Some Specific “Informatics” tools of Bioinformatics

- **Databases**
  - NCBI GenBank - Protein and DNA sequence
  - NCBI Human Map - Human Genome Viewer
  - NCBI Ensembl - Genome browsers for human, mouse, zebra fish, mosquito
  - TIGR - The Institute for Genome Research
  - SwissProt - Protein Sequence and Function
  - ProDom - Protein Domains
  - Pfam - Protein domain families
  - ProSite - Protein Sequence Motifs
  - Protein Data Base (PDB) - Coordinates for Protein 3D structures
  - SCOP Database - Domain structures organized into evolutionary families
  - HSSP - Domain database using Dali
  - CATH Database
  - FlyBase
  - WormBase
  - PubMed / MedLine

- **Sequence Alignment Tools**
  - BLAST
  - Clustal MSAs
  - FASTA
  - PSI-Blast
  - Hidden Markov Models

- **3D Structure Alignments / Classifications**
  - Dali
  - CATH
  - SCOP
  - VAST
  - PRISM
Protein Domains

“Independent Folding Units”

50 - 350 residues
Mean size - 125 residues

Alpha folds; Beta Folds;
Alpha+Beta Folds; Alpha/Beta Folds
COG 272, BRCT family
Sequence Similarity May Miss Functional Homologies Which Can Be Detected by 3D Structural Analysis

Adapted from Chris Sander
Structural Validation of Homology

Adenylate Kinase

Guanylate Kinase

19% Seq ID
Z = 12.2
Classification of Protein Folds
- SCOP
- CATH
- DALI / FSSP
Most proteins in biology have been produced by the duplication, divergence and recombination of the members of a small number of protein families.
Domain Combinations in Genome Sequences

In bacteria close to
1/3 of proteins consist of one domain and
2/3 consist of two or more domains.

In eukaryotes close to
1/4 of proteins consist of one domain and
3/4 consist of two or more domains.

Average Domain Size: 170 residues
courtesy of C. Chothia
**SCOP** the Structural Classification Of Proteins database
This contains all proteins, and protein domains, of known structure classified in terms of their structure and evolutionary relationships.

**SUPERFAMILY** This database contains:
(a) hidden Markov models (HMMs) of all the proteins and protein domains in SCOP
(b) a list of the matches made by these HMMs to the sequences of 56 genomes classified by family.

http://stash.mrc-lmb.cam.ac.uk/SUPERFAMILY/
courtesy of C. Chothia
SCOP - Protein Fold Hierarchy
Manually Curated Database of Domain Structures

Class - 5
Fold - ~500
Superfamily - ~ 700
Family ~ 1000

Family - domains with common evolutionary origin

Homology: Derived by evolutionary divergence
Five Principal Fold Classes

All $\alpha$ folds
All $\beta$ folds
$\alpha + \beta$ folds
$\alpha / \beta$ folds
small irregular folds
SUPERFAMILY matches to genome sequences

courtesy of C. Chothia
## SUPERFAMILY Results for *Buchnera* and Human Genome Sequences

<table>
<thead>
<tr>
<th></th>
<th>Buchnera</th>
<th>Humans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sequences</td>
<td>564</td>
<td>23867</td>
</tr>
<tr>
<td>Sequences matched by SUPERFAMILY</td>
<td>410</td>
<td>12616</td>
</tr>
<tr>
<td>Coverage of genome</td>
<td>61%</td>
<td>41%</td>
</tr>
<tr>
<td>Number of matched domains</td>
<td>609</td>
<td>22548</td>
</tr>
<tr>
<td>Number of families</td>
<td>277</td>
<td>648</td>
</tr>
<tr>
<td>Mean family size</td>
<td>2.2</td>
<td>35</td>
</tr>
<tr>
<td>Number of large families that form half the matched domains</td>
<td>37</td>
<td>17</td>
</tr>
</tbody>
</table>

courtesy of C. Chothia
SUPERFAMILY Results for
*Buchnera* and Human Genome Sequences:
Top Five Domain Families

