Protein Function Prediction

Protein function - field of biochemists

- can it be predicted / guessed from
  - structure ?
  - sequence ?

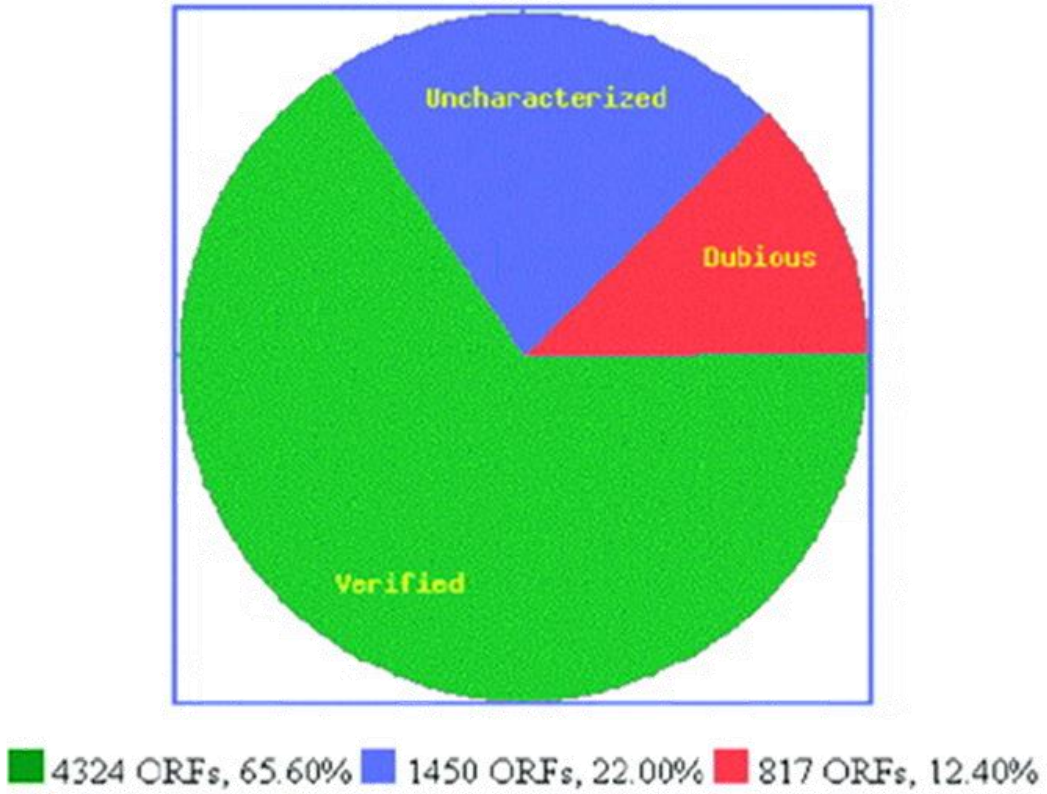
Is this an issue ?

- 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

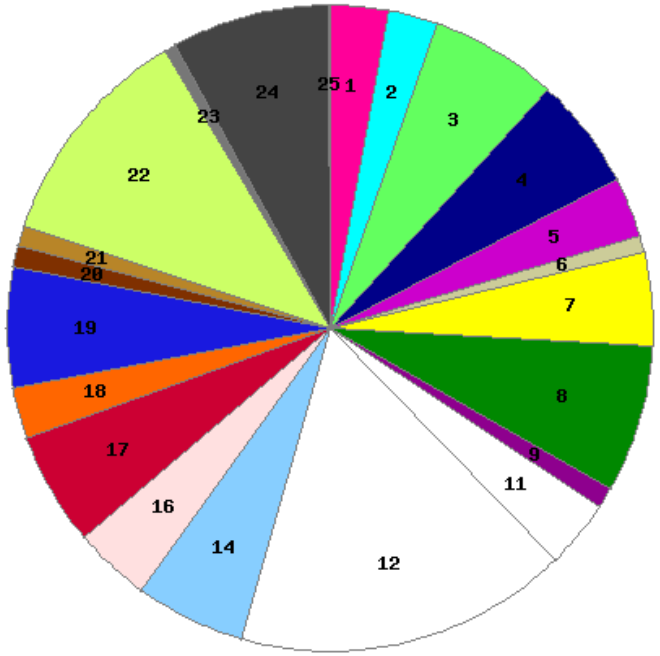
Not really known what many proteins do



Protein function may not be easy

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

e. coli



Color	Gene Role Category	
1	Amino acid biosynthesis	3.16 %
2	Biosynthesis of cofactors, prosthetic groups, and carriers	2.67 %
3	Cell envelope	6.90 %
4	Cellular processes	6.01 %
5	Central intermediary metabolism	3.33 %
6	Disrupted reading frame	0.84 %
7	DNA metabolism	5.07 %
8	Energy metabolism	7.92 %
9	Fatty acid and phospholipid metabolism	1.20 %
	gene/protein expression	0 %
11	Hypothetical proteins	3.65 %
12	Hypothetical proteins - Conserved	18.0 %
	metabolism	0 %
14	Mobile and extrachromosomal element functions	5.96 %
15	Pathogen responses	0 %
16	Protein fate	4.18 %
17	Protein synthesis	6.14 %
18	Purines, pyrimidines, nucleosides, and nucleotides	2.71 %
19	Regulatory functions	6.50 %
20	Signal transduction	1.20 %
21	Transcription	1.02 %
22	Transport and binding proteins	12.2 %
23	Unclassified	0.62 %

From cmr.tigr.org

- very well studied, common bacterium
- 5×10^3 genes

Plan

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

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

Pre-summary

- function prediction is staggeringly important
    - need to know some common terms (this week)
- I dislike it

Philosophy

Sie müssen nicht alles glauben

Function prediction is

- easy (homology)
- act of faith (interesting)

This week

- background and meaning of function

Later

- case studies (Fallstudien)

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

- 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

- biological process
- molecular function
- cellular component

example 1uw0

- blessed by protein data bank

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/

Ashburner et al, Nature Genet. 25, 25-29 (2006) "Gene Ontology: tool for the unification of biology"

