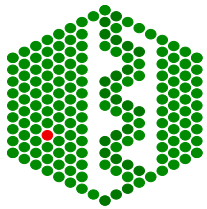
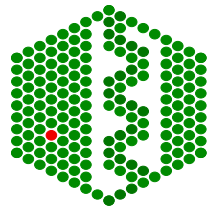
PROTEIN SEQUENCE ANALYSIS



Need good protein sequence analysis tools because:

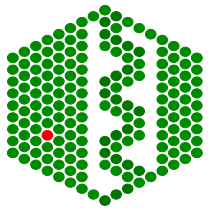
- As number of sequences increases, so gap between seq data and experimental data increases
- But increase number of sequences - increase sequence DB and therefore increased chance of finding similar sequence
- Computer analysis can narrow down number of functional experiments required



UNKNOWN PROTEIN SEQUENCE

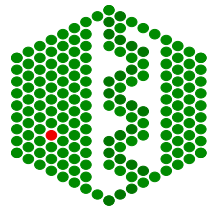
LOOK FOR:

- Similar sequences in databases ((PSI) BLAST)
- Distinctive patterns/domains associated with function
- Functionally important residues
- Secondary and tertiary structure
- Physical properties (hydrophobicity, IEP etc)

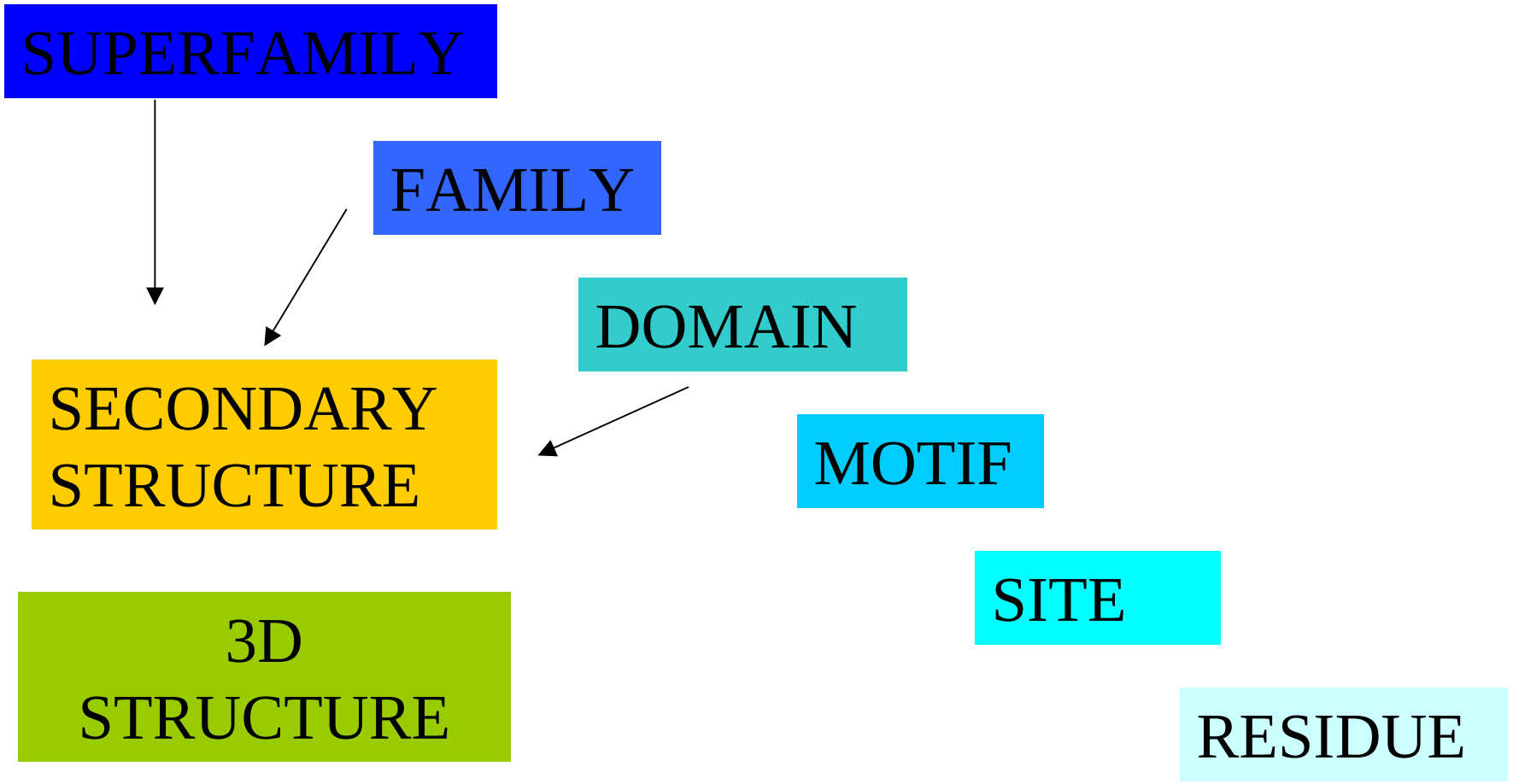


BASIC INFORMATION COMES FROM SEQUENCE

- One sequence- can get some information eg amino acid properties
- More than one sequence- get more info on conserved residues, fold and function
- Multiple alignments of related sequences- can build up consensus sequences of known families, domains, motifs or sites.
- Sequence alignments can give information on loops, families and function from conserved regions

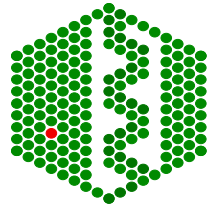


LEVEL OF FUNCTION INFORMATION IN PROTEIN SEQUENCES



AMINO ACID PROPERTIES

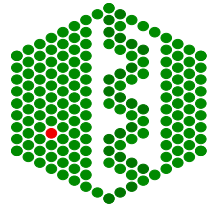
• Small	Ala, Gly
• Small hydroxyl	Ser, Thr
• Basic	His, Lys, Arg
• Aromatic	Phe, Tyr, Trp
• Small hydrophobic	Val, Leu, Ile
• Medium hydrophobic	Val, Leu, Ile, Met
• Acidic/amide	Asp, Glu, Asn, Gln
• Small/polar	Ala, Gly, Ser, Thr, Pro



Protein functions from specific residues

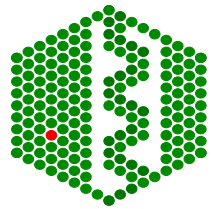
