Patterns in protein sequences and structures

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}

```
...LKHVAYVFQALIYWIK...
...AVEMAGVKYLQVQHGS...
...LYTGAIVTNNDGPYMA...
...KEYKCKVEKELTDICN...
```

PROSITE Database

Current version contains 1038 documentation entries that describe 1379 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

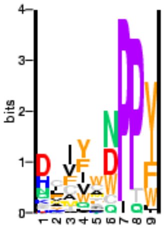
http://www.expasy.ch/prosite

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

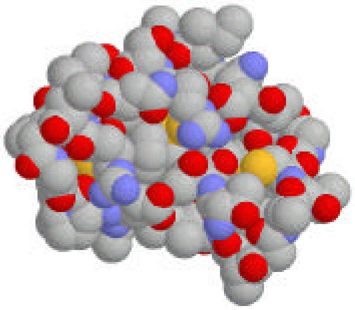
ProDom : a protein domain database
http://protein.toulouse.inra.fr/prodom.html

Pattern Discovery

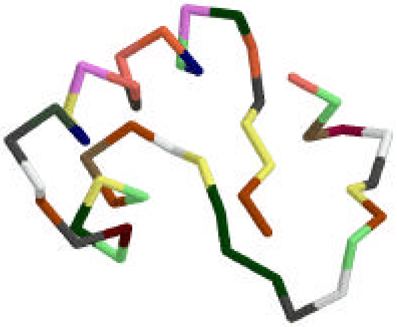
MEME (Multiple EM for Motif Elicitation):
discovery of motifs (highly conserved regions)
in groups of related DNA or protein sequences
http://bioweb.pasteur.fr/seqanal/motif/meme/

PrattWWW : Motif Discovery
http://bioweb.pasteur.fr/seqanal/Pratt/pratt-form.html

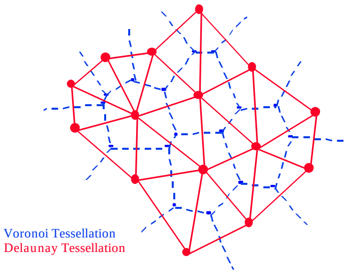
Structural Patterns



Crambin backbone



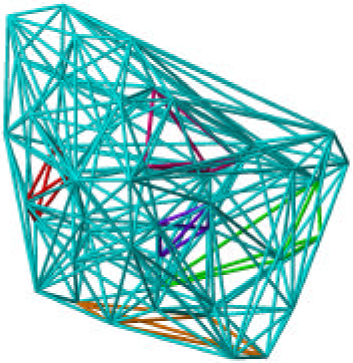
Voronoi/Delaunay Tessellation in 2D



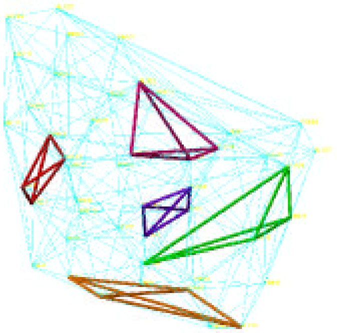
Delaunay simplex is defined by points, whose Voronoi polyhedra have common vertex

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

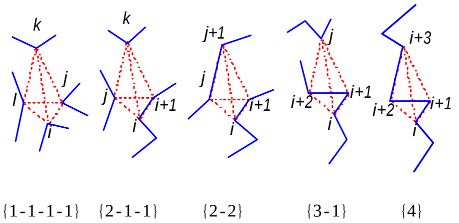
Delaunay tesselation of Crambin (1crn)



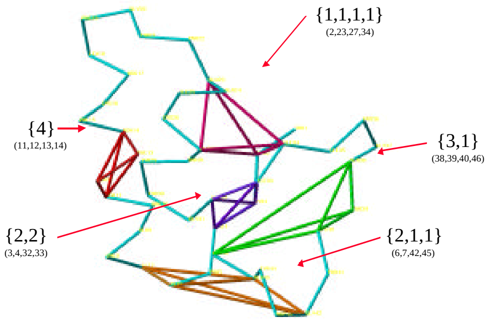
Tessellated Crambin



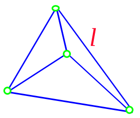
Classification of Delaunay simplices by sequential proximity



