

# Introduction à la bioinformatique

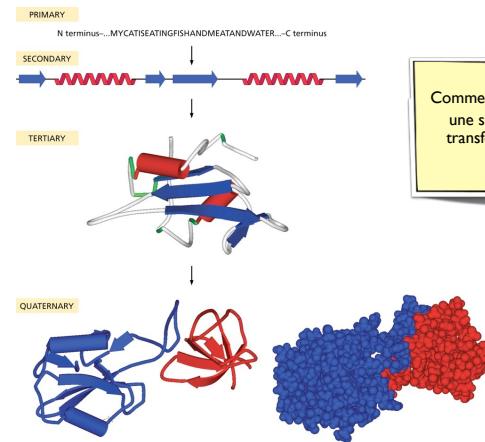
Prédiction de la structure secondaire

1

## La biologie structurale

3-1

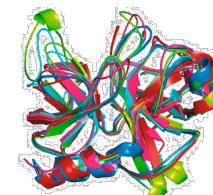
## La structure d'une protéine



Comment est-ce qu'on peut prédire si une séquence d'acides aminés se transforme en  $\alpha$ -hélice,  $\beta$ -brin ou boucle?

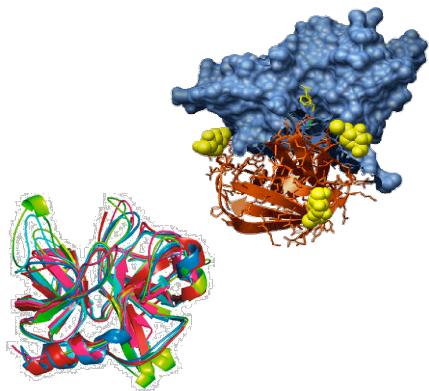
2

## La biologie structurale



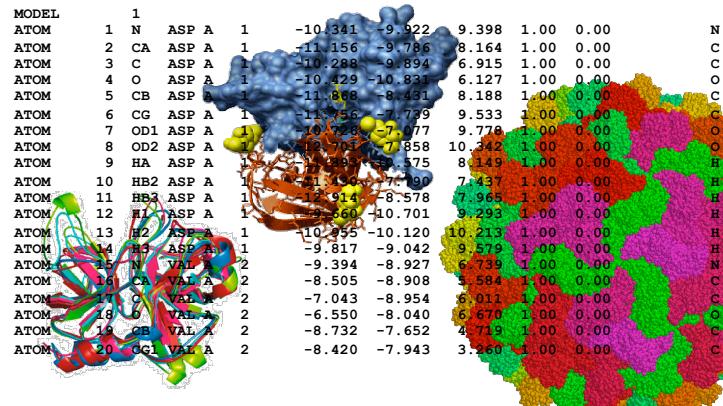
3-2

## La biologie structurale



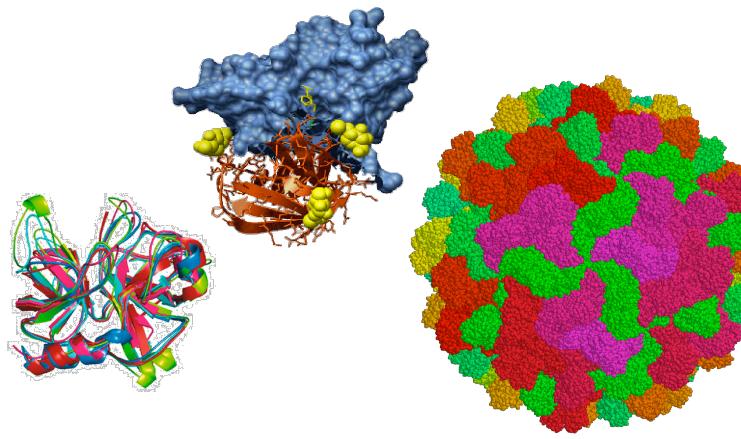
3-3

## La biologie structurale



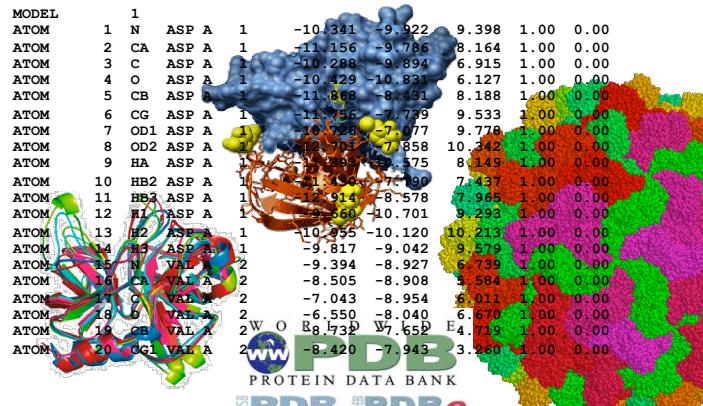
3-5

## La biologie structurale



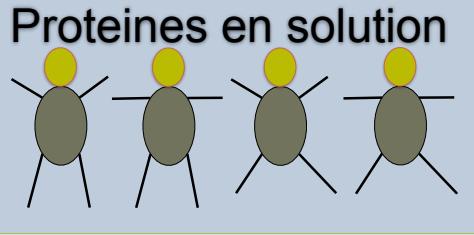
3-4

## La biologie structurale



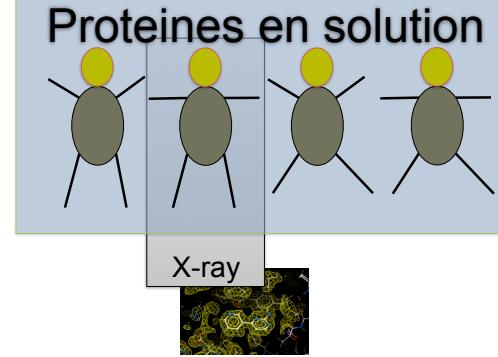
3-6

## La biologie structurale 2



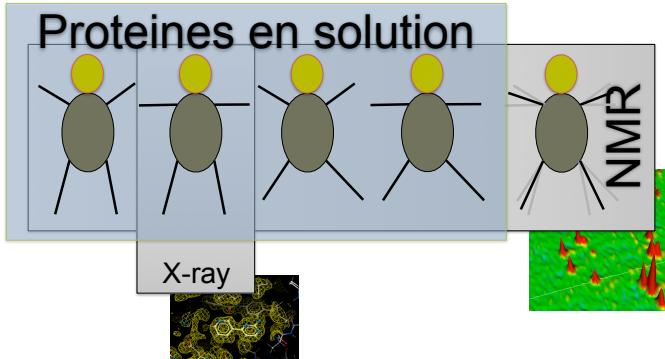
4-1

## La biologie structurale 2



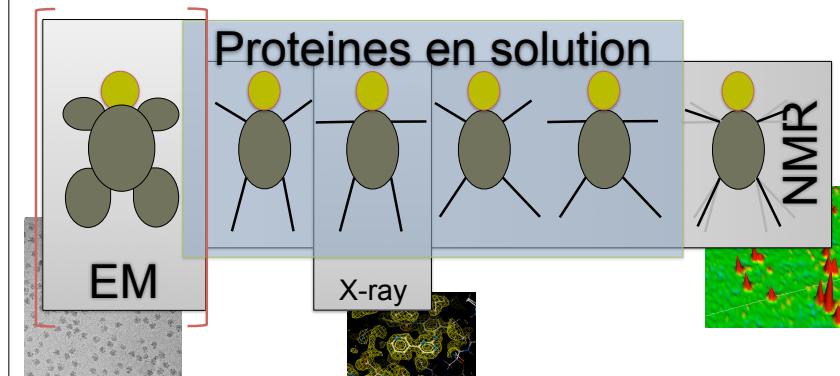
4-2

## La biologie structurale 2



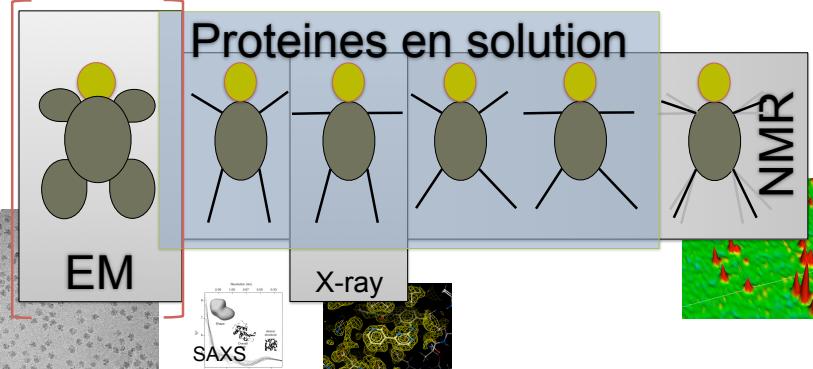
4-3

## La biologie structurale 2



4-4

## La biologie structurale 2



4-5

... les structures ...

