# Finding and Quantifying Protein Monomeric Structural Pseudo-Symmetry


It is well known that protein complexes are often symmetric, made from multiple copies of non-symmetric monomers arranged symmetrically. It is also the case that many single protein chains consist of repeating units of similar structure arranged in a symmetric manner.

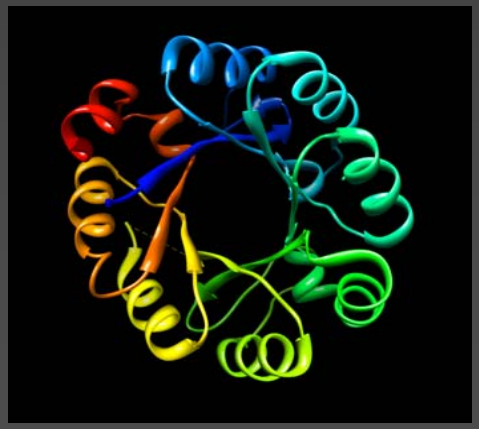
We are interested in the function and evolution of such symmetric monomers and have created an automated procedure to identify them, the type of symmetry present, and to count the number of repeats.

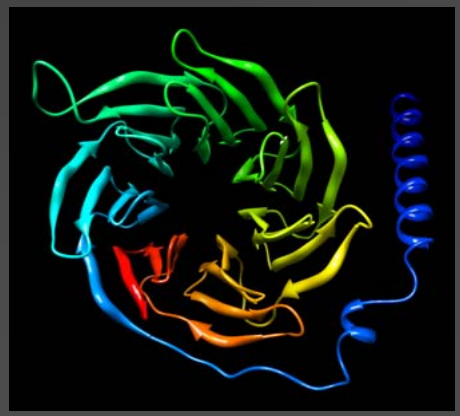
Todd J. Taylor
Molecular Modeling Section, Lab of Molecular Biology, NCI
37 Convent Dr, Bethesda MD 20814
http://binf.gmu.edu/ttaylor/
todd.taylor@nih.gov

# Introduction and Motivation

-Many protein chains are made of repeating units of similar structure arranged in a symmetric manner. Examples are “TIM” barrel structures with 8-fold rotational symmetry, beta-blade propellers (rotational symmetry), alpha-alpha super-helices (screw symmetry), and leucine-rich repeat horseshoe-shaped structures.

-The existence of symmetric structures poses a number of questions. Is there any correlation between symmetry and function? How are they different from the symmetric structures of multimeric complexes? How many symmetric chains and what types of symmetry exist in the protein universe? What is their evolutionary history?





# More Introduction and Motivation

-Internally symmetric protein domains are relatively simple structures, being sequences of relatively small repeating structural units.

-But they perform all kinds of functions: transcription factors, growth factors, enzymes, protein-protein interaction domains, scaffolds, carriers, etc.

-Are single repeating units  prototypes of elementary structures or ‘building blocks’? It may be possible to build complex, non-symmetrical structures by mixing and matching different repeating units.

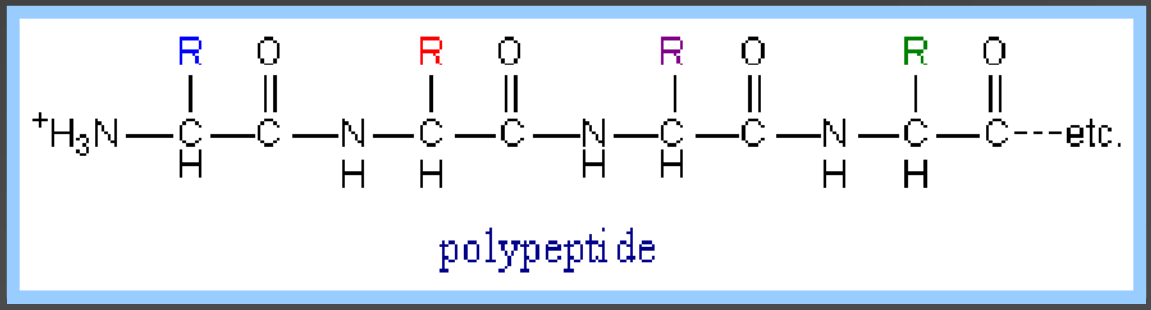
-Before we tackle questions like these, we first need to be able to identify and characterize symmetric protein monomers.

# A Quick Introduction to Proteins

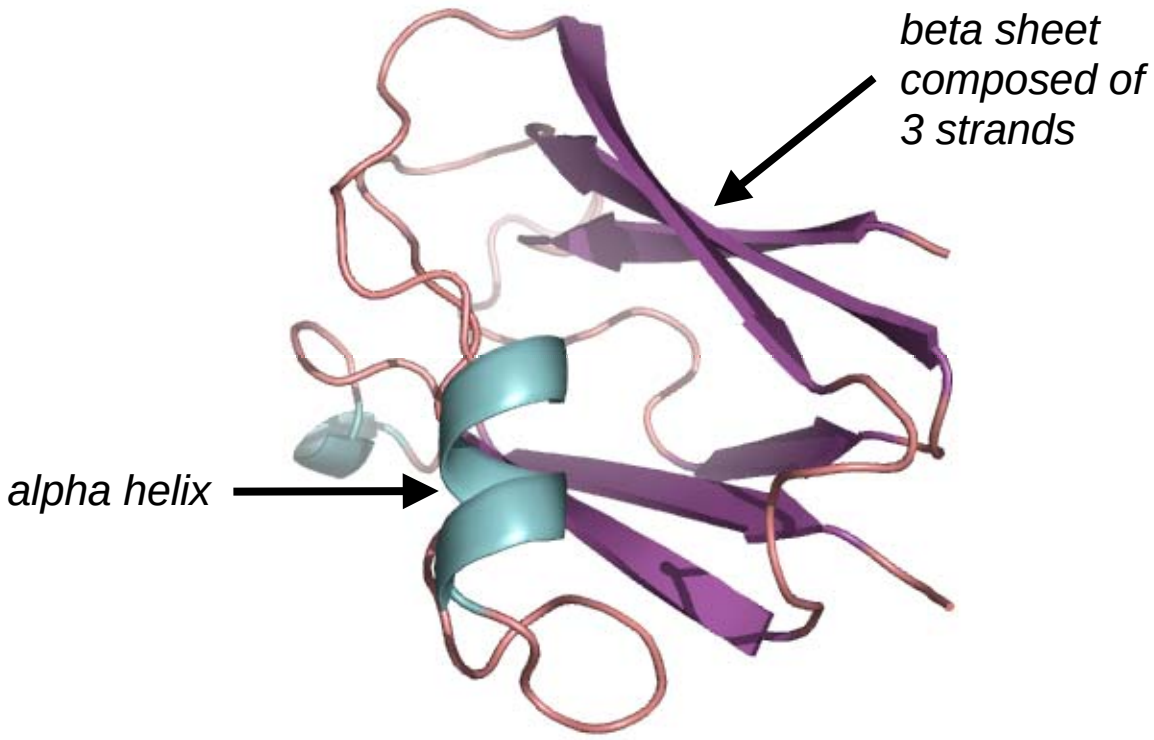
- Proteins are one of the four major classes of biological macro-molecules (proteins, lipids, nucleic acids, and carbohydrates).
- Proteins are linear polymer chains of typically ~50-1200 amino acids. They belong to the class of molecules known as polypeptides.
- There are twenty amino acid types that occur in living things.
- Amino acids are sometimes called *residues* when covalently bonded together to form a protein molecule.
- Some of the functions proteins perform include catalyzing metabolic reactions, chemical signal transduction, and forming the physical skeleton of some cellular components.
- Most types of protein molecule fold into a well defined shape under physiological conditions and this shape is uniquely determined by the amino acid sequence of the protein.
- The shape of the protein molecule facilitates its biochemical function.

# A Quick Introduction to Proteins (continued)

- Aqueous proteins fold into compact, globular shapes in order to sequester nonpolar amino acids away from the surrounding (polar) water molecules.
- Each amino acid consists of a nitrogen atom bonded to a carbon atom (called the *α-carbon* or *Cα*) which in turn is bonded to another carbon (called *C'*) and the sequence repeats N-Cα-C'- N-Cα-C'-etc to form the protein *main chain*.
- A chemical group called the *side chain* (denoted as R below) is attached to each α-carbon. Each amino acid type has a different side chain which gives that type its particular chemical properties.



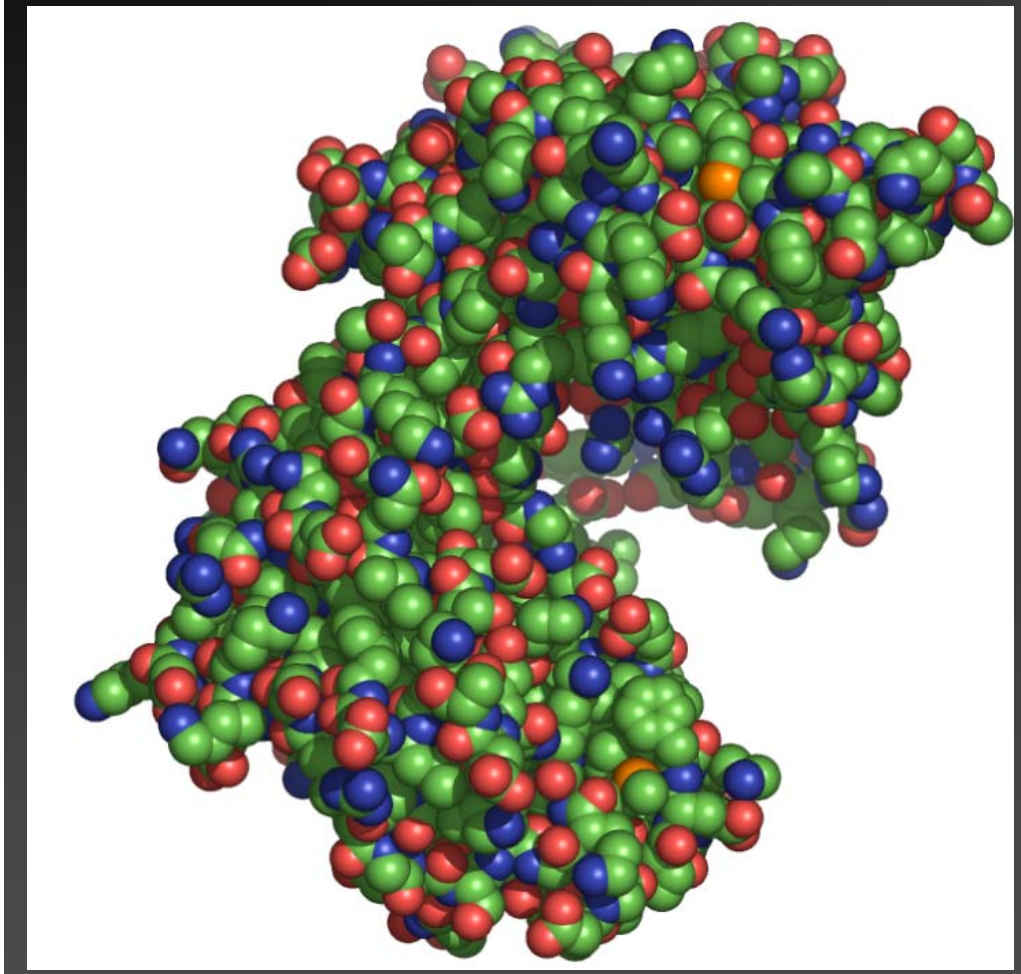
# Secondary Structure




*ribbon diagram for plastocyanin (PDB code1bawA)*

- Regular repeating patterns of hydrogen-bonded contacts between amino acids are called *secondary structure*. One such pattern is the *alpha-helix* where the chain coils into a right handed helix. A second regular pattern is the *beta sheet* where sections of relatively straight protein chain are hydrogen-bonded to each other to form a sheet.

