Protein Function Prediction

Andrew Torda, Wintersemester 2009 / 2010, Angewandte …

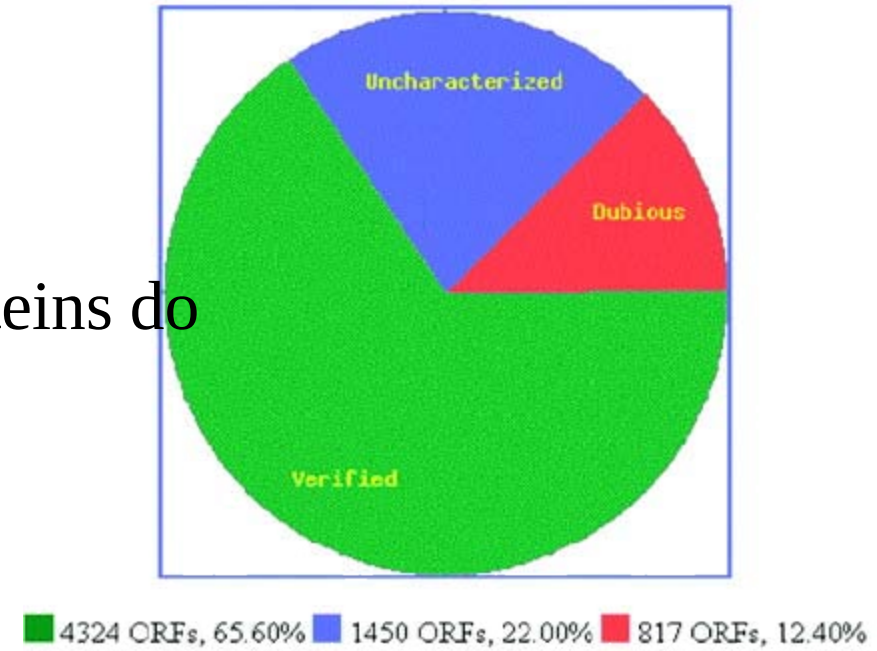
- Protein function - field of biochemists
  - can it be predicted / guessed from
    - structure ?
    - sequence ?

- Is this an issue ?
  - 5 to 10 years ago
    - a protein was of interest, because one knew its function
      - then found its sequence + structure
  - now, lots of proteins unknown

Example yeast genome

- yeast 6.6×10^3 proteins / ORFs
- ≈ decade after sequencing

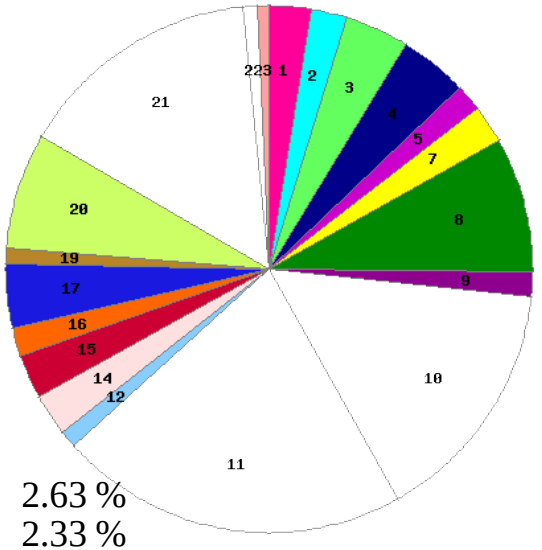
- not really known what many proteins do



- protein function may not be easy
  - extreme case - prions
    - structure lots of effort (X-ray, NMR)
    - function - expression, knockouts
    - function still not really clear

Hirschman, J. E. et al. Nucl. Acids Res. 2006 34, D442-D445

Example e. coli



From cmr.tigr.org

- very well studied, common bacterium
- 4289 genes

?

	Gene Role Category	
1	Amino acid biosynthesis	2.63 %
2	Biosynthesis of cofactors ....carriers	2.33 %
3	Cell envelope	3.98 %
4	Cellular processes	4.38 %
5	Central intermediary metabolism	1.70 %
6	Disrupted reading frame	0 %
7	DNA metabolism	2.40 %
8	Energy metabolism	8.55 %
9	Fatty acid and phospholipid metabolism	1.56 %
10	Hypothetical proteins	15.7 %
11	Hypothetical proteins - Conserved	22.1 %
12	Mobile and extrachromosomal element functions	1.07 %
13	Pathogen responses	0 %
14	Protein fate	2.70 %
15	Protein synthesis	2.79 %
16	Purines, pyrimidines, nucleosides, and nucleotides	1.79 %
17	Regulatory functions	4.08 %
18	Signal transduction	0 %
19	Transcription	0.95 %
20	Transport and binding proteins	7.34 %
21	Unclassified	15.5 %
22	Unknown function	0.88 %
23	Viral functions	0.76 %

Plan

- How could one quantify function ?
- What might one use to predict it ?

  - sequence homology
  - structure homology
  - sequence patterns / motifs
  - structure patterns / motifs

Beliefs

- If two proteins have very similar sequence
  - structure is similar (easy to quantify / true)
  - function should be similar
- Two proteins have rather different sequences
  - structures sometimes similar (many examples)
  - function ? like to be similar
- Consequence
  - find a new protein, look for similarity
  - hope for similarity to well-characterised proteins

- other opinions and examples

Why I do not like function

- Can we quantify / define it ?

```
emb|CAA55527.1| zinc finger protein [Homo sapiens]                    723   0.0
ref|XP_001160877.1| PREDICTED: zinc finger protein 227 isoform 1...   723   0.0
ref|XP_001132303.1| PREDICTED: similar to zinc finger protein 43...   722   0.0
ref|XP_001166123.1| PREDICTED: zinc finger protein 607 isoform 4...   722   0.0
sp|Q8IYB9|ZN595_HUMAN Zinc finger protein 595 >gi|23271315|gb|AA...   722   0.0
ref|XP_523409.2| PREDICTED: hypothetical protein [Pan troglodytes]   722   0.0
ref|NP_082814.1| hypothetical protein LOC73430 [Mus musculus] >g...   722   0.0
dbj|BAA06541.1| KIAA0065 [Homo sapiens]                              722   0.0
[ . . . ]
ref|XP_574335.2| PREDICTED: similar to zinc finger protein 51 [R...   720   0.0
dbj|BAD92323.1| zinc finger protein 493 variant [Homo sapiens]        720   0.0
gb|AAI12347.1| ZNF493 protein [Homo sapiens]                         719   0.0
ref|NP_008886.1| zinc finger protein 33B [Homo sapiens] >gi|6677...   719   0.0
ref|XP_001114064.1| PREDICTED: similar to zinc finger protein 59...   719   0.0
ref|NP_116078.3| zinc finger protein 607 [Homo sapiens] >gi|4707...   719   0.0
dbj|BAD18693.1| unnamed protein product [Homo sapiens]               718   0.0
ref|XP_979055.1| PREDICTED: similar to reduced expression 2 [Mus...   718   0.0
sp|P18751|ZO71_XENLA Oocyte zinc finger protein XLCOF7.1             718   0.0
ref|XP_539908.2| PREDICTED: similar to replication initiator 1 i...   717   0.0
```

What is function ?

- glycogen phosphorylase in muscle acting on ….
  - very clear
- a protein in DNA replication which contains a phosphorylation site ?