Entry	Entry name	Protein names	Gene names	Organism	Length
P10144	GRAB_HUMAN	Granzyme B	GZMB, CGLI, CSPB, CTLA1, GRB	Homo sapiens (Human)	247
Q06141	REG3A_HUMAN	Regenerating islet-derived protein ...	REG3A, HIP, PAP, PAPI	Homo sapiens (Human)	175
P25685	DNJB1_HUMAN	DnaJ homolog subfamily B	DNAJB1, DNAJ1, HDJ1, HSPF1	Homo sapiens (Human)	340
P42694	HEL2_HUMAN	Probable helicase with zinc fingers	HEL2, DRHC, KIAA0054	Homo sapiens (Human)	1,042
P31689	DNJ1_HUMAN	DnaJ homolog subfamily A member 1	DNAJA1, DNAJ2, HDJ2, HSJ2, HSPF4	Homo sapiens (Human)	397
O95273	CCDB1_HUMAN	Cyclin-D1-binding protein 1	CCNDBP1, DIP1, GCIP, HMM	Homo sapiens (Human)	360
O60271	JIP4_HUMAN	C-Jun-amino-terminal kinase-interac...	SPAG9, HSS, KIAA0516, MAPK8IP4, SYD1, HLC6	Homo sapiens (Human)	1,321
Q9UB54	DJB1L_HUMAN	DnaJ homolog subfamily B member 11	DNAJB11, ED1, ER13, HD19, PSEC0121, UNQ537/PRO108	Homo sapiens (Human)	358

6

## Les séquences ...

Entry	Entry name	Protein names	Gene names	Organism	Length
P10144	GRAB_HUMAN	Granzyme B	GZMB, CGLI, CSPB, CTLA1, GRB	Homo sapiens (Human)	247
Q06141	REG3A_HUMAN	Regenerating islet-derived protein ...	REG3A, HIP, PAP, PAPI	Homo sapiens (Human)	175
P25685	DNJB1_HUMAN	DnaJ homolog subfamily B	DNAJB1, DNAJ1, HDJ1, HSPF1	Homo sapiens (Human)	340
P42694	HEL2_HUMAN	Probable helicase with zinc fingers	HEL2, DRHC, KIAA0054	Homo sapiens (Human)	1,042
P31689	DNJ1_HUMAN	DnaJ homolog subfamily A member 1	DNAJA1, DNAJ2, HDJ2, HSJ2, HSPF4	Homo sapiens (Human)	397
O95273	CCDB1_HUMAN	Cyclin-D1-binding protein 1	CCNDBP1, DIP1, GCIP, HMM	Homo sapiens (Human)	360
O60271	JIP4_HUMAN	C-Jun-amino-terminal kinase-interac...	SPAG9, HSS, KIAA0516, MAPK8IP4, SYD1, HLC6	Homo sapiens (Human)	1,321
Q9UB54	DJB1L_HUMAN	DnaJ homolog subfamily B member 11	DNAJB11, ED1, ER13, HD19, PSEC0121, UNQ537/PRO108	Homo sapiens (Human)	358

5

Séquences vers structures

• Comment est-ce qu'on détermine la structure secondaire ou tertiaire d'une séquence?

– Méthodes expérimentales!

• NMR, X-ray, ...

7-1

## Séquences vers structures

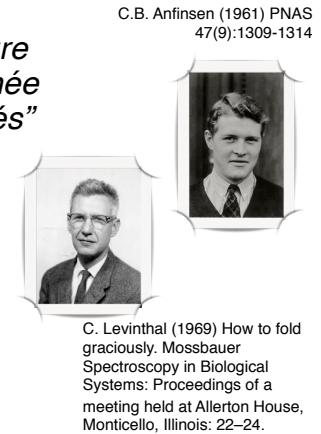
- Comment est-ce qu'on détermine la structure secondaire ou tertiaire d'une séquence?

–Méthodes expérimentales!  
•NMR, X-ray, ...

7-2

## Séquences vers structures 2

- Le dogme d'Anfinsen : "La structure native d'une protéine est déterminée par la séquence des acides aminés"  
– comment?
- Le paradoxe de Levinthal:
  - une protéine de 100 AA
  - 198 angles phi/psi différents
  - 3 conformations pour chaque angle
  - $3^{198}$  possibilités!
- l'échantillonnage séquentiel est impossible



8

## Séquences vers structures

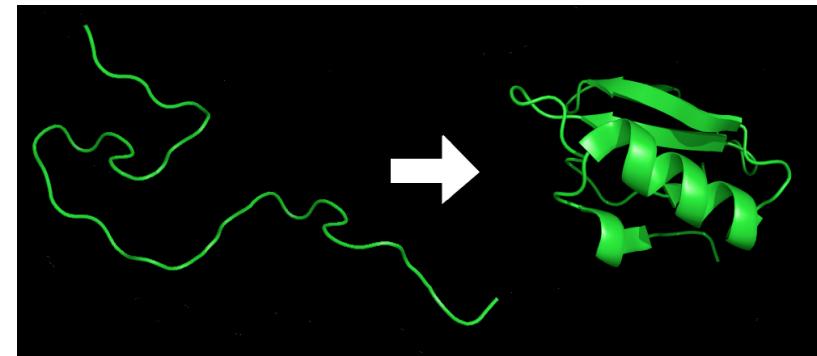
- Comment est-ce qu'on détermine la structure secondaire ou tertiaire d'une séquence?

–Méthodes expérimentales!  
•NMR, X-ray, ...

- Si les données expérimentales ne sont pas disponibles?
  - Prédiction en utilisant les structures qui sont déjà connues → **problème de classification**
    - Détermine si le résidu appartient à la classe *hélix*, *bêta-brin* ou *boucle*

7-3

## Séquences vers structures 3



**PARADOXE:** une protéine se plie dans quelques millisecondes

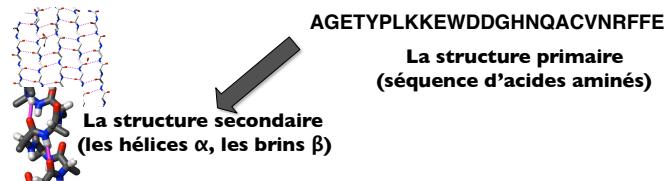
9-1

## Séquences vers structures 3



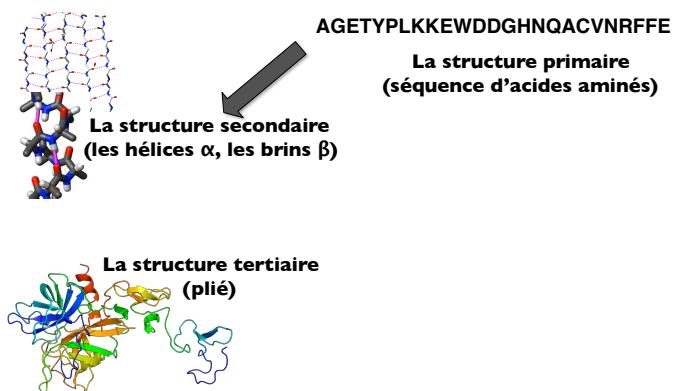
9-2

## Problème de classification



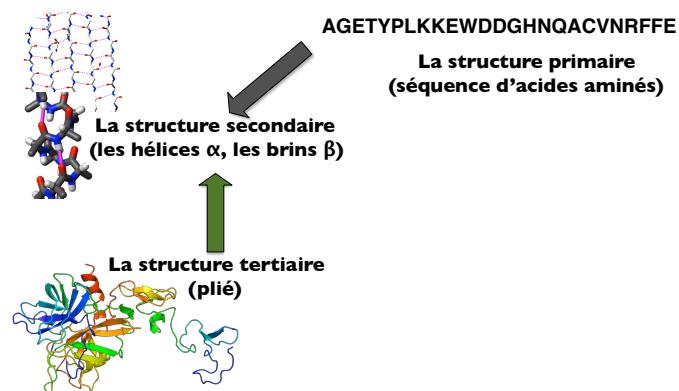
10-1

## Problème de classification



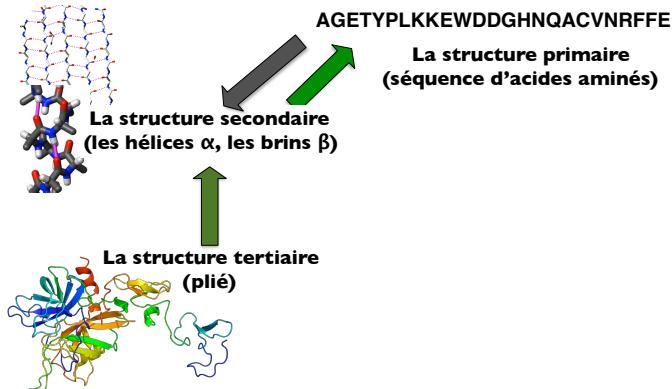
10-2

## Problème de classification



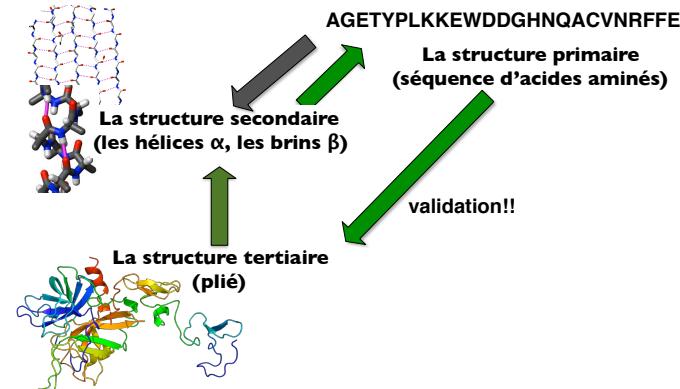
10-3

## Problème de classification



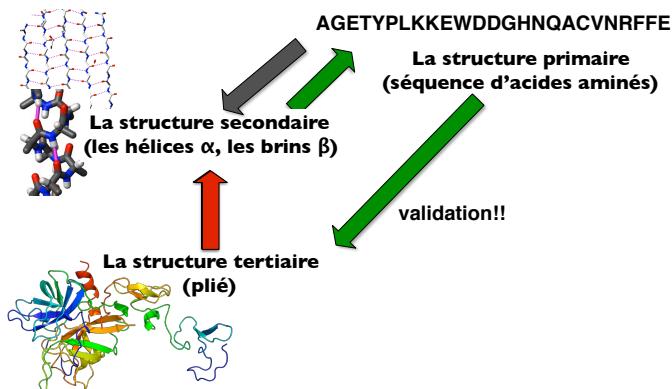
10-4

## Problème de classification



10-5

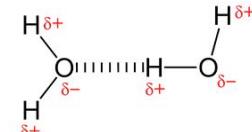
## Problème de classification



10-6

## Tertiaire à secondaire

- Les liens hydrogènes
  - Un hydrogène sur un atome électronégatif (comme N et O) a une charge partiellement positive