# Protein Domains



*Phosphoglycerate Kinase (16pk) – two domains*

Many proteins can be decomposed into *structural domains*.

Domains are physically distinct regions which often also have distinct biochemical functions.

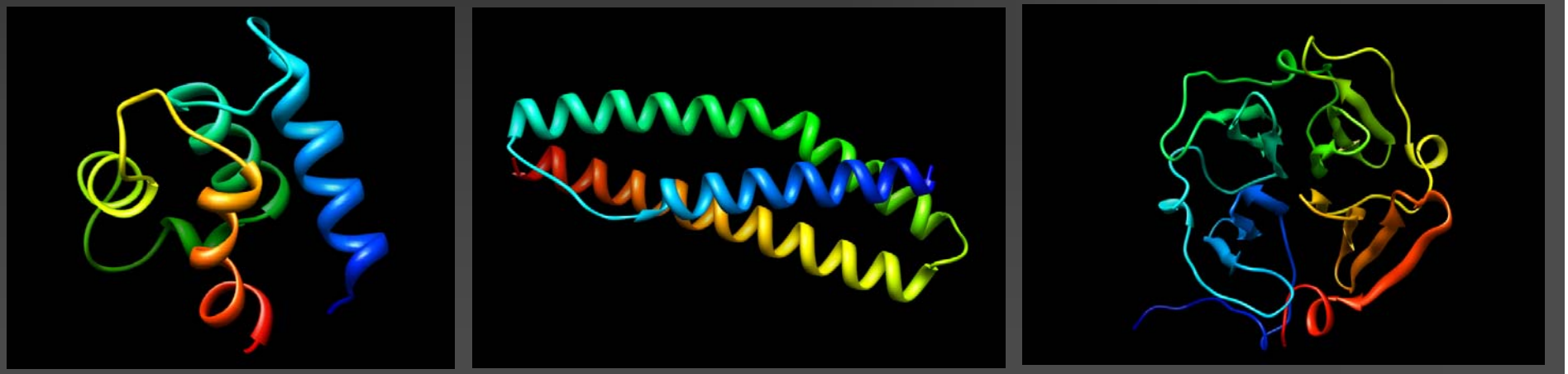
Structural domains in proteins from higher organisms are often lone protein molecules in primitive organisms. A domain can occur in several proteins with different overall biochemical functions. Proteins are modular.

Several large, comprehensive schemes for classifying proteins exist. The domain, not the complete protein, is the fundamental unit of classification in these schemes.

# Protein Folds

Bioinformaticians and structural biologists organize protein domains hierarchically in much the same way biologists organize organisms (kingdom, phylum, class, etc.). Several such hierarchical schemes exist.

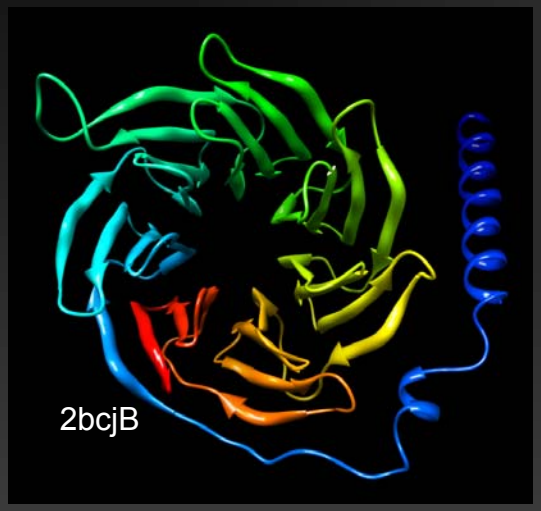
The *fold* is the second highest level in the SCOP hierarchy of protein structure classification. Domains of similar architecture (the same secondary structure elements with the same topology), but not necessarily detectable sequence similarity and evolutionary relatedness, are grouped into a single fold. Function can differ considerably among the members of a fold.



A few SCOP folds (left to right) EF-hand like,  spectrin repeat-like, 4-bladed beta propeller

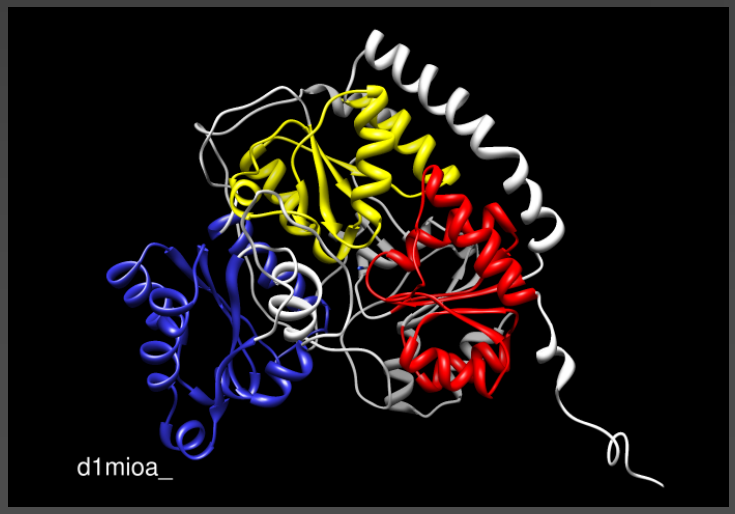
# Molecular Symmetry

## Internally symmetric monomer




## Symmetric oligomer

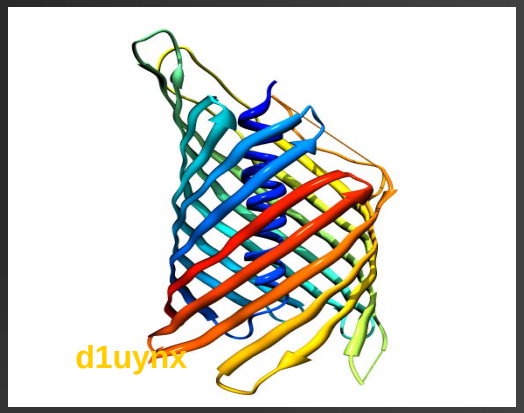

## Repeat-containing, non-symmetric monomer



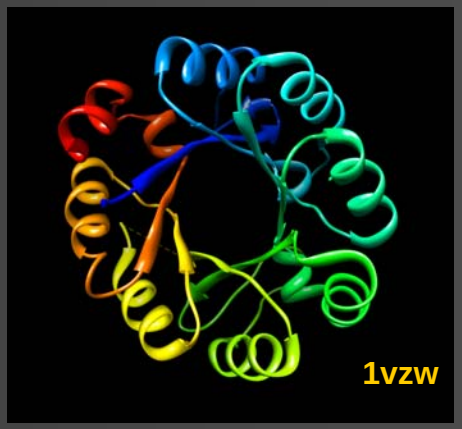
# Symmetry in Single Chain Protein Domains

β-trefoil (FGF)

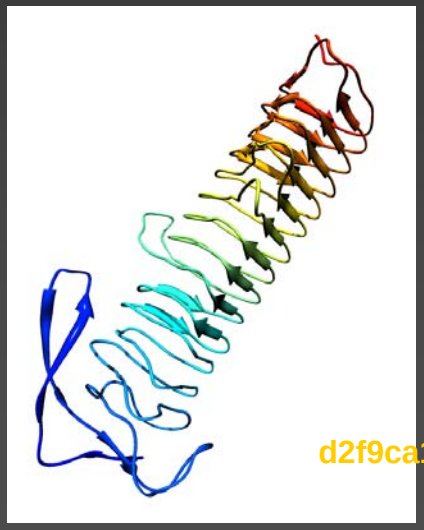

Transmembrane β-barrel



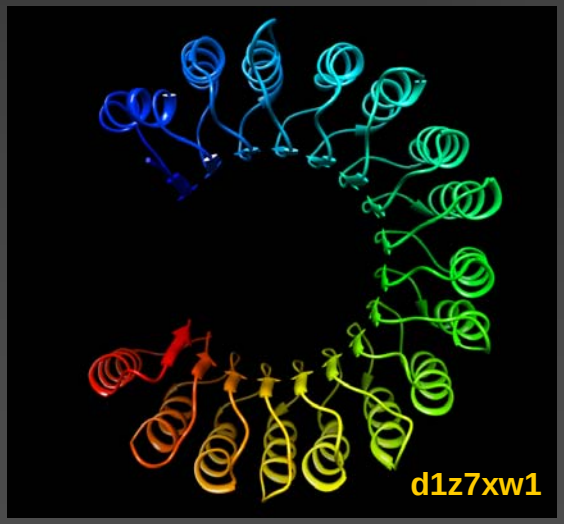

TIM barrel



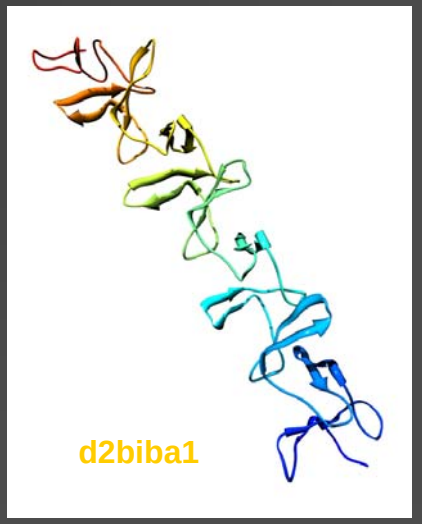

β-helix




Leucine-rich repeat horseshoe




β-hairpin stack



# The SymD Program

-The SymD program detects protein monomeric symmetry.

-SymD assigns an initial correspondence between a protein of length N and a copy of the protein circularly permuted by n residues.

