Understanding Sequence, Structure and Function Relationships and the Resulting Redundancy

many slides by

Philip E. Bourne

Department of Pharmacology, UCSD

Agenda

- Understand the relationship between sequence, structure and function. Consider specifically:
  - sequence-structure
  - structure-structure
  - structure-function

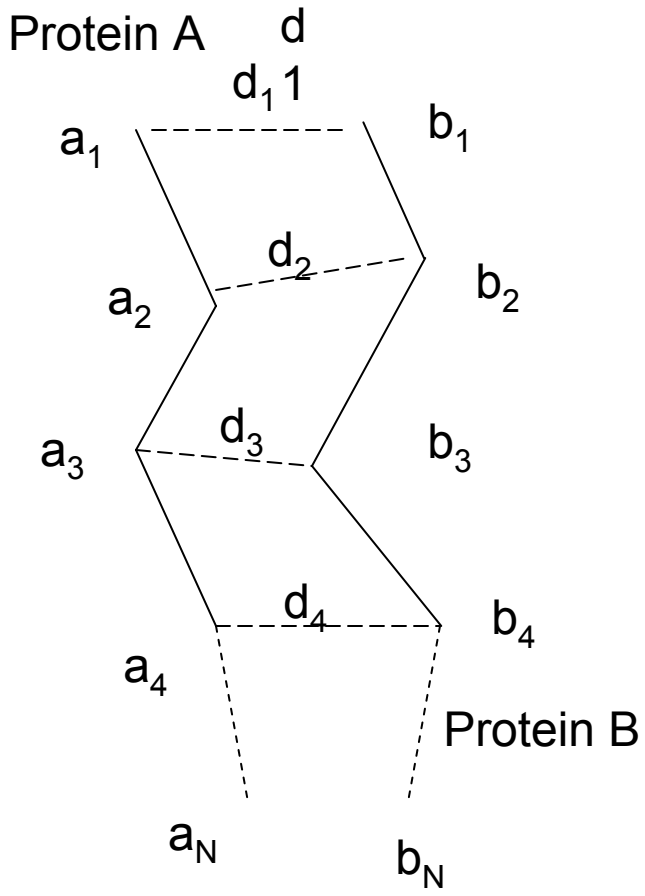
- Take home message: a non-redundant set of sequences is different than a non-redundant set of structures is different than a non-redundant set of functions

Why Bother?

- Biology:
  - A full understanding of a molecular system comes from careful examination of the sequence-structure-function triad
  - Each triad is then a component in a biological process

- Method:
  - Bioinformatics studies invariably start from a non-redundant set of data to achieve appropriate statistical significance

Background – RMSD Defined



Represents the overall distance between two proteins usually averaged over their C-alpha atoms denoted here a and b

$$RMSD = Sqrt\left(1/N \sum_{i=1}^{i=N} |d_i|^2\right)$$

Thus RMSD is the square root of the sum of the squares of the distances between all C-alpha atoms

Rule of thumb:
1-2A - the proteins are close
<6 A - they are likely related

Note: Assumes you know residues correspondences

Some Useful Observations

- Below 30% protein sequence identity detection of a homologous relationship is not guaranteed by sequence alone
- Structure is much more conserved than sequence
- Distinguishing between divergent versus convergent evolution is an issue
- Structure is limited relative to sequence or the order 1:100
- Structure follows a power law with respect to function – each structural template has from 1 to n functions

Relationship Between Sequence and Structure

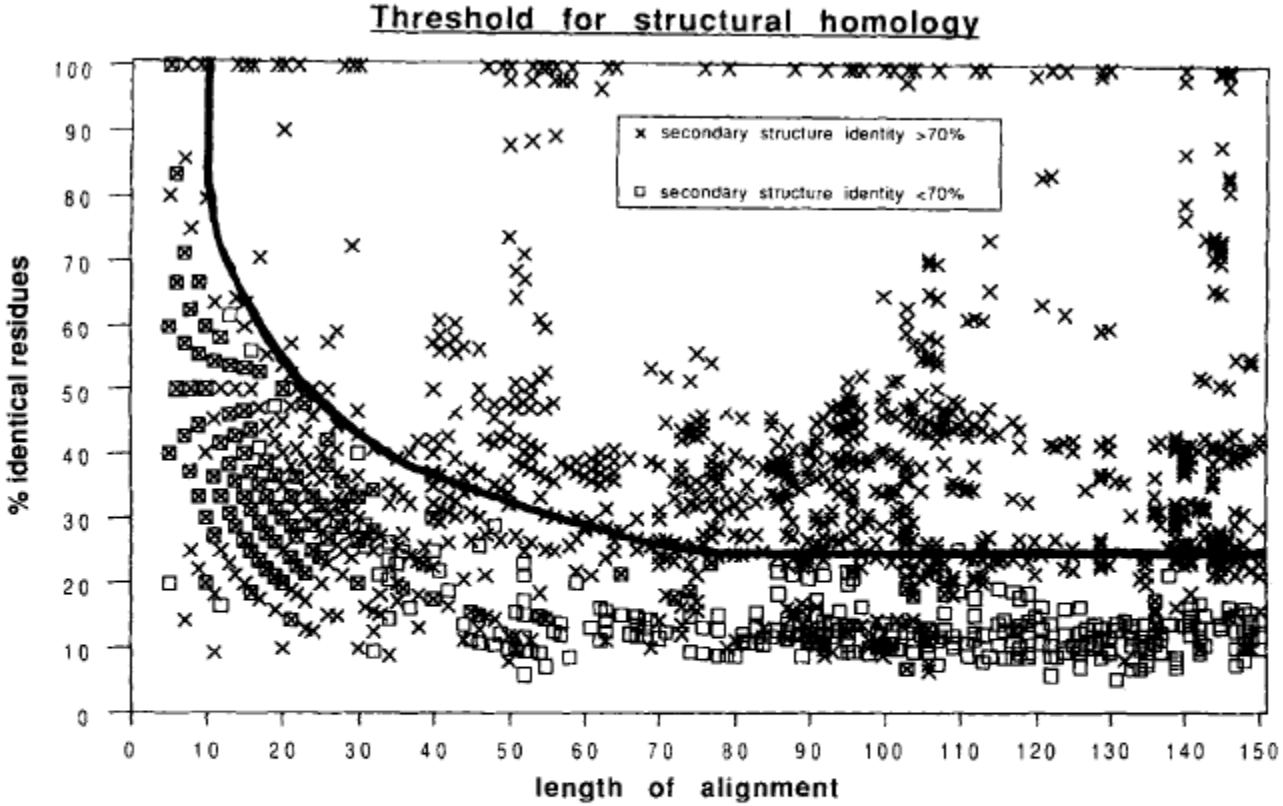


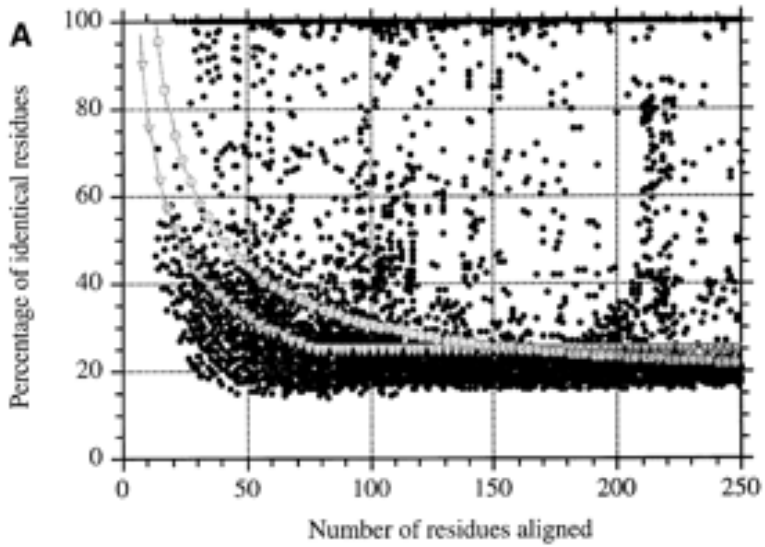
Fig. 4. Homology threshold for structurally reliable alignments as a function of alignment length, a principal result (numerical values in Table I). Each data point represents an alignment between two fragments from proteins of known structure. The graph is a two dimensional projection of Figure 2 onto the plane of sequence similarity/alignment length, with structural similarity collapsed to a one bit yes/no description (crosses/squares). The data points are a subset of the data in Figure 2. The homology threshold (curved line) divides the graph into a region of *safe structural homology* (upper right) where essentially all fragment pairs are observed to have good structural similarity (crosses, secondary structure identity above 70%) and a region of *homology unknown or unlikely* (lower left) where some fragment pairs are structurally similar (crosses) and some are not (squares, secondary structure identity below 70%). The histogram of Figure 3a corresponds to a vertical slice of this graph in the length range 79–150 residues, summing all available data points in that length range.