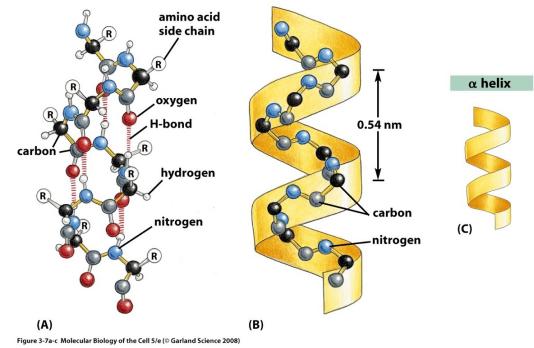


- Il peut créer un lien hydrogène avec un autre

11

## Tertiaire à secondaire 2

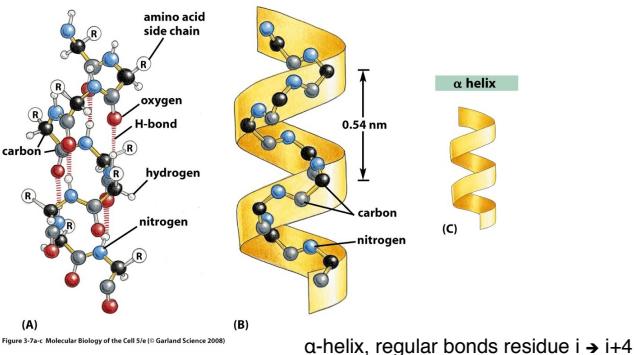
- Motifs de liaisons hydrogènes réguliers au sein des structures tertiaires



12-1

## Tertiaire à secondaire 2

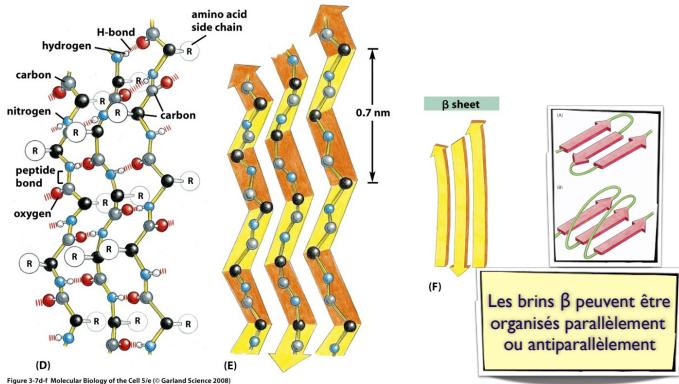
- Motifs de liaisons hydrogènes réguliers au sein des structures tertiaires



12-2

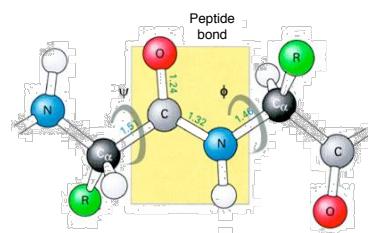
## Tertiaire à secondaire 3

- Motifs de liaisons hydrogènes réguliers au sein des structures tertiaires



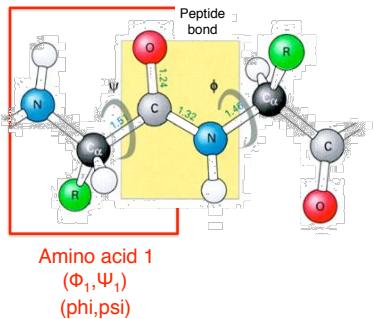
13

## Tertiaire à secondaire 4



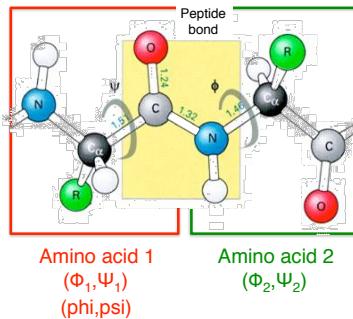
14-1

## Tertiaire à secondaire 4



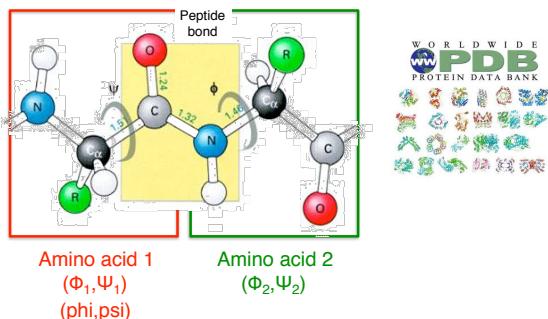
14-2

## Tertiaire à secondaire 4



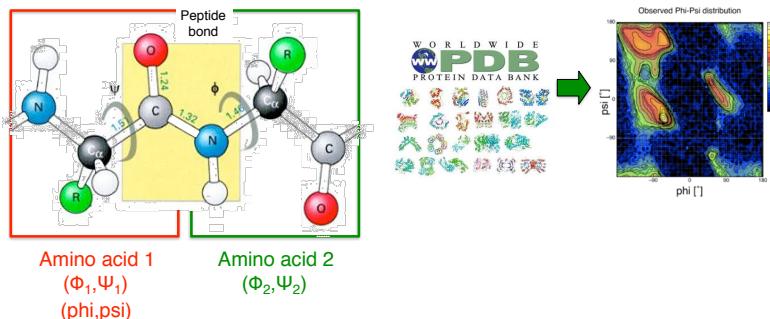
14-3

## Tertiaire à secondaire 4



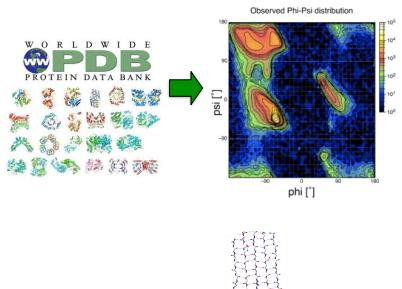
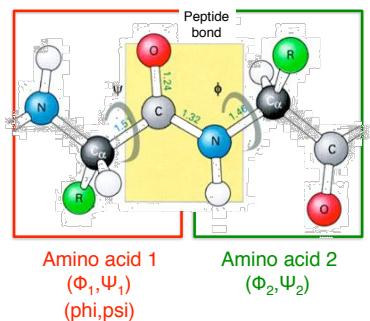
14-4

## Tertiaire à secondaire 4



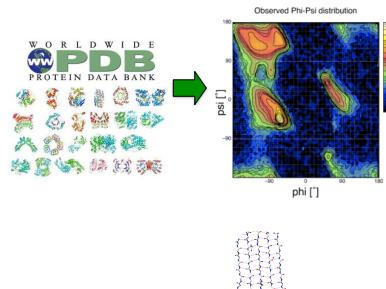
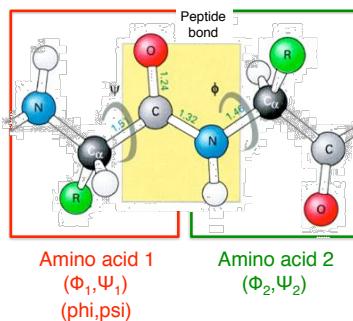
14-5

## Tertiaire à secondaire 4



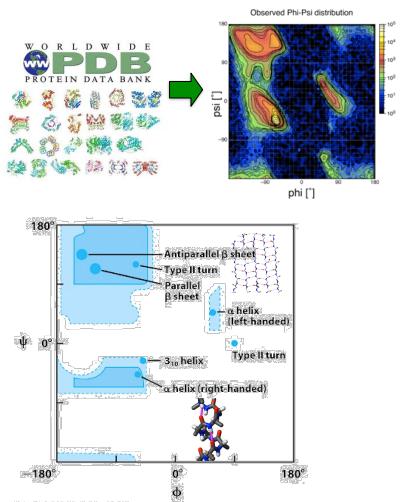
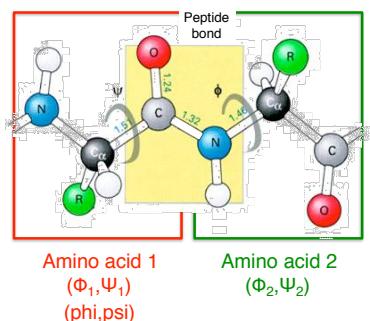
14-6

## Tertiaire à secondaire 4



14-7

## Tertiaire à secondaire 4



14-8

## Tertiaire à secondaire 5

- Define Secondary Structure of Proteins (DSSP)

– 1983, programme écrit en Pascal

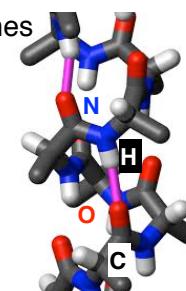
– Une définition simple pour les liens hydrogènes entre les atomes au sein de la chaîne principale

principale

$$E = 0.084 \left\{ \frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}} \right\} \cdot 332 \text{ kcal/mol}$$

•  $r$  est la distance entre ...

- N, H atomes de la chaîne principale du résidu i
- O, C atomes de la chaîne principale du résidu j

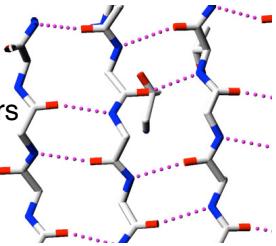


<http://swift.cmbi.ru.nl/gv/dssp/>

15

## Tertiaire à secondaire 6

- DSSP
  - Classification des acides aminés
  - Structure secondaire est déterminée en utilisant des motifs simples:
    - Hélice  $\alpha$  (**H**),  $i \rightarrow i+4$
    - $\beta$ -brin (**E**), un ensemble de plusieurs liaisons hydrogènes
    - Hélice  $3_{10}$  (**G**),  $i \rightarrow i+3$
    - Et autres ...