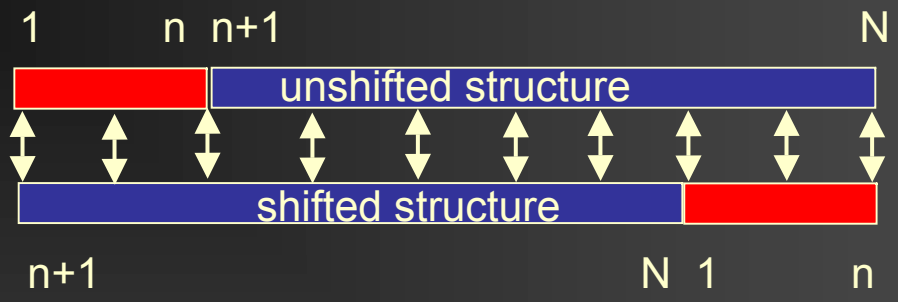
-This initial alignment is refined, using the SE heuristic (described later) to give a gapped alignment.

-The optimal rigid body superposition, in a least squares sense, of the aligned residues from this gapped alignment is calculated using the procedure of Kabsch, which gives a corresponding transformation matrix and rotation axis.

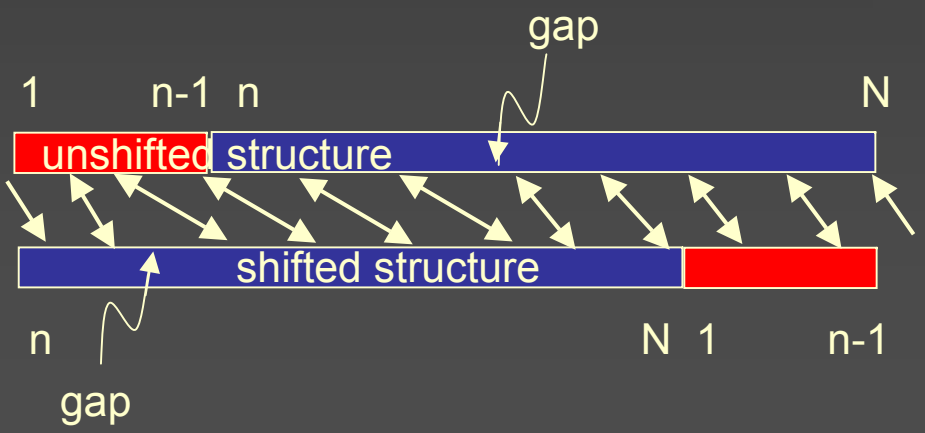
-This procedure is repeated for all shifts n with N-3> |n| >3, calculating new gapped alignments and transformations, and that non-self transformation that transforms the structure so that it is most similar to the original is chosen as the best transformation.

# The SymD program

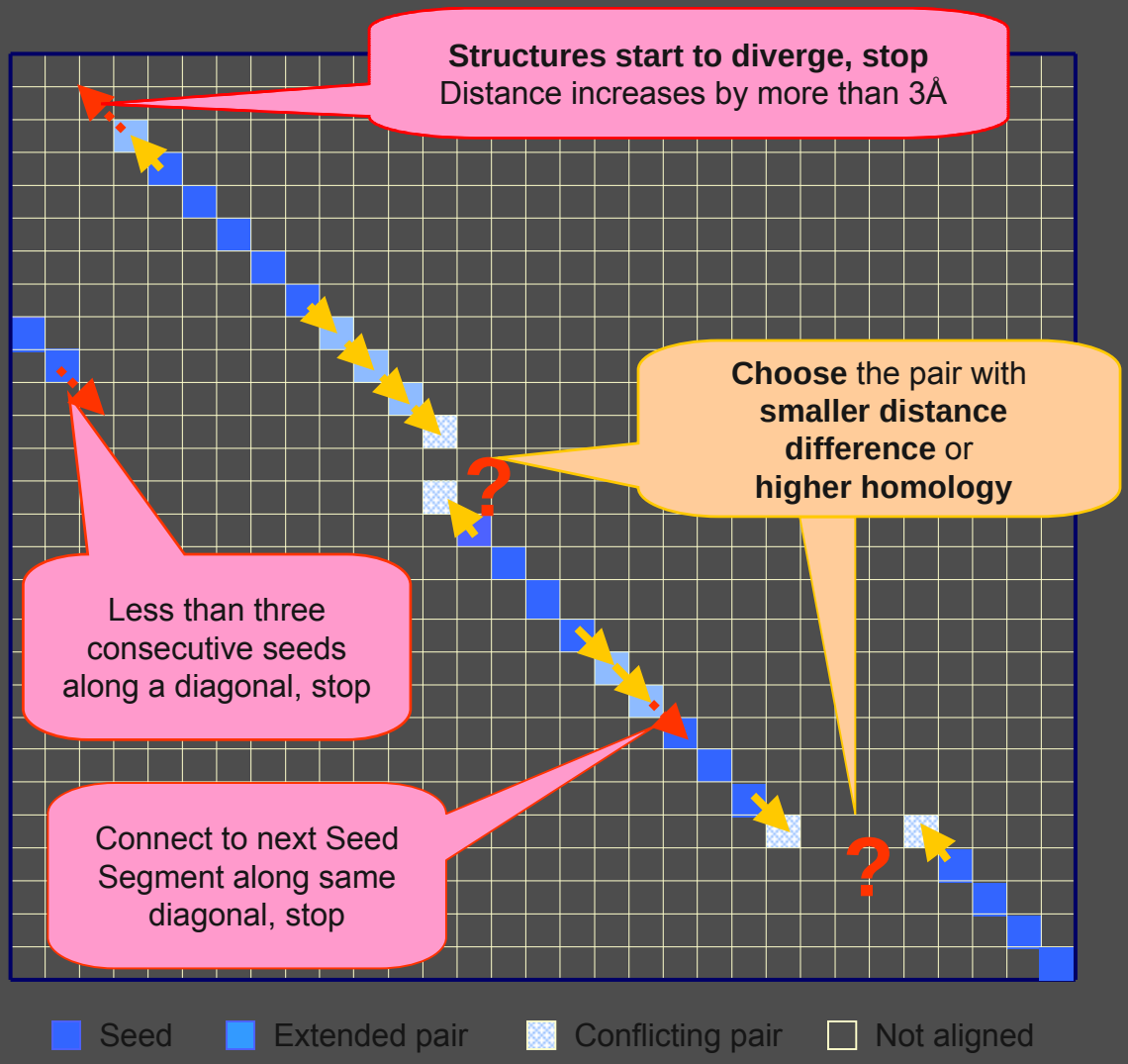
Initial alignment given by circular permutation of offset n, also called the *initial shift*.
Every residue in the unshifted structure aligns to some other residue in shifted.




SymD successively applies the SE heuristic:
-The Kabsch procedure is applied to the aligned pairs from the initial shift to get an optimal superposition.
-Residue pairs that superpose well form seeds that are extended. The extended seeds are joined together to form a new alignment that includes gaps.
-Only the subset of aligned residue pairs are fed to the Kabsch procedure to get a new superposition.



# SE (Seed Extension) Algorithm



# The Template Modeling Score

-The similarity of the protein monomer and the transformed copy of itself is measured using the TM-score (Zhang and Skolnick).

-TM was originally designed to measure the similarity between a real, experimentally determined protein structure, aligned with a structure prediction

-TM requires an alignment–a correspondence between residues in the compared structures. SymD provides one.

$$TM = \frac{1}{N_{res}} \sum_{i}^{N_m} \frac{1}{1 + \left(d_i / d_0\right)^2} \qquad d_0 = 1.24 \cdot \sqrt[3]{N_{res} - 15} - 1.8$$

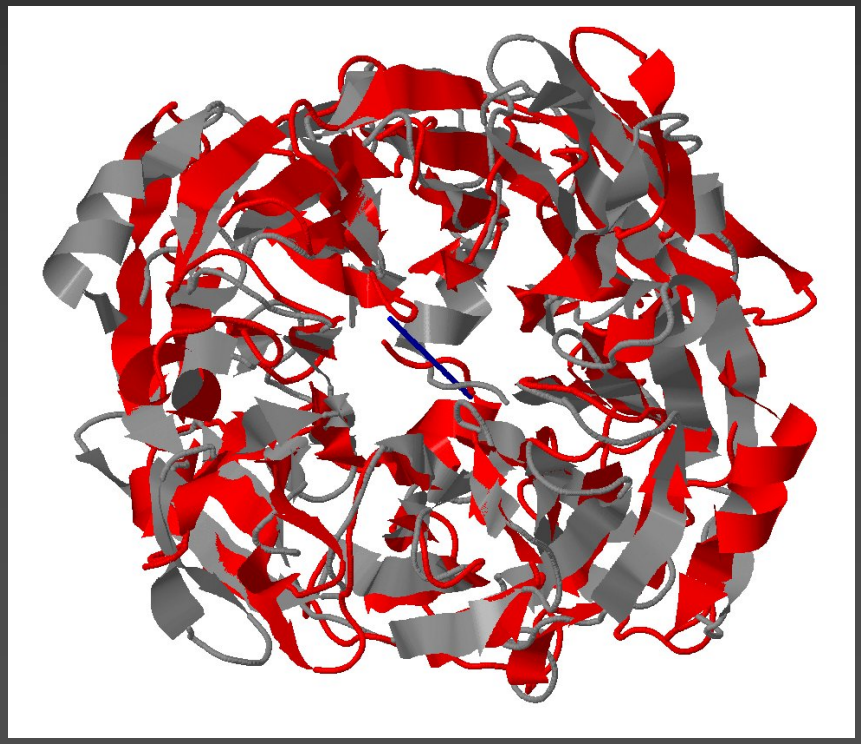
*The sum is taken over aligned pairs. $N_{res}$ is the number of residues in the protein. The distance between the $i^{th}$ pair of aligned residues is $d_i$*

# SymD Output

-SymD aligns pairs of residues, one from the untransformed structure and one from the transformed.

-The residue number differences between the members of these pairs, untransformed minus transformed, are called shifts.

-SymD also return a structural superposition with its corresponding rotation axis and translation.



Aligned Pairs

| unshifted | shifted | shift |
|---|---|---|
| 654 K | 557 T | -97 |
| 655 S | 558 S | -97 |
| 656 D | 559 L | -97 |
| 657 W | 560 G | -97 |
| 658 L | 561 L | -97 |
| 471 Q | 564 D | 93 |
| 472 D | 565 V | 93 |
| 473 I | 566 Q | 93 |
| 474 V | 567 R | 93 |
| 475 F | 568 V | 93 |

# Virtues of the SymD Approach

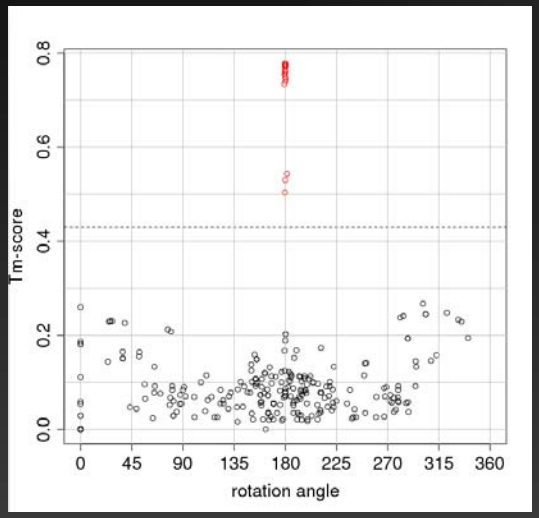
- Because it refines the structural alignment to include gaps, SymD can deal with imperfect structural repeats in which there are insertions and deletions.

