BINF 731

Protein Structure Analysis

Iosif Vaisman

2012

Sequence patterns

KKFAQSTNLKSHILT

KQFSHSAQLRAHIST
GKFSDSNQLKSHMLV
KDISSSESLRTHMFK
KRFSHSGSYSSHISS
KRFSHSGSFSSHMTS
KTLSDRLEYQQHMLK

Regular Expressions

Operation	**Regular Expression**	**Example**
Concatenation	DLIV	DLIV
Alternation	D[LIV]K	DLK
Replication	DL(2,5)K	DLLK

Regular Expressions

Patterns described in a standard way are known as *regular expressions*

x	ANY		
[ ]	OR	[ILV]	I or L or V
{ }	NOT	{DE}	not D or E
( )	repetitions	x(2,3)	x-x or x-x-x
-	separator		
<	N-terminal		
>	C-terminal		
.	END		

Regular Expressions

[AC]-x-V-x(4)-{ED}.

[Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}

...LKHV**A**Y**V**FQALIYWIK...
...AVEM**A**G**V**KYLQVQHGS...
...LYTG**A**I**V**TNNDGPYMA...
...KEYK**C**K**V**EKELTDICN...

PROSITE Database

Current version contains 1079 documentation entries that describe 1459 different patterns, rules and profiles/matrices

[ST]-x(2)-[DE]
Casein kinase II phosphorylation site

[AG]-x(4)-G-K-[ST]
ATP/GTP-binding site motif A (P-loop)

Y-x-[NQH]-K-[DE]-[IVA]-F-[LM]-R-[ED]
Heat shock hsp90 proteins family signature

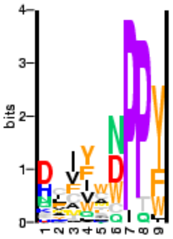
http://www.expasy.ch/prosite

Blocks Database

Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins

N-6 Adenine-specific DNA methylases proteins
width=9 seqs=78

```
DMA_VIBCH|Q08318  (85)  SCTQWWPPF 77
HEMK_MYCLE|P45832 (181) DLFVAQPTL 100
MT57_ECOLI|P25240 (111) DGALGNPPF 13
MTC1_CHVN1|Q01511 (172) NFVFLDPPY 8
MTC1_COREQ|P42828 (71)  QLSFSCPPF 49
MTH2_HAEHA|P00473 (32)  KIAFFDPQY 52
MTH3_HAEIN|P43871 (23)  HAIISDIPY 73
MTM1_MICAM|P50190 (306) AAVLTNPPF 14
MTM2_MORBO|P23192 (25)  QLAVIDPPY 10
MTMU_MYCSP|P43641 (37)  QVIYADPPW 13
MTR1_RHOSH|P14751 (60)  QLIICDPPY 8
...................................
```



http://www.blocks.fhcrc.org/

Pfam Database

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains

Zinc finger, C2H2 type

```
TYY1_HUMAN/383-407  YVCPF.DGCN...KKFAQSTNLKSHILT...H
ZG52_XENLA/61-83    YTCT...QCN...KQFSHSAQLRAHIST...H
KRUP_DROME/306-328  YTCE...ICD...GKFSDSNQLKSHMLV...H
YKQ8_CAEEL/78-102   YKCT...VCR...KDISSSESLRTHMFKQ.HH
DEFI_CHICK/268-292  YECP...NCK...KRFSHSGSYSSHISSK.KC
ZFH1_DROME/389-413  FGCD...NCG...KRFSHSGSFSSHMTSK.KC
YL57_CAEEL/42-65    YLCY...YCG...KTLSDRLEYQQHMLK..VH
ZFA_MOUSE/542-564   FKCD...ICL...LTFSDTKEVQQHALV...H
BASO_HUMAN/719-742  FQCD...ICK...KTFKNACSVKIHHKN..MH
HUNB_DROME/297-319  FQCD...KCS...YTCVNKSMLNSHRKS...H
SFP1_YEAST/598-623  FKCPV.IGCE...KTYKNQNGLKYHRLH..GH
ZG29_XENLA/62-84    FVCT...VCG...KTYKYKHGLNTHLHS...H
```

http://pfam.wustl.edu/

Other Motif Databases

PRINTS : a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family
http://bioinf.man.ac.uk/dbbrowser/PRINTS/