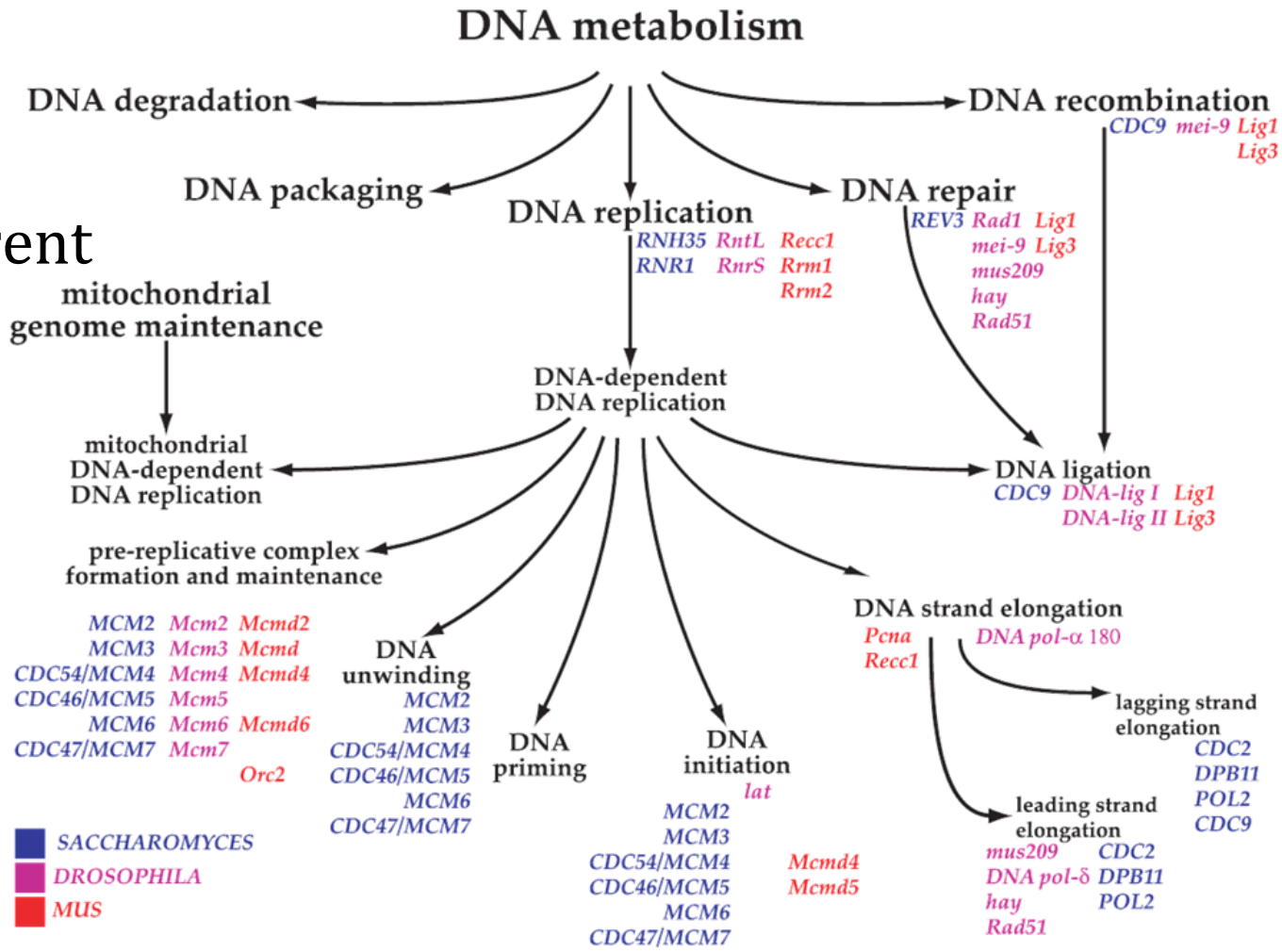
Gene Ontology - biological process

"biological objective"

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

examples of high level

- cell growth and maintenance
- signal transduction



Ashburner et al, Nature Genet. 25, 25-29 (2006) "Gene Ontology: tool for the unification of biology"

