Parts 2 & 3
Fold classification &
Function from structure

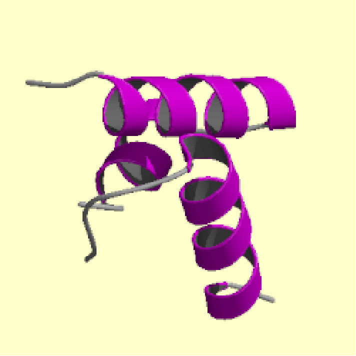
Overview

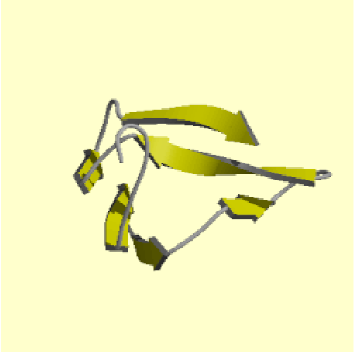
- Protein structure classification
  - Goals & Concepts
  - Methods & Databases
    - Minimum RMSD superposition
    - CATH, SCOP,...
- Function from structure
  - Function from fold
  - Active site based
    - Find a putative active site, and infer function
      - □ Intrinsic methods
      - □ Extrinsic methods

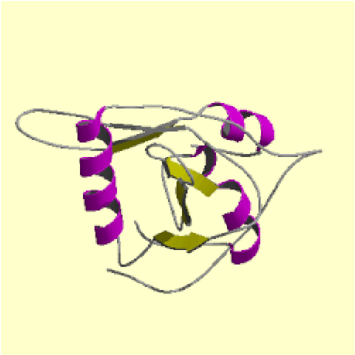
Protein fold classification

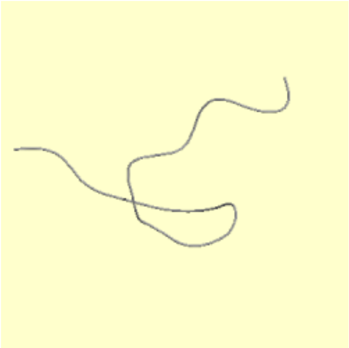
Why is this interesting?

- Understanding structure
- Evolutionary insights
- Creation of data sets
  - PDB is highly redundant!
- Function from structure
- ...







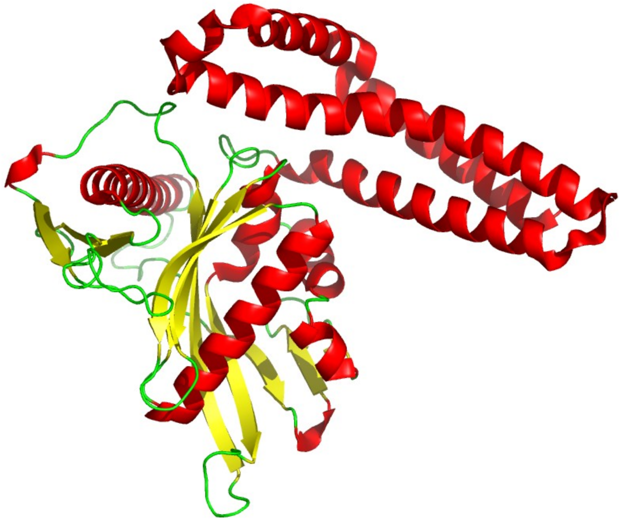


Most Proteins are Multidomain

- 40% of globular protein structures are MD
  - Most have 2 domains
- High proportion of proteins in genomes are MD
    - Ekman et al (2005), JMB, 348, 231-243
  - prokaryotes: 40%
  - eukaryotes: 65%
  - Often not easy to find domains based on sequence alone
- Fold classification is done at the domain level
  - Need a method to recognize domains

What is a domain?

- (Potentially) Independent folding unit
  - Compact, globular structure
  - More intra- than inter domain contacts
  - No shared secondary structures
  - It is an 'evolutionary unit'
- These rules are often fuzzy
- Various methods identify domains



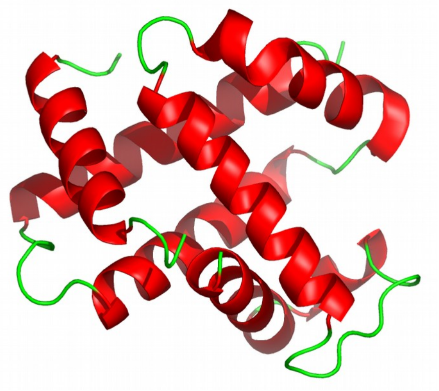
RF2

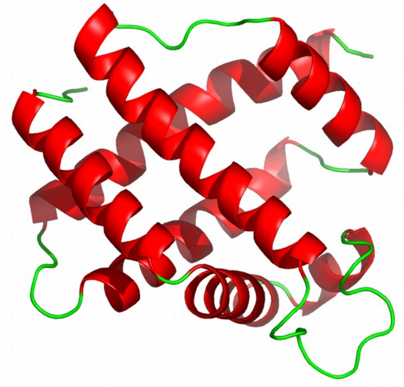
Folds and structure

- Some terminology...
  - Structure
    - A specific protein
  - Fold
    - Global properties of a structure
      - □ Secondary structure elements
      - □ Connections between elements
      - □ Orientations of elements
- Example
  - Triose Phosphate Isomerase
    - TIM barrel

Superfamilies

- Superfamily
  - Set of families
    - Not related judged by sequence
    - ...but adopt the same fold
    - ...and have a common evolutionary origin
  - Most families belong to a previously observed superfamily
  - ...and 25 % of superfamilies have a common function
  - ..so one can often go from a fold to a function for a newly solved protein structure

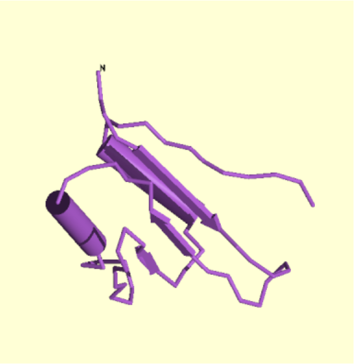




Haemoglobin &
Leghaemoglobin
(11.9% identity)

Superfolds I

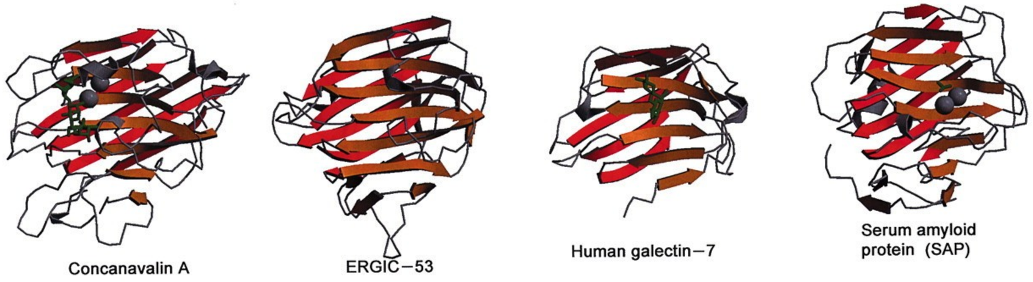
- Superfolds
  - Folds that occur in several superfamilies
  - ...due to convergent evolution (?)
- Examples
  - TIM-barrel: 15 superfamilies
  - αβ-plaits: 12 superfamilies
  - Rossmann-fold: 35 superfamilies
- Sometimes Superfold→binding site
  - TIM barrel
    - Active site is always at the top



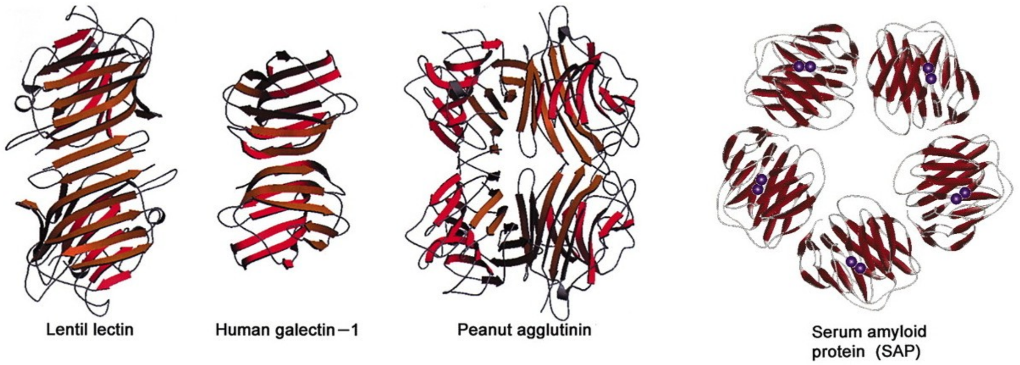
Superfolds II

- Jelly roll fold
  - Sugar binding proteins
    - Lectins
    - Loris (2002), BBA,1572

