

Analyse et prédition de la structure locale des protéines : des structures secondaires aux alphabets structuraux.

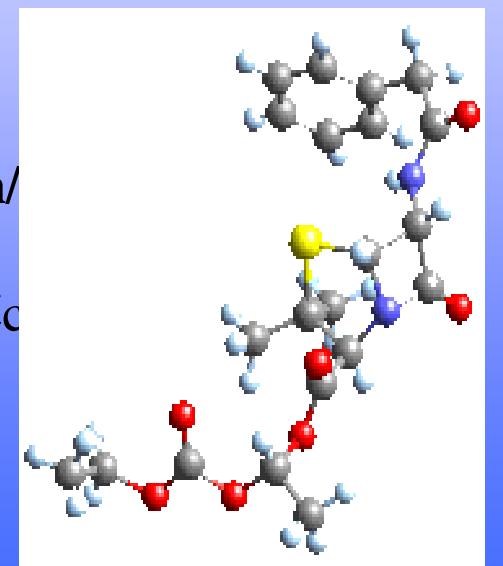


Alexandre G. de Brevern

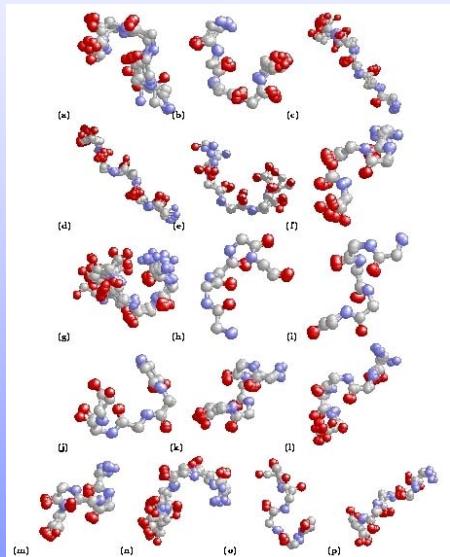
<http://www.ebgm.jussieu.fr/~debrevorn/>

Équipe de Bioinformatique Génomique et Moléculaire
(EBGM)

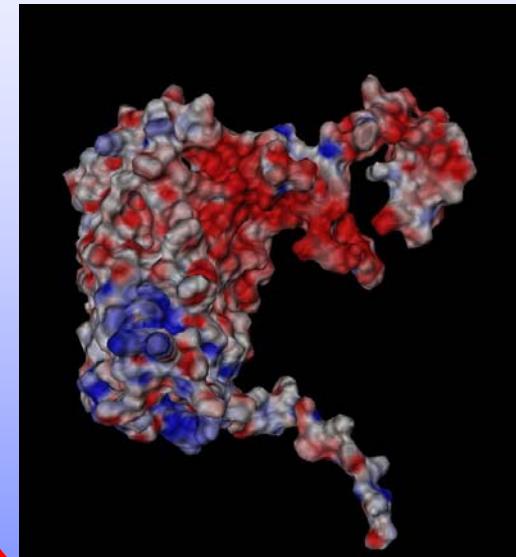
INSERM U 726 / Université Paris VII
75251 PARIS Cedex 05 – FRANCE



Equipe de Bioinformatique Génomique et Moléculaire

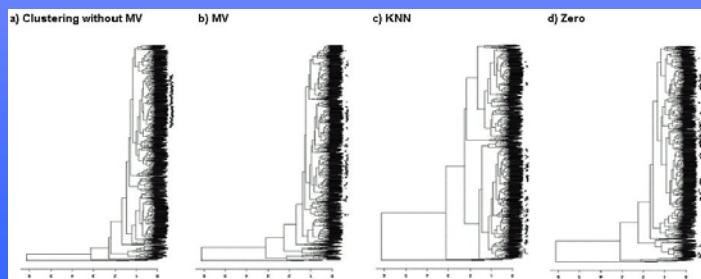


Structure



Séquence

Fonction



Analyse et prédiction de la structure locale des protéines : des structures secondaires aux alphabets structuraux.

**La question : Pourquoi s'intéresser à autre chose
qu'aux structures secondaires définies par DSSP ?**

Les structures secondaires

Un autre regard

Les alphabets structuraux

Intérêt sur le plan de la structure

Séquence → Prédiction

L'analyse des protéines



1D : Séquence

..- Ala - Gly - Leu - Pro - Ala - Met - Pro - Ile - Leu - Arg - Gly - Met -...



2D structures secondaires

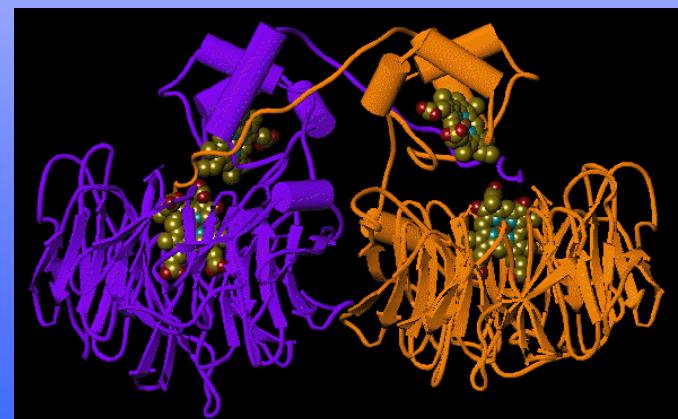
...-hélice--boucle--brin--boucle--brin--boucle--hélice-...



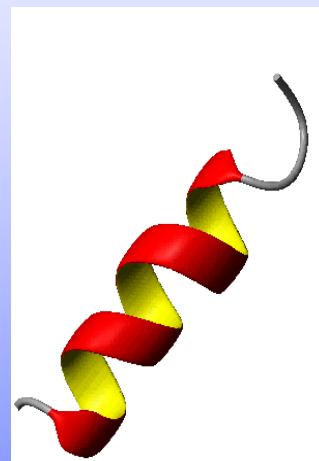
3D la protéine



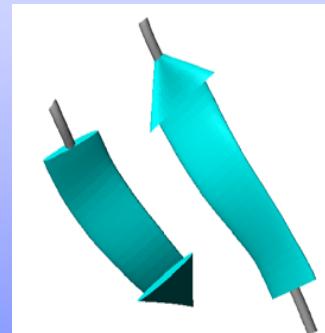
4D les protéines (complexes)



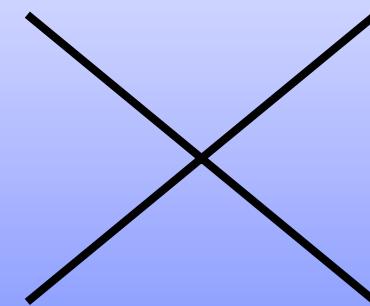
Les structures secondaires



état hélicoïdal
(28% – 35%)

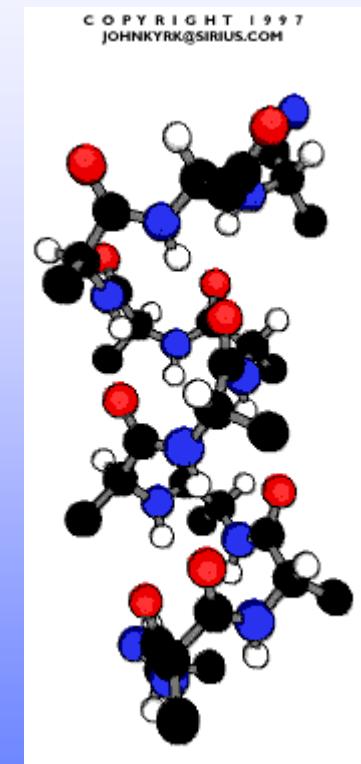
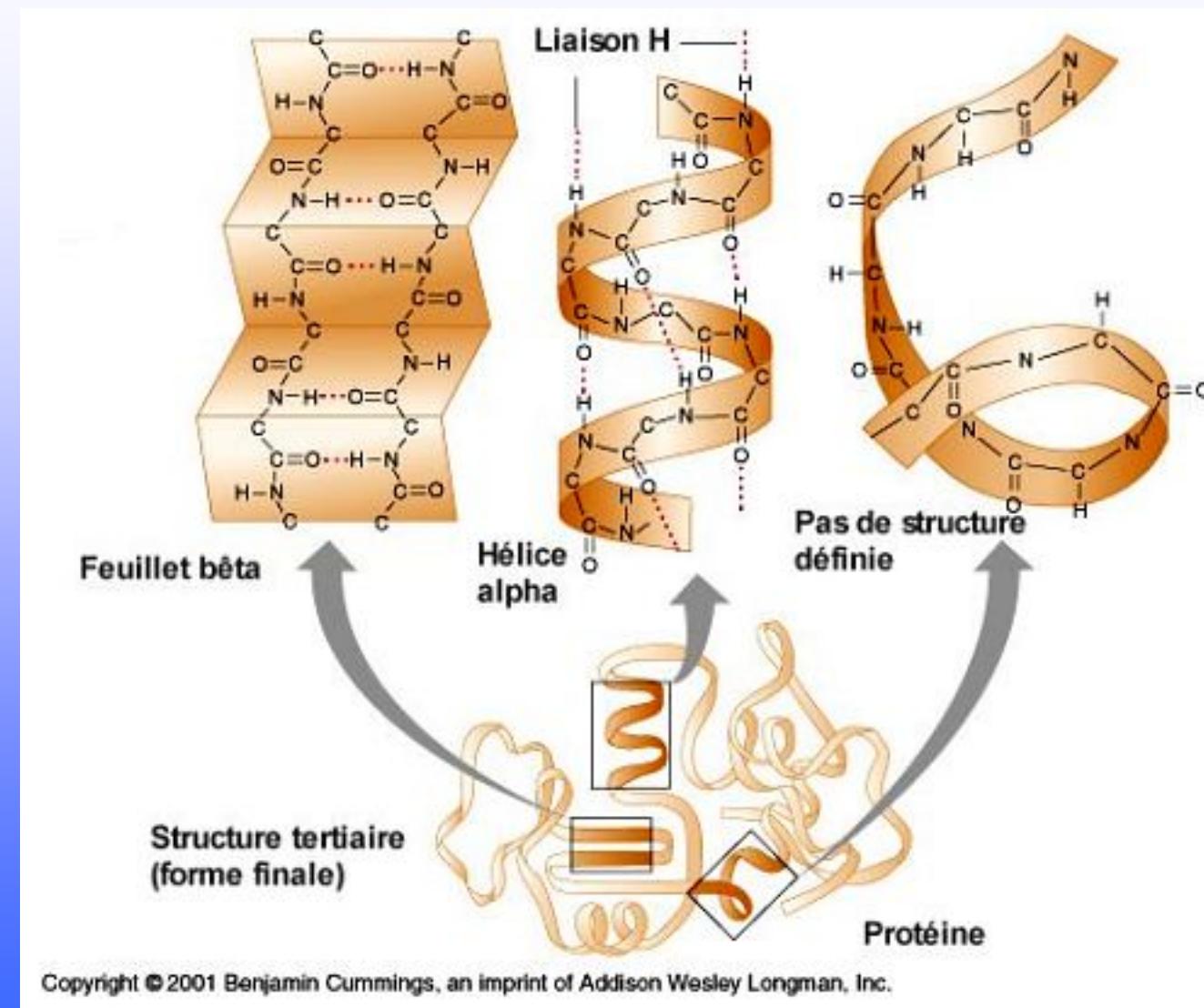