Gene Ontology - molecular function

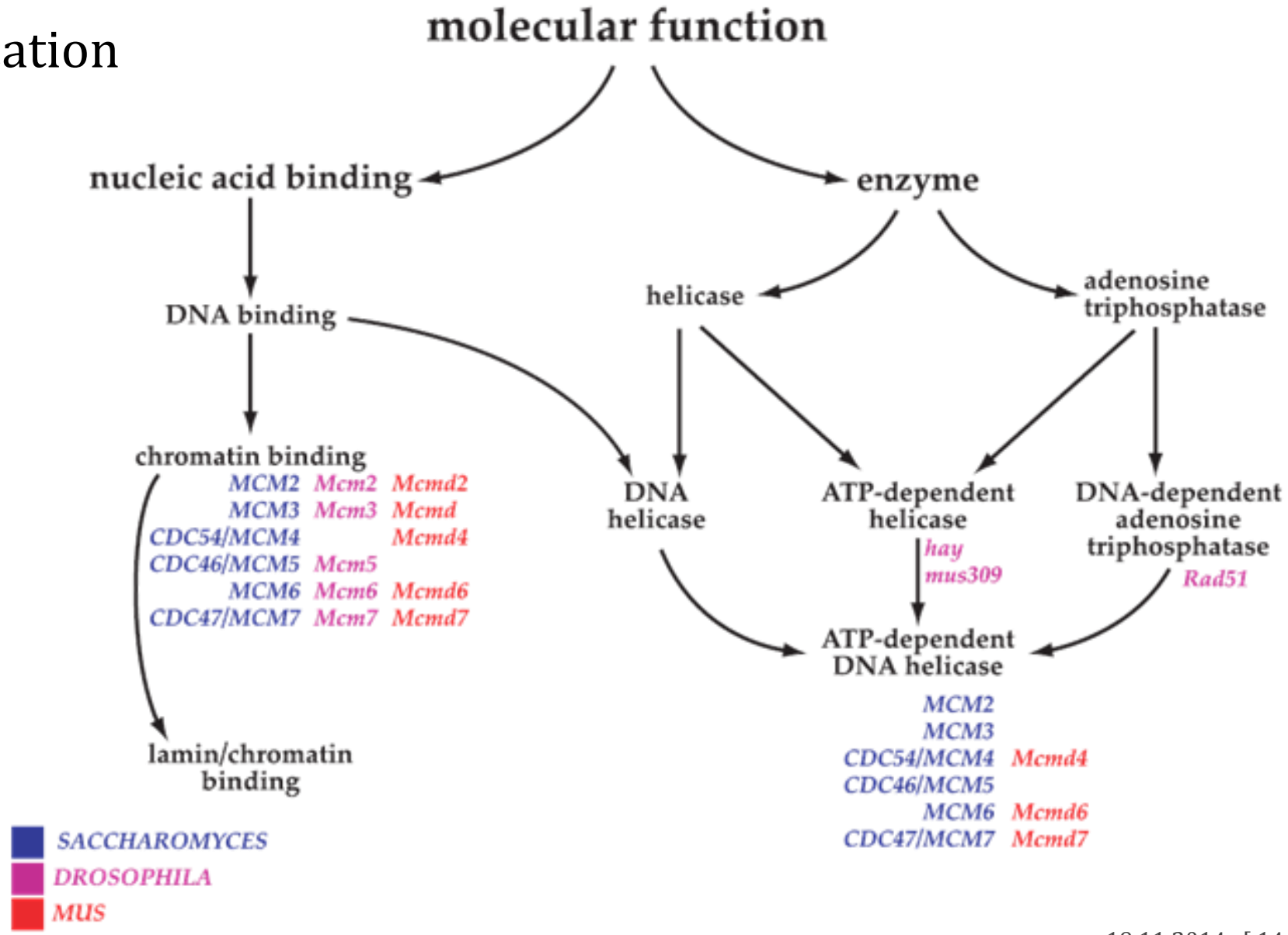
Closer to enzyme classification

Examples of high level

- enzyme
- transporter
- ligand

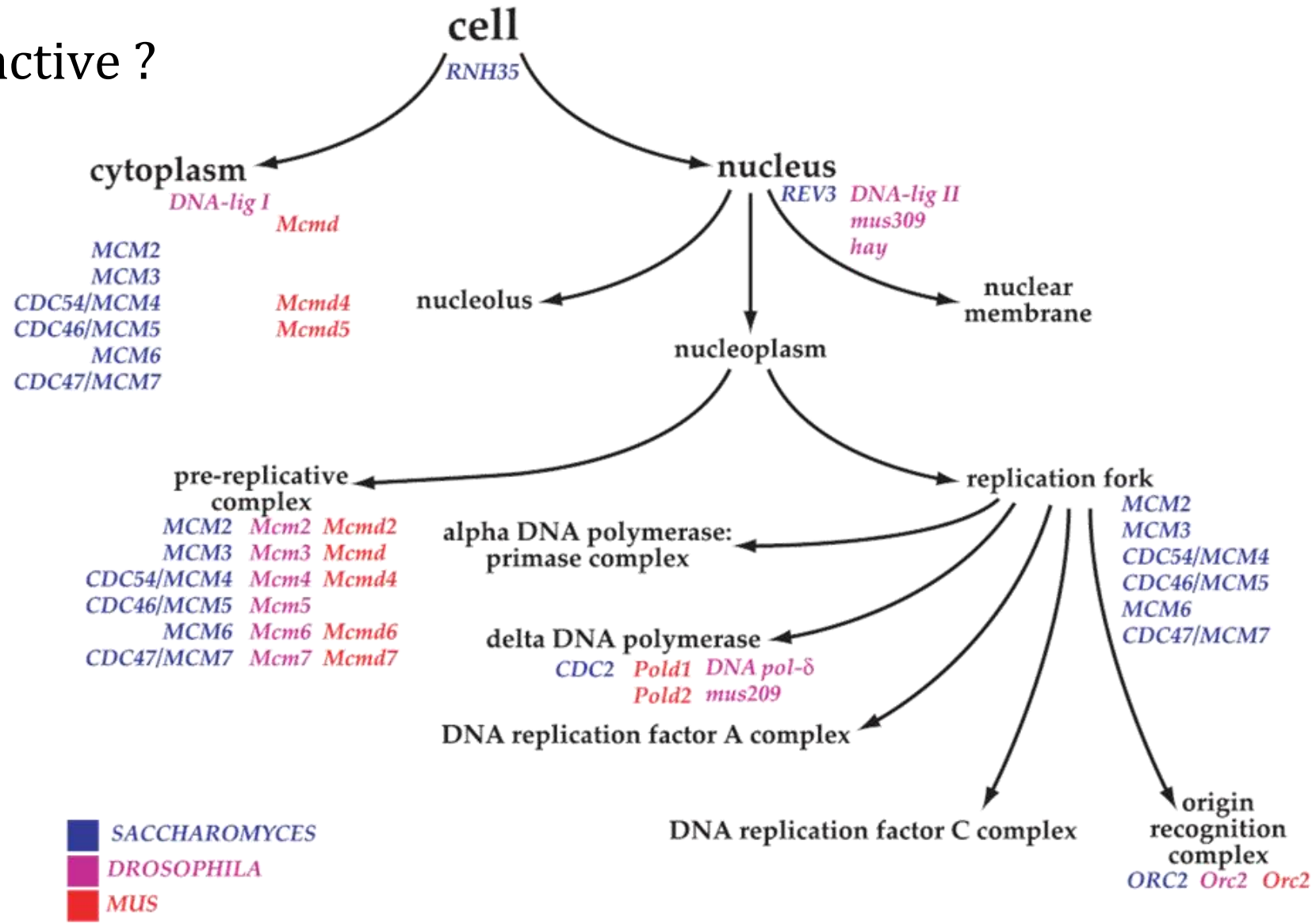
Lower level

- adenylate cyclase

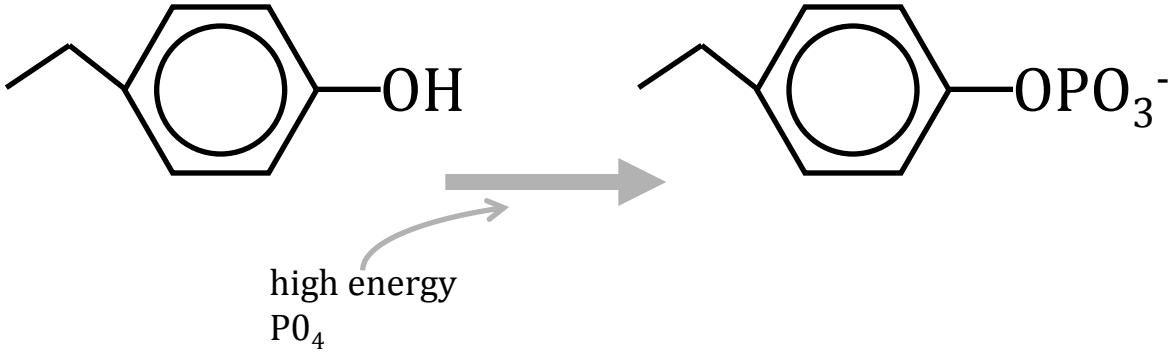


Gene Ontology - Cellular Location

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

- 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. Propagation of errors

Propagation of errors

How does a mis-annotation occur ?