- C disulphide-rich, metallo-thionein, zinc fingers
- DE acidic proteins (unknown)
- G collagens
- H histidine-rich glycoprotein
- KR nuclear proteins, nuclear localisation
- P collagen, filaments
- SR RNA binding motifs
- ST mucins

- Polar (C,D,E,H,K,N,Q,R,S,T) - active sites
- Aromatic (F,H,W,Y) - protein ligand-binding sites
- Zn+-coord (C,D,E,H,N,Q) - active site, zinc finger
- Ca2+-coord (D,E,N,Q) - ligand-binding site
- Mg/Mn-coord (D,E,N,S,R,T) - Mg2+ or Mn2+ catalysis, ligand binding
- Ph-bind (H,K,R,S,T) - phosphate and sulphate binding



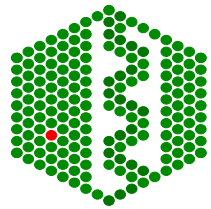
Protein functions from regions

- Active sites- short, highly conserved regions
- Loops- charged residues and variable sequence
- Interior of protein- conservation of charged amino acids



Additional analysis of protein sequences

- transmembrane regions
- signal sequences
- localisation signals
- targeting sequences
- GPI anchors
- glycosylation sites
- hydrophobicity
- amino acid composition
- molecular weight
- solvent accessibility
- antigenicity

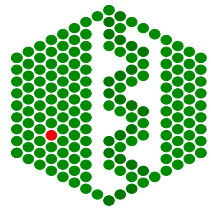


FINDING CONSERVED PATTERNS IN PROTEIN SEQUENCES

- Pattern - short, simplest, but limited
- Motif - conserved element of a sequence alignment, usually predictive of structural or functional region

To get more information across whole alignment:

- Matrix
- Profile
- HMM



PATTERNS

- Small, highly conserved regions
- Shown as regular expressions

  Example:

  [AG]-x-V-x(2)-x-{YW}

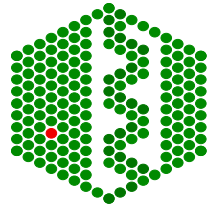
  – [] shows either amino acid

  – X is any amino acid

  – X(2) any amino acid in the next 2 positions

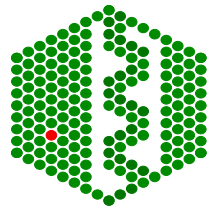
  – {} shows any amino acid except these