The classic hssp curve from Sander and Schneider (1991) *Proteins* 9:56-68

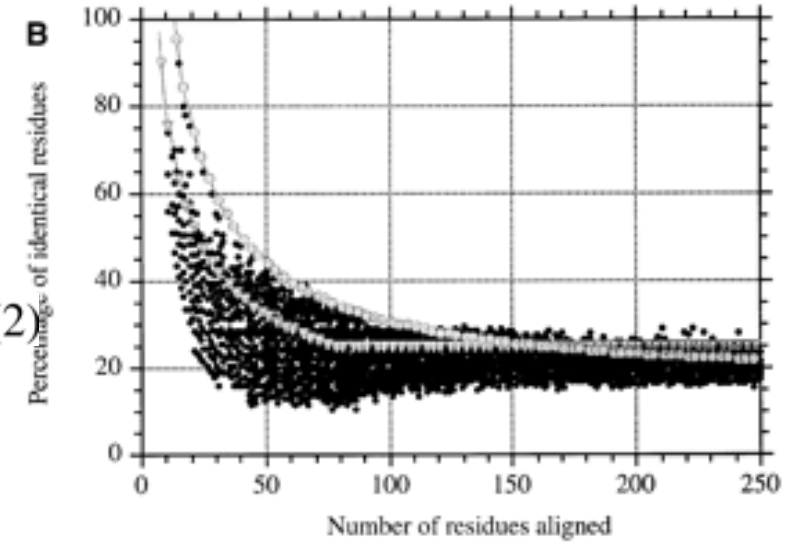
$$p^I(n) = n + \begin{cases} 290.15 \cdot L^{-0.562}, & \text{for } L < 80 \\ 25 & \text{, doe } L \geqslant 80 \end{cases} \qquad (1)$$

This Analysis was Updated by Rost in 1999

http://peds.oupjournals.org/cgi/content/full/12/2/85

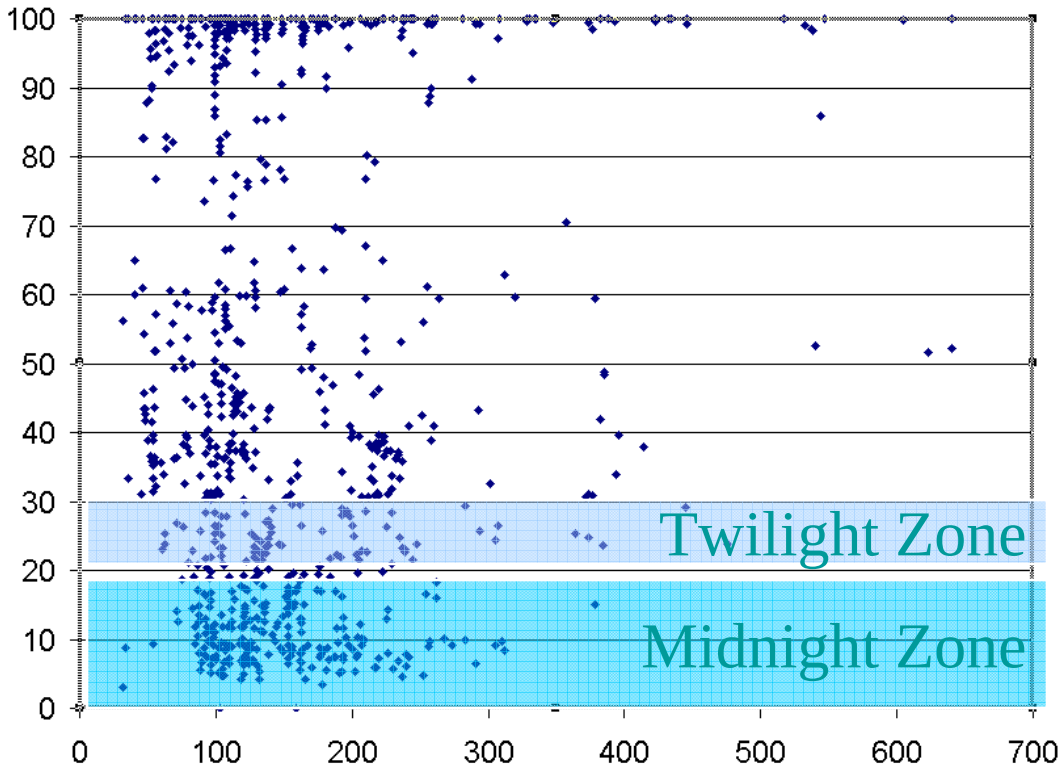


$$p^I(n = n + 480 \cdot L^{-0.32 \cdot (1 + e^{-L/1000})} \quad (2)$$



Sequence vs. Structure – Another Perspective

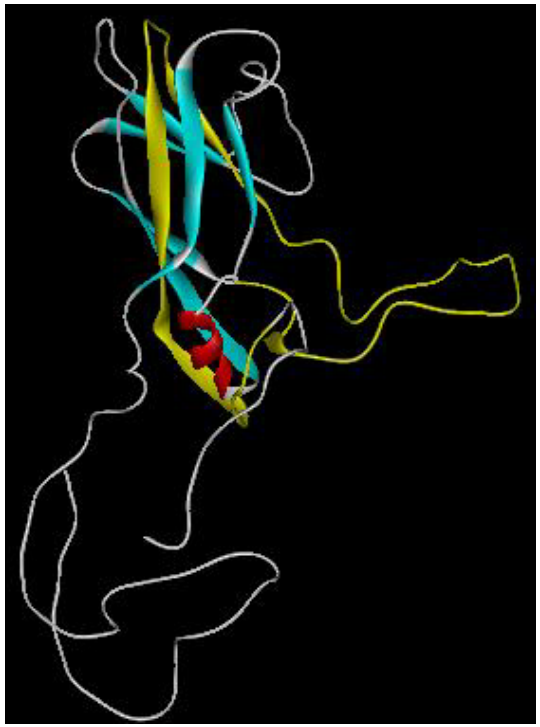
Random 1000 structurally similar PDB polypeptide chains from CE with z > 4.5 (% sequence identity vs alignment length)



