Structural-alphabet motifs in protein loop structures: from structure to function

Leslie Regad

Laboratoire Molécules Thérapeutiques *in Silico*

Université Paris Diderot, Paris 7

Functional annotation of proteins

>1OX1:A|PDBID|CHAIN|SEQUENCE
MSLRIPRIYHPISLENQTQCYLSEDAANHVARVLRMTEGEQLE
LFDGSNHIYPAKIIESNKKSVKVEILGRELADKESHKIHLGQV
ISRGERMEFTIQKSVELGVNVITPLWSERCGVKLDAERMDKKI
QQWQKIAIAACEQCGRNIVPEIRPLMKLQDWCAENDGALKLNL
HPRAHYSIKTLPTIPAGGVRLLIGSEGGLSAQEIAQTEQQGFT
EILLGKRVLRTETASLAAISALQICFGDLGEEGGSHHHHHH

FUNCTION?

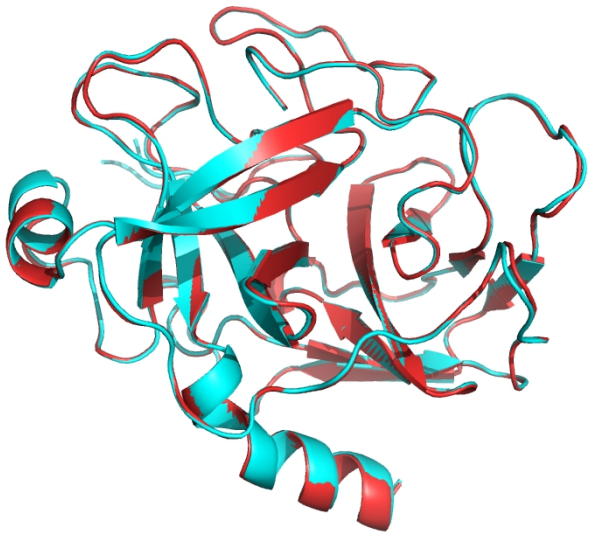
Functional annotation of proteins

>1OX1:A|PDBID|CHAIN|SEQUENCE
MSLRIPRIYHPISLENQTQCYLSEDAANHVARVLRMTEGEQLE
LFDGSNHIYPAKIIESNKKSVKVEILGRELADKESHKIHLGQV
ISRGERMEFTIQKSVELGVNVITPLWSERCGVKLDAERMDKKI
QQWQKIAIAACEQCGRNIVPEIRPLMKLQDWCAENDGALKLNL
HPRAHYSIKTLPTIPAGGVRLLIGSEGGLSAQEIAQTEQQGFT
EILLGKRVLRTETASLAAISALQICFGDLGEEGGSHHHHHH

FUNCTION?

Search of homologous proteins

99% sequence identity
> 1ox1_A
TLPTIPAGGVRLLIGSEGGLSAQEIAQTEQQGFTEILLGK
> 1sfv_A
TLPTIPAEGVRLLIGSEGGLSAQEIAQTEQQGFTEILLGK



serine protease

Functional annotation of proteins

>1OX1:A|PDBID|CHAIN|SEQUENCE
MSLRIPRIYHPISLENQTQCYLSEDAANHVARVLRMTEGEQLE
LFDGSNHIYPAKIIESNKKSVKVEILGRELADKESHKIHLGQV
ISRGERMEFTIQKSVELGVNVITPLWSERCGVKLDAERMDKKI
QQWQKIAIAACEQCGRNIVPEIRPLMKLQDWCAENDGALKLNL
HPRAHYSIKTLPTIPAGGVRLLIGSEGGLSAQEIAQTEQQGFT
EILLGKRVLRTETASLAAISALQICFGDLGEEGGSHHHHHH

FUNCTION?

Search of homologous proteins

Pbl: no homologous proteins with known function

Methods based on the extraction of structural motifs involved in functional site

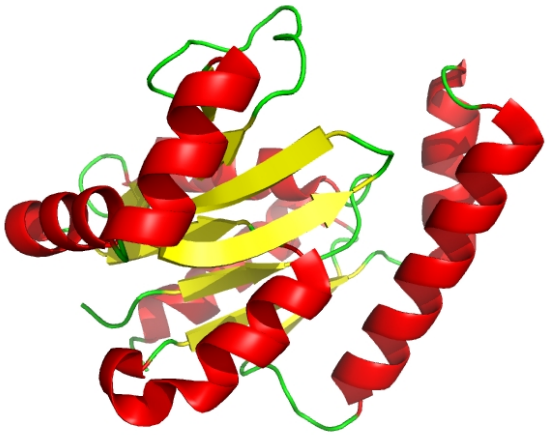
Extraction of functional structural motifs

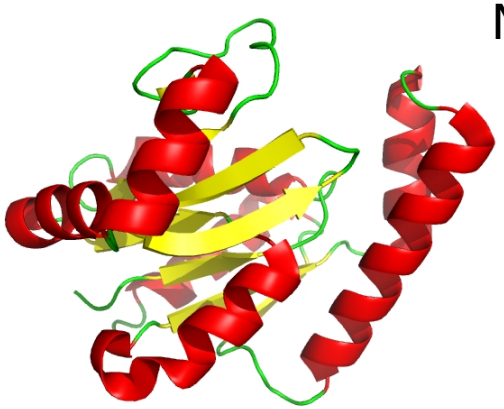
- For a given functional site (Sodhi et al., 2004; Nebel et al., 2007; Bordner, 2008; Halperin et al.; 2008):
  - Structural motifs specific to this functional site
  - Knowledge of the location of this functional site

ATP binding site

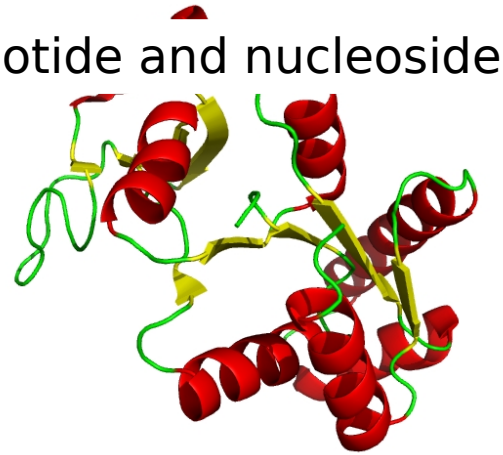
Extraction of functional structural motifs

- For a given functional site (Sodhi et al., 2004; Nebel et al., 2007; Bordner, 2008; Halperin et al.; 2008):
  - Structural motifs specific to this functional site
  - Knowledge of the location of this functional site
- Structural motifs specific to a protein group (Ausiello et al., 2008; Polacco et al.; 2008; Esapdaler et al., 2004; Tendulkar et al., 2004; Manikandan et al., 2008; Wu et al., 2010)
  - Extraction of new functional motifs





Nucleotide and nucleoside kinase



Extraction of functional structural motifs

- For a given functional site (Sodhi et al., 2004; Nebel et al., 2007; Bordner, 2008; Halperin et al.; 2008):
  - Structural motifs specific to this functional site
  - Knowledge of the location of this functional site
- Structural motifs specific to a protein group (Ausiello et al., 2008; Polacco et al.; 2008; Esapdaler et al., 2004; Tendulkar et al., 2004; Manikandan et al., 2008; Wu et al., 2010)
  - Extraction of new functional motifs
- Extraction of structural motifs
  - Structural/sequence Alignments
  - Geometric parameters: RMSd, distance ...