- one little mistake with EC numbers, lab, typing mistake, 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 Giorgio (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

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

L[VI]P[LPI][FA]…
```

Scanning against patterns ?

- regular expressions
- classic sequence searches

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

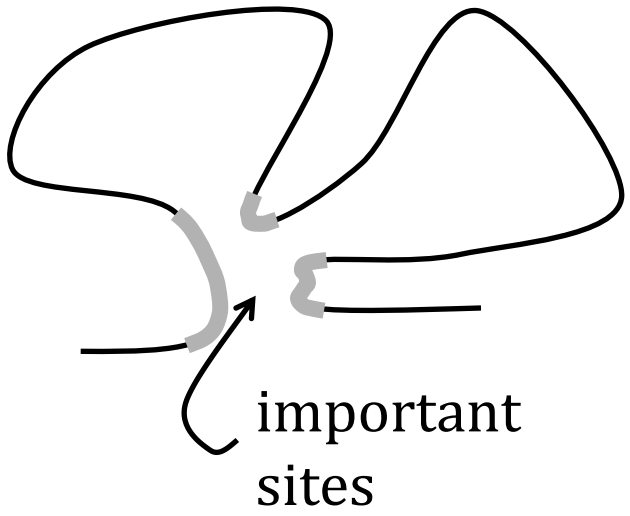
Do we like them ?

- fundamental problem
  - difficult to see how characteristic a pattern is
    - not a causal relationship
  - co-occurrence ≠ causality

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 ?

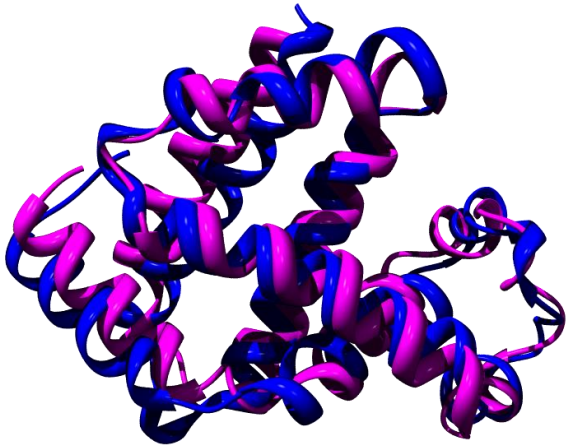


- good reason to start thinking about 3 D

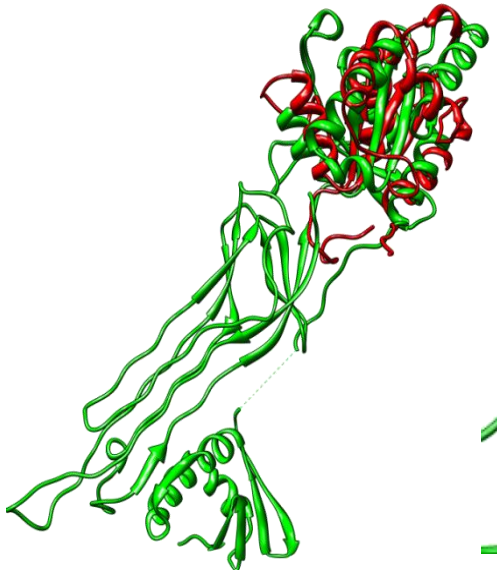
First a little diversion

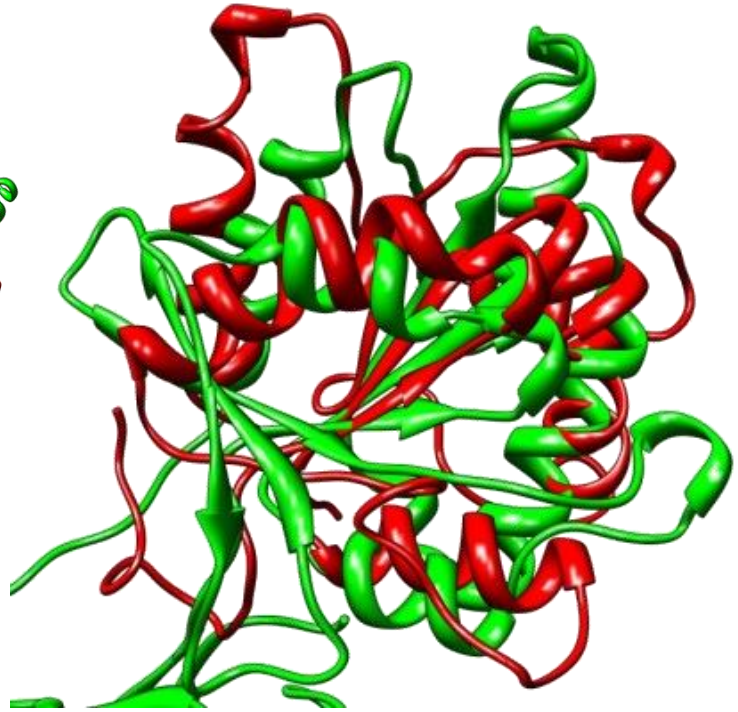
- Often one wants a set of proteins with similar structure
  - to look for patterns / features
  - classification treated more thoroughly later

3D Similarity



haemoglobin & erythrocruorin
14 % sequence id





1fyv & 1udx, TLR receptor and nucleotide binder,
9 % sequence id

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

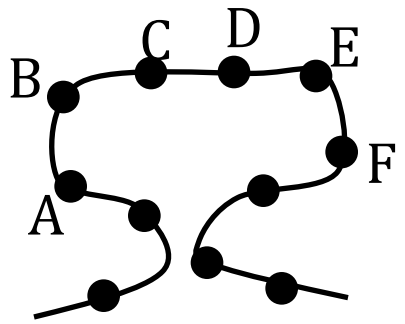
- to score **S**, compare against A,B,C,...

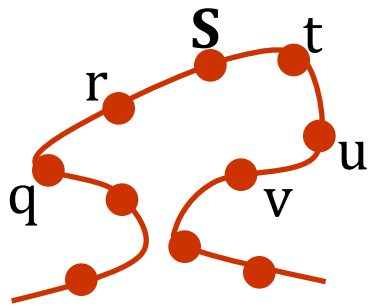
With structures

- what is similarity of **S** with A,B,C,... ?
  - depends on qr..tu

- several approaches