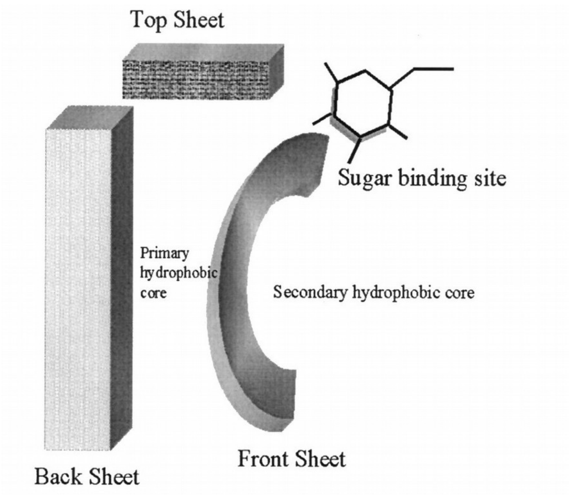
Legume lectins Galectins Pentraxins



Four families with same fold



Similar variation in quaternary structure

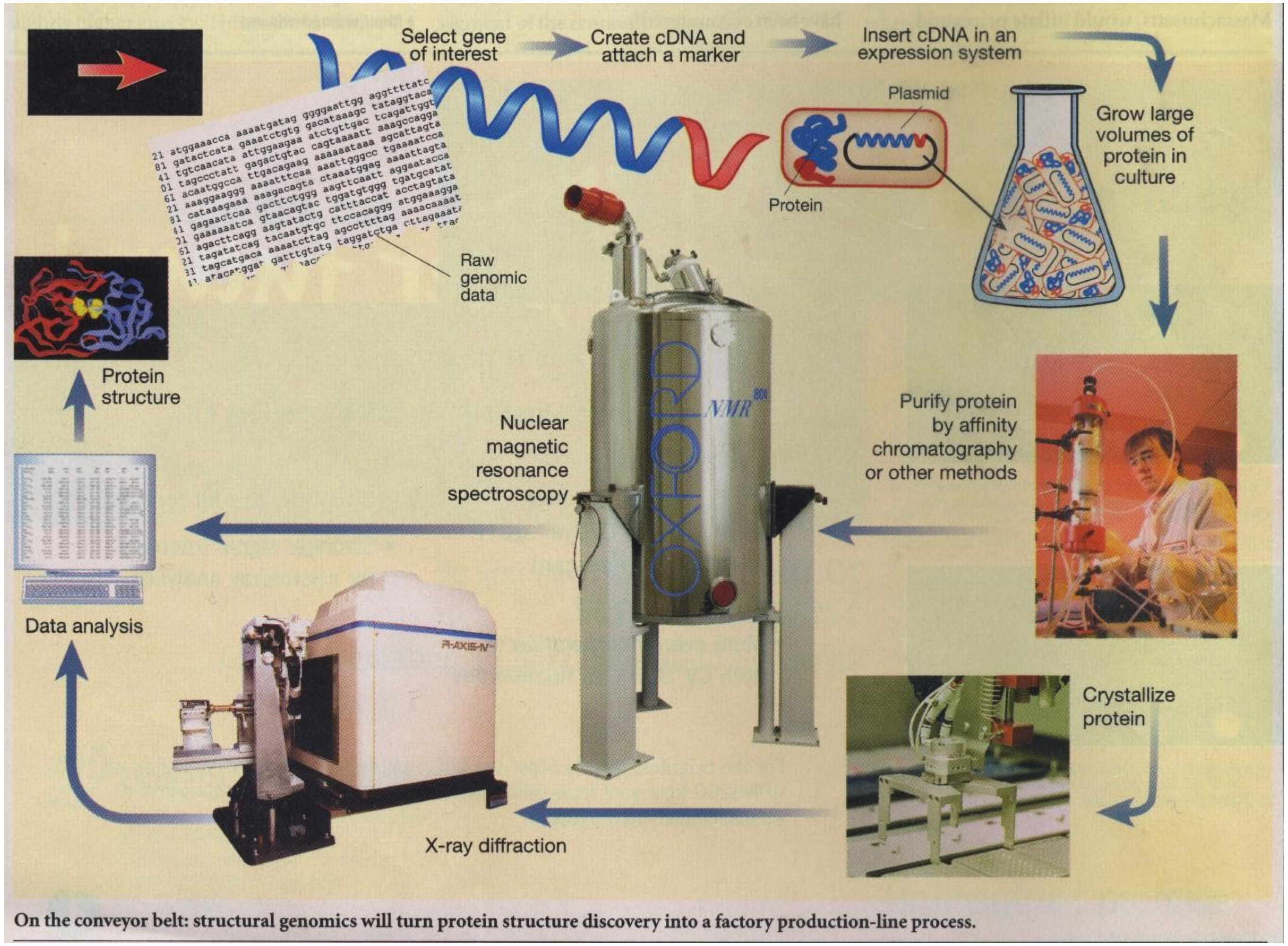


Sugar is bound in the same location

Structural genomics

- Genomics of protein structures
  - These projects can have two different aims
    - Structures for all proteins of an organism
      - Yeast, tuberculose,...
    - Structural representatives for all folds
      - Database for homology modeling
  - Some projects
    - Paris Sud Yeast Structural Genomics, France
    - *M. tuberculosis* Structural Proteomics Project, Germany
    - Center for Eukaryotic Structural Genomics, USA
      - Covers fold space

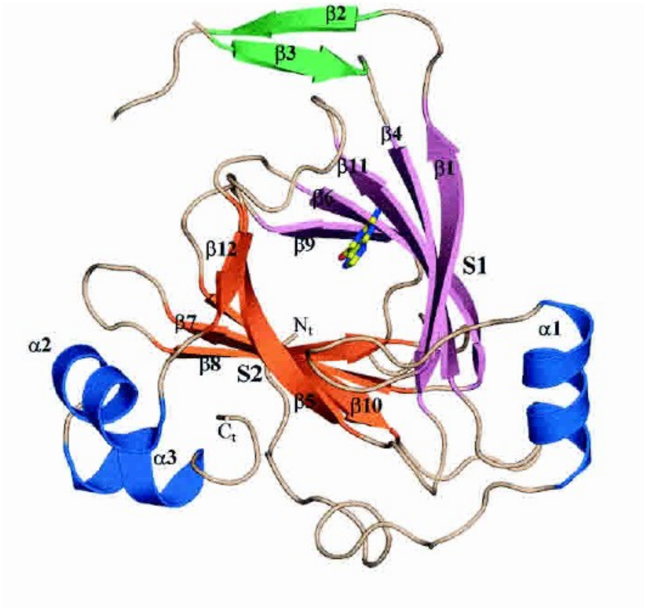
Structural genomics pipeline


Select gene of interest
Create cDNA and attach a marker
Insert cDNA in an expression system
Plasmid
Protein
Grow large volumes of protein in culture
Raw genomic data
Purify protein by affinity chromatography or other methods
Nuclear magnetic resonance spectroscopy
OXFORD
NMR
Protein structure
Data analysis
Crystallize protein
X-ray diffraction
R-AXIS-IV

On the conveyor belt: structural genomics will turn protein structure discovery into a factory production-line process.

Fold to function: YML079w

- Solved by Yeast Structural Genomics Project
  - Proteins, (2005), 14, 209-215
- Sequence did not point to a known fold
  - But YML079w adopts Jelly-roll fold
    - Cupin-superfamily
  - This fold is associated with
    - Storage in plants
      - Nucleotides
    - Bacterial enzymes
    - Lots of leads to work with!
      - YML079 binds Guanine



Measuring protein similarity

RMSD I

- A protein structure is a set of 3D vectors
- How do we measure similarity between two sets?
  - Suppose we have 2 sets of *n* 3D vectors *x* and *y*
    - Assume their centers of mass are at the origin (otherwise translate)
  - Root Mean Square Deviation (RMSD)
    - x_i and y_i are {3,1} column vectors

$$\mathrm{RMSD}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1} \left| x_i - y_i \right|^2}$$

RMSD II

- But this depends on the orientations of *x* and *y*!
  - What we really want is:

$$\mathrm{RMSD}(\boldsymbol{x},\boldsymbol{y}) = \min_U \sqrt{\frac{1}{n}\sum_{i=0}^{n-1} \left|x_i - Uy_i\right|^2}$$

  - U=Rotation matrix
  - Equivalent to minimizing E(U):

$$\mathrm{E}(U) = \sum_{i=0}^{n-1} \left|x_i - Uy_i\right|^2$$

Rotation matrix

- A square matrix U that, by multiplication, changes the direction but not the magnitude of a vector.
- 3D rotation matrices are orthogonal matrices with the following properties:
  - $U^T=U^{-1}$ and thus $UU^T=I_0$
  - det(U)=1
  - The columns AND rows form an orthonormal basis of R^3
    - vectors of length 1, mutually perpendicular
- Roto-reflection
  - If U is orthogonal and det(U)=-1, U is a roto-reflection
  - Roto-reflections are excluded in the case of proteins

RMSD III

- Let's expand E

$$E(U)=\sum_{i=0}^{n-1}\left|x_i-Uy_i\right|^2$$

$$E(U)=\sum_{i=0}^{n-1}\left(\left|x_i\right|^2+\left|y_i\right|^2\right)-2\sum_{i=0}^{n-1}x_i^t U\, y_i$$

$$E(U)=E_0-2\,L(U)$$

E_0 is independent of U

We want to maximize L

$$L(U)=\mathrm{Tr}(X^t U\, Y)$$

Where X and Y are {3,N}

matrices containing the coordinates

(Tr=trace=sum of diagonal elements)

