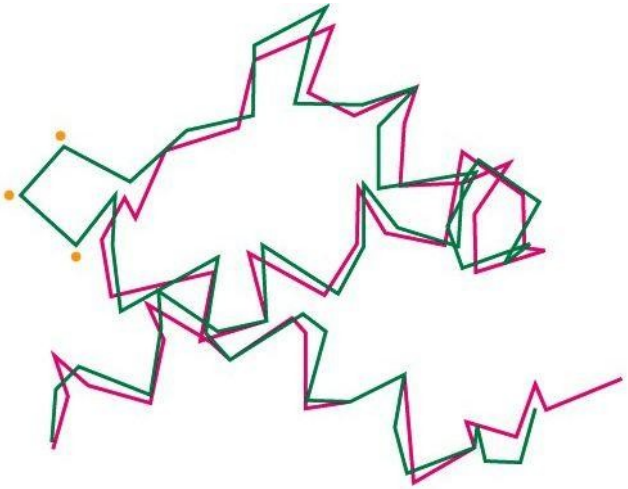
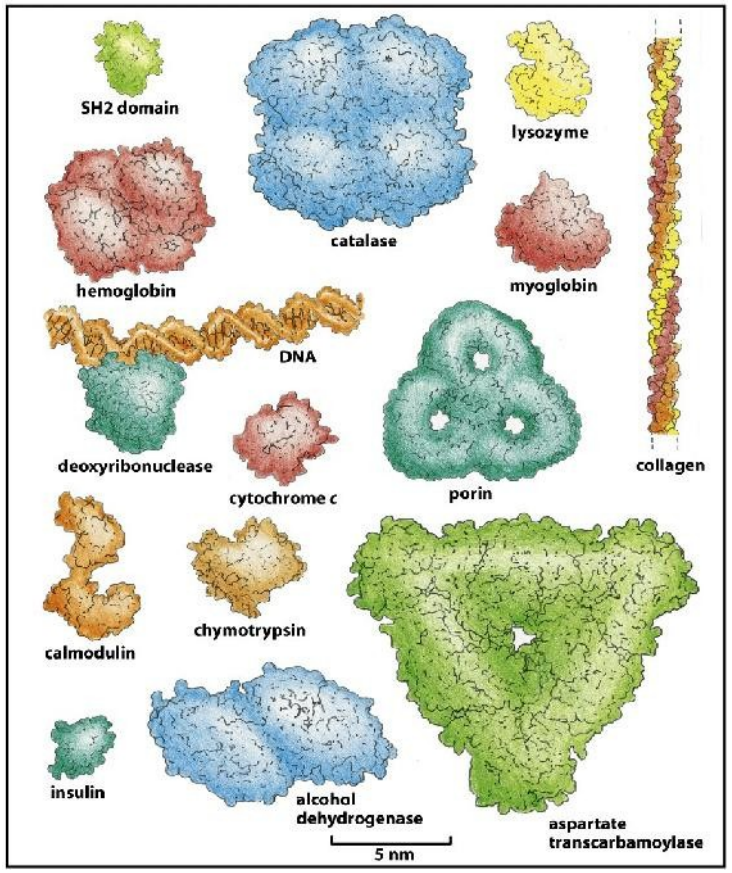
Protein Conservation



Genomics and Bioinformatics

Chapter 8

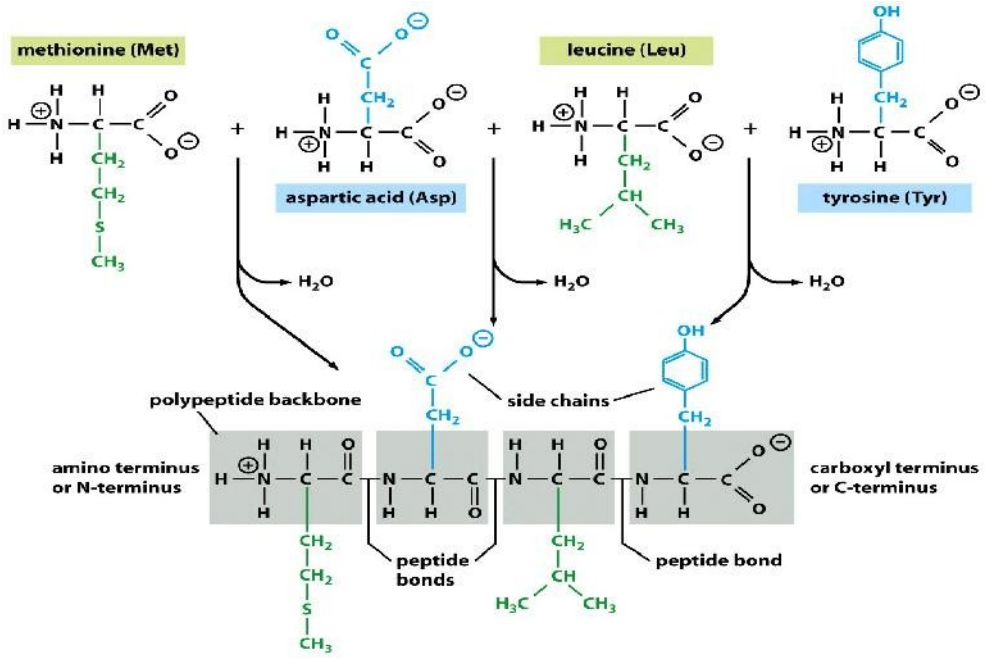
Proteins can have very different 3D structures



Unlike DNA, which always forms a double helix

Figure 3-23 *Molecular Biology of the Cell* (© Garland Science 2008)

Proteins are built from amino acid chains



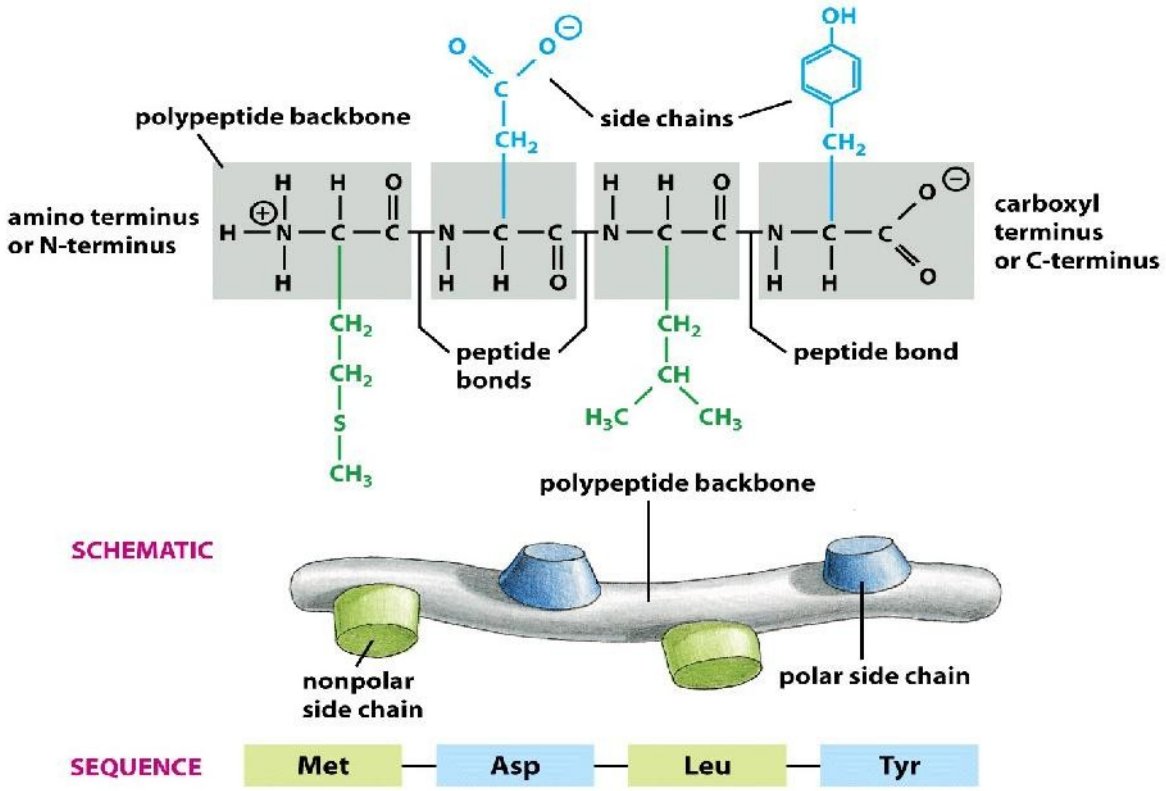
The polypeptide backbone

The side chains that differ between amino acids

All proteins have the same type of backbone. They differ in the sequence of side chains.

