



Detecting unfolded regions in protein sequences

Anne Poupon
Génomique Structurale de la Levure
IBBMC
Université Paris-Sud / CNRS
France

Large proteins and complexes: a domain approach



- **Structural studies large proteins**
 - difficult to produce in sufficient quantities
 - difficult to crystallize (inter domain mobility)
- **Domains are the evolutionary units of proteins**
 - discovery of new domains
 - discovery of new structures/functions
- **Domain approach rewarding for the study of complexes**
 - combined with electron microscopy



objectives

bioinformatics tools for

- the automatic detection of domains
- identification linker regions

genetic engineering for

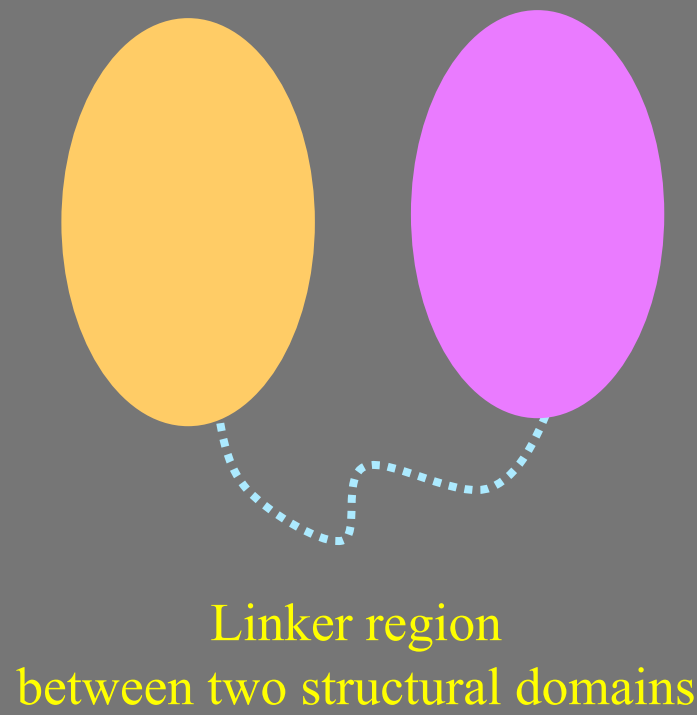
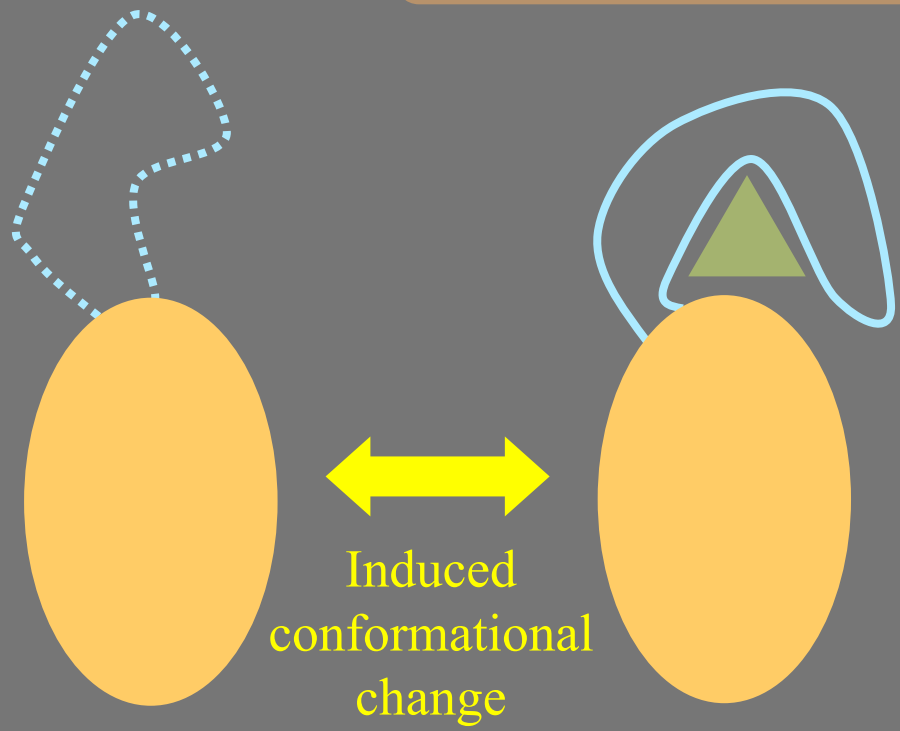
- creation of meaningful fragment libraries
- screening for soluble folded fragments