*Buchnera*
- P-loop containing nucleotide triphosphate hydrolases
- Nucleic acid binding proteins
- NAD-binding Rossman domains
- Nucleotidyl transferases
- Class II aaRS synthetases

*Humans*
- Classic zinc fingers
- Immunoglobulin superfamily
- P-loop containing nucleotide triphosphate hydrolases
- EGF/Laminin
- Cadherin

courtesy of C. Chothia
Eukaryotes

courtesy of C. Chothia
Bacteria

courtesy of C. Chothia
CATH Protein Domain Database
Partially Automatic Fold
Classification

CATH is a hierarchical classification of protein domain structures, which clusters proteins at four major levels, Class(C), Architecture(A), Topology(T) and Homologous superfamily (H).


Class, derived from secondary structure content, is assigned for more than 90% of protein structures automatically.

Architecture, which describes the gross orientation of secondary structures, independent of connectivities, is currently assigned manually.

The topology level clusters structures according to their topological connections and numbers of secondary structures.

The homologous superfamilies cluster proteins with highly similar structures and functions. The assignments of structures to topology families and homologous superfamilies are made by sequence and structure comparisons.
Representations of Protein Structures

a - full atom

b,c - strands / helices

d - Topology diagrams
Structural Alignment of Two Globins
Structure-based Sequence Alignments

Alignment of Individual Structures

Fusing into a Single Fold “Template”

Hb  VLSPADKTNVKAAGKVGHAEGYGAEEALERMFLSFPTTKTYFPHF-DLS-------HGSAQVKGHGKKVADALTNAV
  ||||...|||...|||...|||...|||...|||...|||...|||...|||...|||...|||...
Mb  VLSSEGQVLHVVAKVEADVAGHGQDILRLFKSHPETLEKFDRIAFLKLKTEAMKASEDLKKGVTPLALTGA

Hb  AHVD-DMPNALSDLHALHKLRLVDPVNKLLSHCLLVTLAHLPAEFTPAVHASLDKFLASVTVLTSKYR-------
  |||...|||...|||...|||...|||...|||...|||...|||...|||...|||...|||...|||...|||...
Mb  KK-KGHHEAELKPLAQSHATKHKPIKYLEFISEAIHVHLHSRHPGDFGADAQGAQMNKALELFKDIAAKYLQG

Elements: Domain definitions; Aligned structures, collecting together Non-homologous Sequences; Core annotation

Previous work: Remington, Matthews ‘80; Taylor, Orengo ‘89, ‘94; Artymiuk, Rice, Willett ‘89; Sali, Blundell, ‘90; Vriend, Sander ‘91; Russell, Barton ‘92; Holm, Sander ‘93; Godzik, Skolnick ‘94; Gibrat, Madej, Bryant ‘96; Falicov, F Cohen, ‘96; Feng, Sippl ‘96; G Cohen ‘97; Singh & Brutlag, ‘98
Some Similarities are Readily Apparent others are more Subtle

Easy:
Globins
125 res., ~1.5 Å

Tricky:
Ig C & V
85 res., ~3 Å

Very Subtle: G3P-dehydrogenase, C-term. Domain
>5 Å
Automatically Comparing Protein Structures

Given 2 Structures (A & B),

2 Basic Comparison Operations
1. **Find an Alignment** between A and B based on their 3D coordinates

2. Given an alignment optimally **SUPERIMPOSE** A onto B
   - Find Best R & T to move A onto B
Distance Matrices
Provide a 2D Representation of the 3D Structure
Explain Concept of Distance Matrix on Blackboard

N x N distance matrix

Antiparallel beta strands
Parallel beta strands
Helices

N dimensional space

Metric matrix

\[ M_{ij} = D_{ij}^2 - D_{io}^2 - D_{jo}^2 \]

M Eigenvectors (M = 3 for 3D structure)
DALI: Protein Structure Comparison by Alignment of Distance Matrices

