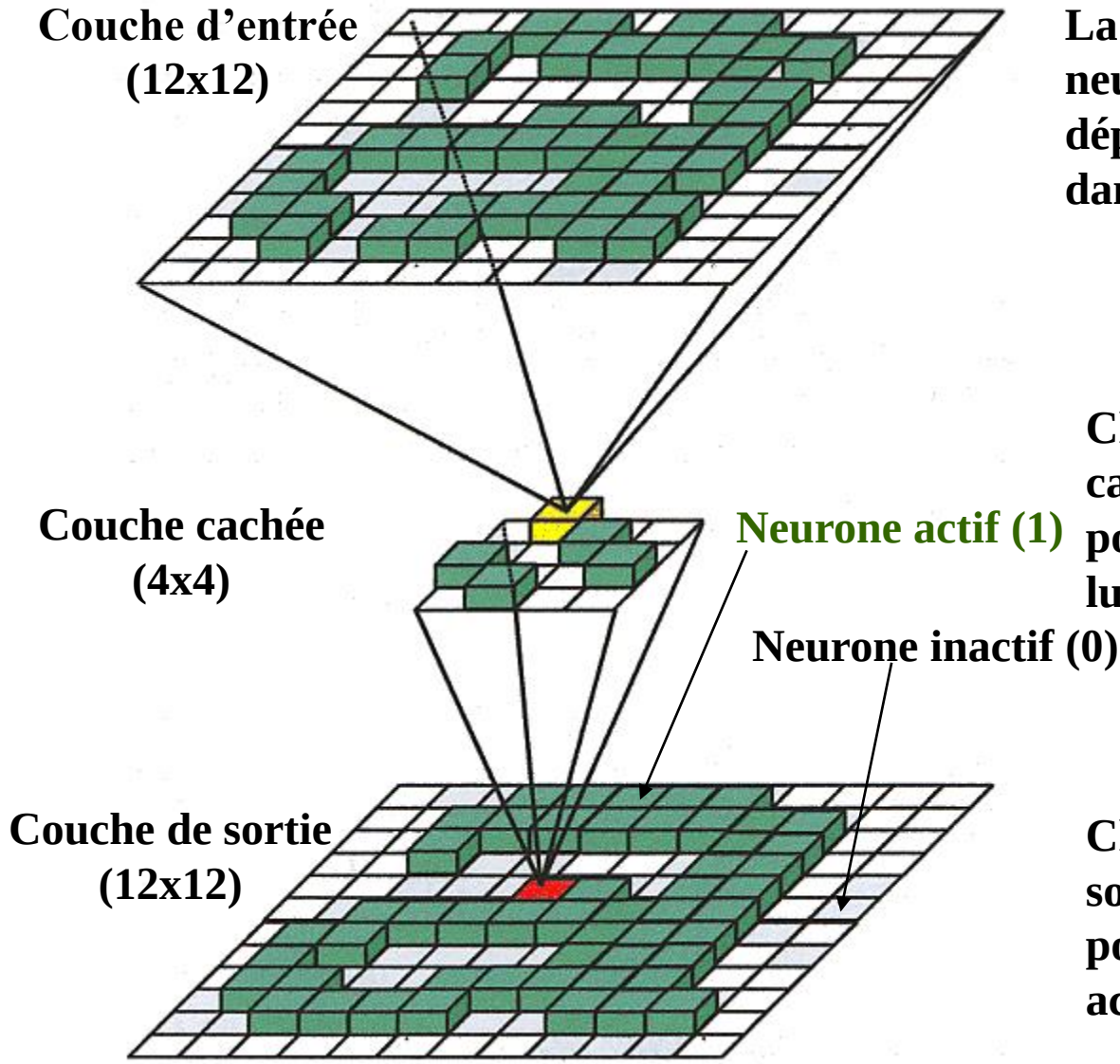


- Division de la base complète (126 protéines) en 2 échantillons
- Apprentissage sur la moitié des exemples
- Test de généralisation sur l'autre moitié

g.deleage@ibcp.fr

Exemple de fonctionnement d'un réseau de neurones binaires

144x16x144= 4608 connexions



La lettre a est codée sur 144 pixels (1 neurone par pixel) les neurones actifs dépendent directement de l'image du a dans la grille d'entrée.

Chacun des 16 neurones de la couche cachée (jaune) fait la moyenne des 144 poids de la couche d'entrée et définit pour lui un état actif ou inactif

Chacun des 144 neurones de la couche sortie (rouge) fait la moyenne des 144 poids de la couche cachée et définit un état actif (vert) ou inactif (blanc)

g.deleage@ibcp.fr

Méthode PHD (Rost & Sander, 1993)

Niveau 1
Séquence/structure
Entrée profils
Sortie a,b,c

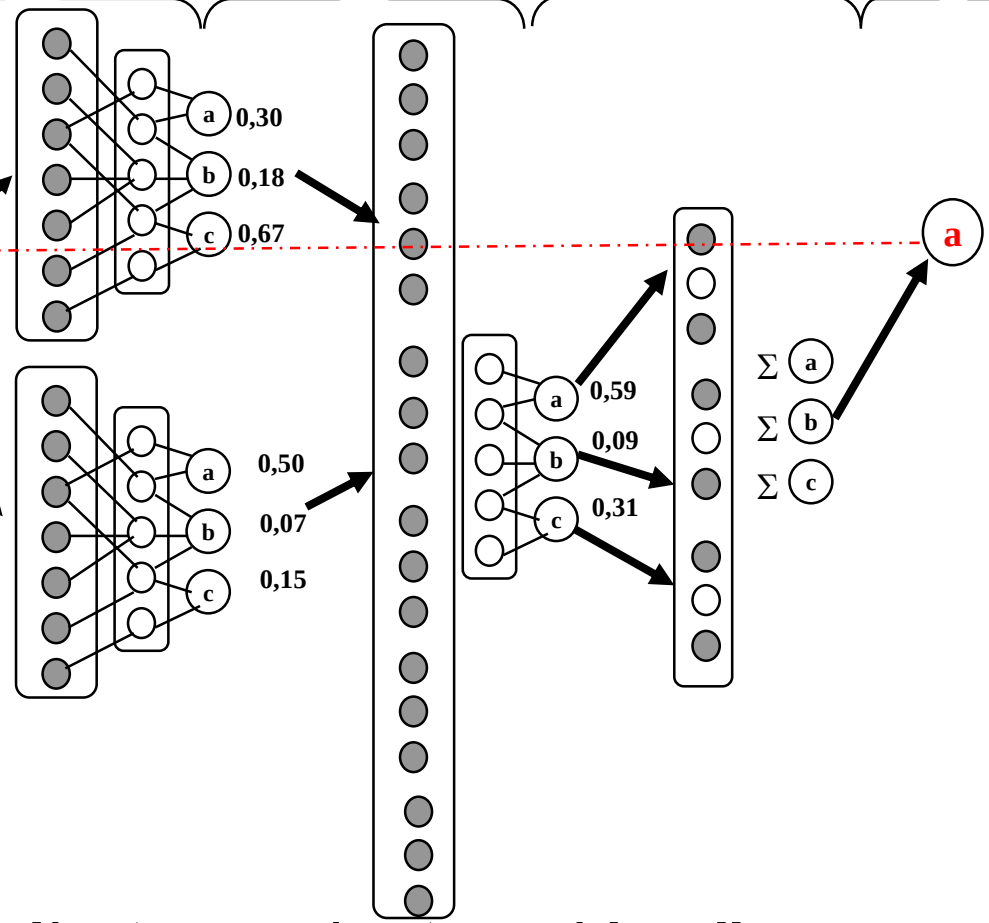
Niveau 2
Structure/structure
Entrée sortie de 1
Sortie a,b,c

Niveau 3
Décision
Entrée sortie de réseaux
Sortie moyenne a,b,c

Niveau 4
Prédiction
Meilleur score

Profils de séquences

```
K     K.HK    1:K=75%/H=25%
E     EDAE    2:E=60%/D=20%/A=20%
L h   FFFF    3:L=20%/F=80%
N h   SAAS    4:N=20%/S=40%/A=40%
D h   QKKQ    5:D=20%/Q=40%/K=40%
L h   LLLL    6:L=100%
E h   EEEE    7:E=100%
K h   KEKK    8:K=80%/E=20%
K h   KQEK
Y h   FFYF
N     DDND
A     AAAA
H b   RKKR
I b   LLLL
G b   GGGG
```



- Division de la base complète (126 protéines) en 2 échantillons
- Apprentissage sur la moitié des exemples
- Test de généralisation sur l'autre moitié

g.deleage@ibcp.fr

Méthode de Rost et Sander (1993)

- Avantages
  - Méthode d'apprentissage performante
  - Bonne qualité de prédiction (69%=>72,5%)
  - La qualité augmente avec la taille des bases de données
  - Instantané en production
  - Apport des alignements multiples

- Inconvénients
  - Pertinence de la base de données d'apprentissage
  - Pas de compréhension des mécanismes prédictifs
  - Paramètres du réseau (nombre de neurones, couches cachées)
  - Réapprentissage long (à refaire en fonction de la base de données)

g.deleage@ibcp.fr

Méthodes statistiques

- **Chou et Fasman (1974 =>1980)**
  - Peu de paramètres (≈120 coefficients)
  - La première méthode utilisable par les biologistes
  - Méthode manuelle
  - Non reproductible
  - Difficile à implémenter
  - Qualité 52%
- **Garnier *et al.* (GOR I, II, III, IV) (1978 =>1989)**
  - Utilise la théorie de l'information
  - Prise en compte de l'environnement séquentiel (17 x 4 x 20 ≈1360 coefficients)
  - Méthode automatique non ambiguë
  - Rapide (instantanée)
  - Méthode insensible à l'homologie
  - Qualité 56% à 65%
- **Double Prédiction DPM (Deléage et Roux, 1987)**
  - Confrontation de la prédiction de la classe structurale et des % de structures
  - Qualité 60%
- **Discrimination linéaire DSC (King et Sternberg, 1996)**
  - Hydrophobie, effets de terminaison, propensions aa, filtrage
  - Utilisation des alignements multiples
  - Qualité 68,5%

Faible évolutivité des méthodes statistiques

g.deleage@ibcp.fr

Méthodes de similarité & neuronales

- **Plus proches voisins (Levin *et al.*, 1986; 1988)**
  - Comparaison de peptides courts (≈ 10^{10} pour 500 acides aminés)
  - Paramètres (Matrice, seuil, longueur des peptides)
  - Méthode automatique
  - Facile à implémenter
  - Temps de calcul
  - Qualité 62%
- **Méthodes auto-optimisées Geourjon et Deléage (SOPM, 1994 ; SOPMA, 1995)**
  - Méthode automatique non ambiguë
  - Temps de calcul assez long (5'/séquence)
  - Méthode sensible à l'homologie
  - Prise en compte des alignements de protéines homologues
  - Qualité 65-72%
- **Réseaux de neurones (PHD, Rost et Sander 1993->1999 ; HNN 1997, Nnpredict, Predator)**
  - Méthodes optimales
  - Paramètrage délicat (couches, neurones, jeu test, ré-apprentissage)
  - Qualité 72-75%
- **Prise en compte des familles de protéines homologues (PHD, SIMPA96, SOPMA)**
  - Méthodes modulaires
    - FASTA, BLAST, CLUSTALW, Prédiction
    - SWISS-PROT, PDB