There are no Absolute Rules - Similar Sequences – Different Structures

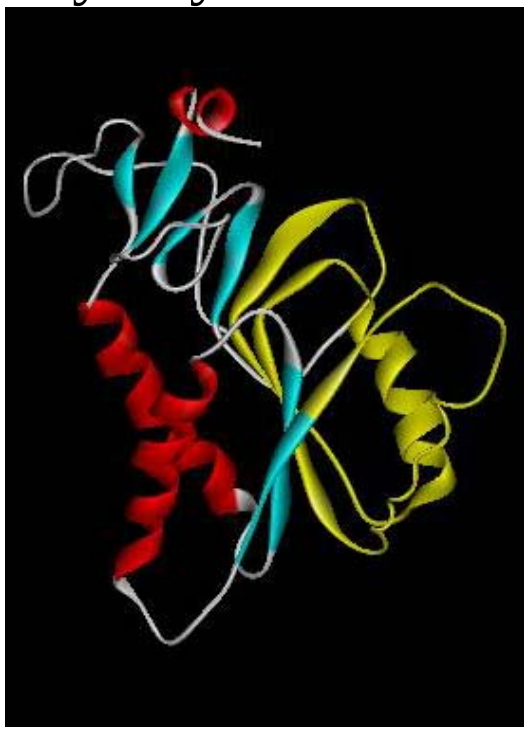
1PIV:1

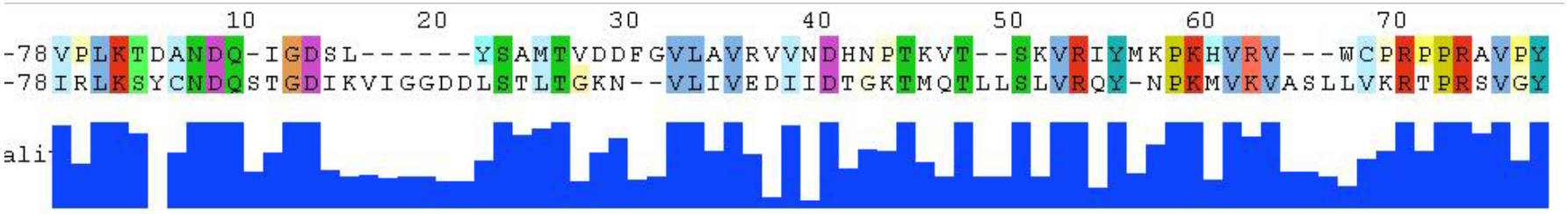
Viral Capsid Protein



1HMP:A

Glycosyltransferase



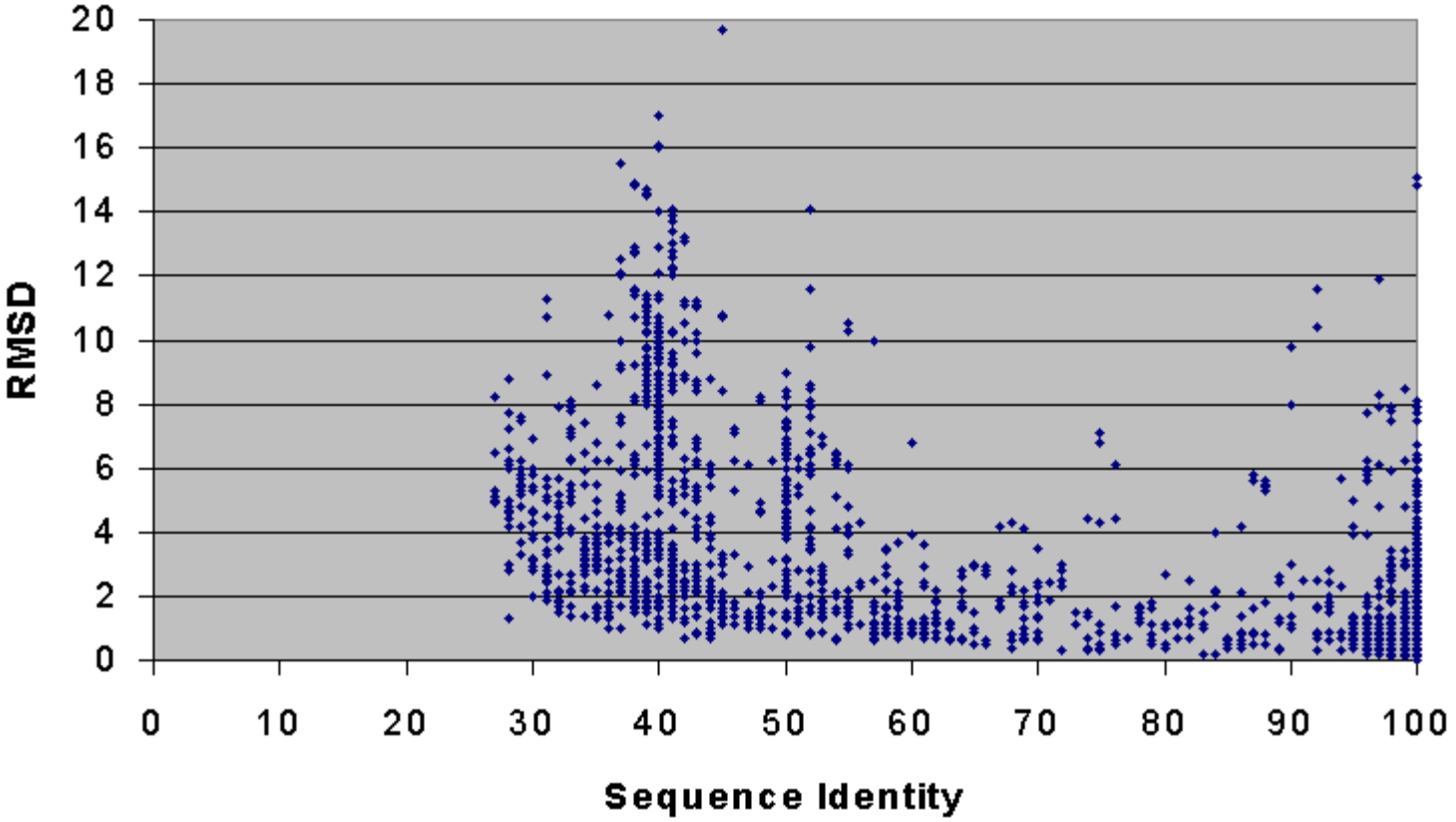


80 Residue Stretch (Yellow) with Over 40% Sequence Identity

Given This Complex Relationship a Non-redundant Set of Sequences Does not Imply a Non-redundant Set of Structures

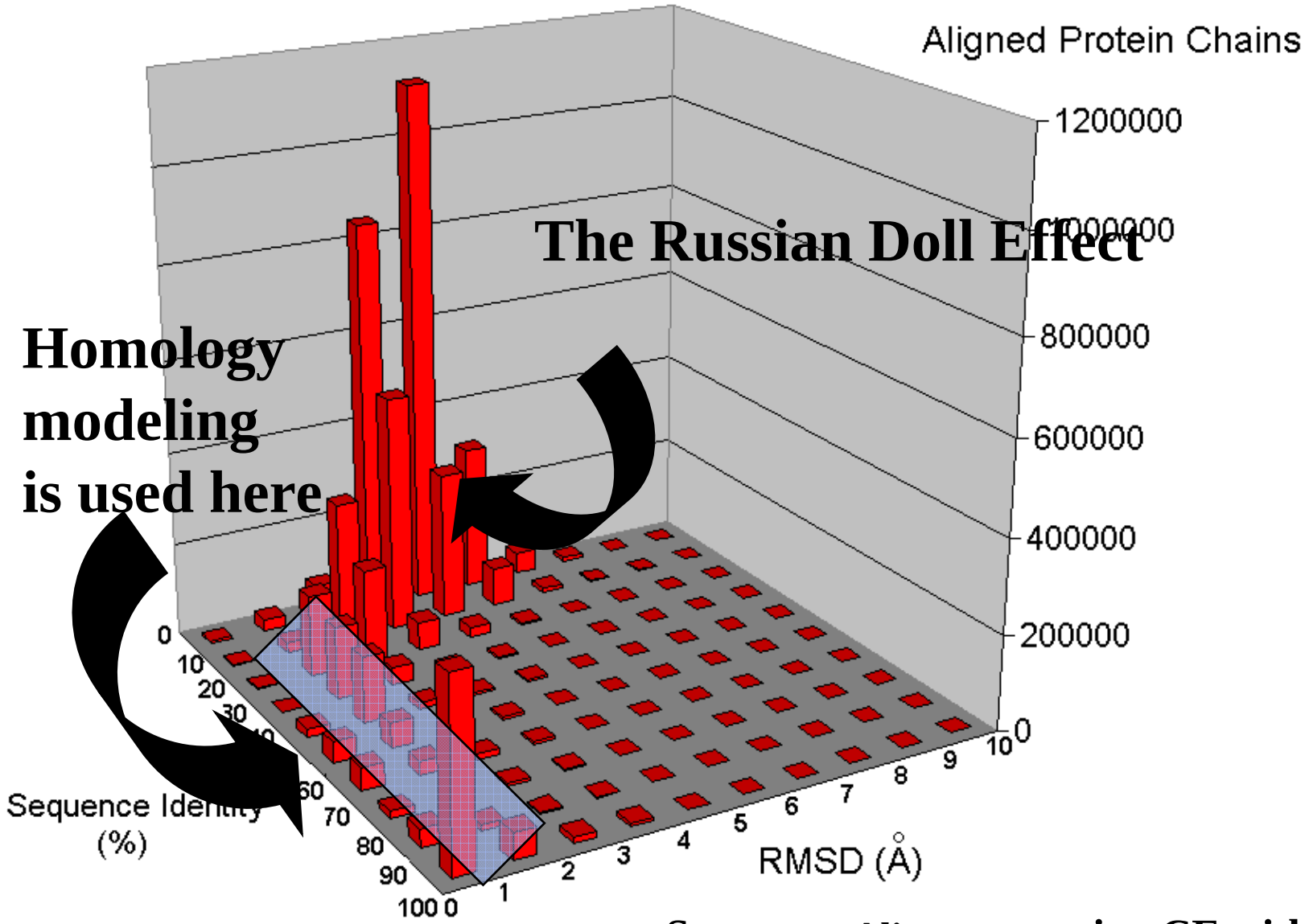
Non-Redundant Sequences Yet Redundant Structures