16

## Tertiaire à secondaire 7

17-1

## Tertiaire à secondaire 7

- Pour les méthodes prédictives on a besoin d'un ensemble de référence
  - Un ensemble de référence contient des protéines assez différentes en structures et séquences
    - moins de 30% d'identité entre les séquences
  - Structures de haute qualité
  - Des familles différentes

17-2

## Tertiaire à secondaire 8

18-1

## Tertiaire à secondaire 8

- Seulement sous-ensemble de toutes les protéines !!

18-2

## Tertiaire à secondaire 8

- Seulement sous-ensemble de toutes les protéines !!

- Les outils:

- DSSP

- Liaisons hydrogènes

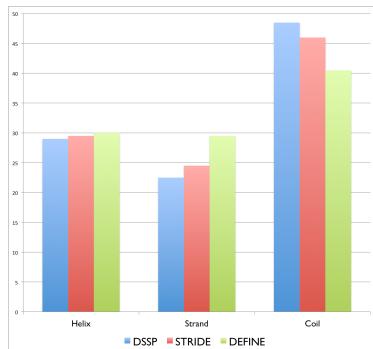
- STRIDE

- Liaisons hydrogènes et ( $\phi, \psi$ )

- DEFINE

- distance interatomique

- NOTION HUMAINE !



18-4

## Tertiaire à secondaire 8

- Seulement sous-ensemble de toutes les protéines !!

- Les outils:

- DSSP

- Liaisons hydrogènes

- STRIDE

- Liaisons hydrogènes et ( $\phi, \psi$ )

- DEFINE

- distance interatomique

- NOTION HUMAINE !

18-3

## Tertiaire à secondaire 9

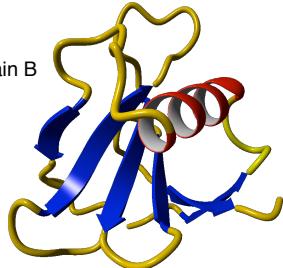
```
>luci_B; molId:1; molType:protein; unp:P05798; molName:Guanyl-specific r...  
DTSCTVCLSLPPEATDTLNLIASDGPFPPYSQDGVVFQNRESVLPTQSYGYYHEYTVITPGARTRGTRRIITGEA  
TQEDYYTGHDHYATFSLIDQTC
```

19-1

## Tertiaire à secondaire 9

>luci\_B; molId:1; molType:protein; unp:P05798; molName:Guanyl-specific r...  
DTSGTVCLSA LPPEATDTLN LIASDGPFY SQDGVVFQNR ESLVPTQSYG YYHEYTVITP  
TQEDYYTGHDHYATFS LIDQTC

PDB code 1uci, chain B



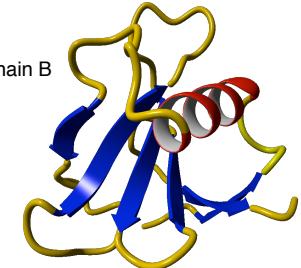
19-2

## Tertiaire à secondaire 9

>luci\_B; molId:1; molType:protein; unp:P05798; molName:Guanyl-specific r...  
DTSGTVCLSA LPPEATDTLN LIASDGPFY SQDGVVFQNR ESLVPTQSYG YYHEYTVITP  
TQEDYYTGHDHYATFS LIDQTC

PDB code 1uci, chain B

Regions  
Secondary structure  
Yellow: Loop  
Blue: Strand  
Red: Helix

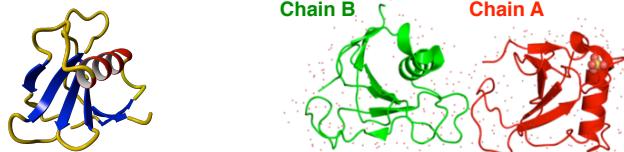


1 DTSGTVCLSA LPPEATDTLN LIASDGPFY SQDGVVFQNR ESLVPTQSYG YYHEYTVITP  
61 GARTRGTRRI ITGEATQEDY YTGDHYATFS LIDQTC

19-3

## Tertiaire à secondaire 10

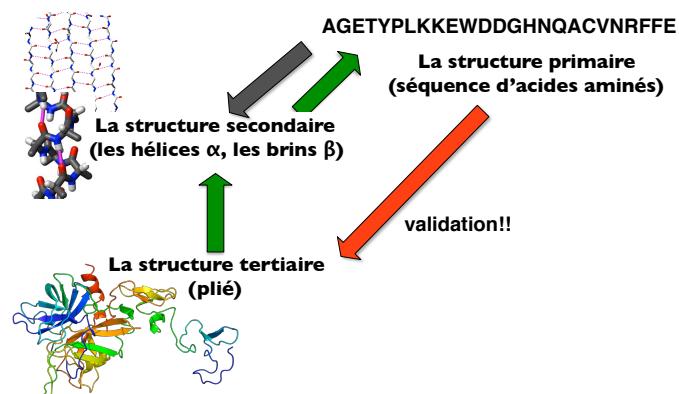
- les chaînes de protéines enregistrées dans des fichiers PDB représentent une seule protéine ...



- L'organisation de plusieurs chaînes ne correspond pas nécessairement à

20

## Problème de classification



21

## La Qualité des prédictions

- Validation
  - Confirmation de la qualité de la méthode
  - **Avant la construction du modèle prédictif**
  - Séparation des données en: une collection d'entraînement et une collection de test
    - utilise des données que la méthode ne connaît pas
  - **Contre-vérification (cross-validation)**
    - Divise la collection d'entraînement dans un nombre de sous-ensembles (p.e. K=10)
    - Entraîne le modèle prédictif en utilisant K-1 sous ensembles
    - Teste la qualité du modèle sur le sous-ensemble restant
    - La qualité du modèle correspond à la qualité moyenne sur tous les sous-ensembles.

22

## La Qualité des prédictions 3

$Q_3 = \frac{N_{\text{residues\_correctly\_predicted}}}{N_{\text{residues\_total}}}$	
Sequence secondaire	THISISIMYNEWPEPTIDESEQUENCE HHHHHCCCCEEEECCCEEECCCCHHHH (Hélice H, Brin E, C coil)
Prédiction 1	CHHHCCCCEEEECCCEEECCCCHHHHC 18 correct, $Q_3 = 69.0\%$
Prédiction 2	HHHHCCCCHHHHCCCCHHCCCCHHHH 18 correct, $Q_3 = 69.0\%$ , mais Mauvaise prédiction Tous les β brins sont identifiés comme hélices α

24

## La Qualité des prédictions 2

- mesure  $Q_3$ 
$$Q_3 = \frac{N_{\text{residues\_correctly\_predicted}}}{N_{\text{residues\_total}}}$$
  - Prédiction aléatoire  $Q_3=1/3$  (1/3 α, 1/3 β, 1/3 coil)
  - Prédiction aléatoire avec des vraies données  $Q_3 \approx 0.38$ 
    - Les fréquences de chaque structure secondaire ne sont pas les mêmes.
  - Mauvaises prédictions donnent aussi des scores plus grands que le score aléatoire
  - Score calculé pour chaque séquence !

23

## La Qualité des prédictions 4

- Quand on fait une *classification* on peut avoir les situations suivantes:
  - **Vrai positif** (true positive - TP) = un acide aminé qui fait partie d'une **hélice**, est assigné à la classe **hélice**
  - **Vrai négatif** (true negative - TN) = un acide aminé qui appartient à un **bêta-brin** est assigné à la classe **bêta-brin**.
  - **Faux positif** (false positive - FP) = **erreur de type 1**= un acide aminé qui appartient à un **bêta-brin** est assigné à la classe **hélice**.
  - **Faux négatif** (false negative - FN) = **erreur de type 2** = un acide aminé qui appartient à une **hélice** est assigné à la classe **bêta-brin**.

25

## La Qualité des prédictions 5

- TP, TN, FP et FN peuvent être regroupés dans une matrice de confusion

		prédictions	
		prédiction positive	prédiction négative
Vraies conditions	condition positive	TP	FN
	condition négative	FP	TN

26

## La Qualité des prédictions 6

- Prédicteurs produisent régulièrement une valeur continue
  - Ordonnez les éléments par score de prédiction décroissant/descendant ( $X_1 \geq X_2 \geq X_3 \geq \dots$ ), e.g. probabilité que l'acide aminé appartient à une hélice
  - On connaît la classe à laquelle l'élément appartient.
- TP, TN, FP et FN dépendent sur le seuil  $T$  qui détermine quand un élément (avec valeur  $X$ ) appartient à une classe
  - $X \geq T \rightarrow \text{positif}$
  - $X < T \rightarrow \text{négatif}$

28

## La Qualité des prédictions 5

- TP, TN, FP et FN peuvent être regroupés dans une matrice de confusion

		prédictions	
		prédiction positive	prédiction négative
Vraies conditions	condition positive	TP	FN
	condition négative	FP	TN

taux de vrai positif (TPR)  
ou sensibilité  $TP / (TP+FN)$

spécificité  $TN / (TN+FP)$

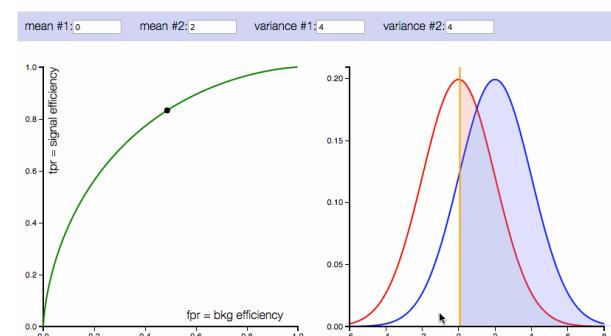
taux de faux positif (FPR) = 1 - spécificité =  $FP / (FP+TN)$