Figure 3-1 (part 1 of 2) *Molecular Biology of the Cell* (© Garland Science 2008)

The two ends of the backbone are chemically different



Protein sequences are always written from N-terminus to C-terminus

Figure 3-1 (part 2 of 2) *Molecular Biology of the Cell* (© Garland Science 2008)

There are 20 amino acids

AMINO ACID			SIDE CHAIN
Aspartic acid	Asp	D	negative
Glutamic acid	Glu	E	negative
Arginine	Arg	R	positive
Lysine	Lys	K	positive
Histidine	His	H	positive
Asparagine	Asn	N	uncharged polar
Glutamine	Gln	Q	uncharged polar
Serine	Ser	S	uncharged polar
Threonine	Thr	T	uncharged polar
Tyrosine	Tyr	Y	uncharged polar

POLAR AMINO ACIDS

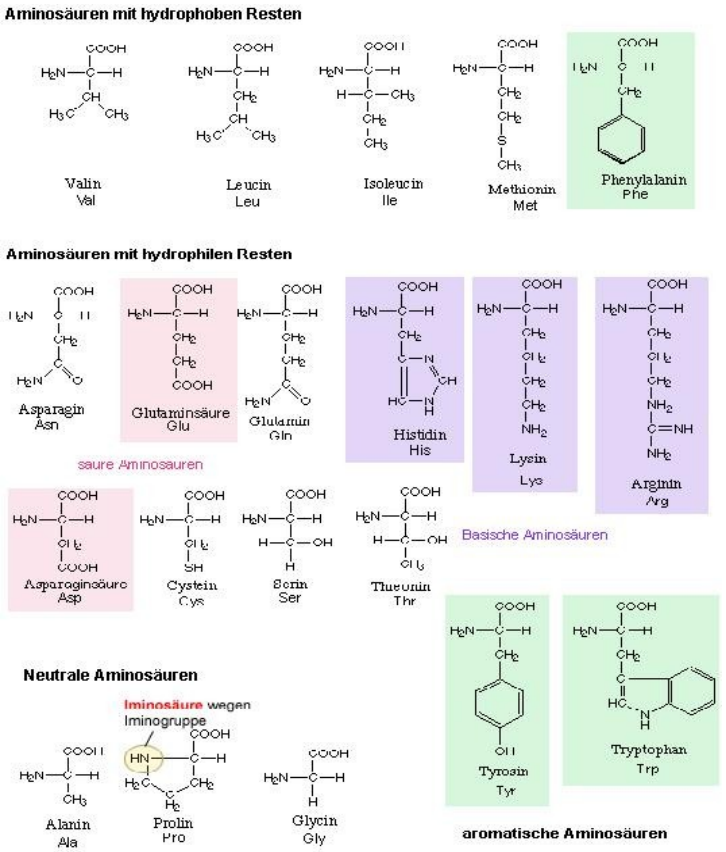
AMINO ACID			SIDE CHAIN
Alanine	Ala	A	nonpolar
Glycine	Gly	G	nonpolar
Valine	Val	V	nonpolar
Leucine	Leu	L	nonpolar
Isoleucine	Ile	I	nonpolar
Proline	Pro	P	nonpolar
Phenylalanine	Phe	F	nonpolar
Methionine	Met	M	nonpolar
Tryptophan	Trp	W	nonpolar
Cysteine	Cys	C	nonpolar

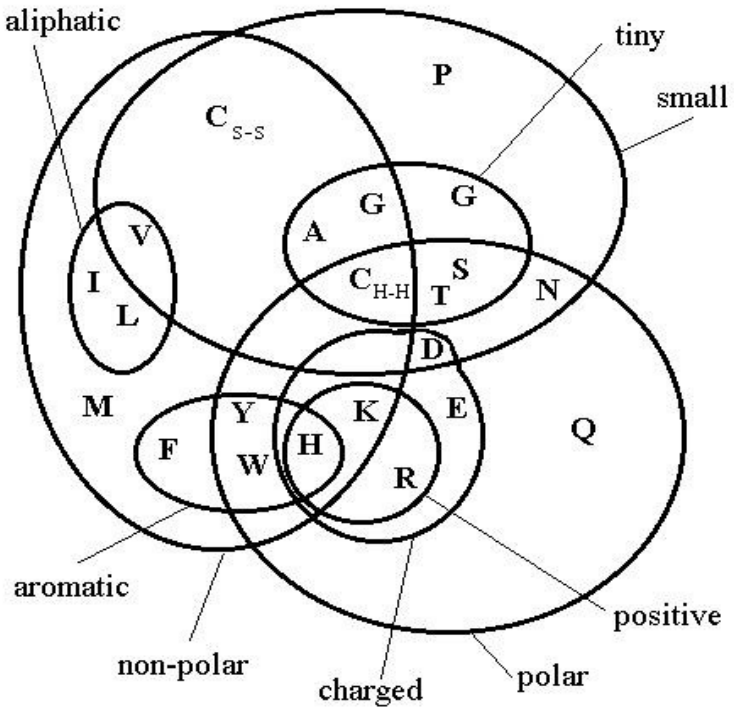
NONPOLAR AMINO ACIDS

Polar side chains can form hydrogen bonds with the surrounding water (hydrophilic). Non polar side chains avoid contact with water (hydrophobic).

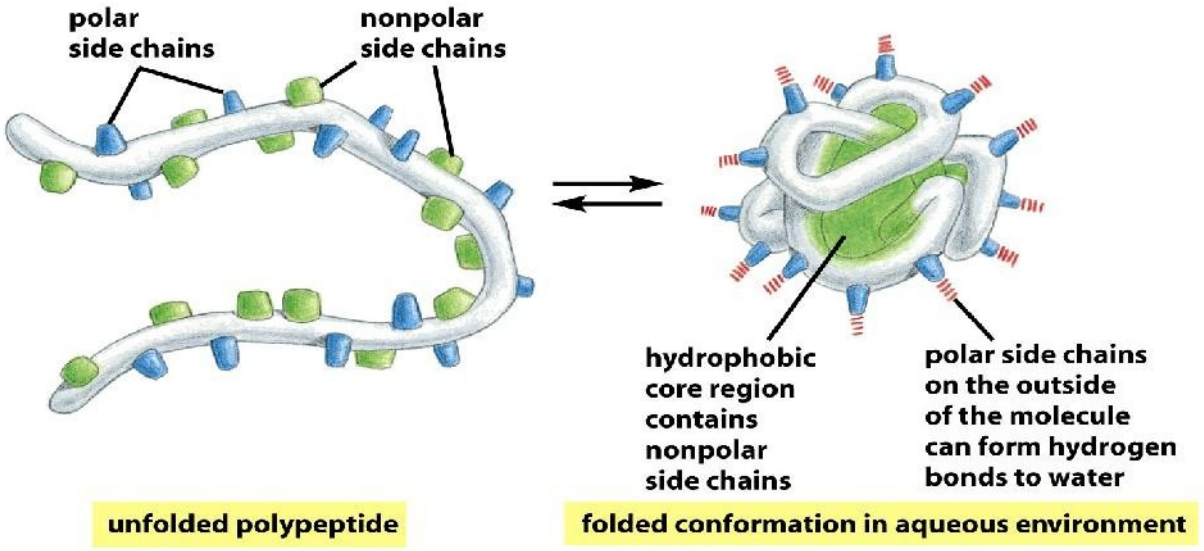
Figure 3-2 *Molecular Biology of the Cell* (© Garland Science 2008)

The chemical properties of the side chains vary substantially





Interactions of the side chains with each other and with the surrounding water fold the amino acid chain into a characteristic 3D structure



Proteins fold into a state of lowest free energy

Figure 3-5 *Molecular Biology of the Cell* (© Garland Science 2008)

Interacting side chains need not be close in the sequence

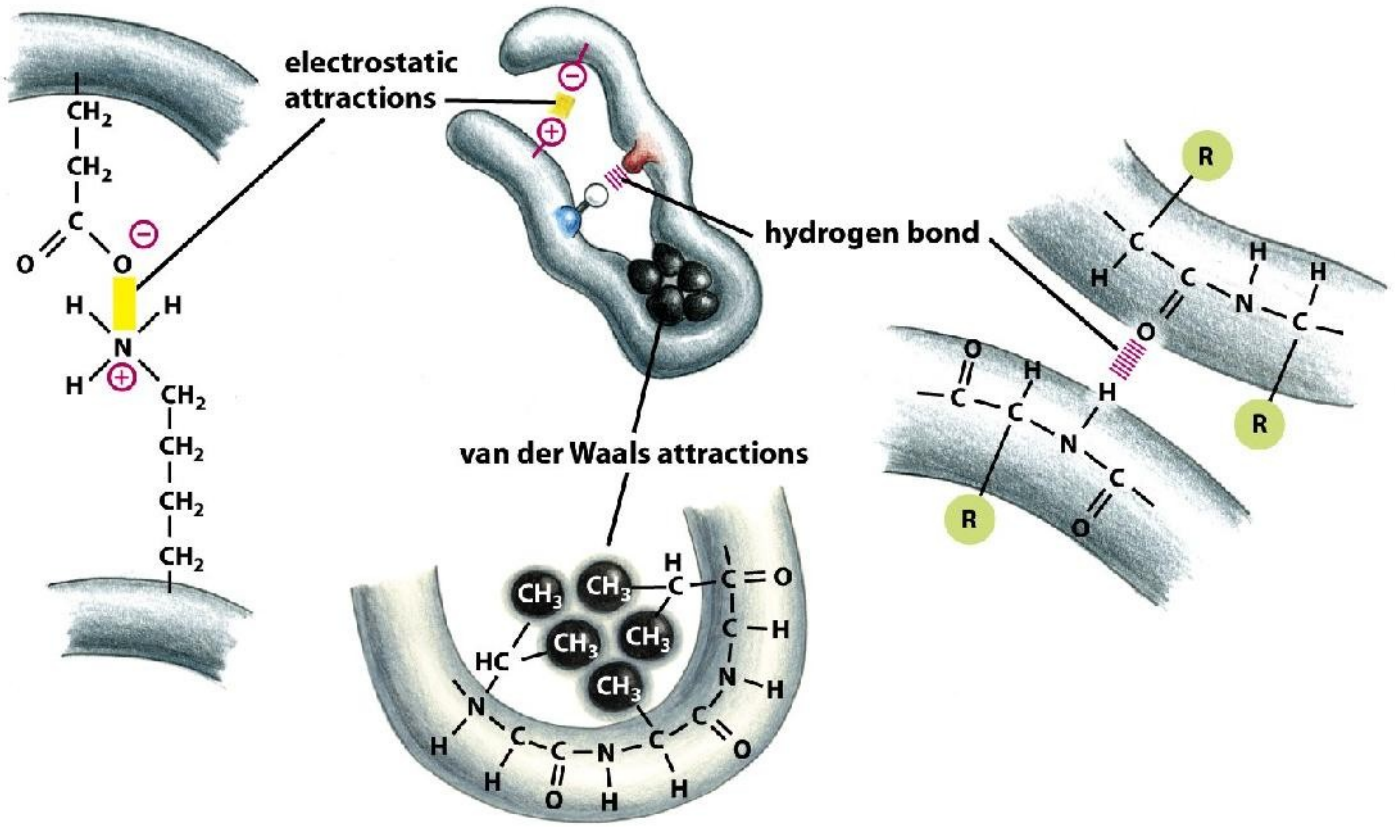
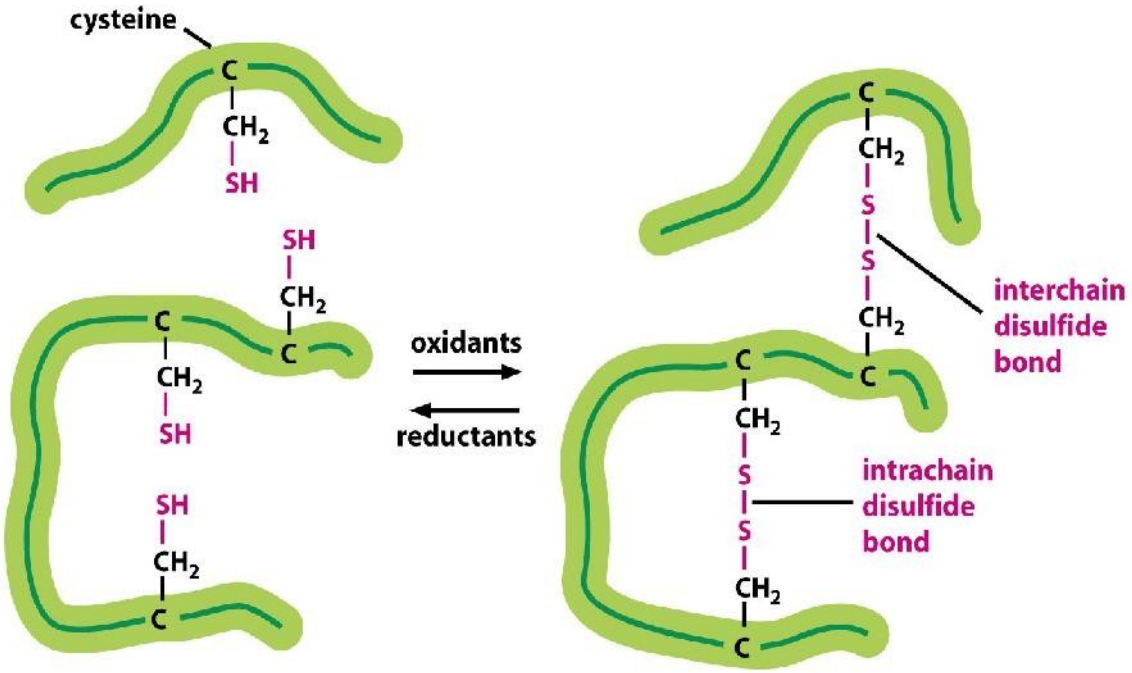