Forte évolutivité des méthodes similaires

g.deleage@ibcp.fr

- ✓ PHD https://www.predictprotein.org/
- ✓ JPRED 4 http://www.compbio.dundee.ac.uk/~www-jpred/
- ✓ PSIpred 4 http://bioinf.cs.ucl.ac.uk/psipred/
- ✓ NPS@ https://npsa-prabi.ibcp.fr/
  - ✓ SOPM (Geourjon and Deléage, 1994)
  - ✓ SOPMA (Geourjon and Deléage, 1995)
  - ✓ HNN (Guermeur, 1997)
  - ✓ MLRC (Guermeur *et al.*, 1999)
  - ✓ DPM (Deléage and Roux, 1987)
  - ✓ DSC (King and Sternberg, 1996)
  - ✓ GOR I (Garnier *et al.*, 1978)
  - ✓ GOR III (Gibrat *et al.*, 1987)
  - ✓ GOR IV (Garnier *et al.*, 1996)
  - ✓ PHD (Rost and Sander, 1993)
  - ✓ PREDATOR (Frishman and Argos, 1996)
  - ✓ SIMPA96 (Levin, 1997)
- ✓ Logiciel client/serveur
  - ✓ ANTHEPROT (http://antheprot-pbil.ibcp.fr)

g.deleage@ibcp.fr

Difficulté d'établir une base de données de référence

- Trouver des protéines représentatives de l'ensemble des protéines
  - Protéines membranaires?
  - Que faire des complexes et des ligands?
  - Structures différentes de la même protéine.
- Analyse automatique des structures 3D
  - Critères géométriques (angles phi et psi , Ramachandran) Levitt & Greer
  - Critères énergétiques (liaisons hydrogènes) DSSP Kabsch & Sander
  - Combinaison Psea (Colloch et al.)
- Fixer un seuil de similarité entre les protéines (50% ou 25% d 'identité)
  - Chou & Fasman, 1978, 29 protéines (totalité)
  - Kabsch et Sander (1983) 60 protéines (Id<50%)
  - Geourjon & Deléage (1994) 234 protéines (Id<50%)
  - Rost & Sander (1993) 126 protéines (Id< 25%)
  - Hobohm et Sander (1998) 700 protéines (Id<25%)
- Comment intégrer les familles de séquences de protéines?
  - Alignements multiples (qualité, exhaustivité, mise à jour)

g.deleage@ibcp.fr

L'assignation des structures secondaires d'après les fichiers PDB

g.deleage@

Pourcentage Q de qualité de prédiction

$$Qk = \frac{\sum_{i=1}^{k} NbC(i)}{\sum_{i=1}^{k} NbO(i)}$$

- NbC(i): nombre d'acides aminés correctement prédits dans l'état i
- NbO(i): nombre d'acides aminés observés dans l'état i
- k : Nombre total de conformations prises en compte
  - k =3 hélice, feuillet, apériodique =>Q 3
  - k =4 hélice, feuillet, turn et apériodique =>Q4

NB: La valeur de Qk ne donne pas d'information sur la fiabilité de prédiction

g.deleage@ibcp.fr

Coefficients de corrélation de prédiction

$$C(i) = \frac{(pi.ni) - (ui.oi)}{[(ni+ui)(ni+oi)(pi+ui)(pi+oi)]^{1/2}}$$

- **I désigne un état parmi les k possibilités (hélice, feuillet, turn et apériodique)**
- **pi : nombre de résidus correctement prédits et observés en i**
- **ni : nombre de résidus correctement prédits et non observés en i**
- **ui : nombre de résidus non prédits et observés en i**
- **oi : nombre de résidus prédits et non observés en i**
- **C(i)=1 corrélation positive totale**
- **C(i)=0 pas de corrélation**
- **C(i)=-1 corrélation négative totale**

g.deleage@ibcp.fr

Facteurs de qualité

- **La qualité dépend de la base de référence (critères utilisés)**
- **La qualité dépend du nombre d'états conformationnels utilisés**
  - Prédiction au hasard selon 2 états : Q=50%
  - Prédiction au hasard selon 3 états : Q=33%
- **La qualité dépend de l'état structural considéré**
  - $Q_{apériodique} > Q_{hélice} > Q_{feuillet}$
- **La qualité dépend de la différence entre le score le plus élevé pour un état structural donné et le second**

- **La qualité dépend aussi de la méthode utilisée, des critères d'estimation de la qualité, de l'accord entre les différentes méthodes et bien sûr de l'expertise de l'utilisateur**

g.deleage@ibcp.fr

Déviation standard en teneur de structures

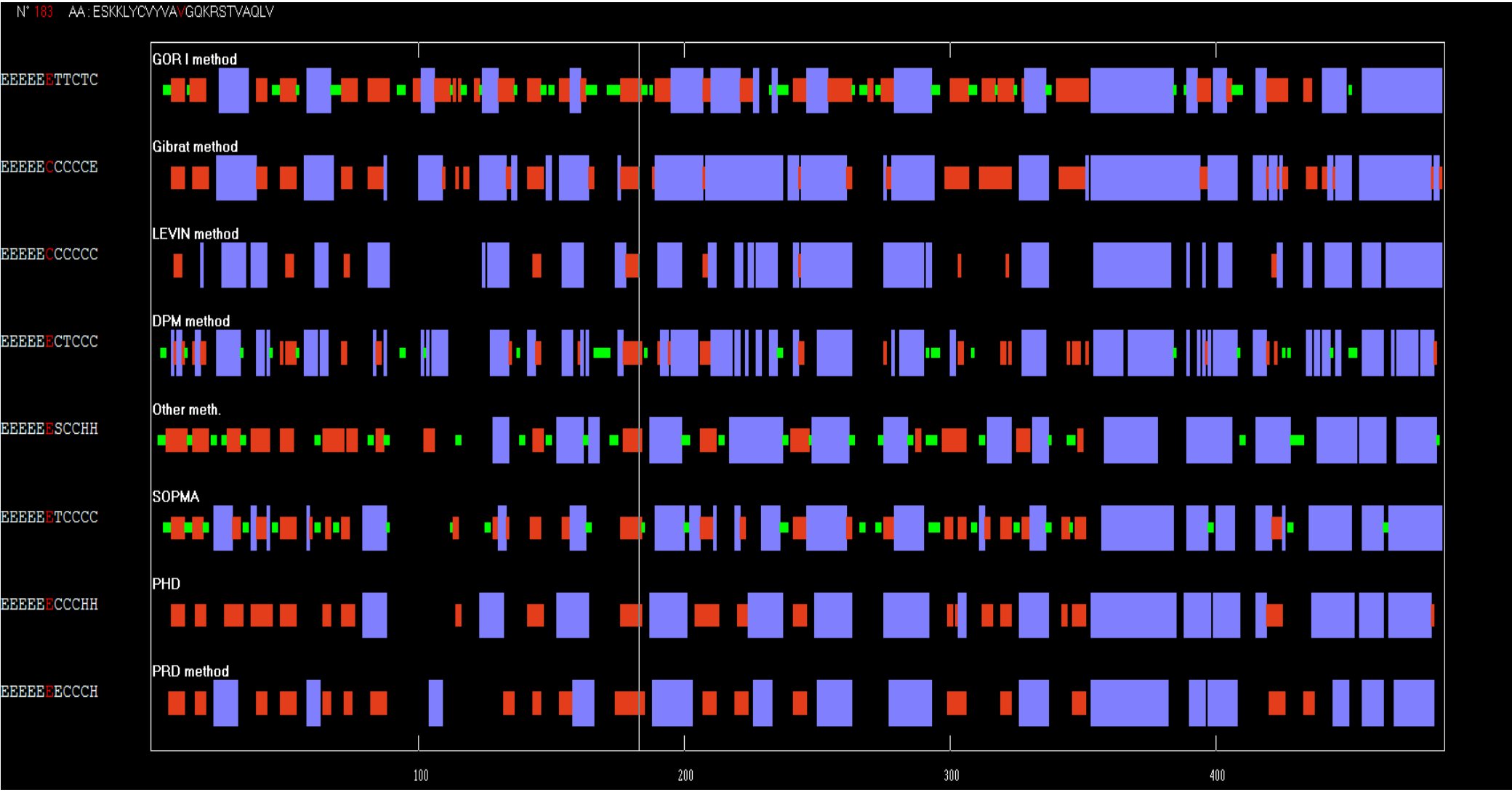
- σ est la déviation standard de la composition en structures secondaires

$$\sigma(i) = \sqrt{\sum_{j=1}^{n} \left(o(j) - p(j)\right)^2}$$

- j est
- i est un des états conformationnels
- O(j) est le pourcentage d'acides aminés observés dans l'état i dans la protéine j
- p(j) est le pourcentage d'acides aminés prédits dans l'état i dans la protéine j