- Since the alignment procedure is carried out many times for many different *initial shifts* (number of residues by which the copy is circularly permuted), a great many alignments and corresponding transformations and symmetry axes are produced.

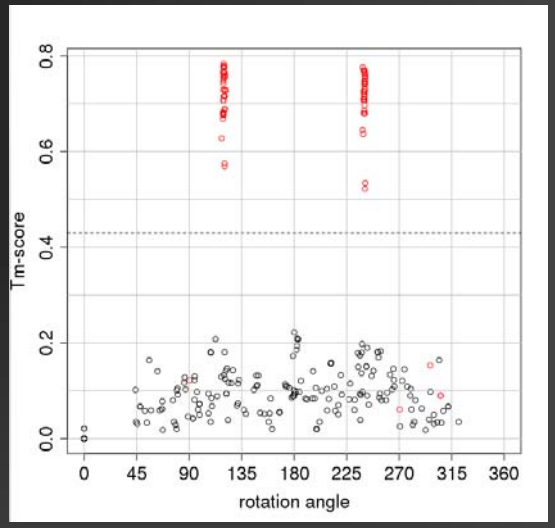
-When an initial shift puts the structural repeats of the copy somewhat out of register with the original, the SymD procedure tends to relax the transformed structure so that they come into register. Therefore many different initial shifts will generally converge to the same correct symmetry axis, although perhaps with different rotation angles (e.g. 90, 180, and 270 for a 4-fold rotationally symmetric domain).

# TM-scores from Alignment Scans of Sample Domains

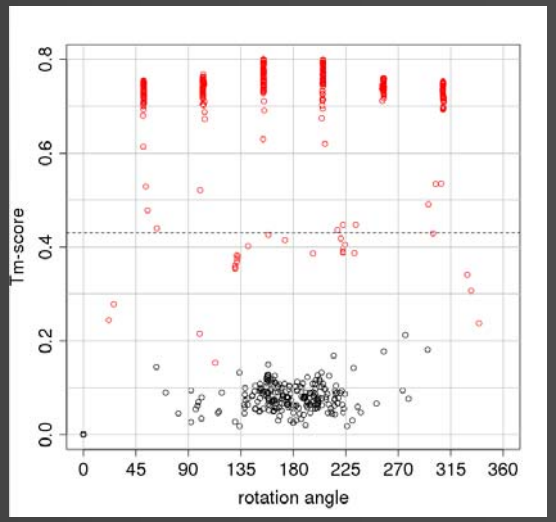
d1s99a_ (ferredoxin)



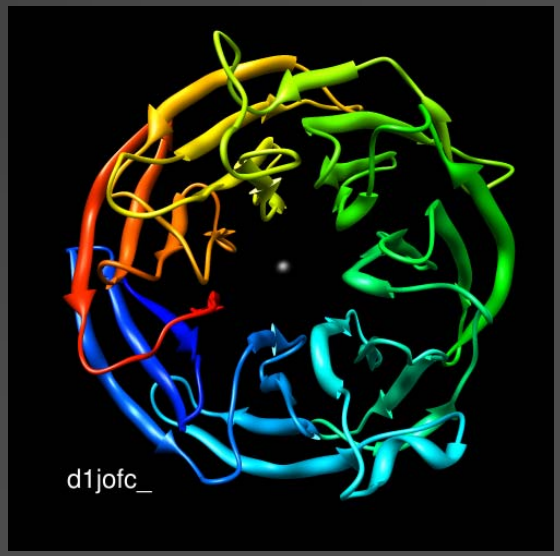

d1wd3a2 (β-trefoil)



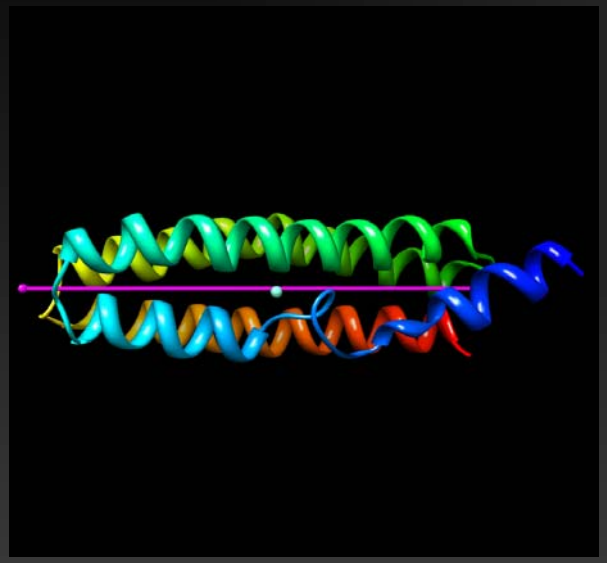

d1jofc_ (β-propeller)

