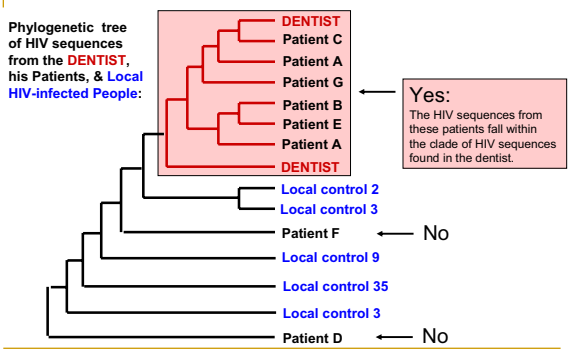
Phylogenetic Trees

Did the *Florida Dentist* infect his patients with HIV?



From Ou et *al.* (1992) and Page & Holmes (1998)

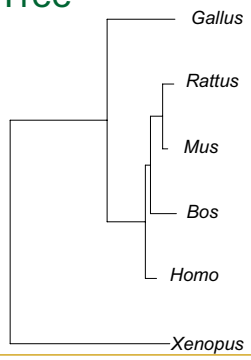
Data for Building Phylogenies

- Characteristics
  - Traits (continuous or discrete)
  - Biomolecular features
  - *character state matrix*
- Numerical distance estimates
  - *distance matrix*

Different Kinds of Trees

- Order of evolution
  - Rooted: indicates direction of evolution
  - Unrooted: only reflects the distance
- Rate of evolution
  - Edge lengths: distance (scaled trees)
    - Assumes constant rate of evolution
  - Unscaled trees

Rooted Tree



Tree building Methods

- Character-based methods
  - Maximum parsimony
  - Maximum likelihood
- Distance-based methods
  - UPGMA
  - NJ
- Will consider only distance-based methods of pyhlogenetics in this course

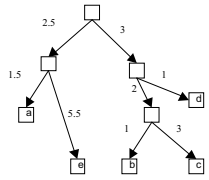
Distance Matrix Methods

- Given a pairwise distace matrix D
- Produce a tree such that the *path distance* between leaves i and j (sum of edge weights in the path between i and j) equals d_{ij}
- Optimize the error between d and D
  - Least square error metric: LSQ
  - LSQ(d,D) = $\Sigma \Sigma (d_{ij} - D_{ij})^2$
  - NP-complete
- Heuristics (usually based on agglomerative (group by group) clustering)
  - UPGMA
  - NJ
  - Both assume additive distances
    - implies that distance is a metric
      - symmetry
      - triangle inequality
      - d(x,y) = 0 iff x = y
      - d(x,y) >= 0

Distance Measures

- Edit distance
- Global/semi-global similarity scores (may be inverted and shifted to get positive distances)

Example Tree and Additive Matrix



	a	b	c	d	e
A	0	10	12	8	7
B		0	4	4	14
C			0	6	16
D				0	12
E					0

There exists a tree with additive distances

Approximating Additive Matrices

In practice, the distance matrix between molecular sequences will not be additive.

An additive tree T whose distance matrix approximates the given one is used.

The methods for exact tree reconstruction provide an inventory for heuristics for tree construction based on approximating additive metrics.

Heuristics give exact results when operating on additive metrics.

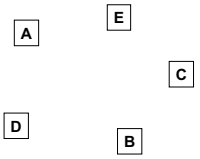
UPGMA

- Unweighted Pair-Group Method with Arithmetic Mean
  - Sokal and Michener 1958
- Agglomerative (group by group) clustering
- Ultrametric tree
  - distances from root to all leaves are equal

UPGMA Step 1

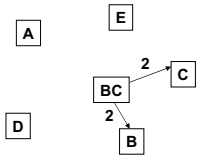
combine B and C

Choose two clusters with minimum distance and combine them



	A	B	C	D	E
A	0	10	12	8	7
B		0	4	4	14
C			0	6	16
D				0	12
E					0

Updating distance matrices



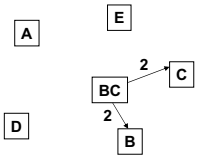
	A	BC	D	E
A	0	11	8	7
BC		0	5	15
D			0	12
E				0

Distance of new cluster to nodes in the cluster is half of original distance

Distance of new cluster to other clusters is weighted mean of individual distances