Figure 3-4 *Molecular Biology of the Cell* (© Garland Science 2008)

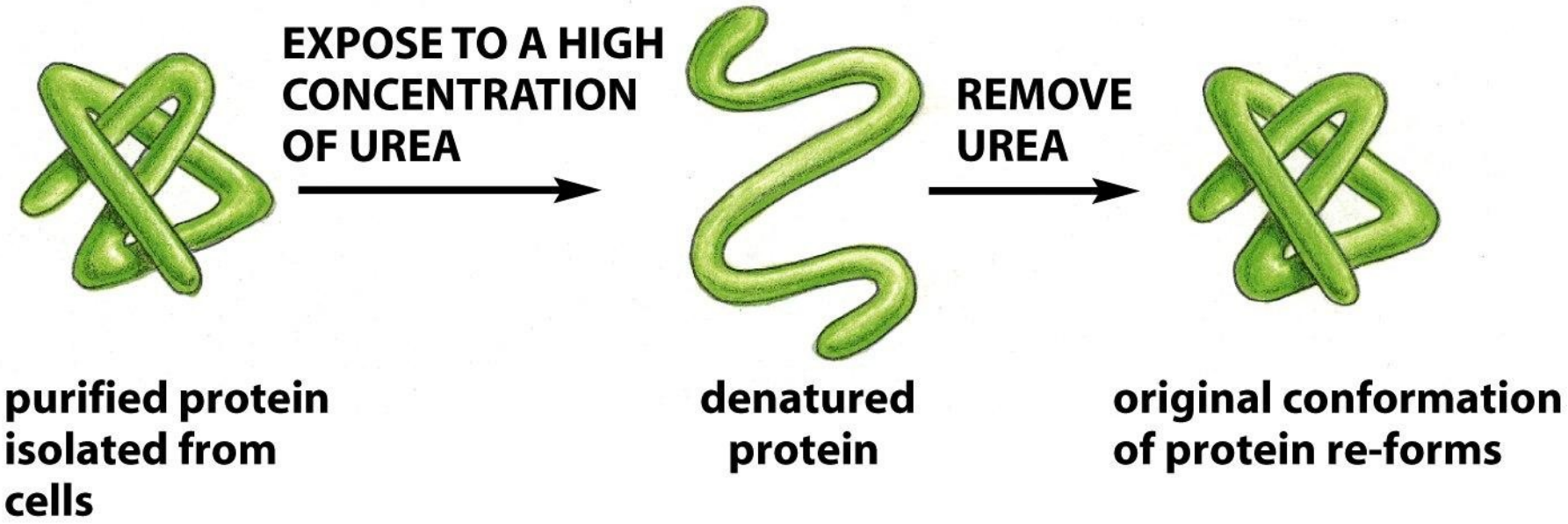
Cysteine pairs can form a special kind of bond that stabilizes many structures



Only cysteines can do this.

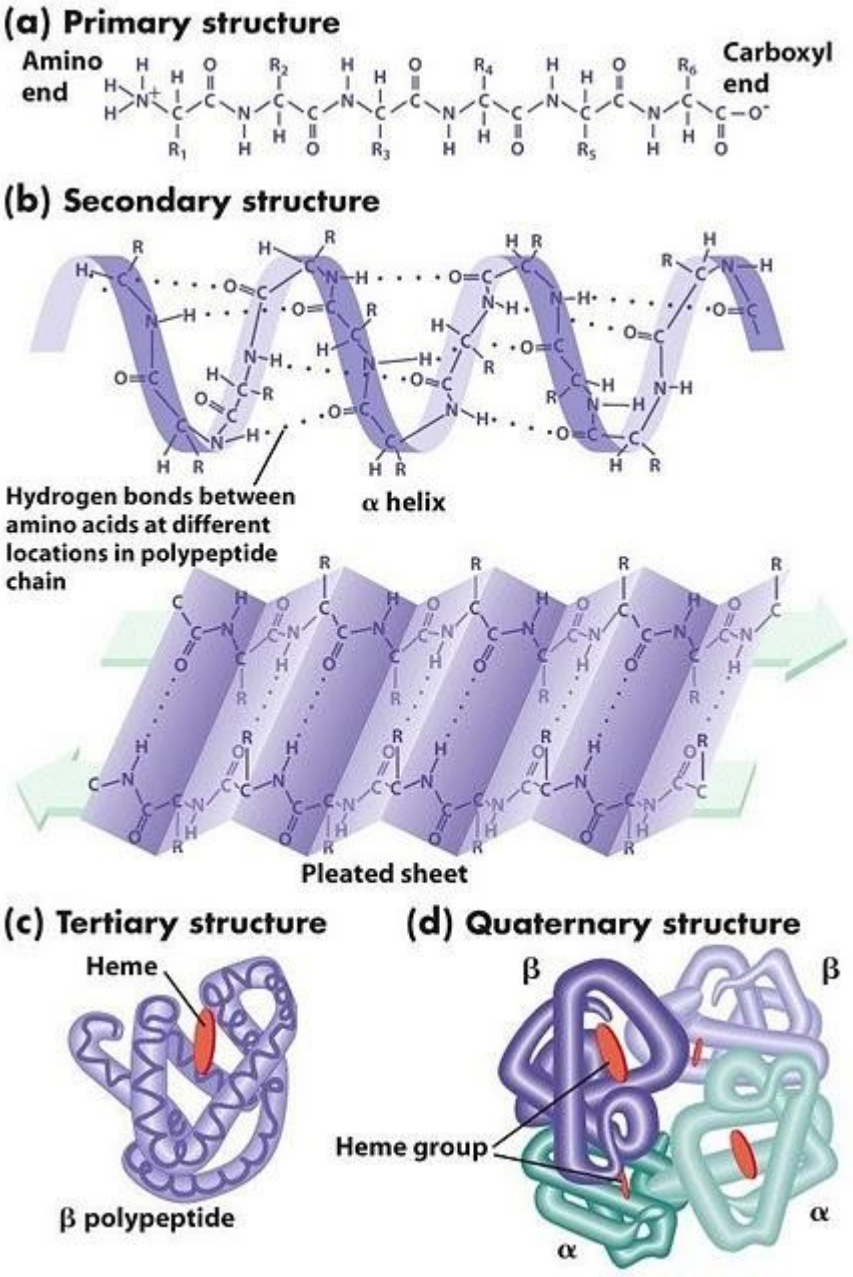
Figure 3-28 *Molecular Biology of the Cell* (© Garland Science 2008)

The amino acid sequence determines the folded 3D structure



… at least mostly …

Figure 3-6a *Molecular Biology of the Cell* (© Garland Science 2008)



Folding involves first the formation of secondary structures that then further fold into tertiary structures

Several amino acid chains can assemble to form a complex with a characteristic quaternary structure.