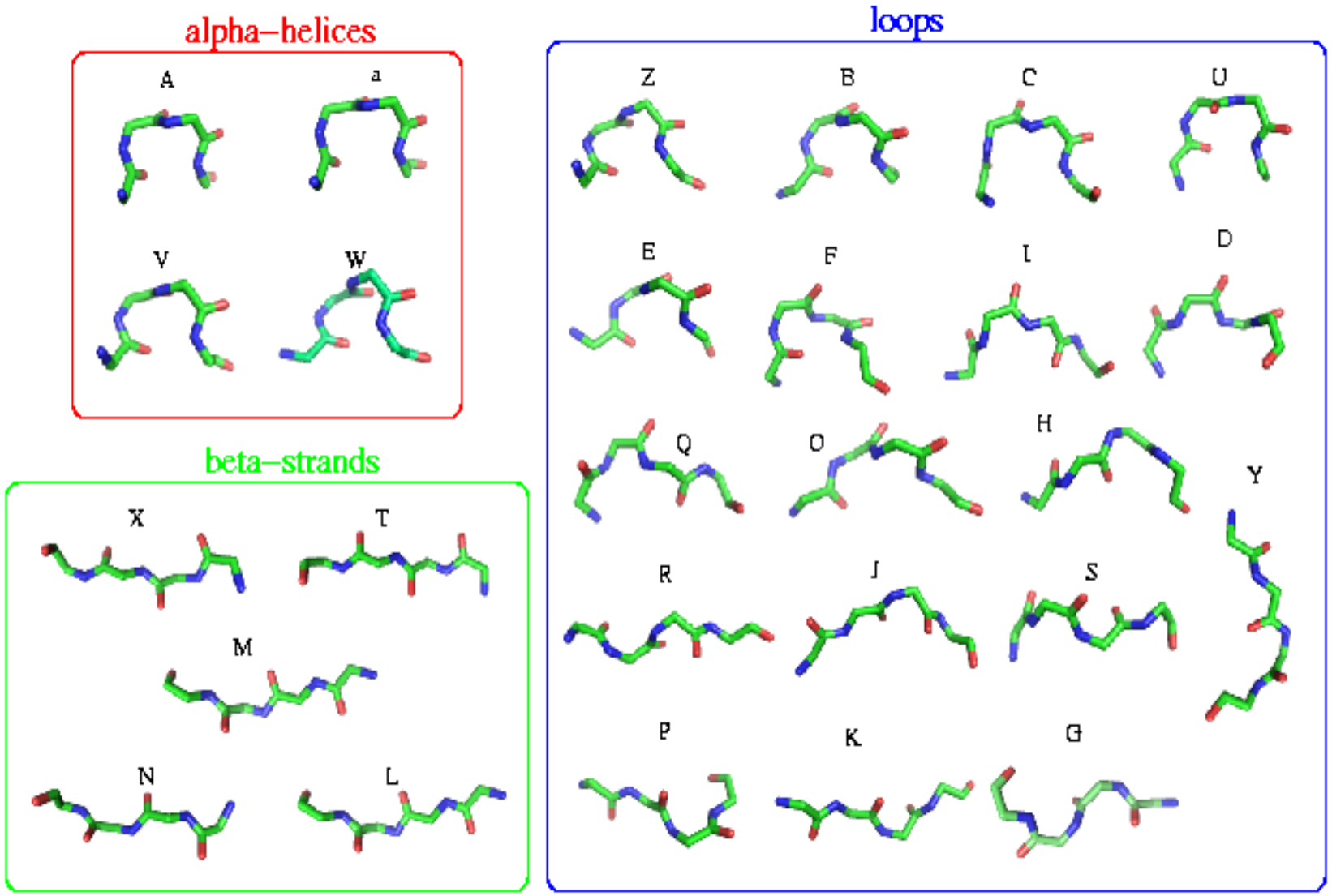
Strategy of extraction of functional structural motifs

STRUCTURAL MOTIFS SPECIFIC to a FUNCTION

Structural words

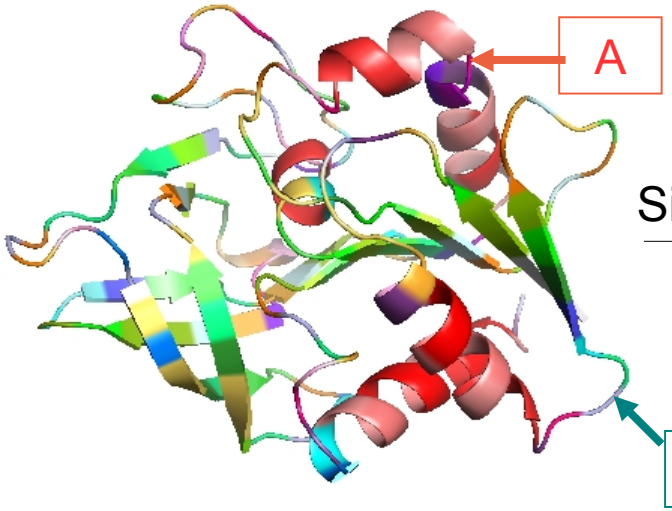
- Simplification of structures into string
- Extraction of recurrent words of 4 letters (Regad et al., 2010)

Structural alphabet HMM-SA (Camproux et al., 2004)



Structural alphabet HMM-SA (Camproux et al., 2004)