état étendu
(18% – 26 %)



boucle
(40% – 50 %)

Les structures secondaires

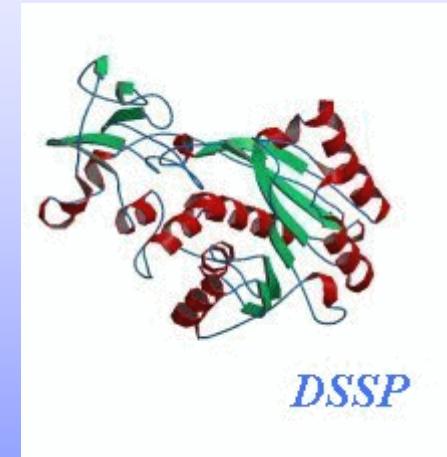


Différentes méthodes d'assignations :

- Greer & Levitt (1977)
 - **DSSP** (Kabsch & Sander, 1983).
- | | |
|--|------------|
| | Distance |
| | Liaisons H |

What is unfortunate is that people use these secondary structure assignments unquestioningly; perhaps the greatest damage the programs do is to create an impression (for which Levitt, Greer, *et al.*, cannot be blamed) that there is **A RIGHT ANSWER**. Provided that the danger is recognized, such programs can be useful (Arthur M. Lesk, *Introduction to protein architecture*, p.80).

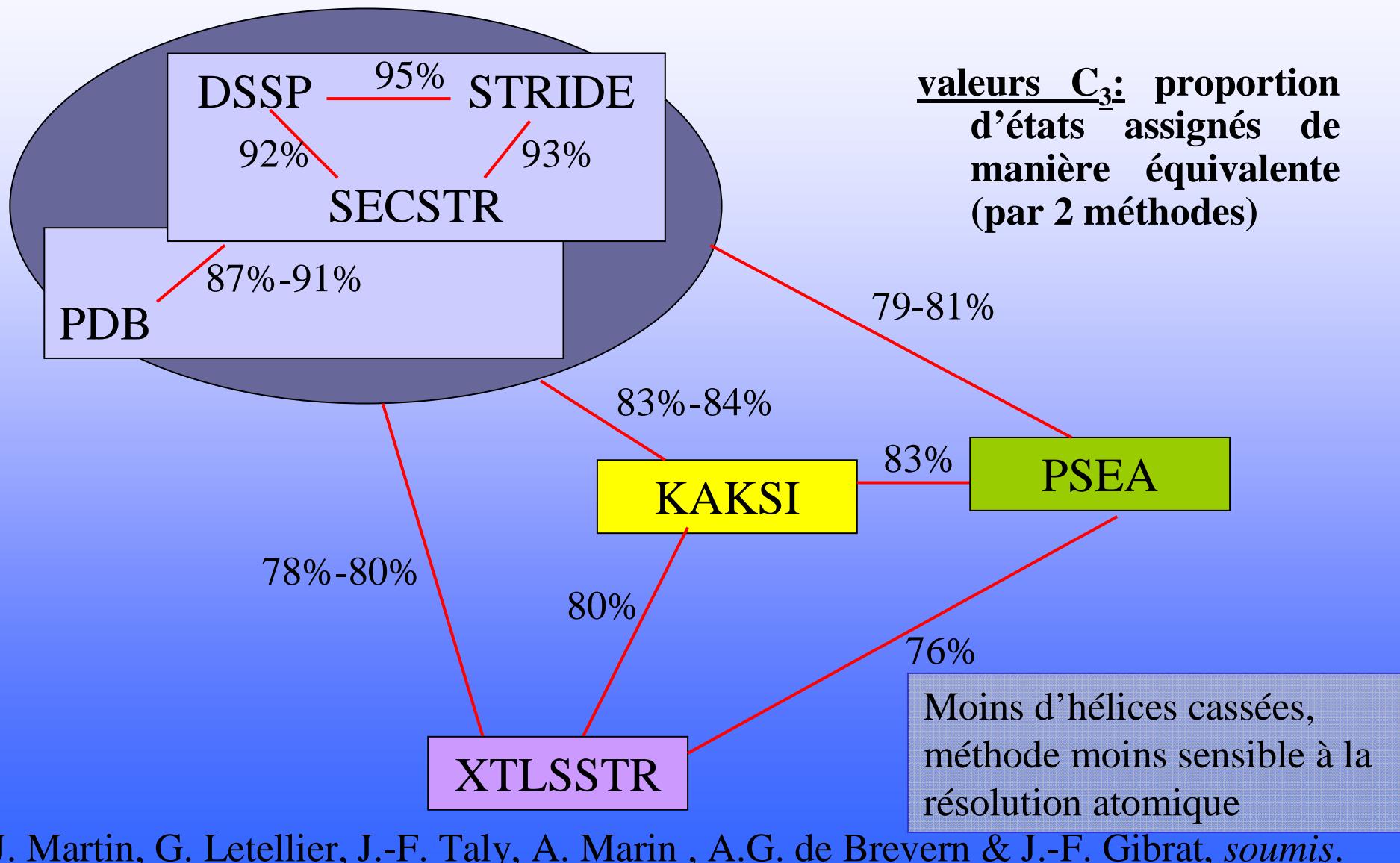
AA	WDKYAQEVYEMNFGEKPEGDITQVNEKTI	PDHDILCAGFP				
DSSP3	CCHHHHHHHHHHCCCCCCC	HHHCCCCCCCCC	EEEEEC			
STRID3	CCHHHHHHHHHHCCCCCCCCCCCCCCCC	CCCCCCCCCCCC	EEEEEC			
PSEA	EEHHHHHHHHHHCCCCCCCCCCCC	CCCCCCCCCCCC	EEECCC			
DEFINE	EEHHHHHHHHHHHEEEEEEEHHHHHHHH	EEEEEEEEE	EEEEEEEEE			
PCURVE	CCHHHHHHHHHHCCCCCCCCCCCC	CCCCCCCCCCCC	EEEEEEE			
cons.*****.....	*****..			
PB	b fk l m m m m m m n o p a c d e d f k l p c f k l p c c d f b d c d d d f					
[C93]	CCHHHHHHHHHHCCCCCCCC	CCCCCCCC	EEEEEEE			
XTLSS.	CHHHHHHHHHHHEEEPPC	NNNC	GGGGPPPCEEEECCPP			
SECSTR	CCHHHHHHHHHHCCCC	CCCC	GGGCCCCCCCC	EEEEEC		
DSSP	CCHHHHHHHHHHHS	CCC	BCCGGGS	CTTTS	CCC	SEEEEEC
STRIDE	CCHHHHHHHHHHCCCC	BCTTTTTTTT	TTT	CCCC	EEEEEC	
HELAN.	--VVVVVVVVVV--					
BETAEX	--	b	--ppppq--			



Exemple d'assignation par différentes méthodes pour la protéine 10MH avec DSSP, STRIDE, PSEA, DEFINE, PCURVE, XTLSSTR and SECSTR.

Fourrier, Benros & de Brevern (2004) *BMC Bioinformatics*, **5**, 58.

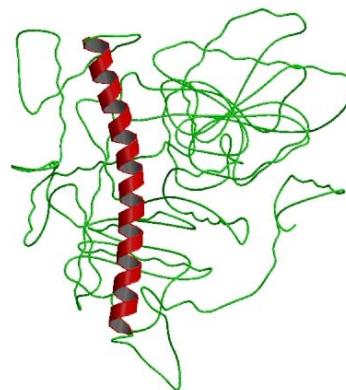
Les structures secondaires



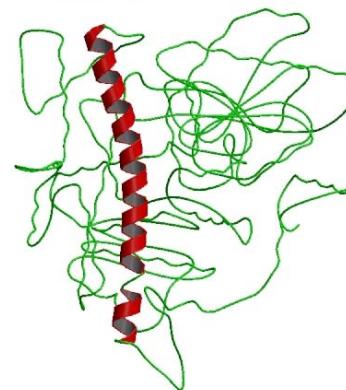
Les structures secondaires



DSSP



KAKSI



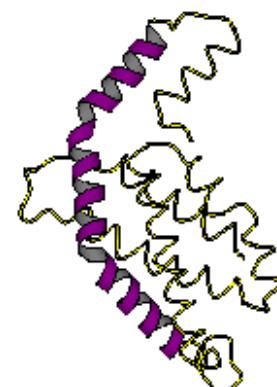
STRIDE



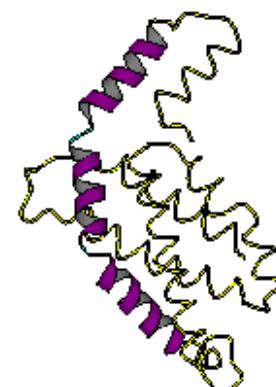
KAKSI



STRIDE



KAKSI



Les hélices trop ‘longues’.