Typical secondary structures are α-helices and ß-sheets

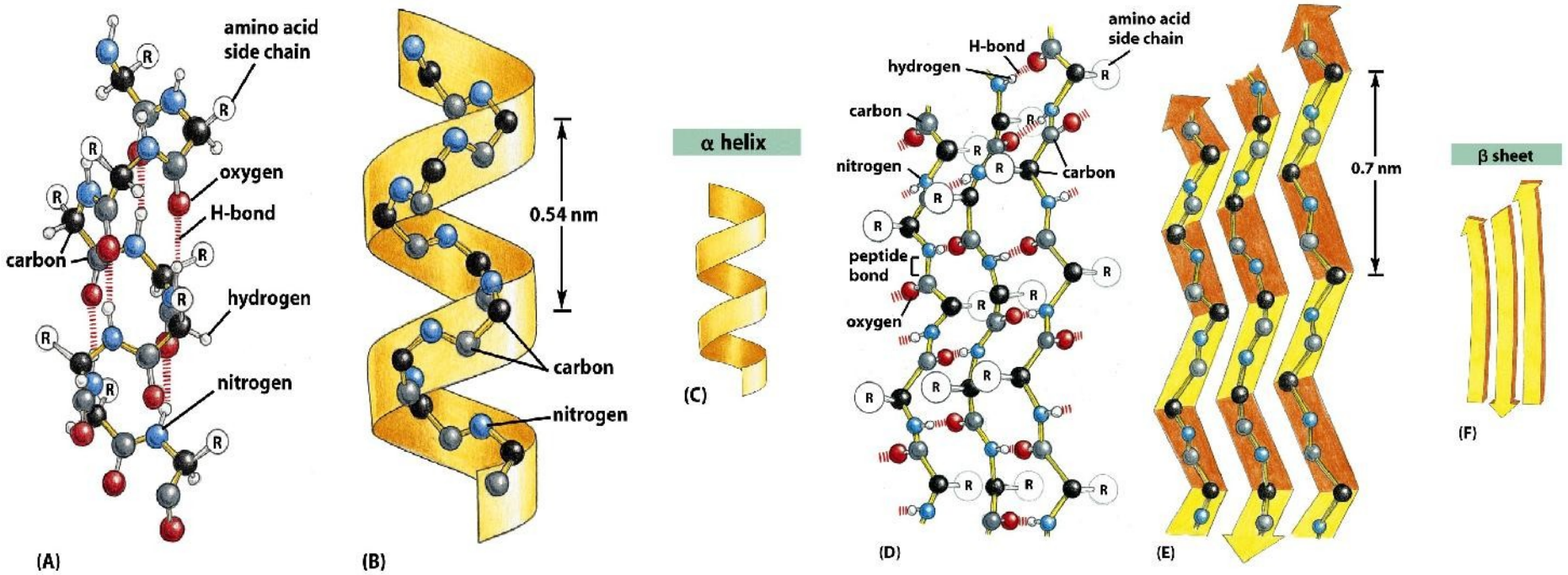
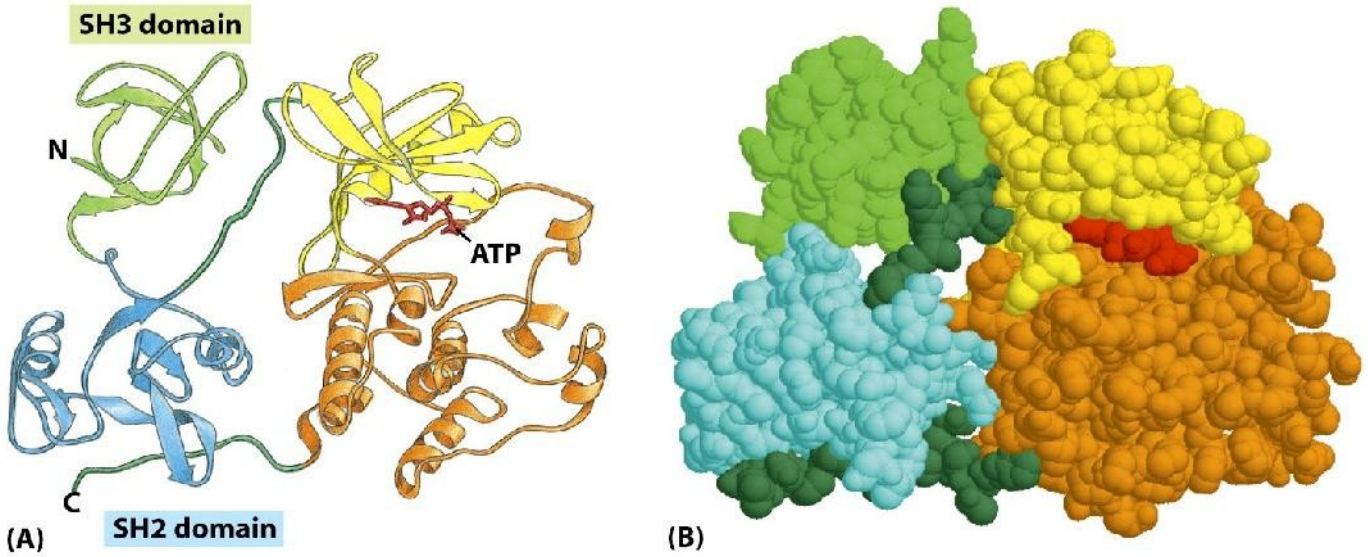


Figure 3-7a,b,c *Molecular Biology of the Cell* (© Garland Science 2008)

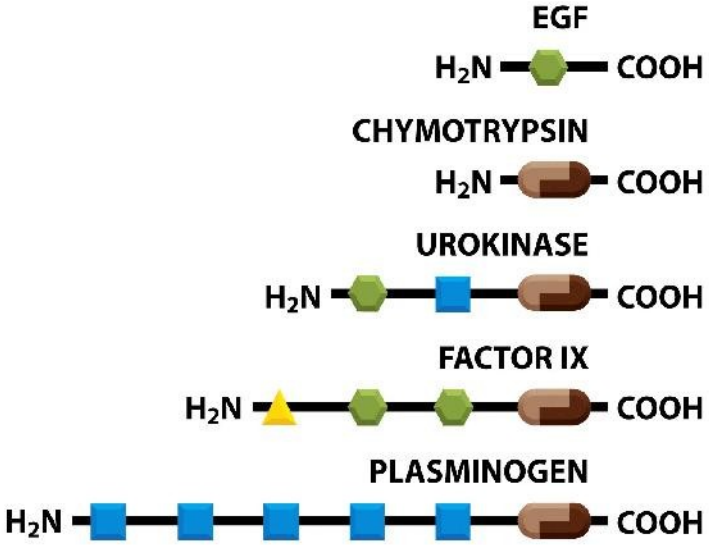
Many protein structures are modular



Protein domains are modular substructures of proteins that fold mostly independently of each other and are loosely connected by loops.

Figure 3-10 *Molecular Biology of the Cell* (© Garland Science 2008)

Protein domains are frequently reused modules of code



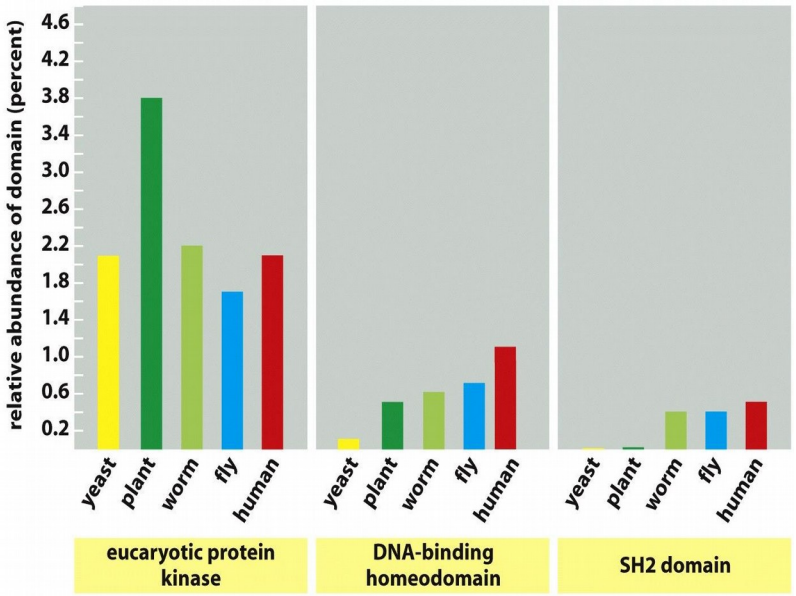


Figure 3-15 *Molecular Biology of the Cell* (© Garland Science 2008)
Figure 3-18 *Molecular Biology of the Cell* (© Garland Science 2008)