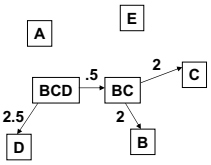
UPGMA step 2

combine BC and D



	A	BC	D	E
A	0	11	8	7
BC		0	5	15
D			0	12
E				0

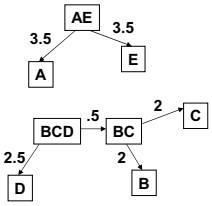
Updating distance matrices



	A	BCD	E
A	0	10	7
BC		0	14
E			0

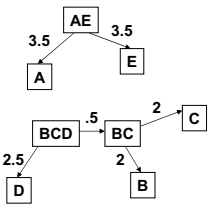
UPGMA step 3

combine A and E



	A	BCD	E
A	0	10	7
BC		0	14
E			0

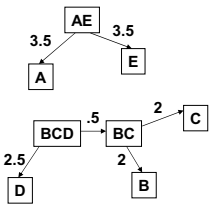
Updating distance matrices



	AE	BCD
AE	0	12
BCD		0

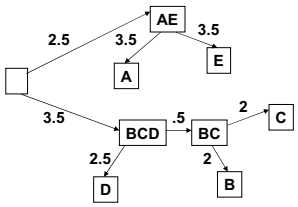
UPGMA step 4

combine AE and BCD



	AE	BCD
AE	0	12
BCD		0

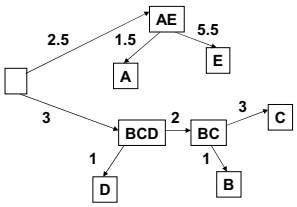
UPGMA Result



produced tree

	A	B	C	D	E
A	0	10	12	8	7
B		0	4	4	14
C			0	6	16
D				0	12
E					0

Actual tree



actual tree

	A	B	C	D	E
A	0	10	12	8	7
B		0	4	4	14
C			0	6	16
D				0	12
E					0

Limitations of UPGMA

- Ultrametric tree
  - Path distance from the root to each leaf is the same
- Ultrametric distance
  - Usual metric conditions
  - d(x,y) <= max[d(x,z), d(y,z)]
    - 2 largest distances in any group of 3 are equal
    - meaning in a tree setting?
- UPGMA works correctly for ultrametric distances

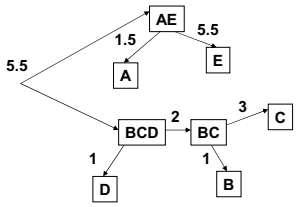
Neighbor Joining (NJ)

- Saitou and Nei, 1987
  - Join clusters that are close to each other and also <u>far from the rest</u>
- Produces unrooted tree
- NJ is a fast method, even for hundreds of sequences.
- NJ always finds the correct tree if distances are additive

Algorithm

- Define $u_i = \sum_{k \neq i} D_{ik} / (n-2)$
  - measure of average distance from other nodes
- Iterate until 2 nodes are left
  - choose pair (i,j) with smallest $D_{ij} - u_i - u_j$
    - close to each other and far from others
  - merge to a new node (ij) and update distance matrix
    - $D_{k,(ij)} = (D_{ik} + D_{jk} - D_{ij})/2$ -- consider the tree paths
    - $D_{i,(ij)} = (D_{ij} + u_i - u_j)/2$ -- similarly
    - $D_{j,(ij)} = D_{ij} - D_{i,(ij)}$ -- similarly
  - delete nodes i and j
- For the final group (i,j), use D_{ij} as the edge weight.

Neighbor-Joining Result



	A	B	C	D	E
A	0	10	12	8	7
B		0	4	4	14
C			0	6	16
D				0	12
E					0

WWW Resources

- PHYLIP : an extensive package of programs for all platforms
  `http://evolution.genetics.washington.edu/phylip.html`
- CLUSTALX : beyond alignment, it also performs NJ
- PAUP* : a very performing commercial package
  `http://paup.csit.fsu.edu/index.html`
- PHYLO_WIN : a graphical interface, for unix only
  `http://pbil.univ-lyon1.fr/software/phylowin.html`
- MrBayes : Bayesian phylogenetic analysis
  `http://morphbank.ebc.uu.se/mrbayes/`
- PHYML : fast maximum likelihood tree building
  `http://www.lirmm.fr/~guindon/phyml.html`
- WWW-interface at Institut Pasteur, Paris
  `http://bioweb.pasteur.fr/seganal/phylogeny`
- Tree drawing
  NJPLOT (for all platforms)
  `http://pbil.univ-lyon1.fr/software/njplot.html`

A fun example: Chaining Chain Letters

- Charles Bennett collected 33 chain letters 1980-1995. Using phylogenetic methods, Ming Li *et al.* reconstructed their history. Answered open questions of chain letter experts. Appeared in *Scientific American, June 2003* issue.
- Like a gene, they are about 2000 characters; like a virus, they have infected millions of people, they mutate just like a genome. Traditional phylogeny methods should also work on them. But they don't: alignment fail--translocated sentences; no model of evolution.
- They used their own method to calculate shared information between each pair of chain letters.