Les structures secondaires



DSSP



KAKSI

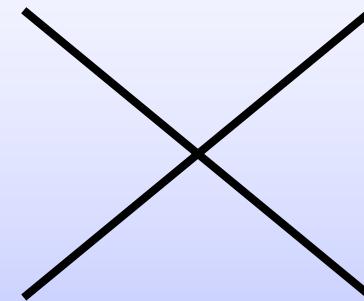
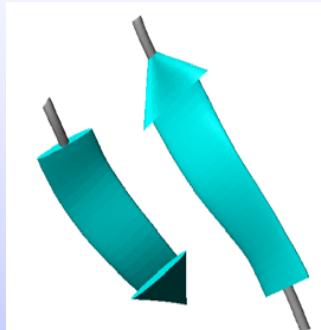
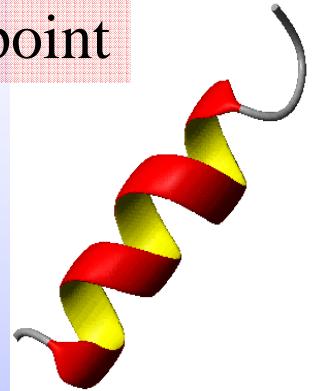


Les surprises de DSSP.

Les structures secondaires



Autre point



3 catégories seulement ?

Hélice α

Hélice β_{10}

Hélice π

Helanal [5]

brins β

feuilles β

β -bulge

Coudes γ , β , π & α

Polyproline II

β -hairpins

(classifications)

En fait plus de 40 catégories

Un alphabet structural est une série (ou librairie) de **petits prototypes** qui approximent **chaque partie** des structures protéiques.

Ils sont composés d'un **nombre limité** d'éléments structuraux **récursifs** des structures protéiques.

Les associations entre ces "**lettres**" **structurales** sont gouvernées par des règles logiques et forment des mots (de structures protéiques).

Un alphabet structural n'a pas *d'a priori* vis-à-vis des structures secondaires, *i.e.* **ce n'est pas une catégorisation des boucles**.

Le dilemme des alphabets structuraux.

Nombre de prototypes	Approximation de la structure 3D locale	Prédiction

Important

Equipe	Année	Nombre	Longueur
Unger	1989 / 93	103 / 83	6
Pretrelski	1992	113	8
Schuchhardt	1996	100	9

Limité

Equipe	Année	Nombre	Longueur
Rooman	1990	4	4, 5, 6 & 7
Fetrow	1993 / 97	6	7
Bystroff	1998 / 2000	13 / 16	5 to 17 (8)
Camproux	1999 / 2004	12 / 27	4
de Brevern	2000	16	5
Hunter	2003	28	7

M. Tyagi, C. Benros, J. Martin, & A.G. de Brevern, Description of the local protein structure. II. Novel approaches, *in preparation*.

Les Blocs Protéiques

Objectifs

Approximation de
la structure
locale 3D

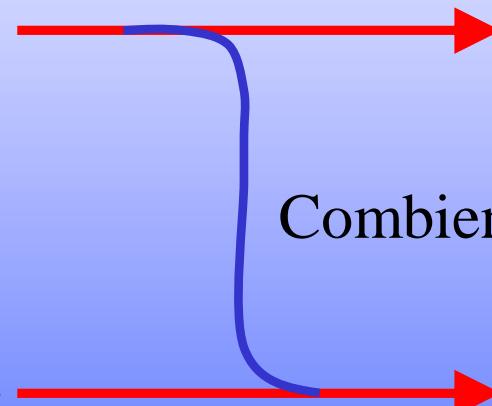
Prédiction de la
structure locale à
partir de la séquence

Méthodes

Classifieur non
supervisé (SOM /
HMM)

Combien de BPs ?

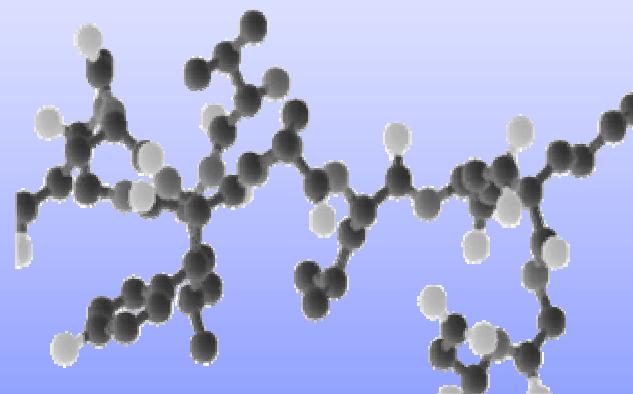
Prédiction
Bayésienne



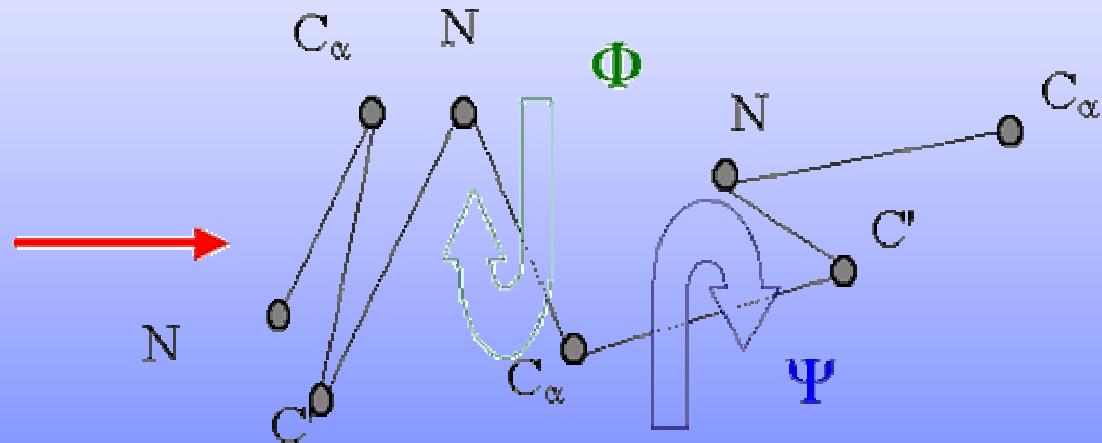
Les Blocs Protéiques

Information 3D.

3D



dihedral angles

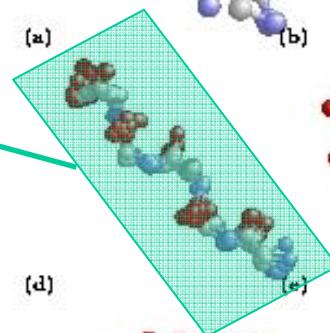


5 résidus => 8 angles dièdres (ϕ, ψ)

Les Blocs Protéiques



BP d

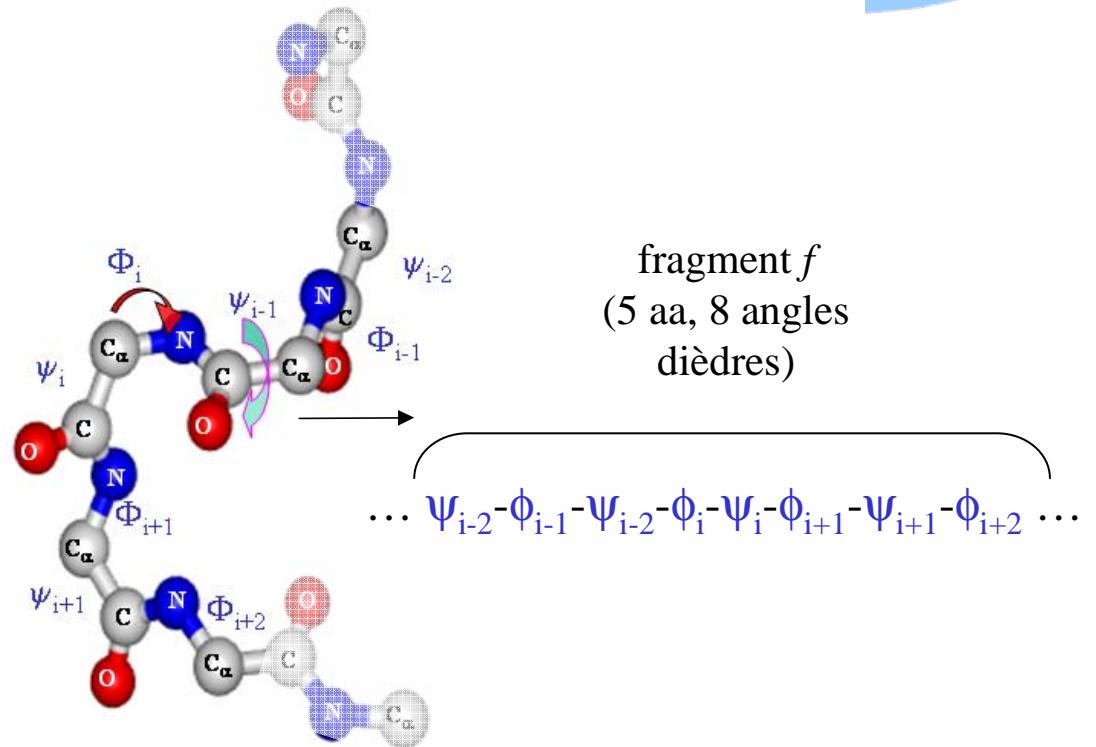
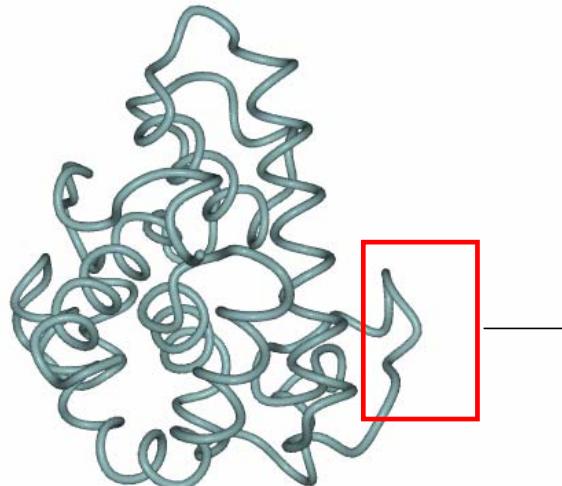


BP m



de Brevern A.G.,
Etchebest C. & Hazout, S.
(2000), *Bayesian
probabilistic approach for
prediction backbone
structures in terms of
protein blocks*, *Proteins*,
41(3):271-287.