DOMO : a protein domain database
http://www.infobiogen.fr/~gracy/domo/home.htm

ProDom : a protein domain database
http://protein.toulouse.inra.fr/prodom.html

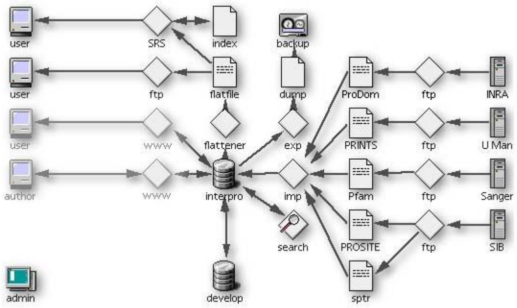
InterPro Database

InterPro : integrated resource for the commonly used signature databases - Pfam, PRINTS, PROSITE, ProDom and SWISS-PROT + TrEMBL.

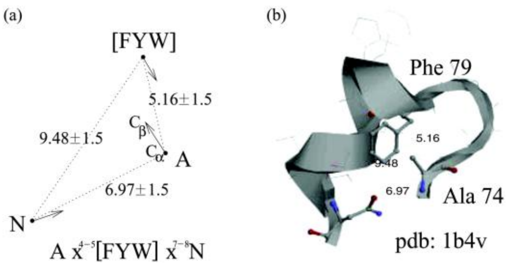
Current release of InterPro (3.2) contains 3939 entries, representing 1009 domains, 2850 families, 65 repeats and 15 post-translational modification sites.

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

InterPro Database



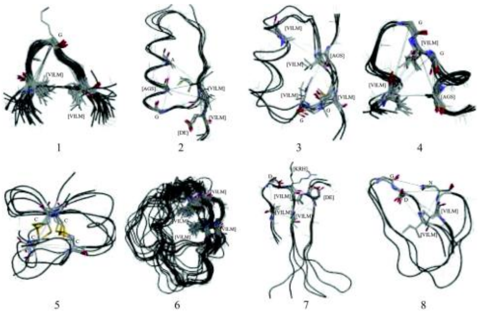
Sequence-structure patterns



A x^{4-5}[FYW] x^{7-8}N

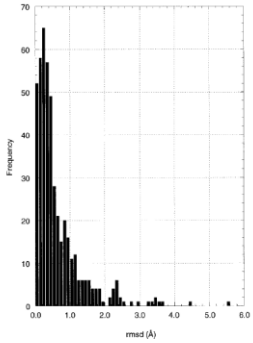
Bradley et al., PNAS, 2002

Sequence-structure patterns



Bradley et al., PNAS, 2002

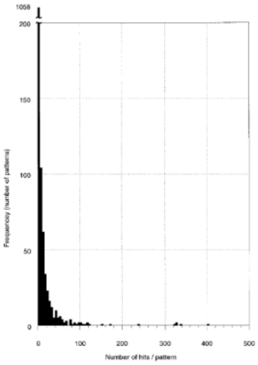
Structural motifs of PROSITE patterns



Distribution of rmsd values for the true hits. The rmsd was calculated from all true hits eliminating false and unidentifed hits for each of the 466 patterns having more than one true hit.