- different methods attempt different tasks

- If we agree on a level, can it be done in a machine-friendly form?

- Oldest attempt for enzymes …

EC Numbers

- 1956 international commission on enzymes
- 1961 first report on names
- regular updates until today

- names - according to reaction catalysed
- hierarchical
  - Class 1. Oxidoreductases
  - Class 2. Transferases
  - Class 3. Hydrolases
  - Class 4. Lyases
  - Class 5. Isomerases
  - Class 6. Ligases
- some examples

EC Numbers

- Lyase example

  "Lyases are enzymes cleaving C-C, C-O, C-N, and other bonds by elimination, leaving double bonds or rings, or conversely adding groups to double bonds"
- subclasses
  - EC 4.1 Carbon-carbon lyases
    - EC 4.1.1 Carboxy-Lyases
      - next page
    - EC 4.1.2 Aldehyde-Lyases
    - EC 4.1.3 Oxo-Acid-Lyases
    - EC 4.1.99 Other Carbon-Carbon Lyases
  - EC 4.2 Carbon-oxygen lyases
  - EC 4.3 Carbon-nitrogen lyases
  - EC 4.4 Carbon-sulfur lyases
  - EC 4.5 Carbon-halide lyases
  - EC 4.6 Phosphorus-oxygen lyases
  - EC 4.99 Other lyases

EC Numbers

- EC 4.1.1.1 pyruvate decarboxylase
- EC 4.1.1.2 oxalate decarboxylase
- EC 4.1.1.3 oxaloacetate decarboxylase
- EC 4.1.1.4 acetoacetate decarboxylase
- EC 4.1.1.5 acetolactate decarboxylase
- EC 4.1.1.6 aconitate decarboxylase
- EC 4.1.1.7 benzoylformate decarboxylase
- EC 4.1.1.8 oxalyl-CoA decarboxylase
- […..]
- EC 4.1.1.84 D-dopachrome decarboxylase
- EC 4.1.1.85 3-dehydro-L-gulonate-6-phosphate decarboxylase
- EC 4.1.1.86 diaminobutyrate decarboxylase

- Problems
  - proteins may have more than one function
  - annotated function may not be the one *in vivo*
  - horror
    - two enzymes - unrelated, no homology, no connection
    - both appear to catalyse the same reaction
      - end in same EC class
- Benefits
  - more correct than incorrect
  - almost suitable for automation and machine recognition

Gene Ontology

- 3 characteristics
  1. biological process
  2. molecular function
  3. cellular component
- example 1uw0