27

## La Qualité des prédictions 7

- Les prédictions pour des seuils  $T$  décroissants sont visualisées avec une courbe ROC (*Receiver Operating Characteristic*)

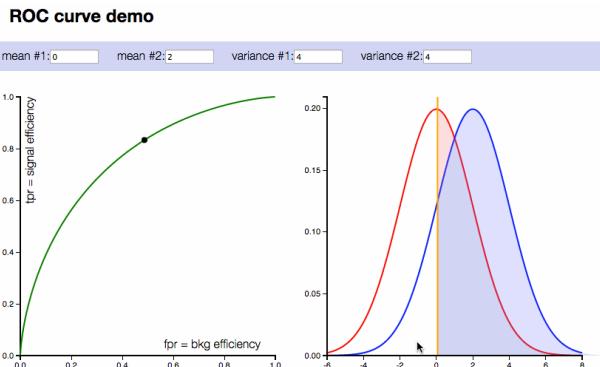
ROC curve demo



29-1

## La Qualité des prédictions 7

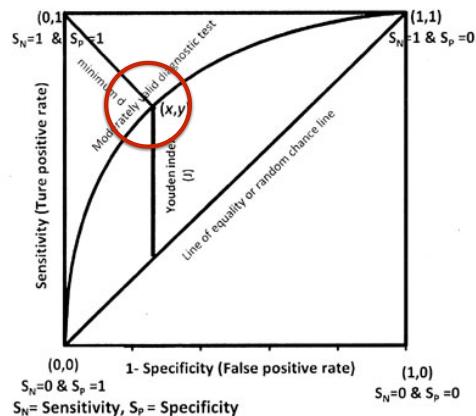
- Les prédictions pour des seuils T décroissants sont visualisées avec une courbe ROC (*Receiver Operating Characteristic*)



29-2

## La Qualité des prédictions 9

- La meilleur prédicteur?



31

## La Qualité des prédictions 8

- Un algorithme simple :
- input** = L, ensemble de test; f(i) fonction qui estime si i appartient à la classe positive;  
P et N le nombre d'exemples positifs et négatifs (P>0 et N>0)
- output** = R, la liste de points sur la courbe ROC en ordre FPR décroissant

```

1:  $L_d \leftarrow L$  decreasing order by
   f scores
2: FP  $\leftarrow$  TP  $\leftarrow$  0
3: R  $\leftarrow$  []
4:  $f_{prev} \leftarrow -\infty$ 
5: i  $\leftarrow$  1
6: while i  $\leq |L_d|$  do
7:   if  $f(i) \neq f_{prev}$  then
8:     push (FP/N, TP/P) onto R
9:      $f_{prev} \leftarrow f(i)$ 
10:  endif
11:  if  $L_d[i]$  is a positive example then
12:    TP  $\leftarrow$  TP+1
13:  else
14:    FP  $\leftarrow$  FP+1
15:  endif
16:  i  $\leftarrow$  i+1
17: endwhile
18: push (FP/N, TP/P) onto R
19: end

```

30

## La Qualité des prédictions 10

- Matthews correlation coefficient (MCC)*

- Prédiction pour chaque état secondaire
- Utilise faux/vrai positifs (FP/TP) et faux/vrai négatifs (FN/TN)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Actual SS	HHHHHCCCCCEEEECCCEEECCCCHHHHH
Prediction 1	CHHHCCCCCEEEECCCEEECCCCHHHHC H 6 TP, 1 FP, 12 TN, 3 FN MCC <sub>H</sub> = 0.623 E 5 TP, 2 FP, 13 TN, 2 FN MCC <sub>E</sub> = 0.581
Prediction 2	HHHHCCCCCCHHHHCCCCHHCCCCHHHH H 8 TP, 6 FP, 10 TN, 1 FN MCC <sub>H</sub> = 0.497 E 0 TP, 0 FP, 18 TN, 7 FN MCC <sub>E</sub> = error! (1 TP, 0 FP, 17 TN, 6 FN MCC <sub>E</sub> = 0.327)

32

## La Qualité des prédictions 11

33-1

## La Qualité des prédictions 11

- Mesure SOV (*Segment-overlap measure*)
  - SOV est basé sur la superposition moyenne entre le segment observé et le segment prévu par la méthode
  - Donne des punitions pour des séquences interrompues

33-2

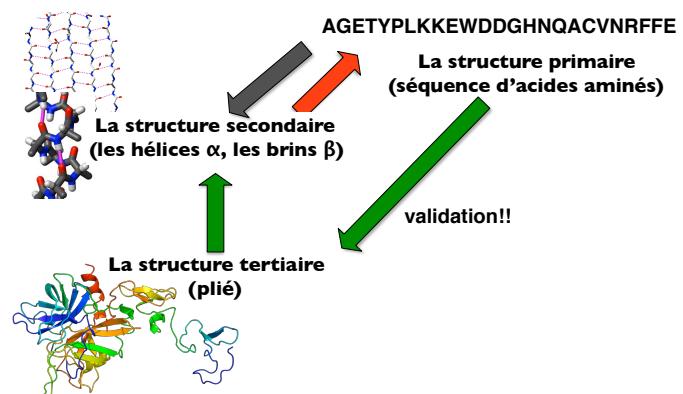
## La Qualité des prédictions 11

- Mesure SOV (*Segment-overlap measure*)
  - SOV est basé sur la superposition moyenne entre le segment observé et le segment prévu par la méthode
  - Donne des punitions pour des séquences inter

		Sov	$Q_3$
Observed	CHHHHHHHHHHHC		
Prediction 1	CHCHCHCHCHCC	12.5	58.3
Prediction 2	CCCCHHHHCCCC	63.2	58.3
Prediction 3	CHHHCHHHCHHC	40.6	83.3
Prediction 4	CHHCCHHHHHCC	52.3	75.0
Prediction 5	CCCCHHHHHCCC	80.6	66.7

33-3

## Problème de classification



34

# La Méthode Chou-Fasman

- Algorithme de Chou-Fasman (1974) Biochemistry 13(2): 211-222 et 222-

– Calculez pour chaque acide aminé la propension que cet acide aminé appartient à : E, H ou C

– Règles de prédictions

- Seulement 15 structures (X-ray) !!

TABLE I: Amino Acid Residues in the Helix, Inner Helix,  
β-Sheet, and Coil Regions of 15 Proteins.

Amino Acid	No. of Residues	Residues in Helix	Residues in Inner Helix	Residues in β Region	Residues in Coil Region
Ala	228	119	62	38	71
Arg	78	22	9	12	44
Asn	133	35	12	15	83
Asp	111	39	10	15	57
Cys	54	15	3	12	27
Gln	95	40	16	20	35
Glu	113	62	28	5	46
Gly	232	45	22	32	155
His	74	33	11	9	32
Ile	106	38	22	29	39
Leu	196	94	64	41	61
Lys	175	67	34	22	86
Met	28	12	6	8	8
Pho	82	33	16	18	31
Pro	85	18	0	9	58
Ser	202	57	24	25	120
Thr	156	47	21	32	77
Trp	44	18	10	9	17
Tyr	100	22	10	22	56
Val	181	74	44	51	56
Total	2473	890	424	424	1159

<sup>a</sup> The three helical end residues on both N- and C-termini of a helical region are omitted.

# La Méthode Chou-Fasman 2

- Les paramètres conformationnels

TABLE II: Frequency of Helical, Inner Helical,<sup>a</sup> β, and Coil Residues in 15 Proteins with Their Conformational Parameters  $P_{\alpha}$ ,  $P_{\beta}$ , and  $P_{\gamma}$ .

Amino Acid	$f_{\alpha}^{\text{a}}$ <sup>b</sup>	$P_{\alpha}^{\text{c}}$	$f_{\alpha}^{\text{i}}$ <sup>b</sup>	$P_{\alpha}^{\text{i}}$ <sup>c</sup>	$f_{\beta}^{\text{a}}$ <sup>b</sup>	$P_{\beta}^{\text{d}}$	$f_{\beta}^{\text{i}}$ <sup>b</sup>	$P_{\beta}^{\text{i}}$ <sup>c</sup>
Ala	0.522	1.45	0.272	1.59	0.167	0.97	0.311	0.66
Arg	0.282	0.79	0.115	0.67	0.154	0.90	0.564	1.20
Asn	0.263	0.73	0.090	0.53	0.113	0.65	0.624	1.33
Asp	0.351	0.98	0.090	0.53	0.137	0.80	0.514	1.09
Cys	0.278	0.77	0.056	0.33	0.222	1.30	0.500	1.97
Gln	0.421	1.17	0.168	0.98	0.211	1.23	0.568	0.79
Glu	0.549	1.53	0.248	1.45	0.044	0.26	0.407	0.87
Gly	0.490	0.53	0.091	0.53	0.138	0.81	0.668	1.42
His	0.446	1.24	0.149	0.87	0.122	0.71	0.432	0.92
Ile	0.358	1.00	0.208	1.22	0.274	1.60	0.368	0.78
Leu	0.480	1.34	0.327	1.91	0.204	1.22	0.311	0.66
Lys	0.383	1.07	0.194	1.13	0.126	0.74	0.491	1.05
Met	0.429	1.20	0.214	1.25	0.286	1.67	0.286	0.61
Phe	0.402	1.12	0.195	1.14	0.219	1.28	0.378	0.81
Pro	0.212	0.59	0	0	0.106	0.62	0.682	1.45
Ser	0.282	0.79	0.119	0.70	0.124	0.72	0.594	1.27
Thr	0.295	0.82	0.128	0.75	0.205	1.20	0.494	1.05
Trp	0.409	1.14	0.227	1.33	0.203	1.19	0.386	0.82
Tyr	0.220	0.61	0.100	0.58	0.220	1.29	0.560	1.19
Val	0.409	1.14	0.243	1.42	0.282	1.65	0.309	0.66

$$\langle f_{\alpha} \rangle^{\text{a}} = 0.339, \langle P_{\alpha} \rangle^{\text{c}} = 1.00, \langle f_{\alpha} \rangle^{\text{i}} = 0.171, \langle P_{\alpha} \rangle^{\text{i}} = 1.00, \langle f_{\beta} \rangle^{\text{a}} = 0.171, \langle P_{\beta} \rangle^{\text{c}} = 1.00, \langle f_{\beta} \rangle^{\text{i}} = 0.469, \langle P_{\beta} \rangle^{\text{i}} = 1.00$$