Types of Delaunay simplices in Crambin

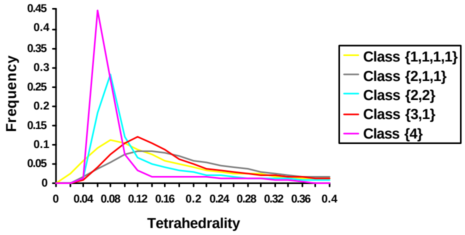


Tetrahedrality of Delaunay simplices

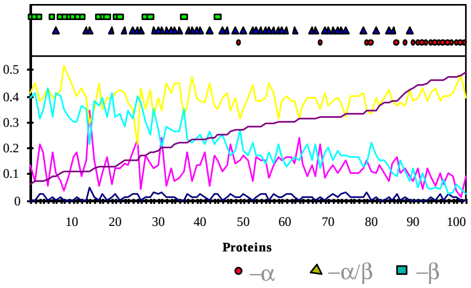


$$T = \sum_{i>j} (l_i - l_j)^2 / 15\bar{l}^2$$

Tetrahedrality distribution of Delaunay simplices



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



Compositional Propensities of Delaunay Simplices

$$q_{ijkl} = \log \frac{f_{ijkl}}{p_{ijkl}}$$

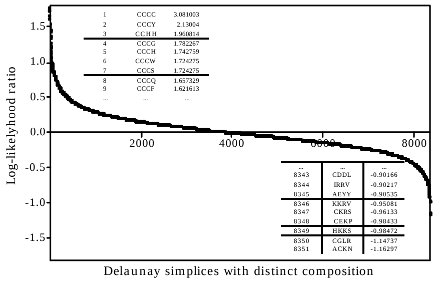
f- *observed quadruplet frequency*

$$p_{ijkl} = C a_i a_j a_k a_l$$

a- *individual AA frequency*

$$C = \frac{4!}{\prod_i^n (t_i!)}$$

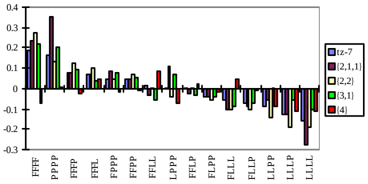
Log-likelihood of amino acid quadruplets with different compositions



1	CCCC	3.081003
2	CCCY	2.13004
3	CCHH	1.960814
4	CCCG	1.782267
5	CCCH	1.742759
6	CCCW	1.724275
7	CCCS	1.724275
8	CCCQ	1.657329
9	CCCF	1.621613
...	...	...

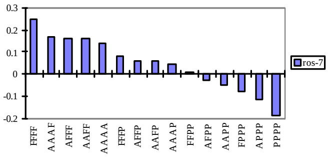
...	...	...
8343	CDDL	-0.90166
8344	IRRV	-0.90217
8345	AEYY	-0.90535
8346	KKRV	-0.95081
8347	CKRS	-0.96133
8348	CEKP	-0.98433
8349	HKKS	-0.98472
8350	CGLR	-1.14737
8351	ACKN	-1.16297

Log-likelihood of amino acid quadruplets (reduced alphabet)



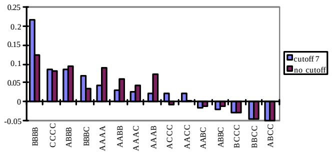
(F) Ala, Val, Phe, Ile, Leu, Pro, Met
(L) Asp, Glu, Lys, Arg
(P) Ser, Thr, Tyr, Cys, Asn, Gln, His, Trp

Log-likelihood of amino acid quadruplets (reduced alphabet)



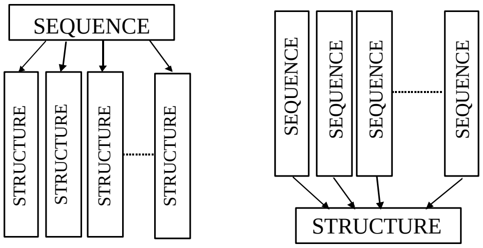
(F) Cys, Phe, Ile, Leu, Met, Val, Trp
(A) Ala, His, Thr, Tyr
(P) Asp, Glu, Gly, Lys, Asn, Pro, Gln, Arg, Ser

Log-likelihood of amino acid quadruplets (reduced alphabet)



(A) Met, His, Val, Tyr, Asn, Asp, Ile
(B) Gln, Leu, Glu, Lys, Phe
(C) Trp, Pro, Arg, Gly, Ser, Ala, Thr, Cys

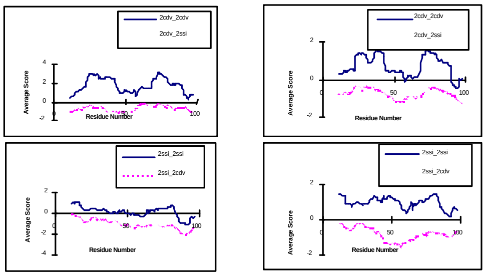
Threading



Sequence-structure compatibility

Structure-sequence compatibility

Delaunay profiles for native and misfolded structures



20L

6L5T