

# Protein Structures: Components and Analysis

## BME 110: Computational Biology Tools

5/24/2007

1

© David Bernick, 2007

## Amino acids -- properties and symbols

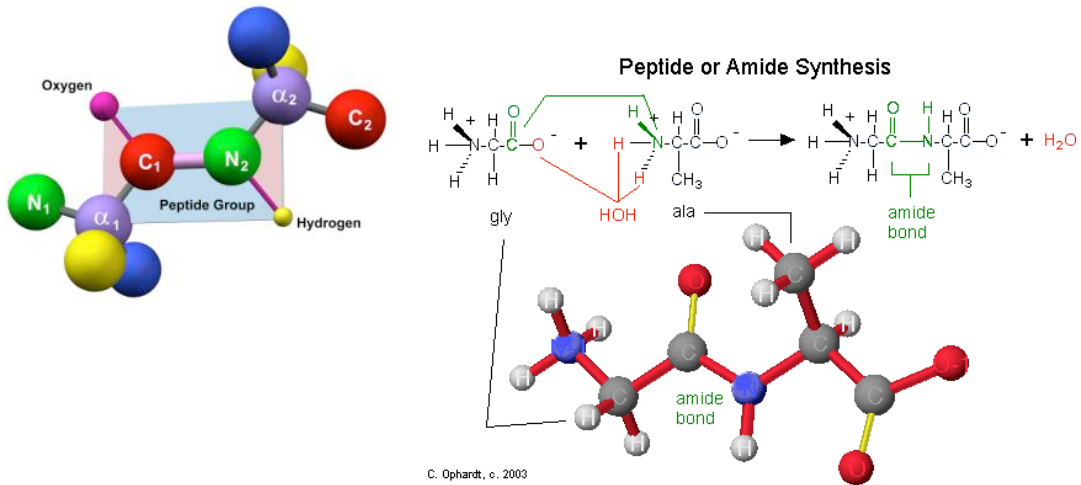
Amino acid				Amino acid			
Alanine	A	Ala	Neutral Non-polar	Methionine	M	Met	Neutral Non-polar
Cysteine	C	Cys	Neutral Slightly Polar	Asparagine	N	Asn	Neutral Polar
Aspartate	D	Asp	Acidic Polar	Proline	P	Pro	Neutral Non-polar
Glutamate	E	Glu	Acidic Polar	Glutamine	Q	Gln	Neutral Polar
Phenylalanine	F	Phe	Neutral Non-polar	Arginine	R	Arg	Basic Polar
Glycine	G	Gly	Neutral Non-polar	Serine	S	Ser	Neutral Polar
Histidine	H	His	Basic Polar	Threonine	T	Thr	Neutral Polar
Isoleucine	I	Ile	Neutral Non-polar	Valine	V	Val	Neutral Non-polar
Lysine	K	Lys	Basic Polar	Tryptophan	W	Trp	Neutral Slightly polar
Leucine	L	Leu	Neutral Non-polar	Tyrosine	Y	Tyr	Neutral Polar

