Sequence-based mutation analysis

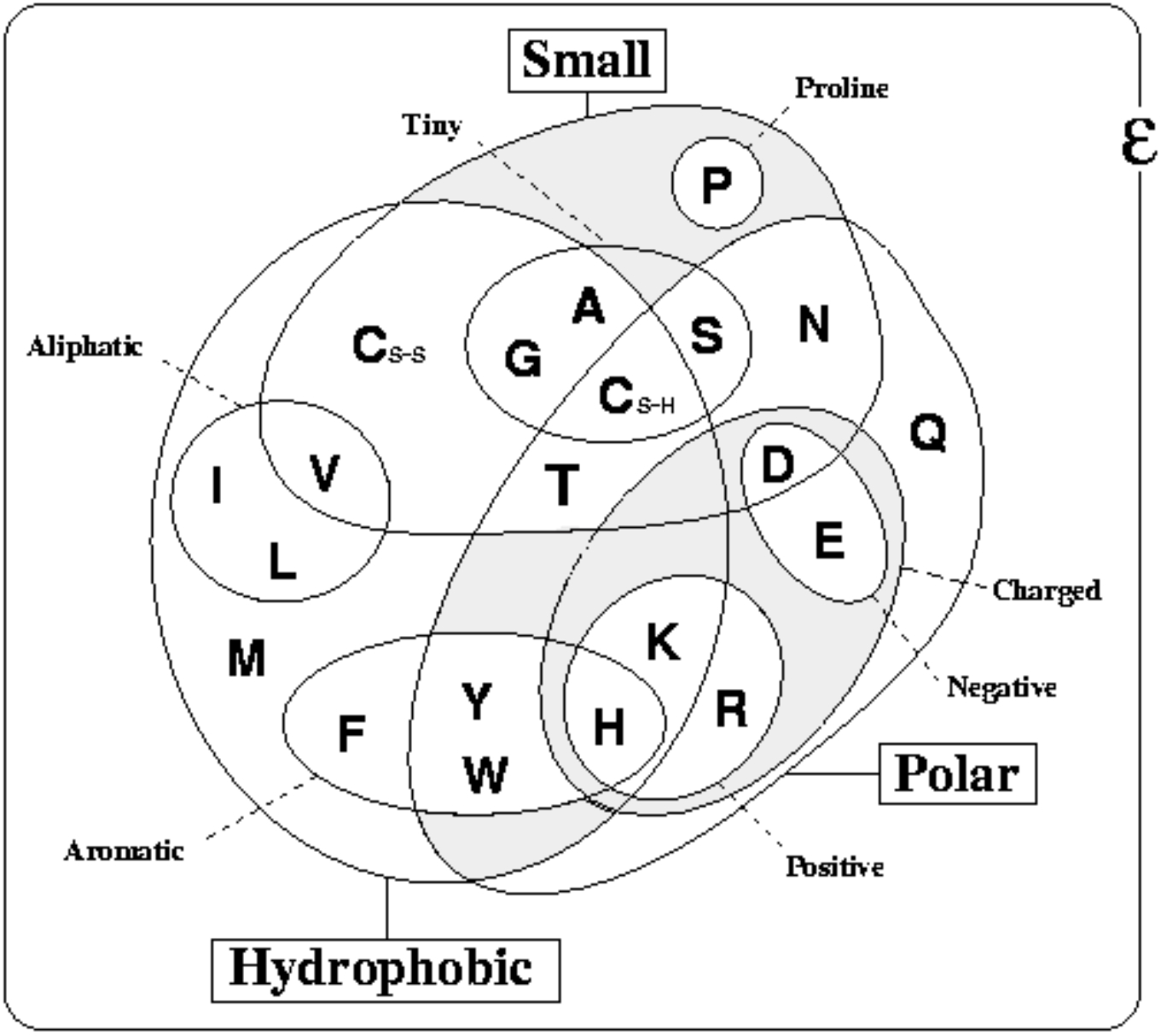
SNP effects prediction from sequence

Sebastian Hollizeck, Robert Wagner

Practical Bioinformatics 'Protein Structure and Function Analysis'