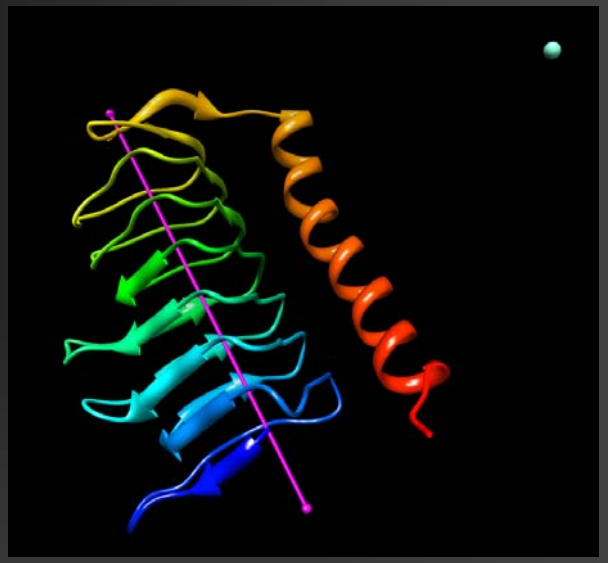


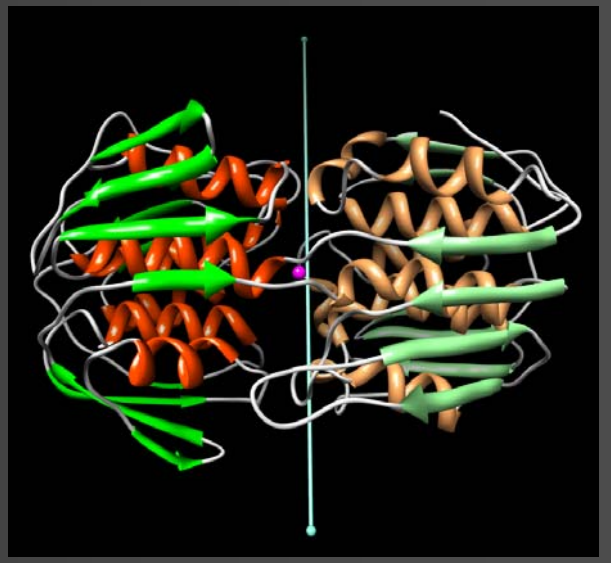


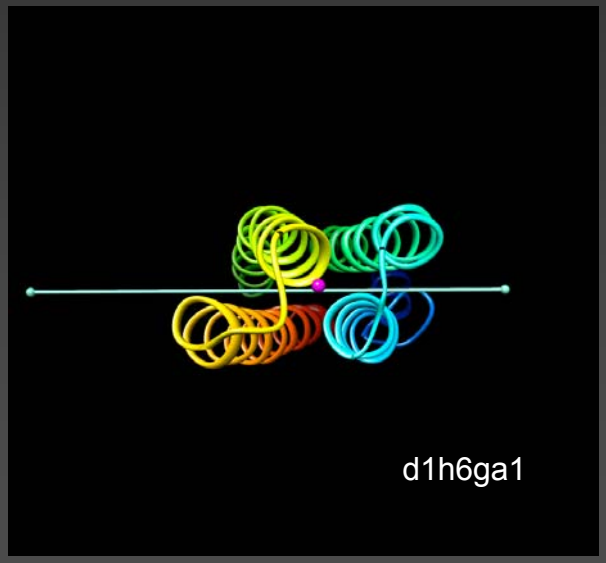



# Structures with More than One Symmetry Element

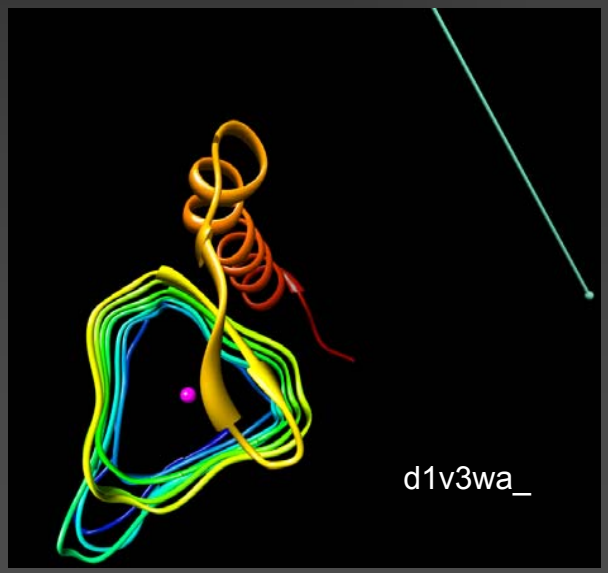




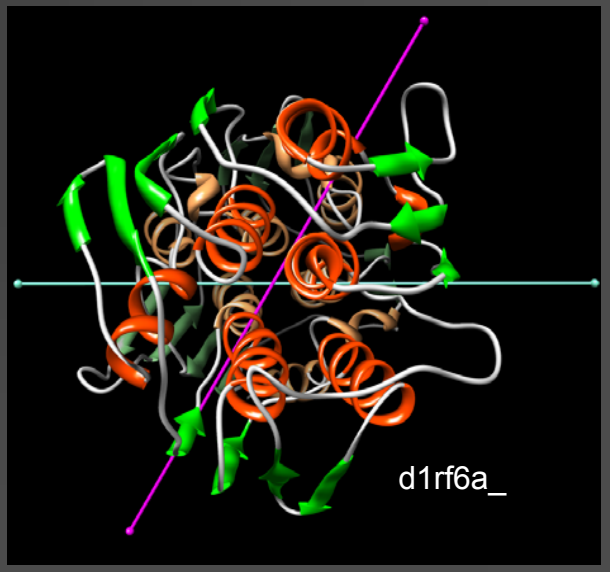





α-catenin M-fragment

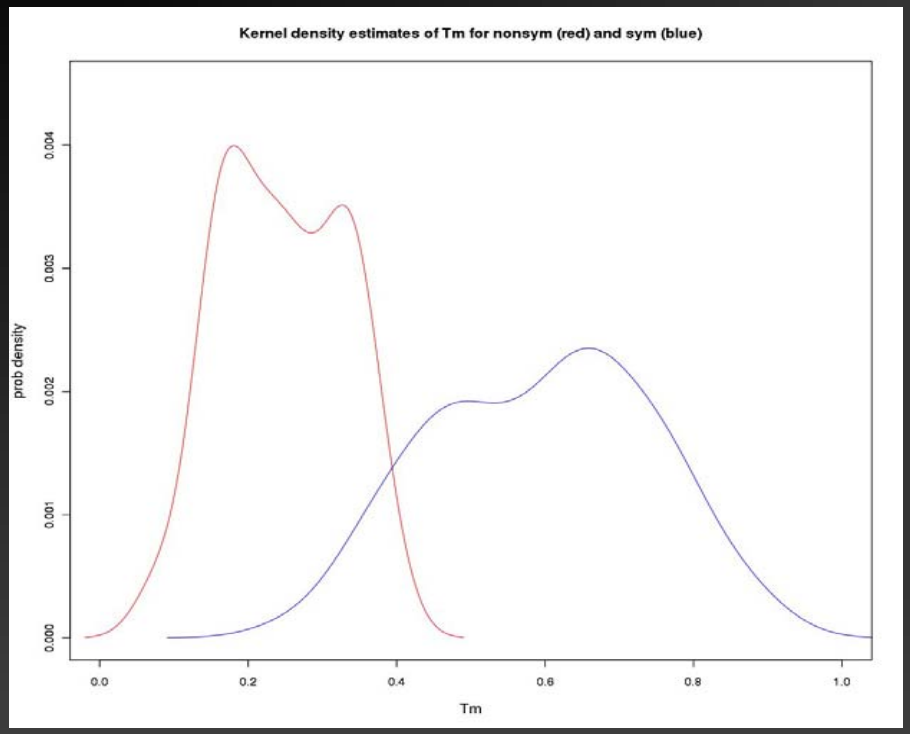


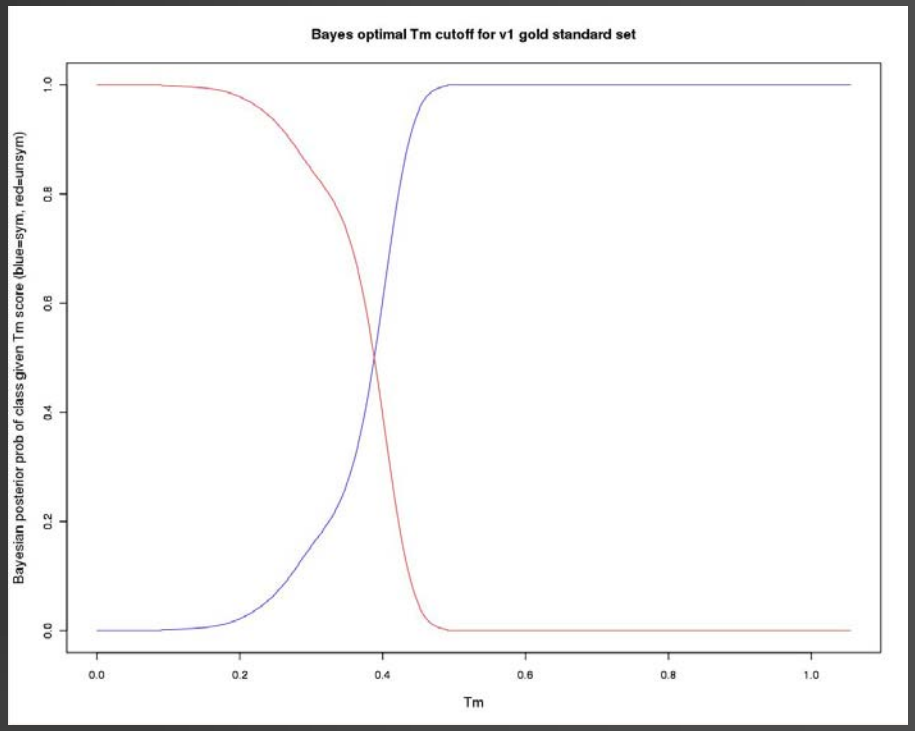

Carbonic anhydrase




EPSP synthase

# Symmetric/Unsymmetric Cutoff



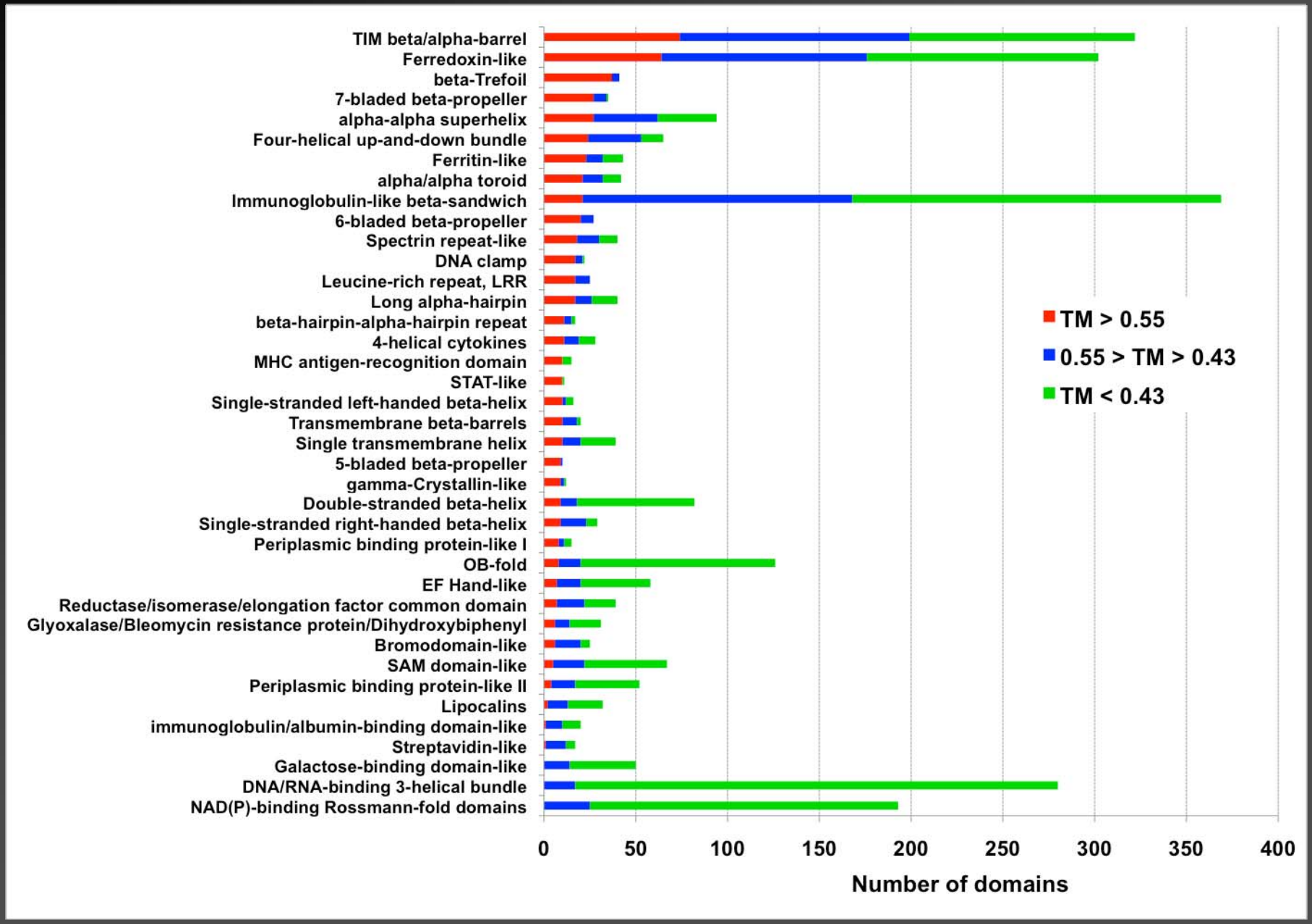




-Using a manually classified gold standard set of 134 domains, we computed optimal cutoffs by several methods (e.g. Bayes, fitting EVD).

-We obtain the probability of Tm given the class (symmetric/ unsymmetric) from our gold standard set. Inverting via Bayes theorem gives the probability of class given Tm and a cutoff (~0.4) between the classes that minimizes the probability of misclassification.

-Fitting an EVD and requiring a very low level of false positives, results in a more conservative cutoff of 0.43, which is what we settled on.

# SCOP Folds with at Least 10 Domains with TM > 0.43



# Counting Repeats in Symmetric Monomers

-To count the number of repeats, we rotate the protein around the symmetry axis (given by the best SymD transformation) and translate it along the axis in increments that are scaled down from the best symmetry operation.

-For example, if the best transformation is a 30 degree rotation and 3A translation along the rotation axis, then one might test the goodness of superposition between original and structures that have been transformed in increments of 1 degree rotation and 0.3A translation.

-At each step, these transformed structures are compared to the original with a simple score: the number of residues for which i, i-1, and i+1 in the original and j, j-1, and j+1 in the transformed copy superpose to better than 3A.

-Peaks in the plot of this score against the number of such incremental transformations applied correspond to repeating units. This procedure also generates the mean repeat length in residues and the mean angle subtended by repeats.

# Counting Repeats in Symmetric Monomers

-Such incremental comparison is done because the best SymD transformation may correspond to a lower symmetry than is actually present in the molecule. For example, most TIM barrels superpose with themselves well with a four or eight fold rotation, but usually superpose better with a 2-fold rotation, and this is often the highest scoring transformation that SymD finds for TIM's.

-For *closed* structures, we quit rotating the copy after one full turn. For *open* structures, we quit when the copy has translated off the original.

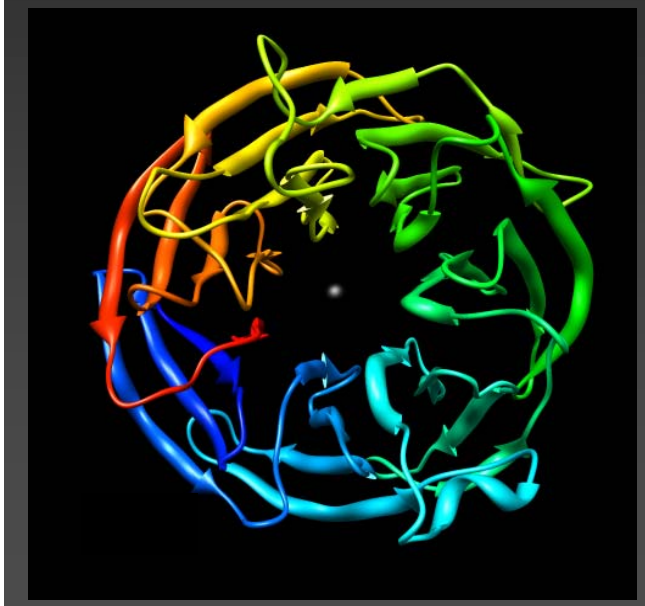
-We take the autocorrelation of the plot of the number or residues that superpose well versus rotation angle. From the first peak in the autocorrelation, we calculate a smoothing window and smooth the plot using a simple moving average with this smoothing window.