g.deleage@ibcp.fr

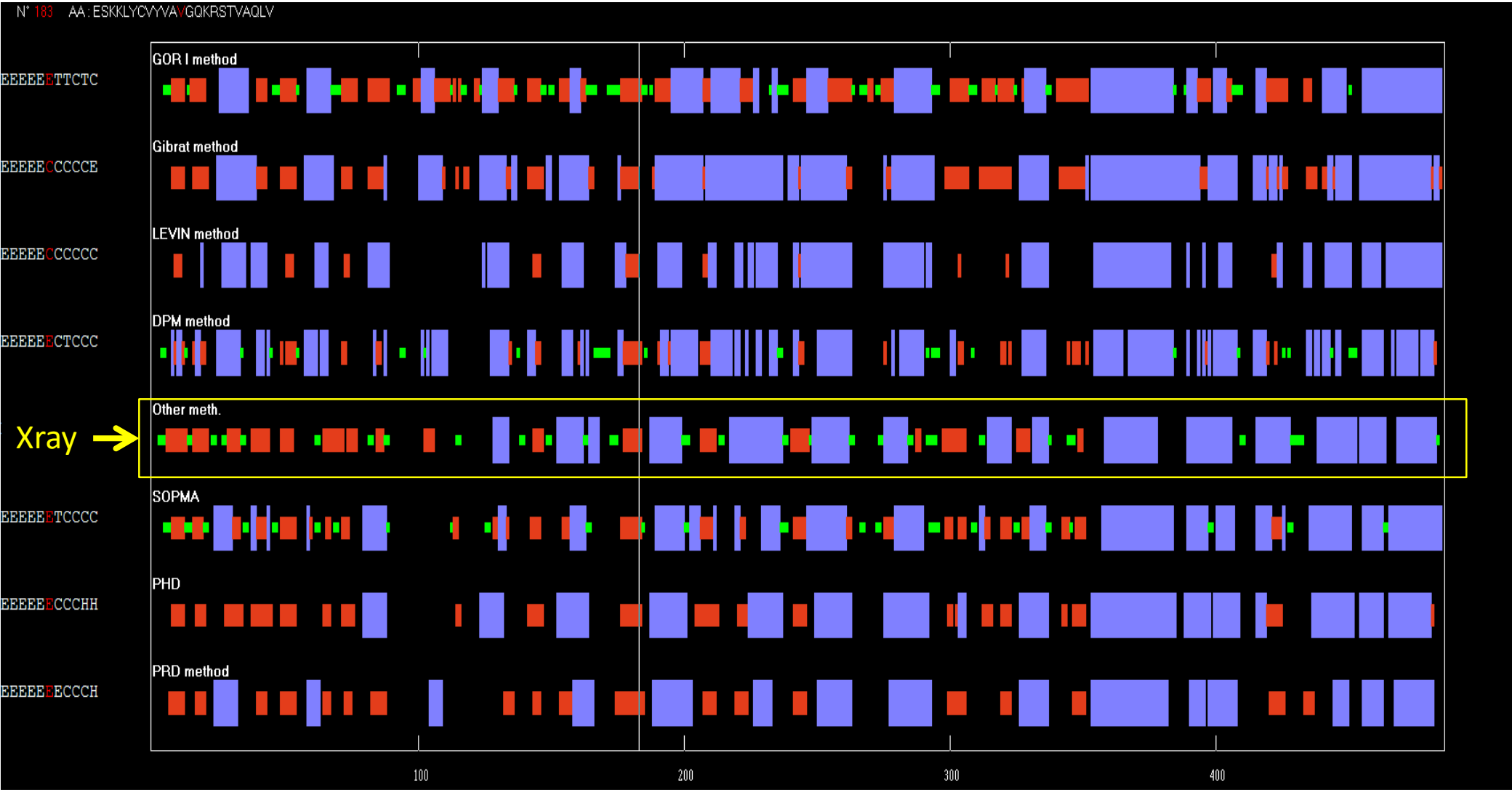
Prédictions de structures secondaires à l'œuvre



Brins β Hélices α

g.deleage@ibcp.fr

Prédictions de structures secondaires à l'œuvre



Brins β Hélices α

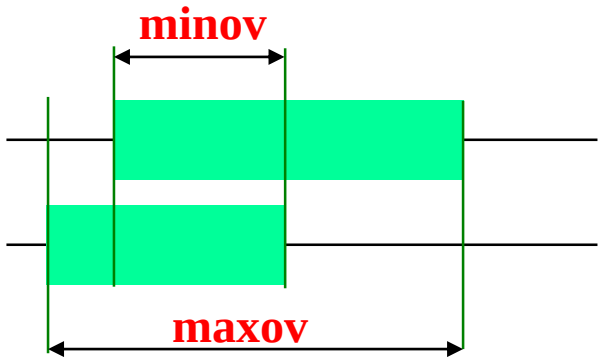
g.deleage@ibcp.fr

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



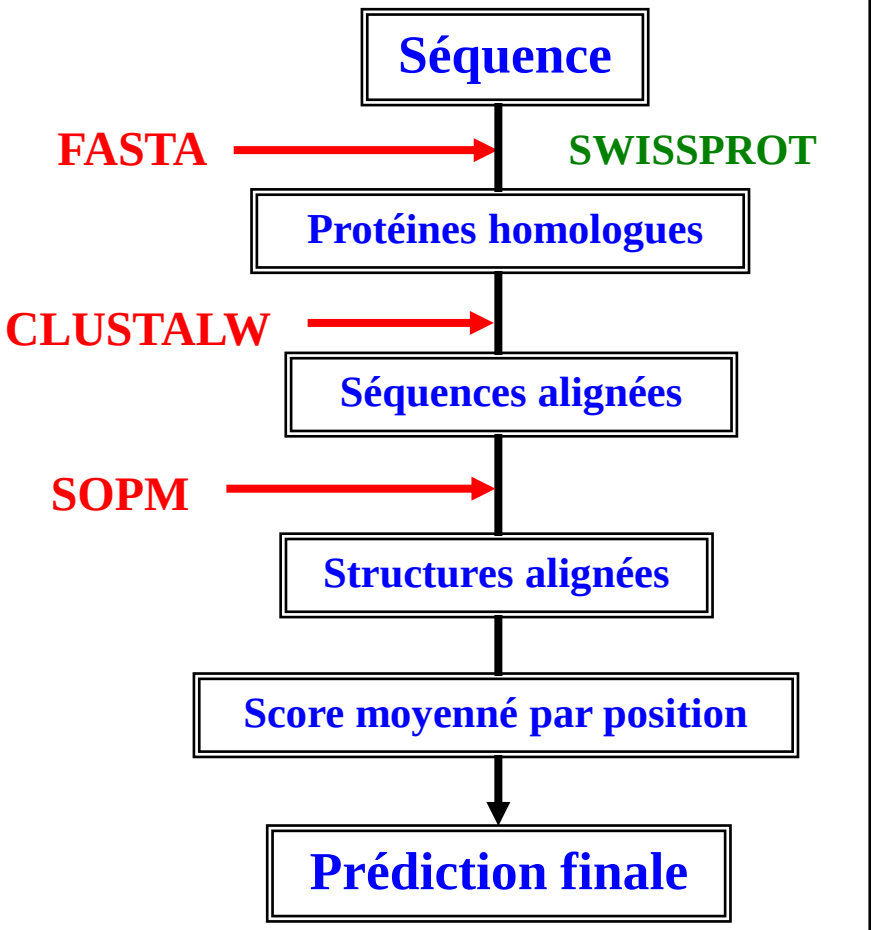
- δ est défini par :

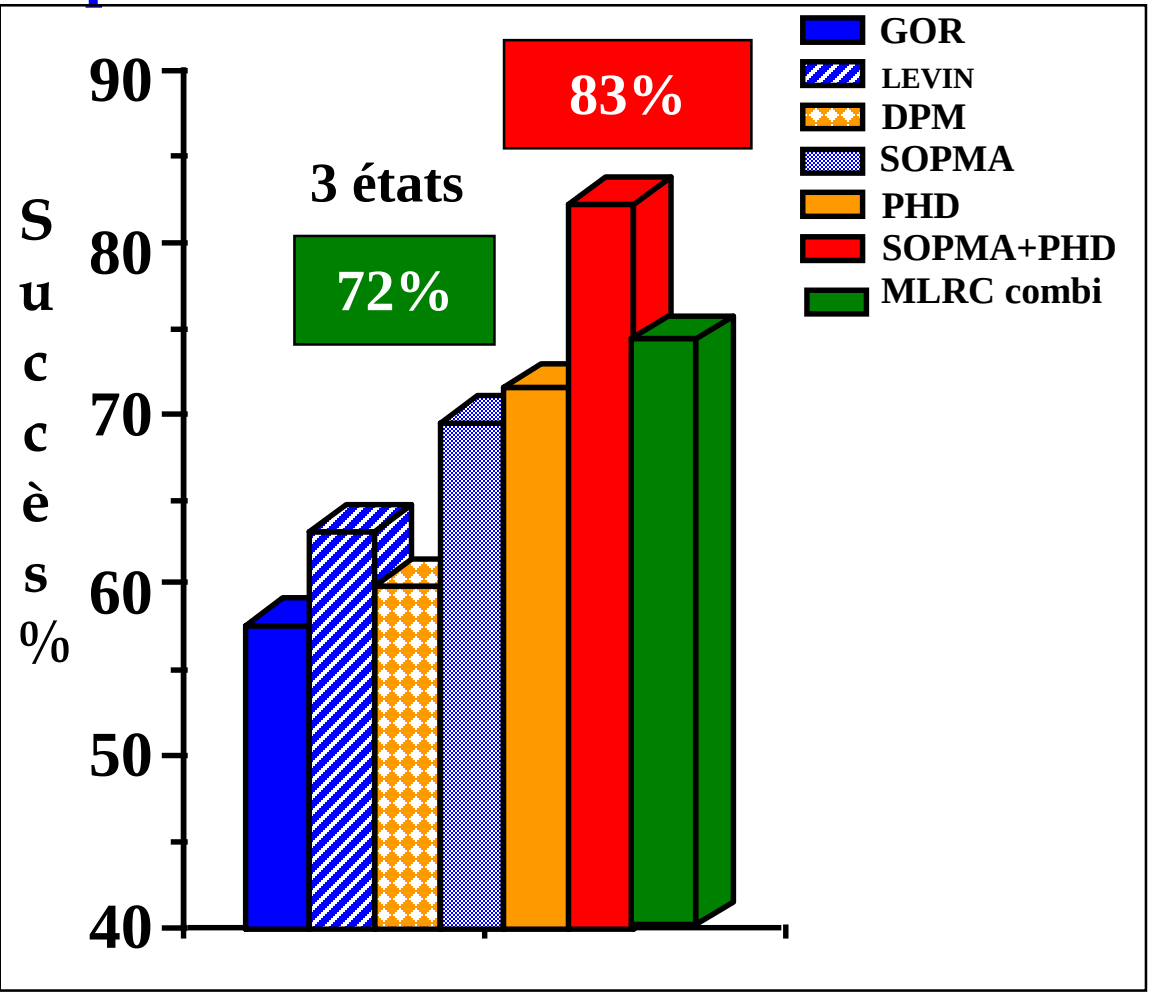
$$\delta(s_q, s_t) = \min \left\{ \begin{array}{l} (maxov(s_q, s_t) - minov(s_q, s_t)); minov(s_q, s_t); \\ int(len(s_q/2)); int(len(s_t/2) \end{array} \right\}$$

g.deleage@ibcp.fr

Prédictions de structures secondaires

- **Méthodes isolées 72%**
- **Prédictions jointes 82% pour 75% des acides aminés**