From Kasuya and Thornton (1999)

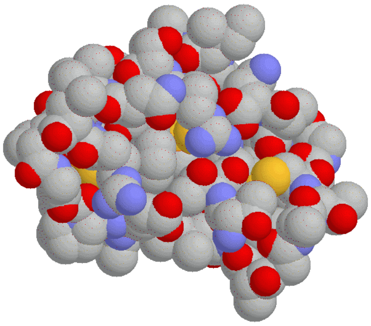
Structural motifs of PROSITE patterns



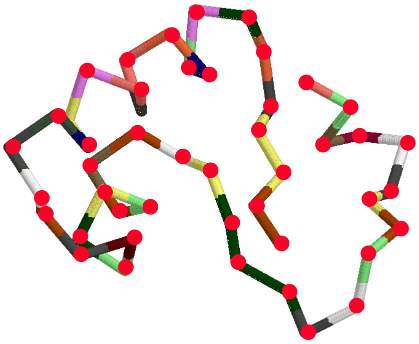
Distribution of number of hits per pattern. Each column represents frequency of patterns having sequence matches in the 3D-sequence library. The bin width is 10. There were 1058 patterns having numbers of hits smaller than 10, including 712 patterns with no hits.

From Kasuya and Thornton (1999)

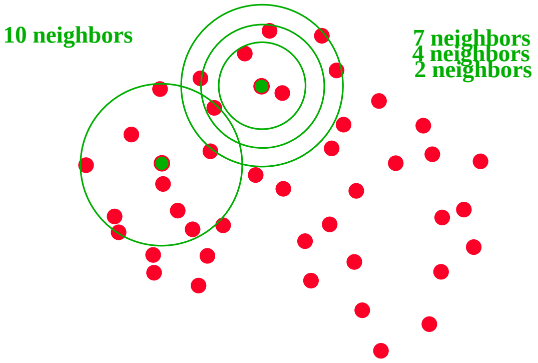
Protein representation (Crambin)



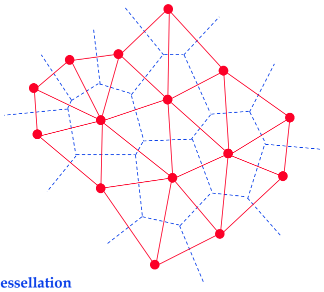
Protein representation (Crambin)



Neighbor identification in proteins



Neighbor identification in proteins: Voronoi/Delaunay Tessellation in 2D



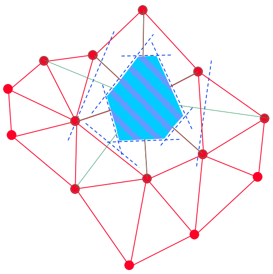
Delaunay simplex is defined by points, whose Voronoi polyhedra have common vertex

Delaunay simplex is always a triangle in a 2D space and a tetrahedron in a 3D space

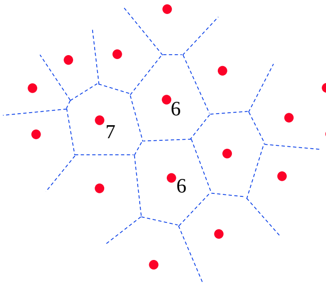
Voronoi Tessellation

Delaunay Tessellation

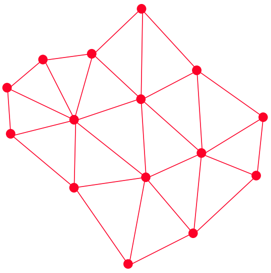
Neighbor identification in proteins: Voronoi/Delaunay Tessellation in 2D