3D structure



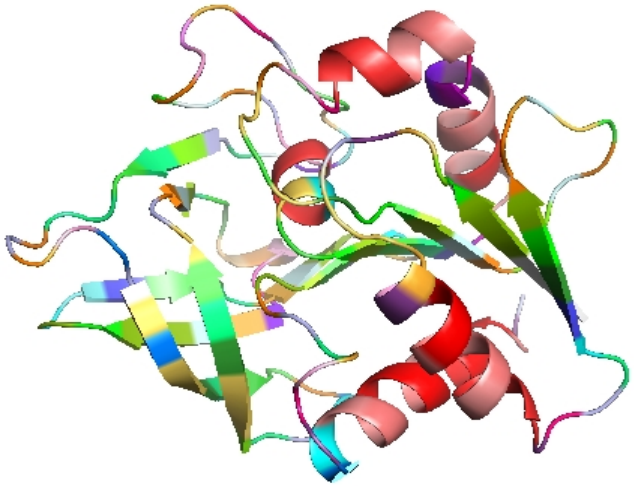
Simplification by HMM-SA

Encodage de 1gpw_B

TTMLGEQJJJGEBAAVWVWaaaEZQKUS
XMTPPGIJNPILLPQKGBZERUITMMXJMJ
JPBZaaaVWVWZDOEWVWAAVWVWDS
KNTMNJYFOIKQHQPBRQPRaa...

Structural alphabet HMM-SA (Regad et al., 2008, 2010)

3D structure



Simplification by HMM-SA

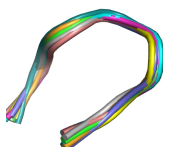
Encodage de 1gpw_B

TTMLGEQJJJGEBAAVWVWaaaEZQKUS
XMTPPGIJNPILLPQKGBZERUITMMXJMJ
JPBZaaaVWVWZDOEWVWAAVWVWDS
KNTMNJYFOIKQHQPBRQPRaa...

Extraction of words: 7 residue fragments

GBZE	PQKG
BZEJ	QKGB
ZEJG	KGBZ
EJGE	GBZE

GBZE



PQKG

Strategy of extraction of functional structural motifs

STRUCTURAL MOTIFS SPECIFIC to a FUNCTION

Mots structuraux

- Simplification of structure into string
- Extraction of recurrent words of 4 letters (Regad et al., 2010)

Statistic over-representation

Functional sites = site with an unexpected occurrence

(Rocha et al., 1998; van Helden et al., 2000)

Over-representation: SPatt (Nuel et al., 2010)

- *Over-representation scoreof a word* w $(Lp(w))$

$$Lp(w) = -log_{10}[P(N^{exp}(w) > N^{obs}(w))]$$

$N^{exp}(w)$: *Expected occurrence of a word* w

$N^{obs}(w)$: *Observed occurrence of a word* w

- $Lp(w) > threshold \Rightarrow over\text{-}represented\ word$

- Exact statistics
- Adapted for data sets composed of large number of sequence
- *Poster n°17*

Strategy of extraction of functional structural motifs

STRUCTURAL MOTIFS SPECIFIC to a FUNCTION

Mots structuraux

- Simplification of structure into string
- Extraction of recurrent words of 4 letters (Regad et al., 2010)

Statistic over-representation

Functional sites = site with an unexpected occurrence (Rocha et al., 1998; van Helden et al., 2000)

SCOP Superfamilies

- Hierarchical classification of proteins (Murzin et al., 1995)
- Superfamilies ⊂ proteins with similar structures and functions