5/24/2007

2

© David Bernick, 2007

# the peptide bond



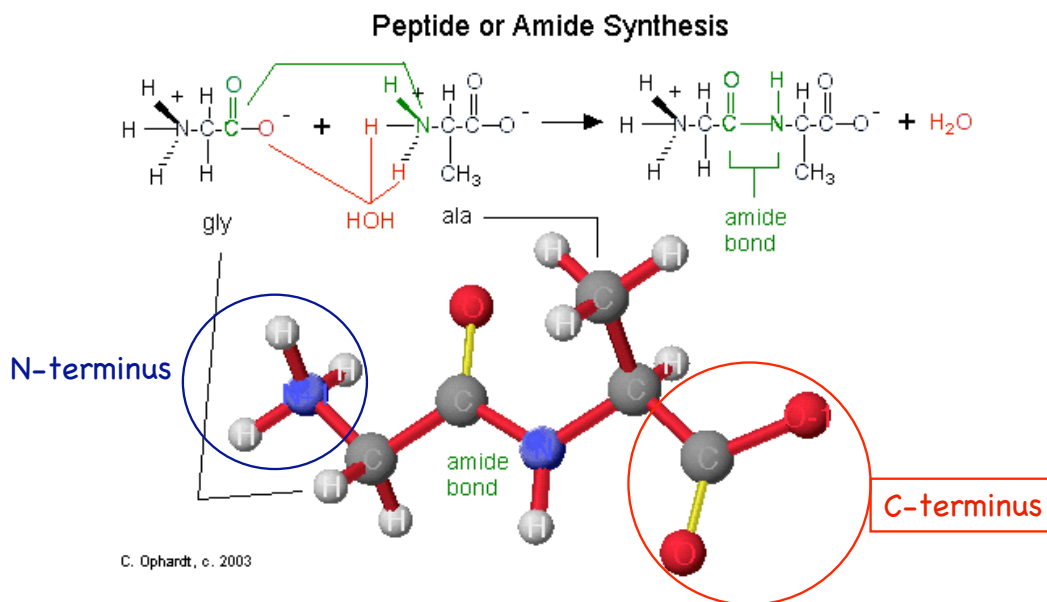
<http://www.codefun.com/Images/Genetic/tRNA/image004.jpg>

5/24/2007

3

© David Bernick, 2007

## Peptides and the peptide bond

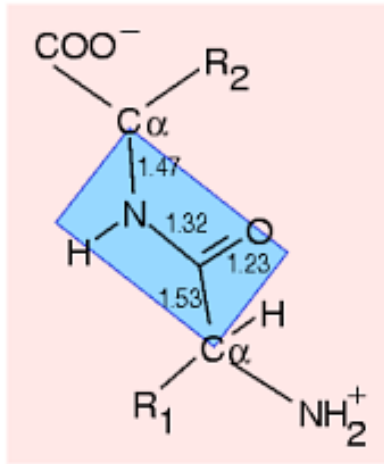


5/24/2007

4

© David Bernick, 2007

# peptide bond distances



|x-H| ~ 1.05 Å  
|N-C $\alpha$ | ~ 1.45 Å  
|N-C| ~ 1.37 Å  
|C-O| ~ 1.23 Å  
|C-C $\alpha$ | ~ 1.49 Å