apply to

- large proteins involved in DNA remodeling/repair
- multiprotein complexes

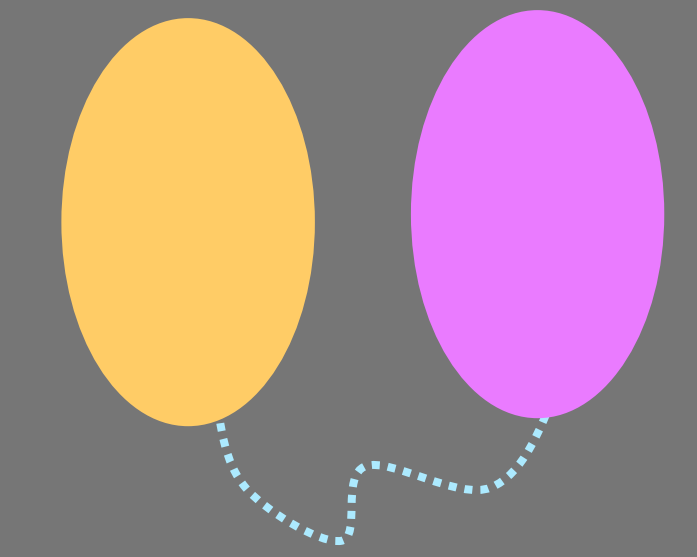
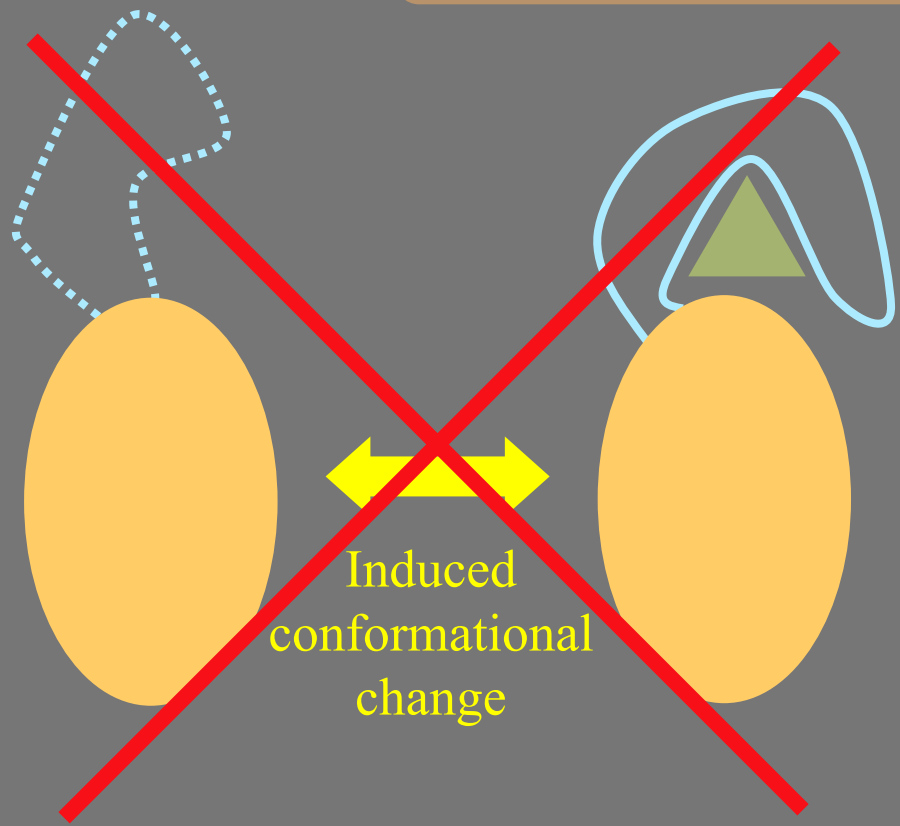