June 11, 2012

Amionacids review



Aminoacids overview

chemical properties of amionacids

- acidic
- aliphatic
- aromatic
- basic
- hydroxyl
- sulfur containing

Aminoacids overview

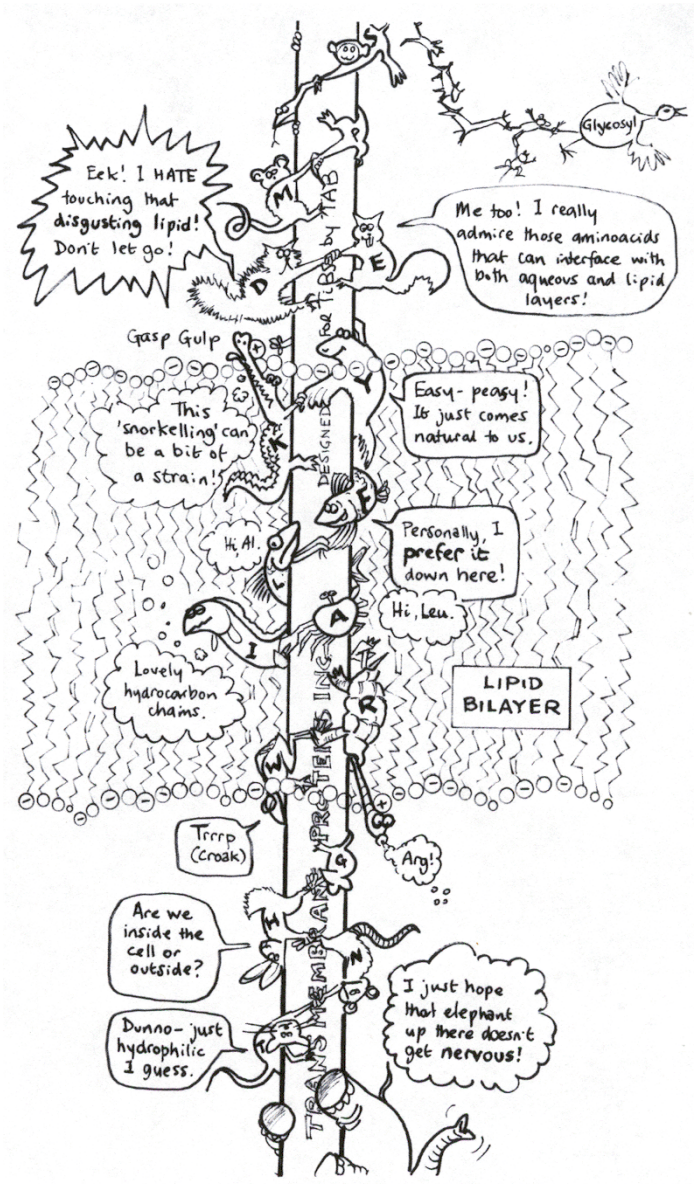


Figure: Killian et al. 2000

Acidic amino acids

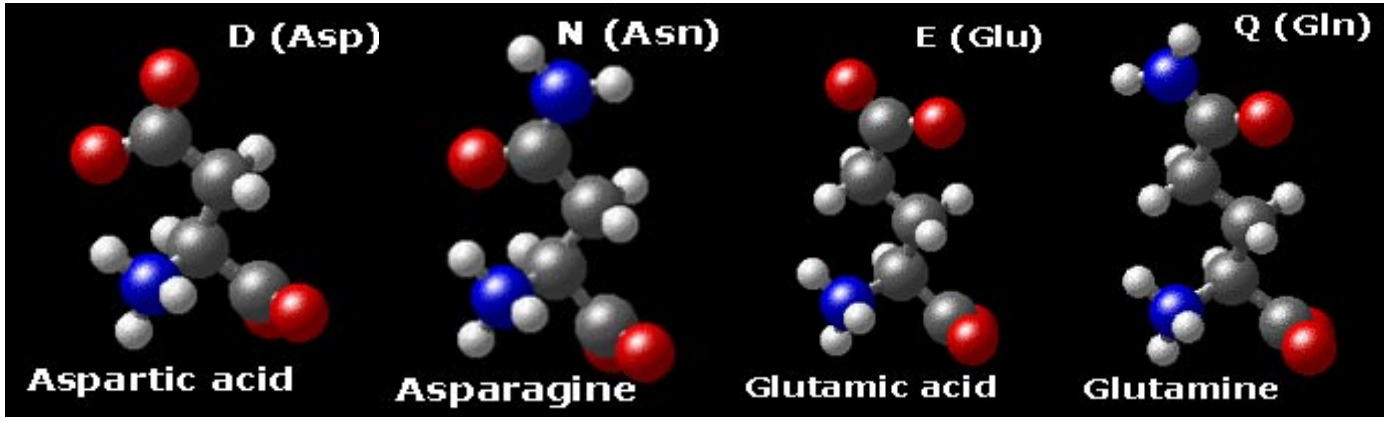


Figure: Acidic amino acids and their amides