  BUT- limited to near exact match in small region



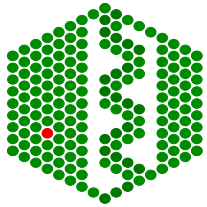
MATRIX

- 210 possible aa pairs (190 different aa, 20 identical aa)
- Start with sequence alignment and build up a table of probabilites of finding each aa in each position of the sequence
- Can be scored in several different ways



Matrix scores can be based on:

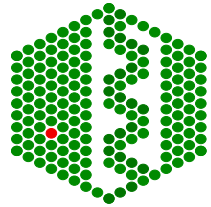
- **Genetic code -**base changes required to convert codons for 2 amino acids
- **Chemical similarity -**polarity, size, shape, charge
- **Observed substitutions -**based on analysing frequencies seen in alignments- inter-reliable
- **Dayhoff mutation data matrix -** likelihood of mutation from one aa to another, but different positions are not equally mutatable, and only useful for close function because sequence alignments are very related proteins



Matrix scoring continued

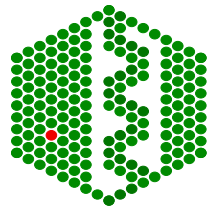
- **BLOSUM -**matrix from ungapped alignments of distantly related sequences -cluster sequences similar at a threshold value of % identity -substitution frequencies for all pairs of aa calculated -used to calculate a log odds BLOSUM (blocks substitution matrix). Can vary threshold values
- **3D structure matrix -**derived from tertiary structure alignment, good, but only used if structure is known

  Best matrices are derived from observed substitution data, it is important to use select scoring appropriate for evolutionary distance interested in.



PROFILES

- Table or matrix containing comparison information for aligned sequences
- Used to find sequences similar to alignment rather than one sequence
- Contains same number of rows as positions in sequences
- Row contains score for alignment of position with each residue

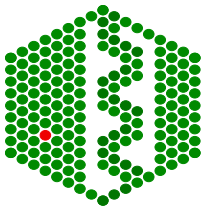


Example of a Profile

F	K	L	L	S	H	C	L	L	V
F	K	A	F	G	Q	T	M	F	Q
Y	P	I	V	G	Q	E	L	L	G
F	P	V	V	K	E	A	I	L	K
F	K	V	L	A	A	V	I	A	D
L	E	F	I	S	E	C	I	I	Q
F	K	L	L	G	N	V	L	V	C

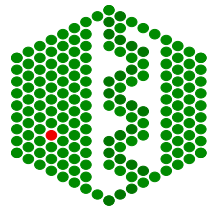
A	-18	-10	-1	-8	8	-3	3	-10	-2	-8
C	-22	-33	-18	-18	-22	-26	22	-24	-19	-7
D	-35	0	-32	-33	-7	6	-17	-34	-31	0
E	-27	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	14	-26	-29	-15	4	12	-29
G	-30	-20	-28	-32	28	-14	-23	-33	-27	-5
H	-13	-12	-25	-25	-16	14	-22	-22	-23	-10
I	3	-27	21	25	-29	-23	-8	33	19	-23
K	-26	25	-25	-27	-6	4	-15	-27	-26	0
L	14	-28	19	27	-27	-20	-9	33	26	-21
M	3	-15	10	14	-17	-10	-9	25	12	-11
N	-22	-6	-24	-27	1	8	-15	-24	-24	-4
P	-30	24	-26	-28	-14	-10	-22	-24	-26	-18
Q	-32	5	-25	-26	-9	24	-16	-17	-23	7
R	-18	9	-22	-22	-10	0	-18	-23	-22	-4
S	-22	-8	-16	-21	11	2	-1	-24	-19	-4
T	-10	-10	-6	-7	-5	-8	2	-10	-7	-11
V	0	-25	22	25	-19	-26	6	19	16	-16
W	9	-25	-18	-19	-25	-27	-34	-20	-17	-28
Y	34	-18	-1	1	-23	-12	-19	0	0	-18