Neighbor identification in proteins: Voronoi/Delaunay Tessellation in 2D

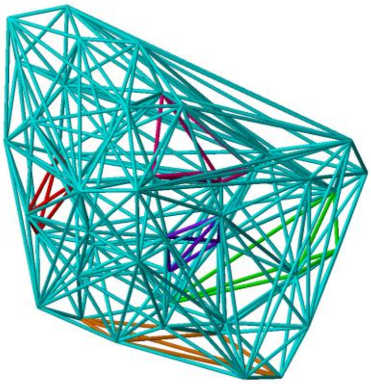


Voronoi Tessellation

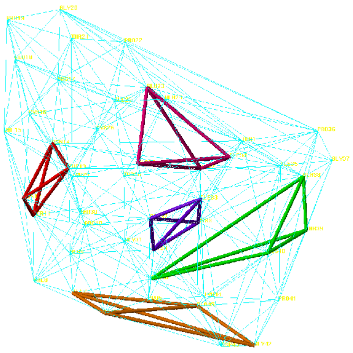


Delaunay Tessellation

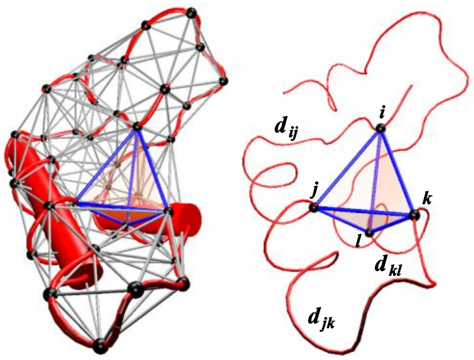
Delaunay tessellation of Crambin



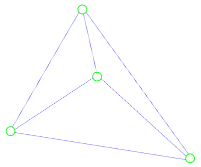
Delaunay tessellation of Crambin



Dealunay simplices classification

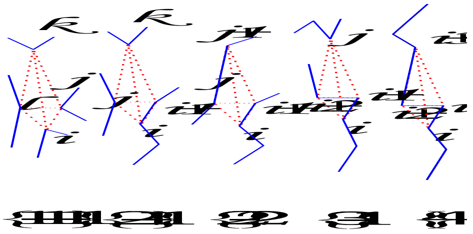


Three views at one object:

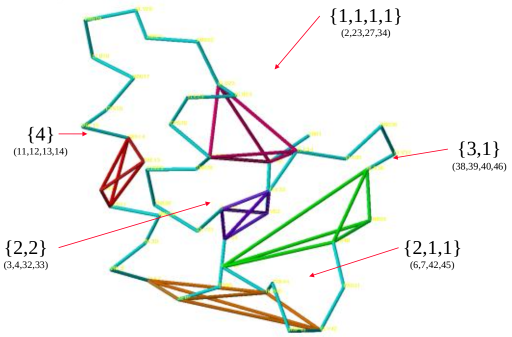


Topological:	**simplex Delaunay**
Geometrical:	**tetrahedron**
Compositional:	**quadruplet of points**

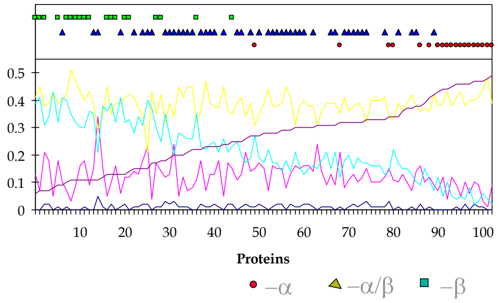
Classification of Delaunay simplices by sequential proximity



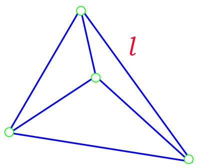
Types of Delaunay simplices in Crambin



Correlations between protein structure family assignment and relative content of classes of Delaunay simplices

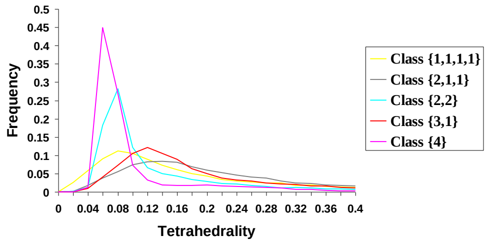


Tetrahedrality of Delaunay simplices



$$T = \sum_{i>j}(l_i - l_j)^2 \,/\, 15\bar{l}^{\,2}$$

Tetrahedrality distribution of Delaunay simplices


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Frequency
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4
Tetrahedrality
Class {1,1,1,1}
Class {2,1,1}
Class {2,2}
Class {3,1}
Class {4}