RMSD IV

- Let's juggle a bit with the matrices in L
  - Note: trace(AB)=trace(BA) for A={m,n} and B={n,m}

$$L(U) = \mathrm{Tr}(X^t\,U\,Y)$$ Trace of {n,n} matrix

$$L(U) = \mathrm{Tr}(U\,Y\,X^t)$$ Trace of {3,3} matrix

$$L(U) = \mathrm{Tr}(UR)$$ with $R = YX^t$

- R is the {3,3} correlation matrix of X and Y

RMSD V

- Now let's write R as a product of 3 matrices

$$R = YX^t = VSW^t$$ Singular value decomposition

  - S is a {3,3} diagonal matrix, all diagonal elements>0
  - V and W are {3,3} orthogonal matrices
    - $VV^t = I_0$
    - $V^{-1} = V^t$
    - Product of two orthogonal matrices is orthogonal
    - Rows (and columns) of V form an orthonormal basis
      - Unit length, mutually perpendicular

RMSD VI

- Now let's take that result to L

$$L=\mathrm{Tr}(UR)=\mathrm{Tr}(UVSW^t)=\mathrm{Tr}(SW^tUV)=\mathrm{Tr}(ST)$$

where $T=W^tUV$

- Because S is diagonal:

$$L=\mathrm{Tr}(ST)=\sigma_1 T_{11}+\sigma_2 T_{22}+\sigma_3 T_{33}$$

RMSD VII

- Recall we want to maximize L, which minimizes E/RMSD

$$L = \mathrm{Tr}(ST) = \sigma_1 T_{11} + \sigma_2 T_{22} + \sigma_3 T_{33}$$

- Now T is orthogonal
  - Because T is a product of W^t, U and V
  - Thus, rows and columns are unit vectors
- Hence $T_{ij} \leq 1$
- As the σ's are positive, L reaches a maximum when $T_{ij}=1$
  - So T must be the identity matrix I_0

RMSD VIII

- Because T=Identity

$$T_{\max} = I_0 = W^t U_{min} V$$

$$U_{\min} = W V^t$$

$$L_{\max} = \mathrm{Tr}(S T_{\max}) = \mathrm{Tr}(S) = \sigma_1 + \sigma_2 + \sigma_3$$

- Now plug this into the RMSD expression

$$RMSD = \sqrt{\frac{1}{n}(E_0 - 2 L_{\max})} = \sqrt{\frac{1}{n}(E_0 - 2(\sigma_1 + \sigma_2 + \sigma_3))}$$

RMSD: SVD Decomposition

- A crucial step was:

$$R = YX^t = VSW^t$$

- Singular Value Decomposition theorem
  - Any real {n,m} matrix A can be written as:

$$A = VSW^t$$

  - V=orthogonal {n,n}, W^t orthogonal {m,m}
  - S=diagonal {n,m}
    - Diagonal elements are called the singular values

RMSD: Reflection catch

- Recall:

$$U_{\min} = W V^t$$

- Sometimes $U_{\min}$ is a roto-inversion!
  - Hence you will superimpose a mirror image
  - Solution:

$$U_{\min} = W Z V^t$$

$$\mathrm{RMSD}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{1}{n}(E_0 - 2(\sigma_1 + \sigma_2 + s\sigma_3))}$$

  - If $\det(WV^t)=-1$ then $Z=\mathrm{diag}(1,1,-1)$, $s=-1$
  - If $\det(WV^t)=1$ then $Z=I_0$, $s=1$

RMSD: Pseudocode

- Put Y on top of X:

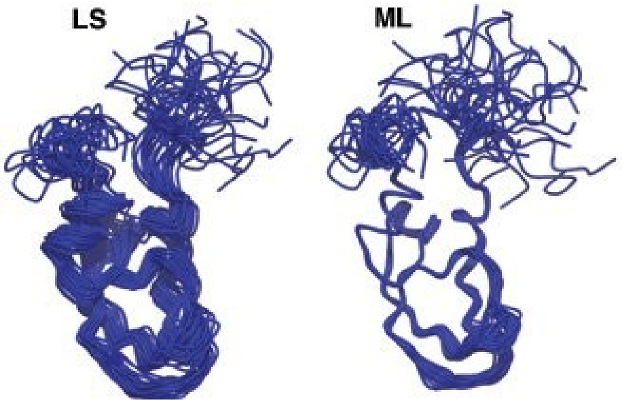
```
# X,Y are {3,N} matrices
Move X, Y to center of mass
$R=YX^T$
# Singular Value Decomposition
$V, S, W^t=SVD(R)$
$Z=diag(1,1,-1)$
$U=WV^T$
# Check for reflection
if det(U)==-1:
        $U=WZV^T$
# Rotate Y by applying U
Y_rotated=UY
# Calculate RMSD (either in real space or by formula on slide 25)
```

Theseus

- Classic LS algorithm assumes that the atom positions
  - Are uncorrelated (despite chemical bonds, errors,...)
  - Have identical variance (homoscedastic) $\Sigma = \sigma I$
  - Gaussian error model $P(U|\mathbf{x}, \mathbf{y}) \propto \prod_i \exp(-|x_i - Uy_i|^2)$
    - Equivalence with RMSD expression
- Maximum likelihood Procrustes formulation
    - Mean shape M (with K atoms)
    - Perturbation E: Matrix Gaussian
    - General covariance matrix ∑

$$X_i = R_i(M + E_i) + T_i$$

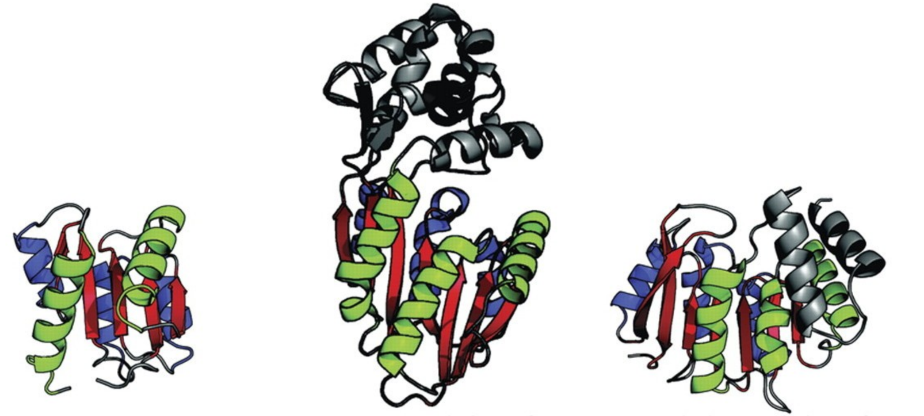
$$E_i \sim N_{K,3}(0, \Sigma, I_3)$$



Theobald & Wuttke, PNAS, 2006

Finding the equivalent positions I

- RMSD algorithm
  - Assumes we have two sets of paired vectors
    - Native/complexed structures
  - Often this is not the case
    - Insertions, deletions, missing residues, variable loops, conformational changes
  - A method is needed to find equivalent pairs!
    - Heuristic methods prevail



Chemotaxis protein Y

Histamine *N*-methyltransferase

Catechol *O*-methyltransferase

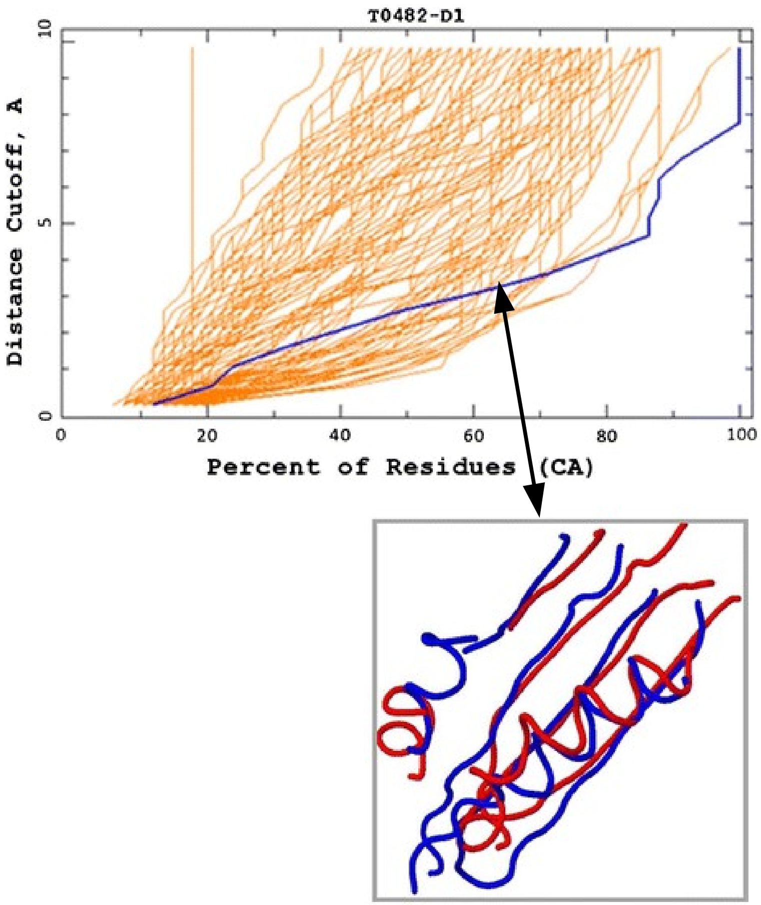
Rossmann fold

Finding the equivalent positions II

- Monte Carlo approach
  - Start with random alignment
  - Try random changes
  - Accept/reject based on the result
    - Dali, Holm & Sander (1996), Science, 273, 595-602
  - Used for fine-tuning by other methods
- Align secondary structure elements
  - Try all combinations
- Many other heuristic methods and variants exist