```
ABCDEF
qrStuv
```

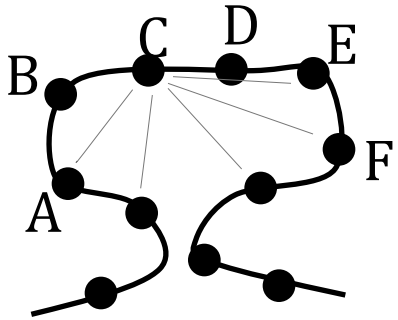




3D similarity

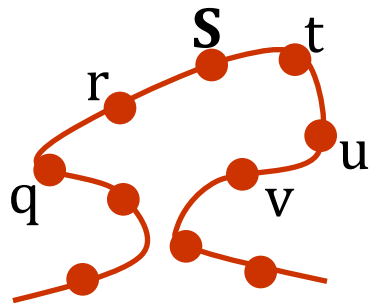
Slide struct 1 over 2

- step wise try to look for match (not good)



Label each site in struct 1 & 2 with some structural information

- distance matrices (local distances)
- 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

- fundamental claim
  - sequence varies more than structure

3D similarity

Even if you have the structure of your protein

1. search for sequence similar proteins

   if that fails

2. search for structural similarity

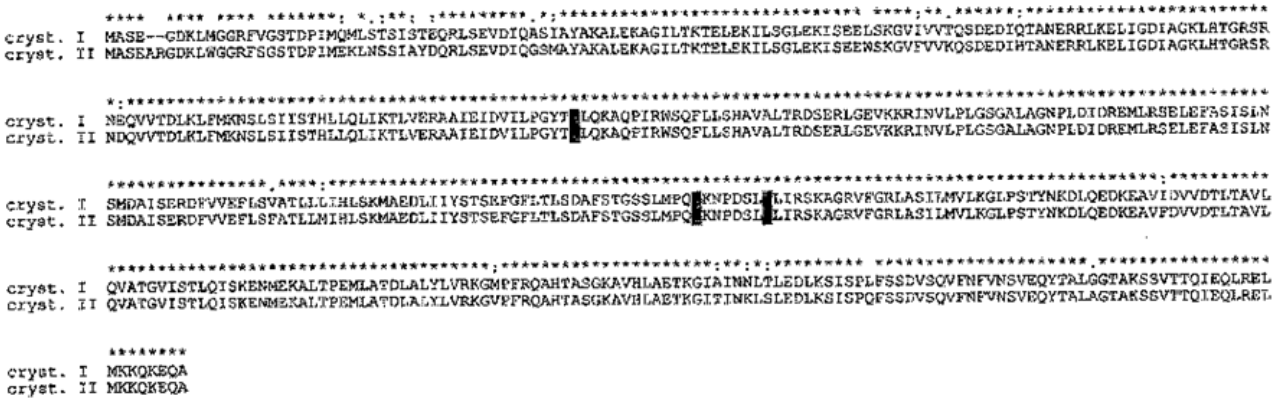
This is best, but even here there are exceptions

Sequence homology ?

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

or

- 40 % identity still not enough

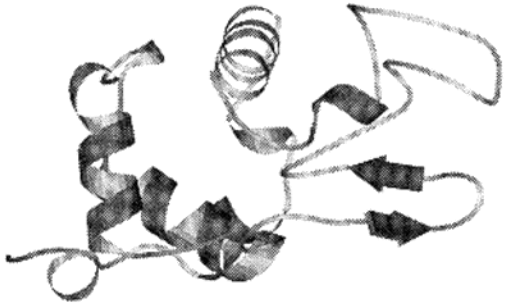


duck crystallin δI *non-enzyme*
duck crystallin δII/argininosuccinate lyase *enzyme*

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"