=
Over-represented word in a superfamily

Over-represented words in SCOP superfamilies

Loops extracted from 4924 proteins
and classified according superfamiiles

48726
YUODQ
FFFI
GYUQ
GIPQHQ
GYUQZD
URNHBBQ

52540
PZCDS
YUODO
PGBZDGUO
YUODO
GYUQ
QYUODO

SPatt

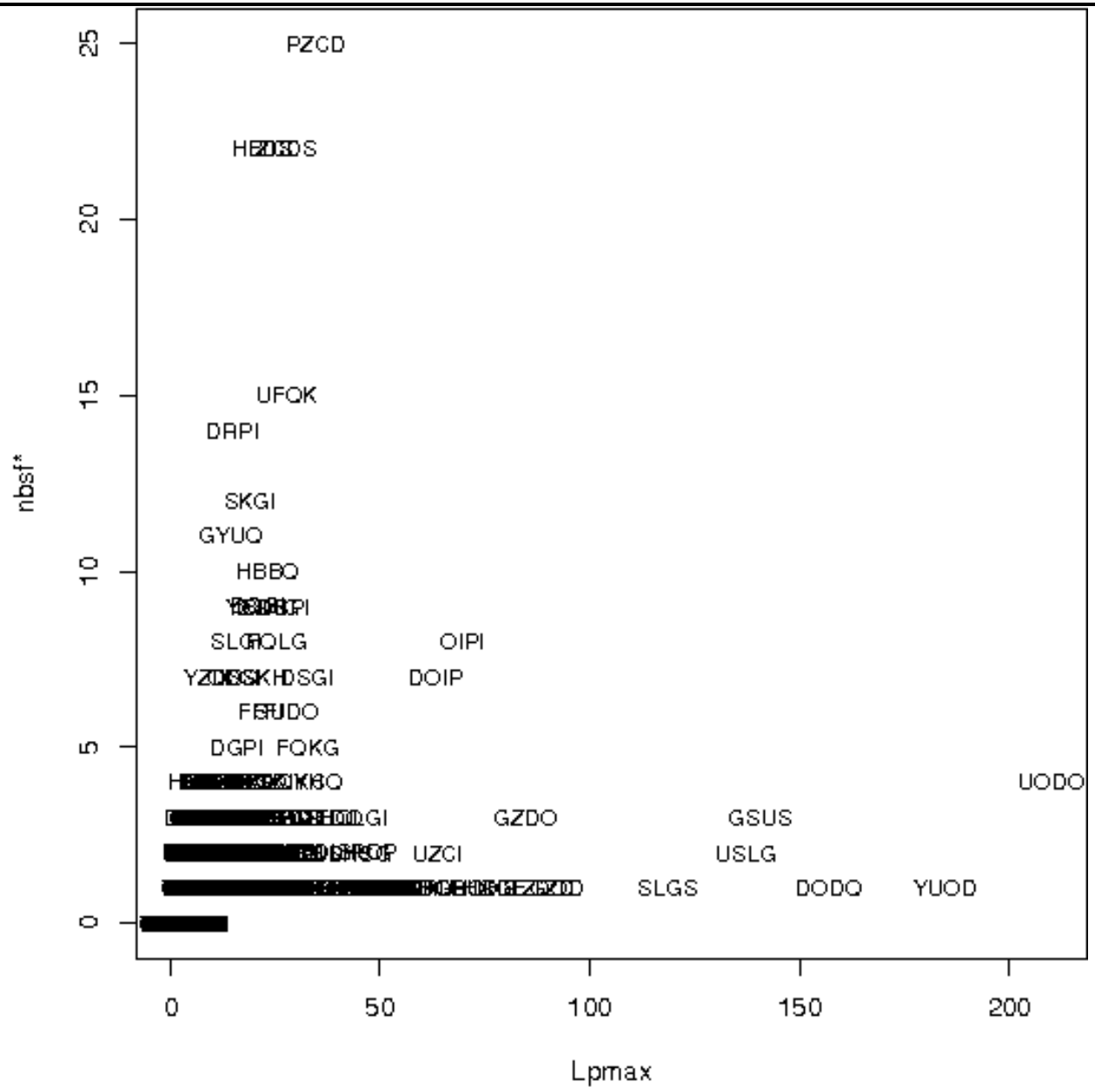
Nuel, Regad et al., 2010

Word	superfamily	Lp
YUOD	48726	0.5
YUOD	52540	*95**
GYUQ	48726	*12**
GYUQ	52540	*7**

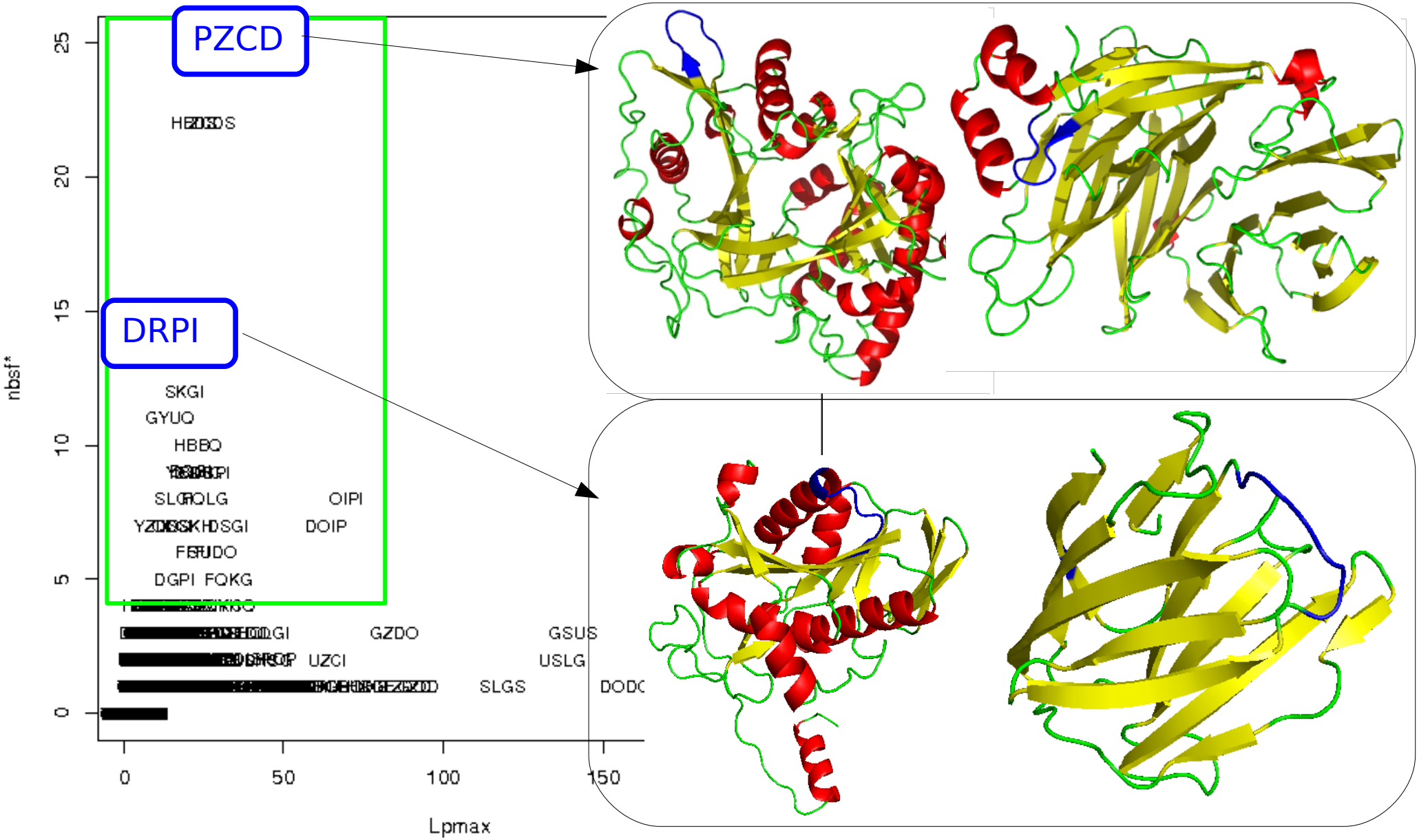
word	Lpmax	nb_sf*
YUOD	95	1
GYUQ	12	2

- Lpmax = maximal over-represented scores in all superfamilies
- nb_{sf*} = number of superfamilies where a word is over-represented