Protein structure is often more conserved than sequence

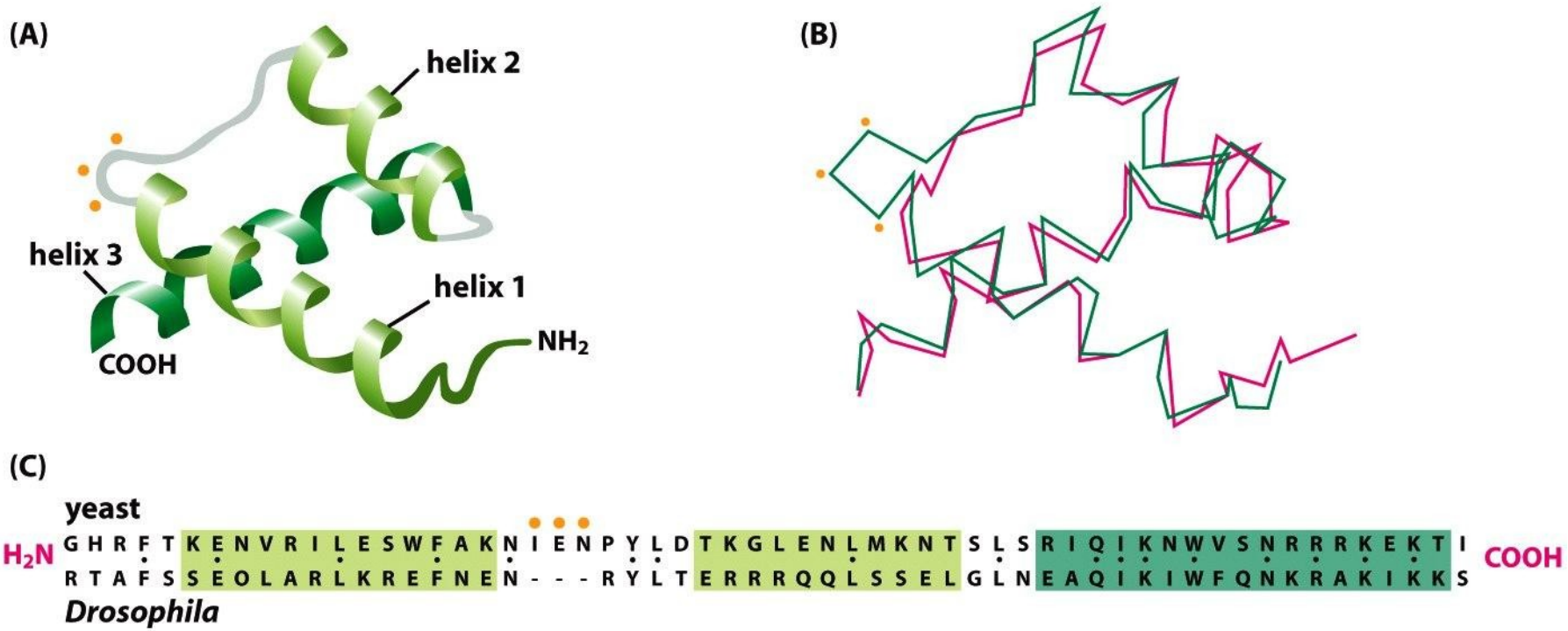


Figure 3-13 *Molecular Biology of the Cell* (© Garland Science 2008)

Nevertheless, local sequence similarities between protein sequences are often found inside domains

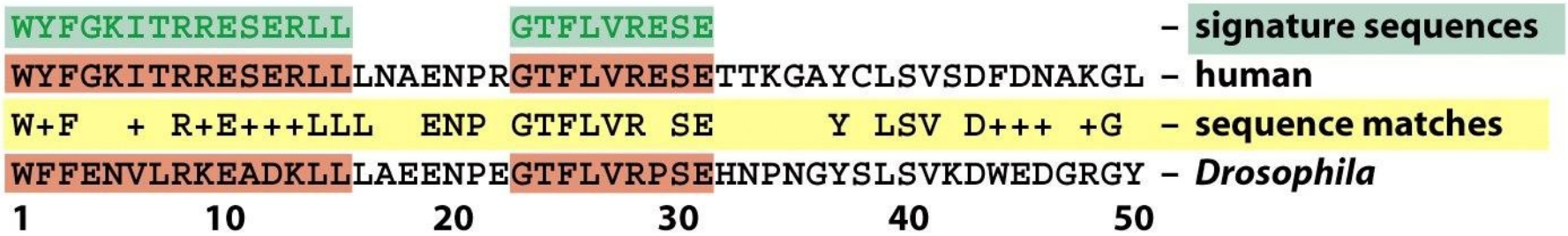


Figure 3-14 *Molecular Biology of the Cell* (© Garland Science 2008)

The three dimensional surface of a protein determines its binding sites

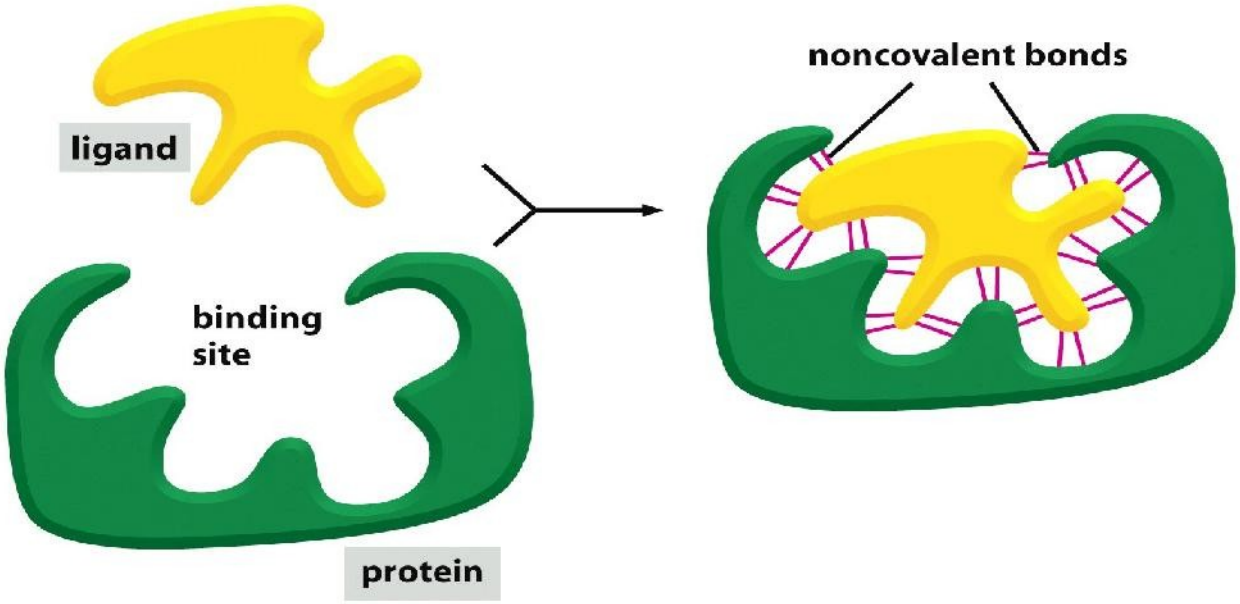
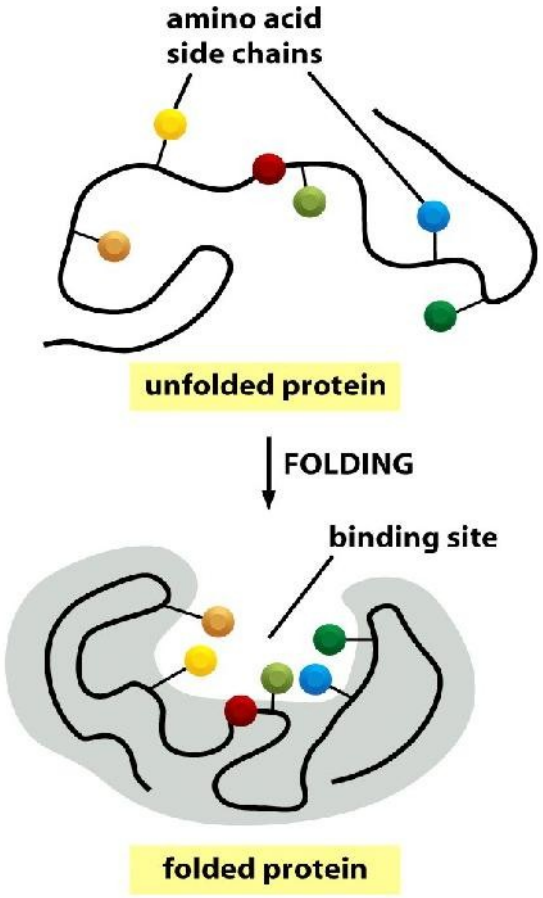


Figure 3-36 *Molecular Biology of the Cell* (© Garland Science 2008)

Side chains that bind a substrate are close in the 3D structure but not necessarily in the sequence



These amino acids and their neighbors are typically strongly conserved (selection pressure).

Since they can be spread out in the sequence we often observe many conserved sequences segments.

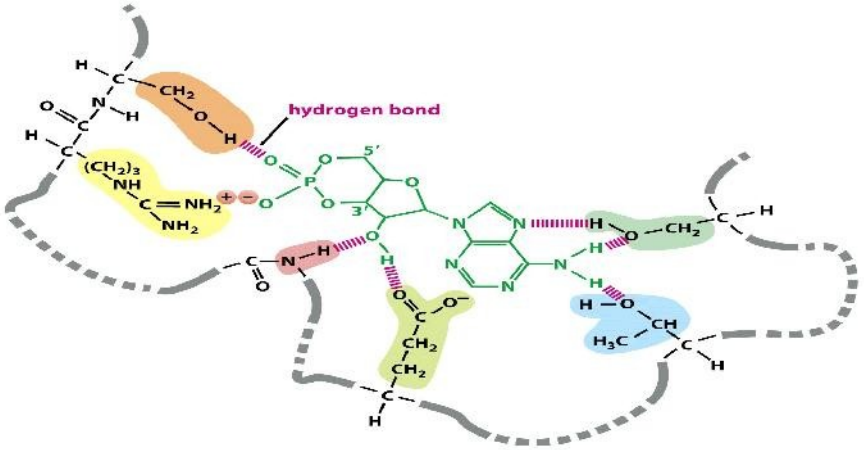
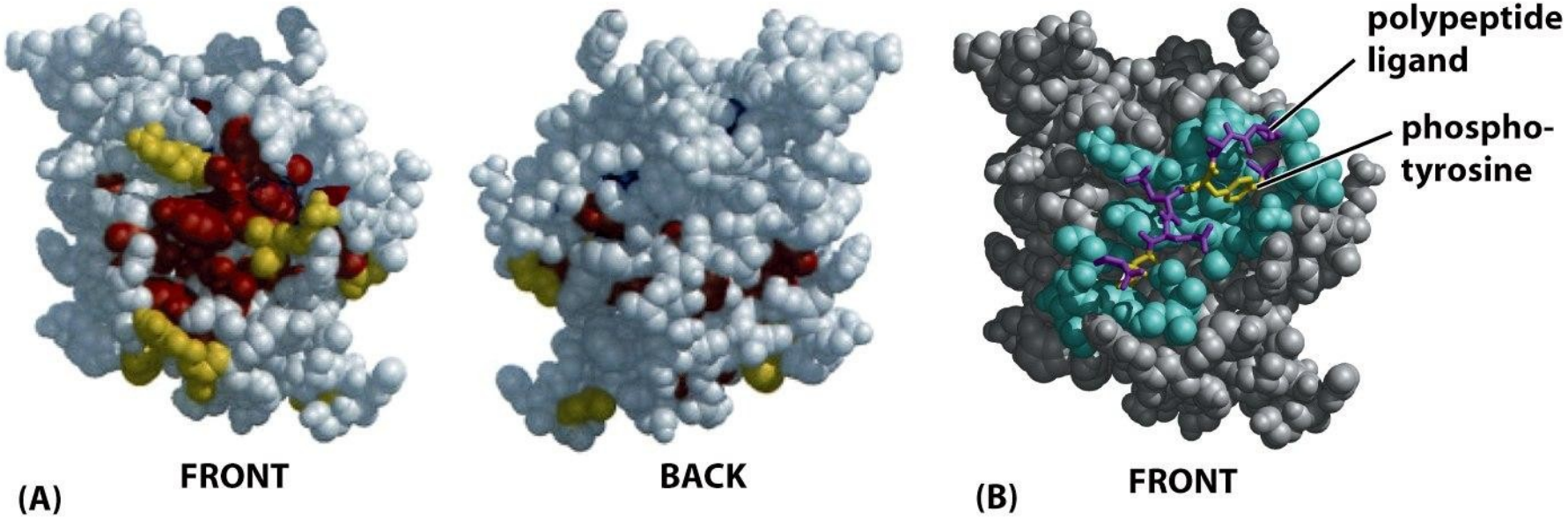