Global Distance Test – Total Score

- GDT_TS is more robust than RMSD
  - RMSD is very sensitive to small deviations, as for example in loops
- GDT=Percentage of Cα atoms that can be aligned to each other within a specified distance A.
  - Right: GDT plots for protein T0482 in CASP8. The aligned structures for the blue curve are shown (native in red) for 67 residues.
  - Ideal: area under curve is minimized
- GDT_TS=average of the GDT for 1, 2, 4 and 8 Å



DOI: 10.1186/s13015-015-0058-0

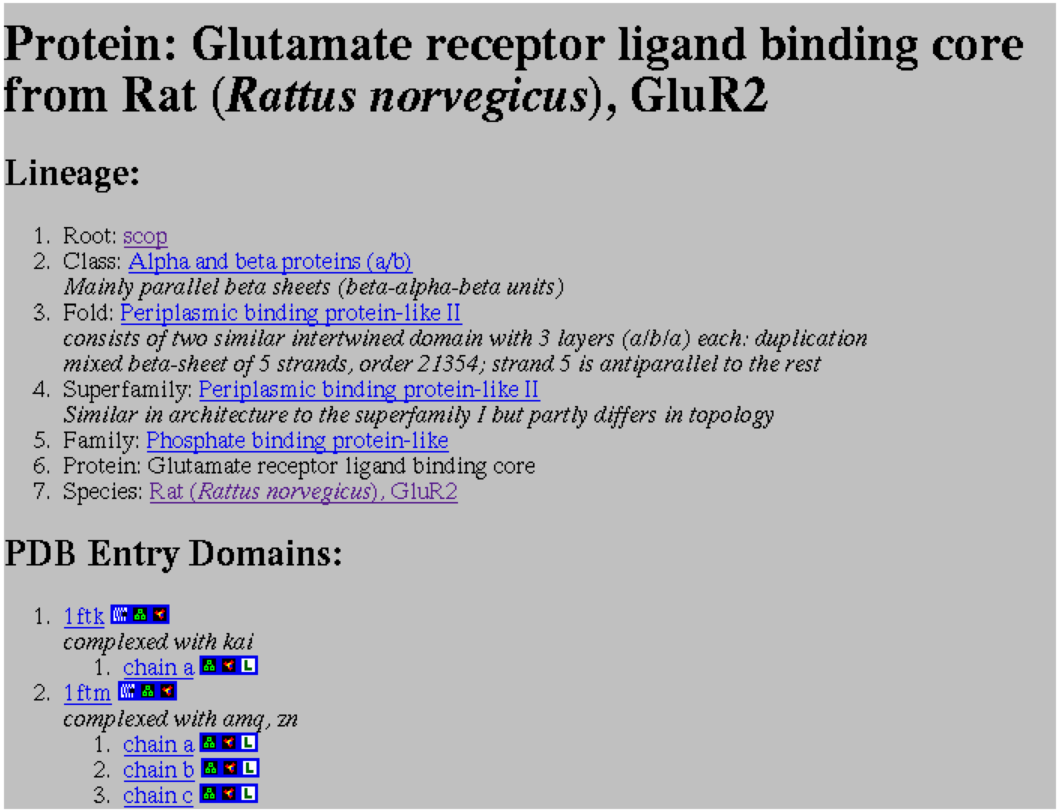
Structure Classification Databases

SCOP

- A. Murzin, Cambridge, UK
  - JMB (1995), 247, 536-540
  - Last update 2009; SCOP2 (beta) launched in 2014
- Classification
  - Class (α, β, $\alpha\beta$, irregular)
  - Fold (1195)
  - Superfamily (1962)
  - Family (3902)
- Manually constructed
  - Gold standard
  - Scalability problems, last update 2009

SCOP example

- http://scop.mrc-lmb.cam.ac.uk/scop/



Protein: Glutamate receptor ligand binding core from Rat (*Rattus norvegicus*), GluR2

Lineage:

1. Root: scop
2. Class: Alpha and beta proteins (a/b)
   Mainly parallel beta sheets (beta-alpha-beta units)
3. Fold: Periplasmic binding protein-like II
   consists of two similar intertwined domain with 3 layers (a/b/a) each: duplication
   mixed beta-sheet of 5 strands, order 21354; strand 5 is antiparallel to the rest
4. Superfamily: Periplasmic binding protein-like II
   Similar in architecture to the superfamily I but partly differs in topology
5. Family: Phosphate binding protein-like
6. Protein: Glutamate receptor ligand binding core
7. Species: Rat (*Rattus norvegicus*), GluR2

PDB Entry Domains:

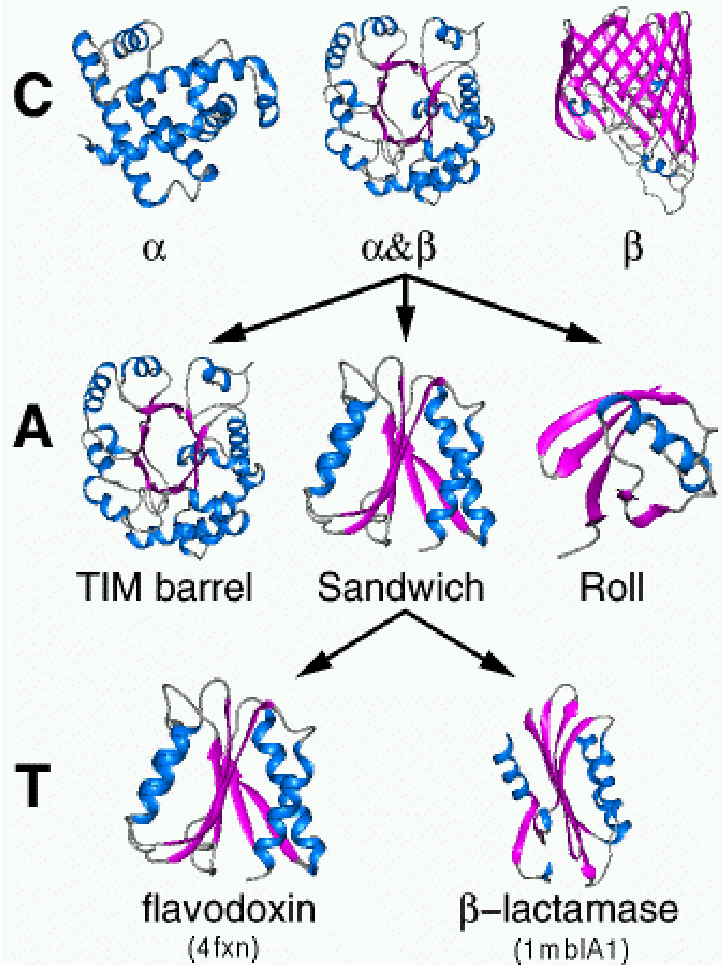
1. 1ftk
   complexed with kai
   1. chain a
2. 1ftm
   complexed with amq, zn
   1. chain a
   2. chain b
   3. chain c

CATH

- Thornton/Orengo group, UCL, UK
  - Structure (1997), 5, 1093-1108
- Class Architecture Topology Homology
- Much more automated than SCOP
  - More objective, but some 'failures'
  - Pairwise superposition
    - Still scalability problems!
- http://www.cathdb.info/

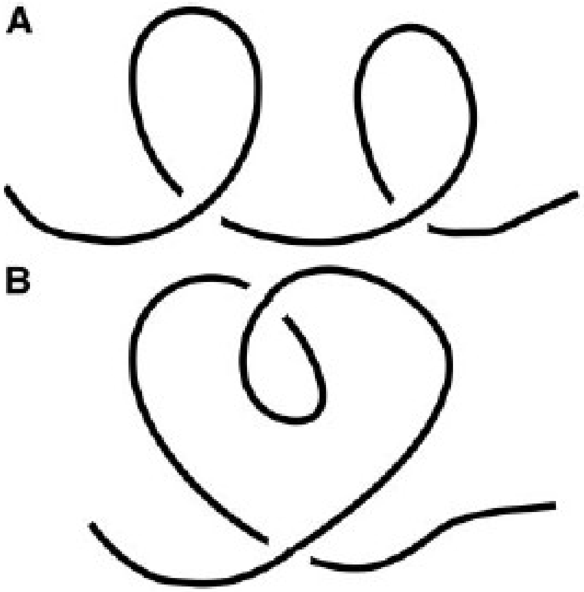
CATH classification

- Class (α, β, αβ, no SS)
    - Secondary structure
        - Statistics of July 2017
- Architecture (41)
    - Packing of sec. structures
- Topology (1391)
    - Connection of sec. structure
    - 1391=total number of folds
- Homology
    - Superfamily (6119)
    - Family, with 35% cut off (31289)



Knot theory

- Røgen & Fain, DTU/Stanford
  - PNAS (2002), 100, 119-124
- Uses generalized Gauss integrals
  - Backbone=curve in space
    - Crossing number
      - □ Average over all observer positions...
      - □ ...of the number of crossings
    - Writhe number
      - □ Uses signed crossings
  - Characterized by a 30-Dimensional vector
- Classification by clustering of vectors
- Fully automated, fast, scales well & objective