Structure Comparison of 30% of PDBSelect Set



Structure vs Structure

Structure Is Highly Redundant

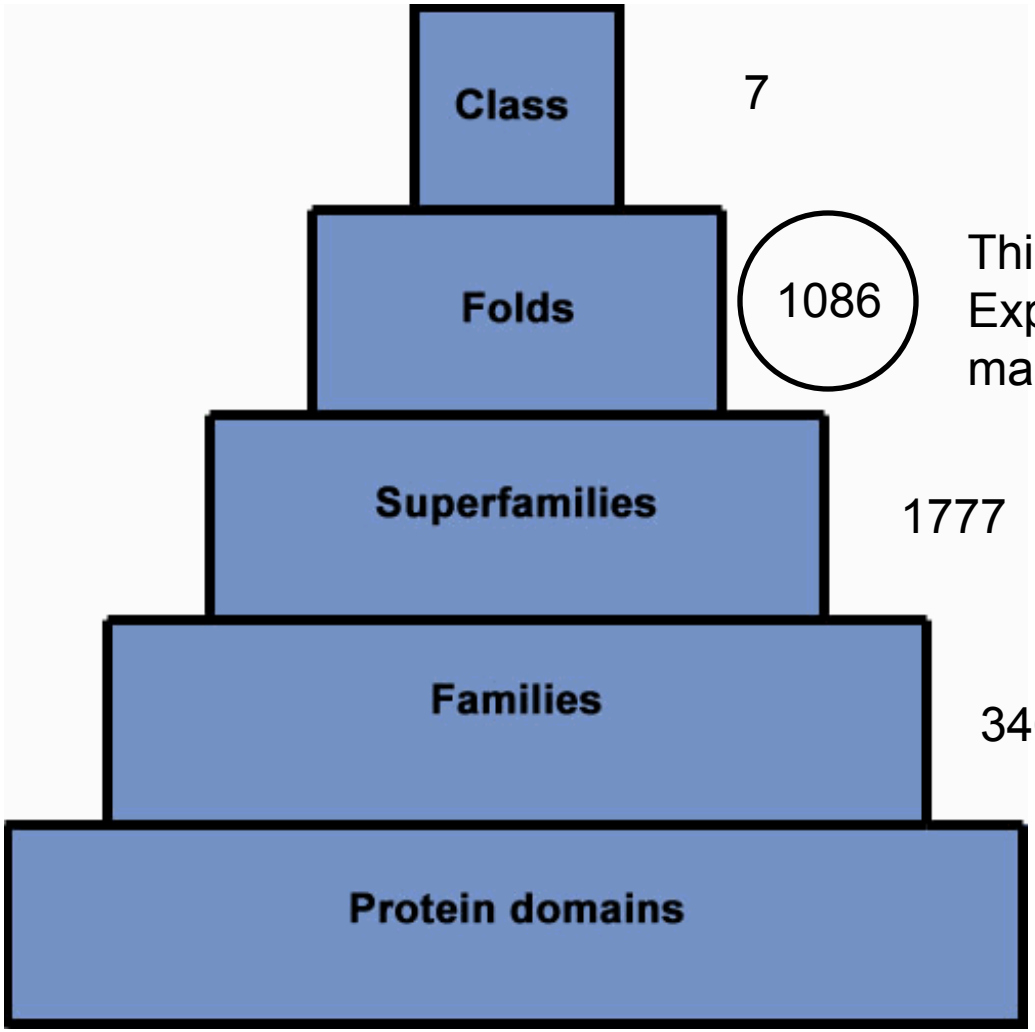


Structure Alignments using CE with z>4.0

We will be revisiting this later!

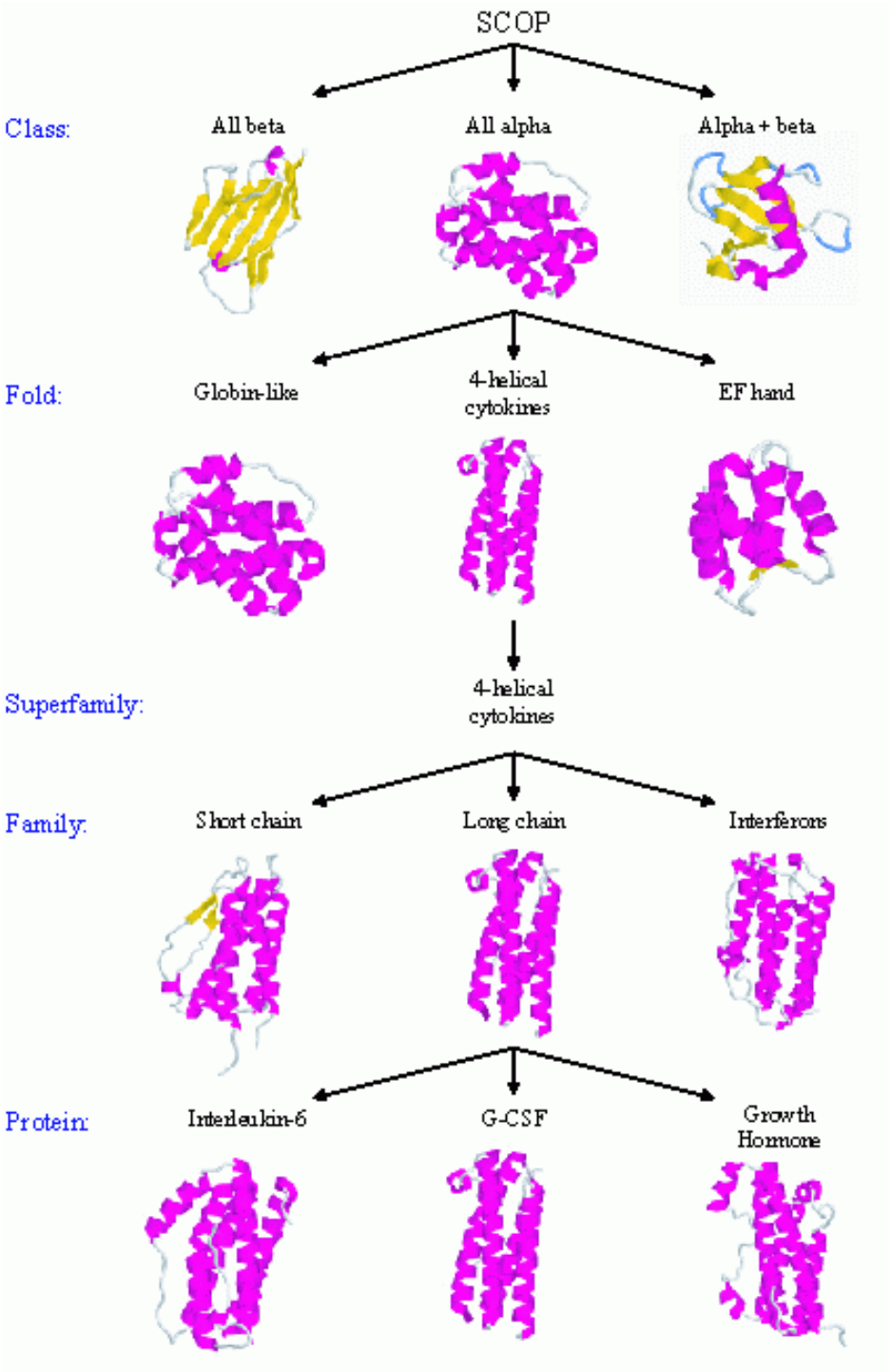
- Specifically:
  - How do we capture this redundancy?
  - What systems are commonly used to express this redundancy and what do they bring to our understanding of biology?
- For now consider what this means using the most popular structure classification scheme - SCOP

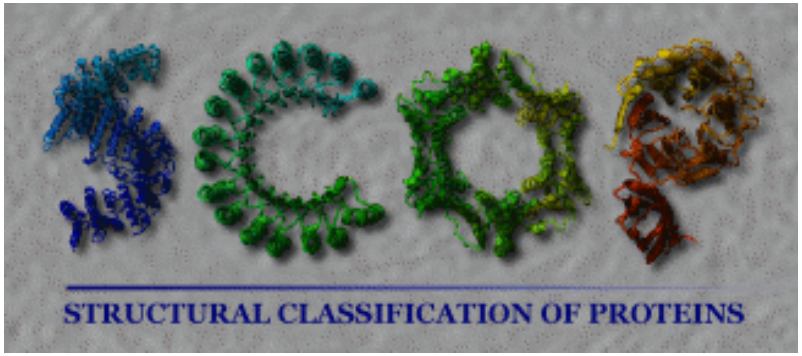
The SCOP Hierarchy v1.73



The SCOP Hierarchy v1.73

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	259	459	772
All beta proteins	165	331	679
Alpha and beta proteins (a/b)	141	232	736
Alpha and beta proteins (a+b)	334	488	897
Multi-domain proteins	53	53	74
Membrane and cell surface proteins	50	92	104
Small proteins	85	122	202
Total	1086	1777	3464