Les Blocs Protéiques



Les Blocs Protéiques



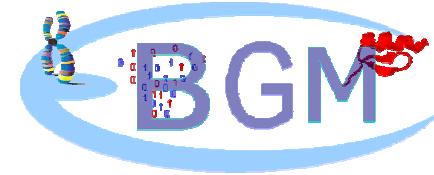
	Ψ_{i-2}	ϕ_{i-1}	Ψ_{i-1}	ϕ_i	Ψ_i	ϕ_{i+1}	Ψ_{i+1}	ϕ_{i+2}
PB <i>a</i>	41.14	75.53	13.92	-99.80	131.88	-96.27	122.08	-99.68
PB <i>b</i>	108.24	-90.12	119.54	-92.21	-18.06	-128.93	147.04	-99.90
PB <i>c</i>	-11.61	-105.66	94.81	-106.09	133.56	-106.93	135.97	-100.63
PB <i>d</i>	141.98	-112.79	132.20	-114.79	140.11	-111.05	139.54	-103.16
PB <i>e</i>	133.25	-112.37	137.64	-108.13	133.00	-87.30	120.54	77.40
PB <i>f</i>	116.40	-105.53	129.32	-96.68	140.72	-74.19	-26.65	-94.51
PB <i>g</i>	0.40	-81.83	4.91	-100.59	85.50	-71.65	130.78	84.98
PB <i>h</i>	119.14	-102.58	130.83	-67.91	121.55	76.25	-2.95	-90.88
PB <i>i</i>	130.68	-56.92	119.26	77.85	10.42	-99.43	141.40	-98.01
PB <i>j</i>	114.32	-121.47	118.14	82.88	-150.05	-83.81	23.35	-85.82
PB <i>k</i>	117.16	-95.41	140.40	-59.35	-29.23	-72.39	-25.08	-76.16
PB <i>l</i>	139.20	-55.96	-32.70	-68.51	-26.09	-74.44	-22.60	-71.74
PB <i>m</i>	-39.62	-64.73	-39.52	-65.54	-38.88	-66.89	-37.76	-70.19
PB <i>n</i>	-35.34	-65.03	-38.12	-66.34	-29.51	-89.10	-2.91	77.90
PB <i>o</i>	-45.29	-67.44	-27.72	-87.27	5.13	77.49	30.71	-93.23
PB <i>p</i>	-27.09	-86.14	0.30	59.85	21.51	-96.30	132.67	-92.91



Calcul de distance => 16 scores →

Le plus faible

Les Blocs Protéiques



Exemple BP *a*

>153L

ZZmnopfklpccebjafk1mmmnop



Le plus faible

BPs : fréquence → PB *d* (19%) et PB *m* (30%),
ensuite > 6%

transition → nombre limité de transition d'un BP à un autre

approximation → 0.41 Å (médiane = 0.26 Å)

forte discrimination entre BPs
(le second plus proche est ... loin)

de Brevern A.G. (2005), *New assessment of a structural alphabet*, **In Silico Biology**, 5, 26.

Les Mots Structuraux

1 Mot Structural (MS) = série de **5 PBs consécutifs** (les + fréquents)

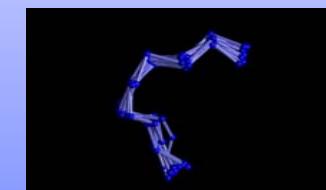
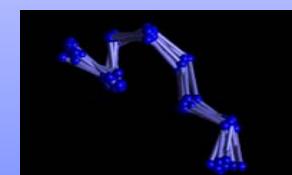
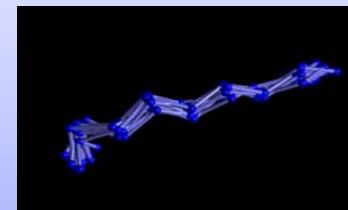
Défini en fait une forme de grammaire,

De successions très fréquentes.

5 C α → 9 C α

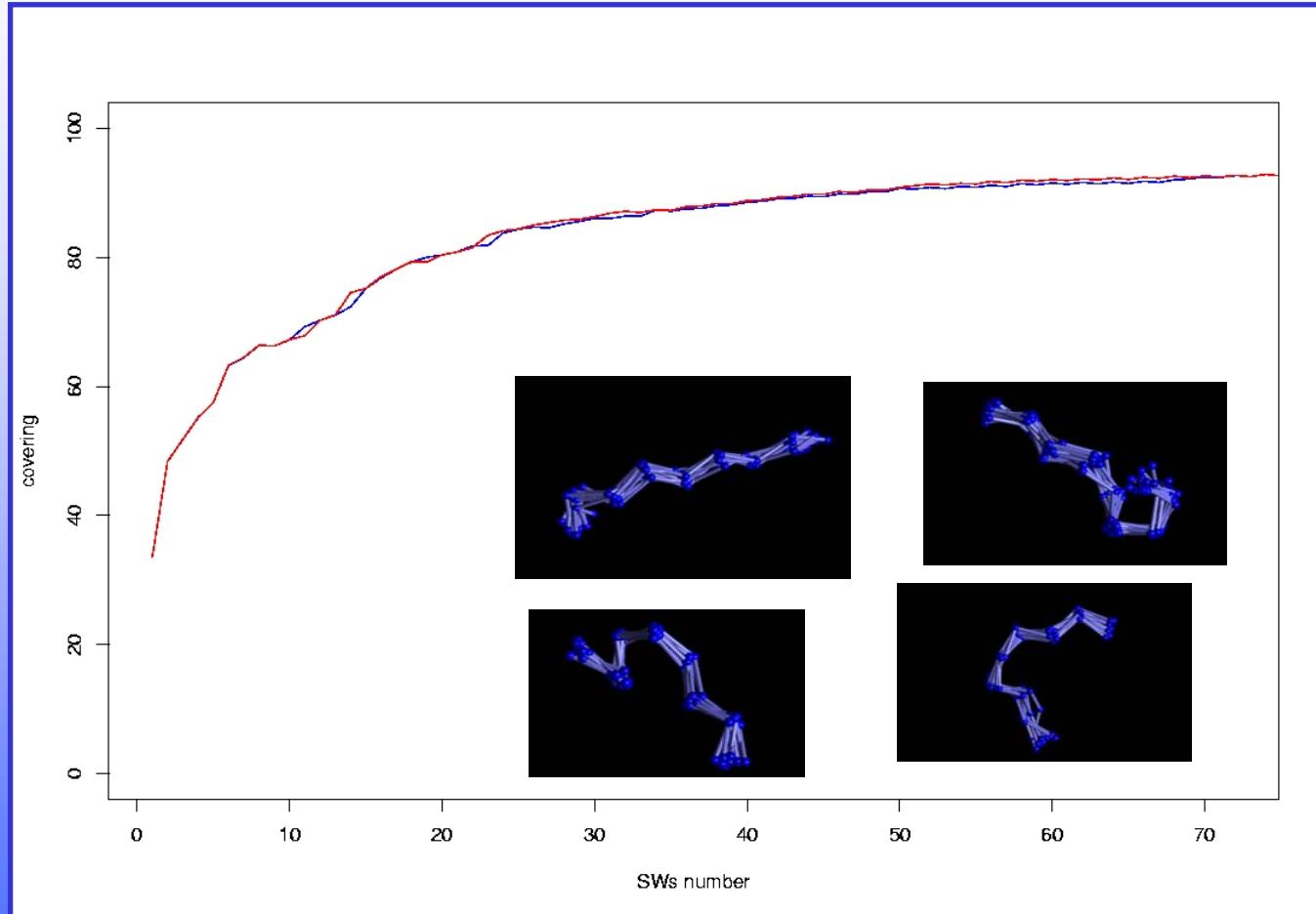
72 MSs les + fréquents → 0.9 Å

mmmnop
mmnopa
mnopac
opacd



Recouvrement permet d'avoir des chemins (création d'un graphe)

Les Mots Structuraux



Recouvrement : 72
MS => 92 % des résidus

de Brevern A.G., Valadié, H., Hazout H. & Etchebest C. (2002), *Extension of a local backbone description using a structural alphabet. A new approach to the sequence-structure relationship.*, *Protein Science*, 11(12):2871-2886.