all very hydrophilic

Aliphatic amino acids

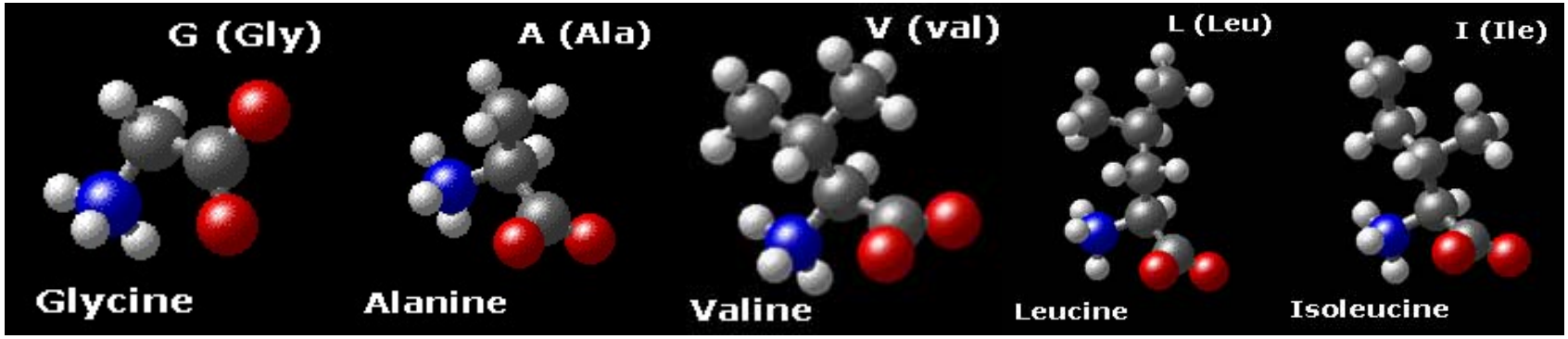


Figure: Aliphatic amino acids
increase in hydrophobicity from left to right

Aromatic amino acids

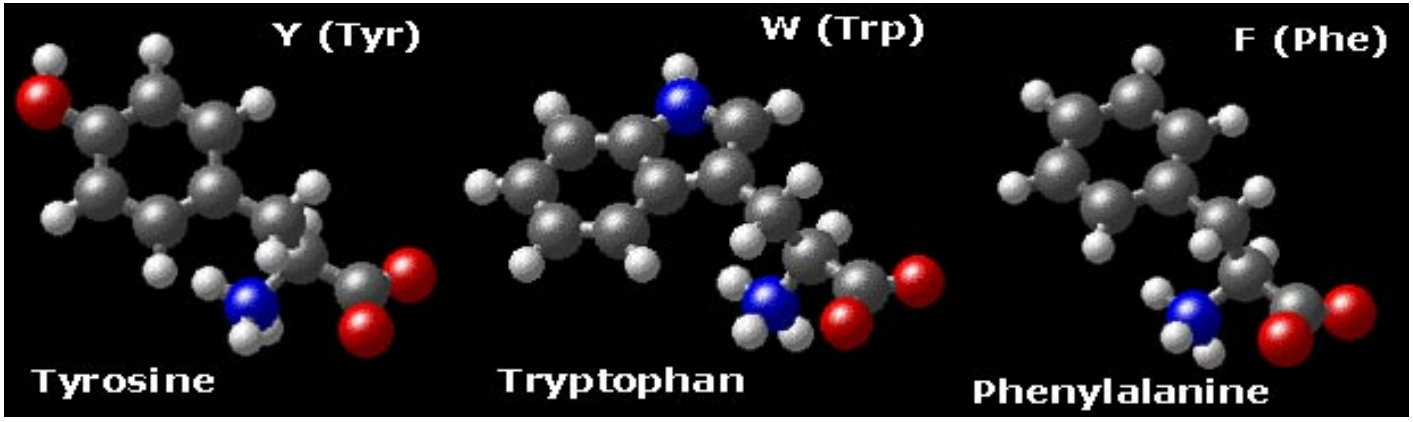


Figure: Aromatic amino acids
increase in hydrophobicity from left to right

Basic amino acids

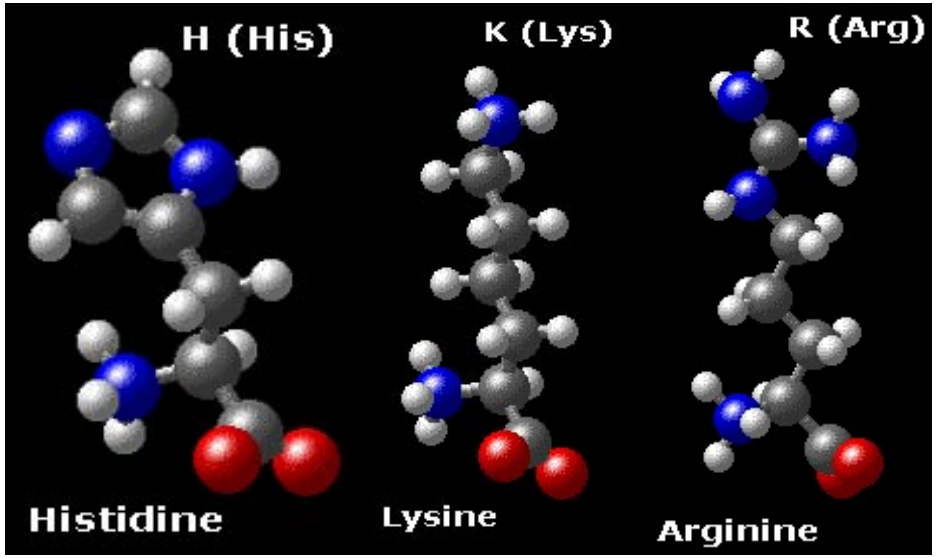