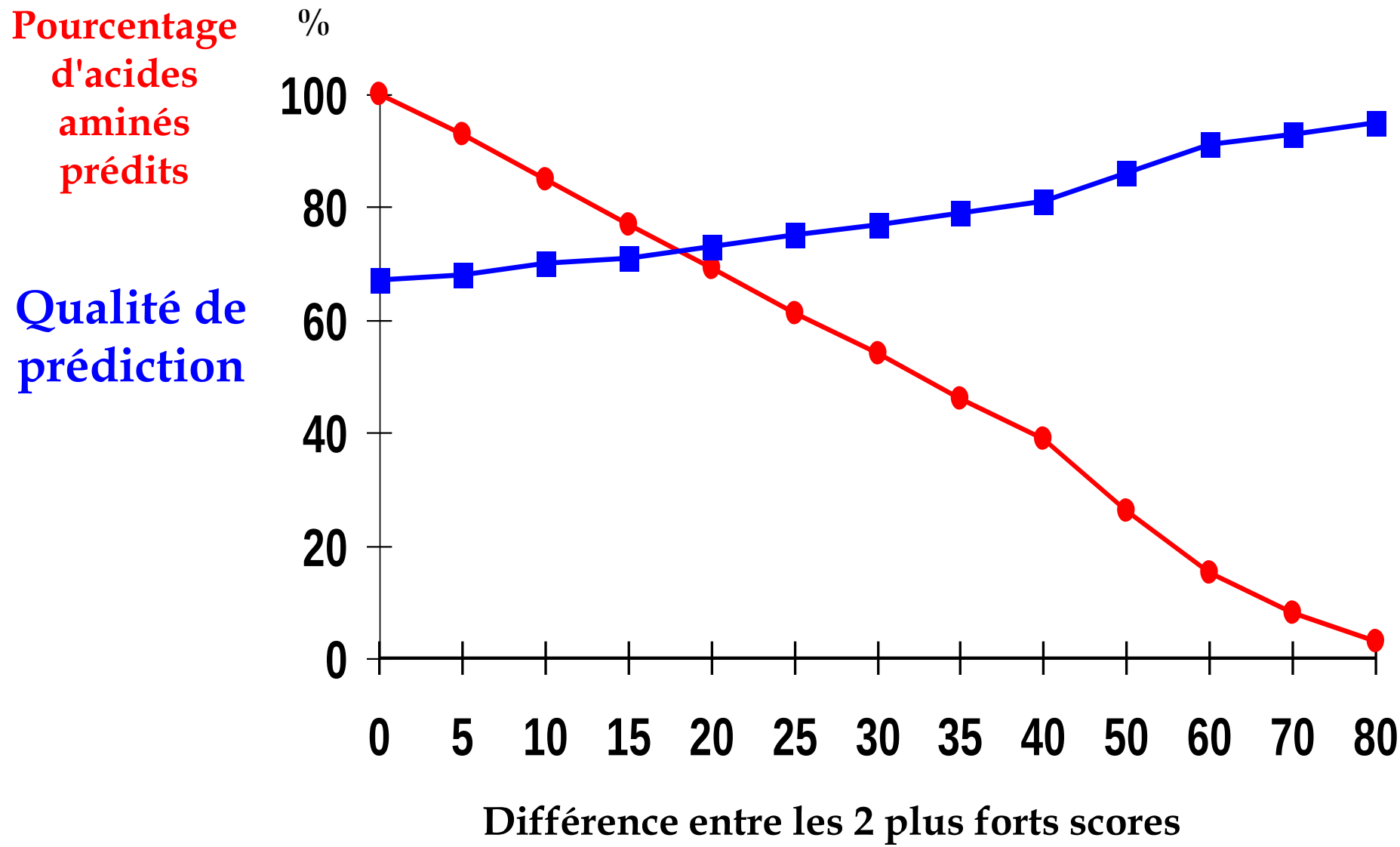




- ✓ Les meilleures méthodes sont modulaires et utilisent les familles de séquences
- ✓ Les méthodes actuelles ne prennent en compte que les informations locales
- ✓ La qualité de prédiction est au mieux de 75% (Q_3 HEC base de données <25% d'identité)
- ✓ Au mieux 75% des structures secondaires sont prédictibles et 25% dépendent du repliement 3D

g.deleage@ibcp.fr

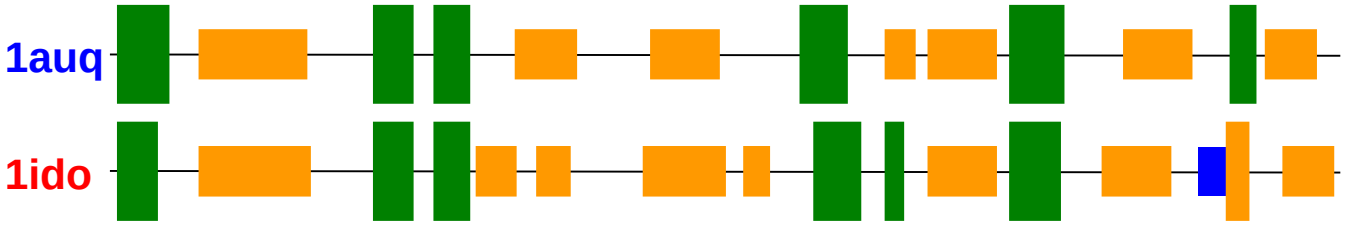
Calibrage de la qualité de prédiction

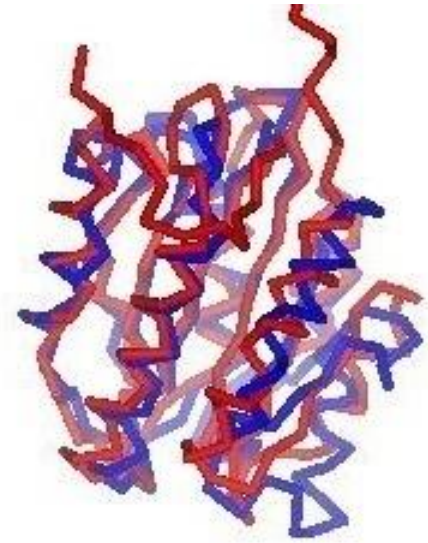


g.deleage@ibcp.fr

Pourquoi les prédictions de structures secondaires sont utiles?

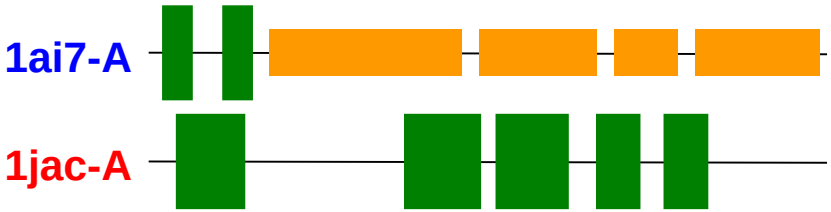
15,9% identity ; Close structures (RMSD 2Å)

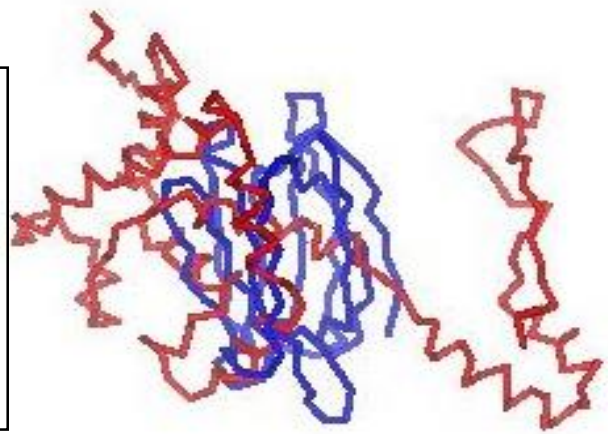




α helix β sheet

16% identity ; Different structures (RMSD = 20 Å)





g.deleage@ibcp.fr

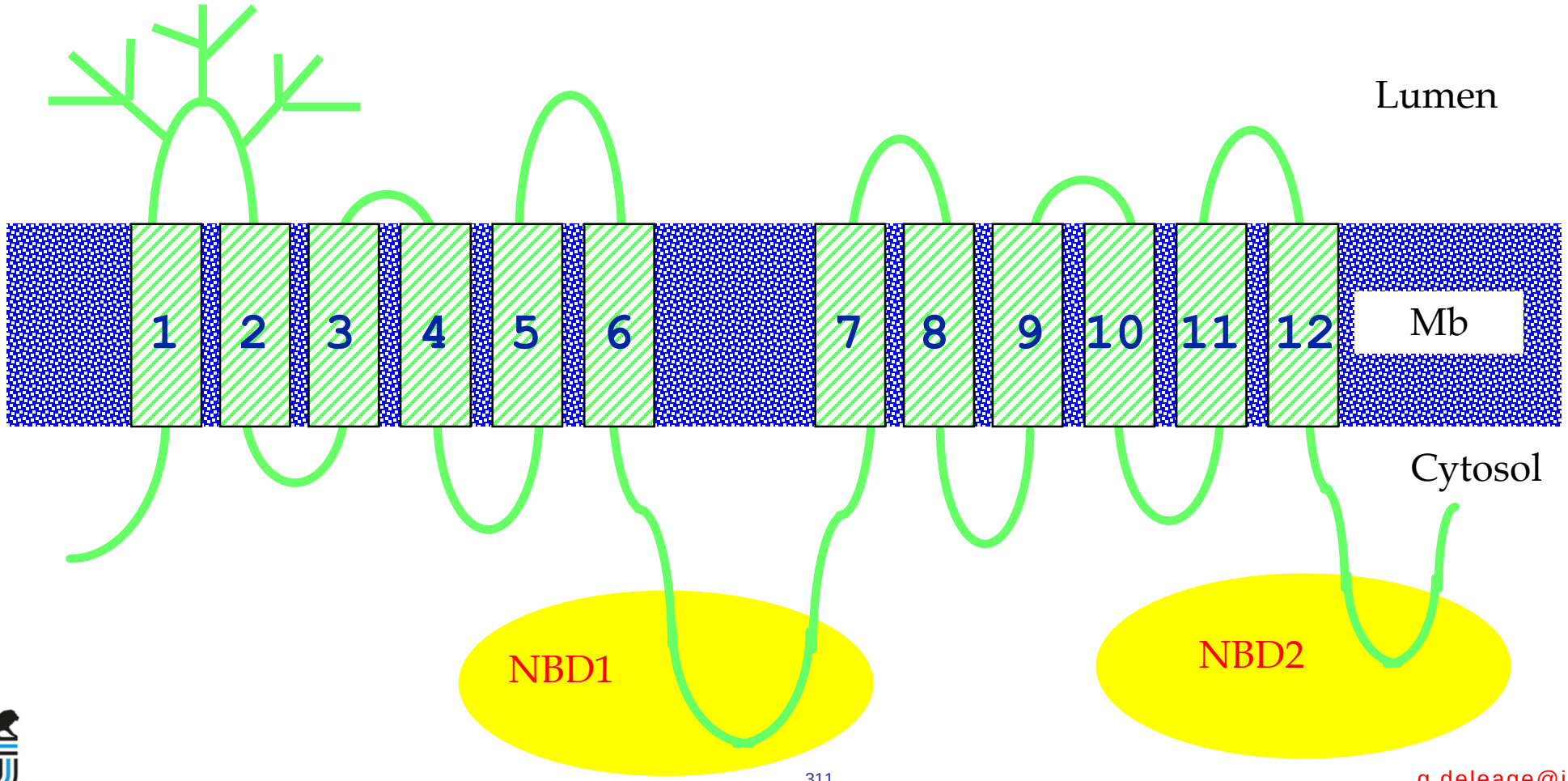
Applications - Exemples

- Structures secondaires plus conservées que les séquences
  - Glycoprotéine P
- Prédictions, structure X et expérimentation
  - Transcriptase Inverse de HIV
- Prédictions de limites de domaines structuraux et RMN
  - FruR (Répresseur de l'opéron fructose)
  - Heat Shock Factor de maïs
- Phylogénie structurale => mécanisme d'interaction
  - FruR (Répresseur de l'opéron fructose)
- Modélisation par analogie
  - gp 41 d'HIV
  - Protéine apoptotique NR-13

g.deleage@ibcp.fr

Glycoprotéine P

- **Protéine membranaire**
  - Rôle de résistance aux agressions chimiques
  - Rejette les agents chimiques hors de la cellule
  - Transport actif peu spécifique dépendant de l'hydrolyse d 'ATP
- **Cause de la résistance à la chimiothérapie au long cours**



g.deleage@ibcp.fr

Recherche de la région de fixation de nucléotides

Domaine de fixation de nucléotides

ATP/GTP-binding site motif A (P-loop) PS00017

[AG]-x(4)-G-K-[ST]