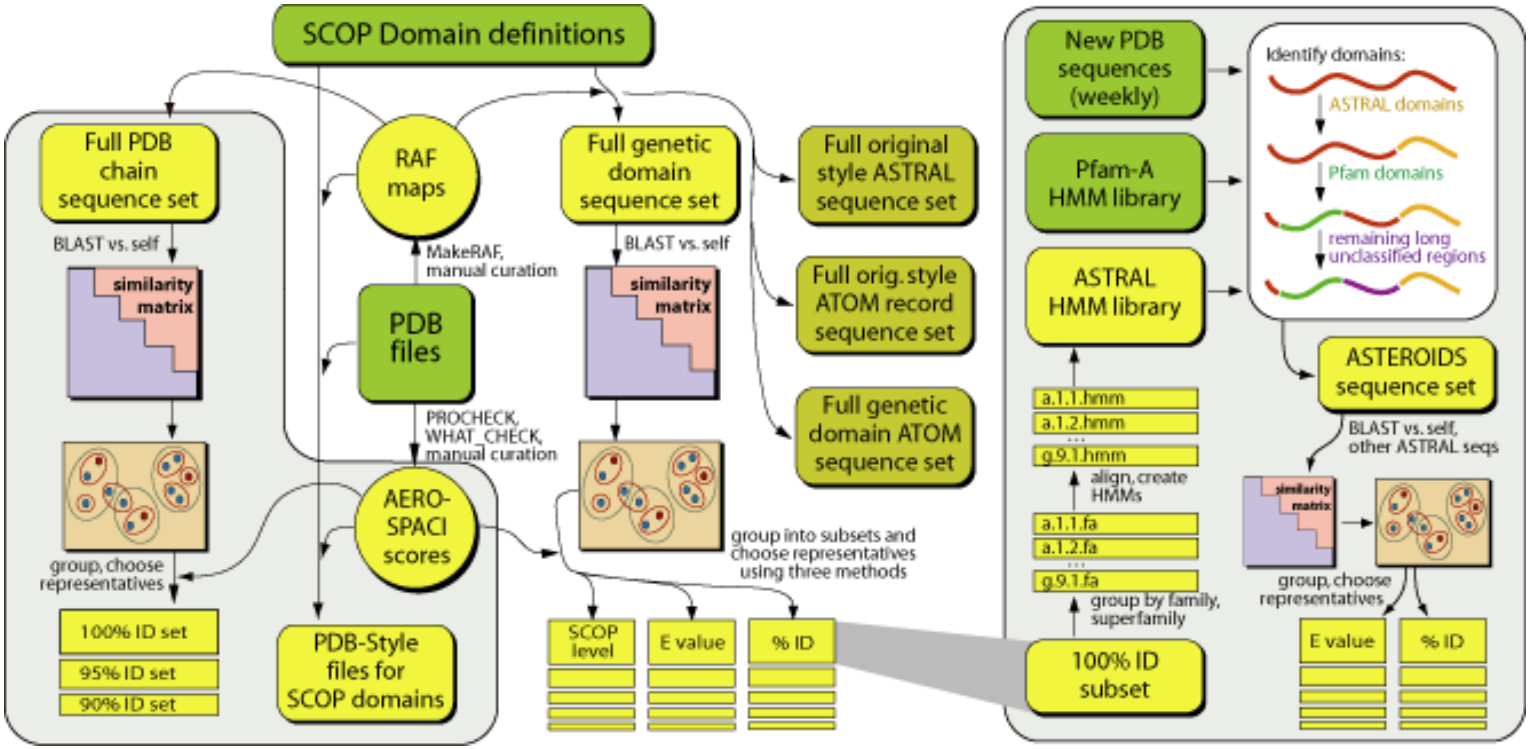
Specific Examples From the SCOP Hierarchy

New Features in SCOP

- **Species**
  - a distinct protein sequence and its naturally occurring or artificially created variants
- **Protein**
  - similar sequences of essentially the same functions that either originate from different biological species or represent different isoforms within the same organism
- **Family**
  - proteins with related sequences (>30% sid) but typically distinct functions
- **Superfamily**
  - protein families with common functional and structural features inferred to be from a common evolutionary ancestor
- **Fold**
- **Class**

ASTRAL Compendium

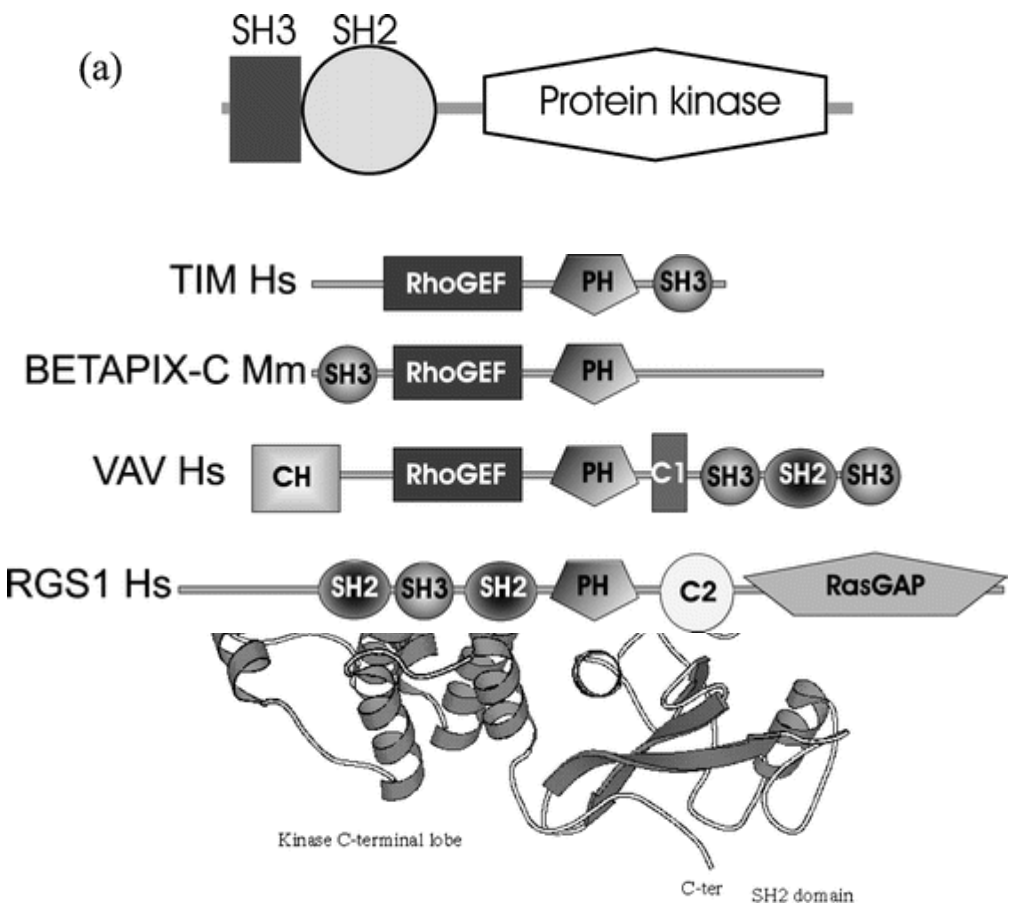
- **Selections from SCOP intended to provide high-quality subsets with low redundancy at desired degrees of similarity**



http://astral.berkeley.edu/

Protein Domains

- Definition
  - Compact, spatially distinct
  - Fold in isolation
  - Recurrence



Structure vs Function

Some Basic Rules Governing Structure-Function Relationships …

- The golden rule is there are no golden rules – George Bernard Shaw
- Above 40% sequence identity sequences tend to have the same structure and function – But there are exceptions
- Structure and function tend to diverge at the same level of sequence identity

Structure vs Function

This is even more complicated than the relationship between sequence and structure and not as well understood

Complication Comes from One Structure Multiple Functions

- We saw this from GO already
- phosphoglucose isomerase acts as a neuroleukin, cytokine and a differentiation mediator as a monomer in the extracellular space and as a dimer in the cell involved in glucose metabolism

Consider an Example Relative to SCOP

- lysozyme and alpha-lactalbumin:
  - Same class alpha+beta
  - Same superfamily – lysozyme-like
  - Same family C-type lysozyme
  - Same fold – lysozyme-like
  - different function at 40% sequence identity
    - Lysozyme – hydrolase EC 3.2.1.17
    - Alpha lactalbumin – Ca binding lactose biosynthesis

More Details…

Lysozyme is an O-glycosyl hydrolase, but α-lactalbumin does not have this catalytic activity. Instead it regulates the substrate specificity of galactosyl transferase through its sugar binding site, which is common to both α-lactalbumin and lysozyme. Both the sugar binding site and catalytic residues have been retained by lysozyme during evolution, but in α-lactalbumin, the catalytic residues have changed and it is no longer an enzyme.