Example 1uw0

molecular function

- DNA binding
- DNA ligase (ATP) activity
- ATP binding
- zinc ion binding

biological process

- DNA replication
- DNA repair
- DNA recombination

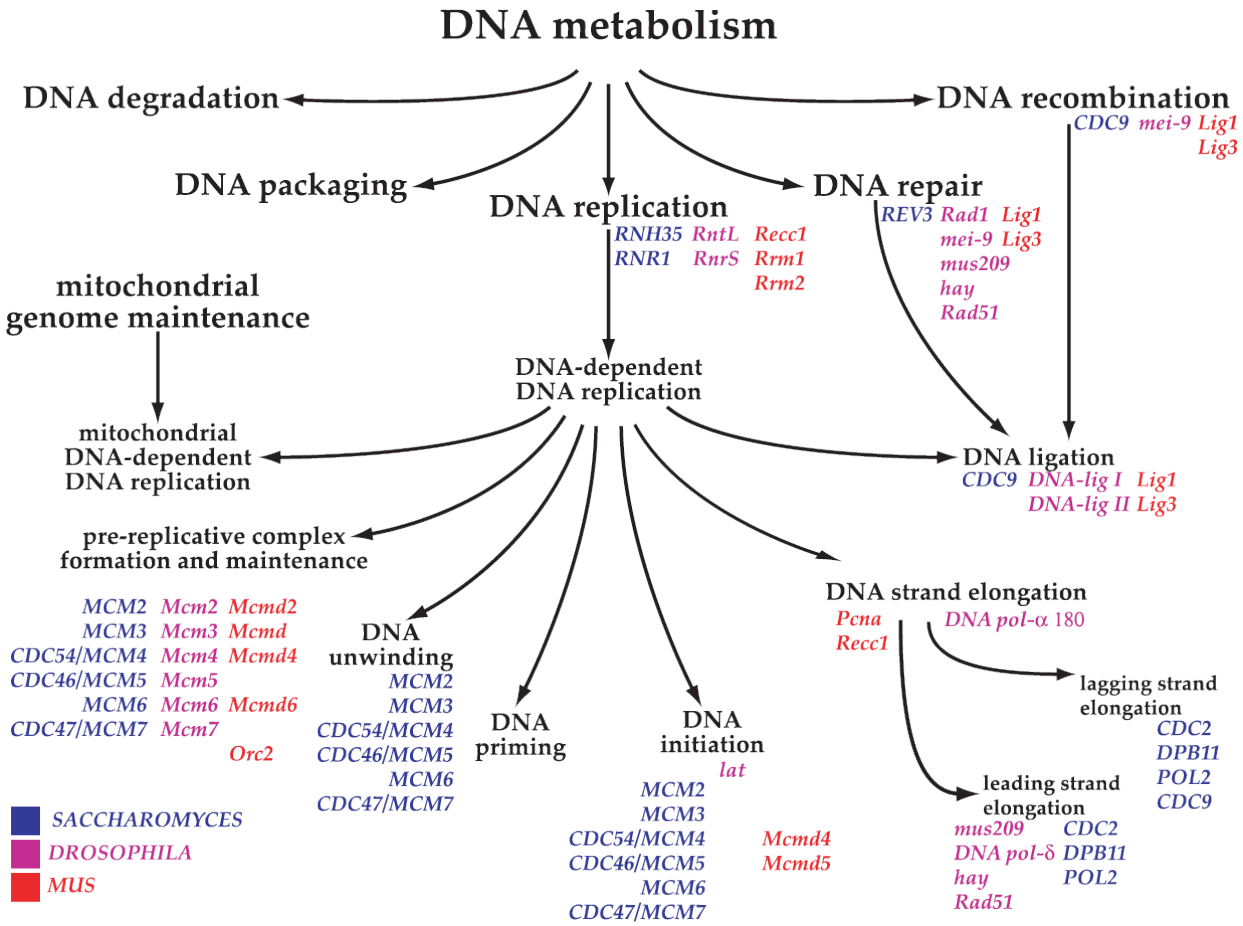
cellular component

- nucleus

http://www.geneontology.org/

Gene Ontology - biological process

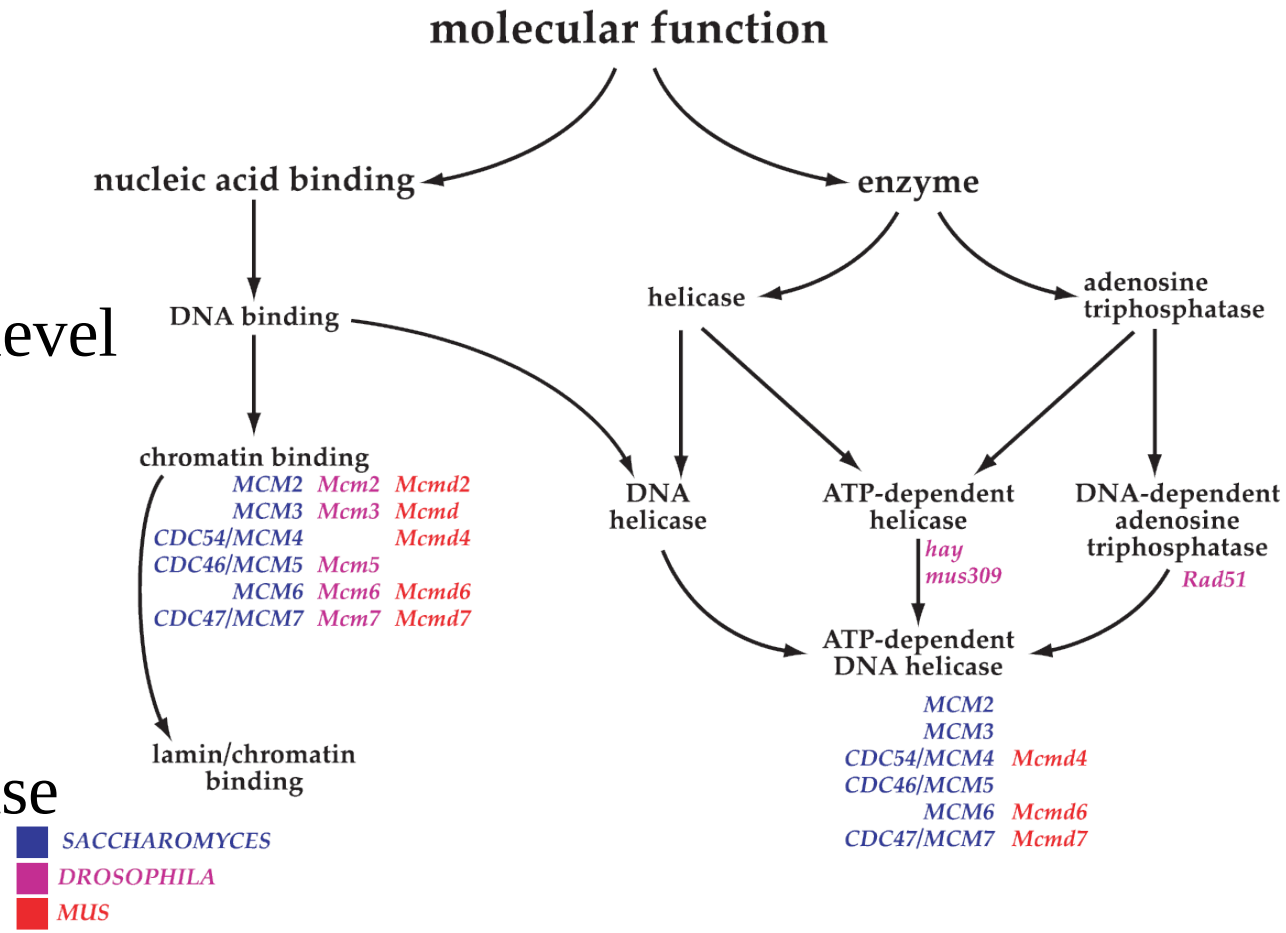
- "biological objective"
- not strictly chemistry
- nodes can have more than one parent
  - DNA ligation



- examples of high level
  - cell growth and maintenance
  - signal transduction

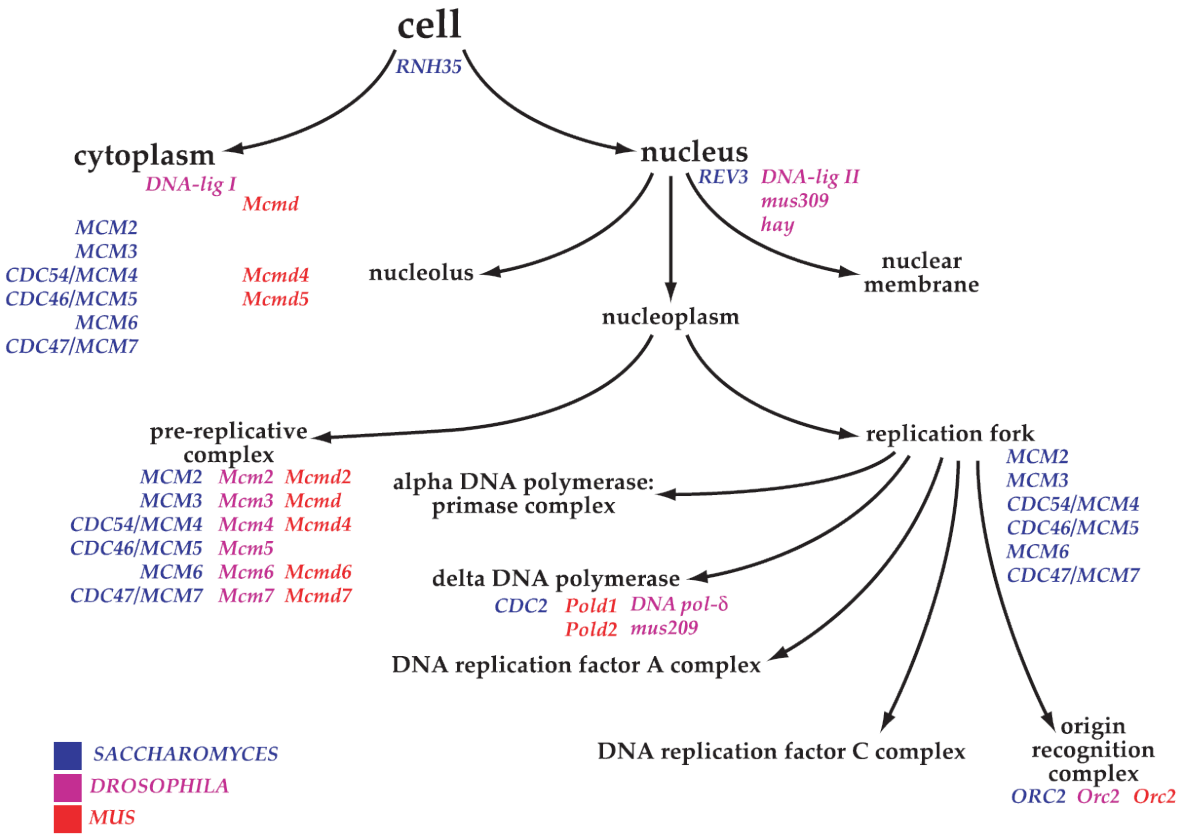
Gene Ontology - molecular function

- closer to enzyme classification
- examples of high level
  - enzyme
  - transporter
  - ligand
- lower level
  - adenylate cyclase

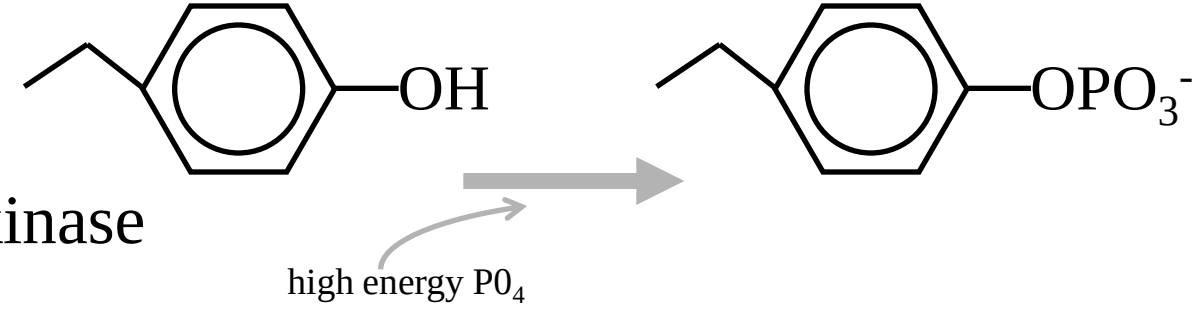


Gene Ontology - Cellular Location

- where is the gene active ?