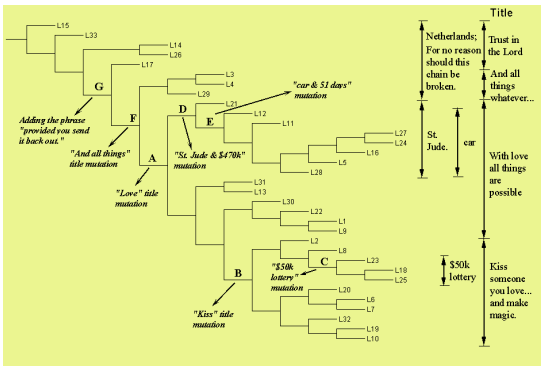
A sample letter:

Trust in the lord with all your heart and he will acknowledge and He will light the way. This Prayer has been sent to you for good luck. The original copy is from the Netherlands. It has been around the world nine times. The luck has been brought to you. You are to receive good luck within four days of receiving this letter. This is no joke. You will receive it in the mail. Send copies of this letter to people you think need good luck. Do not send money. Do not keep this letter. It must leave your hands within ninety six hours after you receive it. An RAF officer received $70,000. Joe Elliott received $50,000 and lost it because he broke the chain. While in the Philippines, George Welch lost his life six days after he received this letter. He failed to circulate the Prayer. However, before his death, he received $775,000. Please send twenty copies and see what happens to you on the fourth day. This chain comes from Venezuela and was written by Sal Anthony De Cgrif, a missionary from South America. Since this chain must make a tour of the world, you must make twenty copies identical to this one and send it to your friends, parents, and acquaintances. After a few days you will get a surprise. This is true, even if you are not superstitious. Take note of the following. Constantine Dias received the chain in 1953. He asked his secretary to make twenty copies and send them. A few days later he won a lottery for two million dollars in his country. Carlo Crabdit, and office employee, received the chain. He forgot it and in a few days lost his job. He found the chain and sent it to twenty people. Five days later he got an even better job. Delin Fairchild received the chain and not believing in it, threw it away. Nine days later he died. For no reason what so ever should this chain be broken

Another typical chain letter:

with love all things are possible this paper has been sent to you for good luck. the original is in new england. it has been around the world nine times. the luck has been sent to you. you will receive good luck within four days of receiving this letter. provided, in turn, you send it on. this is no joke. you will receive good luck in the mail. send no money. send copies to people you think need good luck. do not send money as faith has no price. do not keep this letter. It must leave your hands within 96 hours. an r.a.f. (royal air force) officer received $470,000. joe elliot received $40,000 and lost them because he broke the chain. while in the philippines, george welch lost his wife 51 days after he received the letter. however before her death he received $7,755,000. please, send twenty copies and see what happens in four days. the chain comes from venezuela and was written by saul anthony de grou, a missionary from south america. since this letter must tour the world, you must make twenty copies and send them to friends and associates. after a few days you will get a surprise. this is true even if you are not superstitious. do note the following: constantine dias received the chain in 1953. he asked his secretary to make twenty copies and send them. a few days later, he won a lottery of two million dollars. carlo daddit, an office employee, received the letter and forgot it had to leave his hands within 96 hours. he lost his job. later, after finding the letter again, he mailed twenty copies; a few days later he got a better job. dalan fairchild received the letter, and not believing, threw the letter away, nine days later he died. in 1987, the letter was received by a young woman in california, it was very faded and barely readable. she promised herself she would retype the letter and send it on, but she put it aside to do it later. she was plagued with various problems including expensive car repairs, the letter did not leave her hands in 96 hours. she finally typed the letter as promised and got a new car. remember, send no money. do not ignore this. it works. st. jude

Phylogeny of 33 Chain Letters



Another example: A language tree created using UN's The Universal Declaration Of Human Rights

Protein structures

Protein Structure

- Why protein structure?
- The basics of protein
- Basic measurements for protein structure
- Levels of protein structure
- Prediction of protein structure from sequence
- Finding similarities between protein structures
- Classification of protein structures

Why protein structure?