<sup>a</sup> The three helical end residues on both N- and C-termini of a helical region are omitted. <sup>b</sup>  $f_{\alpha}$ ,  $f_{\alpha}^{\text{i}}$ ,  $f_{\beta}$ , and  $f_{\beta}^{\text{i}}$  are respectively the frequency of residues in the helix, inner helix, β, and coil regions. <sup>c</sup>  $P_{\alpha}$ ,  $P_{\alpha}^{\text{i}}$ ,  $P_{\beta}$ , and  $P_{\beta}^{\text{i}}$  are respectively the conformational parameters for the helix ( $f_{\alpha}/\langle f_{\alpha} \rangle^{\text{a}}$ ), the inner helix ( $f_{\alpha}^{\text{i}}/\langle f_{\alpha} \rangle^{\text{i}}$ ), the β region ( $f_{\beta}/\langle f_{\beta} \rangle^{\text{a}}$ ), and the coil region ( $f_{\beta}^{\text{i}}/\langle f_{\beta} \rangle^{\text{i}}$ ). <sup>d</sup>  $\langle f_{\alpha} \rangle^{\text{a}}, \langle f_{\alpha} \rangle^{\text{i}}, \langle f_{\beta} \rangle^{\text{a}}, \langle f_{\beta} \rangle^{\text{i}}$  are respectively the average frequency of residues in helical, inner helical, β, and coil regions. <sup>e</sup>  $\langle P_{\alpha} \rangle^{\text{a}}, \langle P_{\alpha} \rangle^{\text{i}}, \langle P_{\beta} \rangle^{\text{a}}, \langle P_{\beta} \rangle^{\text{i}}$  are respectively the average conformational parameter for the helix, inner helix, β, and coil regions.

35

# La Méthode Chou-Fasman 3

- L'algorithme utilise ses propensions pour la prédition

Residue	$P_{\alpha}$	Residue	$P_{\beta}$
Glu	1.51	Val	1.70
Met	1.45	Hα	1.50
Ala	1.42	Ile	1.47
Leu	1.21	Tyr	1.38
Lys	1.16	Trp	1.37
Phe	1.13	Leu	1.30
Gln	1.11	Cys	1.19
Trp	1.08	Thr	1.19
Ile	1.08	Gln	1.10
Val	1.06	Met	1.05
Asp	1.01	Arg	0.93
His	1.00	Asn	0.89
Arg	0.96	His	0.87
Thr	0.83	Ala	0.83
Ser	0.77	Ser	0.75
Cys	0.70	Gly	0.75
Tyr	0.69	Lys	0.74
Asn	0.67	Pro	0.55
Pro	0.57	Asp	0.54
Gly	0.57	Glu	0.37

L'algorithme est composé de deux règles, une règle pour les hélices α et une autre règle pour les β-brins

H formation forte  
h formation faible  
i indifférent  
b briseur faible  
B briseur fort

37

# La Méthode Chou-Fasman 4

- Assignez chaque acide aminé la propension  $P_{\alpha}/P_{\beta}$

- Règles heuristiques<sup>1</sup>

– Hélices: Retrouvez les régions qui contiennent 4 des 6 résidus connectés (H<sub>a</sub> ou H<sub>b</sub>) avec  $P_{\alpha}$  moyenne > 1.03

– Les I (indifférent) comptent pour 0.5 respectivement – 3 h (ou H) et 2 I sont suffisants pour la formation d'une hélice

– Prolongez la région jusqu'à:

- ce que 4 résidus consécutifs aient une  $P_{\alpha}$  moyenne < 1.0
- Ou si le segment total est plus long que 6 résidus et  $P_{\alpha}$  moyenne >  $P_{\beta}$  moyenne, assignez la région comme hélice α

– Répétez cela pour la séquence complète

<sup>1</sup> les détails dans l'article Chou et Fasman (1974) Prediction of protein conformation. Biochemistry 13(2):222-245 sur la page 224

38

## La Méthode Chou-Fasman 5

- Formation Hélices :

Residue	P <sub>α</sub>	Residue	P <sub>β</sub>
Glu	1.51	Val	1.70
Met	1.45	Ile	1.60
Ala	1.42	Tyr	1.47
Leu	1.21	Phe	1.38
Lys	1.16	Trp	1.37
Phenylalanine	1.13	Leu	1.30
Gln	1.11	Cys	1.19
Trp	1.08	Thr	1.19
Ile	1.06	Gln	1.05
Val	1.06	Met	1.05
Asp	1.01	Arg	0.93
His	1.00	Asn	0.89
Arg	0.98	His	0.87
Thr	0.83	Ala	0.83
Ser	0.77	Ser	0.75
Cys	0.70	Gly	0.75
Ter	0.68	Lys	0.74
Asn	0.67	Phe	0.74
Pro	0.57	Asp	0.54
Gly	0.57	Glu	0.37

	P <sub>α</sub>	P <sub>β</sub>	a	b
T	0.82	1.2	i	0 1
S	0.79	0.72	i	0 2
G	0.53	0.81	B	0 2
T	0.82	1.2	i	0 3
V	1.14	1.65	H	1 4
C	0.77	1.3	i	0 3
L	1.34	1.22	H	1 3
S	0.79	0.72	b	0 1
A	1.45	0.97	i	0 1
E	1.53	0.26	H	1 2
T	0.82	1.2	i	0 1
D	0.98	0.8	i	0 1
Q	1.17	1.23	h	1 1
S	0.79	0.72	i	0 1
Y	0.61	1.29	h	1 3
G	0.53	0.81	i	0 2
Y	0.61	1.29	h	1 3
H	1.24	0.71	b	0 3
E	1.53	0.26	H	1 2
Y	0.61	1.29	h	1 3
T	0.82	1.2	h	1 1
D	0.98	0.8	i	0 2
Q	1.17	1.23	h	1 1
S	0.79	0.72	i	0 1
Y	0.61	1.29	h	1 3
G	0.53	0.81	i	0 2
Y	0.61	1.29	h	1 3
H	1.24	0.71	h	1 3
E	1.53	0.26	H	1 2
Y	0.61	1.29	h	1 5
T	0.82	1.2	h	1 4
V	1.14	1.65	H	1 3
I	1	1.6	H	1 2
T	0.82	1.2	h	1 1
P	0.59	0.62	b	0 0
G	0.53	0.81	i	0 0

39

## La Méthode Chou-Fasman 7

- Formation brins :

Residue	P <sub>α</sub>	Residue	P <sub>β</sub>
Glu	1.51	Val	1.70
Met	1.45	Ile	1.60
Ala	1.42	Tyr	1.47
Leu	1.21	Phe	1.38
Lys	1.16	Trp	1.37
Phenylalanine	1.13	Leu	1.30
Gln	1.11	Cys	1.19
Trp	1.08	Thr	1.19
Ile	1.08	Gln	1.05
Val	1.06	Met	1.05
Asp	1.01	Arg	0.93
His	1.00	Asn	0.88
Arg	0.98	His	0.87
Thr	0.83	Ala	0.83
Ser	0.77	Ser	0.75
Cys	0.70	Gly	0.75
Tyr	0.69	Lys	0.74
Asn	0.67	Phe	0.74
Pro	0.57	Asp	0.54
Gly	0.57	Glu	0.37

	P <sub>α</sub>	P <sub>β</sub>	b	<P>
T	0.82	1.2	i	0 1
S	0.79	0.72	b	0 3
G	0.53	0.81	B	0 2
T	0.82	1.2	h	1 4
V	1.14	1.65	H	1 3
C	0.77	1.3	i	0 3
L	1.34	1.22	h	1 2
S	0.79	0.72	b	0 1
A	1.45	0.97	i	0 1
L	1.34	1.22	h	1 1
P	0.59	0.62	b	0 1
P	0.59	0.62	b	0 1
E	1.53	0.26	H	1 3
A	1.45	0.97	H	1 2
T	0.82	1.2	i	0 1
D	0.98	0.8	i	0 2
Q	1.17	1.23	h	1 1
S	0.79	0.72	i	0 1
Y	0.61	1.29	h	1 3
G	0.53	0.81	i	0 2
Y	0.61	1.29	h	1 3
H	1.24	0.71	b	0 3
E	1.53	0.26	H	1 2
Y	0.61	1.29	h	1 5
T	0.82	1.2	h	1 4
V	1.14	1.65	H	1 3
I	1	1.6	H	1 2
T	0.82	1.2	h	1 1
P	0.59	0.62	b	0 0
G	0.53	0.81	i	0 0

41

## La Méthode Chou-Fasman 6

- β-brins

– Retrouvez les régions qui contiennent 3 sur 5 résidus connectés ( $h_\beta$  ou  $H_\beta$ ) avec  $P_\beta$  moyenne  $> 1.05$

• Les I (indifférent) comptent pour 0.5 respectivement

• Prolongez sur les deux extrémités jusqu'à :

– ce que 4 résidus consécutifs aient une  $\langle P_\beta \rangle < 1.0$

– Ou si le segment total est plus long que 5 résidus et  $P_\beta$  moyenne  $> P_\alpha$  moyenne, assignez la région comme β-brin

– Répétez cela pour la séquence complète

– s'il y a une superposition entre α et β, prenez la structure avec la plus grande propension