Same writhe/crossing number
Different higher order numbers

Example: writhe calculation

- C is a smooth curve, **r**$_1$ and **r**$_2$ are points on C

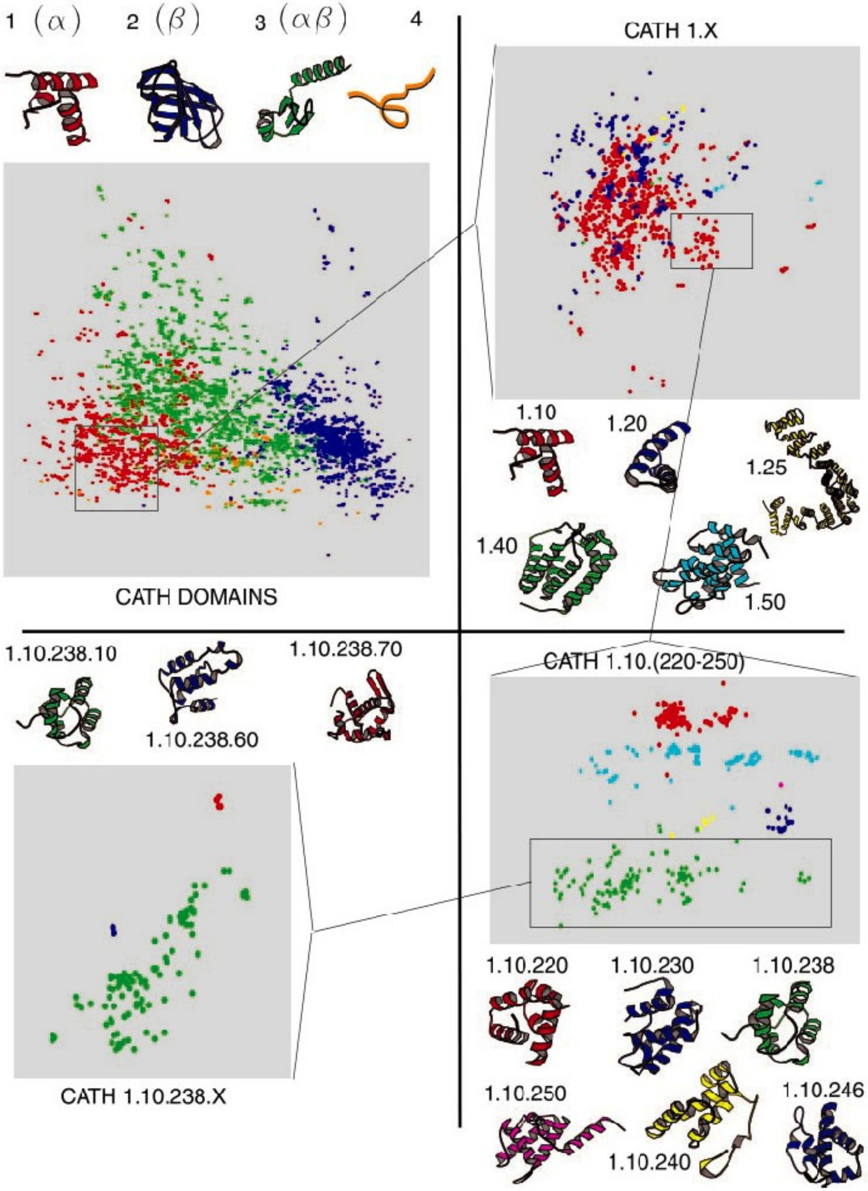
$$Wr = \frac{1}{4\pi}\int_C \int_C d\mathbf{r}_1 \times d\mathbf{r}_2 \cdot \frac{\mathbf{r}_1 - \mathbf{r}_2}{|\mathbf{r}_1 - \mathbf{r}_2|^3}$$

- Approximating C as a finite chain of N line segments

$$Wr = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\Omega_{ij}}{4\pi} = 2\sum_{i=2}^{N}\sum_{j<i}\frac{\Omega_{ij}}{4\pi}$$

$$n_1 = \frac{r_{13} \times r_{14}}{|r_{13} \times r_{14}|},\; n_2 = \frac{r_{14} \times r_{24}}{|r_{14} \times r_{24}|},\; n_3 = \frac{r_{24} \times r_{23}}{|r_{24} \times r_{23}|},\; n_4 = \frac{r_{23} \times r_{13}}{|r_{23} \times r_{13}|}$$

$$\Omega^* = \arcsin(n_1 \cdot n_2) + \arcsin(n_2 \cdot n_3) + \arcsin(n_3 \cdot n_4) + \arcsin(n_4 \cdot n_1)$$



Architectures

Classes

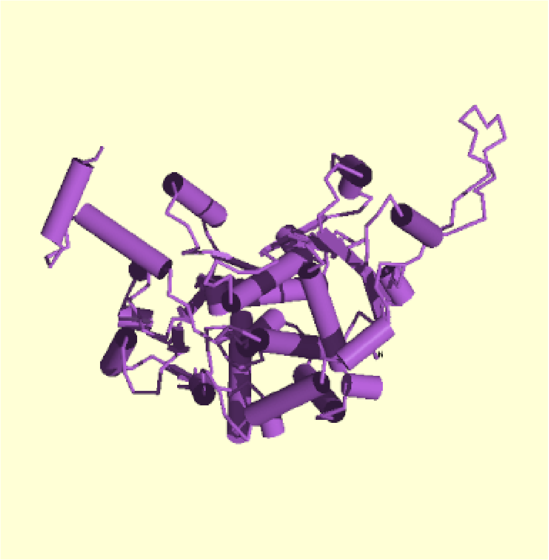
Topologies

Superfamilies

Part 3. Function from Structure

Function from structure

- Infer function by locating active sites
- Structural genomics projects
  - Structures without a story
- Uncomplexed structures
- Moonlighting proteins
  - Phosphoglucose isomerase (PGI)
    - Glycolysis
    - Maturation of B-cells
    - Nerve growth factor
    - Stimulates cell migration



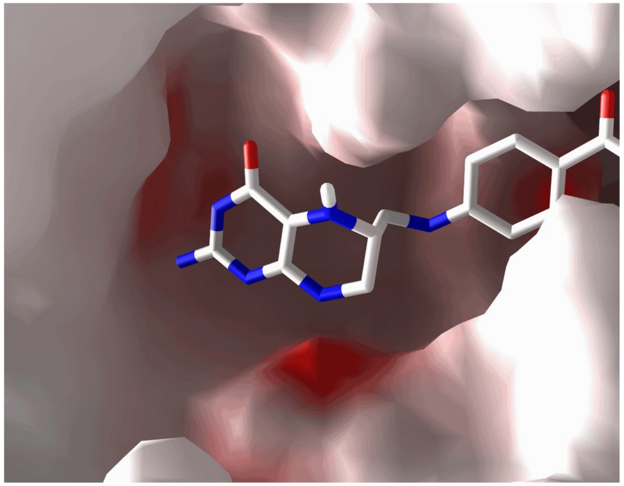
Strategies

- Intrinsic
  - Based on general properties of active/binding sites
    - Charge, shape, sequence....
  - Does not identify function itself
- Extrinsic
  - By comparison with other structures
  - Can identify function

Intrinsic methods

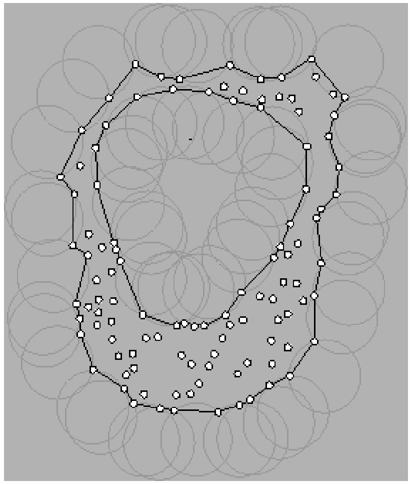
Using geometry

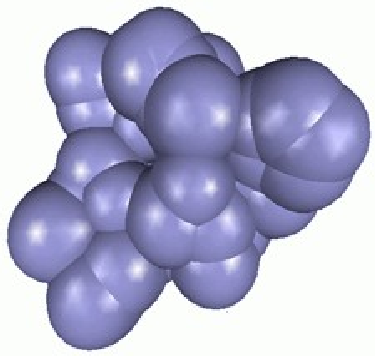
- Using surface cavities
  - Peters *et al.* (1996), JMB, 256, 201-213
- Very high efficiency
  - Calculate molecular surface
    - α-shapes
  - Identify 'cavities'
  - Select largest cavity
  - In 95 % of the cases correct