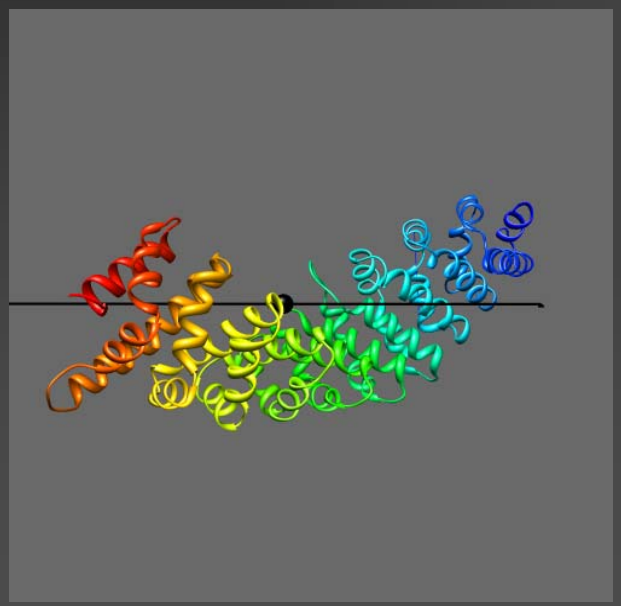
-Local maxima in the smoothed plot correspond to repeats.

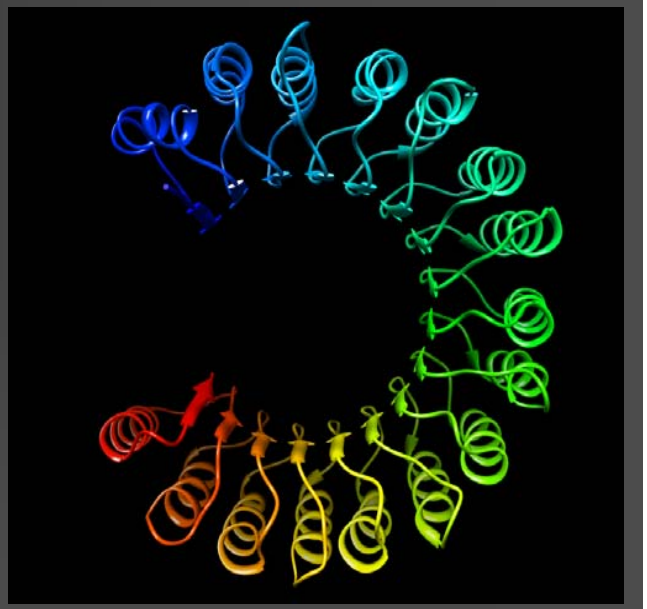
# Open Versus Closed Symmetric Structures

-Ring-shaped structures like beta-propellers and trefoils we call *closed* (below left). Structures with helical symmetry like alpha-alpha superhelices, or planar but not closed like leucine rich repeats (LRR) we call *open* structures (below middle and right).

-Closed structures subtend 360 degrees with respect to the symmetry axis. The angle subtended by open structures can vary.







# Algorithm for Determining Closed or Open

-Denote by the term *shift sequence* ($S_1$, $S_2$, $S_3$, …, $S_n$) the sequence of shifts ordered by corresponding untransformed serial number. Define $N_b$ as the number of the $S_i$ for which the absolute values are less than 13. Define $N_a$ as the number of the $S_i$ for which the absolute values are at least 13. A structure is defined to be *closed* if it meets the following criteria:

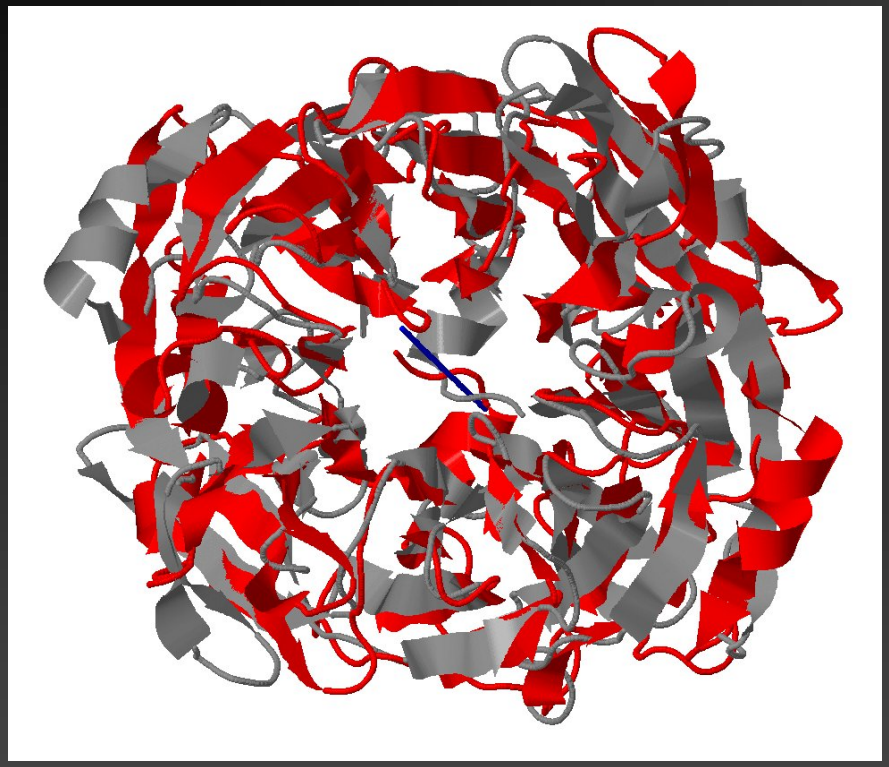
1) The sign of the shift sequence changes exactly once, i.e. $S_{i-1} \times S_i < 0$ for exactly one i, and furthermore $|S_{i-1}| > 20$ and $|S_i| > 20$ at this value of i

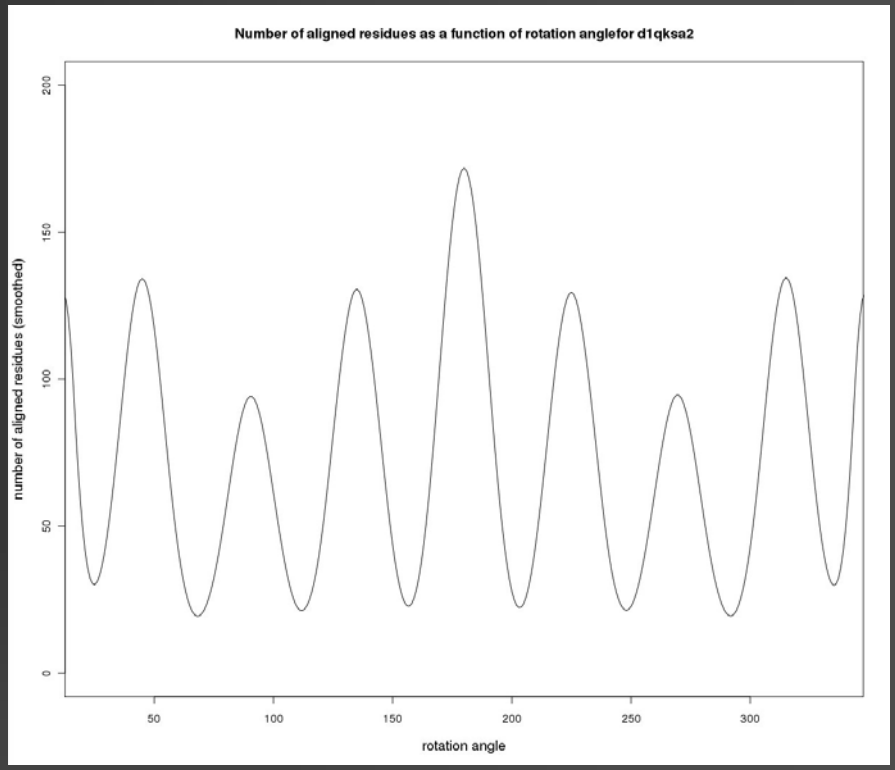
2) the ratio $R = N_a / (N_a + N_b) > 0.6$

-If a structure is not *closed*, it is *open*.

-For closed structures, when rotating a copy of the protein monomer around the symmetry axis to count repeats, we completely neglect the translational component of the SymD transformation when counting repeats.

# Counting Repeats in an 8-Blade Beta-Propeller



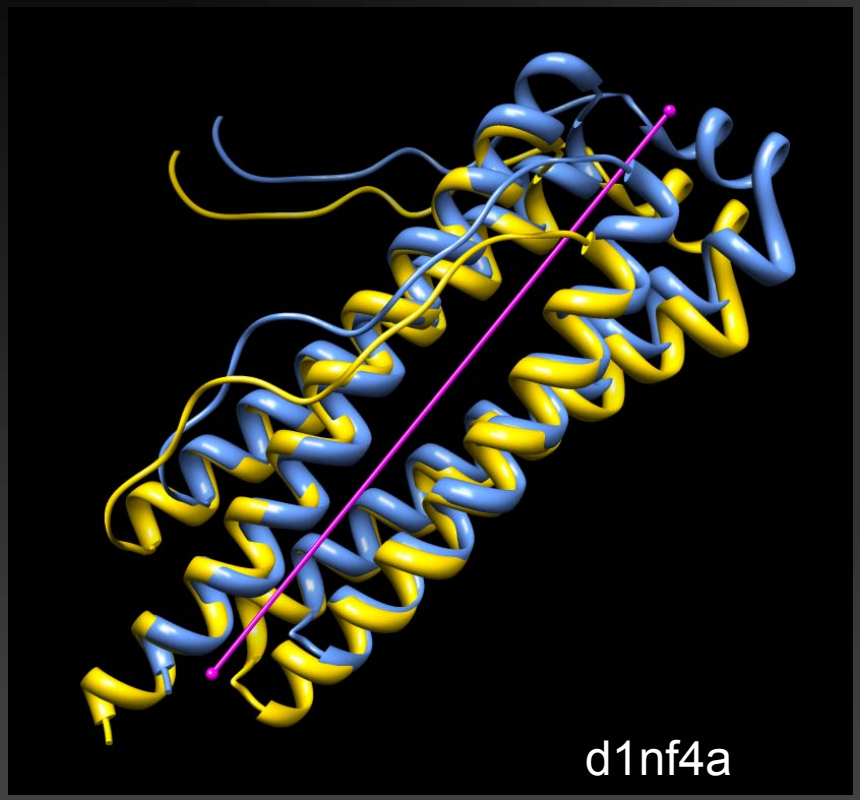



Best Symd superposition (left) with original structure in grey and transformed in red for the SCOP domain d1qksa2, an 8-bladed beta propeller. The rotation/symmetry axis is in blue. Notice that this best superposition is a 180 degree rotation that corresponds to the peak at 180 in the plot of number of superposed residues versus rotation angle (right). Also notice the 'self' peaks near 0 and 360 degrees.