Theoretical frequency: 1,09 E-4

Site : 427 to 434 GNSGCGKS Observed frequency: 2,72 E-5

Site : 1070 to 1077 GSSGCGKS Observed frequency: 2,72 E-5

Recherche du motif

[AG]-x(4)-G-K-[ST]

Protein Data Bank

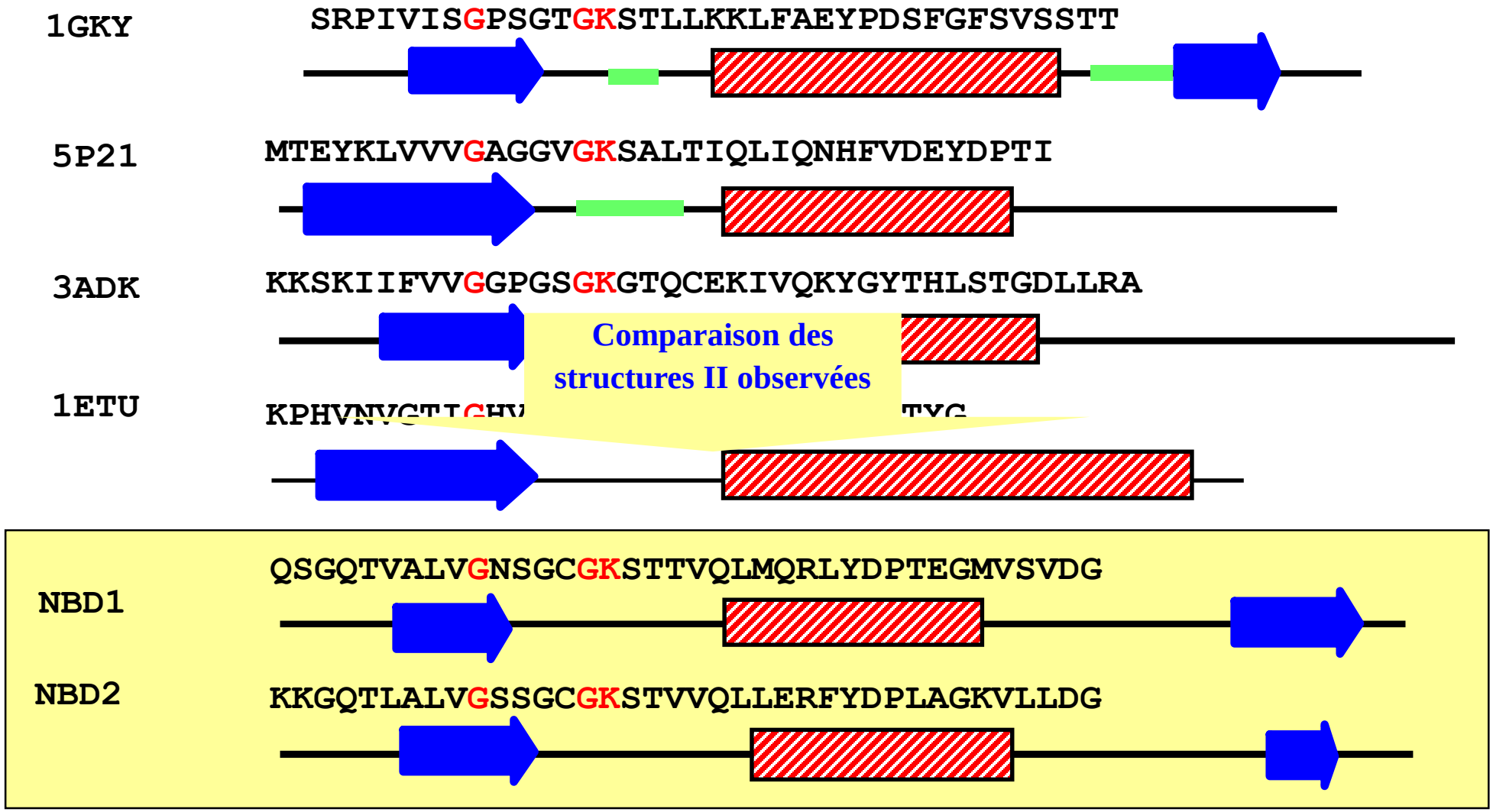
4 familles de protéines détectées alignées

Guanylate kinase, Facteurs d'élongation, Oncogène RAS, adénylate kinase

```
          1GKY   SRPIVISGPSGTGKSTLLKKLFAEYPDSFGFSVSSTT
          5P21 MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTI
          3ADK KKSKIIFVVGGPGSGKGTQCEKIVQKYGYTHLSTGDLLRA
          1ETU KPHVNVGTIGHVDHGKTTLTAAITTVLAKTYG
          NBD1 QSGQTVALVGNSGCGKSTTVQLMQRLYDPTEGMVSVDG
          NBD2 KKGQTLALVGSSGCGKSTVVQLLERFYDPLAGKVLLDG
     Consensus              *    * *.
```