Figure: Basic amino acids

Cyclic amino acids

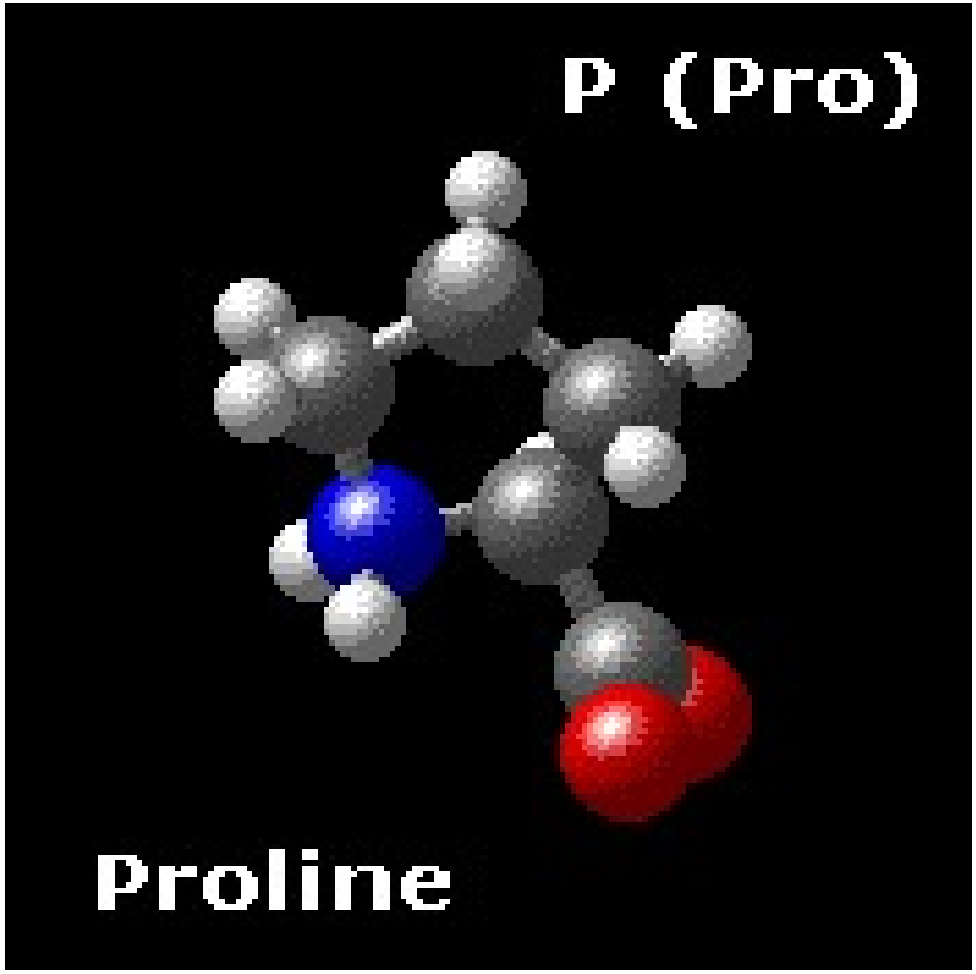


Figure: Proline - the helixbreaker

Hydroxyl amino acids

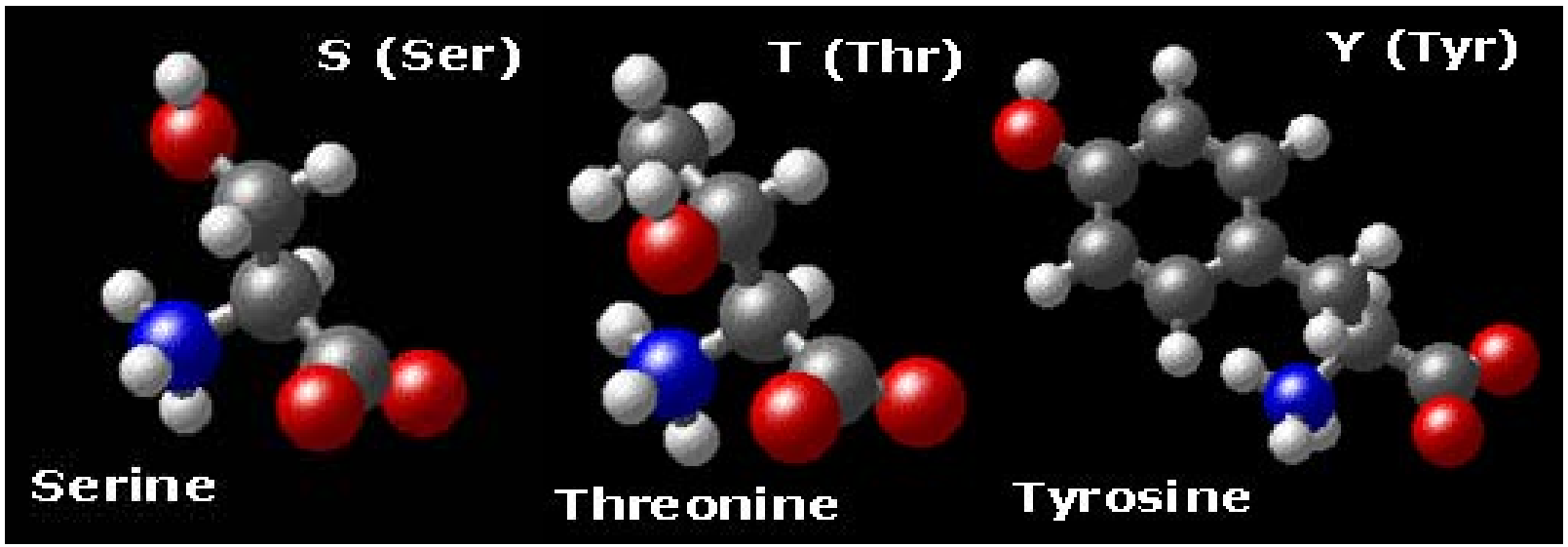


Figure: Hydroxyl amino acids

Sulfur containing amino acids

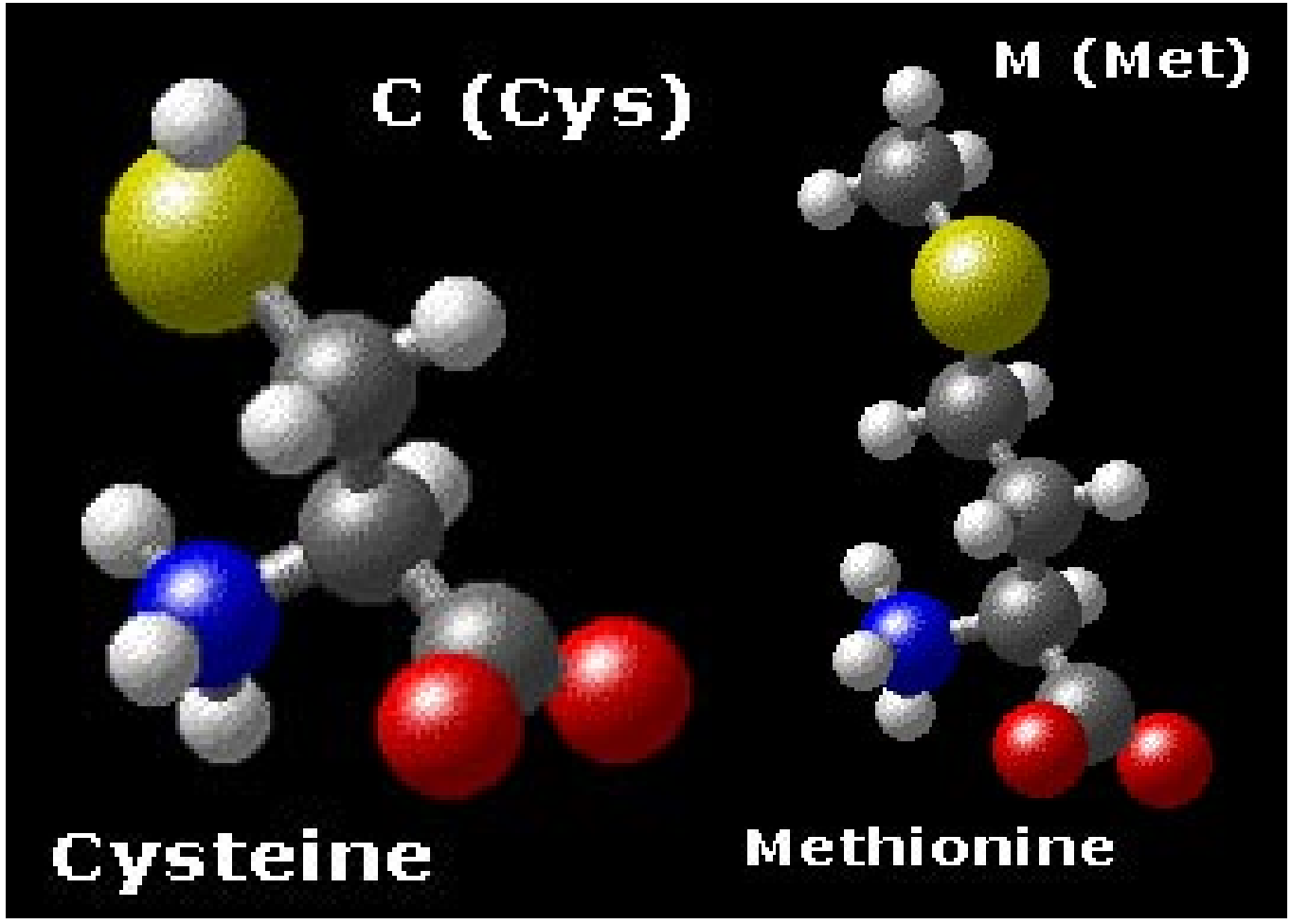


Figure: Sulphur containing amino acids

Comparing amino acids

From Glycine to Leucine
From Alanine to Tryptophan
From Asparagine to Glutamine

Substitution matrices

common substitution matrices

- PAM
- BLOSUM

substitution matrices for special occasions

- secondary structure based
- context based

PAM means **P**oint **A**ccepted **M**utations

- assumes independence of mutations
- for closely related sequences
- for more divergent sequences power the original PAM-matrix

most common PAM1, PAM30, PAM70 and PAM250

BLOSUM

BLOSUM means **BLO**ck **SU**bstitution **M**atrix
approach for more distant related sequences

- derived from distantly related sequences
- finding conserved regions in the multiple alignment
- clustering of sequences to reduce impact of high similarity

most common BLOSUM62 (used in BLAST)