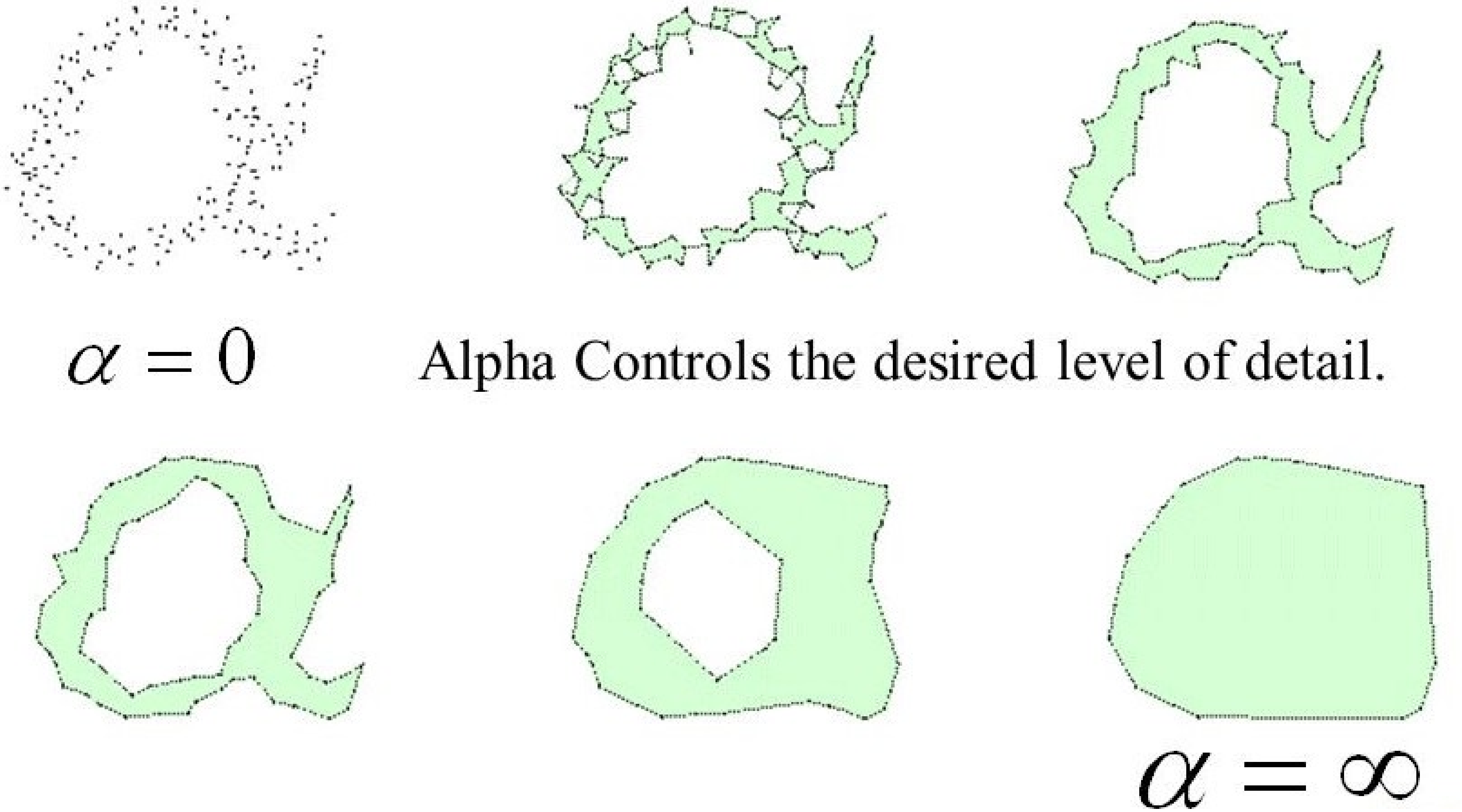
α-shapes

- Formalizes the "shape" of a point set
  - Generalization of the convex hull
  - Styrofoam-eraser analogy
  - Eraser radius α determines level of detail
- From a set of points to a volume
  - Related to the space filling model
    - CPK models and α-shapes are duals
- Finding cavities
  - Surface difference
    - $\alpha=\infty$ and $\alpha=4.5$ Å



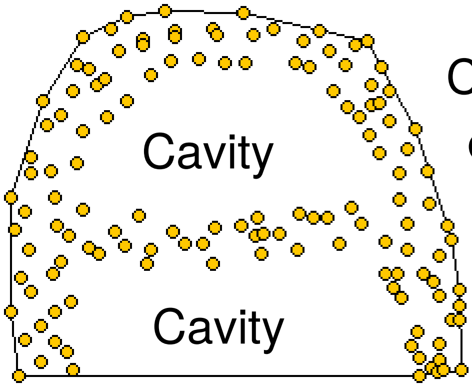


α-shape example



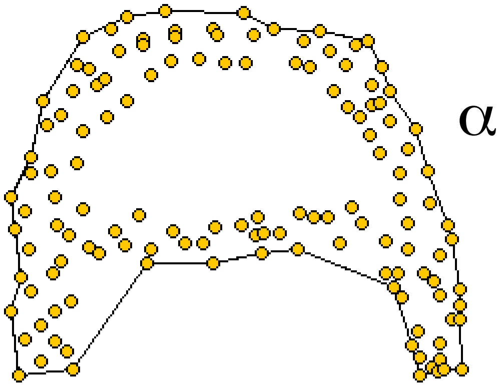
Alpha Controls the desired level of detail.

Varying α to find cavities

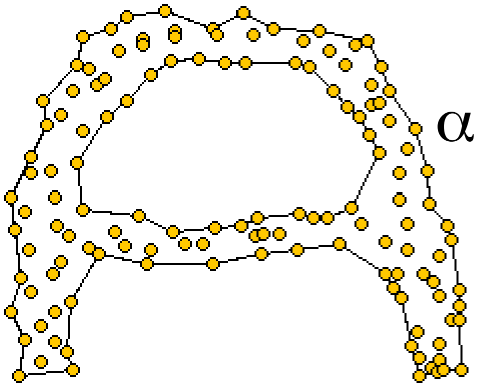


Convex hull

$\alpha=\infty$



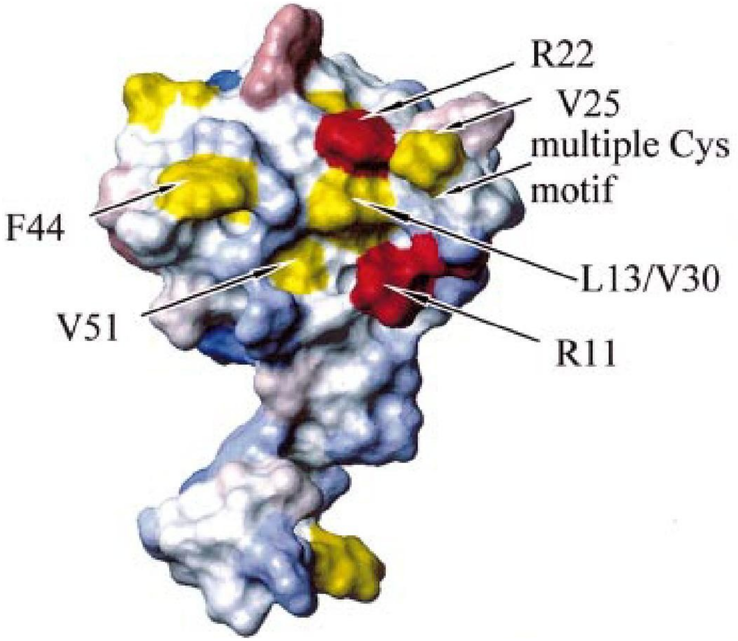
$\alpha=10.0$



$\alpha=4.5$

Using charge

- Elcock (2001), JMB, 312, 885-896
- Identify unfavorable charge concentrations
  - Needed for catalysis
- Continuum electrostatics
  - Solvent!
- Example
  - MTH1184 from structural genomics

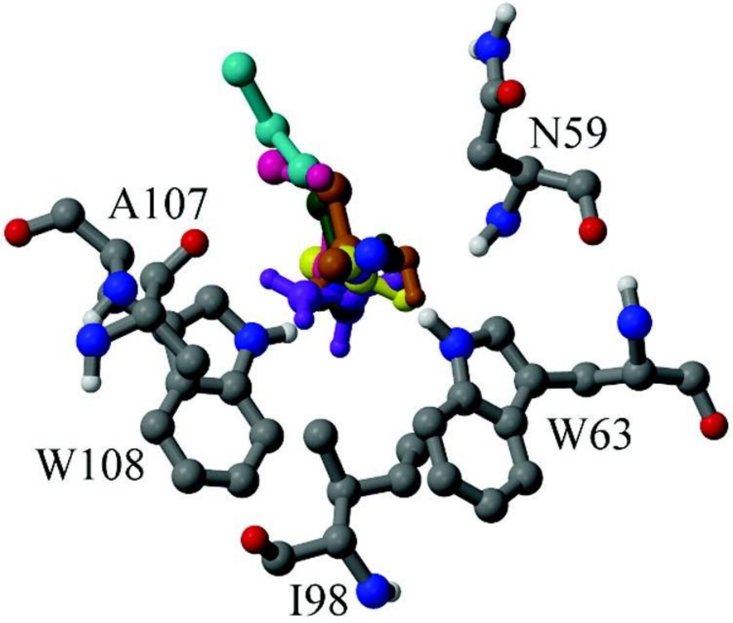


Molecular probes

- Mattos & Ringe, Nature Biot. (1996),14, 595-99
- Small molecular probes
  - Methanol, isopropanol, acetone, urea, acetonitrile, butanol, methylene chloride, DMSO...
  - These small probes often bind in similar sites
    - Determined using X-ray crystallography
  - Consensus binding sites are often active sites!
- Not very practical
  - Can this be simulated *in silico*?