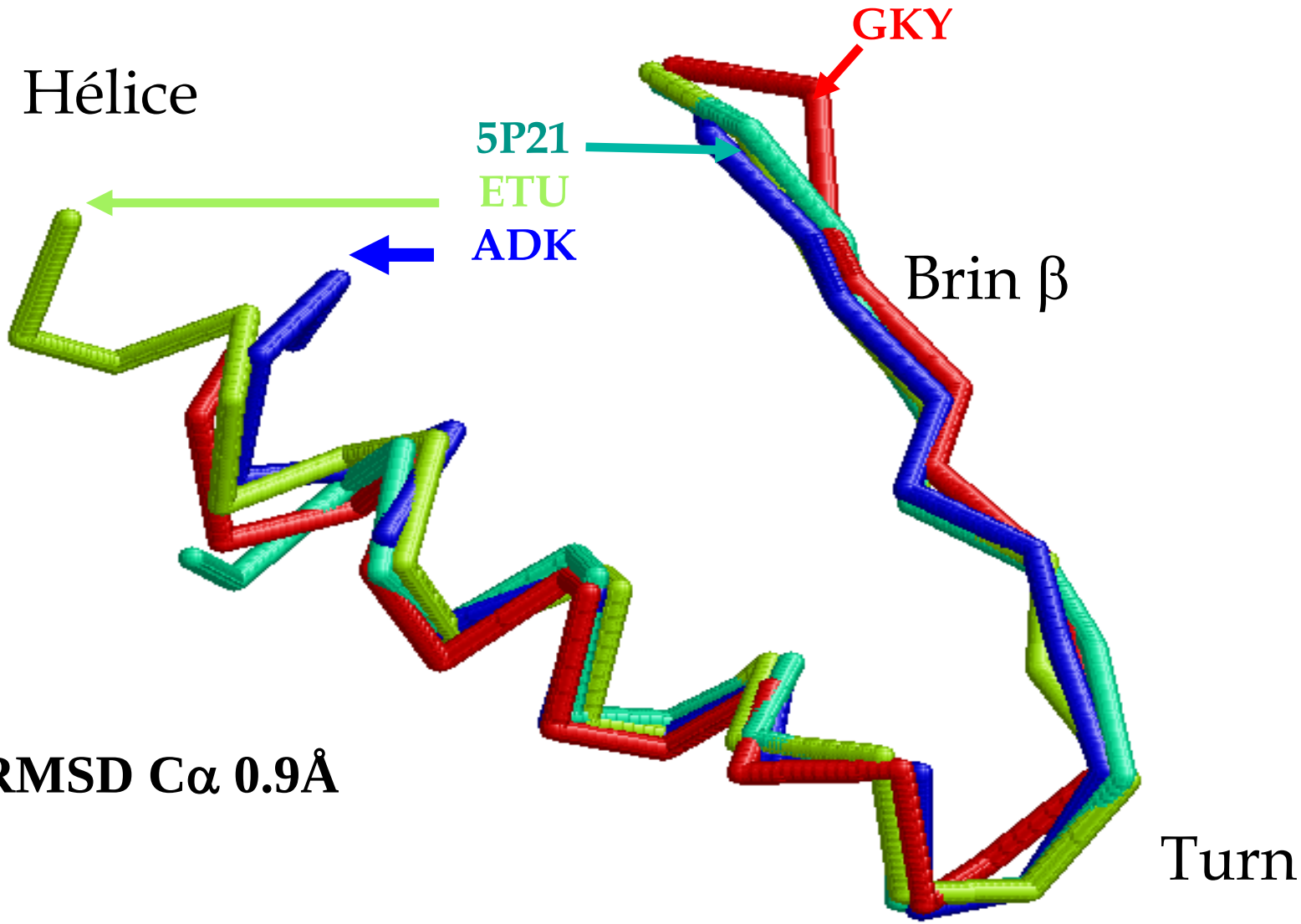

g.deleage@ibcp.fr

Domaine de fixation de nucléotides



g.deleage@ibcp.fr

Domaine de fixation de nucléotides - Glycoprotéine P



Dong, Ladavière, Penin, Deléage et Baggetto BBA **1371** (1998) 317-334

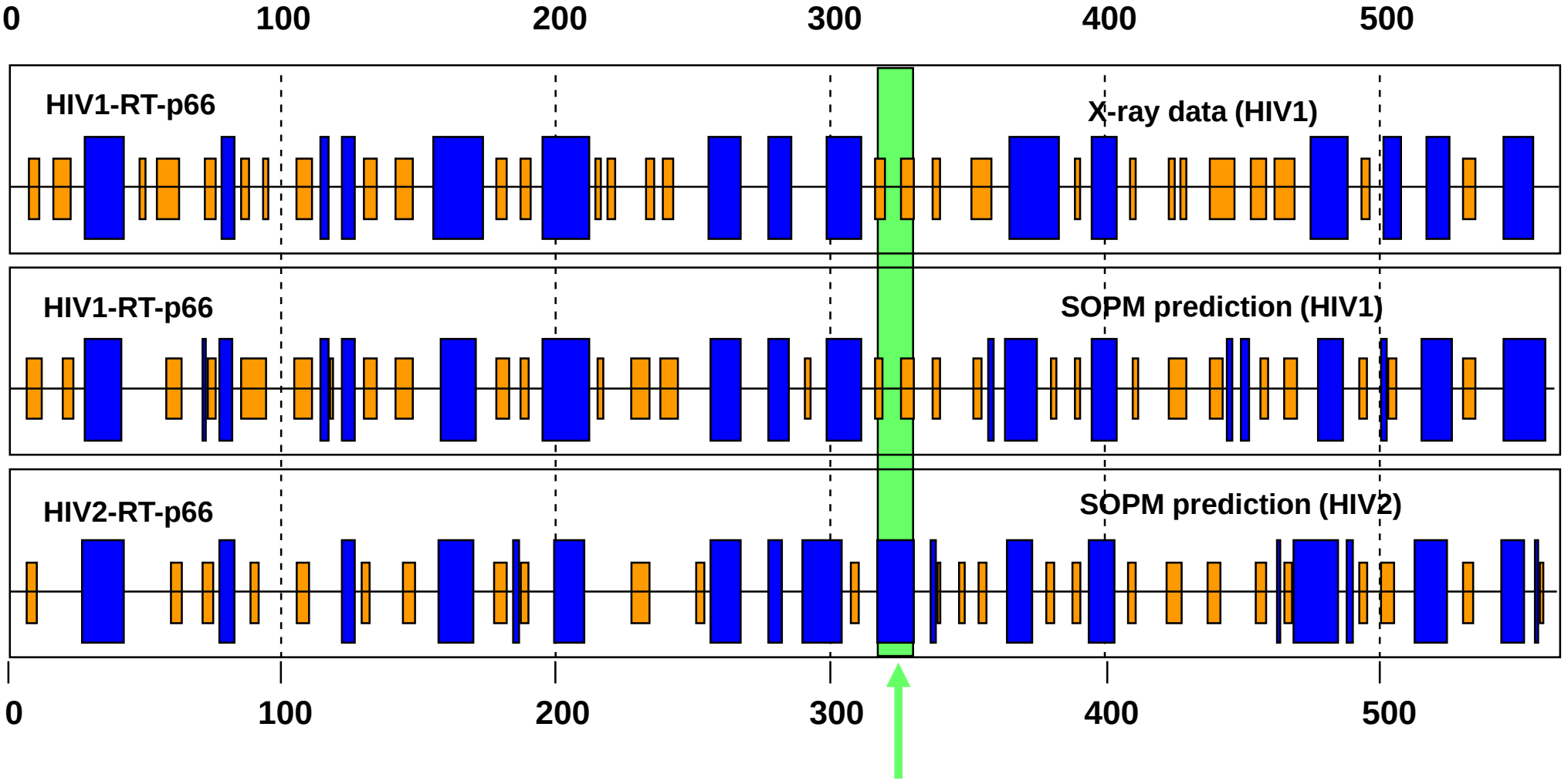
g.deleage@ibcp.fr

Transcriptase inverse du VIH

- **Données sur VIH**
  - 2 souches (VIH 1 et VIH 2)
- **Données sur transcriptase inverse**
  - Hétérodimère RTp51-RTp66.
  - Seul le dimère est actif
  - Différence de la vitesse (facteur 10) de dimérisation de RT entre les souches HIV1 et HIV2
- **Données structurales sur Transcriptase inverse**
  - Structure RT-VIH 1 connue par cristallographie Dr Arnold (USA)
  - Pas de structure 3D de RT-VIH 2
- ***Comment expliquer cette différence au niveau structural?***

g.deleage@ibcp.fr

HIV-RT



α Helix

β sheet

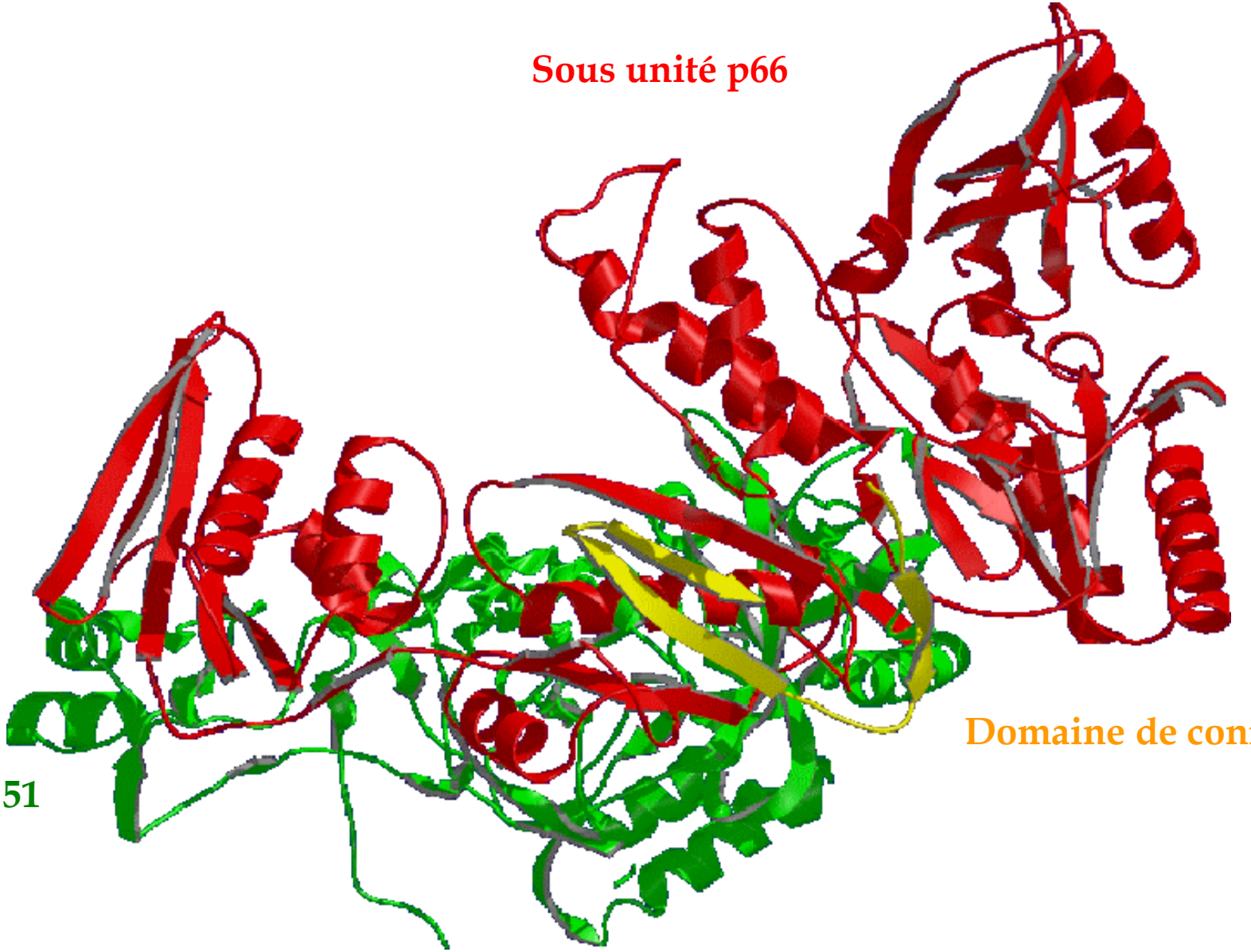
Difference in secondary structure

amino acid number

Divita, Rettinger, Geourjon, Deléage & Goody, J. Mol. Biol. (1995) 245, 508-521

g.deleage@ibcp.fr

Structure 3D de RT VIH 1



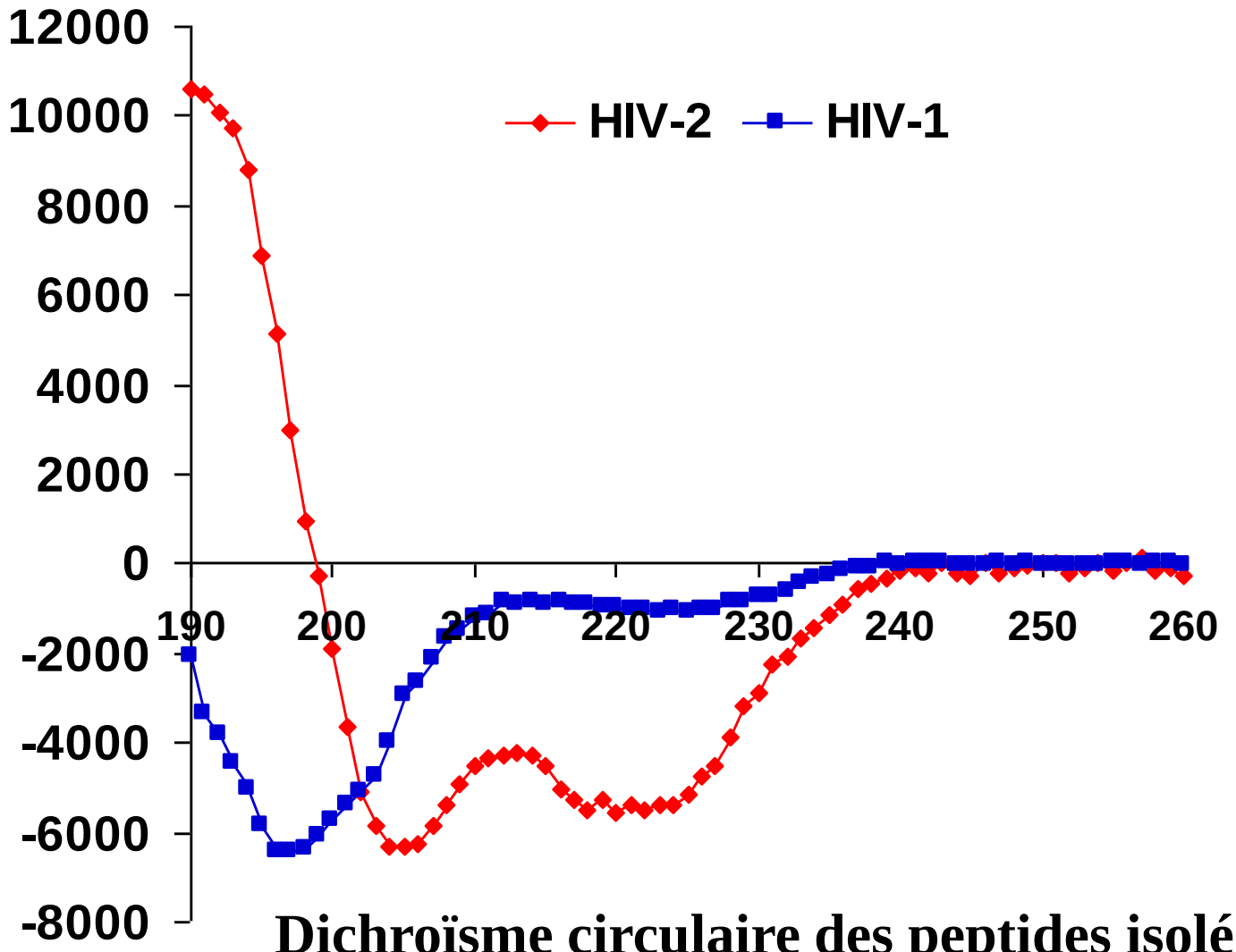
g.deleage@ibcp.fr

Synthèse et test de peptides de connexion VIH1 /VIH2

- **Synthèse de peptides pour les 2 souches dans la zone de connexion**
- **Expérience de compétition entre les sous-unités et les peptides**
  - Peptides isolés inhibent la dimérisation
- **PEPTIDES EN PHASE 1 DE TESTS PHARMACOLOGIQUES**
  - Brevets déposés
  - Etude structurale des peptides isolés