from Pauling, L. 1951

## primary structure -- 1TIM

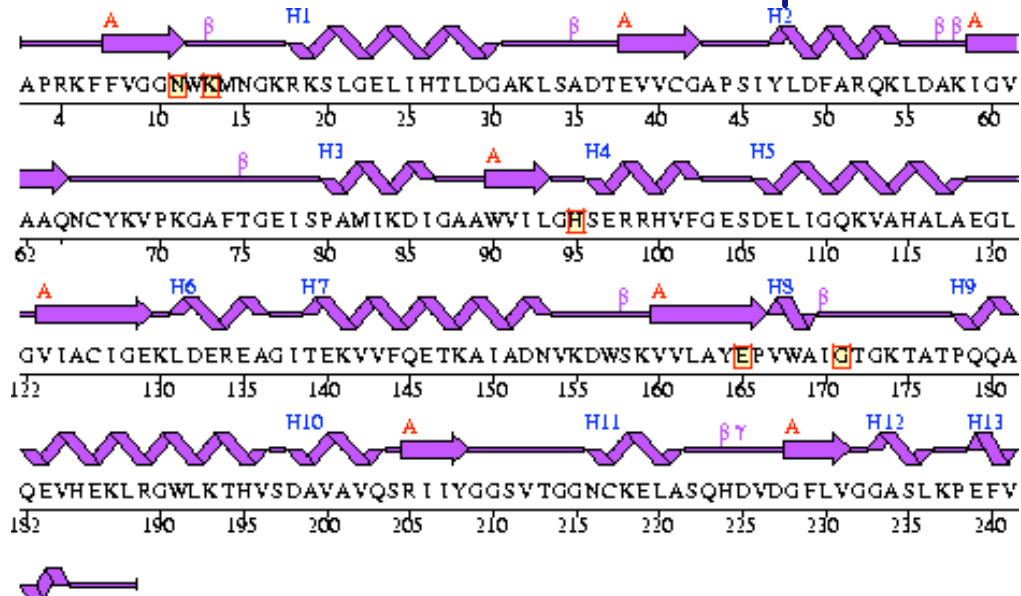
- primary -- sequence

>1TIM:A|PDBID|CHAIN|SEQUENCE

```
APRKFFVGGNWKMNKGRKSLGELIHTLDGAKLSADTEVVCGAPSIYLDFAEQKLDK  
IGVAAQNCYKVPKGAFTGEISPAMIKDIGAAWVILGHSERRHVFGEDELIGQKVAH  
ALAEGLGVIACIGEKLDEREAGITEKVVFOETKAIADNVKDWKVVLAYEPVWAIGT  
GKTATPQQAQEVHEKLRGWLKTHVSDAVAVQSRIIYGGSVTGGNCKELASQHDVDGF  
LVGGASLKPEFVDIINAKH
```

# secondary structure - 1TIM

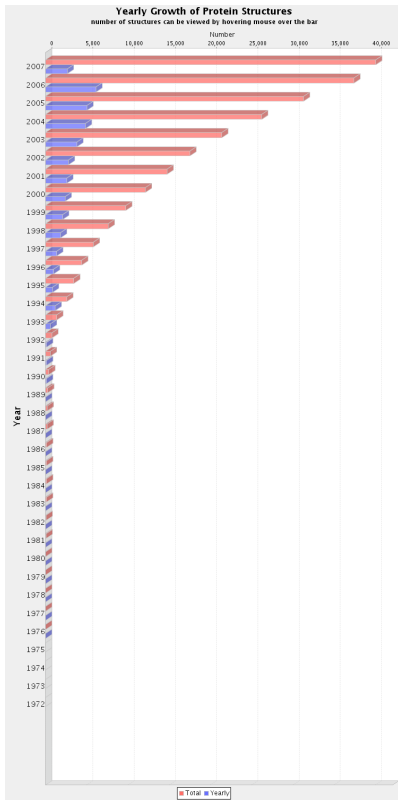
## helix, strand or loop



<http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum>

# tertiary structure -- 1TIM





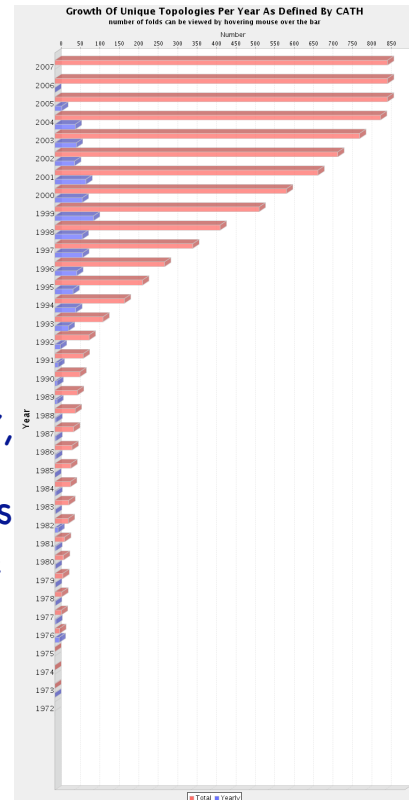
5/24/2007

# Protein Data Bank

www.pdb.org

- as of 5/23/2007, there are 43633 stored structures
- with 1054 unique folds(SCOP)