Les Blocs Protéiques



Ne correspond pas à des
'structures secondaires'

	1st	2nd	3rd	4th	5th
<i>a</i>	0.12	1.21	0.10 (<i>a</i>)	0.00 (<i>b</i>)	
<i>b</i>	18.85	2.70	50.4 (<i>f</i>)	26.3 (<i>c</i>)	19.9 (<i>e</i>)
<i>d</i>	18.85	2.70	50.4 (<i>f</i>)	26.3 (<i>c</i>)	19.9 (<i>e</i>)
<i>e</i>	2.45	1.12	81.1 (<i>h</i>)	8.6 (<i>d</i>)	
<i>f</i>	6.68	1.01	61.5 (<i>k</i>)	35.0 (<i>b</i>)	
<i>g</i>	1.15	1.04	37.5 (<i>h</i>)	29.6 (<i>c</i>)	16.1 (<i>o</i>)
<i>h</i>	2.40	1.02	68.0 (<i>i</i>)	13.8 (<i>j</i>)	8.5 (<i>k</i>)
<i>i</i>	1.86	1.01	82.8 (<i>a</i>)	6.2 (<i>l</i>)	
<i>j</i>	0.83	1.01	21.7 (<i>b</i>)	14.8 (<i>a</i>)	14.7 (<i>k</i>)
<i>k</i>	5.45	1.01	77.2 (<i>l</i>)	10.5 (<i>b</i>)	6.2 (<i>o</i>)
<i>l</i>	5.46	1.01	68.2 (<i>m</i>)	8.6 (<i>p</i>)	7.1 (<i>c</i>)
<i>m</i>	30.22	7.00	34.9 (<i>n</i>)	15.7 (<i>p</i>)	11.3 (<i>k</i>)
<i>n</i>	1.99	1.01	92.4 (<i>o</i>)		
<i>o</i>	2.77	1.02	78.2 (<i>p</i>)	6.5 (<i>m</i>)	5.6 (<i>i</i>)
<i>p</i>	3.47	1.00	58.6 (<i>a</i>)	23.7 (<i>c</i>)	7.6 (<i>m</i>)

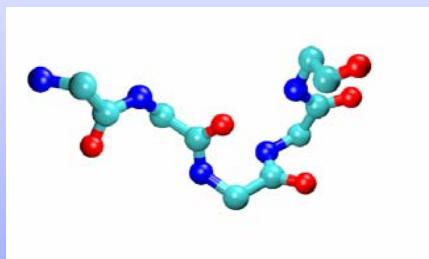
	rmsda (°)	rmsd (Å)	
	mean	dif.	mean
	45.2	29.3	0.46
	42.5	20.3	0.47
	38.4	21.4	0.51
	29.7	27.2	0.41
	40.9	23.5	0.71
	37.5	22.1	0.40
	50.6	14.9	0.60
	47.0	20.9	0.46
	43.4	25.0	0.41
	49.0	19.6	0.83
	35.9	25.4	0.30
	32.5	27.3	0.53
	15.0	40.1	0.31
	26.8	31.2	0.31
	38.3	27.1	0.48
	43.8	25.9	0.47
	30.1	29.5	0.41

Les Blocs Protéiques

Nouvelle séquence

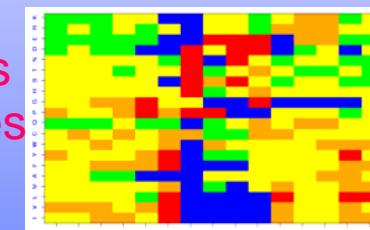
Prédiction Bayésienne

Repliement local (PB)



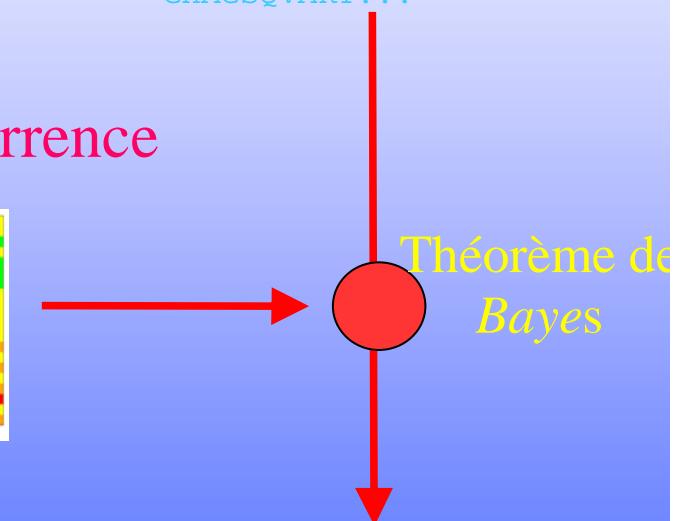
SYARMDIGTTTHDDYARMDIGTTTHDDYANDV
IHEVLAPGCLDAFPLGRDTSVEGSEMVPGK
VIGLLEPMKKSMVPVCVMLKSRGSRGHVRFGRLGLGEGAEEKSTP
HLWVHQEGIYRDEYQLMWQLYPEERYMDNNMWQLYPEERYMDNNS
QIAKYFDRKQIGNAM ...

Matrice d'occurrence



Positions

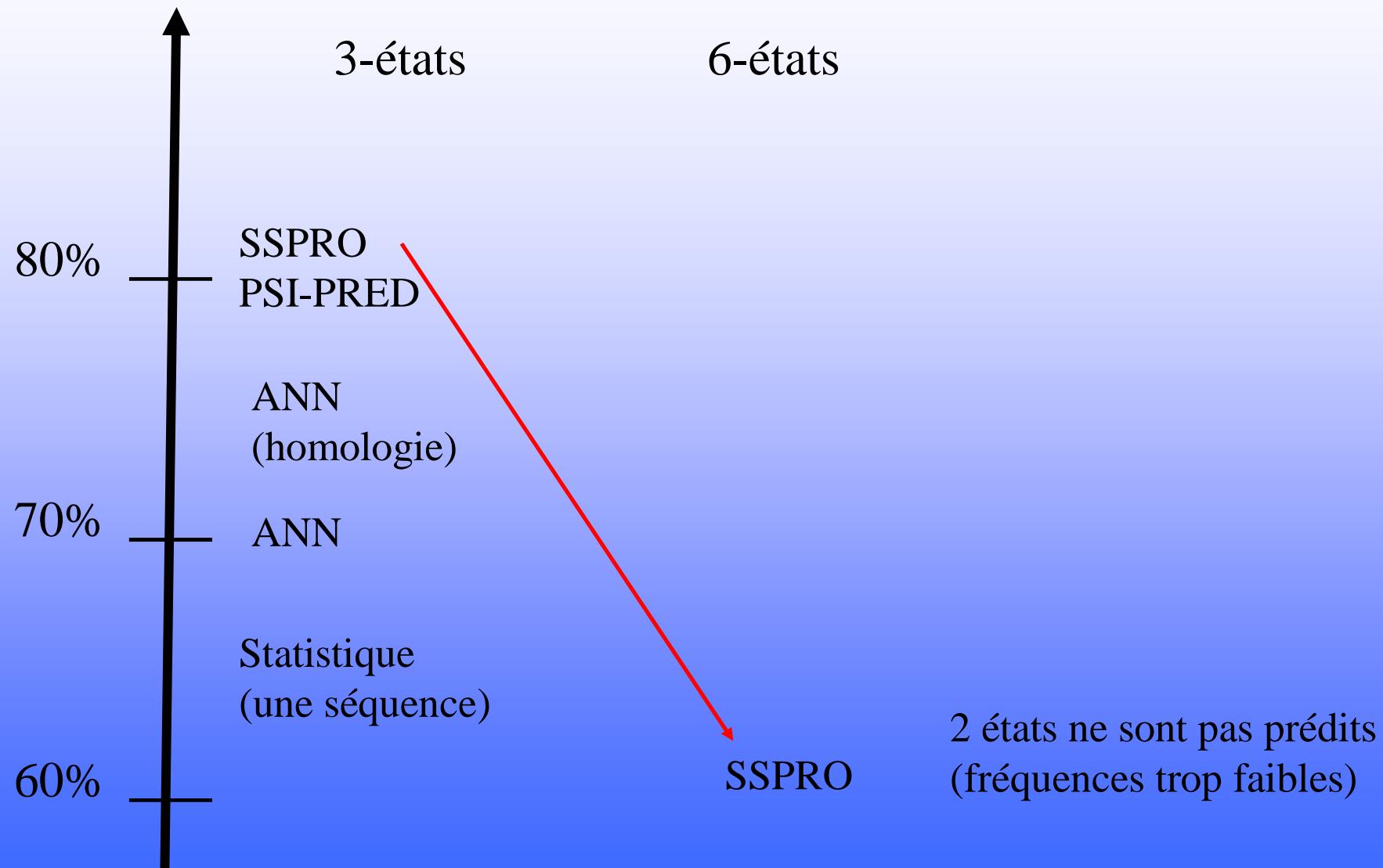
acides
aminés



Prédiction
Index de confiance

SFITPVPGVGPMVTFL...
NKNVIFVADKRKGPGGI...
CVHTFNSWLDVEPRAV...
GAIWKLDLAIWKL...
WWDSHIGAFLDKPKM...
NGLRYGLSSDAHTAVI...
ESAVIGLPSGLESWSFF...
GHAGSQVAKY ...

La prédiction est-elle ‘simple’ ?



Hunter & Subramaniam (2003) *Proteins.*

28 prototypes de 7 résidus de longs.

Prédiction Bayésienne => 40 % de bonne prédiction

8 avec une prédiction > 20 % (les plus fréquents)

11 avec une prédiction = 0%.