Gene Ontology - flexibility



- Example - tyrosine kinase

- very common
- act on tyrosines in specific proteins
- 2 tyr kinase in me (different cells, processes)
  - molecular function same
  - biological process different
  - may have related sequences
- my tyr kinase / bacterial kinase
  - probably like above
- what about two different enzymes in same pathway ?

Gene Ontology - flexibility

- Imagine
  - protein 1 phosphorylates protein 2
  - protein 2 binds to protein 3 (which then binds to DNA)

- proteins 1, 2, or 3 may be coded on nearby genes
  - makes sense in terms of regulation / protein production

- different metabolic functions
- part of same "cellular process"

- useful ?
  - maybe one can predict the biological process
    - even without knowing exact function

Gene Ontology good / bad ?

- Much more flexible than EC numbers BUT
- Aim :
  - use a restricted / finite set of key terms
- PDB web site gives "GO" terms (**`www.rcsb.org`**)
  - lots of proteins without assignments
- the three descriptors (ontologies) are independent
  - should better fit to nature
- definitely better for non-enzyme proteins
- better able to handle badly characterised proteins
  - biological role - something to do with …x

Predicting Function - homology

- Truth
  - two proteins have high sequence similarity
  - structures are similar
- Hope
  - they have similar functions
- Truth
  - proteins with little sequence similarity can have similar structures
    - do they have similar function ? (address this later)

Function via homology

- pure sequence problem
- strategy obvious
  - take sequence + blast, psi-blast, HMMs, …

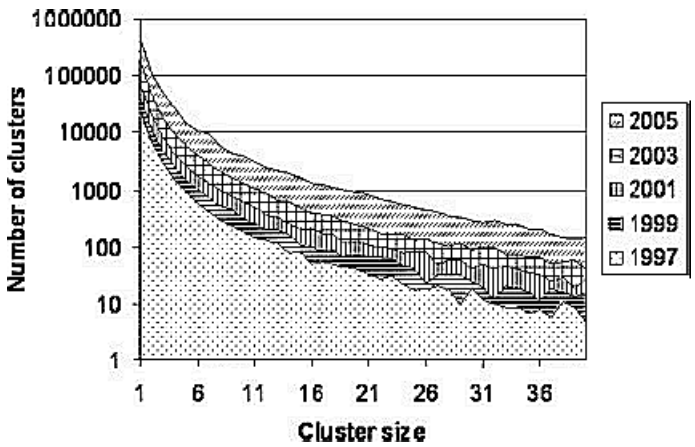
Problems

1. Are functions transferable ? Details later
2. Database growth leads to more mystery (next slide)
3. Propagation of errors

Database growth

- as more sequences are found, things should be more reliable
- number of mysteries also increases

- take a big databank
- cluster at 60 % identity
  - a cluster size 1 is a lonely protein
  - cluster size 6 has five friends



- Number of lonely sequences grows each year

Propagation of errors

- How does a mis-annotation occur ?
  - one little mistake with EC numbers, lab, typo, bug
- How does it propagate ?
  - every successive, similar sequence will inherit mistake
- Does it happen ?
  - often
- Often seen ?
  - only when there are gross inconsistencies
  - work is independently repeated

Motifs and Pieces of Proteins

- more on this topic from Frau Willhoeft (ASE)
- Belief…
  - in a protein, small fragments are recognised
  - Names
    - motifs, patterns, sequence logos
  - one method to find them
    - collect proteins you believe have a feature
      - align
      - look at preferences within each file
  - scanning against patterns ?
    - regular expressions
    - classic sequence searches

```
LVPLFYKTC
LVPlFYKTC
LVPLFYKTC
LVPLFYKTC
LVPLFYKTC
LIPPFYKTC
LVPPFWKTC
LVPPFWKTC
LVPIAHKTC
LIPIAHKTC

L[VI]P[LPI][FA]…
```

Motifs and Pieces of Proteins - Example Patterns

- Acetyl-CoA carboxylase carboxyl transferase alpha subunit signature
- Acetate kinase family signature
- Fish acetylcholinesterase signature
- Insect acetylcholinesterase signature
- Acetyl-CoA biotin carboxyl carrier protein signature
- AMP-binding signature
- Chitin-binding domain signature
- Cholinesterase signature
- Citrate synthase signature
- CLC-0 chloride channel signature
- Carbamoyl-phosphate synthase protein CPSase domain signature
- Snake cytotoxin signature
- + 10 000 more

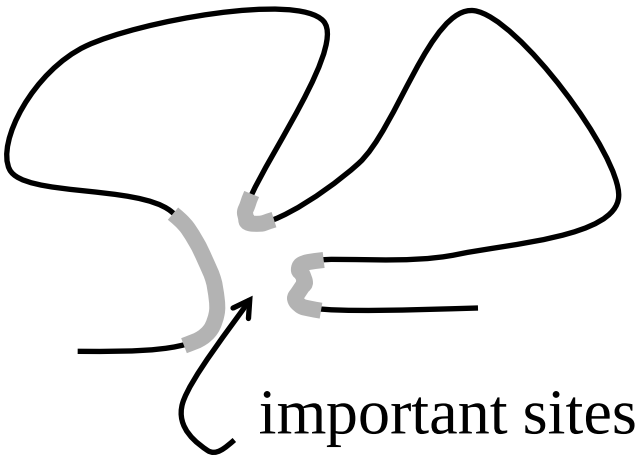
- is this a function prediction ?
  - maybe (a bit)

Motifs and Pieces of Proteins - reliability

- how reliable ?
  - Übung on topic
  - good servers
    - calculate how often a match will be seen by chance
    - should be able to give reliable statistics
- do we like them ?
  - fundamental problem
  - difficult to see how characteristic a pattern is
    - not a causal relationship
- structural versus local sequence properties…

Motifs and Pieces of Proteins - reliability

- function reflects 3D arrangement of residues
- how often will that be reflected by a short range sequence pattern ?