Over-representation of the 11294 words in SCOP superfamilies



ubiquitous words: over-represented in lot of superfamilies

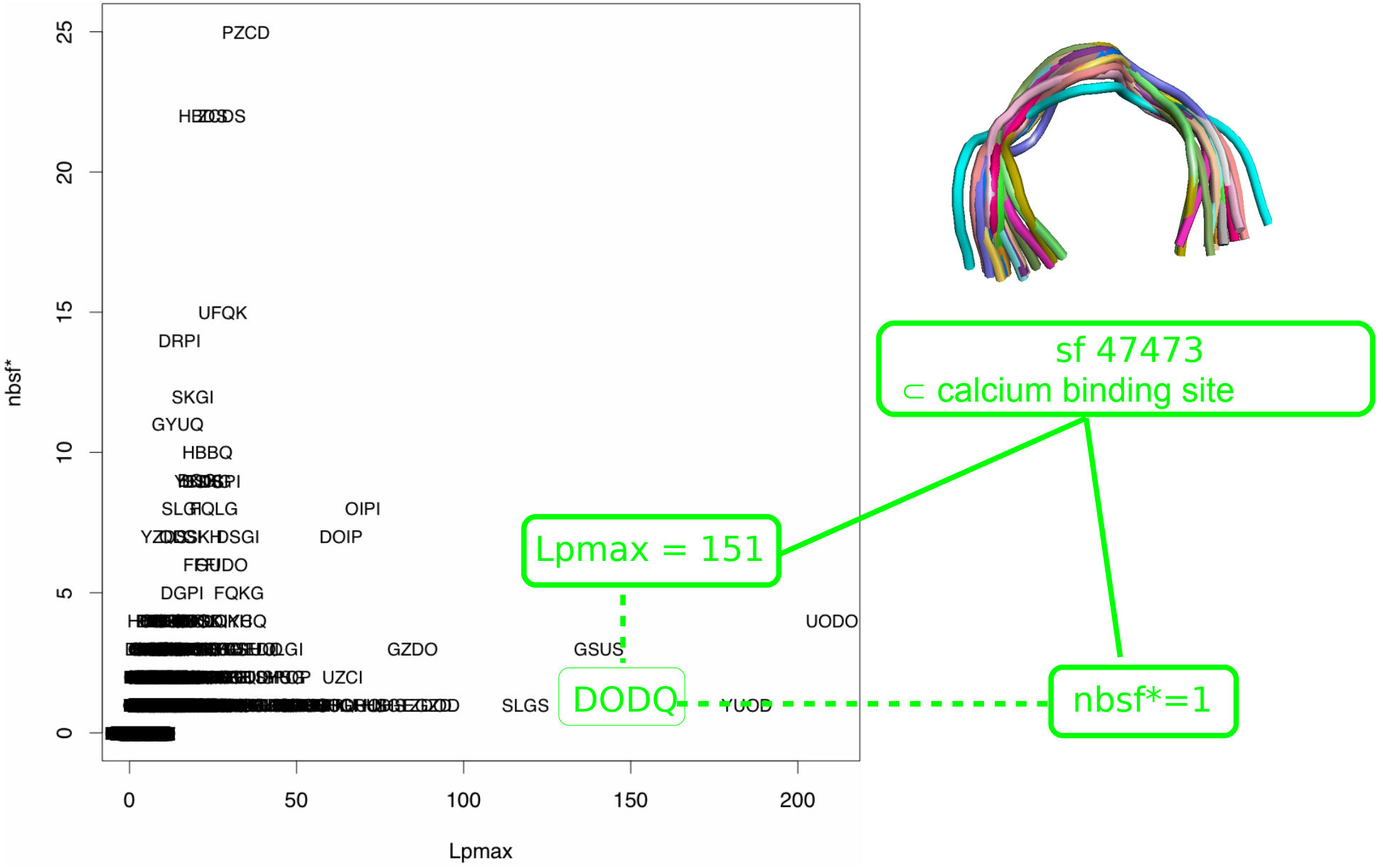


Conclusion motifs ubiquitaires

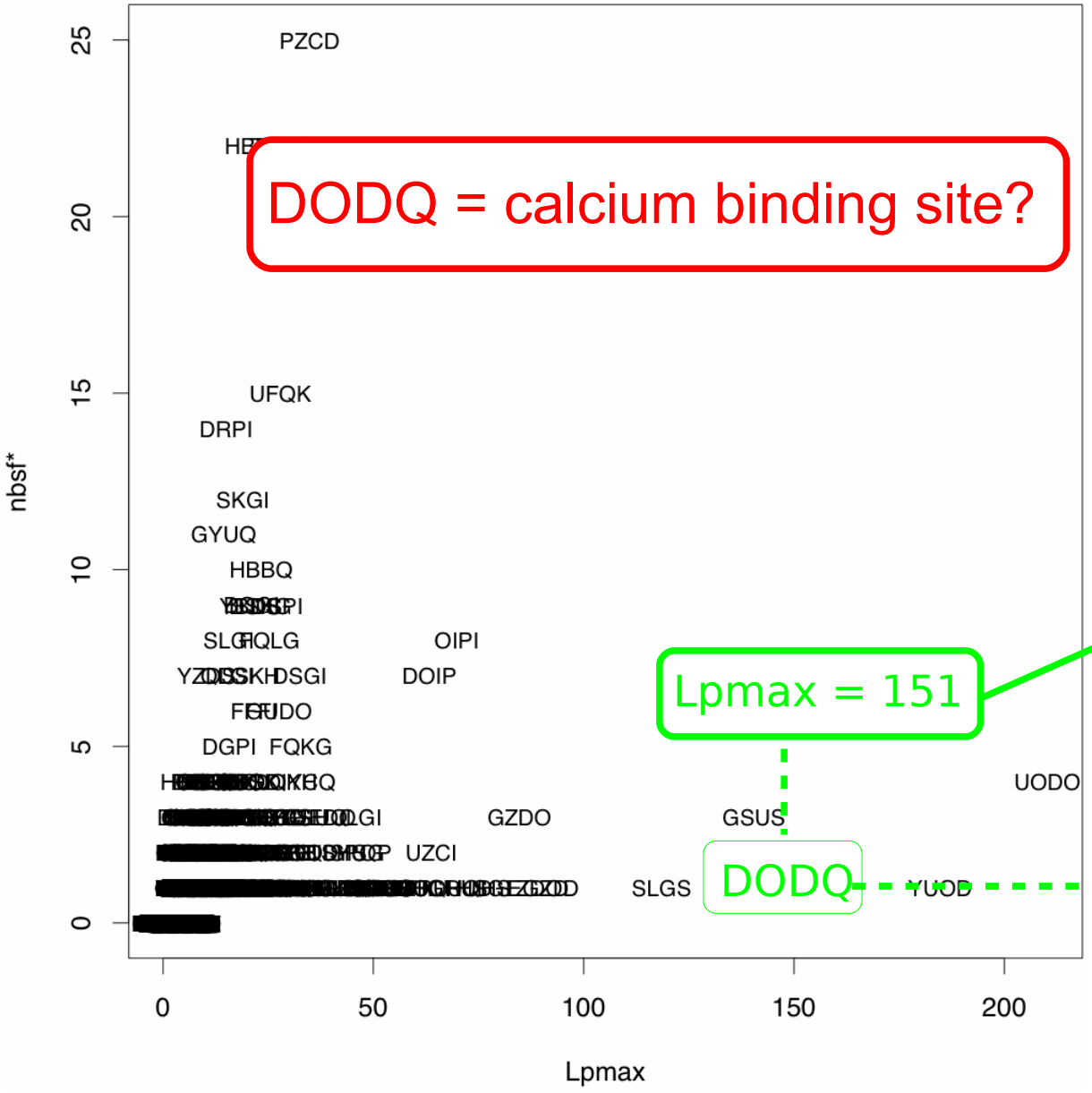
- Structural words could include some small known motifs
  - Turns (Hutchison et al., 1994)
  - Niche (Torrance et al., 2009)
  - Nest (Watson et al.,2002)

- Conserved motifs in proteins seen in different superfamilies
  - Structural role
  - Important for protein folding ?