Unfolded fragments





Unfolded fragments



Linker region
between two structural domains

Unfolded fragments and crystallography



Most prediction methods predict regions that **might be unfolded under certain conditions**

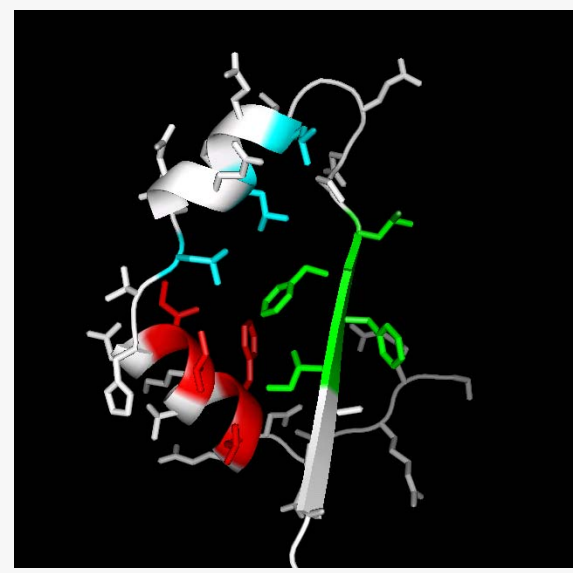
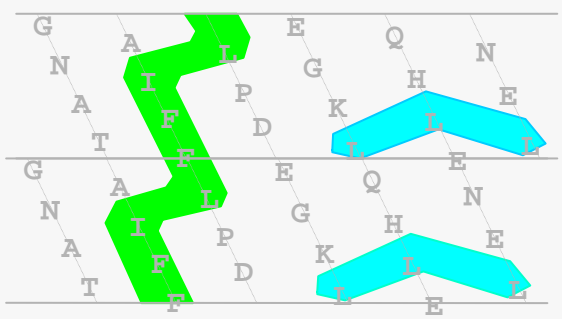
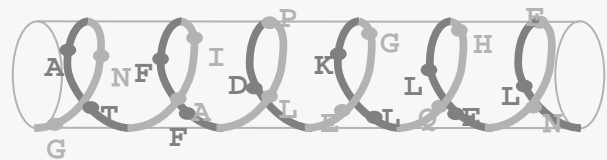
We want to predict regions that are **always unfolded**



Principles and limits of HCA

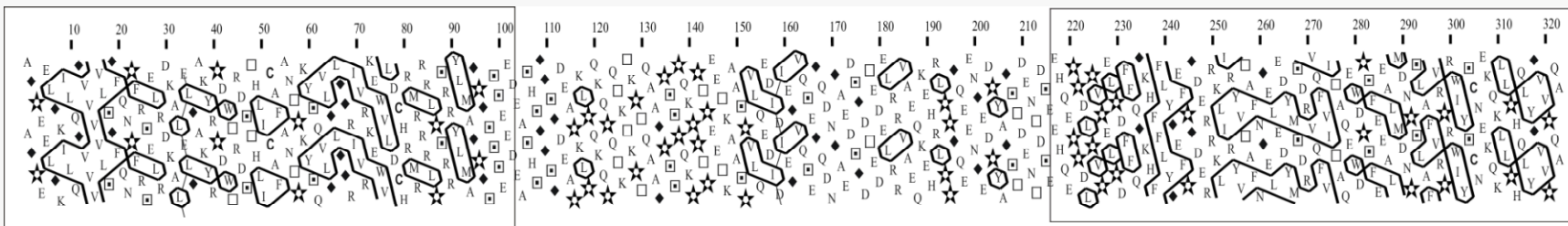
246 ... GNATIFFLPDEGKLGKLENELETHDIIITKFLNEDRRS... 283

246 ... 00001111100000100100010001100110000000... 283





Principles and limits of HCA



Folded

Unfolded

Folded

Requires a lot of manpower and expertise
Not automated
No statistics



Principles and limits of HCA

« Why » do we see the unfolded regions on the HCA diagram ?