- In the factory of living cells, proteins are the workers, performing a variety of biological tasks.
- Each protein has a particular 3-D structure that determines its function.

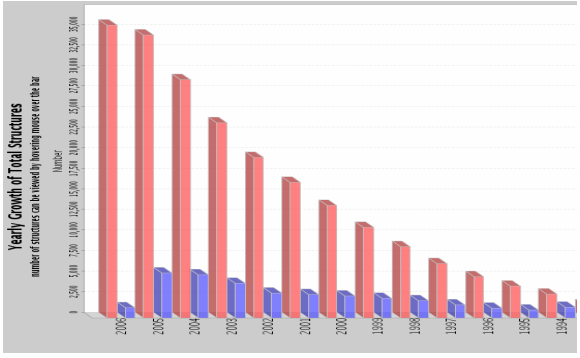
Sequence -> Structure -> Function

- Protein structure is more conserved than protein sequence, and more closely related to function.

Structural information

- Protein Data Bank: maintained by the Research Collaboratory of Structural Bioinformatics(RCSB)
  - http://www.rcsb.org/pdb/
  - > 35000 structures of proteins
  - Also contains structures of Protein/Nucleic Acid Complexes, Nucleic Acids, Carbohydrates
- Most structures are determined by X-ray crystallography. Other methods are NMR and electron microscopy(EM). Theoretically predicted structures were removed from PDB a few years ago.

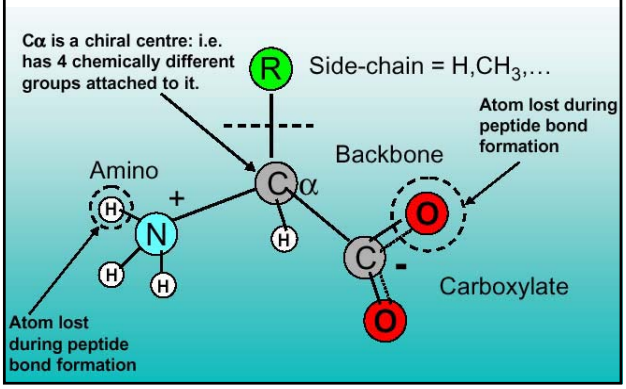
PDB Growth



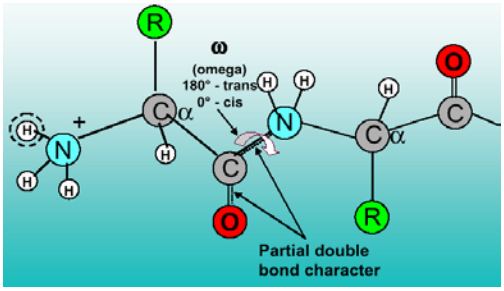
The basics of protein

- Proteins are linear heteropolymers: one or more polypeptide chains
- Building blocks: 20 types of amino acids.
- Range from a few 10s-1000s
- Three-dimensional shapes (“fold”) adopted vary enormously.

Common structure of Amino Acid

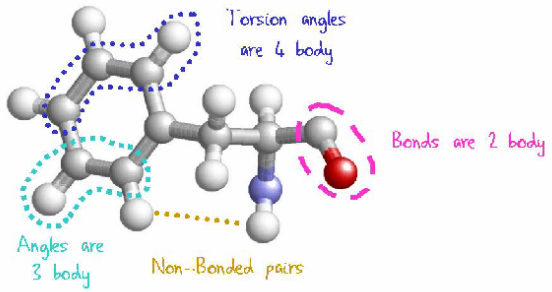


Formation of polypeptide chain



Basic Measurements for protein structure

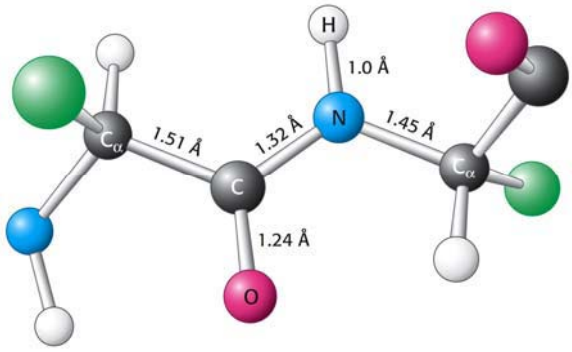
- Bond lengths
- Bond angles
- Dihedral (torsion) angles



Bond Length

- The distance between bonded atoms is constant
- Depends on the “type” of the bond
- Varies from 1.0 Å(C-H) to 1.5 Å(C-C)
- BOND LENGTH IS A FUNCTION OF THE POSITIONS OF TWO ATOMS.