Comparsion PAM vs BLOSUM

aspects	PAM	BLOSUM
evolutionary model	explicit	implicit
based on	global alignment	conserved regions
weighting of mutations	equal	clustering
naming for similarity	lower number	higher number

Prediction tools

For the practical we will use the following tools to predict effects of SNPs:

- SIFT
- PolyPhen2
- SNAP

Sorting **i**ntolerant **f**rom **t**olerant substitutions

- Uses sequence homology
- Searches similar sequences
- Choses closely related sequences
- Obtains multiple sequence alignment

- Calculates normalized possibilities for all possible substitutions for each position
- Every substitution below a given cutoff is assumed to be deleterious
- Based on BLOSSUM Matrix

Polymorphism **Phen**otyping v2

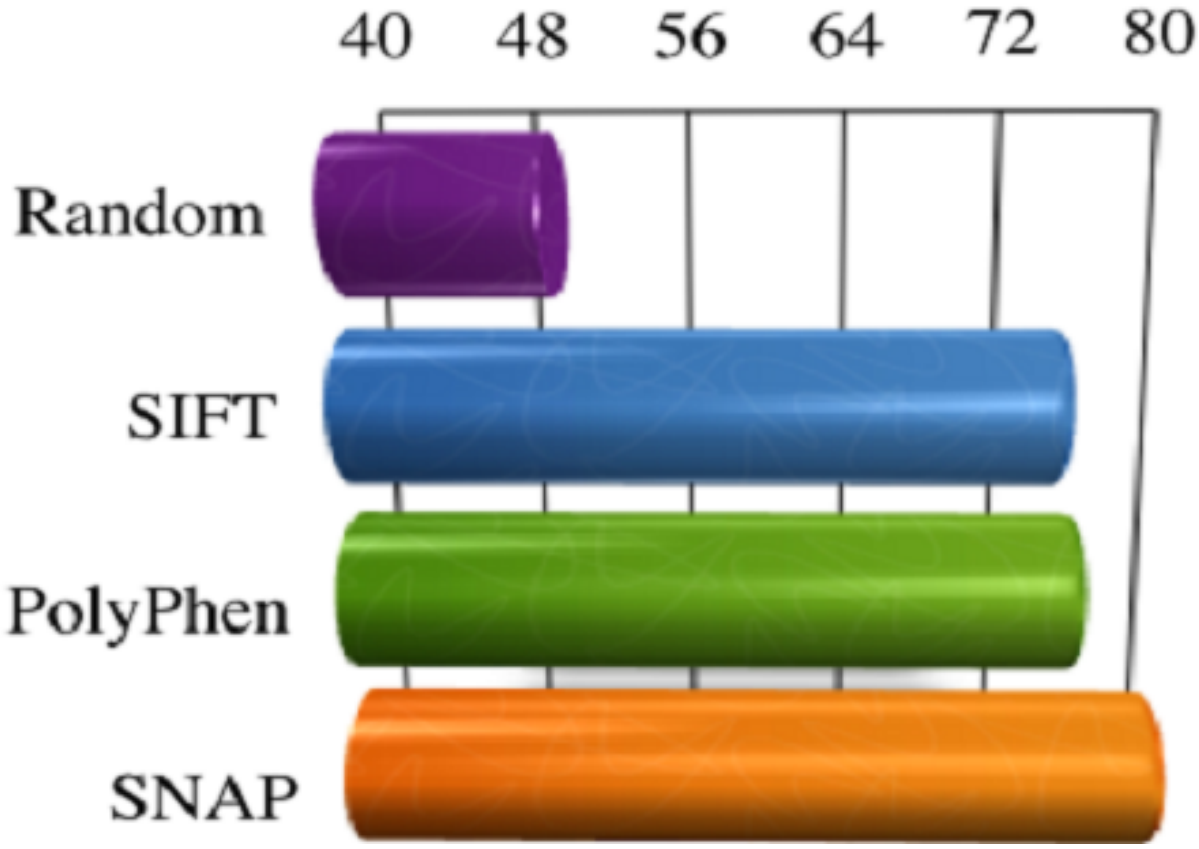
- Basic idea very similar to SIFT - but using machine learning this time (Naive Bayes classifier)
- Searches homologous sequences
- Creates multiple sequence alignment
- Predicts how likely two alleles are to occupy a site, given the amino acid substitution pattern from the alignment