Taux de prédiction:

Taux initial : 34.4 %

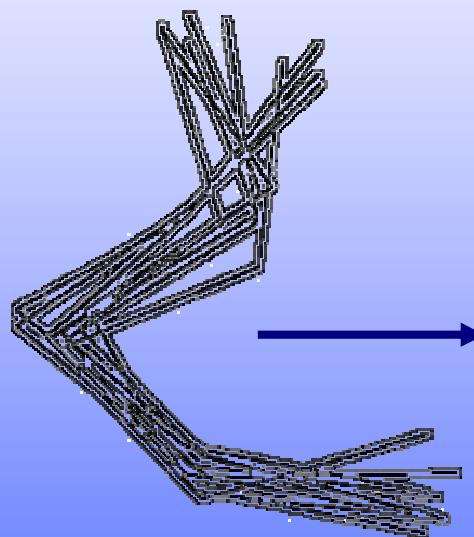
[aléatoire = 7.5%]

Taux de prédiction Minimal / maximal par BP :
initial : [13.1 % - 60.3 %]

Les Blocs Protéiques

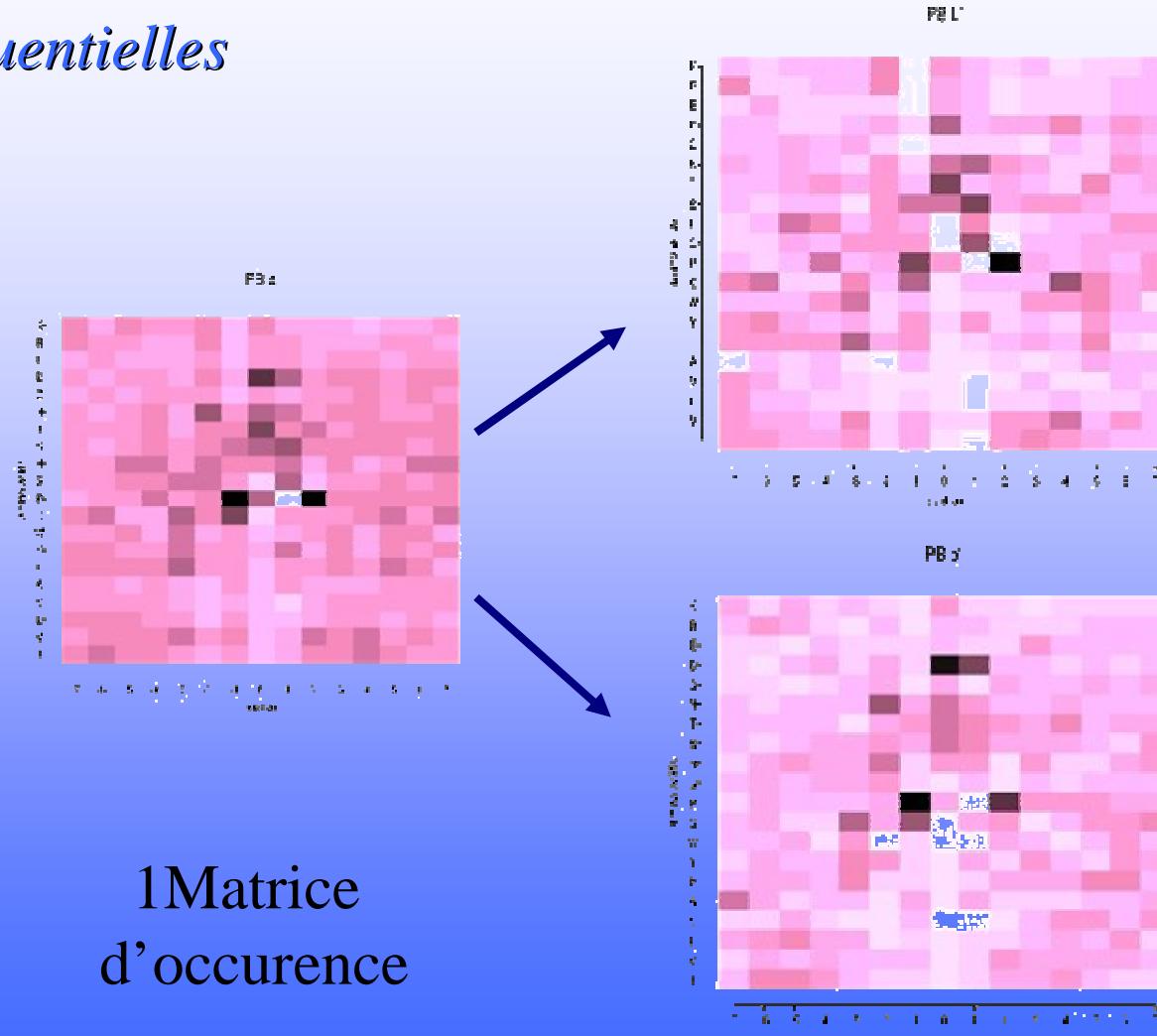


Familles Séquentielles



1 BP

1 Matrice
d'occurrence



Taux de prédition:

Taux initial : 34.4 % [aléatoire = 7.5%]

Prédiction avec les Familles Séquentielles = 40.7%

Taux de prédition Minimal / maximal par BP :

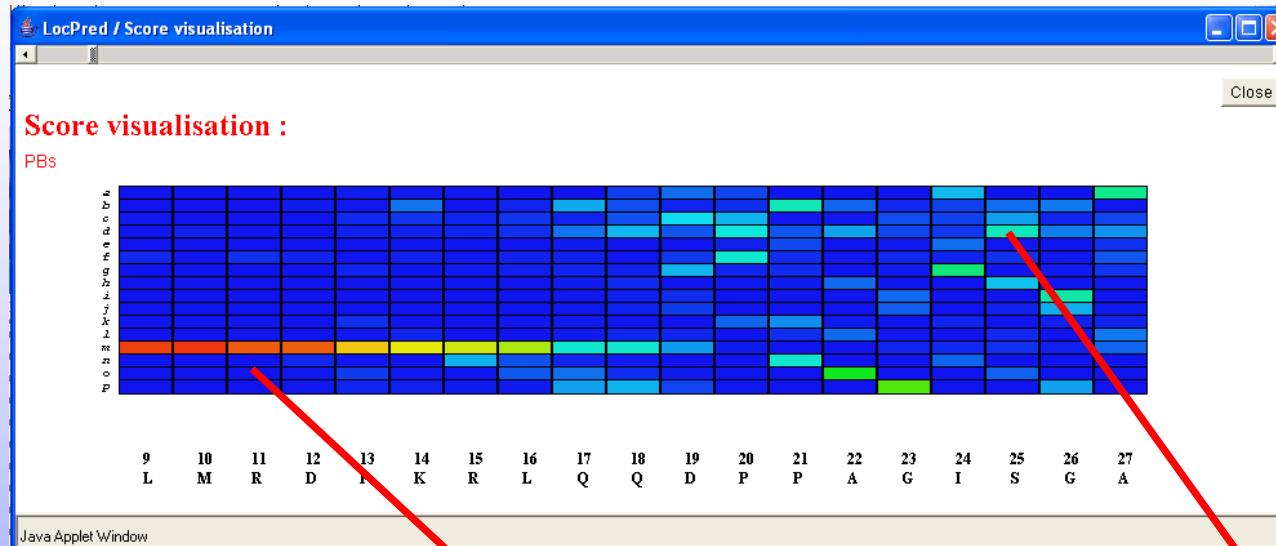
initial : [13.1 % - 60.3 %]

final : [27.0% -53.2%]

Gain pour 95% des protéines

de Brevern A.G., Etchebest C. & Hazout, S. (2000), *Bayesian probabilistic approach for prediction backbone structures in terms of protein blocks*, **Proteins**, 41(3):271-287.

Les Blocs Protéiques



Loc Pred

Indice de confiance

Basé sur les probabilités

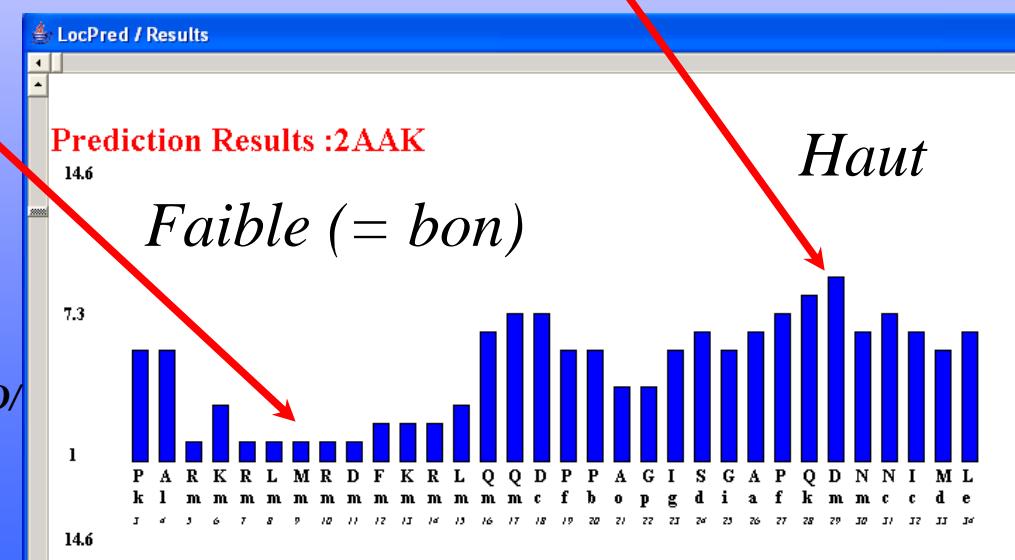
Des scores de prédiction

EBGM :

<http://www.ebgm.jussieu.fr/~debrevern/LOCPRED/>

RPBS:

<http://bioserv.rpbs.jussieu.fr/LocPred/>



de Brevern A.G., Benros C., Gautier R., Valadié H., Hazout S. & Etchebest C. (2004), *Local backbone structure prediction of proteins*, In Silico Biology, 4, 34.

Améliorations des Familles Séquentielles



Nouvelle stratégie ~ recuit simulé avec mémoire des minimums locaux.

Nouveau critère d'apprentissage → les BPs non répétitifs ne doivent pas être noyés dans le gain.

Tentative de couplage avec PSI-PRED (prédiction de structures secondaires).