• Generate Cα-Cα distance matrix for each protein A and B
• Decompose into elementary contact patterns; e.g. hexapeptide-hexapeptide submatrices
• Systematic comparisons of all elementary contact patterns in the 2 distance matrices; similar contact patterns are stored in a “pair list”
• Assemble pairs of contact patterns into larger consistent sets of pairs (alignments), maximizing the similarity score between these local structures
• A Monte-Carlo algorithm is used to deal with the combinatorial complexity of building up alignments from contact patterns
• Dali Z score - number of standard deviations away from mean pairwise similarity value
Figure 2. How to maximize the structural overlap of 2 proteins? The algorithm can be followed from top to bottom in 3 schematic representations: left, 3D chain trace; middle, 2D distance matrices; right, 1D sequence alignments. Two topologically different 3-stranded β-sheet proteins (idealized) are being compared. The structurally equivalent fragments are labeled a, b, c in one protein (protein 1) and a', b', c', respectively, in the other (protein 2). In the distance matrices, similar contact patterns are filled with the same pattern, boxes on the main diagonal correspond to intra-fragment distances and off-diagonal boxes to inter-fragment distances. The similarity score of an alignment is calculated from the pairwise differences of all equivalent elements of the 2 distance matrices. Top row: an alignment is initiated from matching contact patterns that equivalence the hexapeptide–hexapeptide pair a-b with a'-b'. Second row: fragments b and b', which are part of the previous alignment, are used to look for additional fragments by which to extend the alignment. The fragments c and c' are identified because the contact patterns (b,c) and (b',c') are similar. Third row: (a,b)-(a',b') and (b,c)-(b',c') are merged into the alignment (a,b,c)-(a',b',c'). Although the search builds on substructures, the similarity score of the alignment depends on the fitness of each of the equivalence fragments in the context of all others in an alignment. In this sense, there is a co-operative effect built into the optimization. Bottom row: the final agreement of hexapeptide–hexapeptide contact patterns after the removal of insertions/deletions and reordering of the aligned segments b' and c' in the 2nd protein. The resulting 1D alignment is at the lower right. The comparison of contact patterns is independent of sequence gaps or shuffling of segment order and can also identify matches with reversed chain direction.
Dali Domain Dictionary
Deitman, Park, Notredame, Heger, Lappe, and Holm
Nucleic Acids Res. 29: 5557 (2001)

• Dali Domain Dictionary is a numerical taxonomy of all known domain structures in the PDB

• Evolves from Dali / FSSP Database

• Dali Domain Dictionary Sept 2000
  ◊ 10,532 PDB entries
  ◊ 17,101 protein chains
  ◊ 5 supersecondary structure motifs (attractors)
  ◊ 1375 fold types
  ◊ 2582 functional families
  ◊ 3724 domain sequence families
Explain Concept of Distance Matrix on Blackboard

N x N distance matrix

N dimensional space

Metric matrix

\[ M_{ij} = D_{ij}^2 - D_{io}^2 - D_{jo}^2 \]

Eigenvectors of metric matrix

Principal component analysis
A Global Representation of Protein Fold Space  

Database of 498 SCOP “Folds” or “Superfamilies”

The overall pair-wise comparisons of 498 folds lead to a 498 x 498 matrix of similarity scores $S_{ij}$s, where $S_{ij}$ is the alignment score between the ith and jth folds.

An appropriate method for handling such data matrices as a whole is metric matrix distance geometry. We first convert the similarity score matrix $[S_{ij}]$ to a distance matrix $[D_{ij}]$ by using $D_{ij} = S_{max} - S_{ij}$, where $S_{max}$ is the maximum similarity score among all pairs of folds.

We then transform the distance matrix to a metric (or Gram) matrix $[M_{ij}]$ by using $M_{ij} = D_{ij}^2 - D_{i0}^2 - D_{j0}^2$

where $D_{i0}$, the distance between the ith fold and the geometric centroid of all $N = 498$ folds. The eigen values of the metric matrix define an orthogonal system of axes, called factors. These axes pass through the geometric centroid of the points representing all observed folds and correspond to a decreasing order of the amount of information each factor represents.
A Global Representation of Protein Fold Space