9



© David Bernick, 2007

## structures

<http://www.pdb.org/pdb/explore.do?structureId=1TIM>



type X-RAY DIFFRACTION

Resolution[Å]	R-Value	R-Free	Space Group
2.50	n/a	n/a	P 2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>

- **Banner, D.W., Bloomer, A., Petsko, G.A., Phillips, D.C., Wilson, I.A.**  
Atomic coordinates for triose phosphate isomerase from chicken muscle.  
*Biochem.Biophys.Res.Commun.*  
v72 pp.146-155 , 1976

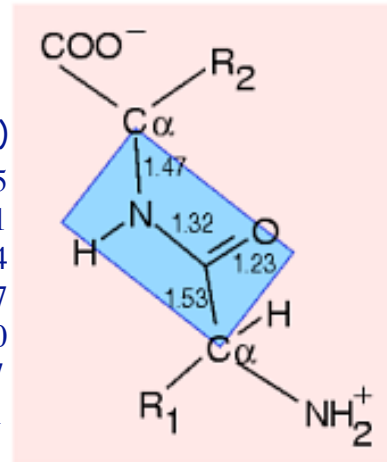
5/24/2007

10

© David Bernick, 2007

# PDB structure records (1TIM)

record	atom	residue	coordinates (x, y, z)
ATOM	1 N	ALA A 1	43.240 11.990 -6.915
ATOM	2 CA	ALA A 1	43.888 10.862 -6.231
ATOM	3 C	ALA A 1	44.791 11.378 -5.094
ATOM	4 O	ALA A 1	44.633 10.992 -3.937
ATOM	5 CB	ALA A 1	44.722 10.051 -7.240
ATOM	6 N	PRO A 2	45.714 12.244 -5.497
ATOM	7 CA	PRO A 2	46.689 12.815 -4.561



$$|C\alpha_{ALA}, N_{ALA}| = \sqrt{(\bar{X})^2 + (\bar{Y})^2 + (\bar{Z})^2}$$

$$= \sqrt{(43.240 - 43.888)^2 + (11.990 - 10.862)^2 + (-6.915 + 6.231)^2}$$

$$\approx 1.4697$$

5/24/2007

11

© David Bernick, 2007

## Why Examine Protein Structures?

- Structure more conserved than sequence
  - Similar folds often share similar function
  - Remote similarities may only be detectable at structure level
- Interpreting experimental data
  - Locating sites of interesting mutations
  - Locating splice sites
- Designing experiments
  - *In silico* mutagenesis

# Structure Analysis

- Identify interesting sites on protein
- Measure distances, angles, etc.
- Examine surface properties (shape, charge)
- Compare two structures
  - Homologs
  - Mutants
  - With and Without Ligands

## Comparing Protein Structures

- Defined alignment
  - Mutant-wildtype, model-native, two different conformations.
  - Unique solution exists -- we know the true alignment
- Derived alignment
  - Unknown query
  - Known parent (assumed homolog)
  - calculate a computationally 'Optimal' alignment
  - infer annotation from parent to query

# What do we want from an Alignment?

- 'Optimal alignment'
  - Important parts of protein should associate (align) with each other
    - Catalytic residues and their position in 3-space
    - Important structures (hinges, binding sites)
    - Protein interface residues and their position in 3-space
    - History
  - Natural selection only selects for successful Function
  - Alignments are assumed to be sequential
- Sequence alignments can be improved when we have structural information
  - No unique solution (more residues or closer match?)
  - Structural alignment implies a sequence alignment

## Tools and Databases

- Structure Databases and search tools
  - NCBI Structure (VAST and MMDB)
    - <http://www.ncbi.nlm.nih.gov/Structure/>
    - Molecular Modeling Database
      - Experimentally derived structures from PDB (not theoretical)
  - FSSP (DALI)
    - <http://www.ebi.ac.uk/dali/>
    - <http://ekhidna.biocenter.helsinki.fi/dali/start>
    - Families of Structurally Similar Proteins
      - Maintains database of Protein Neighbors organized by PDB code
  - CE
    - <http://cl.sdsc.edu/>
    - Combinatorial Extension
      - Maintains database of Protein Neighbors by PDB code