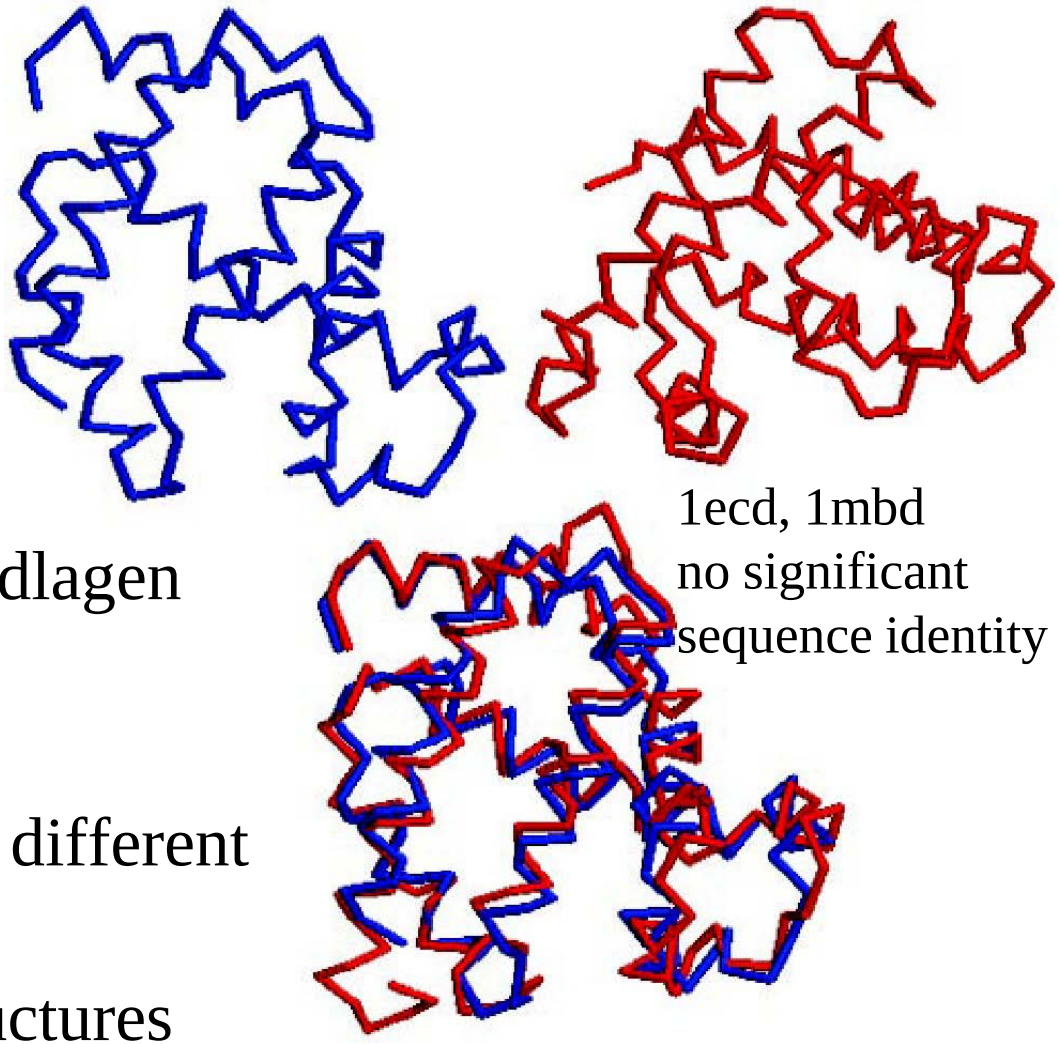


- good reason to start thinking about 3 D

First a little diversion

- Often one wants a set of proteins with similar structure
  - to look for patterns / features
  - classification treated more thoroughly later

3D Similarity



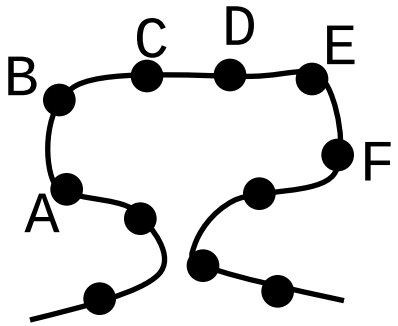
1ecd, 1mbd
no significant
sequence identity

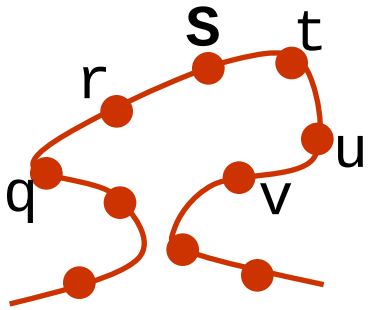
- More thorough in the grundlagen Kurs
- True:
  - proteins may have very different sequences
  - surprisingly similar structures

3D similarity

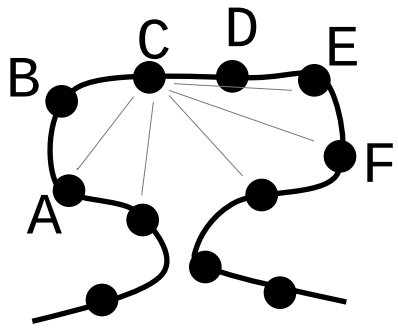
- Aligning two structures (without sequence)
  - fundamentally much harder than sequence alignment (NP complete)
- sequence version - calculate an alignment
  - to score **S**, compare against ABC..
- with structures
  - what is similarity of **S** with ABC.. ?
    - depends on qr..tu
- several approaches

```
ABCDEF
qrStuv
```



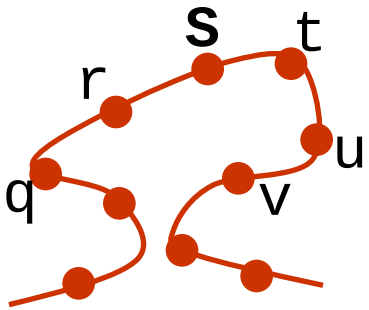


3D similarity



- slide struct 1 over 2
  - step wise try to look for match (not good)
- label each site in struct 1 & 2 with some information

- I can now compare the distance matrix around `C` to each in second structure



- I could also label each site
  - with secondary structure
  - any representation of structural properties

3D similarity

Result - we can take any structure and find similar ones

- without sequence similarity

Important ?

- belief - evolution
- you have a functioning enzyme
  - constantly suffering mistakes, mutations, deletions, insertions
  - if the shape changes - you die
  - if the function is lost - you die
- eventually evolution will explore all sequences which have not killed you

- not such a good model for evolution - but fundamental belief
  - sequence varies more than structure

3D similarity

- If you have the structure of your protein
  1. search for sequence similar proteins

  if that fails

  2. search for structural similarity

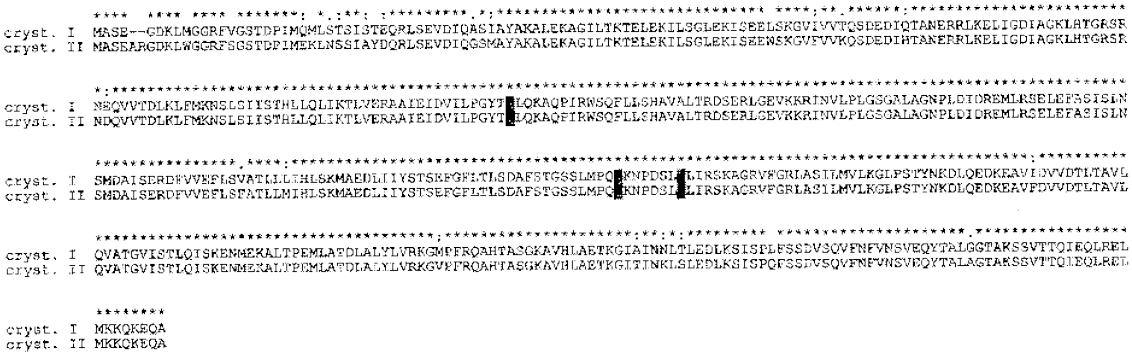
- How reliable is this philosophy ?

Sequence homology ?