40

## La Méthode Chou-Fasman 8

P <sub>α</sub>	P <sub>β</sub>	a	b	<P <sub>α</sub> >	<P <sub>β</sub> >	St real
T	0.82	1.2	i	0 1	3	0.81
S	0.79	0.72	i	0 2	3	0.90
G	0.53	0.81	B	0 2	4	0.90
T	0.82	1.2	i	0 3	h	1 4
V	1.14	1.65	h	1 4	1	1.05 *
C	0.77	1.3	i	0 3	H	1 3
L	1.34	1.22	h	1 3	h	1 2
S	0.79	0.72	i	0 3	b	0 1
A	1.45	0.97	H	1 4	i	0 1
L	1.34	1.22	H	1 3	h	1 1
P	0.59	0.62	B	0 2	b	0 1
P	0.59	0.62	B	0 3	b	0 1
E	1.53	0.26	H	1 3	B	0 2
A	1.45	0.97	H	1 2	i	0 2
T	0.82	1.2	i	0 1	h	1 3
D	0.98	0.8	i	0 1	i	0 2
Q	1.17	1.23	h	1 1	h	1 3
S	0.79	0.72	i	0 1	i	0 1
Y	0.61	1.29	h	0 2	h	1 3
G	0.53	0.81	B	0 2	i	0 2
Y	0.61	1.29	b	0 2	h	1 3
Y	0.61	1.29	b	0 3	h	1 3
H	1.24	0.71	h	1 3	b	0 3
E	1.53	0.26	H	1 2	5	0.99
Y	0.61	1.29	b	0 1	5	0.72
T	0.82	1.2	i	0 1	5	0.73
V	1.14	1.65	h	1 3	1	1.09 *
I	1	1.6	i	0.5	5	0.77
T	0.82	1.2	i	0 1	5	1.12 *
P	0.59	0.62	B	0 0	5	1.14 *
G	0.53	0.81	B	0 0	5	1.24 *
						Q3
						17
						54,8%
						31

42

## La Méthode Chou-Fasman 9

43-1

## La Méthode Chou-Fasman 10

Utilisant <http://cib.cf.ocha.ac.jp/bitool/MIX/>

44-1

## La Méthode Chou-Fasman 9

- L'algorithme Chou-Fasman
  - Utilise la statistique (15 - 29 protéines)
  - Depuis 1978 aussi les tours  $\beta$ .
  - Extrêmement simple
    - Calcul par acide aminé; la méthode ne prend pas en compte les interactions entre les acides aminés
    - Algorithme est arbitraire; beaucoup de variations dans les règles heuristiques
  - Mais il donne un meilleur résultat que aléatoire
    - $Q_3$  moyenne entre 42-52% dépendant de la famille de protéines.

43-2

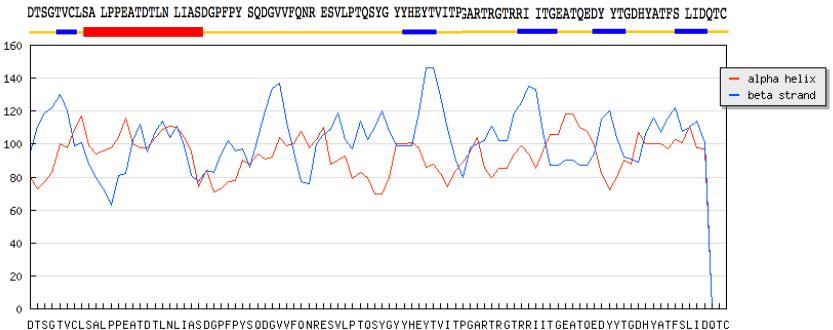
## La Méthode Chou-Fasman 10

DTSGTVCLSA LPPEATDTLN LIASDGPPPY SQDGVVFQNR ESQLPTOSYG YYHEYTVITP<sub>3</sub>ARTRGTRRI ITGEATQEDY YTGDHYATFS LIDQTC

Utilisant <http://cib.cf.ocha.ac.jp/bitool/MIX/>

44-2

## La Méthode Chou-Fasman 10



Utilisant <http://cib.cf.ocha.ac.jp/bitool/MIX/>

44-3

## La méthode GOR 2

- L'information individuelle (~ l'algorithme de Chou-Fasman)**

S = structure secondaire (E, H ou C)  
R = acide aminé

$$I(S|R) = \log[P(S|R) / P(S)] \quad (1)$$

$$P(S|R) = P(SR) / P(R) \quad (2)$$

$$P(SR) = f_{SR} / N \quad P(R) = f_R / N \quad P(S) = f_S / N$$

$f_{SR}$  = le nombre de résidus R au sein de la conformation S

$$I(S|R) = \log[(f_{SR} / f_R) / (f_S / N)] \quad (3)$$

Les valeurs sont estimées en utilisant des bases de données

46

## La méthode GOR

- Construite par Jean Garnier, David Osguthorpe, Barry Robson en 1978
  - La version en cours est GOR-V
- La méthode utilise la théorie d'information
  - l'information individuelle
  - l'information directionnelle
  - l'information par paire
- Modèle est facile à implémenter

45

## La méthode GOR 3

- L'information individuelle, plus générale

n-S = structures secondaires différentes de S  
Donc si S = E, n-S = {H,C}

$$I(\Delta SR) = I(S|R) - I(n-S|R) \quad (1)$$

$$I(\Delta SR) = \log(f_{SR} / f_{n-SR}) + \log(f_{n-S} / f_S) \quad (2)$$

Les valeurs sont estimées en utilisant des bases de données

## La méthode GOR 4

- Ajoutez l'information locale autour du résidu j

$$I(\Delta S_j; R_1, \dots, R_n) = \log[P(S_j, R_1, \dots, R_n) / P(n - S_j, R_1, \dots, R_n)] + \log[P(n - S_j) / P(S_j)]$$

Il est difficile de calculer ces probabilités directement à partir de la base de données

On a besoin d'une approximation ...

48

## La méthode GOR 6

- Implémentation de GOR I et II

-8                  j                  8  
THIS IS MY NEW PEPTID

50-1

## La méthode GOR 5

- l'approximation dans GOR I et II ...
  - L'information individuelle et directionnelle
    - 26 protéines, ~4500 résidus (1978, GOR I)
  - Utilisez une fenêtre de 17 résidus
    - pour chaque résidu à position j, prenez les résidus entre la position j-8 jusqu'à j+8

$$I(\Delta S_j; R_1, \dots, R_n) \approx I(\Delta S_j; R_j) + \sum_{m=-8}^{m=8} I(\Delta S_j; R_{j+m})$$

information individuelle      information directionnelle

49

## La méthode GOR 6

- Implémentation de GOR I et II

-8                  j                  8  
THIS IS MY NEW PEPTID       $I(\Delta S_j; R_j)$   
information individuelle

50-2

## La méthode GOR 6

- Implémentation de GOR I et II

-8                  j                  8  
 THIS IS MY NEWPEPTID       $I(\Delta S_j; R_j)$   
information individuelle  
 $I(\Delta S_j; R_j) = \log(f_{S_j, R_j} / f_{n-S_j, R_j}) + \log(f_{n-S} / f_S)$

50-3

## La méthode GOR 6

- Implémentation de GOR I et II

-8                  j                  8  
 THIS IS MY NEWPEPTID       $I(\Delta S_j; R_j)$   
information individuelle  
 $I(\Delta S_j; R_j) = \log(f_{S_j, R_j} / f_{n-S_j, R_j}) + \log(f_{n-S} / f_S)$

Asparagine N at position j

S	$f_{S,R}$	$f_{n-S,R}$	$f_{n-S}$	$f_S$	I
Helix	200	800	13000	7000	-0,33
Sheet	300	700	15000	5000	0,11
Coil	500	500	12000	8000	0,18

50-4

## La méthode GOR 7

- Implémentation de GOR I et II

-8                  j                  8  
 THIS IS MY XEWPEPTID

51-1

## La méthode GOR 7

- Implémentation de GOR I et II

-8                  j                  8       $\sum_{m=-8}^{m=8} I(\Delta S_j; R_{j+m})$   
THIS IS MY XEWPEPTID  
Information

51-2

## La méthode GOR<sub>7</sub>

- Implémentation de GOR I et II

-8 j 8  
**THISISMYXEWPEPTID**  
Information

$$\sum_{m=-8}^{m=8} I(\Delta S_j; R_{j+m})$$

$$I(\Delta S_j; R_{j+m}) = \log(f_{S_j, R_{j+m}} / f_{n-S_j, R_{j+m}}) + \log(f_{n-S_j} / f_S)$$

51-3

## La méthode GOR<sub>7</sub>

- Implémentation de GOR I et II

-8 j 8  
**THISISMYXEWPEPTID**  
Information

$$\sum_{m=-8}^{m=8} I(\Delta S_j; R_{j+m})$$

$$I(\Delta S_j; R_{j+m}) = \log(f_{S_j, R_{j+m}} / f_{n-S_j, R_{j+m}}) + \log(f_{n-S_j} / f_S)$$

Threonine T at position j-8, any residue at position j in conformation S

S	f <sub>S,R</sub>	f <sub>n-S,R</sub>	f <sub>n-S</sub>	f <sub>S</sub>	I
Helix	100	300	13000	7000	-0,21
Sheet	200	200	15000	5000	0,48
Coil	100	300	8000	12000	-0,3

51-4

## La méthode GOR<sub>8</sub>

52-1

## La méthode GOR<sub>8</sub>

$$\sum_{m=-8}^{m=8} I(\Delta S_j; R_{j+m})$$

S, 0	R, -8	R, -7	R, -6	R, -5	R, -4	R, -3	R, -2	...
T	H	I	S	I	S	M	...	...
Helix	-0,21	...	...	...	...	...	...	...
Sheet	0,48	...	...	...	...	...	...	...
Coil	-0,3	...	...	...	...	...	...	...

