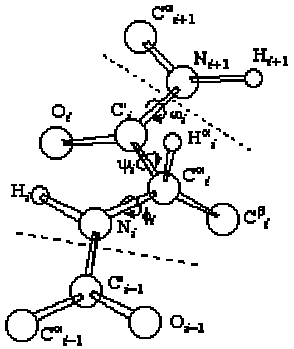
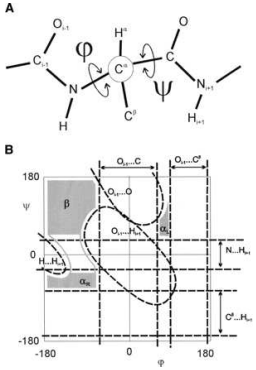
Polypeptide backbone (or the main chain)



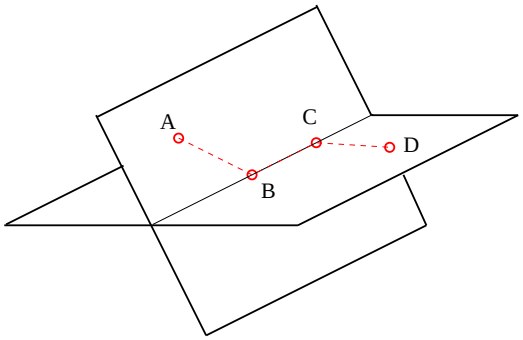
[IUPAC-IUB Commission on Biochemical Nomenclature, Abbreviations and Symbols for the Description of the Conformation of Polypeptide Chains. Eur. J. Biochem., 1969, 17, 193-201]

Ramachandran steric map



[Ho, K.H., Thomas, A. and Brasseur, R., Protein Science, 2003, 12:2508-2522]

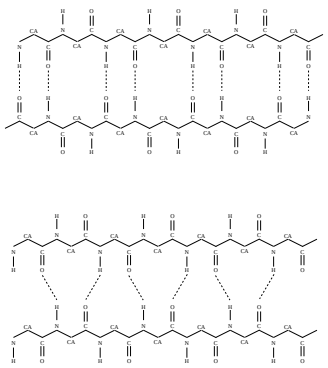
Dihedral angle



Protein Data Bank entry (extract)

```
COMPND    TRIOSE PHOSPHATE ISOMERASE (E.C.5.3.1.1)
SOURCE    CHICKEN (GALLUS GALLUS) BREAST MUSCLE
AUTHOR    D.W.BANNER,A.C.BLOOMER,G.A.PETSKO,D.C.PHILLIPS,
AUTHOR   2 I.A.WILSON
  :
JRNL        AUTH   D.W.BANNER,A.C.BLOOMER,G.A.PETSKO,D.C.PHILLIPS,
JRNL        AUTH 2 I.A.WILSON
JRNL        TITL   ATOMIC COORDINATES FOR TRIOSE PHOSPHATE ISOMERASE
JRNL        TITL 2 FROM CHICKEN MUSCLE
JRNL        REF    BIOCHEM.BIOPHYS.RES.COMM.     V.  72   146 1976
JRNL        REFN   ASTM BBRCA9  US ISSN 0006-291X                146
  :
REMARK   2 RESOLUTION. 2.5 ANGSTROMS.
  :
SEQRES   1 A  247  ALA PRO ARG LYS PHE PHE VAL GLY GLY ASN TRP LYS MET
SEQRES   2 A  247  ASN GLY LYS ARG LYS SER LEU GLY GLU LEU ILE HIS THR
  :
ATOM      1  N   ALA A   1      43.240  11.990  -6.915  1.00  0.00
ATOM      2  CA  ALA A   1      43.888  10.862  -6.231  1.00  0.00
ATOM      3  C   ALA A   1      44.791  11.378  -5.094  1.00  0.00
ATOM      4  O   ALA A   1      44.633  10.992  -3.937  1.00  0.00
ATOM      5  CB  ALA A   1      44.722  10.051  -7.240  1.00  0.00
ATOM      6  N   PRO A   2      45.714  12.244  -5.497  1.00  0.00
ATOM      7  CA  PRO A   2      46.689  12.815  -4.561  1.00  0.00
ATOM      8  C   PRO A   2      46.042  13.601  -3.411  1.00  0.00
ATOM      9  O   PRO A   2      46.030  13.141  -2.267  1.00  0.00
  :
```