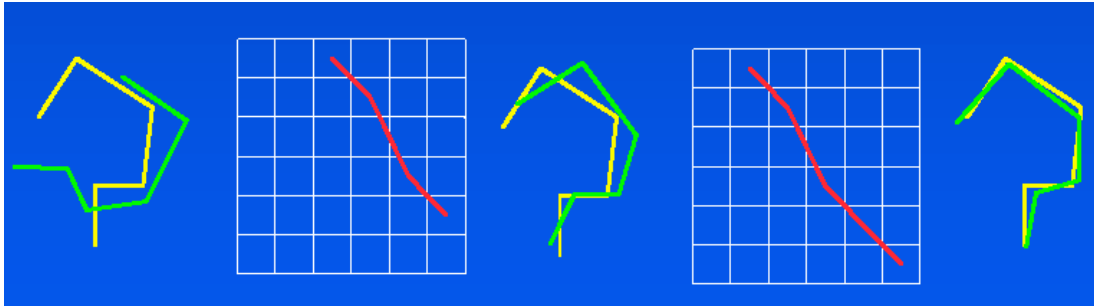
# Tools and Databases(2)

- Structure classification by domain
  - Classifications based on Secondary structure
  - SCOP Structural Classification of Proteins
    - <http://scop.berkeley.edu/>, Alexsi Mursin et al.
    - Last release 18 January 2005
  - CATH Class Architecture Topology Homology
    - <http://www.cathdb.info/>, Automated and manual classification
    - Last release Jan 2007, v. 3.1.0
- CEMC - Multiple Structure Alignment
  - <http://bioinformatics.albany.edu/~cemc/>

## How Structure alignments work

- Methods
  - Structural
  - DALI
  - VAST
- Structure similarity measures
  - RMSD
  - Pvalues

# Iterative Dynamic Programming

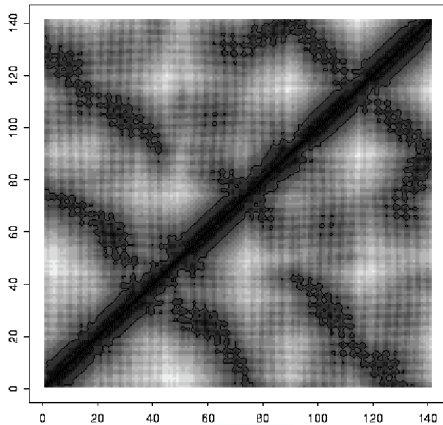


- Algorithm:
  1. Make an initial guess for the superposition
  2. Calculate all pairwise CA-CA distances and generate a scoring matrix.
  3. Find optimal alignment according to this scoring matrix by dynamic programming.
  4. Re-superimpose structures using this alignment
  5. Repeat step 2-4 until convergence.
- No guarantee of optimal solution, final result depends on the initial alignment selected.
- Structural: Subbiah et al, 1993 Curr. Biol 3:141)

## Structural Alignment

- Many methods other than dynamic programming are used.
- Most methods use some sort of heuristics to speed things up and make good initial guesses:
  - Sheba Sequence alignment
  - Mammoth Local structure alignment
  - VAST aligns secondary structure element vectors
  - DALI Distance matrix alignment