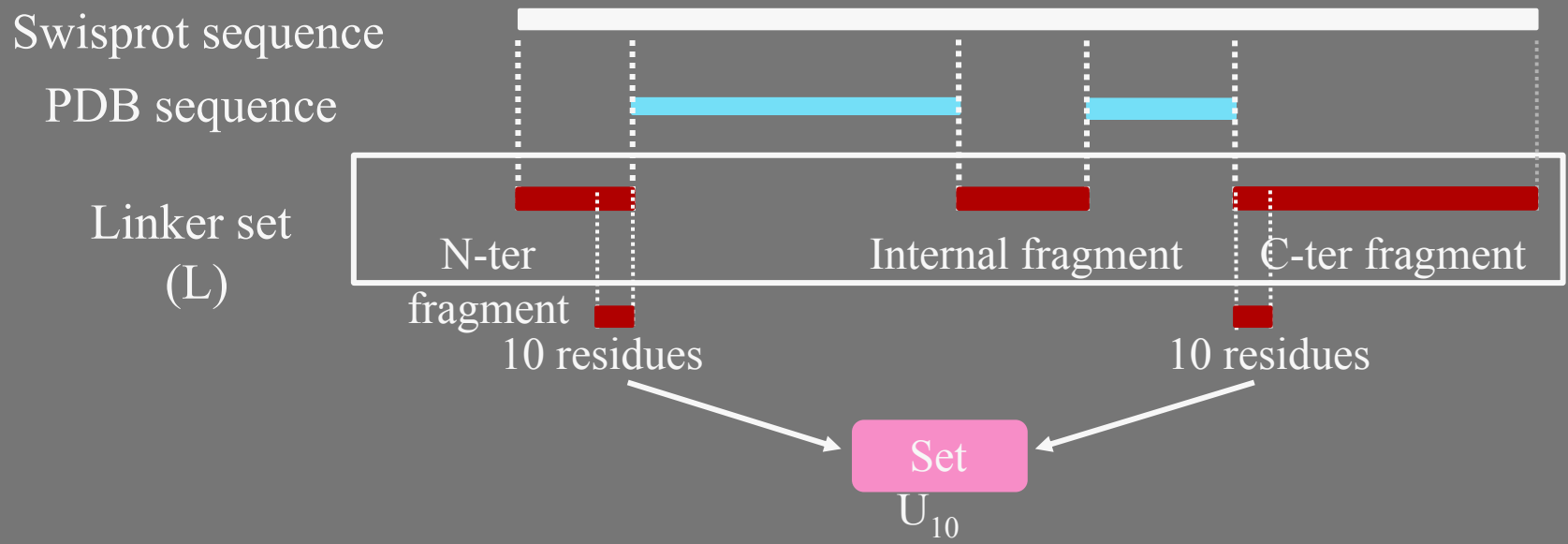
Because the « textures » of folded and unfolded regions are different

- The amino acid composition is different
- The hydrophobic clusters are different

These differences have to be quantified and exploited



Data set



Prelink	Disopred
24686 chains	750 chains
35438 aa	4590 aa



Data set

PDB seq 1



PDB seq 2



PDB seq 3



PDB seq 3



Only sequences that are unfolded in all homolog PDB sequences are retained



Amino acid compositions



246 ... GNATIFFFLPD**E**GKQLHLENELTHTDIIITKFLNEDRRS ... 283

← 21aa →

Probabilities of occurrence of this 21 aa sequence in linker (PL) and structured (PS) regions