# Counting Repeats for a Test Set

| SCOP fold | num domain | num slip | num correct | num repeat | dominant type of symmetry | num closed |
|---|---|---|---|---|---|---|
| Acid protease (b.50) | 32 | 0 | 32 | 2 | 2-fold rotation* | 32 |
| Alpha alpha superhelix (a.118)**** | 93 | 14 | 63 | varies | helical | 0 |
| Alpha alpha toroid (a.102) | 44 | 0 | 41 | varies | rotation** | 41 |
| Beta hairpin alpha hairpin (d.211) | 20 | 0 | 20 | varies | helical | 0 |
| Beta trefoil (b.42) | 74 | 0 | 74 | 3 | 3-fold rotation* | 74 |
| Bromodomain (a.29) | 38 | 3 | 35 | 2 | 2-fold rotation* | 35 |
| Double stranded beta helix (b.82) | 33 | 0 | 33 | 2 | 2-fold rotation* | 31 |
| EF hand (a.39) | 69 | 0 | 64 | 2 | 2-fold rotation or screw* | 57 |
| Ferredoxin (d.58) | 97 | 0 | 97 | 2 | 2-fold rotation* | 97 |
| Ferritin (a.25) | 65 | 9 |  | 2 | 2-fold rotation* | 56 |
| Four helical up down (a.24) | 80 | 11 | 68 | 2 | 2-fold rotation* | 69 |
| Immunoglobulin (b.1) | 78 | 0 | 78 | 2 | 2-fold rotation* | 78 |
| Leucine rich repeat (c.10) **** | 35 | 0 | 23 | varies | rotation or screw | 0 |
| MHC antigen recognition (d.19) | 43 | 0 | 95 | 2 | 2-fold rotation* | 43 |
| Pentein (d.126) | 12 | 0 | 12 | 5 | 5-fold rotation* | 12 |
| B-propeller 8-blade (b.70)*** | 13 | 0 | 13 | 8 | 8-fold rotation*** | 13 |
| B-propeller 5-blade (b.67) | 14 | 0 | 14 | 5 | 5-fold rotation** | 14 |
| B-propeller 4-blade (b.66) | 7 | 0 | 7 | 4 | 4-fold rotation** | 7 |
| B-propeller 7-blade (b.69)*** | 37 | 0 | 37 | 7 | 7-fold rotation** | 37 |
| B-propeller 6-blade (b.68) | 30 | 0 | 30 | 6 | 6-fold rotation** | 29 |
| B-propeller 10-blade | 1 | 0 | 1 | 10 | 10-fold rotation** | 1 |
| Rossman (c.2) | 60 | 0 | 59 | 2 | 2-fold rotation* | 60 |
| Spectrin (a.7) | 49 | 31 | 14 | 2 | 2-fold rotation* | 7 |
| SSLH beta helix (b.81) | 16 | 1 | 14 | varies | helical | 0 |
| SSRH beta helix (b.80)**** | 32 | 0 | 17 | varies | helical | 0 |
| TIM barrel (c.1) | 97 | 0 | 93 | 8 | 8-fold rotation* | 96 |
| Transmembrane beta barrel (f.4) | 20 | 0 | 14/18 | varies | rotation** | 18 |

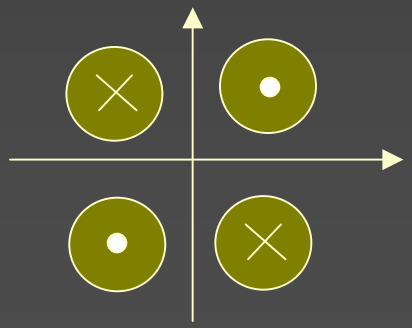
# Slip Symmetry




-Often when the repeat counting procedure fails, the best transformed structure is as in the figure at left.

-We call this sort of pseudo translation *slip symmetry.* It usually occurs with helical proteins.

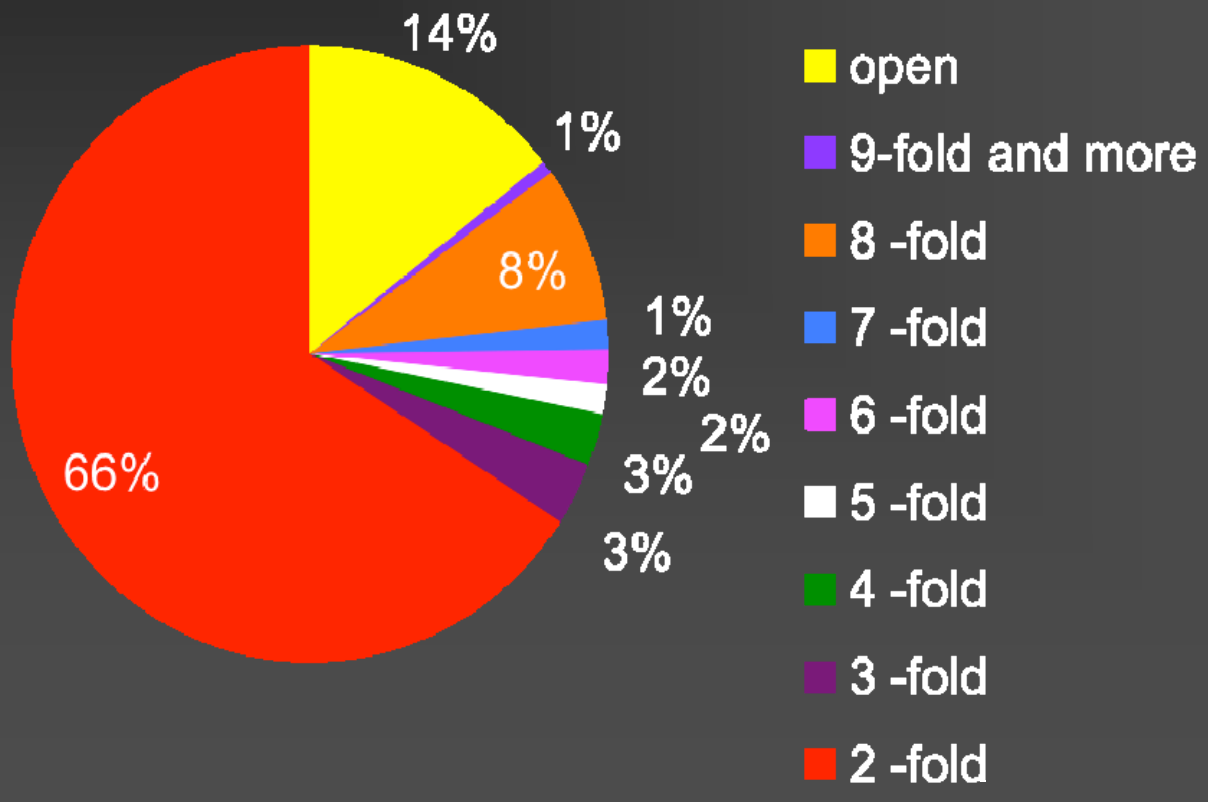
-With slip symmetry, most secondary structure elements from the transformed structure superpose with themselves in the untransformed structure.



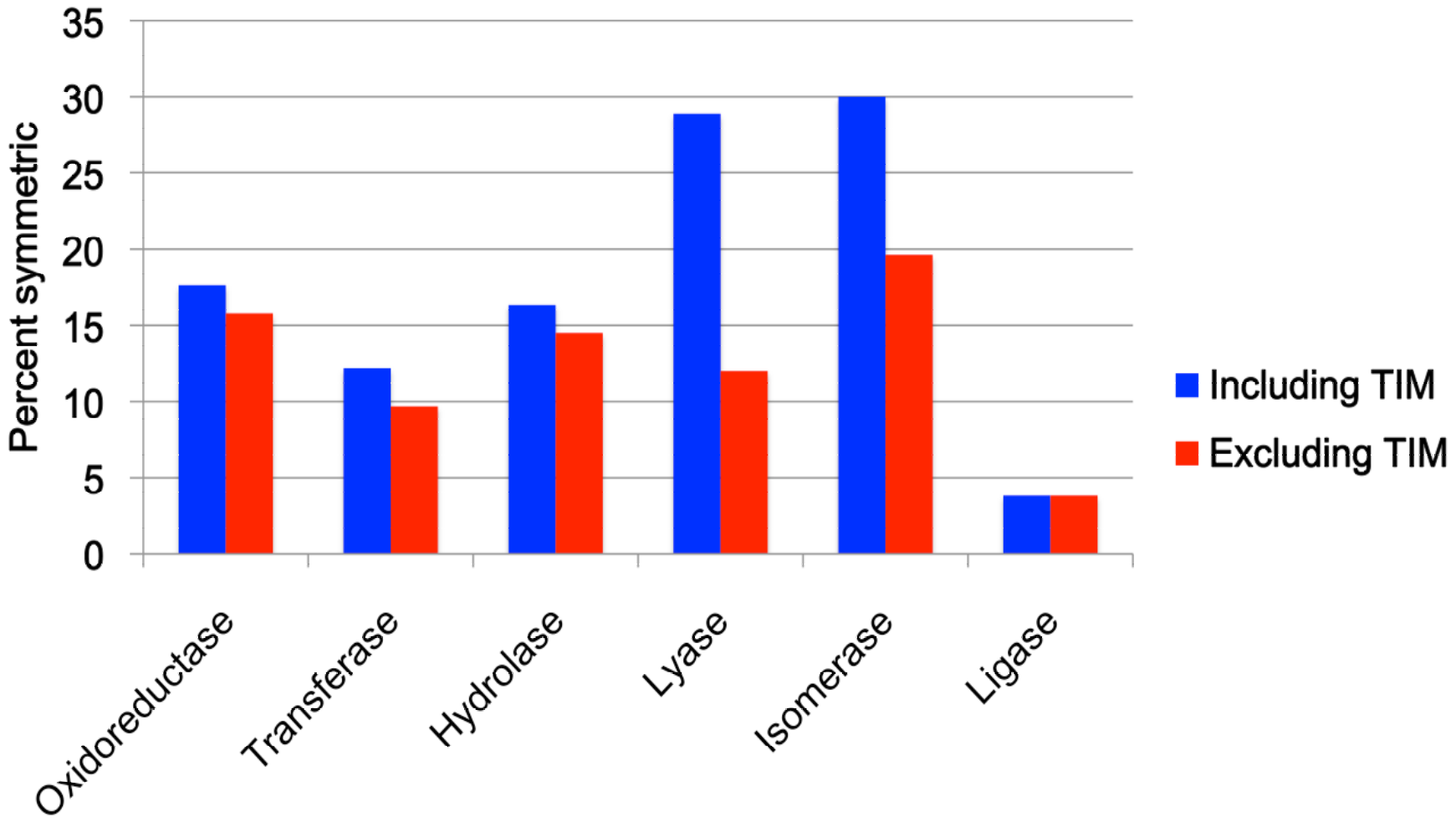
-A cross section from a generic 4-helical bundle is shown at left. Helices with N to C direction going into the screen are marked with X and with N to C direction coming out of the screen with a dot. They typically have one and sometimes two 2-fold axes perpendicular to the helices. d1nf4a shown above instead has the peculiar slip pseudo translation axis.