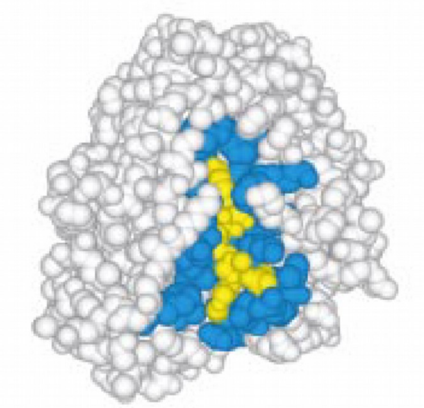
Molecular probes *in silico*

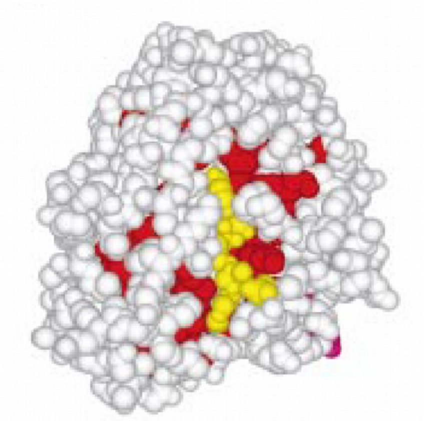
- Dennis *et al.*, PNAS, 99, 2002
- Simulate binding of molecular probes
  - *In silico* consensus binding sites
- Example
  - HEW Lysozyme



Evolutionary trace method

- Madabushi *et al.*, JMB, 316, 2002
- Active site residues are conserved
- ET-method:
  - Determine conserved residues
  - Project on a structure
  - Identify clusters
  - ET server
    - mammoth.bcm.tmc.edu/ETserver.html
- Example:
  - 2,5-diketo-D-gluconic acid reductase A
    - ligand yellow, active site residues blue, conserved residues red

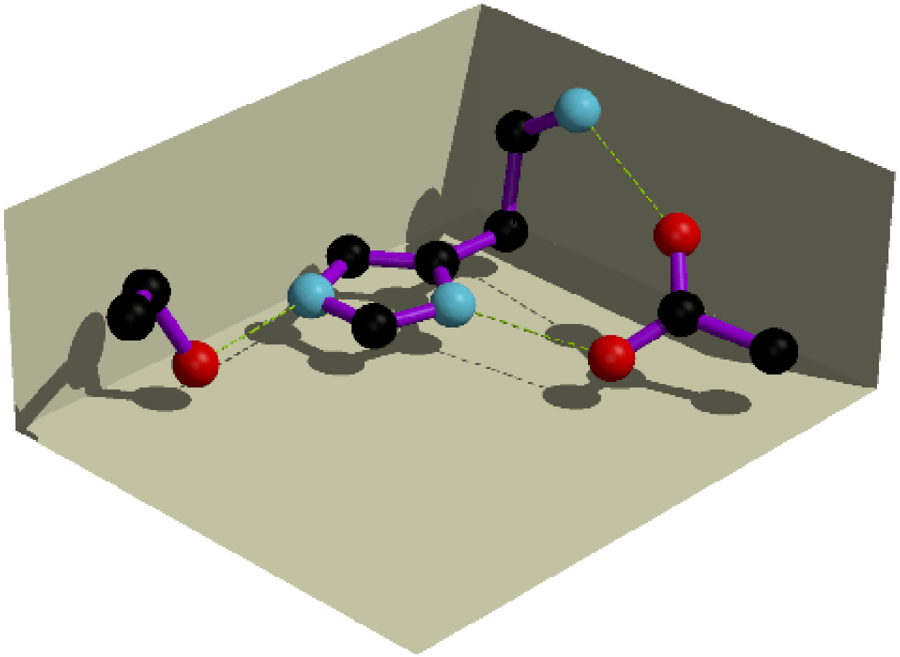




Extrinsic methods

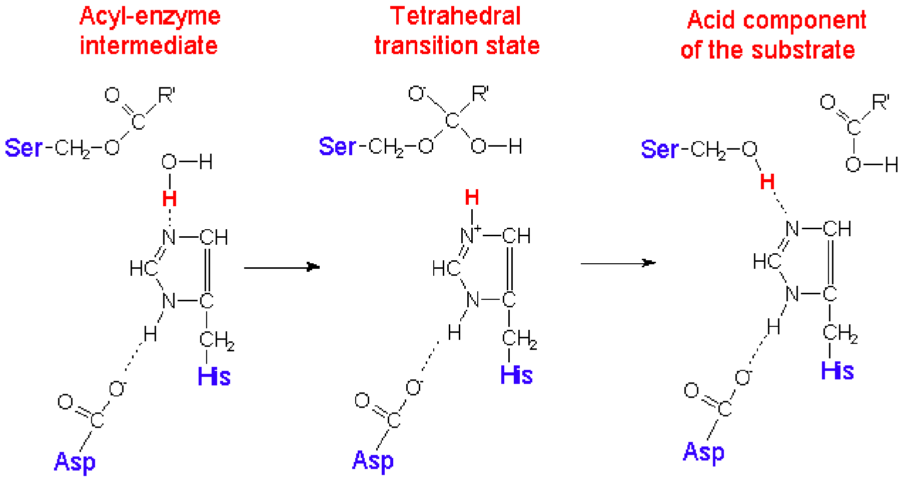
Active site similarities

- Similar active sites arise by convergent evolution
- Ser-His-Asp catalytic triad
  - Serine proteases
    - Trypsin
    - Hydrolyze proteins
  - Subtilisin
    - Hydrolyze proteins
  - α/β-hydrolases
    - Lipases



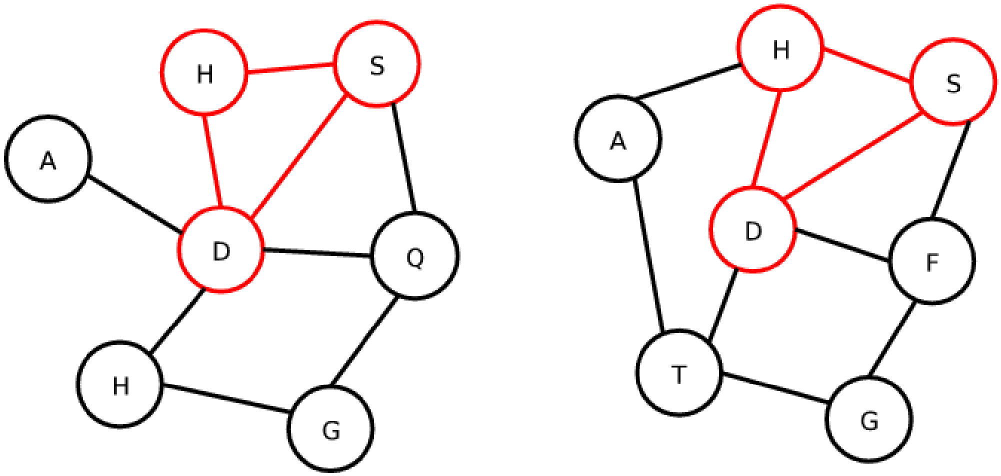
Searching for similar sites

- You can get a lot of info!
  - Position
  - Mechanism
  - Function
- This is not trivial!
  - Combinatorial explosion
    - 200 residues, 60.000+ structures
    - Catalytic site typically 2-5 residues



Graph theory

- Artymiuk et al., JMB (1994), 243, 327-344
- Present a protein as a graph
  - Nodes=residues
  - Edges=contacts



- Find similar subgraphs
  - Ullmann's subgraph isomorphism algorithm
  - Slow, pairwise comparison

Depth first search

- General idea: stop when you know the sites are different
  - Russell, JMB (1998), 279, 1211-1227
- Example: Ser-His-Asp triad

If the Ser-Asp pair in Model is different from the Ser-Asp pair in Target we can already stop here: we already know the triads are geometrically different.