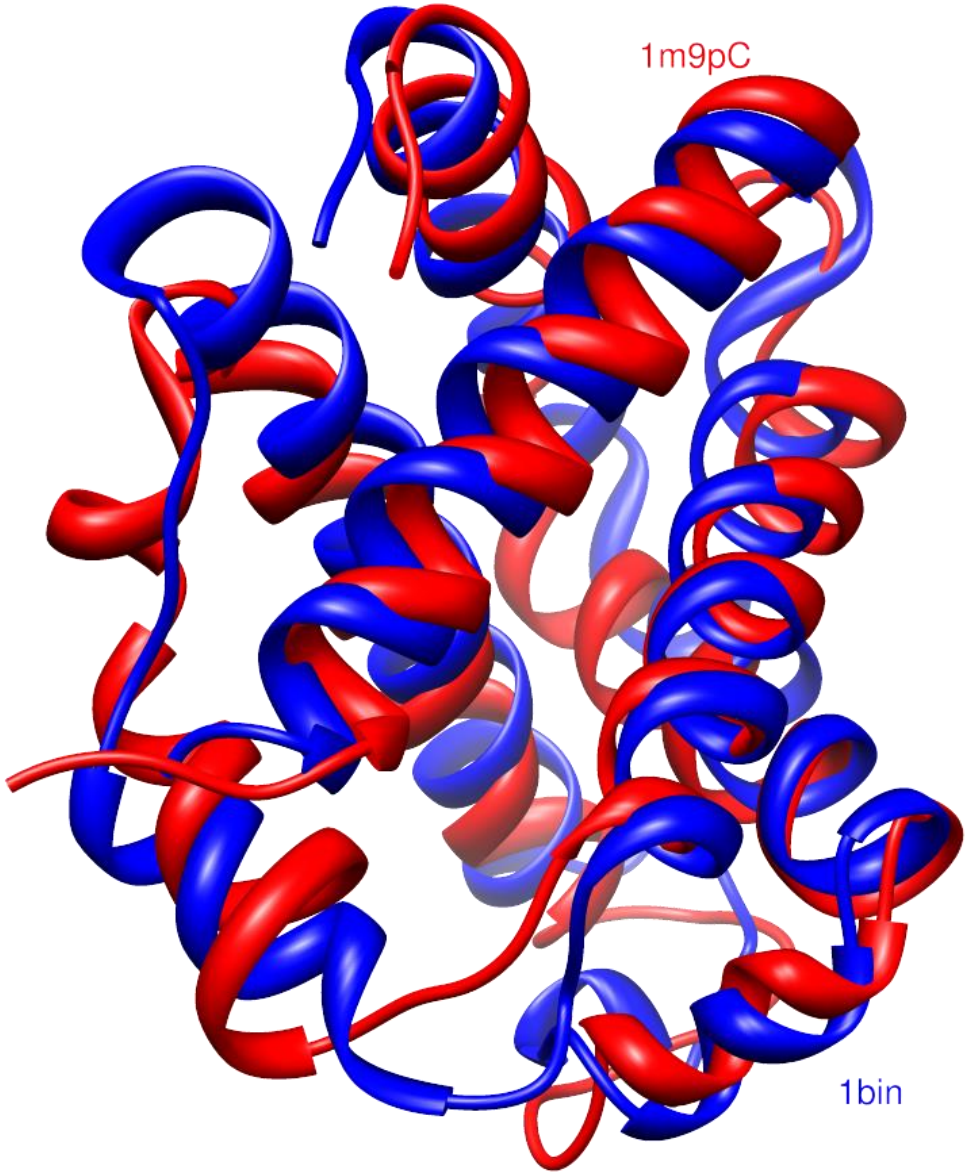
Homology

What one normally expects

- sequence is less conserved than function

- basis of all methods discussed so far

- human haemoglobin
  soybean haemoglobin
  12 % (very low) sequence identity
  same function

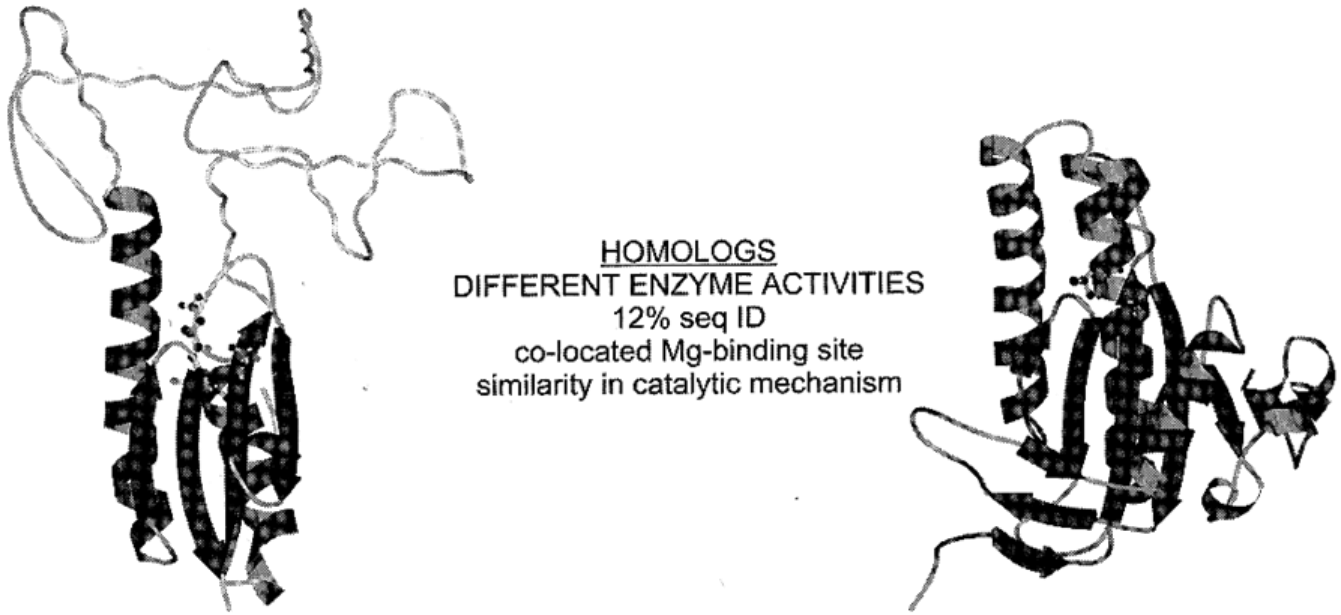


Homology

Sometimes function will change

- not totally unrelated

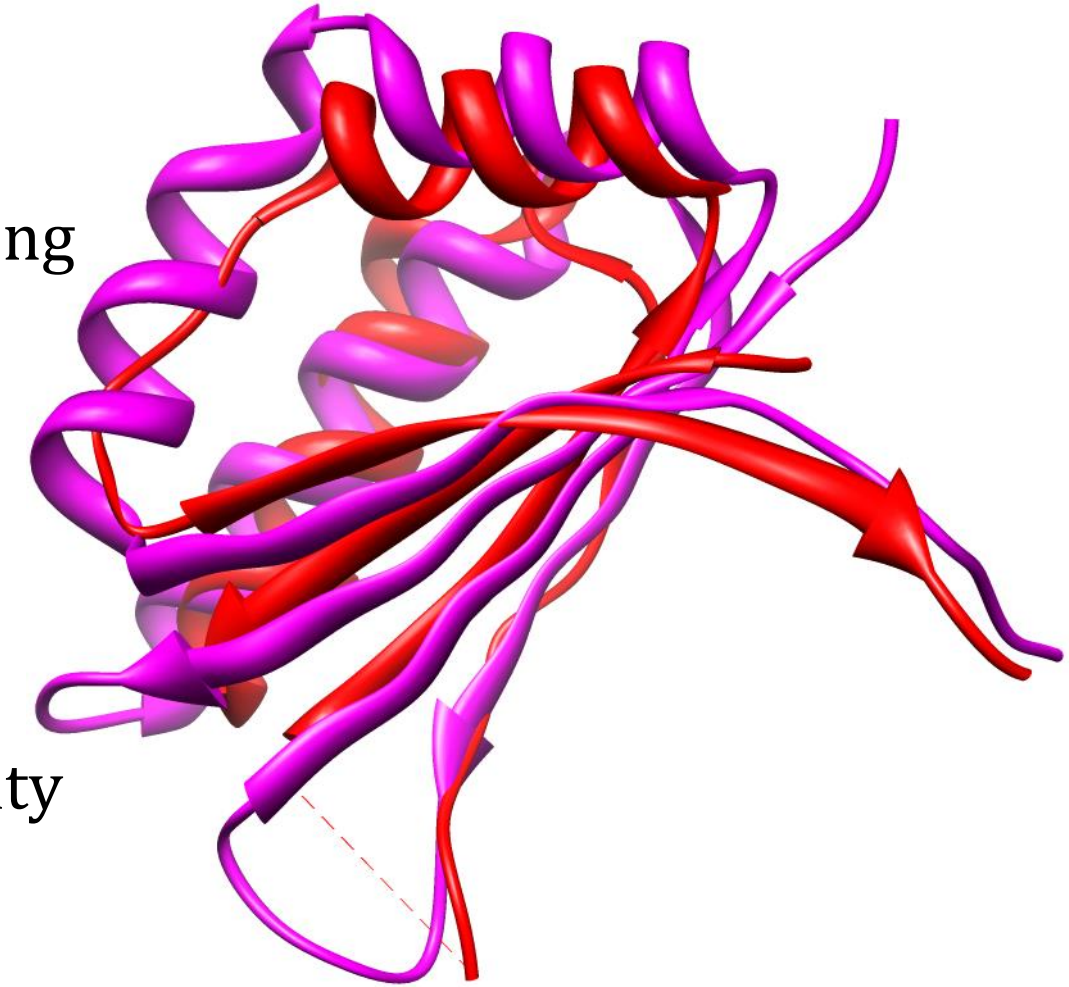
- example where function is not yes / no



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

Nasty case

- structural similarity
- seq similarity 5%
  - 1jjh papillomavirus DNA binding
  - 3kg0 streptomyces oxygenase



very similar structures

- no evidence of functional similarity

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

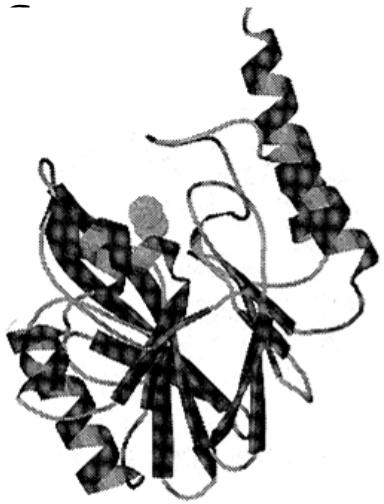
Philosophy - with evolution

- sequences change + structures change

What really dictates enzyme function ?

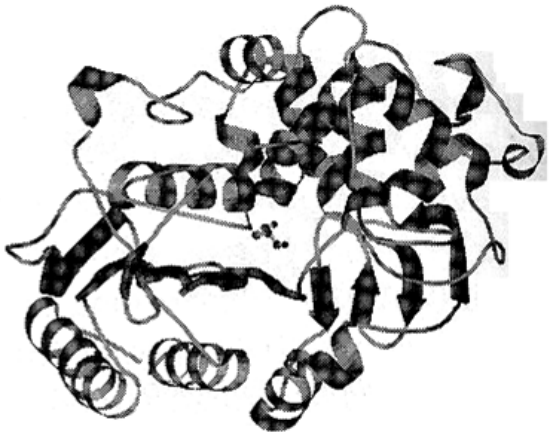