Example of specific word: DODQ



Over-represented words in SCOP superfamilies



Function of motifs specific to a superfamily

- Swiss-Prot = Database of protein annotations

Positions of important fragments (structure, sequence, function)

Pairing between specific words and Swiss-Prot annotation

DODQ:
23 fragments

1acc	178 184
1alv_A	151 157
1alv_A	181 187
1aui_B	36 42
1aui_B	68 74
1aui_B	105 111
1aui_B	146 152
1bx4_A	250 256
1dab_A	558 564

Calcium binding	109 – 117	9	1
Calcium binding	150 – 161	12	2
Calcium binding	180 – 191	12	3

Calcium binding	31 – 42	12	1
Calcium binding	63 – 74	12	2
Calcium binding	100 – 111	12	3
Calcium binding	141 – 152	12	4

Function of motifs specific to a superfamily

- Swiss-Prot = Database of protein annotations

Positions of important fragments (structure, sequence, function)

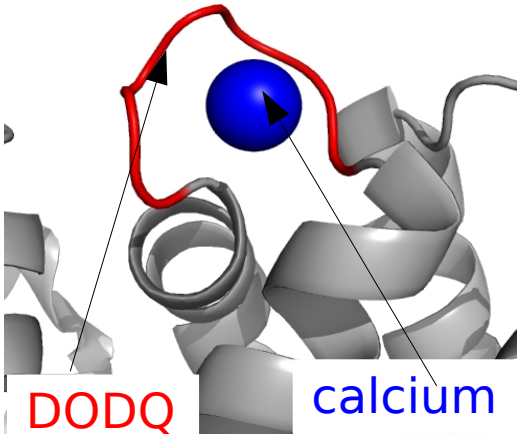
Pairing between specific words and Swiss-Prot annotation

DODQ:
23 fragments

1acc	178 184
1alv_A	151 157
1alv_A	181 187
1aui_B	36 42
1aui_B	68 74
1aui_B	105 111
1aui_B	146 152
1bx4_A	250 256
1dab_A	558 564

14 fragments (=61%) annotated by "*calcium binding*"

→ DODQ = calcium binding site



Function of motifs specific to a superfamily

- Swiss-Prot = Database of protein annotations

Positions of important fragments (structure, sequence, function)

Pairing between specific words and Swiss-Prot annotation

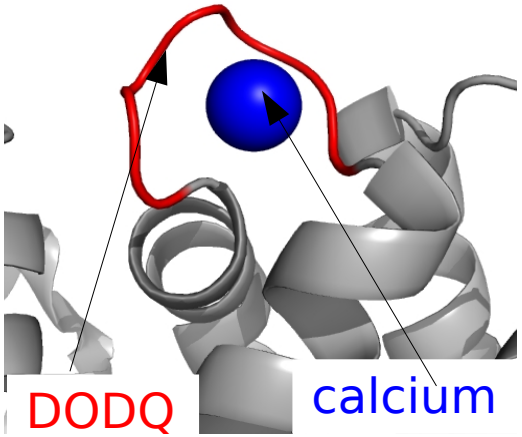
DODQ:
23 fragments

1acc	178 184
1alv_A	151 157
1alv_A	181 187
1aui_B	36 42
1aui_B	68 74
1aui_B	105 111
1aui_B	146 152
1bx4_A	250 256
1dab_A	558 564

14 fragments (=61%) annotated by "*calcium binding*"

→ DODQ = calcium binding site

candidats for calcium binding site

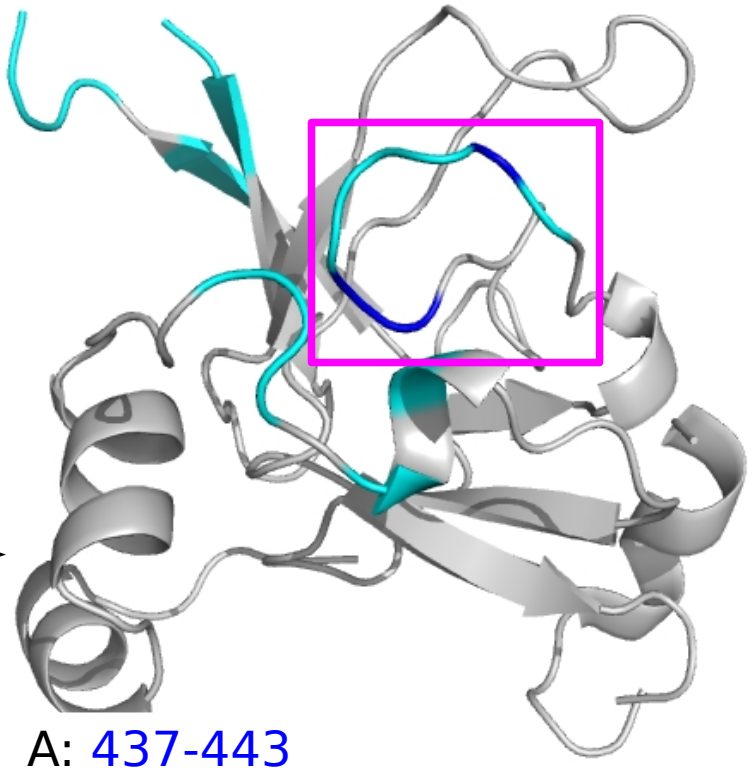


Validation of suggested functional sites

- SitePredict (Bordner et al., 2008) : Prediction of suggested calcium binding sites
  - Based on the conservation of structural properties (RMSD, ASA...)