# Classification of Symmetric Proteins

There are 10,568 domains in the SCOP 1.75/Astral40 non-redundant protein domain database. Of these, 2,047 are symmetric by our criteria, or about 19%. The breakdown of these symmetric domains by symmetry sub-type (n-fold rotation or open/helical) is given below.

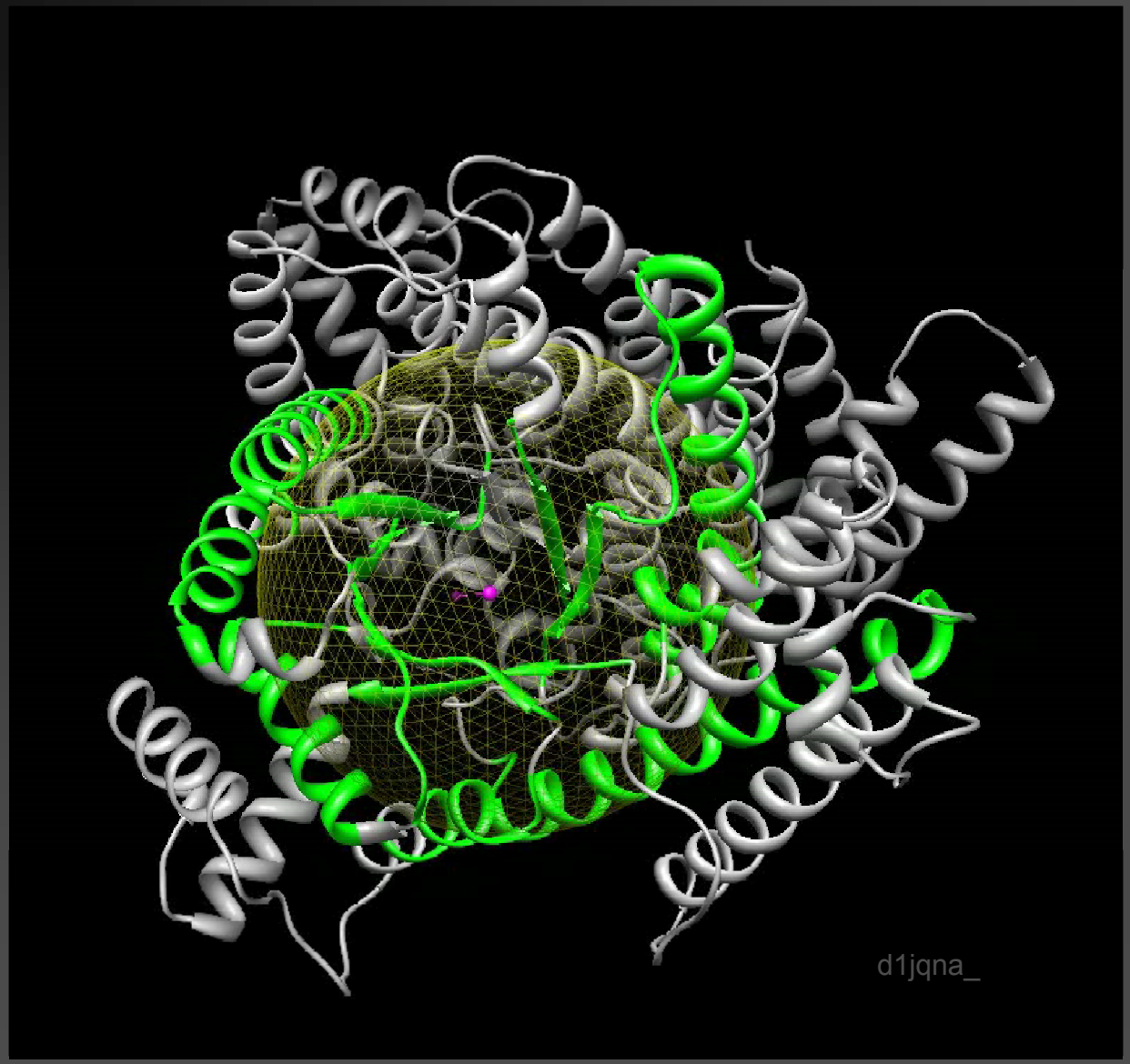


# Symmetry and Enzyme Function

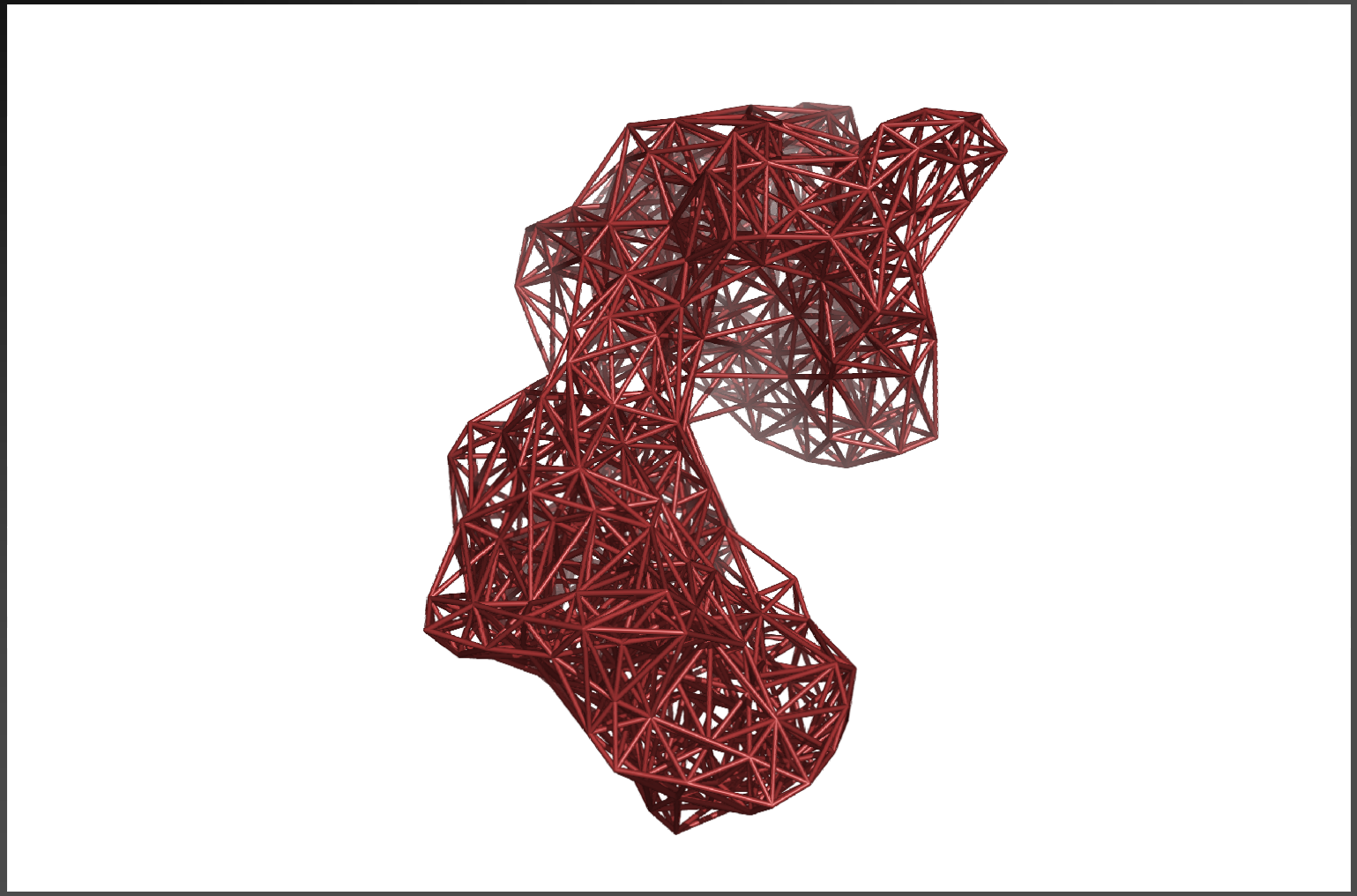



From EC numbers assigned to domains in the PROCOGNATE database

# Detecting Locally Symmetric Region-Spherical Probe



# Detecting Locally Symmetric Region-Delaunay Tetrahedra



*Delaunay tessellation of Phosphoglycerate Kinase (16pk) with 10A simplex edge cutoff imposed.*

## Conclusions and Further Work

We can find and quantify monomeric pseudo-symmetry in proteins using an automated method.

We can accurately count repeats in such symmetric monomers, particularly for closed structures, using an automated method.

It appears that there is a weak correlation of symmetry with some enzymatic functions.

A logical follow on project is to find local symmetry—to find several different symmetric substructures of a single structure, or a symmetric fragment of a non-symmetric structure.

## Acknowledgements

Dukka KC

Changoon Kim

Matt Jenny

BK Lee

Funding:  NIH Intramural

# Selected References


KC D, Taylor TJ, Tai E, Lee B (2012) SymD2.0: Improved Symmetry Detection Algorithm and its application to protein domain universe (submitted)

Kim C, Basner J, Lee BK (2010). Detecting internally symmetric protein structures. BMC Bioinformatics 11:303.

Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. Acta Crystallogr A 32(5):922-923.

Kim C, Tai CH, Lee B (2009) Iterative refinement of structure-based sequence alignments by Seed Extension. BMC Bioinformatics 10:210.

Tai CH, Vincent JJ, Kim C, Lee B (2009) SE: an algorithm for deriving sequence alignment from a pair of superimposed structures. BMC Bioinformatics 10 Suppl 1:S4.

Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57(4):702-710.

Andrade MA, Perez-Iratxeta C, Ponting CP (2001) Protein repeats: structures, functions, and evolution. J Struct Biol 134(2-3):117-131.

Kinoshita K, Kidera A, Go N (1999) Diversity of functions of proteins with internal symmetry in spatial arrangement of secondary structural elements. Protein Sci 8(6):1210-1217.

Taylor WR, Heringa J, Baud F, Flores TP. A Fourier analysis of symmetry in protein structure. Protein Eng 2002;15(2):79-89.