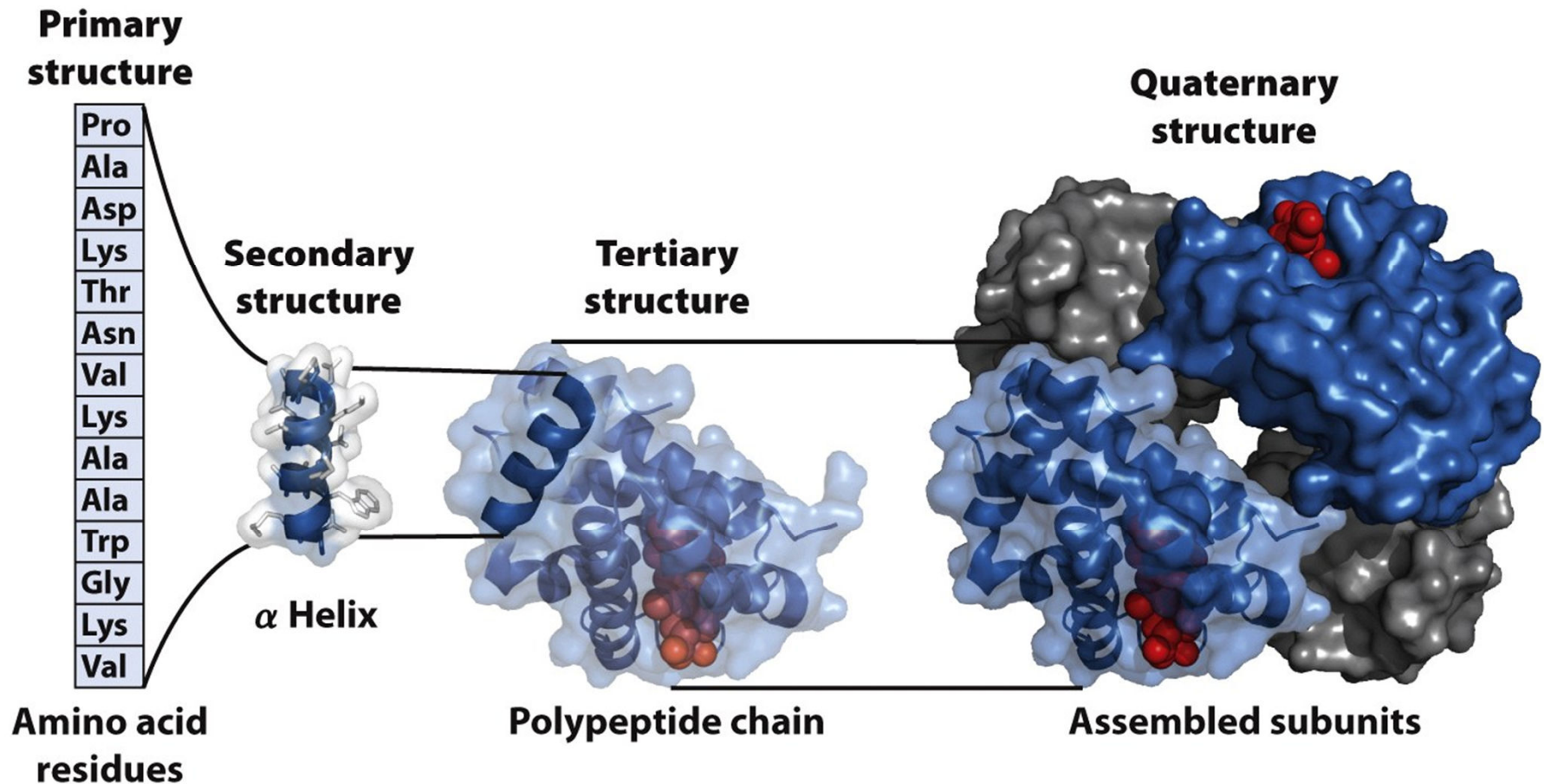


# There are four levels of structure in proteins



We know that sequence  $\rightarrow$  structure, but we can't (yet) predict a structure from a sequence. So what good is knowing the sequence?

We can learn about proteins by *comparing* their sequences:

- **Structure** – similar sequences form similar structures
- **Function** – motifs (sequence patterns) can indicate particular functions
- **Location** – signal sequences (at the protein's N-terminus) direct proteins to specific locations in the cell, or to be excreted
- **Modification** – signal sequences or motifs can indicate sites for modification
- **Evolution** – differences in related sequences reflect evolutionary distance
- **Dysfunction** – changes in sequence can lead to disease

# How do we determine a protein's sequence?

- Directly from the protein
  - Chemical sequencing
  - Physical methods for analyzing structure
- From the gene sequence

**Amino acid**

**sequence (protein)**

**Gln – Tyr – Pro – Thr – Ile – Trp**

**DNA sequence (gene)**

**CAGTATCCTACGATTGG**



## Important protein sequence databases

- RefSeq (NCBI) – “Reference Sequence”
  - Non-redundant database of well-annotated sequences (gene, transcript, protein)
  - <http://www.ncbi.nlm.nih.gov/projects/RefSeq/>
  - Can also use ‘Entrez’ at NCBI site to search multiple sequence databases
- UniProt (EBI) – “Universal Protein Resource”
  - Protein sequences and annotations, plus links to other databases
  - <http://www.uniprot.org/>

# Proteins are aligned for comparison

Sequence alignments:

<b><i>E. coli</i></b>	TGNRTIAVYDLGGGTFDISIIEIDEVDGEKTFEVLATNGDTHLGGEDFDSRLIHLYL
<b><i>B. subtilis</i></b>	DEDQTILLYDLGGGTFDVSILELGDGTFEVRSTAGDNRLGGDDFDQVIIDHL



			Signature sequence	
Archaea	{	<i>Halobacterium halobium</i>	IGHVDHGKSTMVGRLLYETGSVPEHV	IEQH
		<i>Sulfolobus solfataricus</i>	IGHVDHGKSTLVGRLLMDRGFIDEKT	VKEA
Eukaryotes	{	<i>Saccharomyces cerevisiae</i>	IGHVDSGKSTTTGHLIYKCGGIDKRT	IEKF
		<i>Homo sapiens</i>	IGHVDSGKSTTTGHLIYKCGGIDKRT	IEKF
Gram-positive bacterium		<i>Bacillus subtilis</i>	IGHVDHGKSTMVGR	ITTV
Gram-negative bacterium		<i>Escherichia coli</i>	IGHVDHGKTTLTAA	ITTV

Consensus sequences, showing conserved sites, may be represented in different ways