# Distance Matrix ALignment

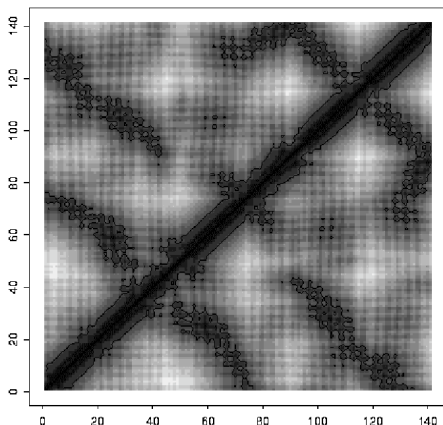


Myoglobin

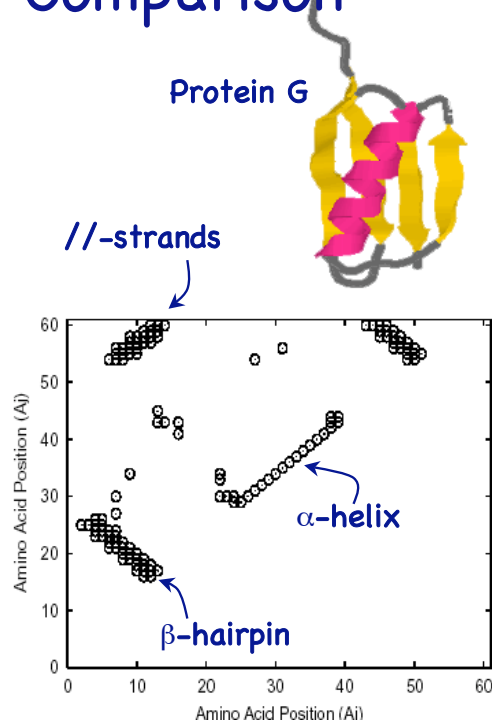


- Matrix of all pair-wise distances
- Characteristic patterns:
  - Main diagonal runs correspond to helix (i.e local contacts)
  - Hairpins - start on main diagonal, run perpendicular
  - Parallel pairs run parallel to main diagonal
  - Others are long range contacts.
- Converts 3D alignment problem to a 2D problem.
  - Find best subset of rows and columns such that the distance matrices of two proteins are optimally similar

# Contact Map Comparison



Myoglobin

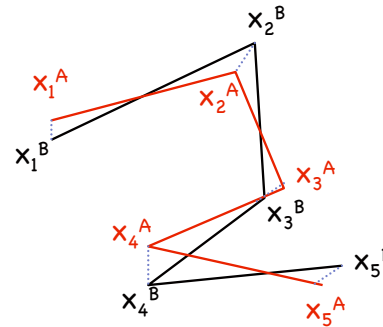


# Similarity Measures: RMSD

- RMSD = root mean square deviation

$$\sqrt{\langle \|x_i^A - x_i^B\|^2 \rangle}$$

1. Superimpose optimally
2. Pair up residues
3. Calculate RMSD

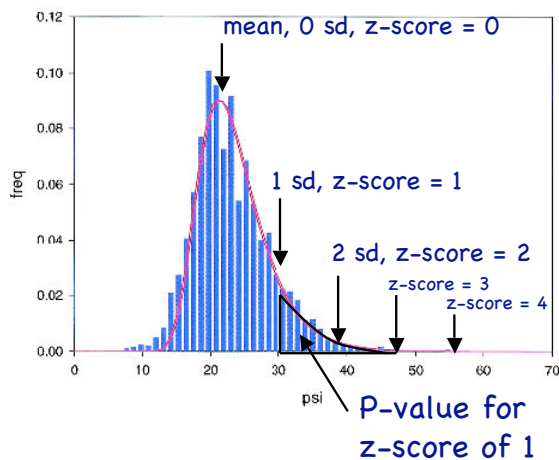


Sensitive to outliers

Depends on number of pairs compared

A better measure is the significance of this RMSD for similar sized matches

## Z-scores & P-values



- Z-score: # of standard deviations above the mean:
  - $\pm 1$  sd  $\sim 66\%$
  - $\pm 2$  sd  $\sim 95\%$
  - If we have a histogram, we can just count; Or integrate a function fitted to the histogram.
- P-value
  - Probability of obtaining  $\geq$  this score under the null model (normally distributed data -- "by chance")

Histogram of scores for random matches