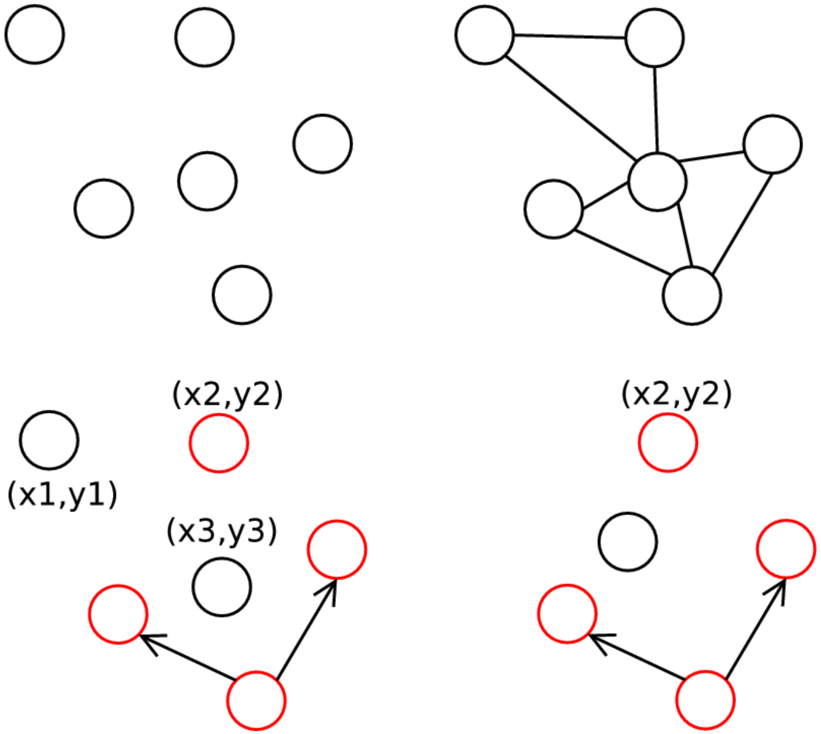
```
for Ser in Target:
    for Ser in Model:
        for Asp in Target:
            for Asp in Model:
                if Asp, Ser similar in Model and Target:
                    for His in Target:
                        for His in Model:
                            if Ser,His,Asp similar in Model and Target:
                                report(Ser, His, Asp)
```

PINTS server

- A server that offers depth first pattern search
  - Compare a protein against a database of patterns
    - Find known sites
  - Compare a protein against a set of known structures
    - Set of representatives from SCOP
    - Find new similarities
  - Compare two proteins and find similar sites
- http://www.russelllab.org/cgi-bin/tools/pints.pl

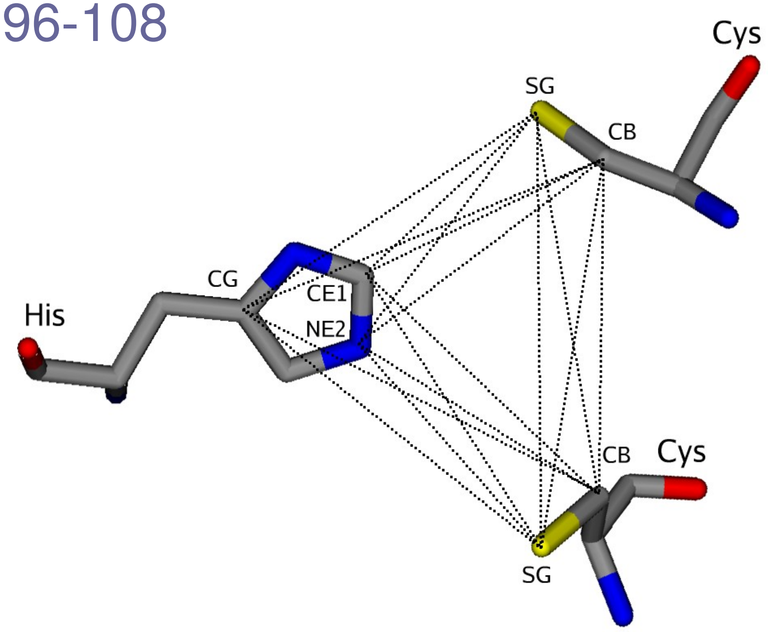
Geometric hashing

- Wallace *et al.*, 1996
- A computer vision method
- Use sets of three atoms (triplets) to create coordinate systems
- Identify cases of....
  - Similar coordinate systems
    - Find a target triplet that matches a motif triplet
    - Done by hash table look up
  - Similar sets of coordinates
    - Example: (x2, y2)



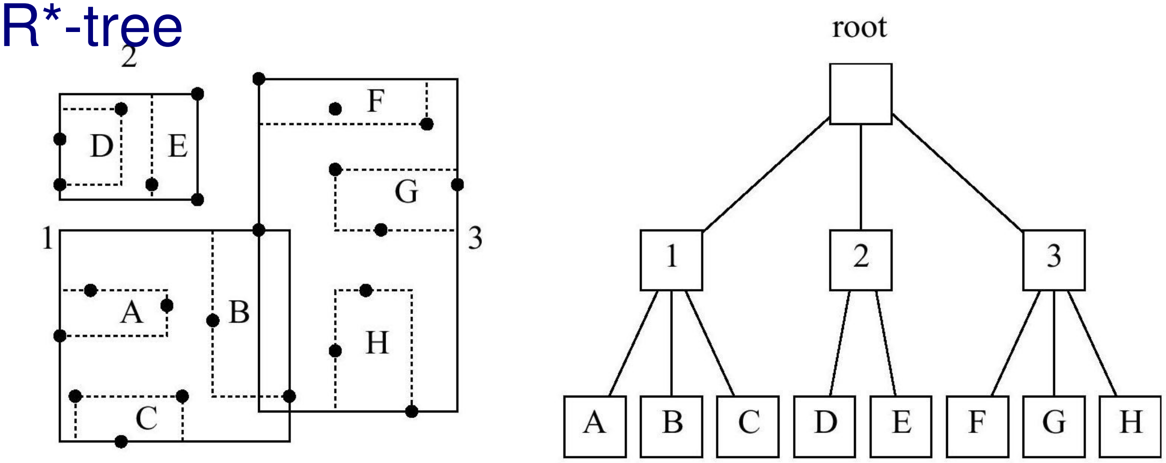
Triad method

- Look at residue triads
  - Hamelryck, 2003, Proteins, 51, 96-108
  - Close together
  - “Interesting” residues
- Represent as vector
  - Atom distances
    - $(d_1, d_2,\ldots,d_N)$
  - Mirror image insensitive!
- Finding similar vectors = finding similar triads
  - How can this be done efficiently?



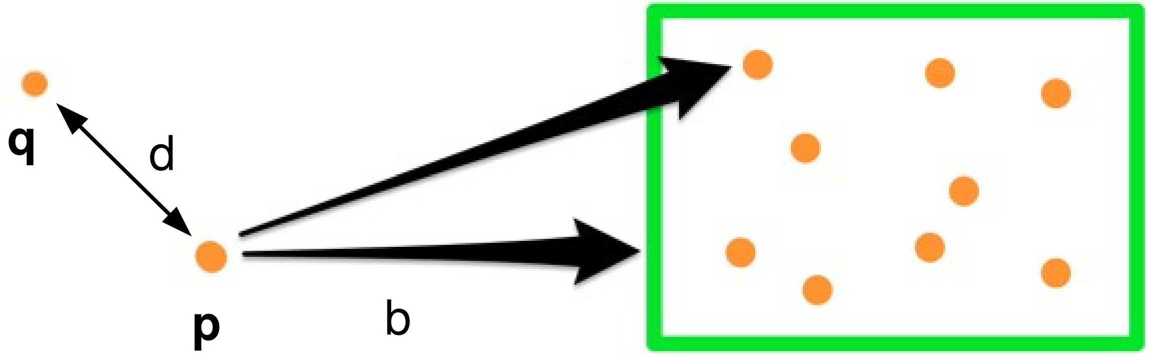
Multidimensional index trees

- High-Dim Neighbor Queries
  - Large multimedia databases
    - 20-60 Dim
    - Given a picture/movie, find similar ones
- MIT's subdivide space using nested 'volumes'
- We used an R*-tree

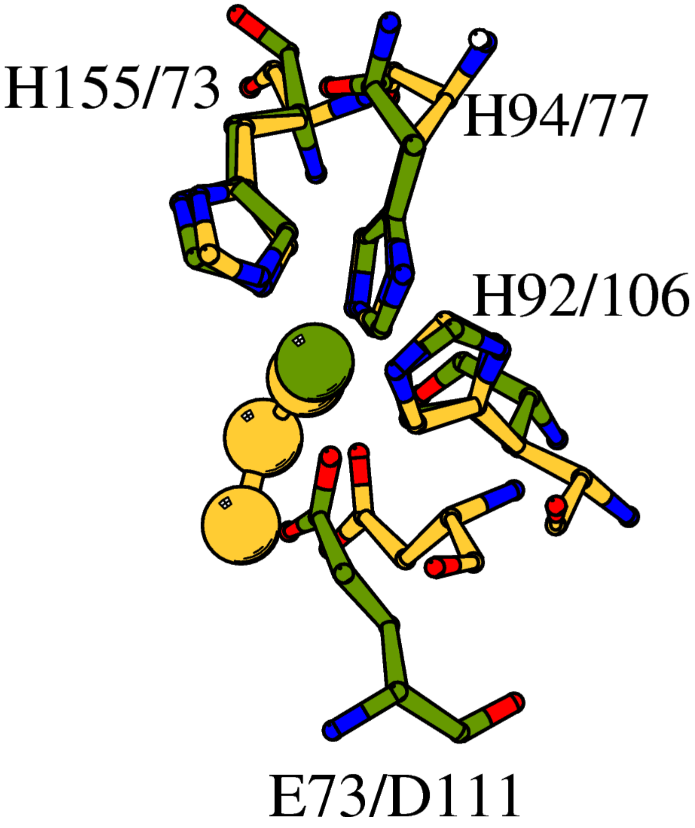


Fast nearest neighbor query

- Given a point *p*, find its nearest neighbor
  - Brute force is slow if many many points
- Rough idea: prune search using the R*-tree.
  - Once you have found a point *q* at distance *d* from *p*, you can exclude all boxes at distance *b*>*d*...
  - ...because b is a **lower bound** of distances to points in the box



Example



- L-fuculose-1-phosphate aldolase (green, mirrored), myohemerythrine (yellow)
- Zn^{2+}/Fe_2O binding sites