Figure 3-37a *Molecular Biology of the Cell* (© Garland Science 2008)

Conserved positions in a protein often point towards binding sites



Red and yellow: conserved positions detected by sequence alignment
Green: positions with contact to the ligand

Figure 3-39 *Molecular Biology of the Cell* (© Garland Science 2008)

Similarity of surface structure translates quantitatively into binding affinity

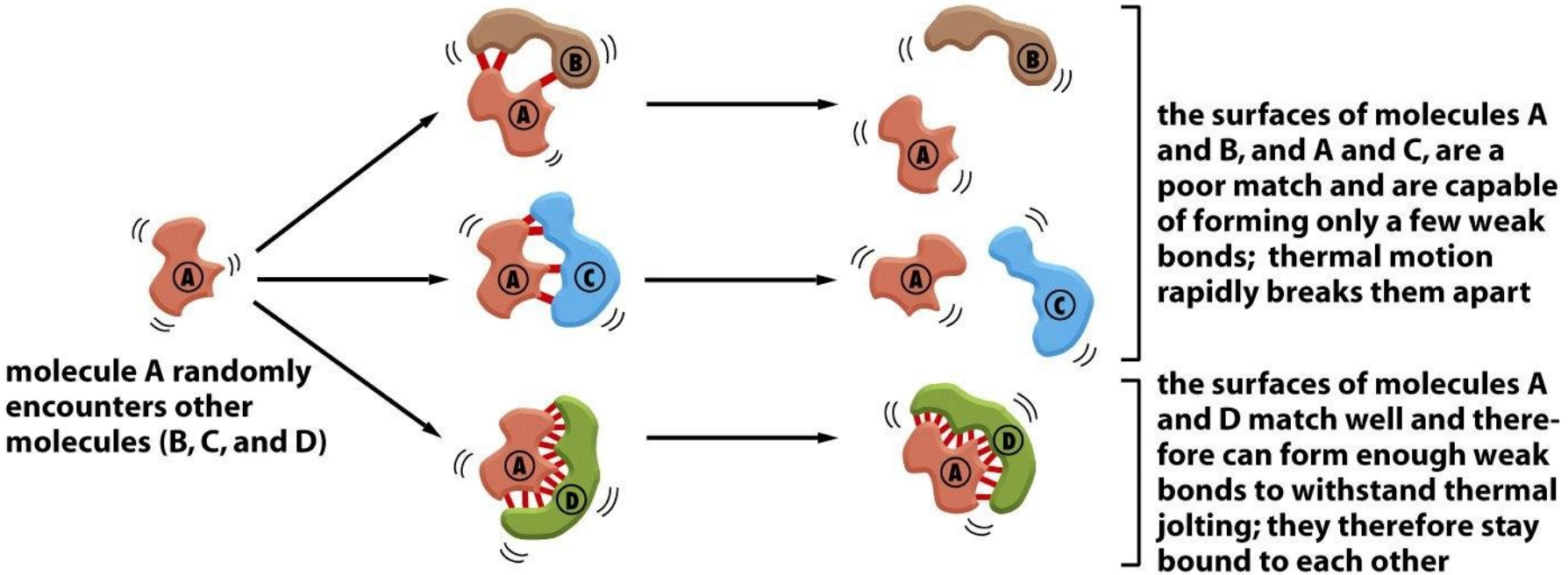
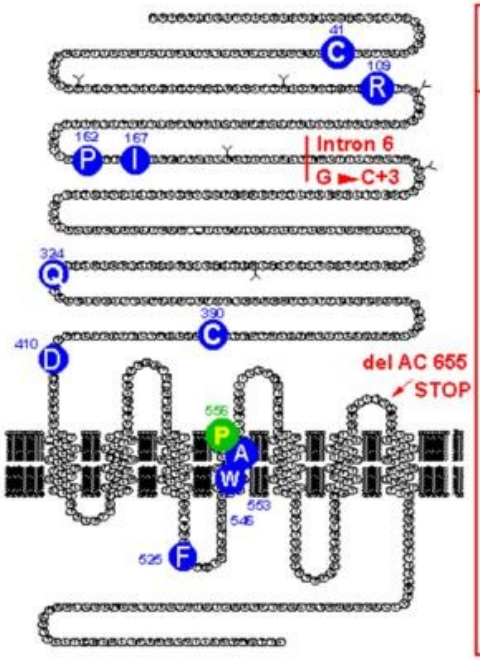


Figure 3-42 *Molecular Biology of the Cell* (© Garland Science 2008)

Single amino acid mutations of the wrong type at the wrong position can destroy the function of a protein



Inactivating mutations in TSH receptor
Cys 41 Ser
Arg 109 Gln
Pro 162 Ala
Ile 167 Asn
Gln 324 *Stop*
Cys 390 Trp
Asp 410 Asn
Phe 525 Leu
Trp 546 *Stop*
Ala 553 Thr
del AC 655
G►C + 3IVS6
Pro 556 Leu (mouse)

www.thyroidmanager.org/Chapter16/chapter16a.html

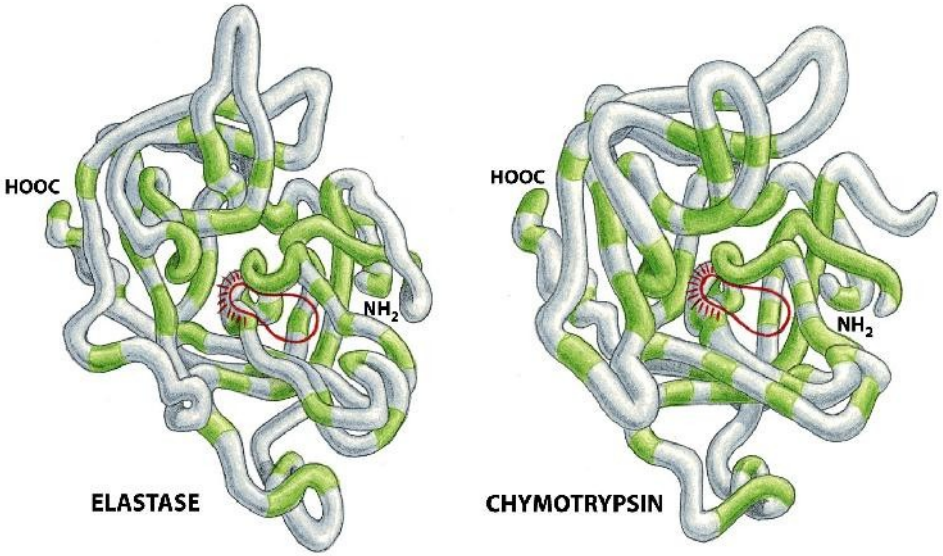
Loss of function mutations

Often found in samples from patients with disease

Cause strong selective pressure

Might not survive over longer periods of evolution

On the other hand mutations can accumulate without affecting the structure and function very much



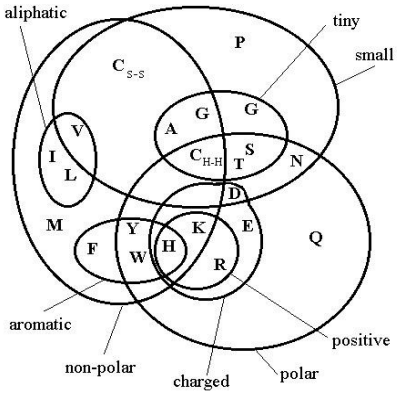
Two homologous human serine proteases

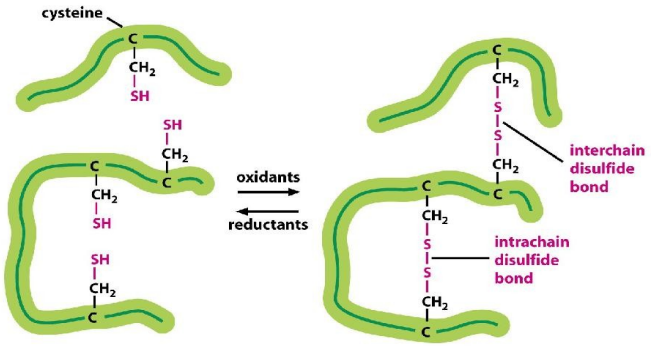
Only the green positions are identical on the sequence level

The mutations are at positions and of a type that does not affect the structure very much

Figure 3-12 *Molecular Biology of the Cell* (© Garland Science 2008)

There are types of mutations that change the 3D structure and function of proteins more than others





Replacing a polar by a non polar amino acid

Replacing a small by a large amino acid

Replacing a small polar by a large non polar amino acid

Replacing a cysteine that was involved in a cysteine bridge …

In protein alignments we should penalize mismatches that put stress on the 3D structure more than others