g.deleage@ibcp.fr

Structure 2D des peptides de connexion VIH1 /VIH2

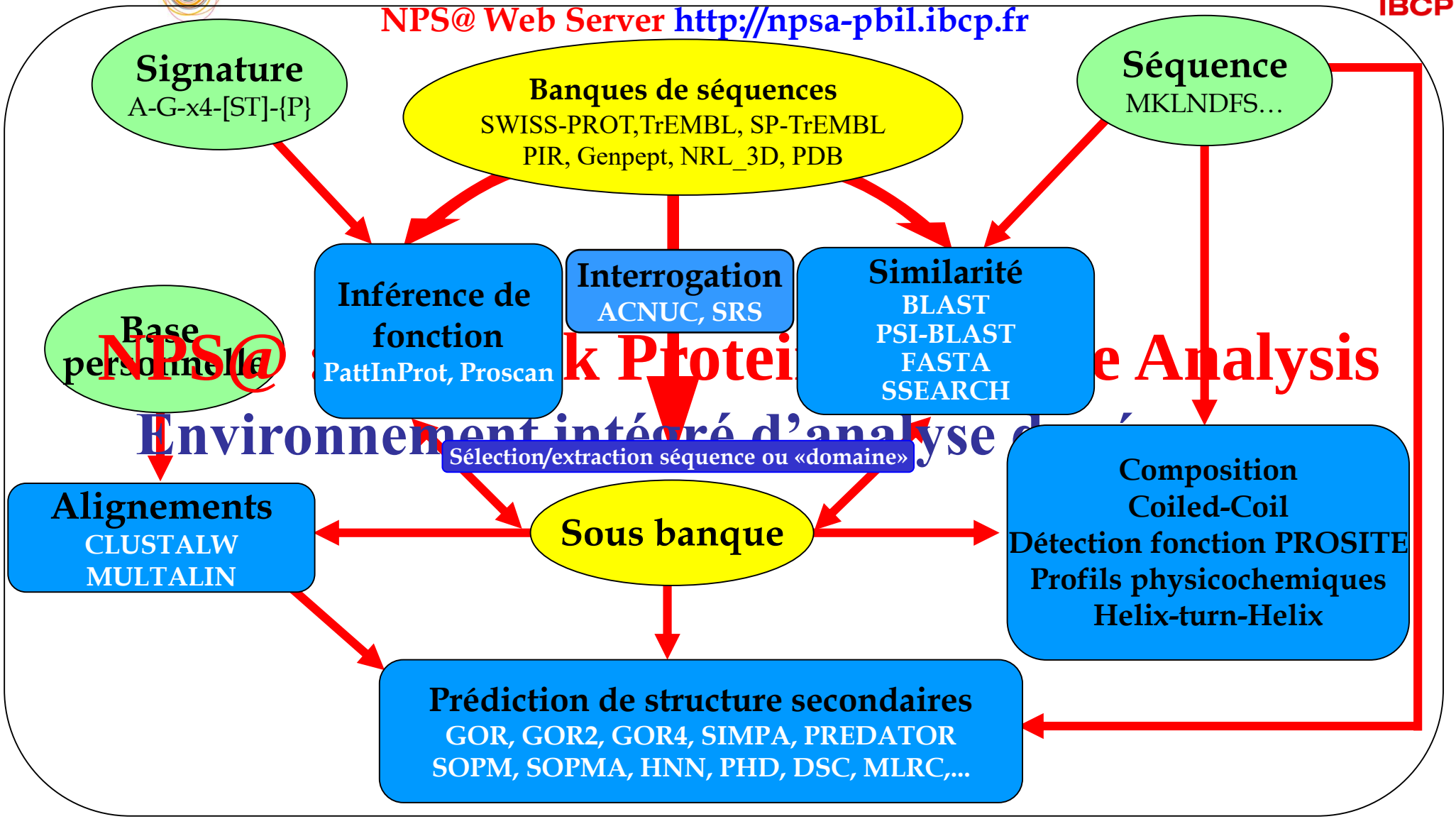


Dichroïsme circulaire des peptides isolés

HIV2 est en hélice

HIV1 n'est pas en hélice

g.deleage@ibcp.fr



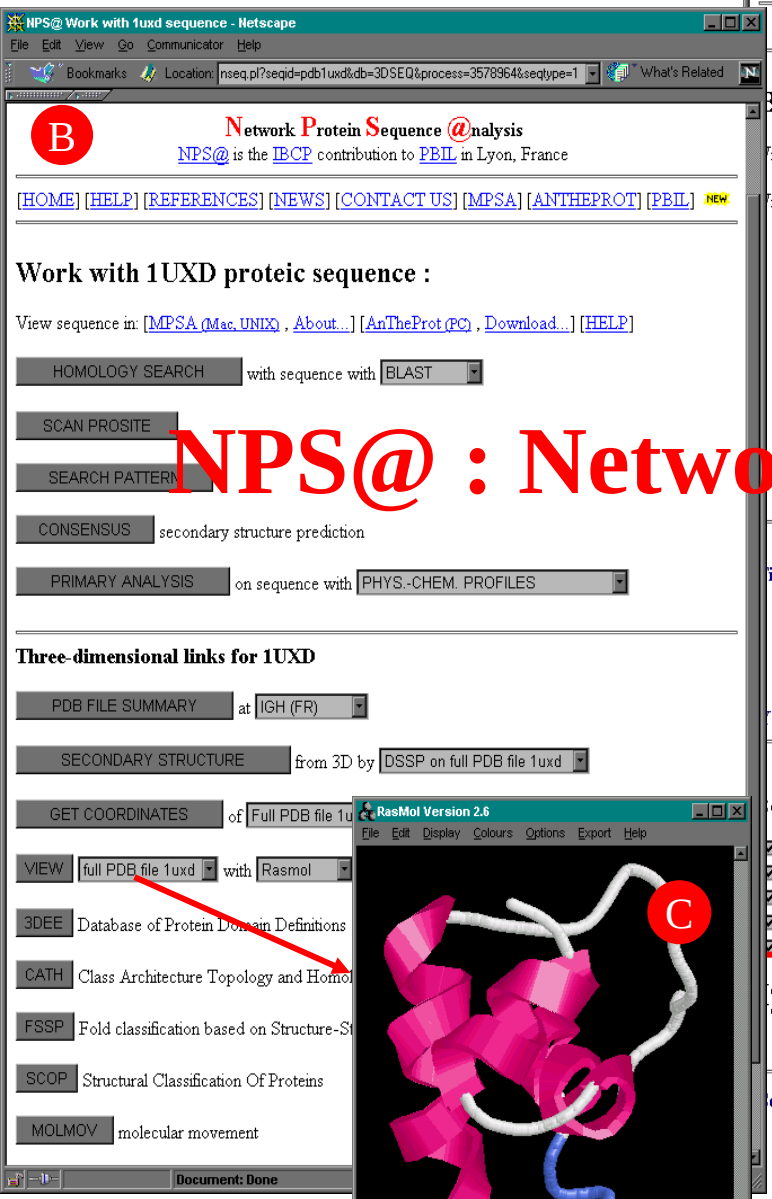
INTERNET

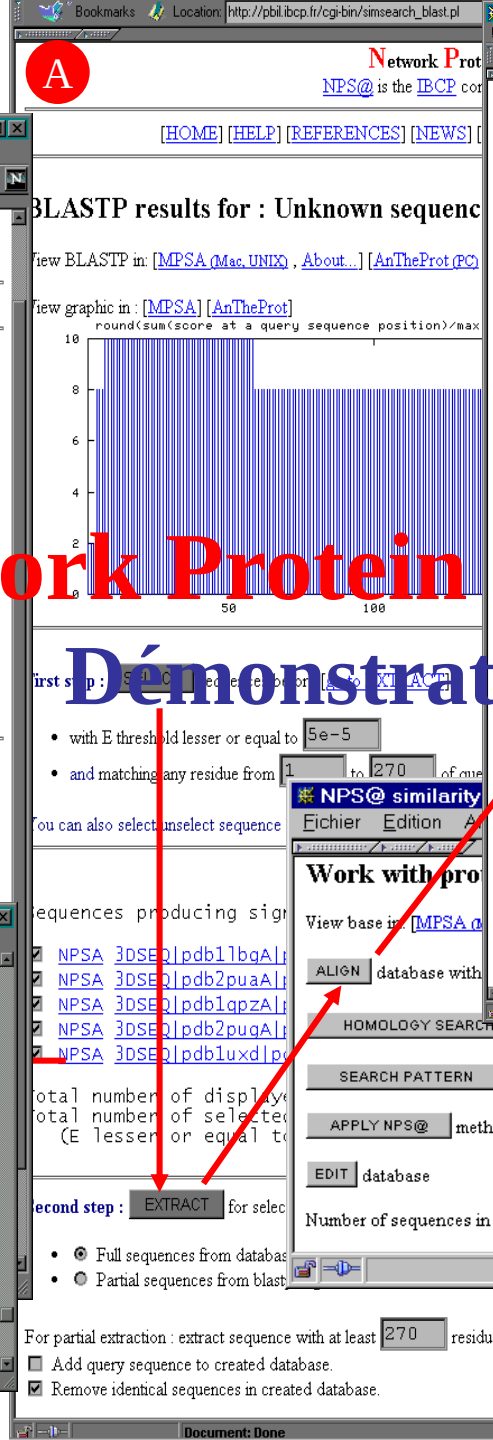
Logiciels clients/serveurs : ANTHEPROT, MPSA