DODQ

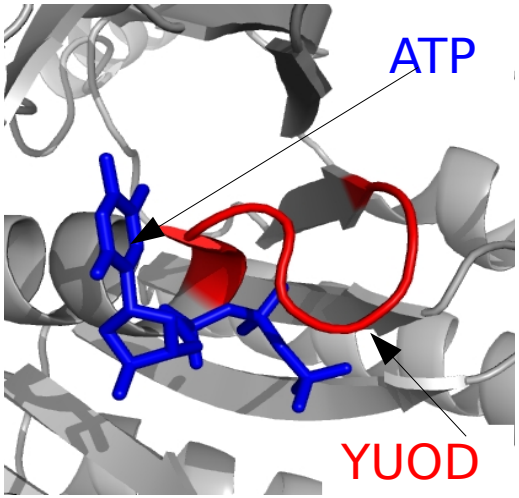
1a12_A	376 382
1alv_A	151 157
1alv_A	181 187
1aui_A	36 42
1aui_B	68 74
1aui_B	105 111
1aui_B	146 152
1qus_A	238 244
1l1d_A	437-443

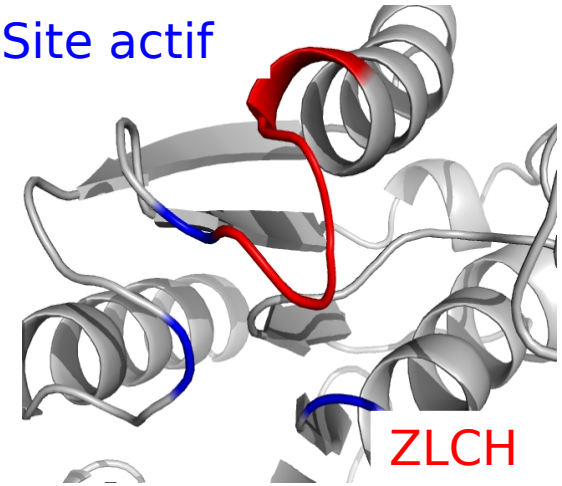


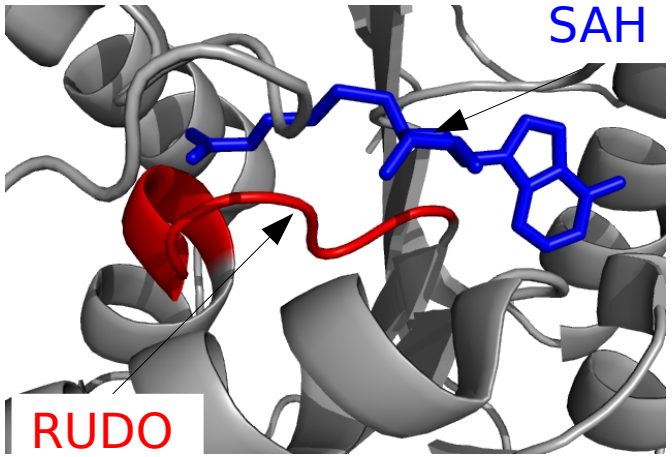
1l1d_A: 437-443

Specific words correspond to functional sites

- Structural words over-represented in a superfamily = structural motifs involved in
  - Binding to to small ligands: ATP/GTP, NAD(P), calcium, SAH/SAM
  - Actif sites
  - Repeat: repetition of 20-30 amino acids







Applications

- HMM-SA + over-representation in SCOP => Extraction of functional structural motifs
  - Without knowledge of the locate of functional sites

- Suggest new occurrences of some functional sites

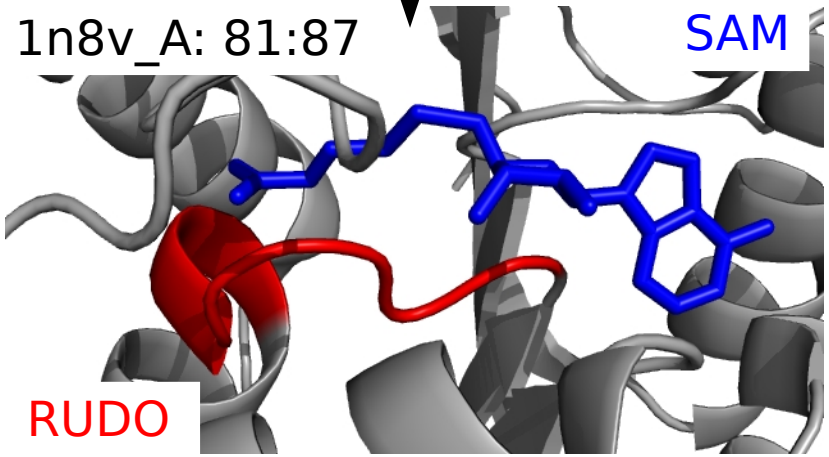
Suggestion of new functional sites

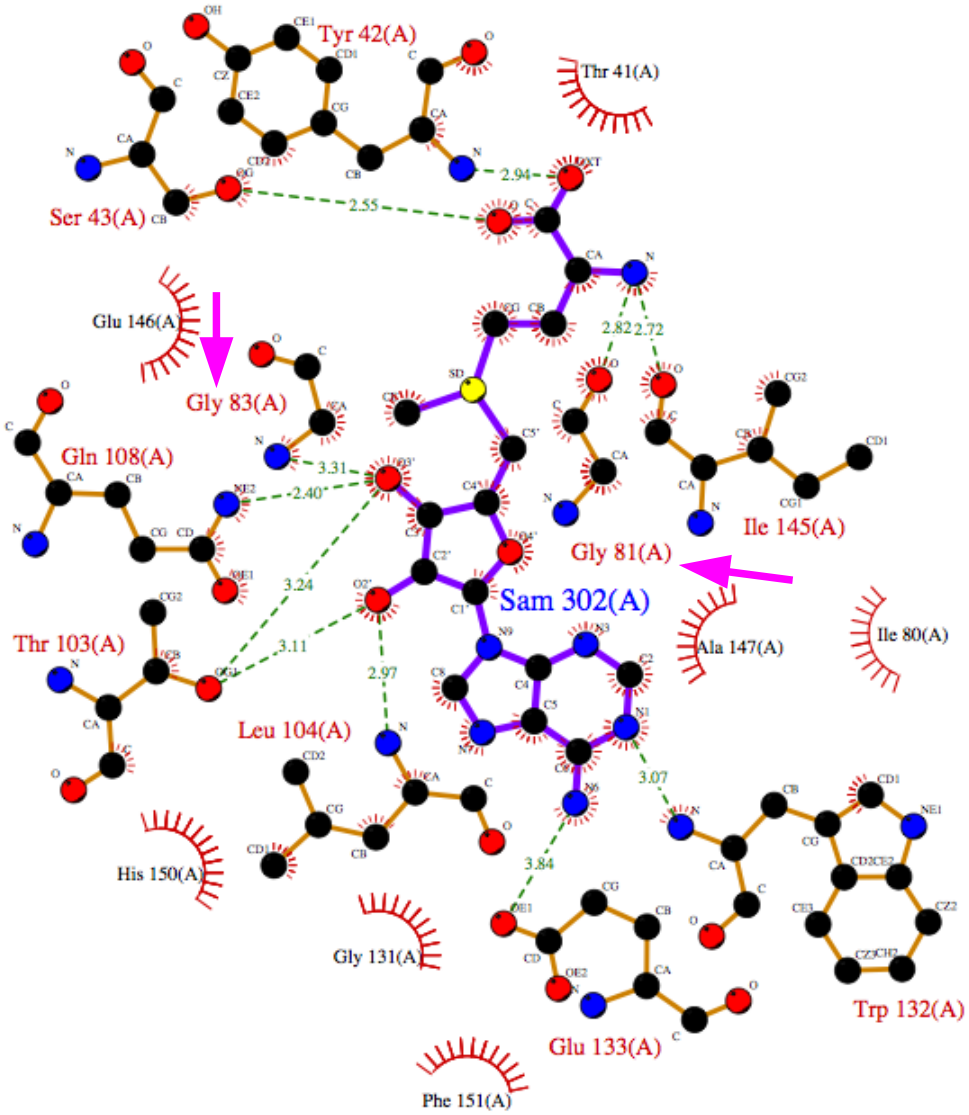
> 2fk8_A.pdbA 268 -152.890196
...BBZSNLHBVWBBCQKYBDSGQYNLG
FFDZSKHAAAAAAAAAAVWBBEDRNYU
SLNMMNUG**RUDO**WVWAAAVZCDSKN
MNMXPQPBAAVWAAVWVWVQPQPQK
TXMNPQLHBZQGSKXPIMTTTFOEBB...