$$PL = \frac{n!}{n_V! n_I! \dots n_G!} pl_V^{n_V} pl_I^{n_I} \dots pl_G^{n_G}$$

$$PS = \frac{n!}{n_V! n_I! \dots n_G!} ps_V^{n_V} ps_I^{n_I} \dots ps_G^{n_G}$$

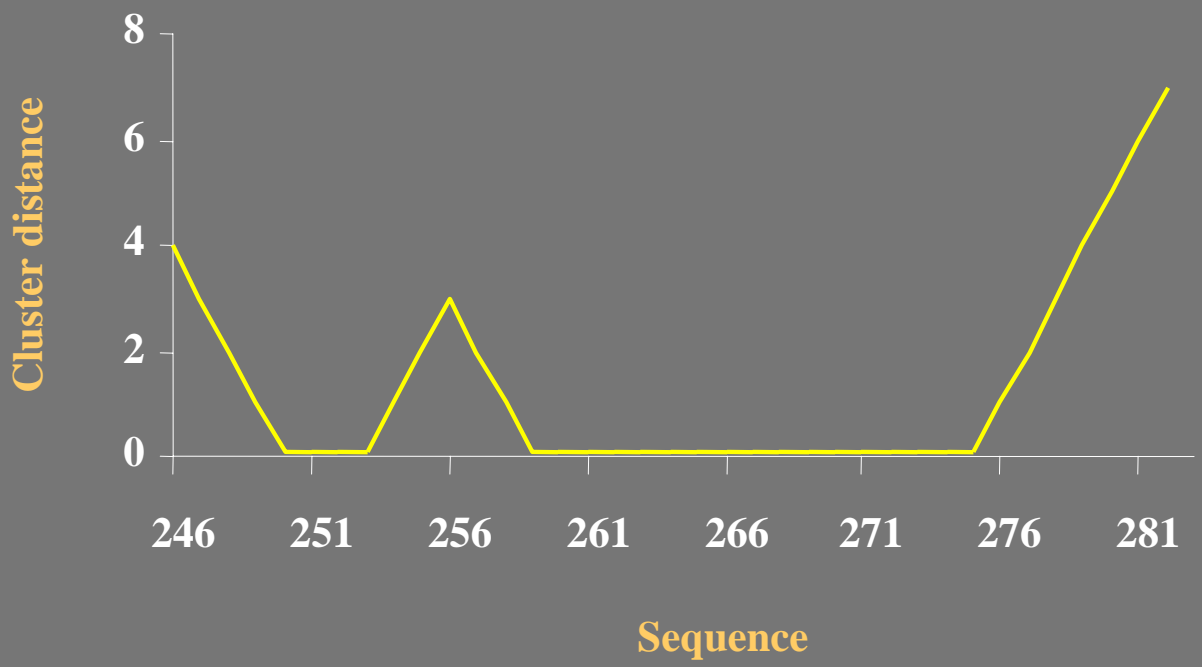


Differences in hydrophobic clusters

246 ... GNAT I FFLPDEGKLQHLENELTHD I I TKFLENEDRRS... 283

246 ... 0 0 0 0 **1 1 1 1** 0 0 0 0 0 **1 0 0 1 0 0 0 1 0 0 0 1 1 0 0 1 1** 0 0 0 0 0 0 0 ... 283

246 ... 4 3 2 1 **0 0 0 0** 1 2 3 2 1 **0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0** 1 2 3 4 5 6 7 ... 283