- the sequence hardly changes
- complete loss of enzyme activity
- different function

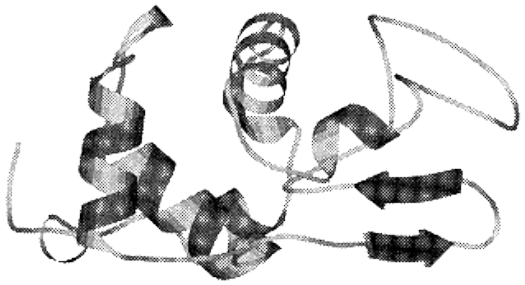
or

- 40 % identity still not enough



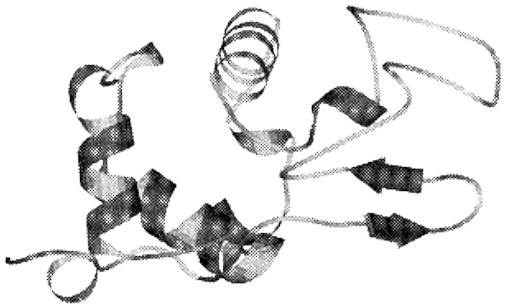
duck crystallin δI *non-enzyme*
duck crystallin δII/argininosuccinate lyase *enzyme*

HOMOLOGS
LOSS OF ENZYME ACTIVITY
94% seq ID
conserved active site



human lysozyme
enzyme

HOMOLOGS
ENZYME / NON-ENZYME
40% seq ID
disruption of active site

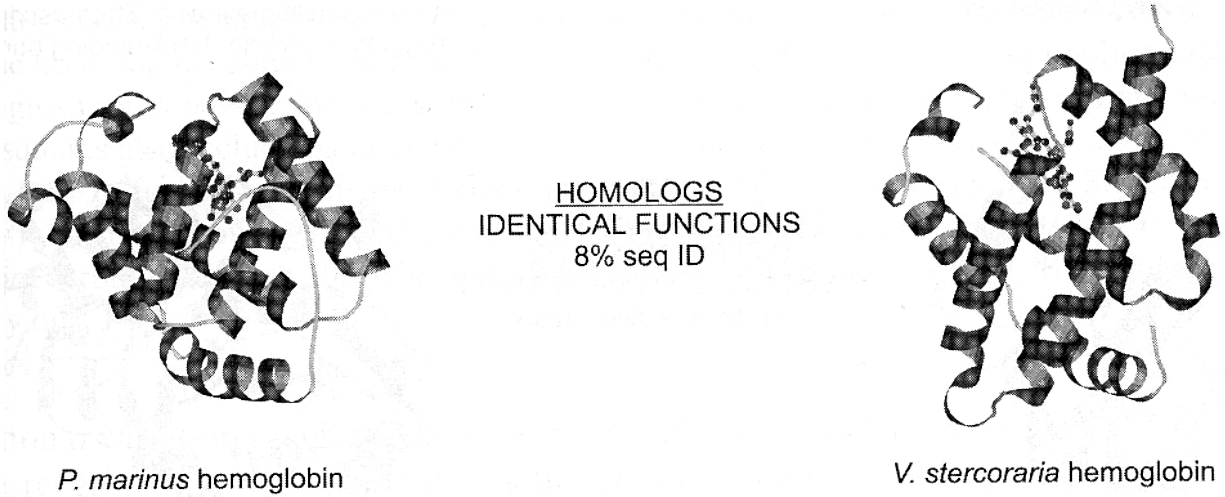


human α-lactalbumin
non-enzyme

Todd, A in The Proteomics Handbook ed Walker, J.M., Humana Press 2005, Totowa, "Deriving Function From Structure: Approaches and Limitations"

Homology

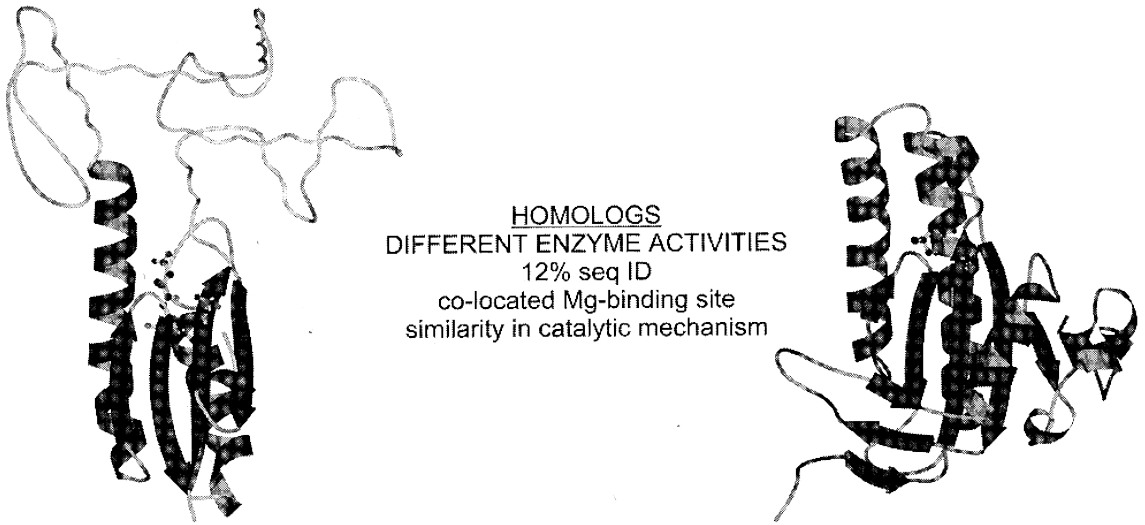
- What one normally expects
  - sequence is less conserved than function
- basis of all methods discussed so far



Homology

- sometimes function will change
  - not totally unrelated

- example where function is not yes / no



Homology

- Worst case

acylphosphatase

STRUCTURAL ANALOGS
SIMILAR FOLDS
DIFFERENT FUNCTIONS
no shared functional attributes

bovine papillomavirus-1 E2 transcription regulation protein, DNA-binding domain

Imagine

- search by sequence - fails
- search by structure - produces an impressive similarity

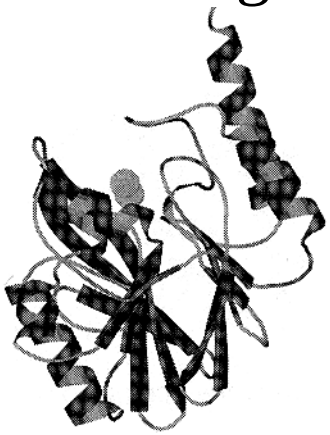
Todd, A in The Proteomics Handbook ed Walker, J.M., Humana Press 2005, Totowa, "Deriving Function From Structure: Approaches and Limitations"

Protein Structure Classifications

- Names are for completeness only
- Nothing on this Folien examinable
- Protein alignments are difficult
- Classifications are made, put in boxes to be played with
- Pure structure similarity
  - program dali, classification FSSP
- Some very much hand made
  - "SCOP" – ex Russian looks at new structures and puts them in classes
  - "CATH" – English group (Orengo) mixes automatic decisions and hand "curation"
- Claim
  - if we can automatically find a "SCOP" class, we have predicted function

3D Motifs

- Philosophy - with evolution
  - sequences change + structures change
- what really dictates enzyme function ?
  - the set of residues around the "active site"
  - even when the fold changes
- need methods to find similar arrangements of residues