Unannotated by Swiss-Prot

Suggestion of new functional sites

> 2fk8_A.pdbA 268 -152.890196
...BBZSNLHBVWBBCQKYBDSGQYNLG
FFDZSKHAAAAAAAAAAVWBBEDRNYU
SLNMMNUG**RUDO**WVWAAAVZCDSKN
MNMXPQPBAAVWAAVWVWVQPQPQK
TXMNPQLHBZQGSKXPIMTTTFOEBB...





Applications

- HMM-SA + over-representation in SCOP => Extraction of functional structural motifs
  - Without knowledge of the locate of functional sites

- Suggest new occurrence of some functional sites

- Annotation of uncharacterized proteins: known structure

Annotation of uncharacterized proteins

PSI Structural Genomic Knowledgebase

Code pdb: 1vhy_B

Encodage by HMM-SA

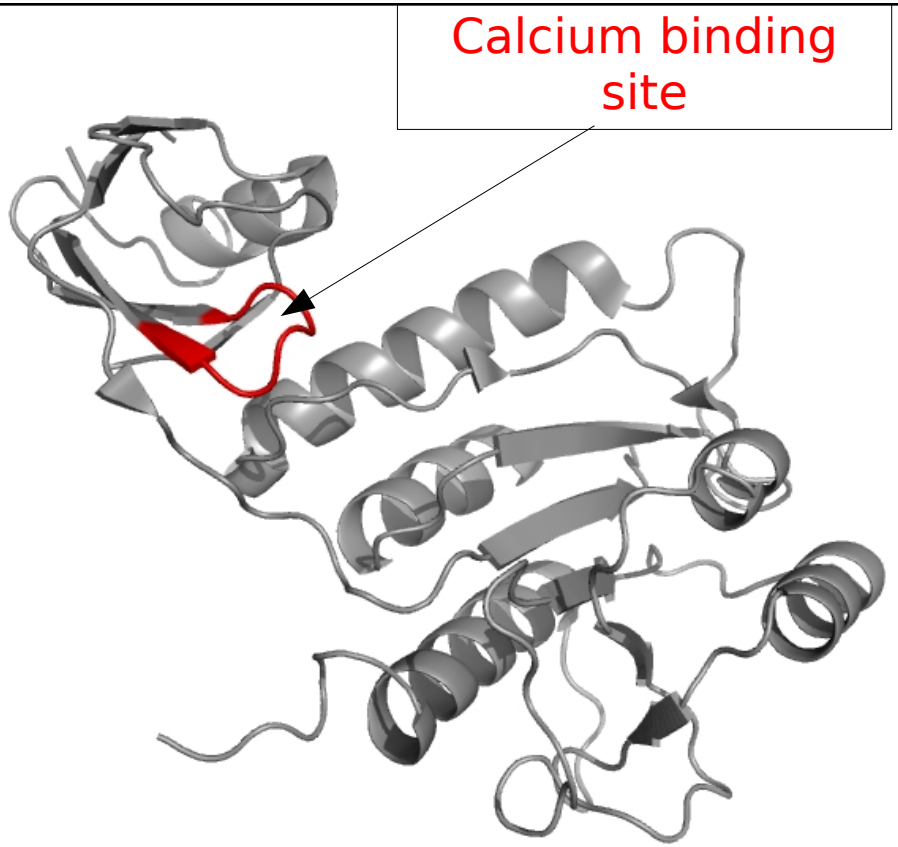
...LNMNPIKGOCSGITLLKHV
WVWVZCCCZDRNYUSKTM
NTGDODQMMMXML..

Functional word search

...LNMNPIKGOCSGITLLKHV
WVWVZCCCZDRNYUSKTM
NTG**DODQ**MMMXML..

SitePredict

No calcium binding sites



Applications

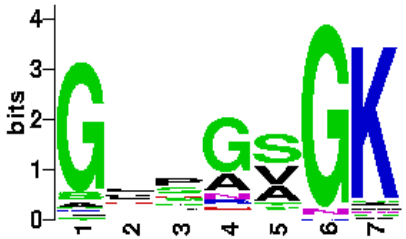
- HMM-SA + over-representation in SCOP => Extraction of functional structural motifs
  - ✔ Without knowledge of the locate of functional sites

- Suggest new occurrence of some functional sites

- Annotation of uncharacterized proteins: known structure

- Prediction of function of protein with unknown structure
  - ✔ Functional words: sequence specificity: spécificité de séquence
    → method of prediction of these functional motifs based on their am acid sequence
    Poster n°20

YUOD



Thank you for your attention!!!!

Thanks to

Pr. Anne-Claude Camproux, MT*i* UMR INSERM S-973, Université Paris Diderot - Paris7

Dr. Reynès Christelle, MT*i* UMR INSERM S-973, Université Paris Diderot - Paris7

Dr. Juliette Martin, IBCP URM CNRS 5086 , Université de Lyon

Dr. Grégory Nuel, MAP5 - UMR CNRS 8145, Université Paris Descartes

Posters of works developed in MT*i*,

- P17: SPatt
- P20: Prediction of functional motifs from amino acid sequence
- P79: MobyleNet
- P89: 3D printer service
- P98: Anatomy of druggable pockets