Hydrogen bonds in β-sheets



DSSP summary codes

H 4-helix (α-helix)
B residue in isolated β-bridge
E extended strand, participates in β-ladder
G 3-helix
I 5-helix
T H-bonded turn
S bend

Crambin (1CRN)

```
TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIIIPGATCPGDYAN
 EE SSHHHHHHHHHHHHTT  HHHHHHHHS EE SSS   TTS
```


DSSP

Hydrogen bond energy

$$E = q_1 q_2 \left(\frac{1}{d(ON)} + \frac{1}{d(CH)} - \frac{1}{d(OH)} - \frac{1}{d(CN)} \right) \times f$$

Antiparallel bridge:

[hbond(i,j) and hbond(j,i)]
or
[hbond(i-1,j+1) and hbond(j-1,i+1)]

Parallel bridge:

[hbond(i-1,j) and hbond(j,i+1)]
or
hbond(j-1,i) and hbond(i,j+1)]

Protein stability

- good stereochemistry; no steric clashes;
- buried charged atoms must be paired;
- enough hydrophobic surface must be buried, and the interior must be sufficiently densely packed, to provide thermodynamic stability.

Modular proteins

- multi-domain proteins, often with many copies of related domains;
- domains recur in many proteins in different structural contexts.

Is the similarity significant, or could it be due to chance?

Even if two proteins are unrelated, we would expect some similarity simply by chance.

Is the alignment score significantly higher than random?

Align random permutations of the sequences, and find the mean and standard deviation of the resulting distribution.

The z-score reflects the significance of a global similarity score.

$$z\text{-}score = \frac{score - mean}{standard\ deviation}$$

Larger values imply greater significance.

e-values and p-values

The expected number of HSPs with a score of at least S is given by the formula:

$$E = Kmne^{-\lambda S}$$

Doubling the length of the query sequence (m) or the size of the database (n) should double the number of HSPs.

To obtain score $2x$, score x must be obtained twice in a row.
So one expects E to decrease exponentially with score.

The probability of observing a score $\geq S$ is:

$$1 - \exp(-Kmne^{-\lambda S})$$

This is the p-value.

BLAST

Basic Local Alignment Search Tool

Less accurate than Smith-Waterman, but over 50 time faster.

1. Find ungapped matches of a small fixed length, w, that score at least T.

2. Extend matches in both directions in an attempt to find an alignment with a score exceeding S.

Segment pairs whose scores cannot be improved by extending or trimming are called high scoring pairs (HSPs).

Typical values for w are 3 when aligning proteins and 11 when aligning nucleic acids.

FASTA

k-tuples, strings of length k.

k = 1 - 2 for proteins and 4-6 for nucleic acids.

Construct a look-up table with all k-tuples in the database.

Look up all k-tuples from the query string and mark matching database k-tuples. Sort matches by the difference in their indices (i-j).

Nearby matches on the same diagonal are joined to form an ungapped local alignment region.

Join nearby high scoring regions on different diagonals.

For the best regions, perform dynamic programming in a window around the region.

3D Transformations

Translation

$$T(d_x, d_y, d_z) = \begin{bmatrix} 1 & 0 & 0 & d_x \\ 0 & 1 & 0 & d_y \\ 0 & 0 & 1 & d_z \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Scaling

$$S(s_x, s_y, s_z) = \begin{bmatrix} s_x & 0 & 0 & 0 \\ 0 & s_y & 0 & 0 \\ 0 & 0 & s_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

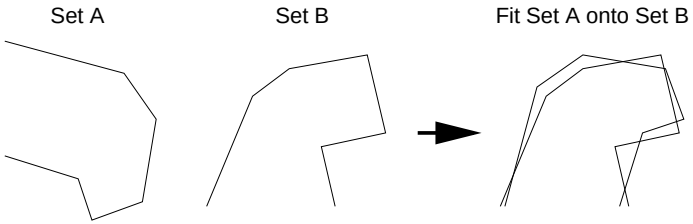
Rotation

$$R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) & 0 \\ 0 & \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R_y(\theta) = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) & 0 \\ 0 & 1 & 0 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R_z(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 & 0 \\ \sin(\theta) & \cos(\theta) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Comparing molecular fragments



- 3-D transformation to map Set A onto Set B
- Root Mean Square (RMS) distance