52-2

## La méthode GOR<sub>8</sub>

- Implémentation de GOR I et II

$$\sum_{m=-8}^{m=8} I(\Delta S_j; R_{j+m})$$

S, 0	R, -8	R, -7	R, -6	R, -5	R, -4	R, -3	R, -2	...
T	H	I	S	I	S	M	...	
Helix	-0,21	...	...	...	...	...	...	...
Sheet	0,48	...	...	...	...	...	...	...
Coil	-0,3	...	...	...	...	...	...	...

- Au total 20 résidus \* 16 positions \* 3 = 960 valeurs
- et 60 valeurs pour l'information individuelle

52-3

## La méthode GOR<sub>9</sub>

- Résultats pour GOR 1

- $Q_3 \approx 49\%$ ,
- Il a des problèmes avec les prédictions des  $\beta$ -brins, pourrait être amélioré jusqu'à 60%

53-2

## La méthode GOR<sub>9</sub>

- Résultats pour GOR 1

- $Q_3 \approx 49\%$ ,
- Il a des problèmes avec les prédictions des  $\beta$ -brins, pourrait être amélioré jusqu'à 60%

DTSGTVCLSA IPPEATDTLN LIASDGPFY SQDGVVFPQR ESLVPTQSYG YYHEYTVITPGARTRGTRRI ITGEATQEDY YTGDHYATFS LIDQTC

53-3



## La méthode GOR<sub>10</sub>

- l'approximation de GOR III

–68 protéines, 12000 résidus (1987)

$$I(\Delta S_j; R_1, \dots, R_n) \approx I(\Delta S_j; R_j) + \sum_{m=-8}^{m=8} I(\Delta S_j; R_{j+m} | R_j)$$

54-3

## La méthode GOR<sub>10</sub>

- l'approximation de GOR III

–68 protéines, 12000 résidus (1987)

$$I(\Delta S_j; R_1, \dots, R_n) \approx I(\Delta S_j; R_j) + \sum_{m=-8}^{m=8} I(\Delta S_j; R_{j+m} | R_j)$$

information individuelle      information par Pairs

54-5

## La méthode GOR<sub>10</sub>

- l'approximation de GOR III

–68 protéines, 12000 résidus (1987)

$$I(\Delta S_j; R_1, \dots, R_n) \approx I(\Delta S_j; R_j) + \sum_{m=-8}^{m=8} I(\Delta S_j; R_{j+m} | R_j)$$

information individuelle

54-4

## La méthode GOR<sub>10</sub>

- l'approximation de GOR III

–68 protéines, 12000 résidus (1987)

$$I(\Delta S_j; R_1, \dots, R_n) \approx I(\Delta S_j; R_j) + \sum_{m=-8}^{m=8} I(\Delta S_j; R_{j+m} | R_j)$$

information individuelle      information par Pairs

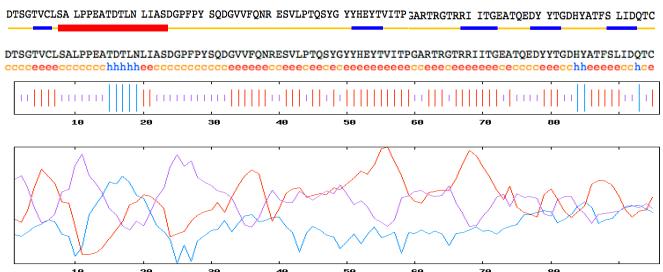
$$\begin{aligned} I(\Delta S_j; R_{j+m} | R_j) &= \log(f_{S_j, R_{j+m}, R_j} / f_{n-S_j, R_{j+m}, R_j}) \\ &\quad + \log(f_{n-S_j, R_j} / f_{S_j, R_j}) \end{aligned}$$

54-6



## La méthode GOR 11

- Les résultats
  - GOR III a un  $Q_3$  de ≈63%



- Comment est-ce qu'on peut encore l'améliorer?

55-5

## La méthode GOR 12

- L'implémentation de GOR IV
  - 267 protéines, 63000 résidus (1996)
  - Calculez l'information pour toutes les paires de

$$\log \frac{P(S_j, LocSeq)}{P(n - S_j, LocSeq)} = \frac{2}{17} \sum_{\substack{m=-8 \\ n>m}}^{+8} \log \frac{P(S_j, R_{j+m}, R_{j+n})}{P(n - S_j, R_{j+m}, R_{j+n})}$$

$$- \frac{15}{17} \sum_{m=-8}^{+8} \log \frac{P(S_j, R_{j+m})}{P(n - S_j, R_{j+m})}$$

56-2

## La méthode GOR 12

$$\log \frac{P(S_j, LocSeq)}{P(n - S_j, LocSeq)} = \frac{2}{17} \sum_{\substack{m=-8 \\ n>m}}^{+8} \log \frac{P(S_j, R_{j+m}, R_{j+n})}{P(n - S_j, R_{j+m}, R_{j+n})}$$

$$- \frac{15}{17} \sum_{m=-8}^{+8} \log \frac{P(S_j, R_{j+m})}{P(n - S_j, R_{j+m})}$$

56-1

## La méthode GOR 14

57-1

## La méthode GOR 14

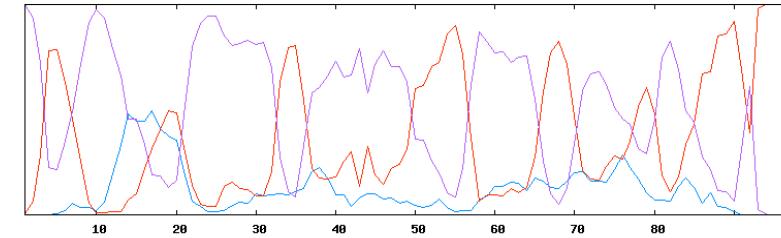
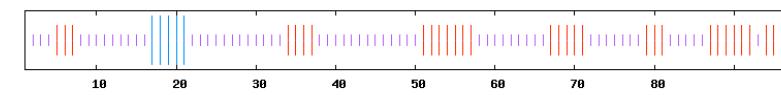
DTSGTVCLSA LPPEATDTLN LIASDGPFPPY SQDGVVVFQNR ESVLPTQSYG YYHEYTVITPGARTRGTRRI ITGEATQEDY YTGDHYATFS LIDQTC



57-2

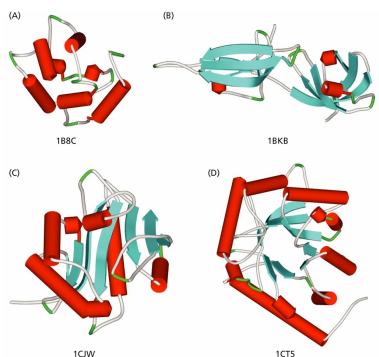
## La méthode GOR 14

DTSGTVCLSA LPPEATDTLN LIASDGPFPPY SQDGVVVFQNR ESVLPTQSYG YYHEYTVITPGARTRGTRRI ITGEATQEDY YTGDHYATFS LIDQTC



57-3

## La méthode GOR 15



- GOR IV a un Q3 moyenne de 64.4%
- Pour les différentes structures : 39.2% ( $\alpha+\beta$ ) to 67.2% ( $\alpha/\beta$ )

58

## La méthode GOR 16

59-1

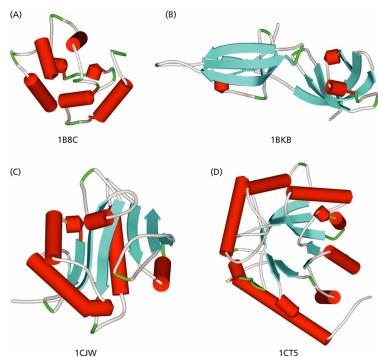
## La méthode GOR 16

- Finalement GOR V : utilise un alignement de plusieurs séquences
  - La structure est plus conservée que la séquence
  - Les insertions et délétions sont plus communes dans boucles
  - Retrouvez des séquences en utilisant PSI-BLAST
    - Limitez à 30% d'identité s'il y en a beaucoup
    - Calculez la structure secondaire pour chaque séquence
    - Utilisez la moyenne des prédictions pour chaque

59-2

## La méthode GOR 17

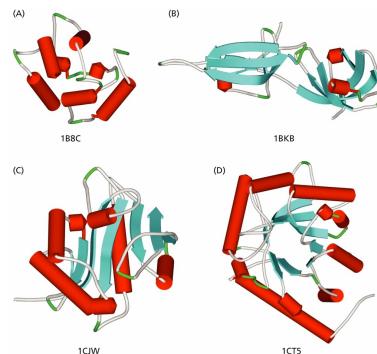
- La base de données est plus grande qu'avant



- Les résultats :
  - $Q_3$  moyenne de 73.5% ( $>$  GOR IV)
  - Pour les différentes structures: 60.3% (all $\beta$ ) to 84.3% (all $\alpha$ ) ( $>$ GOR IV)

60-2

## La méthode GOR 17



- Les résultats :
  - $Q_3$  moyenne de 73.5% ( $>$  GOR IV)
  - Pour les différentes structures: 60.3% (all $\beta$ ) to 84.3% (all $\alpha$ ) ( $>$ GOR IV)

60-1

## Autres Méthodes

## Autres Méthodes

- Zpred : basé sur GOR + information sur les propriétés biochimiques de tous les acides aminés

61-2

## Autres Méthodes

- Zpred : basé sur GOR + information sur les propriétés biochimiques de tous les acides aminés
- Nearest-neighbor methods (Predator,...)

61-4

## Autres Méthodes

- Zpred : basé sur GOR + information sur les propriétés biochimiques de tous les acides aminés

61-3

## Autres Méthodes

- Zpred : basé sur GOR + information sur les propriétés biochimiques de tous les acides aminés
- Nearest-neighbor methods (Predator,...)

61-5

## Autres Méthodes

- Zpred : basé sur GOR + information sur les propriétés biochimiques de tous les acides aminés
- Nearest-neighbor methods (Predator,...)
- Méthodes basées sur de réseaux neurones (PHD, PROF, ...)–Les meilleurs résultats !

61-6