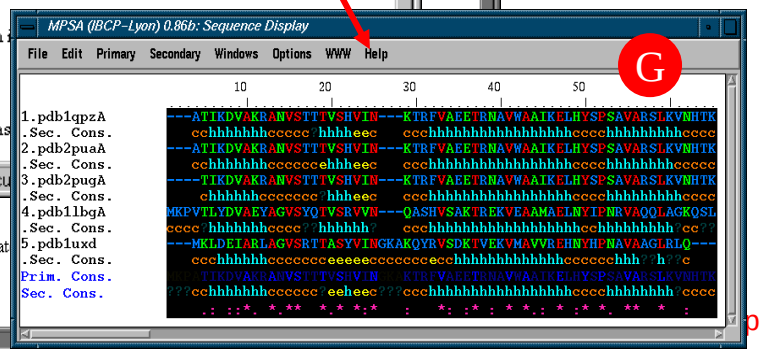
g.deleage@ibcp.fr

NPS@ : Network Protein Sequence Analysis

Démonstration







Structures 3D

g.deleage@ibcp.fr

Fichiers PDB
http://www.rcsb.org/pdb

- Archive mondiale des données structurales des macromolécules biologiques
- Créée en 1971 aux Brookhaven National Laboratories : 7 structures
- Depuis 1999, la PDB est sous la responsabilité du RCSB (Research Collaboratory for Structural Bioinformatics)
  - Rutgers univeristy
  - San Diego Supercomputer Center (SDSC)
  - National Institute of Standards and Technology (NIST)
- Conséquences :
  - Utilisation des technologies les plus modernes
  - Capture efficace de données
  - Curation des données
  - Introduction d'un nouveau format : mmCIF (système relationnel de représentation des données)

g.deleage@ibcp.fr

Fichiers PDB format

Overview

The HEADER record uniquely identifies a PDB entry through the idCode field. This record also provides a classification for the entry. Finally, it contains the date the coordinates were deposited at the PDB.

Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"HEADER"	
11 - 50	String(40)	classification	Classifies the molecule(s)
51 - 59	Date	depDate	Deposition date. This is the date the coordinates were received by the PDB
63 - 66	IDcode	idCode	This identifier is unique within PDB

Example

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
HEADER    COMPLEX (PROTEASE/INHIBITOR)              02-OCT-97   1AWF
```


g.deleage@ibcp.fr

Fichiers PDB format atomique

Overview

The ATOM records present the atomic coordinates for standard residues. They also present the occupancy and temperature factor for each atom. Heterogen coordinates use the HETATM record type. The element symbol is always present on each ATOM record; segment identifier and charge are optional.

Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"ATOM "	
7 - 11	Integer	serial	Atom serial number.
13 - 16	Atom	name	Atom name.
17	Character	altLoc	Alternate location indicator.
18 - 20	Residue name	resName	Residue name.
22	Character	chainID	Chain identifier.
23 - 26	Integer	resSeq	Residue sequence number.
27	AChar	iCode	Code for insertion of residues.
31 - 38	Real(8.3)	x	Orthogonal coordinates for X in Angstroms.
39 - 46	Real(8.3)	y	Orthogonal coordinates for Y in Angstroms.
47 - 54	Real(8.3)	z	Orthogonal coordinates for Z in Angstroms.
55 - 60	Real(6.2)	occupancy	Occupancy.
61 - 66	Real(6.2)	tempFactor	Temperature factor.
73 - 76	LString(4)	segID	Segment identifier, left-justified.
77 - 78	LString(2)	element	Element symbol, right-justified.
79 - 80	LString(2)	charge	Charge on the atom.

g.deleage@ibcp.fr

Fichier PDB des protéines


```
REMARK FILENAME="V_MINI_5.PDB"
ATOM      1  CA  MET     1      -0.399  20.462 -11.874  1.00  0.00
ATOM      2  HA  MET     1      -0.827  19.859 -11.089  1.00  0.00
ATOM      3  CB  MET     1       0.010  19.571 -13.053  1.00  0.00
ATOM      4  HB1 MET     1       1.081  19.428 -13.041  1.00  0.00
ATOM      5  HB2 MET     1      -0.279  20.044 -13.980  1.00  0.00
ATOM      6  CG  MET     1      -0.683  18.212 -12.934  1.00  0.00
ATOM      7  HG1 MET     1      -1.020  17.895 -13.910  1.00  0.00
ATOM      8  HG2 MET     1      -1.532  18.296 -12.271  1.00  0.00
ATOM      9  SD  MET     1       0.483  16.996 -12.273  1.00  0.00
ATOM     10  CE  MET     1      -0.650  16.201 -11.107  1.00  0.00
ATOM     11  HE1 MET     1      -0.128  15.414 -10.579  1.00  0.00
ATOM     12  HE2 MET     1      -1.484  15.778 -11.644  1.00  0.00
ATOM     13  HE3 MET     1      -1.014  16.936 -10.402  1.00  0.00
ATOM     14  C   MET     1       0.810  21.246 -11.346  1.00  0.00
ATOM     15  O   MET     1       0.910  22.444 -11.535  1.00  0.00
ATOM     16  N   MET     1      -1.425  21.397 -12.428  1.00  0.00
ATOM     17  HT1 MET     1      -1.044  21.874 -13.270  1.00  0.00
ATOM     18  HT2 MET     1      -1.671  22.108 -11.709  1.00  0.00
ATOM     19  HT3 MET     1      -2.277  20.863 -12.692  1.00  0.00
ATOM     20  N   ASP     2       1.727  20.570 -10.686  1.00  0.00
ATOM     21  HN  ASP     2       1.615  19.607 -10.554  1.00  0.00
ATOM     22  CA  ASP     2       2.943  21.252 -10.133  1.00  0.00
ATOM     23  HA  ASP     2       3.567  20.537  -9.623  1.00  0.00
ATOM     24  CB  ASP     2       3.676  21.804 -11.352  1.00  0.00
ATOM     25  HB1 ASP     2       3.322  22.800 -11.564  1.00  0.00
ATOM     26  HB2 ASP     2       3.477  21.165 -12.195  1.00  0.00
```


g.deleage@ibcp.fr

Visualisation, construction, optimisation de structures 3D

- ✓ Fil de fer, sphères ombrées, carbones α, chaîne principale, surfaces, «cartoons»
- ✓ Sélection par type d'atomes, acides aminés, chaîne, segment, SS-bonds, Hbonds
- ✓ Codage de couleur associé : atomes, chaîne, propriétés, etc..
- ✓ Etiquettes (atomes, acides aminés, ligands)

Différents logiciels de modélisation

✓ Rasmol	multi-plateforme	http://www.umass.edu/microbio/rasmol/
✓ AntheProt	Windows	http://antheprot-pbil.ibcp.fr
✓ ViewerLite	MacOS et Windows	http://www.accelrys.com
✓ Swiss-PDB viewer	Windows, Linux, MacOS, IRIX	http://www.expasy.ch/spdbv
✓ PyMol	multi-plateforme	http://pymol.sourceforge.net/
✓ VMD	Windows, Unix, MacOS	http://www.ks.uiuc.edu/Research/vmd/
✓ Modeller	Web	http://salilab.org/modeller/modeller.html
✓ Geno3D	Web	http://geno3d-pbil.ibcp.fr
✓ SuMo	Web	http://sumo-pbil.ibcp.fr

Géométrie, Ramachandran, chiralités

- ✓ Distances, angles, Φ,Ψ
- ✓ Diagrammes de Ramachandran
- ✓ Empêchements stériques
- ✓ Affichage des voisins
- ✓ Affichage des liaisons hydrogènes

Comparaison/superposition et évaluations des structures

- ✓ Mesure du RMSD (Ecart quadratique moyen des distances entre atomes)
- ✓ Superposition LOCALE ou GLOBALE
  - ✓ CE = Combinatorial Extention — http://cl.sdsc.edu/ce
  - ✓ CL = Compound Likeness — http://cl.sdsc.edu/cl
  - ✓ VAST (Vector Alignment Search Tool) — http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml
- ✓ Evaluation Procheck — http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html
- ✓ Whatif — http://www.cmbi.kun.nl/whatif/

Classification des structures

- ✓ CATH — http://www.biochem.ucl.ac.uk/bsm/cath_new
- ✓ SCOP — http://scop.mrc-lmb.cam.ac.uk/scop/
- ✓ FSSP/HSSP — http://www.ebi.ac.uk/dali/fssp/

Mémoire associative

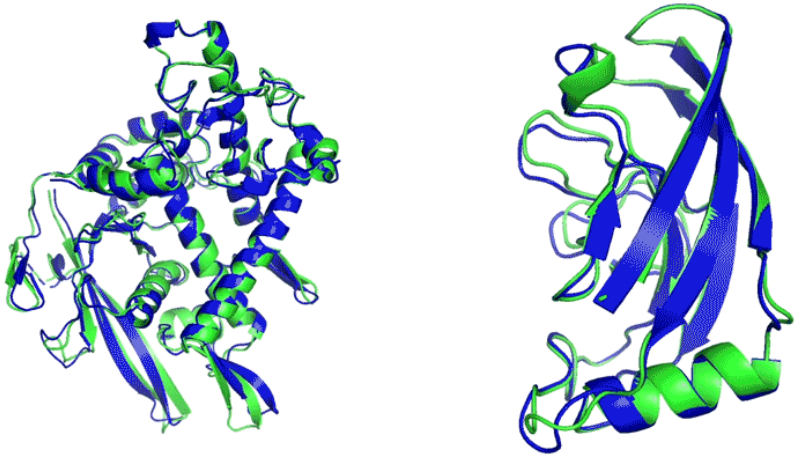
LA CDECANE DE CET OTRIRAOO SEBLME
ECVSSEIXE
PUOR UN CEITVALICNSE NOTEPYHE.

Le réseau de neurone →

Réseaux de neurones
Machine learning
Méthode d'apprentissage
Intelligence articielle
Deep learning

↓

https://deepmind.com/ ←



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

30/11/2020

g.deleage@ibcp.fr

Organisation modulaire des protéines

Type		Libellé	Nature	Coef.
CT	Contrôle Terminal	CT : Initiation Bio-Inform structurale	Ecrit session 1 / Ecrit session 2	2.1
CT	Contrôle Terminal	CT : Initiation Bio-Inform structurale	TP session 1 / Ecrit session 2	0.9

Initiation à La Bio-Informatique Structurale : DISTANCIEL
BCH3005L- ECRIT + EPREUVE TP

14/01/2021:14h00 Durée :

Ecrit Initiation à La Bio-Informatique Structurale : 14H00 - 15H30
Epreuve de TP Initiation à La Bio-Informatique Structurale 16H00 - 17H30

g.deleage@ibcp.fr

Questions?