Bond Length

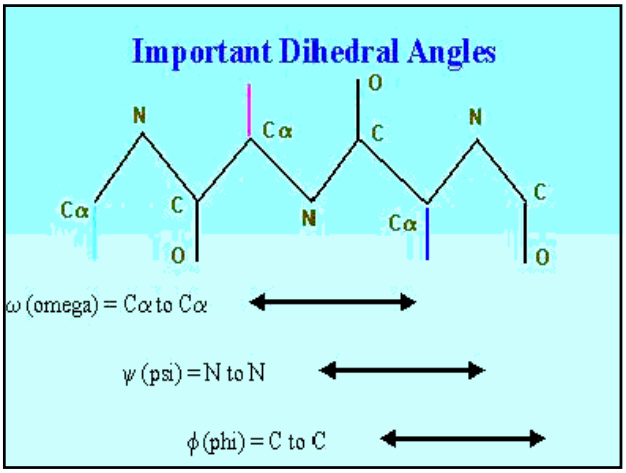


Bond Angles

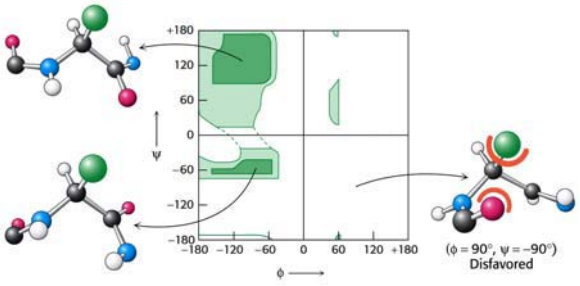
- All bond angles are determined by chemical makeup of the atoms involved, and are constant.
- Depends on the type of atom, and number of electrons available for bonding.
- Ranges from 100° to 180°
- BOND ANGLES IS A FUNCTION OF THE POSITION OF THREE ATOMS.

Dihedral Angles

- These are usually variable
- Range from 0-360° in molecules
- Most famous are ϕ, ψ, ω and χ
- DIHEDRAL ANGLES ARE A FUNCTION OF THE POSITION OF FOUR ATOMS.



Ramachandran plot



Levels of protein structure

- Primary structure
- Secondary structure
- Tertiary structure
- Quaternary structure

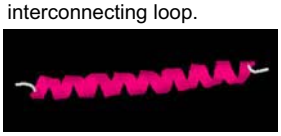
Primary structure

- This is simply the amino acid sequences of polypeptides chains (proteins).

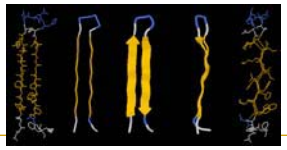
MHGAYRTPRSKTDAYGCQILETRAS

Secondary structure

- Local organization of protein backbone: α-helix, β-strand (groups of β-strands assemble into β-sheet), turn and interconnecting loop.

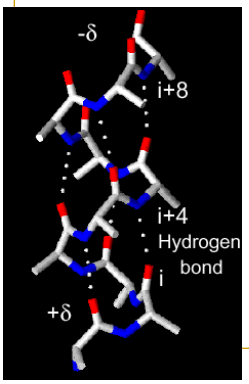


an α-helix



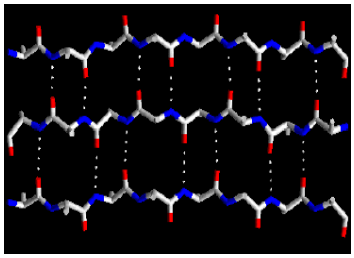
various representations and orientations of a two stranded β-sheet.