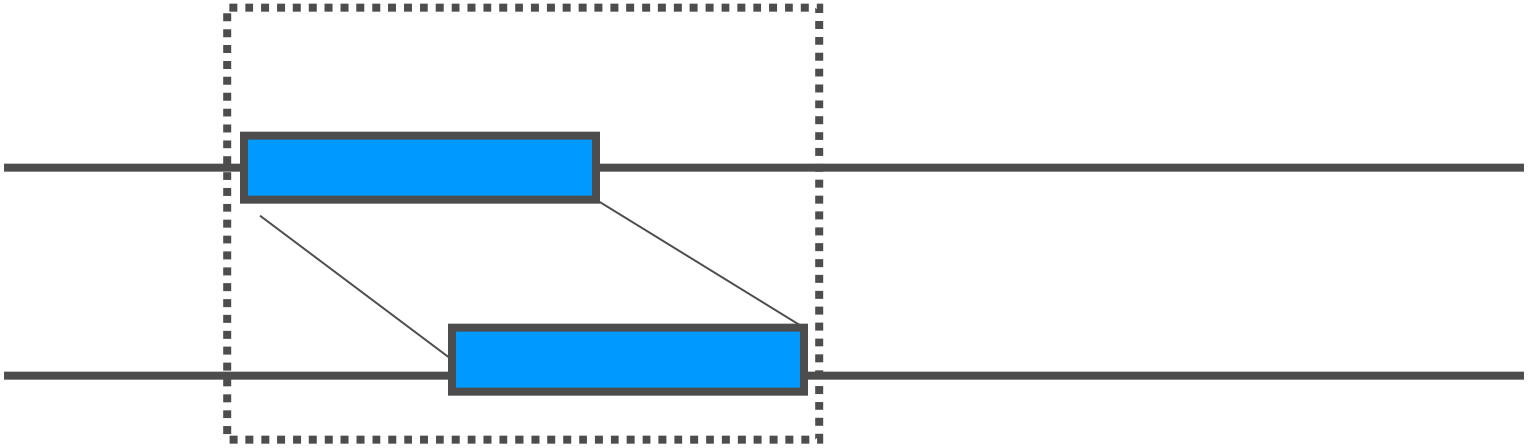
```
WYFGKITRRESERLLLNAENPRGTFLVRESETTKGAYCLSVSDFDNAKGL – human
W+F   + R+E+++LLL   ENP GTFLVR SE      Y LSV D+++ +G  – sequence matches
WFFENVLRKEADKLLLAEENPEGTFLVRPSEHNPNGYSLSVKDWEDGRGY – Drosophila
1         10          20          30          40          50
```

The + points to mismatch positions where the two aligned amino acids are very similar.

In fact these mismatches are taken as evidence for conservation and not divergence. The DNA code changed, but selective pressure on the 3D structure conserved the type of amino acid at this position.

They get positive scores.

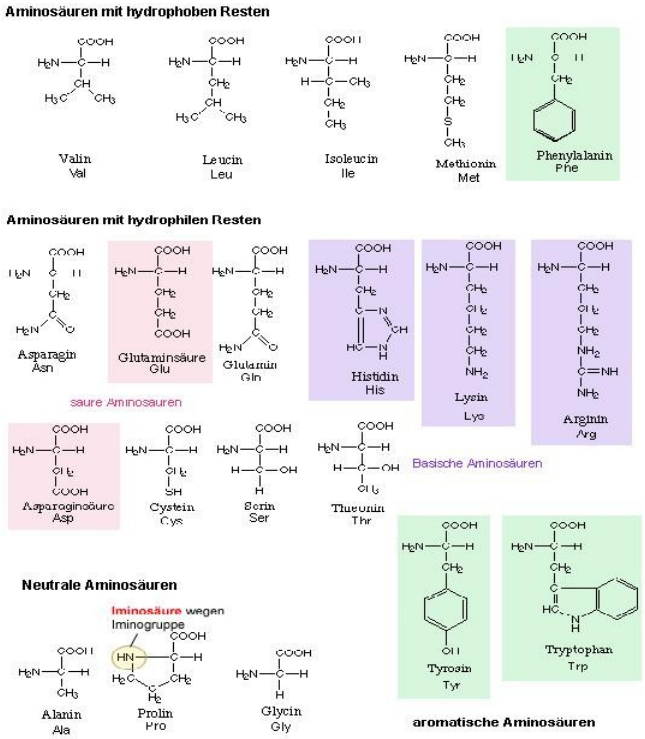
A 20x20 score matrix S reflecting amino acid similarity can be used in protein alignment



```
... KLYMCWA ...
... KAWMCYL ...
```

S(K,K)+S(L,A)+...+S(A,L)

How do we quantify amino acid similarity?



<u>Goal:</u>

Very low scores for pairs of amino acids who's replacement often poses a lot of stress on the 3D structure

Positive scores for amino acid pairs corresponding to mutations that in average do not stress a structure very much

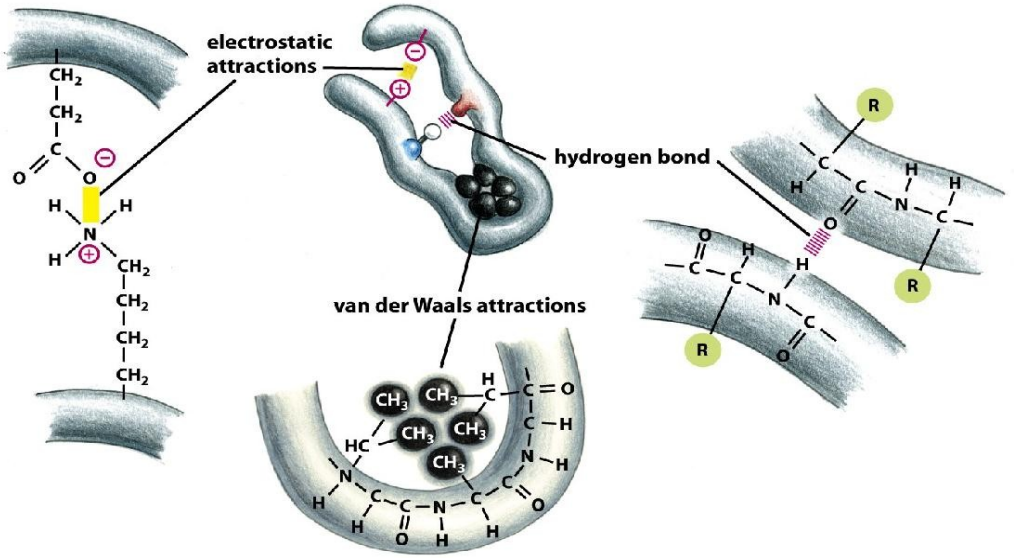
How do we quantify the stress a mutation causes in a structure?

This sounds like a biochemical problem

Changes in polarity cause stress by changing hydrogen bonds

Changes in size cause stress

Changes in electrostatic attraction cause stress



How can we balance the importance of size, polarity and electrostatic potential in a single score function?

We could do a huge experiment

Introduce all possible (20x19)/2 mutations in many proteins at many different positions and evaluate how harmful they are to the structure in average.

This certainly would be an interesting experiment but it is day dreaming.

Doing the mutations is hard but may be possible. Reevaluating the structures is harder and probably not feasible.

And by the way …

How would we measure stress on the 3D structure?

Evolution did the experiment for us

```
WYFGKITRRESERLLLNAENPRGTFLVRESETTKGAYCLSVSDFDNAKGL – human
W+F   + R+E+++LLL   ENP GTFLVR SE      Y LSV D+++ +G  – sequence matches
WFFENVLRKEADKLLLAEENPEGTFLVRPSEHNPNGYSLSVKDWEDGRGY – Drosophila
1         10          20          30          40          50
```

All (20x19)/2 types of mutations happened millions of times in all kind of proteins at all kind of positions.

Natural selection did the job of evaluating stress on 3D structures for us: If there was too much stress the protein lost its function and the gene disappeared from the gene pool.

Amino acid similarity is amino acid exchangeability

Start from aligned homologous sequence pairs.

Collect all mismatch pairs.

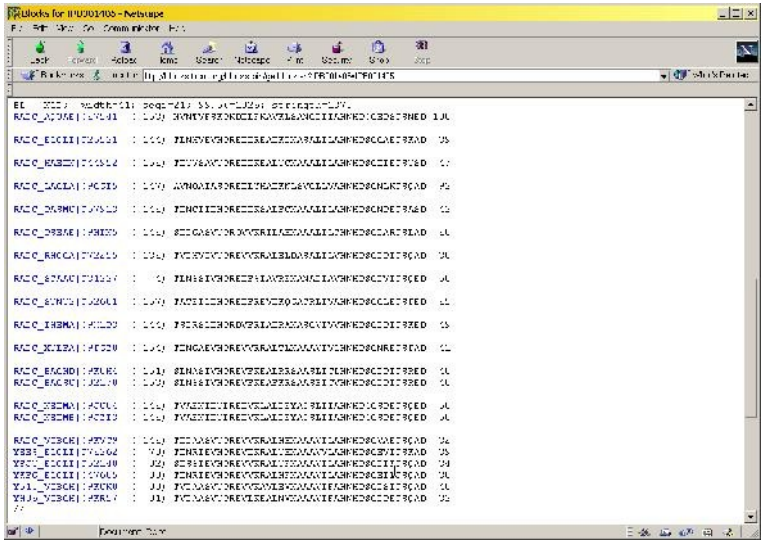
Those that occur frequently consist of similar amino acids.

Those that are rare, are rare because of the 3D structure stress the substitution typically causes. They are dissimilar.

There is a hen and egg problem here: For the score matrix we need alignments, for the alignment we need a score matrix.