Match values are higher for conserved residues



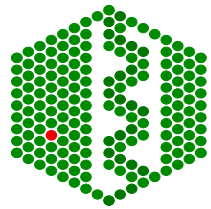
Building a Profile

- To get good profile need good, hand-curated alignment
- Use alignment to build up position-specific scoring matrix
- Use matrix (profile) to do PSI-BLAST with several iterations



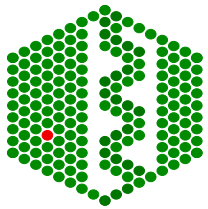
SCORES

- E-value is chance of a random sequence sequence hitting. E-value 1.0 not significant, 0.1 possibly significant,< 0.01 most likely to be significant. All depends on database size



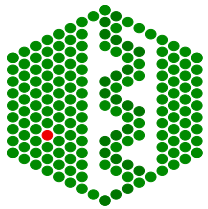
HIDDEN MARKOV MODELS (HMM)

- An HMM is a large-scale profile with gaps, insertions and deletions allowed in the alignments, and built around probabilities
- Package used HMMER (http://hmmer.wusd.edu/)
- Start with one sequence or alignment -HMMbuild, then calibrate with HMMcalibrate, search database with HMM
- E-value- number of false matches expected with a certain score
- Assume extreme value distribution for noise, calibrate by searching random seq with HMM build up curve of noise (EVD)



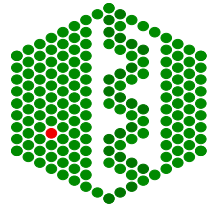
REPEATS

- Structural and evolutionary entities found in 2 or more copies
- Often assemble into elongated “rods”, “superhelices” or “barrel” structures
- Specialised cases when building profiles



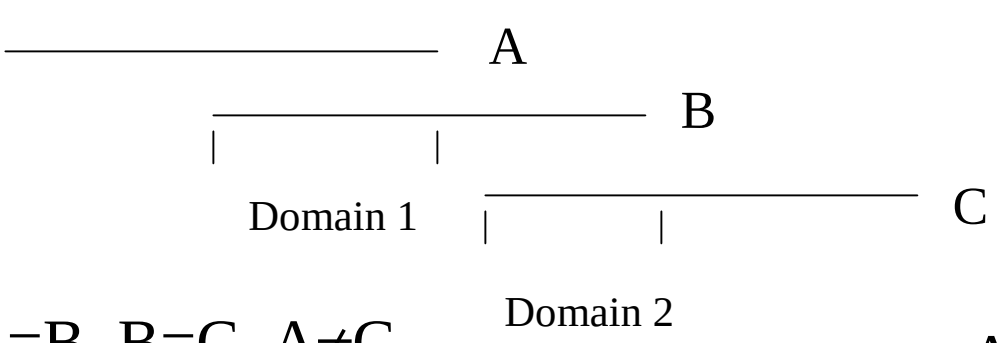
PITFALLS OF METHODS

- **BLAST -** only pick up homologues, not distant, divergent family members
- **PSI-BLAST -** fine for superfamilies, not very good for small very conserved motifs
- **Patterns -** small, localised and need to be highly conserved regions
- **HMMER -** slow process for searching database
- **Profiles -** if false positive picked up, pulls in its companions, in large families members can be missed
- **Alignment methods -** automatic, less biological significance

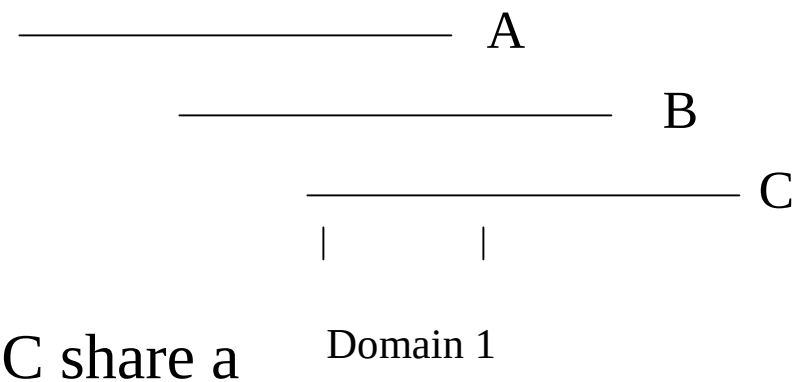


Big problem in protein sequence analysis- multidomain proteins:

- Most conserved domain will score highest in sequence similarity searches, may overlook lower scoring domains
- Iterative searching of multi-domain proteins could pick up unrelated proteins