β-lactamase class B
EC 3.5.2.6
metal-dependent

FUNCTIONAL ANALOGS
DIFFERENT FOLDS
IDENTICAL ENZYME ACTIVITY
different active sites

β-lactamase classes A, C, D
EC 3.5.2.6
catalytic Ser nucleophile

Todd, A in The Proteomics Handbook ed Walker, J.M., Humana Press 2005, Totowa, "Deriving Function From Structure: Approaches and Limitations"

3D Motifs

- Ingredients
- definition of a 3D pattern / motif
- collection of data from proteins
  - library / database of patterns
- method to search for patterns

- start with collection

3D Motifs - data collection

- do we always know the catalytic residues ?
- horrible manual method …
- One approach
  - for each enzyme in an existing database
  - if PDB file exists
    - if (authors have marked active site residues)
      - if (plausible - agrees with chemical literature)
        - store residues + enzyme function in database
- another approach and example

3D Motifs - another approach

Scheme

- definition of interesting groups
- for each protein in some database
  - find all interesting groups which are near each other
  - store the relationships

- for a new protein
  - look for sets of interesting groups
  - compare against the list for proteins in database

- what are interesting groups ?

Jambon, M., Imberty, A., Deléage, G., Geourjon, C (2003), Bioinformatics, 52, 137-145, "… detect common 3D Sites in Protein Structures"

3D Motifs - Interesting Groups

- for each amino acid, think about what is likely to be important
- slightly arbitrary
- emphasis on soluble groups (not exclusively)

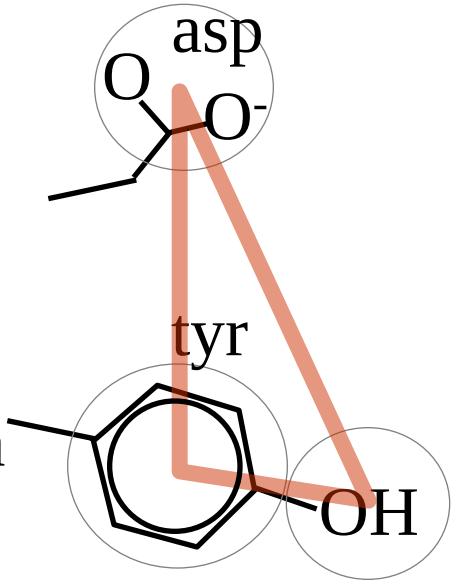
- how are relationships defined ? stored

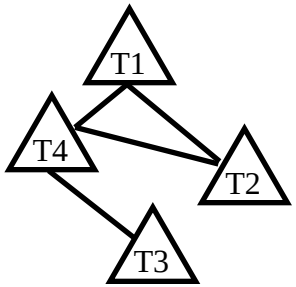
Amino acid	chemical groups
Alanine	
Arginine	guanidinium
Asparagine	amide
Aspartate	carboxyl
Cysteine	thiol
Glutamate	carboxyl
Glutamine	amide
Glycine	glycine
Histidine	aromatic, ammonium
Isoleucine	
Leucine	
Lysine	ammonium
Methionine	thioether
Phenylalanine	aromatic
Proline	proline
Serine	hydroxyl
Threonine	hydroxyl
Tryptophan	aromatic,aromatic, amino
Tyrosine	aromatic, hydroxyl
Valine	

Jambon, M., Imberty, A., Deléage, G., Geourjon, C (2003), Bioinformatics, 52, 137-145, "… detect common 3D Sites in Protein Structures"

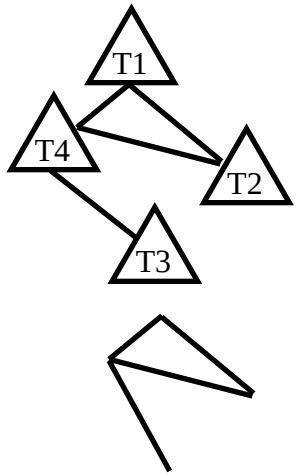
3D Motifs - relationships

- for each group
  - centre of mass of group *i* is c_i
- walk over protein and find all pairs of groups $c_i c_j < 8$ Å
- find every triangle
  - store triangle
    - types of groups (OH, carboxyl, …)
    - buried / surface information
- connections of triangles
  - find every pair of triangles with a common edge - join them





3D Motifs - relationships



- From chemistry to a little graph
  - representation of which groups are most close to other groups
- Do this for every protein in library
  - each protein is represented by a graph

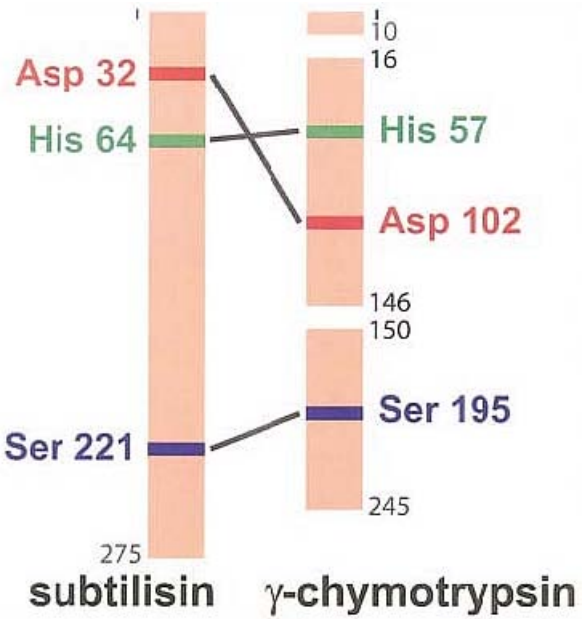
- Query protein
  - turn this into a graph
- Query procedure
  - look for common subgraphs (arrangements of groups)

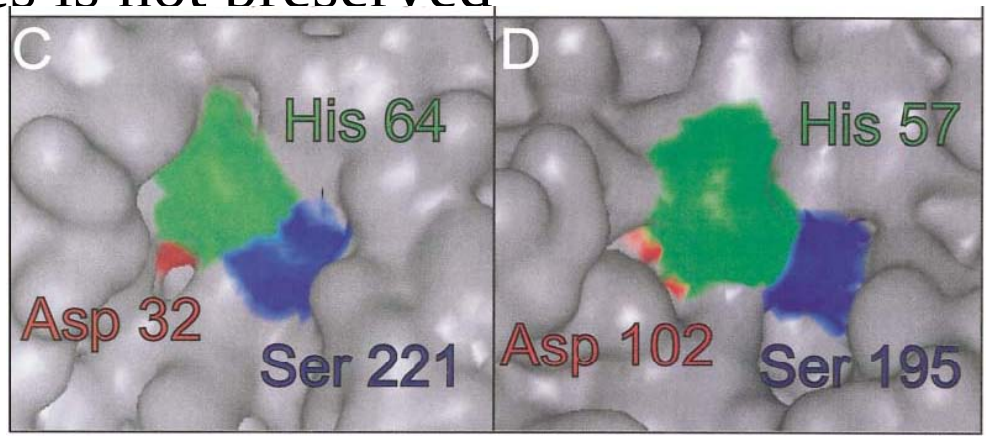
- Does this work ? Examples from authors

Example result

"serine proteases"

- more than one family of proteins
  1. subtilisins
  2. chymotrypsins
     - no sequence similarity
     - no structural similarity
     - active sites are similar
- the order of important residues is not preserved
  - the structure is:





- Is this the best / only approach ?