Améliorations des Familles Séquentielles



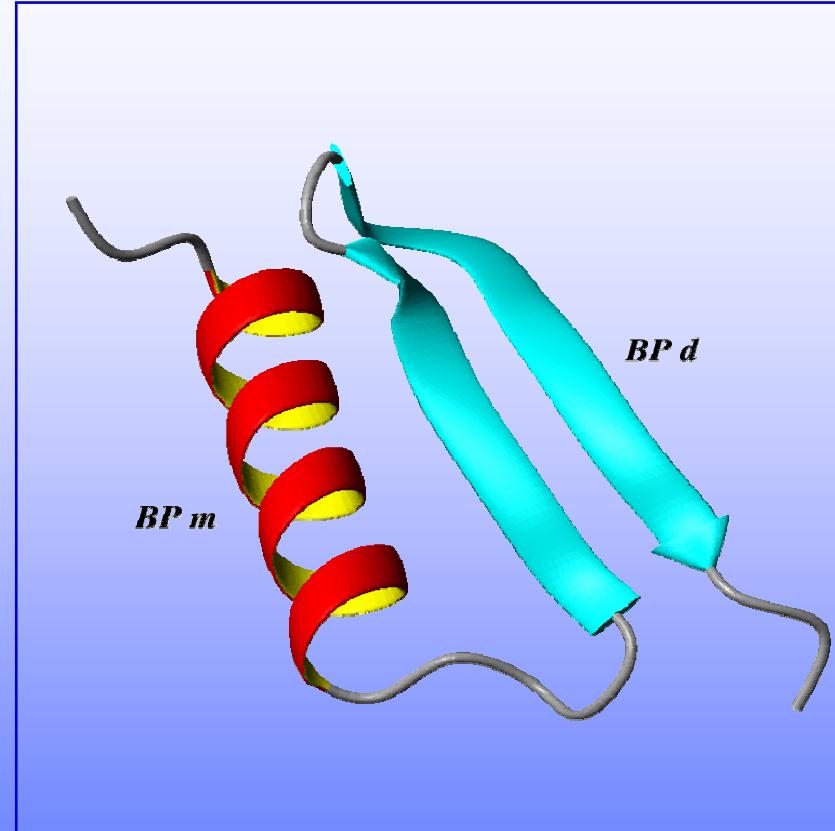
Amélioration des taux de prédictions

Global : Q_{16} (+8%) → 48.7%

répétitifs : 70% (vs. 40%) et 47 % (vs. 29%)

Q_{14} : + 1% [les non - répétitifs]

Prédiction des boucles courtes

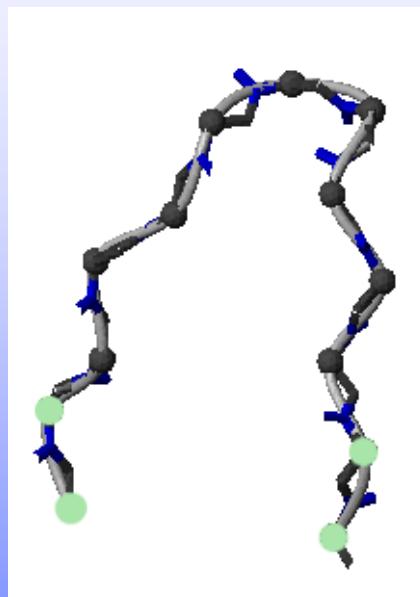


Séries de 2 to 6 PBs entre
deux séries de BPs *mm* et/ou *dd*

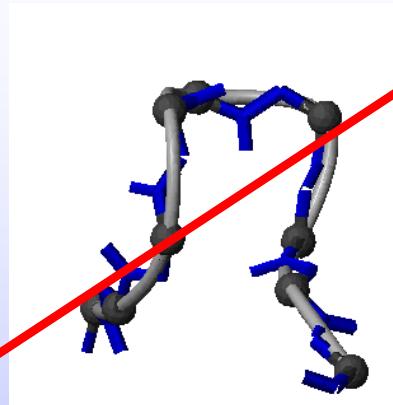
Prédiction des boucles courtes



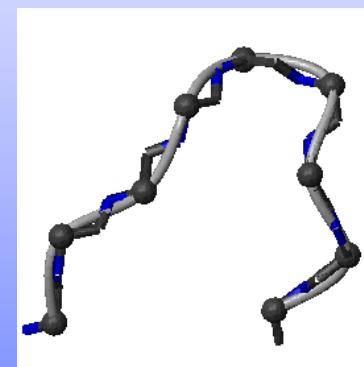
Permet une combinatoire facile



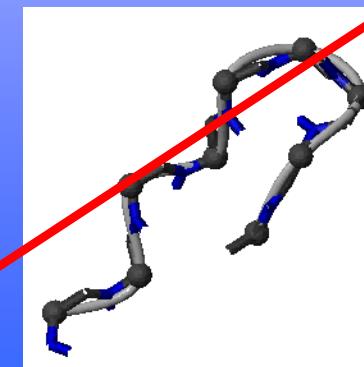
ddehiacdd
HVKVGDTVT
1BXA (34-43)



eojac
CSDHTGTRK
1AWU (53-62)
 $rmsd = 2.2 \text{ \AA}$



ehiac
MSTSVGDRV
12E8 (10-19)
 $rmsd = 0.3 \text{ \AA}$



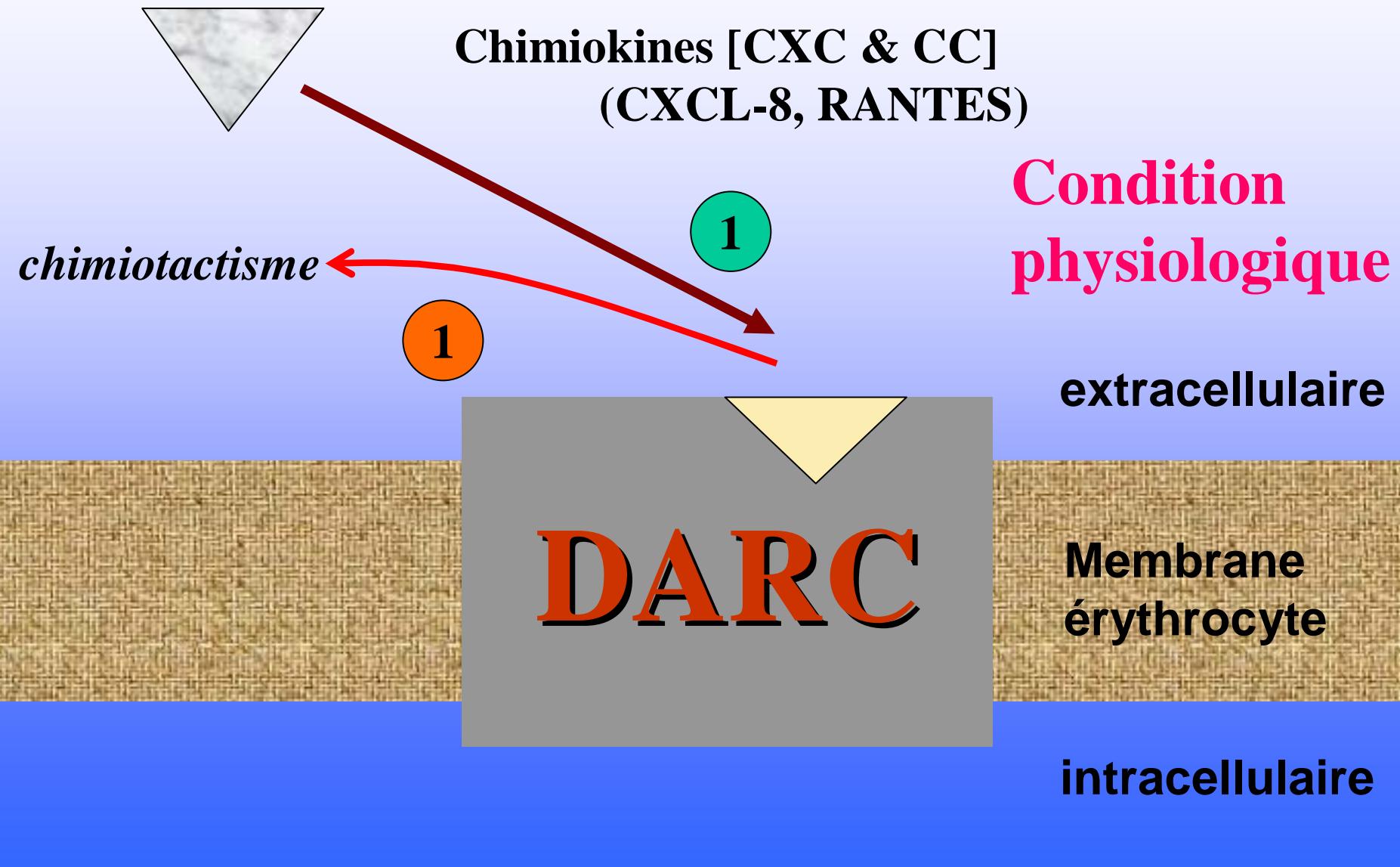
ehjac
TWTMEGNKL
1A57 (65-74)
 $rmsd = 2.5 \text{ \AA}$