- the set of residues around the "active site"
- even when the fold change

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

- CASE STUDY / Example
- there is no gold standard

3D Motifs

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 ?

3D Motifs - Interesting Groups

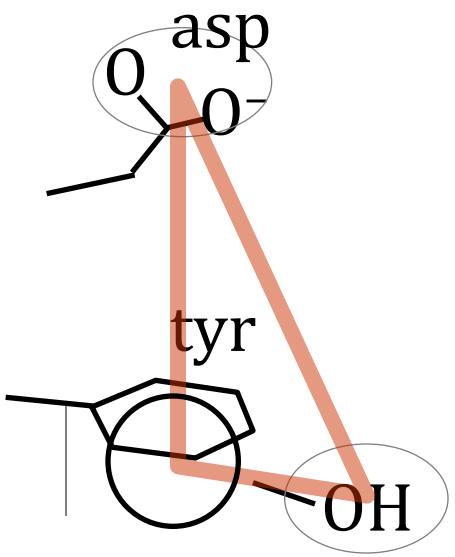
- for each amino acid, think about what is likely to be important
- slightly arbitrary
- emphasis on soluble groups (not exclusively)

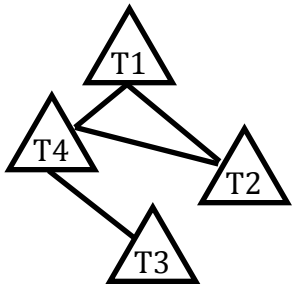
- how are relationships defined ? stored

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	

3d Motifs - relationships

- for each group
  - centre of mass of group i is c_i
- walk over protein and find all pairs with $d_{c_i c_j} < 8$ Å
- find every triangle
  - store triangle with
    - types of groups (OH, carboxyl, ..)
    - buried surface information
- connections of triangles





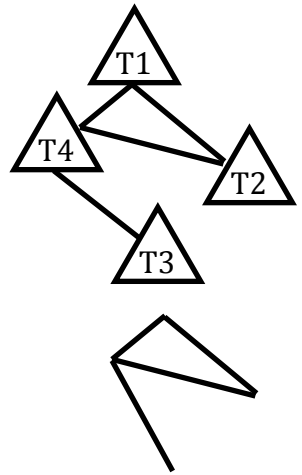
3d Motifs - relationships