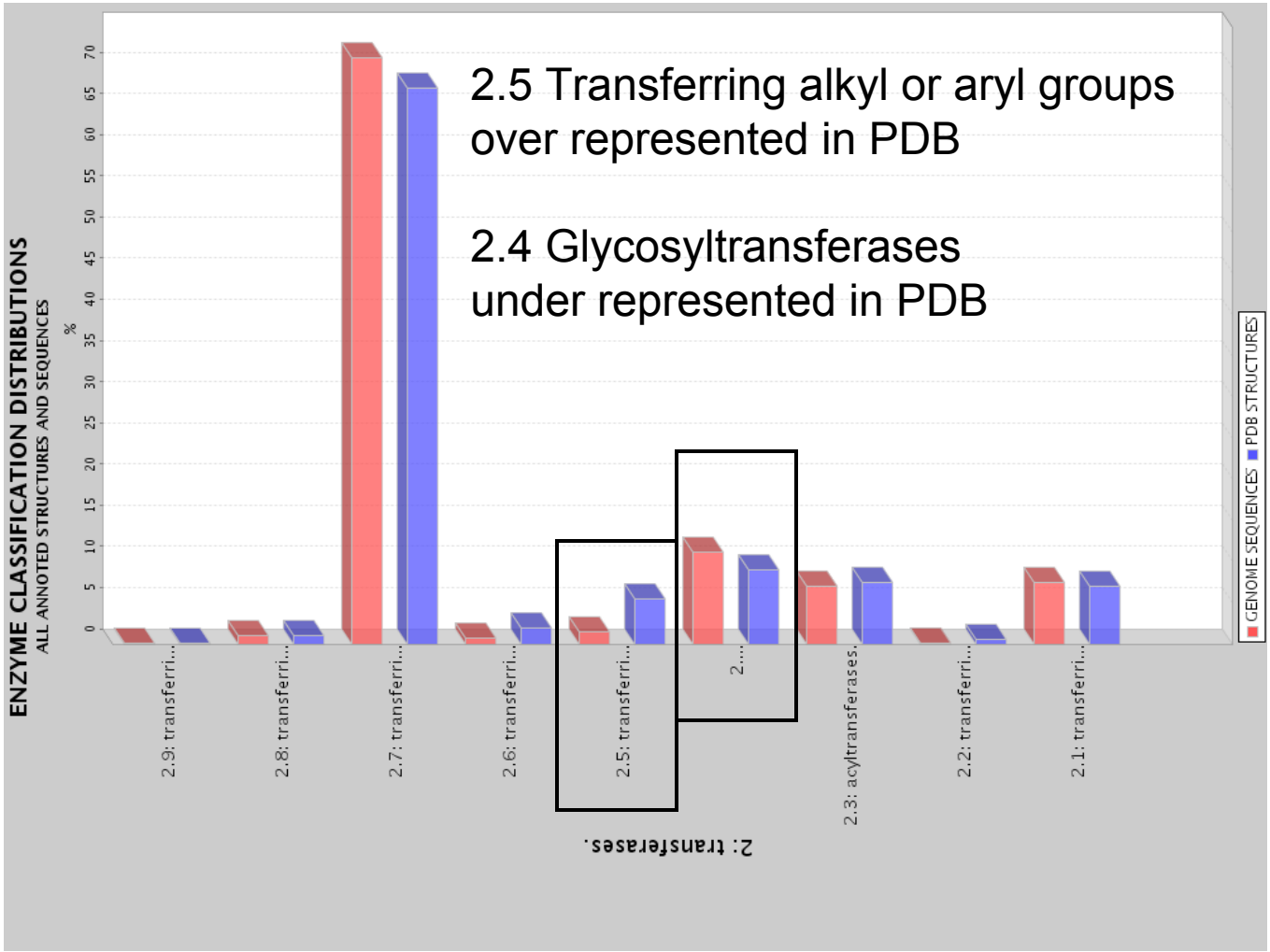
Why is It Not so Well Understood?

1. Function is often ill-defined e.g., biochemical, biological, phenotypical
2. The PDB is biased – it does not have a balanced repertoire of functions and those functions are ill-defined
3. There are a number of functional classifications eg EC, GO that have differing coverage and depth

Point 2 PDB Bias
PDB vs Human Genome
EC – Hydrolases – Begins to Illustrate the Bias in the PDB

PDB

Ensembl Human Genome Annotation



Xie and Bourne 2005 *PLoS Comp. Biol. 1(3) e31* http://sg.rcsb.org

Structure Follows a Power Law Distribution

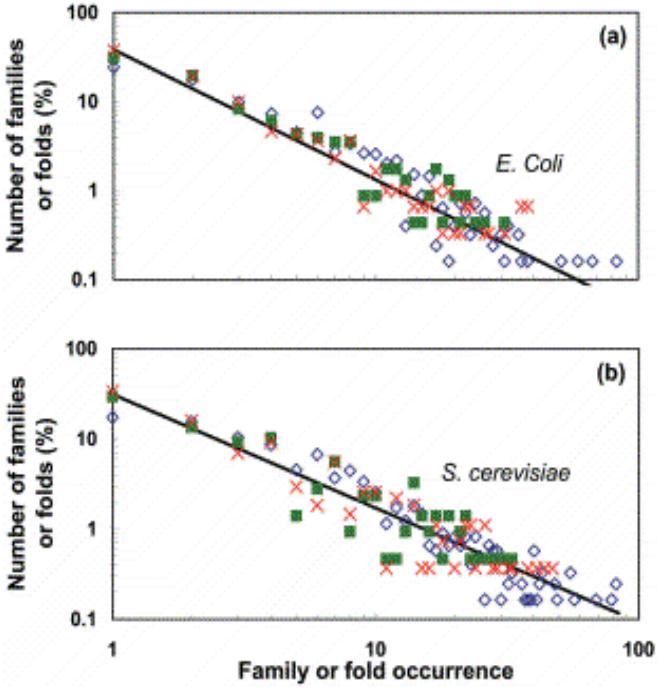


Figure 2. The occurrence of InterPro families (◇), SCOP superfamilies (×) and folds (■) (a) *E. coli* and (b) *S. cerevisiae*. The distributions for the three gene classifications are similar and follow power-law behaviour (—).

- Some folds are promiscuous and adopt many different functions - superfolds

Qian J, Luscombe NM, Gerstein M. JMB (2001) 313(4):673-81

Examples of Superfolds..

TIM barrel fold

The structure consists of an eightfold repeat of beta/alpha units. Eight parallel beta strands on the inside are covered by eight alpha helices on the outside. The fold was first seen in triose phosphate isomerase. All known TIM barrel structures are enzymes, except for the narbonin family. Many of these enzymes are glycosyl hydrolases (EC 3.2.x.x). The fold is higly versatile, being found in single domain monomeric enzymes and as the catalytic domain of larger enyzmes. The active site is found at the C−terminal end of the barrel in a series of loops, hence it is very easy to alter the function and / or specificity without altering the core structure.

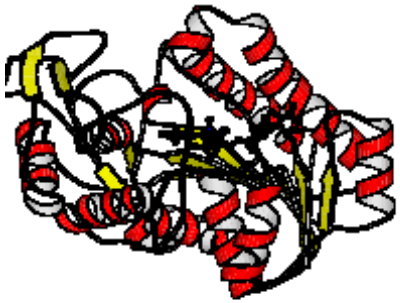
Number of EC numbers associated with this fold (to the third level): 29

Alpha/beta hydrolase fold

The structure is an eight stranded, mostly parallel alpha/beta structure. The fold is tolerant to large insertions and is a very plastic. All proteins known so far containing this fold are enzymes. The enzymatic properties of this fold are formed by a catalytic triad of a nucleophile, acid and a histidine residue. The nucleophile is found in a "nucleophilic elbow" turn located just after the fifth beta strand.

Number of EC numbers associated with this fold (to the third level): 17

Examples of Superfolds



NAD binding domain

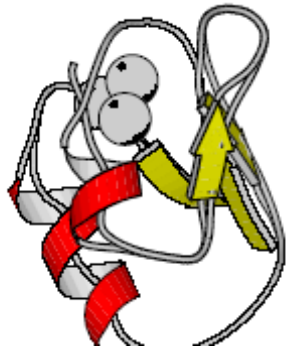
This is a double beta−alpha−beta−alpha−beta motif, and is a common structural motif of enzymes binding NAD, NADP and other related cofactors, for example, NAD is found in dehydrogenases as the hydrogen acceptor. The domain is found as a common core unit in many structures, with other structural units at the periphery.