We can start with easy to align sequences

If we have more than 90% identical positions between two proteins their alignment is easy and does not need a sophisticated alignment score



The BLOCKS database collects gap free alignments of closely related and hence easy to align sequences

These alignments still have 10% mismatch positions

We can count types of mismatches in the BLOCKS alignments

Our goal is a 20x20 score matrix that gives us a score for every pair of amino acids (x,y)

Does it make sense to give the pair (L,C) a different score than the pair (C,L)?

Remember:
C: Cysteine can build cysteine bridges. If a cysteine is replaced by another amino acid the bridge is gone. That does not happen when a lycine is replaced by a cysteine

The fact that some pairs of amino acids are more often observed then others can be formalized by probabilities

Our goal is a 20x20 score matrix that gives us a score for every pair of amino acids (x,y)

Let p_{xy} be the relative frequency of the aligned pair (x,y)

Is it possible that the pair p_{LC} is different from p_{CL}?

Remember:
C: Cysteine can build cysteine bridges. If a cysteine is replaced by another amino acid the bridge is gone. That does not happen when a lycine is replaced by a cysteine

There is a difference between a mutation and the mismatches in BLOCKS

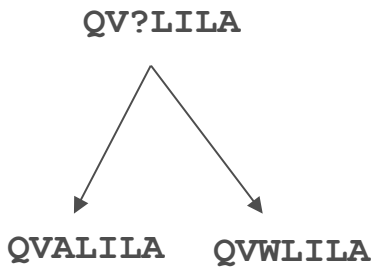
Mutations are directed

```
Ancestor:          ... KALMY ...

Descendent:        ... KALPY ...
```

Here an M mutated to a P

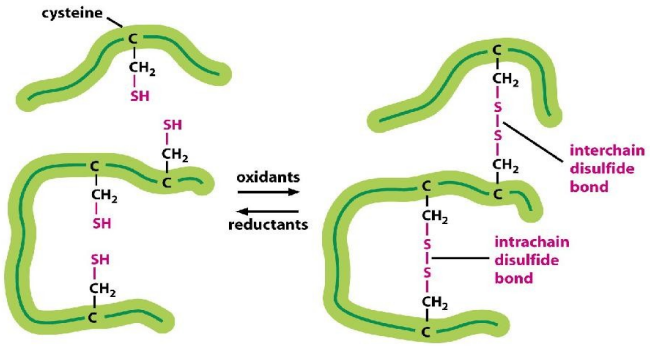
The mismatch positions in BLOCKS are not directed



The mutation process might be non symmetric, but we can not observe this because we only observe symmetric mismatches: *Hence* $p_{xy} = p_{yx}$

Does it make sense to set $S(x,x) \neq S(y,y)$ for different amino acids x and y ?

The diagonal entries of the score matrix contain the self-similarity of amino acids



There are amino acids that can more easily be replaced by others.

While others are so unique that they hardly ever mutate at all without severe problems for the protein's function.

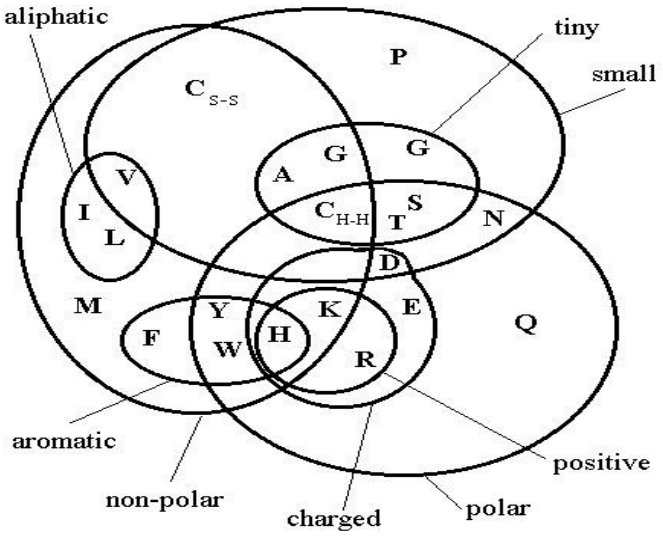
Cysteine is unique in that it builds cysteine bridges with other cysteines in the protein. The bridges greatly stabilize the structure.

In fact we see mismatch positions involving a cysteine only rarely.

Do mismatch frequencies directly translate into amino acid exchangeability?

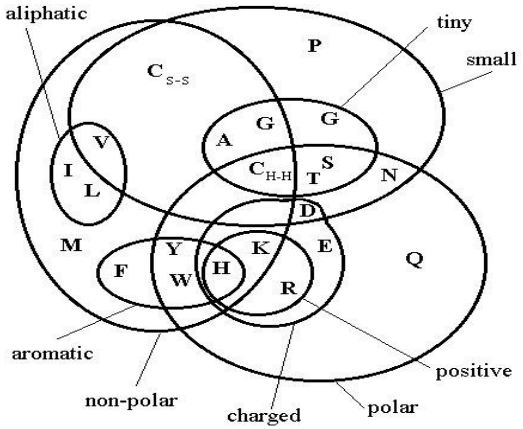
We observe the pair (L,A) much more often in BLOCKS than the pair (W,Y).

Does this mean L is more similar to A than W is to Y?



Amino acids usage in proteins has a non uniform distribution

AA	rel H'keit	% L H'keit
L	0.098	100
A	0.077	78.9
S	0.071	72.4
V	0.067	68.5
G	0.066	67.1
E	0.063	64.9
K	0.059	60.6
I	0.059	60.6
T	0.057	58.1
D	0.054	55.1
R	0.05	51.3
P	0.048	48.8
N	0.046	47.4
F	0.041	42.3
Q	0.041	41.9
Y	0.032	33.2
M	0.022	22.6
H	0.022	22.4
C	0.014	14.8
W	0.013	13



The substitution (L,A) is observed frequently because proteins have many L and many A.

The (Y,W) count was low not because the amino acids are dissimilar but because both Y and W are rare amino acids.

We can use a background model to correct for non uniform amino acid usage

Like with the splice site profile matrix we generate a background model of aligned pairs of amino acids

Let p_x and p_y be the relative frequencies of x and y.

What is the probability to observe the pair (x,y) by chance?

If we assume independence: $p_x \times p_y$

Foreground model: p_{xy}
Background model: $p_x \times p_y$

The log odds of foreground vs. background model yield a score matrix for protein alignments

$$S(x,y) = \log\left(\frac{p_{xy}}{p_x\, p_y}\right)$$

Or for a gap free local alignment

$\ldots x_1\ x_2\ \ldots\ x_n\ \ldots$

$\ldots y_1\ y_2\ \ldots\ y_n\ \ldots$

S(alignment)=S(x_1,y_1)+…+S(x_n,y_n)

$$= \sum \log\left(\frac{p_{x_i\, y_i}}{p_{x_i}\, p_{y_i}}\right)$$

This modeling approach embeds alignment scoring into statistics

Model for homologous sequences (M1):

i.i.d. sequence X_i of amino acid pairs

with $P(X_i=(x,y)) = p_{xy}$

Model for random sequence pairs (M2):

i.i.d. sequence Y_i of amino acid pairs

with $P(Y_i=(x,y)) = p_x \times p_y$

$$\log\left(\frac{\prod p_{x_i y_i}}{\prod p_{x_i} p_{y_i}}\right)$$

Likelihood of M1

Likelihood of M2

The degree of conservation in the BLOCKS alignments influences the counts and hence the scores

If we count pairs of aligned amino acids in closely related sequences we obtain.

High positive values for S(x,x)
And negative values for S(x,y), $x \neq y$

This is only good to align highly conserved sequences because it hardly allows for any mismatch at all.

However, aligning closely related sequences is easy and can be done with almost every score function without problems.

Scores derived from more diverged sequences have positive non-diagonal entries

If we count on more diverse sequence pairs the scores become closer to each other and some
S(x,y) will become positive although x ≠ y

This score takes the alignment of different but similar amino acids as evidence for conservation and not divergence.

```
WYFGKITRRESERLLLNAENPRGTFLVRESETTKGAYCLSVSDFDNAKGL – human
W+F   + R+E+++LLL   ENP GTFLVR SE      Y LSV D+++ +G  – sequence matches
WFFENVLRKEADKLLLAEENPEGTFLVRPSEHNPNGYSLSVKDWEDGRGY – Drosophila
1         10          20          30          40          50
```

Alignments from divergent sequences are less reliable than those of closely related ones.

The BLOSUM matrices use more divergent sequence pairs to count

BLOSUMx:

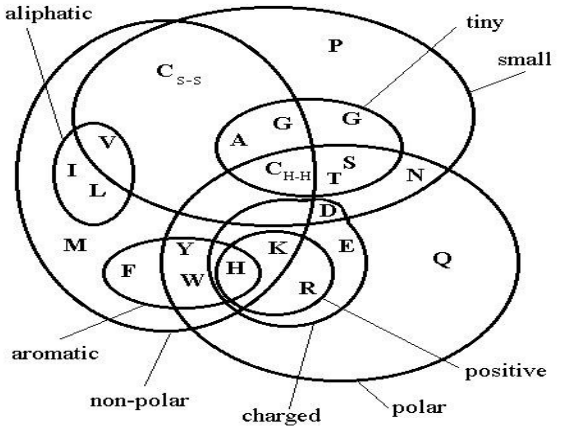
Use only sequence pairs with about x% identity in the alignment

Calculate p_{xy} and go straight for a log odds score without any normalization or extrapolation

BLOSUM62 refers to about 62% identity
BLOSUM80 refers to about 80% identity etc.

BLOSUM62 is a general purpose matrix and the default matrix of most database search engines

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4



BLOSUM62 is derived from sequences with 62% identity. Its practical use however is in alignments of much weaker homologies (30%-35% identity is challenging)

Percent identity is not an all purpose measure for sequence similarity

Why did we develop complicated scores if sequence divergence can be quantified by %identity?

The percent identity is typically higher if the local alignment is short.

We can not align two sequences by optimizing the percent identity.

Otherwise we just match two identical amino acids and get a local alignment of length 1 with 100% identity.

End of Chapter 8