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

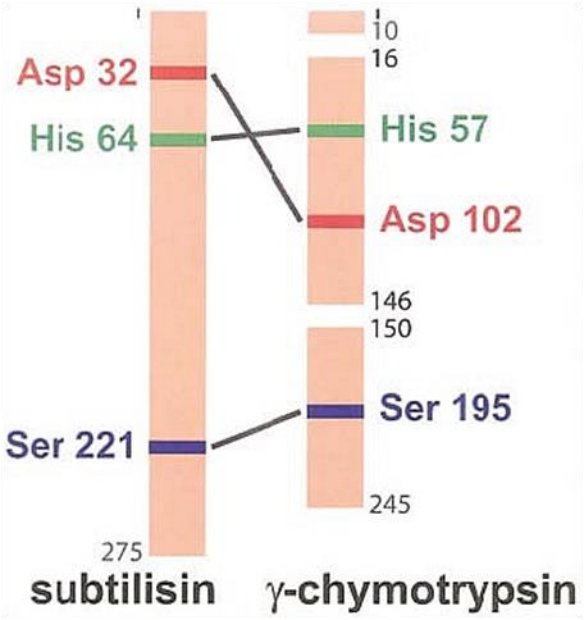
- look for common subgraphs (arrangements of groups)

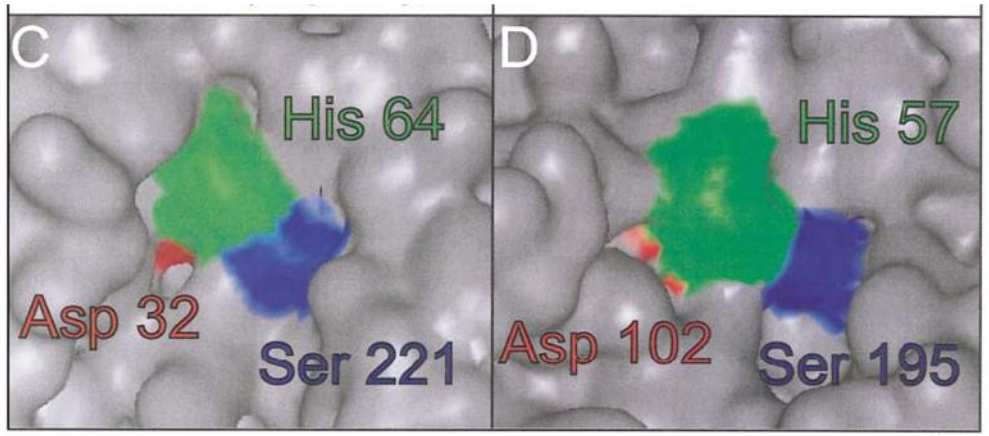
Does this work ? Examples from authors

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

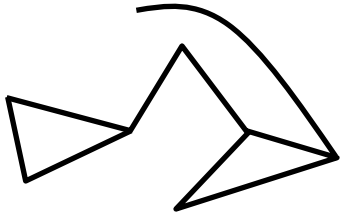
Make $C^{\alpha}C^{\alpha}$ distance matrix

- each edge is put into distance class:
  - nodes are C^{α}

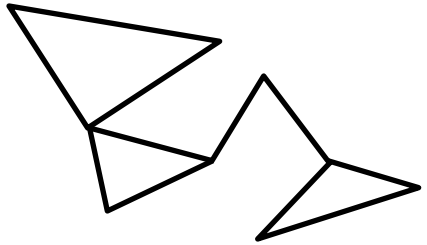
For family (typically 5 to 50 proteins)

- look for common subgraphs

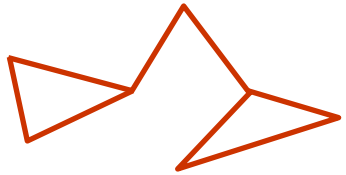
distance Å
0 - 4
4 - 6
6 - 8.5
8.5 - 10.5
10.5 -12.5
12.5 - 15



prot 1



prot 2



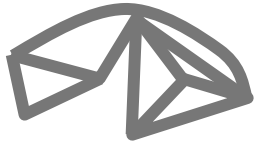
common subgraph

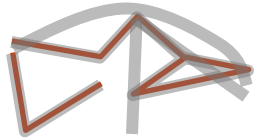
- not finished yet

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

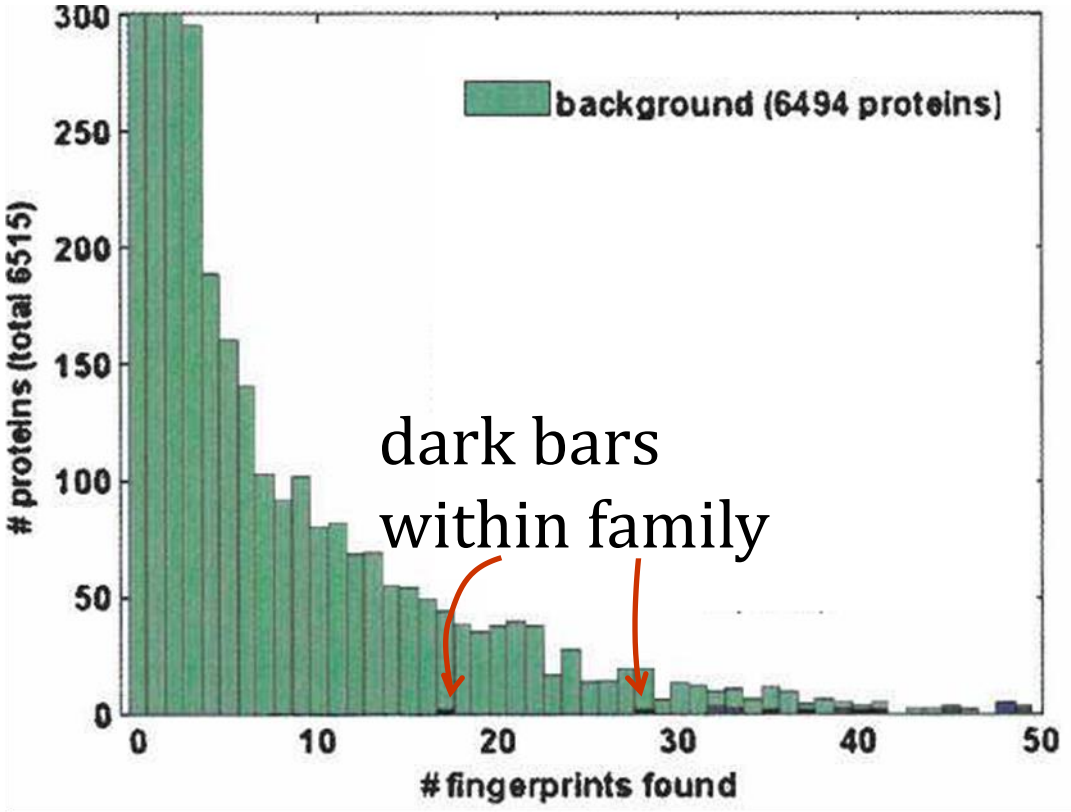
- maybe part of family

- 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