Computed parameters



246 ... GNATIFFLPD**E**GKQLHLENELTHDII TKFLENEDRRS ... 283



21aa



For this residue

Ratio $R = PL/PS$

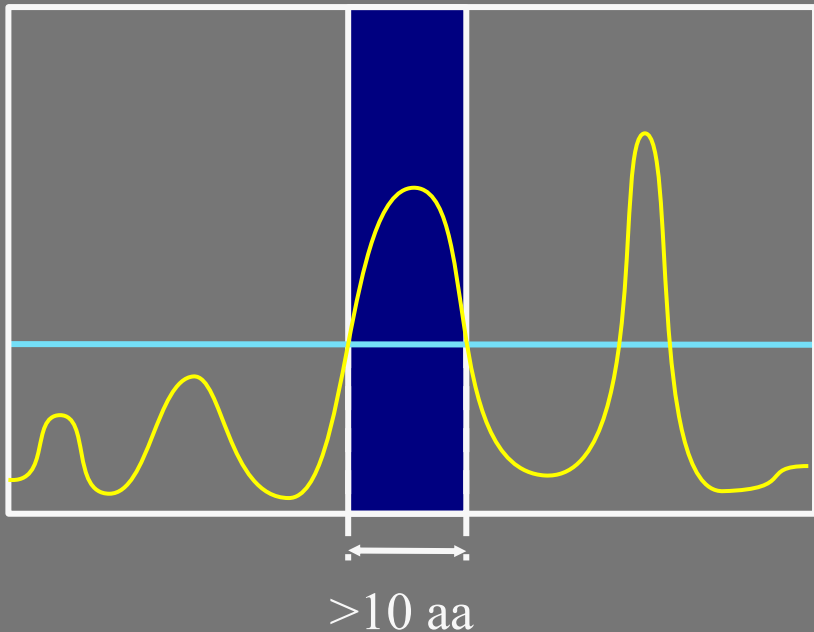
Cluster distance d

Product $P = R.d$

These 3 parameters are plotted for all residues in a protein sequence



Calibrating on PDB



— ratio

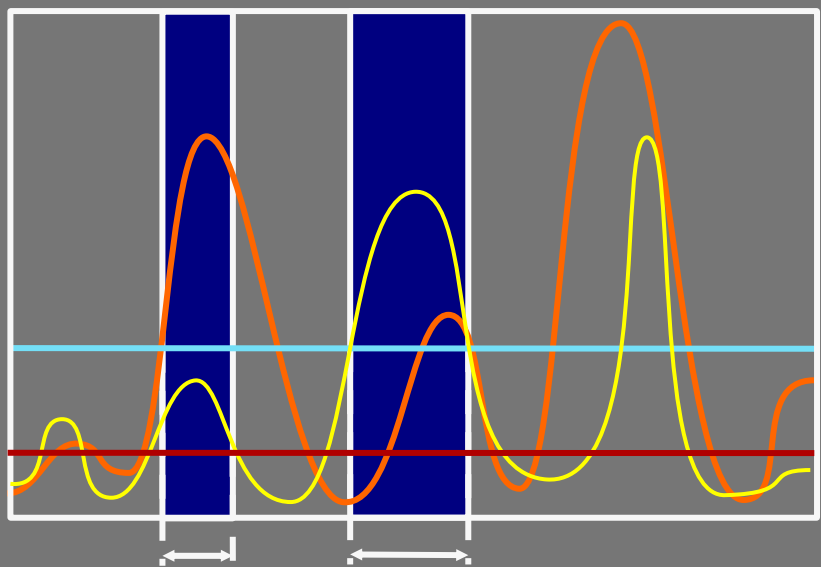
FP : false positives
real FP : the regions that are unfolded under certain conditions are removed

Rule 1

	$R \geq 10$	$R \geq 10$ ≥ 10	R
FP	3	6	34
real FP	0	0	1



Calibrating on PDB



— ratio R
— product P

10
5

>10 aa >10 aa
5 < ratio < 10
product > 10

Rule 1

Rule 2

				5 ≤ R ≤ 10	
	R ≥ 30	R ≥ 20	R ≥ 10	P ≥ 10	P
	≥ 5				
FP	3	6	34	116	49
real FP	0	0	1	6	1