The α-helix



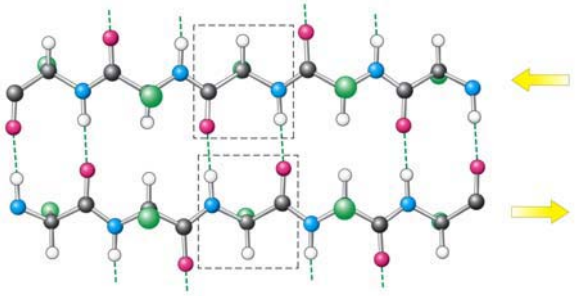
- One of the most closely packed arrangement of residues.
- Turn: 3.6 residues
- Pitch: 5.4 Å/turn

The β-sheet

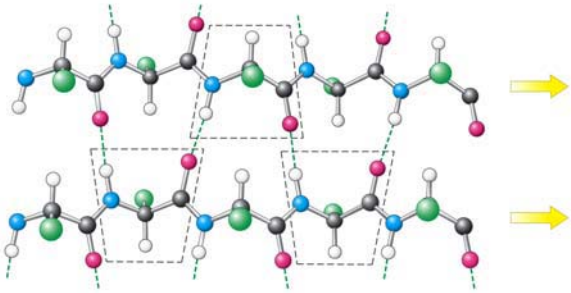


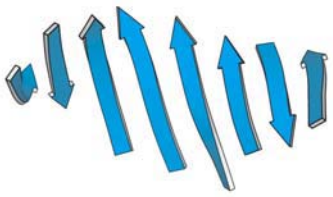
- Backbone almost fully extended, loosely packed arrangement of residues.

Anti-parallel beta sheet



Parallel beta sheet

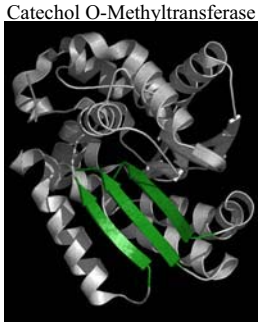




β-Sheet (parallel)

All strands run in the same direction

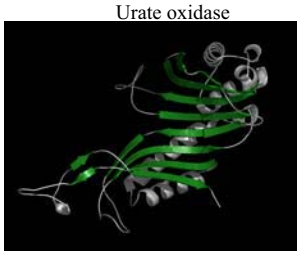
Catechol O-Methyltransferase



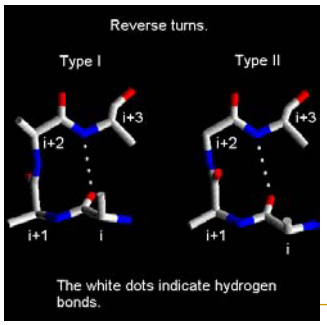
β-Sheet (antiparallel)

All strands run in the opposite direction, more stable

Urate oxidase



Loops and Turns



Loops: often contain hydrophilic residue on the surface of proteins

Turns: loops with less than 5 residues and often contain G, P

TABLE 3.3 Relative frequencies of amino acid residues in secondary structures

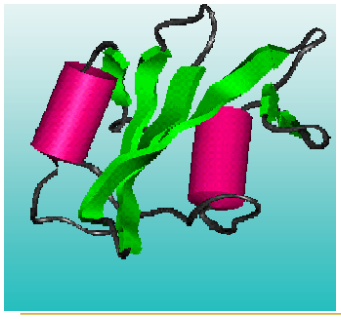
Amino acid	α helix	β sheet	Turn
Ala	**1.29**	0.90	0.78
Cys	**1.11**	0.74	0.80
Leu	**1.30**	1.02	0.59
Met	**1.47**	0.97	0.39
Glu	**1.44**	0.75	1.00
Gln	**1.27**	0.80	0.97
His	**1.22**	1.08	0.69
Lys	**1.23**	0.77	0.96
Val	0.91	**1.49**	0.47
Ile	0.97	**1.45**	0.51
Phe	1.07	**1.32**	0.58
Tyr	0.72	**1.25**	1.05
Trp	0.99	**1.14**	0.75
Thr	0.82	**1.21**	1.03
Gly	0.56	0.92	**1.64**
Ser	0.82	0.95	**1.33**
Asp	1.04	0.72	**1.41**
Asn	0.90	0.76	**1.28**
Pro	0.52	0.64	**1.91**
Arg	0.96	0.99	0.88

Note: The amino acids are grouped according to their preference for α helices (top group), β sheets (second group), or turns (third group). Arginine shows no significant preference for any of the structures.
After T. E. Creighton, *Proteins: Structures and Molecular Properties*, 2d ed. (W. H. Freeman and Company, 1992), p. 256.

Tertiary structure

- Description of the type and location of SSEs is a chain's *secondary structure.*
- Three-dimensional coordinates of the atoms of a chain is its *tertiary structure*.
- *Quaternary structure*: describes the spatial packing of several folded polypeptides

Tertiary structure



Packing the secondary structure elements into a compact spatial unit

"Fold" or domain– this is the level to which structure prediction is currently possible.