Jambon, M., Imberty, A., Deléage, G., Geourjon, C (2003), Bioinformatics, 52, 137-145, "… detect common 3D Sites in Protein Structures"

3D Motifs

- This was an example
  - starting from triangles is arbitrary
  - thresholds (points < 8 Å)
- Are results believable ?
  - false positives ? false negatives ?

3D Motifs – more examples and more details

- A different definition of 3D motifs
- how to search for them
- judging their significance

3D Motifs – skeletons / graphs

Ingredients and philosophy

- require a classification of families
- whole proteins turned into simple graphs

- look for common regions in families
  - call these fingerprints
  - a "family" may have several "fingerprints"
- look for fingerprints in new proteins
- assess significance

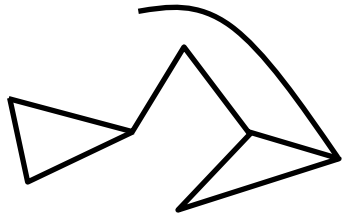
- Steps

Scheme from Bandyopadhyay, Huan, J… Wang, W and Tropsha, A., Protein Sci (2006) 15, 1537-1543

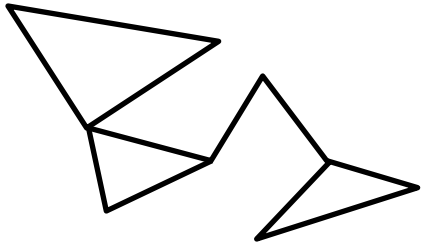
3D Motifs – skeletonising a protein

- make $C^{\alpha}C^{\alpha}$ distance matrix
  - each edge is put into distance class:
    - nodes are C^{α}
- for family (typically 5 to 50 proteins)
  - look for common subgraphs

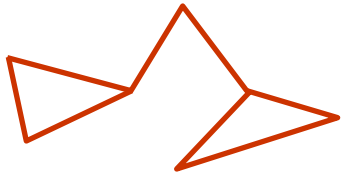
distance Å
0-4
4-6
6-8.5
8.5 – 10.5
10.5-12.5
12.5 – 15



prot 1



prot 2



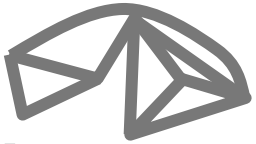
common subgraph

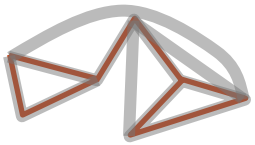
- not finished yet

3D Motifs – "fingerprint identification"

- for a family - we have subgraphs
- repeat graph calculation for large set of proteins (unrelated)
- fingerprint subgraphs
  - in > 80 % of family
  - in < 5 % of background

Query protein ?



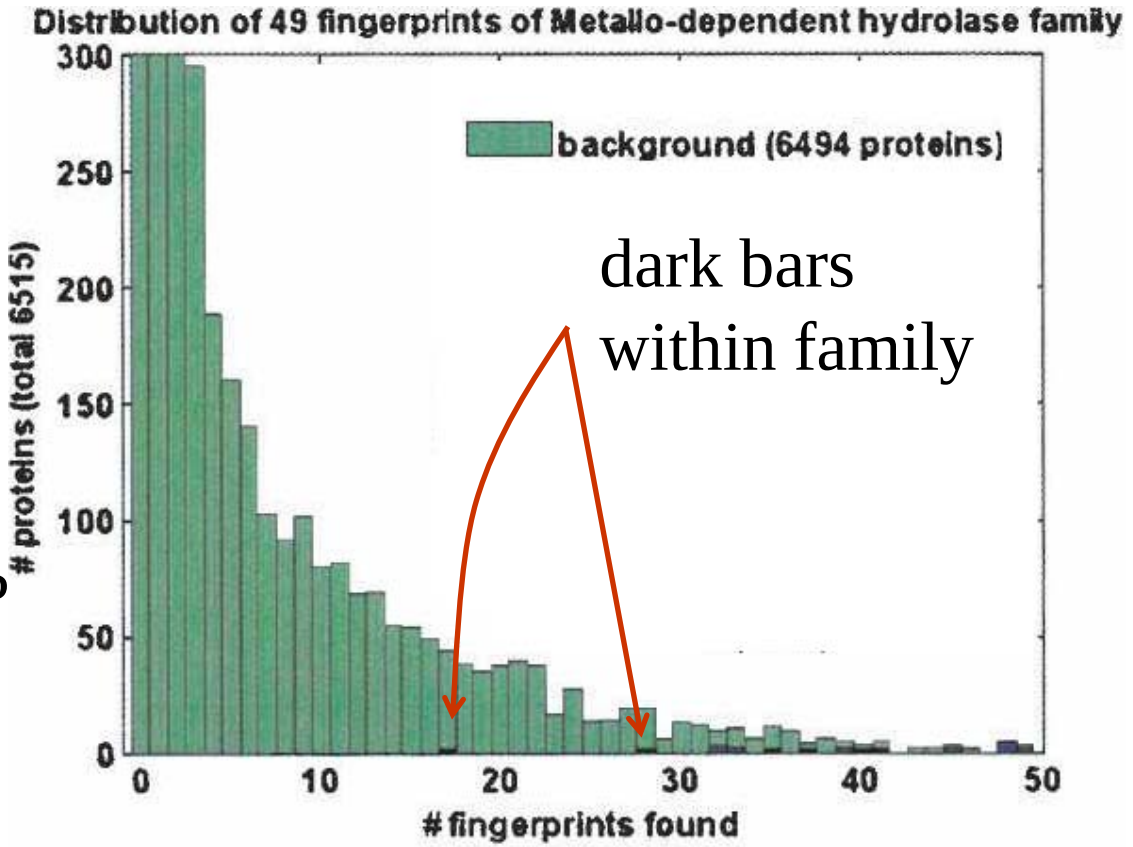


- protein → graph
  - compare query + family graphs
- if query contains the "fingerprint" of a family
  - maybe part of family
- quantify this

3D Motifs – significance of matches

- A family has more than one fingerprint
- some fingerprints are unique, some often seen
- for each –calibrate the significance

- family has 49 fingerprints
- for 6515 proteins check
  - how many have 1 fingerprint, 2, 3,…
- they are specific
- do they miss examples ?
  - rarely



from Bandyopadhyay, Huan, J… Wang, W and Tropsha, A., Protein Sci (2006) 15, 1537-1543

Summary of fingerprints

- Find classes (from literature)
- For each class
  - get 10's of "fingerprints" (distance information + residue type)
  - these are spatially conserved residues across a family
- For queries – look for how many fingerprints are present

- Claim
  - this is not just like structure comparison
    - "SCOP" families are usually functionally the same
  - looks for patterns of matching residues

Summary of fingerprints

- Is method perfect ?
  - the distance definitions are rigid
  - relies on a database from literature
- graph matching
  - very expensive to do rigorously
  - "maximal common subgraph problem"

Summary of function prediction

- Function is difficult to define
  - best if turned into machine readable form

- Transfer of belief via homology dominates annotations
- Homology found / errors transferred
  - via sequence
  - via structure
- Motifs / patterns
  - via sequence or structure
  - rather arbitrary definitions
- Examples here (data collection, recognition)
  - only examples / case studies