Number of EC numbers associated with this fold (to the third level): 5

P−loop NTP hydrolase fold

This fold consists of alpha/beta/alpha, parallel or mixed beta sheets of variable size. The fold binds the phosphate of ATP or GTP and is found in ATP and GTP binding proteins such as adenylate kinase. The P−loop is a phosphate binding loop which binds the phosphate groups of ATP and GTP,and is a glycine−rich sequence with the consensus sequence (A,G)xxxxGK(T,S). The P−loop residues are shown in detail (left) in guanylate kinase.

Number of EC numbers associated with this fold (to the third level): 5



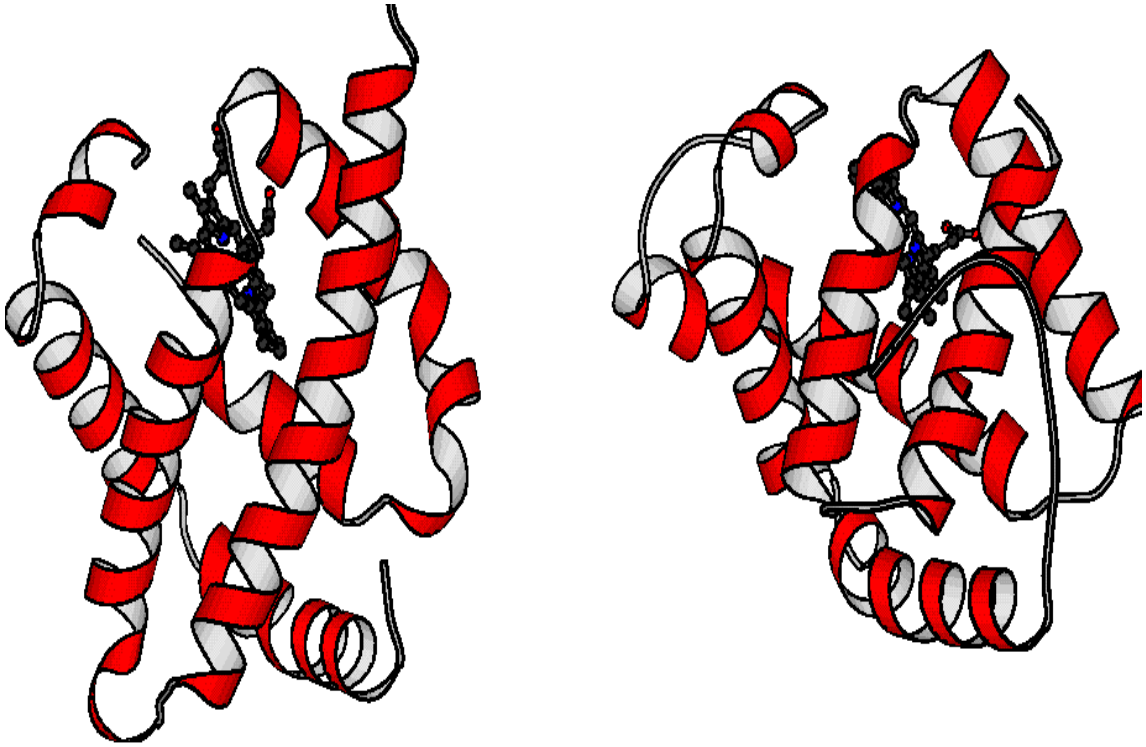
Ferredoxin−like fold

This fold consists of an alpha/beta sandwich with an antiparallel beta sheet. The ferredoxin−like fold is associated with predominantly with non−enzymatic ferredoxins, like the example shown (Ferredoxin ii from *D. gigas*, left). Ferre−doxins are iron−sulphur clusters invovled in electron transport, and often form part of multi−subunit assemblies. An example of an enzyme with this fold is muconolactone isomerase (EC 5.3.3.4).

Number of EC numbers associated with this fold (to the third level): 5

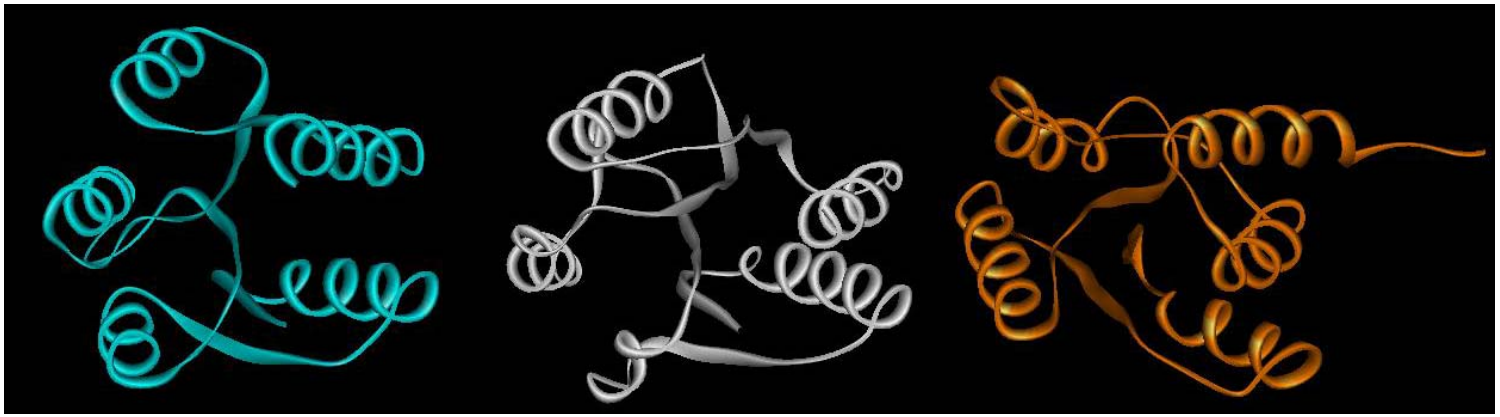
Specific Examples of the Relationship Between Structure and Function

Same Structure and Function Low Sequence Identity

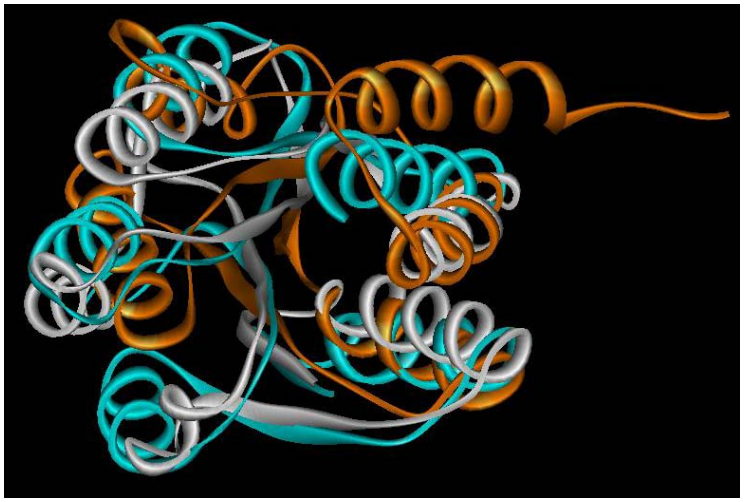


The globin fold is resilient to amino acid changes. *V. stercoraria* (bacterial) hemoglobin (left) and *P. marinus* (eukaryotic) hemoglobin (right) share just 8% sequence identity, but their overall fold and function is identical.