DARC : Duffy Antigen / Receptor for Chemokine

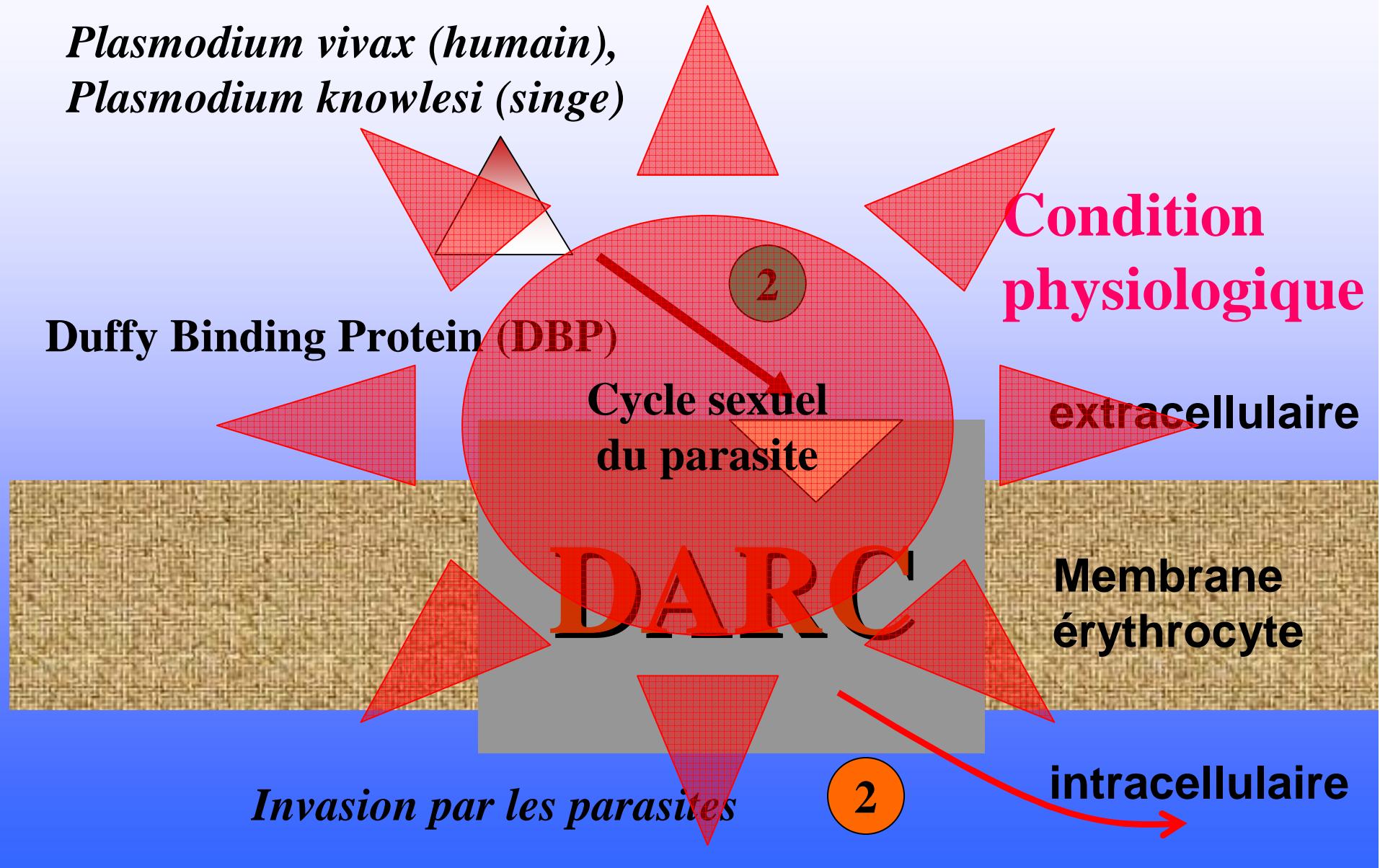
Présent à la surface des cellules sanguines (érythrocytes).

Récepteur silencieux aux chimiokines → pas de transduction du signal !!!

Protéine transmembranaire à 7 segments transmembranaires (hypothèse).



*Plasmodium vivax (humain),
*Plasmodium knowlesi (singe)**



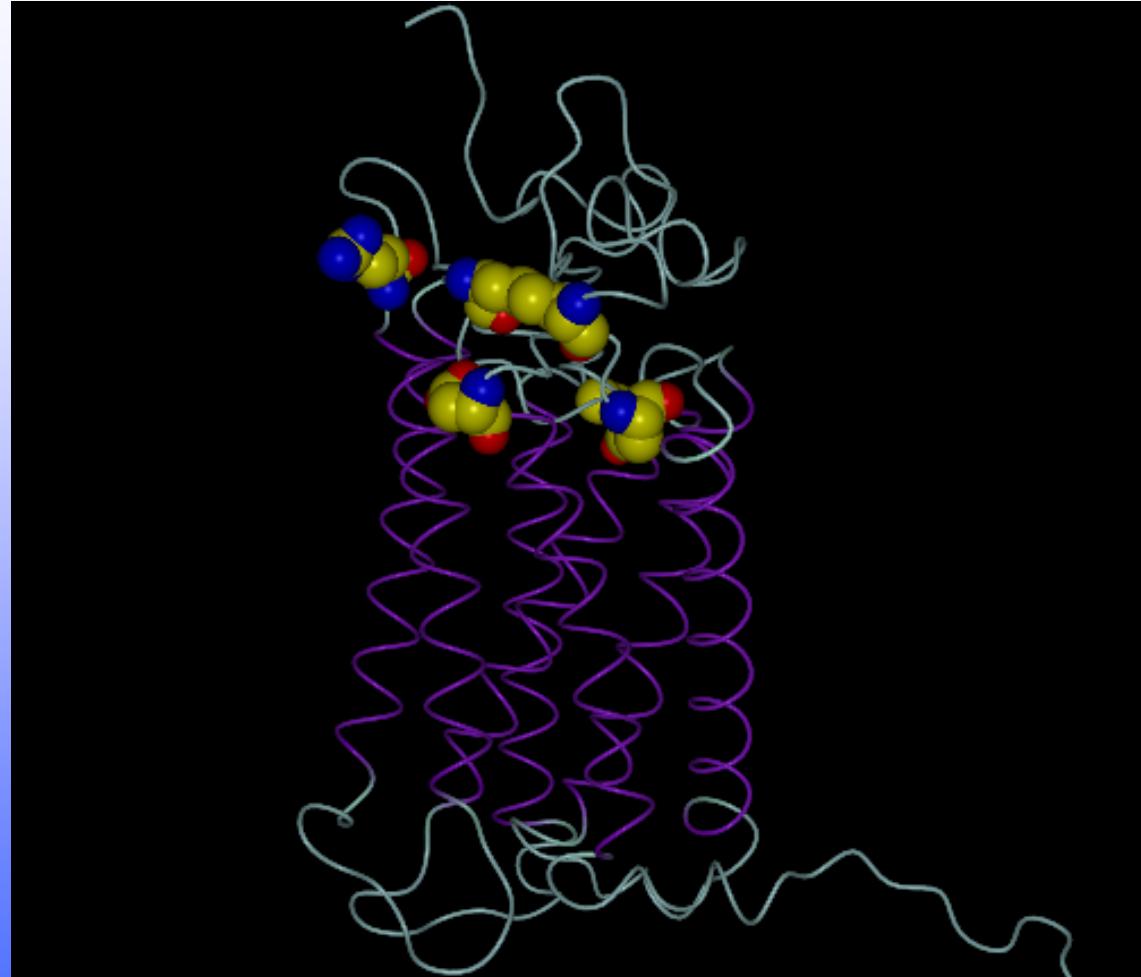
Construction de modèles par modélisation comparative, en utilisant des résultats d'expériences de mutagenèse dirigé.

Méthodes de prédiction de segments transmembranaires,
threading, de novo.

Les Blocs Protéiques ont été utilisé pour analyser la longue extrémité Nterminale (prédiction), puis analyser les résultats de recuit simulés effectués sur la protéine.

→ A permis de mieux analyser la conformation très flexible (moins simple avec DSSP).

Modèle correspondant
aux données
expérimentales.



de Brevern A.G., Wong H., Tournamille C., Colin Y., Le Van Kim C. & Etchebest C.,
A structural model of a seven transmembrane helices receptor: The Duffy Antigen / Receptor for Chemokine (DARC), BBA, en révision.

Conclusion



Les structures secondaires → pas aussi simple que cela

Les alphabets structuraux → autre vision
→ autres possibilités

Remerciements



EBGM

C. Etchebest,
S. Hazout,
C. Benros,

L. Fourrier,
H. Wong,

H. Valadié (CEA Grenoble),
R. Gautier (Univ. Nice).

U726 INSERM (INTS)

Y. Colin,
C. Le Van Kim,
C. Tournamille

LBGM (Réunion)

B. Offmann,
M. Tyagi

MIG INRA

J. Martin
J.-F. Gibrat
A. Marin

Indian Institute of Science (Bangalore)

N. Srinivasan



Institut national
de la santé et de la recherche médicale



Institut national
de la santé et de la recherche médicale



Remerciements



Institut National de la Santé & de la Recherche Médicale

(Health & Medical Care)

University Paris VII

Fondation pour la Recherche Médicale

Groupe de Graphisme & de Modélisation Moléculaire (GGMM)

Appel d'offre Bioinformatique inter EPST
(2001 – 2002) & (2003 – 2004)

- Bystroff, C. and Baker, D. (1998) Prediction of local structure in proteins using a library of sequence-structure motif. *J Mol Biol.* **281**, 565-577. *
- Camproux, A. C. *et al.* (1999). "Hidden Markov model approach for identifying the modular framework of the protein backbone." *Protein Eng* 12(12): 1063-73. *
- de Brevern, A.G., Etchebest, C. and Hazout, S. (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*. **41**, 271-287. *
- de Brevern A.G., Valadié H., Hazout H. & Etchebest C. (2002) Extension of a local backbone description using a structural alphabet. A new approach to the sequence-structure relationship, *Protein Science*, **11**, 2871-2886.
- Karchin, R., Cline, M., Mandel-Gutfreund, Y. and Karplus, K. (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins*. **51**, 504-514. *
- de Brevern, A.G. and Hazout, S. (2003). Improvement of "Hybrid Protein Model" to define an optimal repertory of contiguous 3D protein structure fragments. *Bioinformatics*. **19**, 345-353.
- Camproux, A.C. *et al.* (2004). A hidden markov model derived structural alphabet for proteins. *J Mol Biol* 339(3): 591-605.
- Etchebest C., Benros C., Hazout S. & de Brevern A.G. (2005) A structural alphabet for local protein structures: improved prediction methods *Proteins* in press. *
- de Brevern A.G. (2005) New assessment of a structural alphabet, *In Silico Biology*, **5**:26.
- de Brevern A.G., Wong H., Tournamille C., Colin Y., Le Van Kim C. & Etchebest C., "A structural model of a seven transmembrane helices receptor: The Duffy Antigen / Receptor for Chemokine (DARC)". *