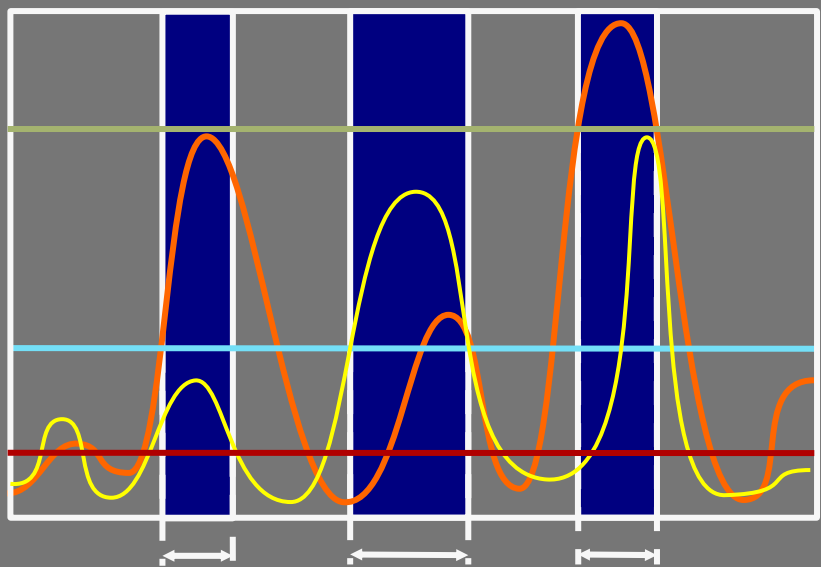
Calibrating on PDB

Length <10 10-20 20-30 >30
total

N-ter Fragments					
Nb fragments	485	340	479	1234	
2538					
Predicted (%)	22.1	28.8	35.5	59.9	
43.8					
C-ter Fragments					
Nb fragments	684	222	102	911	
1919					
Predicted (%)	23.7	26.1	36.3	70.5	
46.8					
Internal fragments					
Nb fragments	301	126	30	18	



Calibrating on PDB



— ratio R
— product P

Rule 3 : $P \geq 30$

>10 aa
5 < ratio < 10
product > 10

product length

<10 10-20 20-30 >30

total	<10	10-20	20-30	>30
Nbfragments	301	126	30	18
475				
Rules 1+2	35.9	41.3	51.3	61.1
40.4				
Rules 1+2+3	53.2	79.4	90	94.4
64.2				



Testing on CASP5

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \rightarrow 1$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \rightarrow 1$$

Group	Sensitivity (%)	Sensibility (%)
20	56	97
68	57	92
355	67	77
454	50	90
Prelink	87	99.7



Testing on Yeast proteins

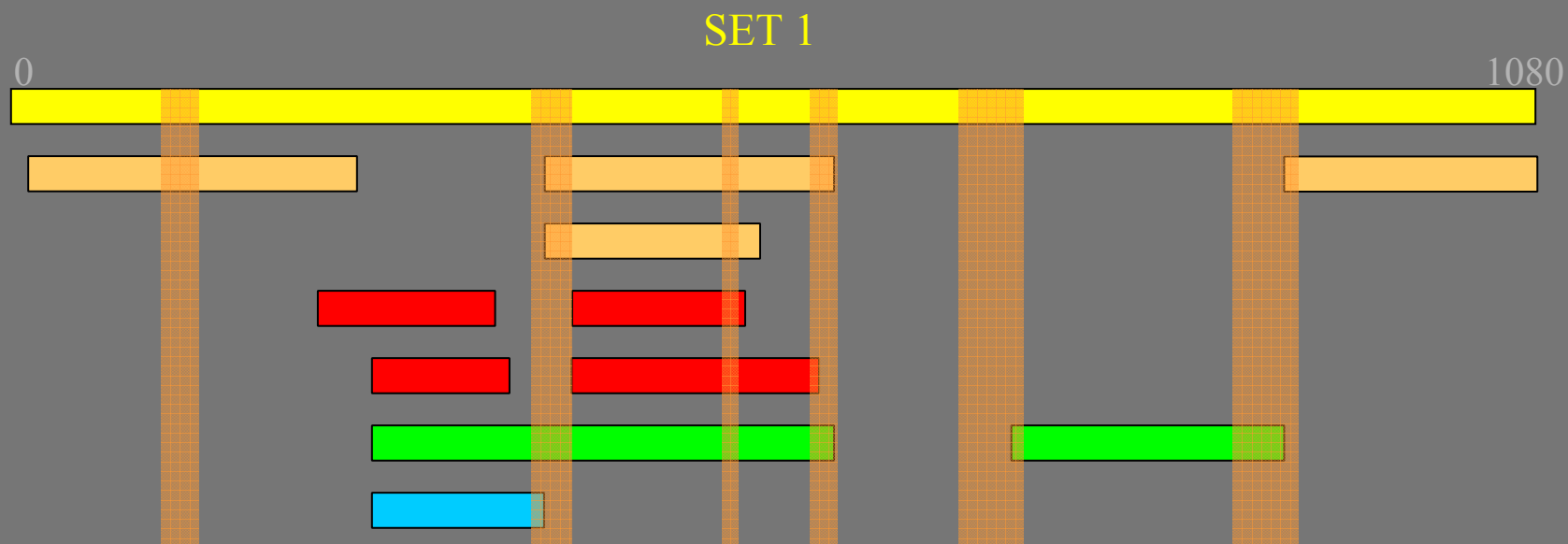
	Nb proteins	Nb with linker predicted	False positives	False negative
Unfolded		14	7	0
Structured	16	1	0	0
Soluble not crystallized	45	21	-	-