PolyPhen2

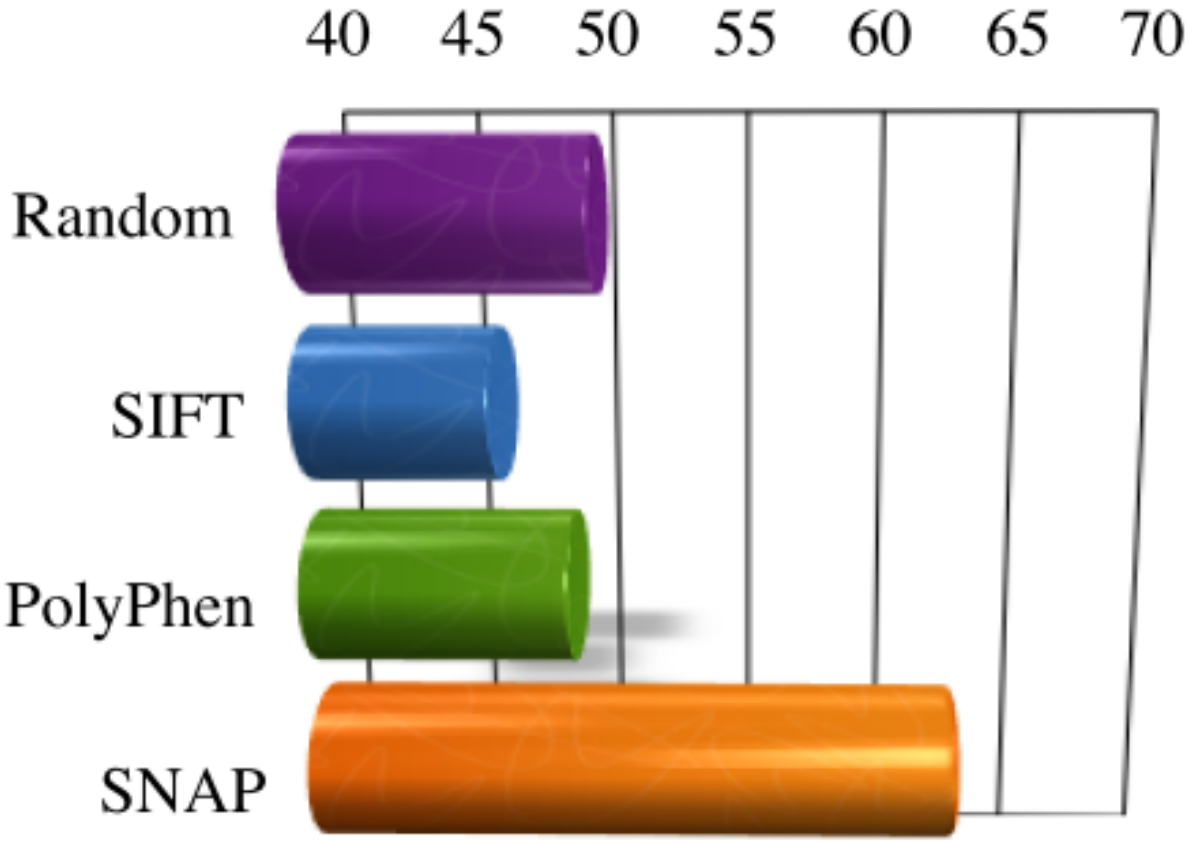
- Also uses other sources of information, like UniProt FT entries, whether or not the mutated position is annotated trans-membrane, signal, active site or binding site for example
- Amino acid substitution calculations based on PSIC score (psi blast)
- According to their paper they perform better than SIFT and SNAP

Screening for **n**on-**a**cceptable **p**olymorphisms

- Uses a neural network to predict SNP effects
- Some of the features used:
  - psi-blast frequency profile
  - relative solvent accessibility prediction (PROFace)
  - secondary structure prediction (PROFsec)
  - Pfam information
  - PSIC scores
  - predicted residue flexibility (PROFbval)
  - window around mutation side (five amino acids)
- According to their paper they work better than SIFT and PolyPhen, especially on tough cases



SNAP evaluation on complete dataset.



SNAP evaluation on cases where methods disagree.

Usage Example

Usage example of SIFT, PolyPhen and SNAP.

End

Thank you for listening!

For further information about the methods (especially SNAP) please see Protein Prediction 2 Lecture 15 - Slides or Video