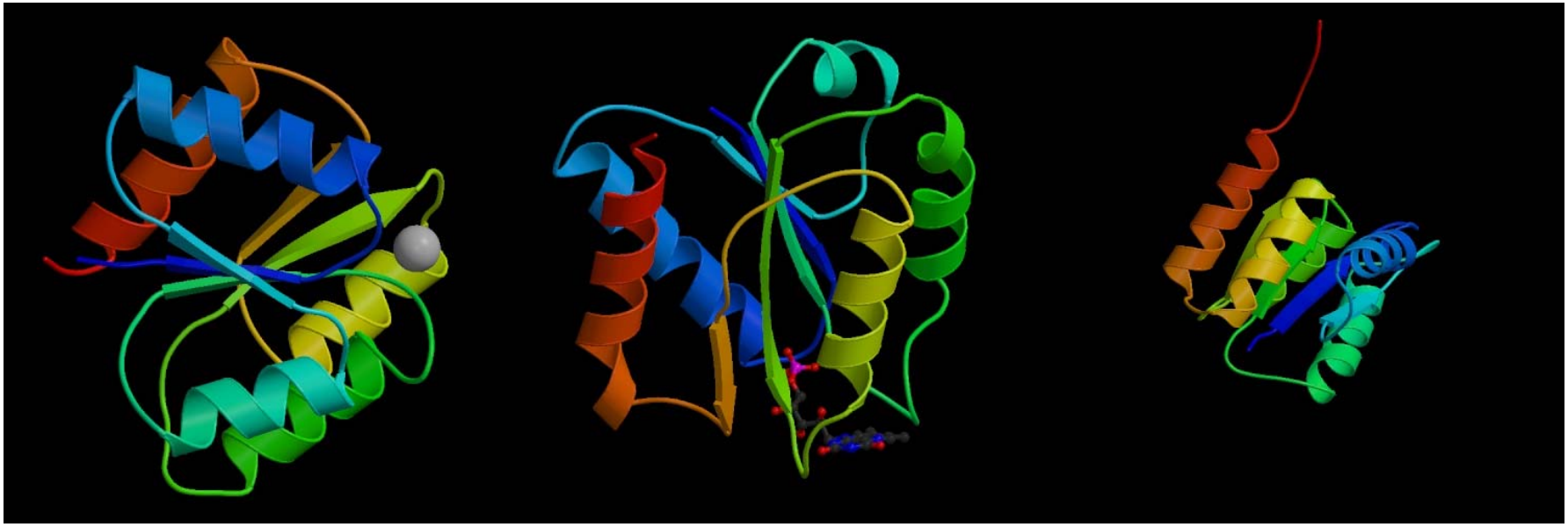
Same Structure Different Function - Alpha/beta proteins characterized as different superfamilies



1ymv 1fla 1pdo



Example – Same Structure Different Function



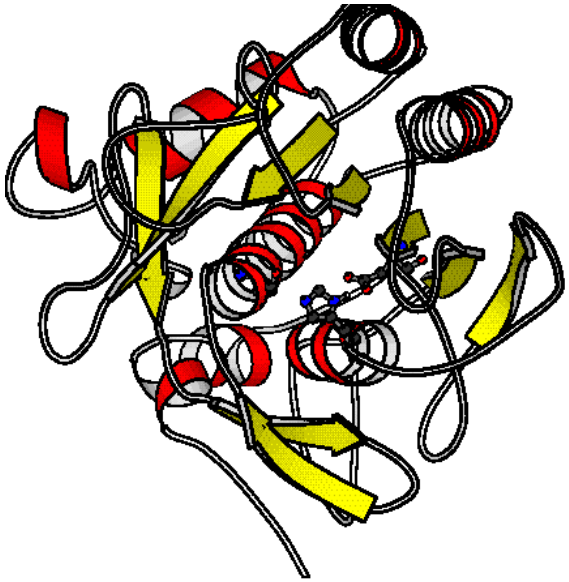
1ymv
CheY
Signal Transduction

1fla
Flavodoxin
Electron Transport

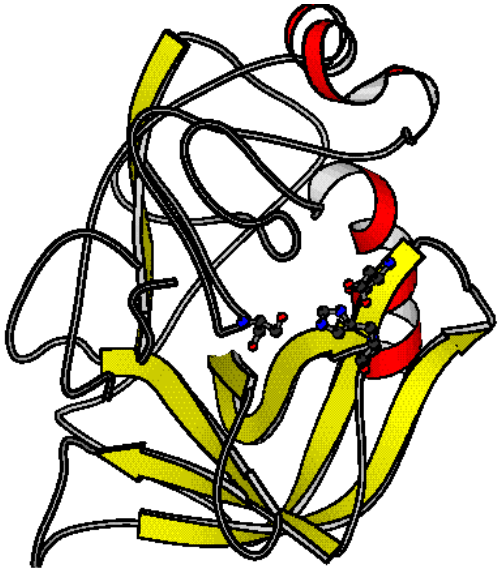
1pdo
Mannose Transporter

Less than 15% sequence identity

Convergent Evolution



a. Subtilisin EC 3.4.21.62



b. Chymotrypsin EC 3.4.21.1

Subtilisin and chymotrypsin are both serine endopeptidases. They share no sequence identity, and their folds are unrelated. However, they have an identical, three-dimensionally conserved Ser-His-Asp catalytic triad, which catalyses peptide bond hydrolysis. These two enzymes are a classic example of convergent evolution.

```
          150                                                       200
Ilk____PSS   .......... .......... ........CC ....CEEEHH HHCCCCCCEE
Ilk____Seq   .......... .......... ........FK ....QLNFLT KLNENHSGEL
------------                                 -+     +L-+++ KL-+---GE-
1fmk--_Seq   KHADGLCHRL TTVCPTSKPQ TQGLAKDAWE IPRESLRLEV KLGQGCFGEV
1fmk--_SS    HCCCCCCCCC CEECCCCCCC CCCCCCCCCE CCHHHEEEEE EEEECCCEEE
                                                          * *  *


           200                                                      250
Ilk____PSS   EEEECCCCE. EEEEEEECCC CCCCCHHHHH HHHHHHHHHC CCCEEEEEEE
Ilk____Seq   WKGRWQGND. IVVKVLKVRD WSTRKSRDFN EECPRLRIFS HPNVLPVLGA
------------ W+G+W-G+-  +-+K+LK-    +T+++-+F- +E---++-++ H++++-++++
1fmk--_Seq   WMGTWNGTTR VAIKTLKP.. .GTMSPEAFL QEAQVMKKLR HEKLVQLYAV
1fmk--_SS    EEEEECCCEE EEEEEECC.. .CCCCHHHHH HHHHHHHHCC CCCECCEEEE
                         *                    *


           250                                                      300
Ilk____PSS   EECCCCEEEE EEHHHHCCCC HHHHHHCCCC CCCCHHHHHH HHHHHHHHHH
Ilk____Seq   CQSPPAPHPT LITHWMPYGS LYNVLHEGTN FVVDQSQAVK FALDMARGMA
------------ ++++P   -- ++T--M++GS L-++L-+-T+ --+--+Q-V+ +A+++A+GMA
1fmk--_Seq   VSEEP...IY IVTEYMSKGS LLDFLKGETG KYLRLPQLVD MAAQIASGMA
1fmk--_SS    ECCCC...EE EEEECCCCCE HHHHHCCCCC CCCCHHHHHH HHHHHHHHHH


           300                                                      350
Ilk____PSS   HHHCCCCCEE CCCCCCCCEE ECCCCEEEEC CCCCEEECCC CCCCCCCCCC
Ilk____Seq   FLHTLEPLIP RHALNSRSVM IDEDMTARIS MADVKFSFQC PGRMYAPAWV
------------ ++++---  - ---L-+++++ ++E+-+++++ ---+--         +---W-
1fmk--_Seq   YVERMNY..V HRDLRAANIL VGENLVCKVA DFGLAR.... ....FPIKWT
1fmk--_SS    HHHHHCC..C CCCCCHHHEE EECCCEEEEC CCCCCC.... ....CCHHHC
                        *    *               *
                       Cat. Loop
           350                                                      400
Ilk____PSS   HHHHHHCCCC CCCCEEEEEE EEHHHHHHHH H.CCCCCCCC CHHHHHHHHH
Ilk____Seq   APEALQKKPE DTNRRSADMW SFAVLLWELV T.REVPFADL SNMEIGMKVA
------------ APEA++++-    ---++D+W SF++LL+EL+ T -+VP+-++ +N-E+-++V
1fmk--_Seq   APEAALYGR. ..FTIKSDVW SFGILLTELT TKGRVPYPGM VNREVLDQV.
1fmk--_SS    CHHHHHHCC. ..CCHHHHHH HHHHHHHHHH CCCCCCCCCC CHHHHHHHH.
             ***
```

Example: Same Fold but Not Function

- "Integrin-linked kinase" (Ilk) is a novel protein kinase fold with strong sequence similarity to known structures (Hannigan et al. 1996 *Nature* 379, 91-96)
- Aligns to Src kinases with BLAST e-value of 10^{-19} and 27% identity (alignment shown is to a known Src kinase structure)
- Several key residues are conserved, but residues important to catalysis, including catalytic Asp, are missing
- Recent experimental evidence suggests that Ilk lacks kinase activity (Lynch et al. 1999 *Oncogene* 18, 8024-8032)

Non-Redundant Sets: Sequences

- NR dataset (NCBI) - All non-redundant GenBank CDS translations+RefSeq Proteins+PDB+SwissProt+PIR+PRF
- Refseq (NCBI) – Annotated
- CDhit http://bioinformatics.org/cd-hit/ - popular algorithms for fast clustering of sequences

Non-Redundant Sets: Sequences with Structure

- PDBselect - http://bioinfo.tg.fh-giessen.de/pdbselect/
- Astral http://astral.berkeley.edu/
- Pisces http://dunbrack.fccc.edu/Guoli/PISCES_OptionPage.php
- RCSB PDB queries