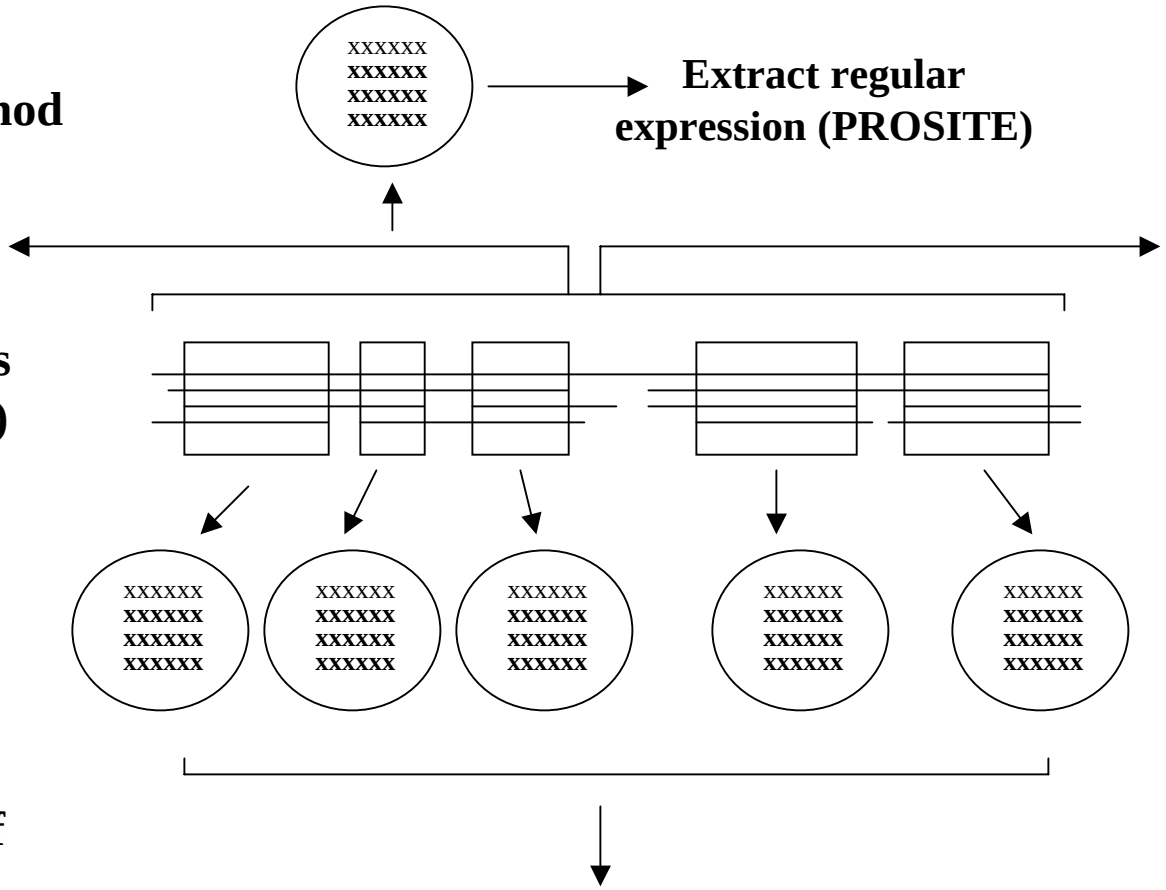


A=B, B=C, A≠C



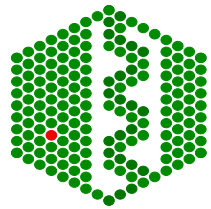
A,B & C share a common domain

SUMMARY OF PATTERN METHODS



COMMON PROTEIN PATTERN DATABASES

- Prosite patterns
- Prosite profiles
- Pfam
- SMART
- Prints
- ProDom
- DOMO
- BLOCKS



SOFTWARE FOR PROTEIN SEQUENCE ANALYSIS

- GCG (http://www.gcg.com/)
- EMBOSS (ftp:ftp.sanger.ac.uk/pub/EMBOSS)
- PIX- HGMP (http://www.hgmp.mrc.ac.uk)
- ExPASy Proteomics tools (http://www.expasy.org/tools)
- PredictProtein (http://www.embl-heidelberg.de/predictprotein/)