*Chorismate synthase
(saccharomyces cerevisiae)*





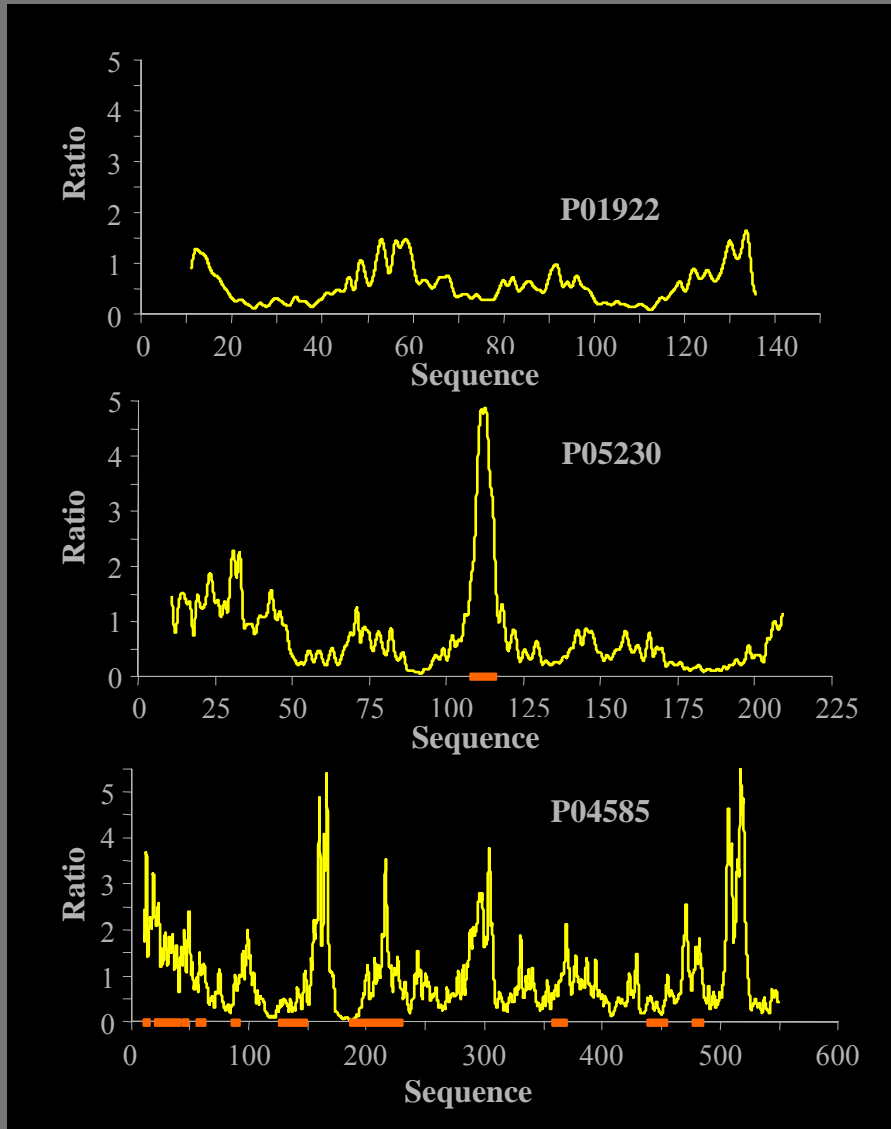
Yeast DNA repair protein SET1



Not expressed
Not soluble

Soluble
Crystallized (structure)

Limitations



How to account for the « shape » of the curve ?

Perspectives



- Not all parameters have been optimized:
 - window length
 - thresholds
- More parameters are needed
 - neighborhood
- The rules have been determined empirically
 - We should use statistical learning algorithms
 - We should also try tree learning algorithms

Thanks to ...



Karen Coeytaux

Herman van Tilbeurgh
Joël Janin

Julie Bernauer
Sophie Cheruel

... and the rest of the Yeast Structural Genomics team.