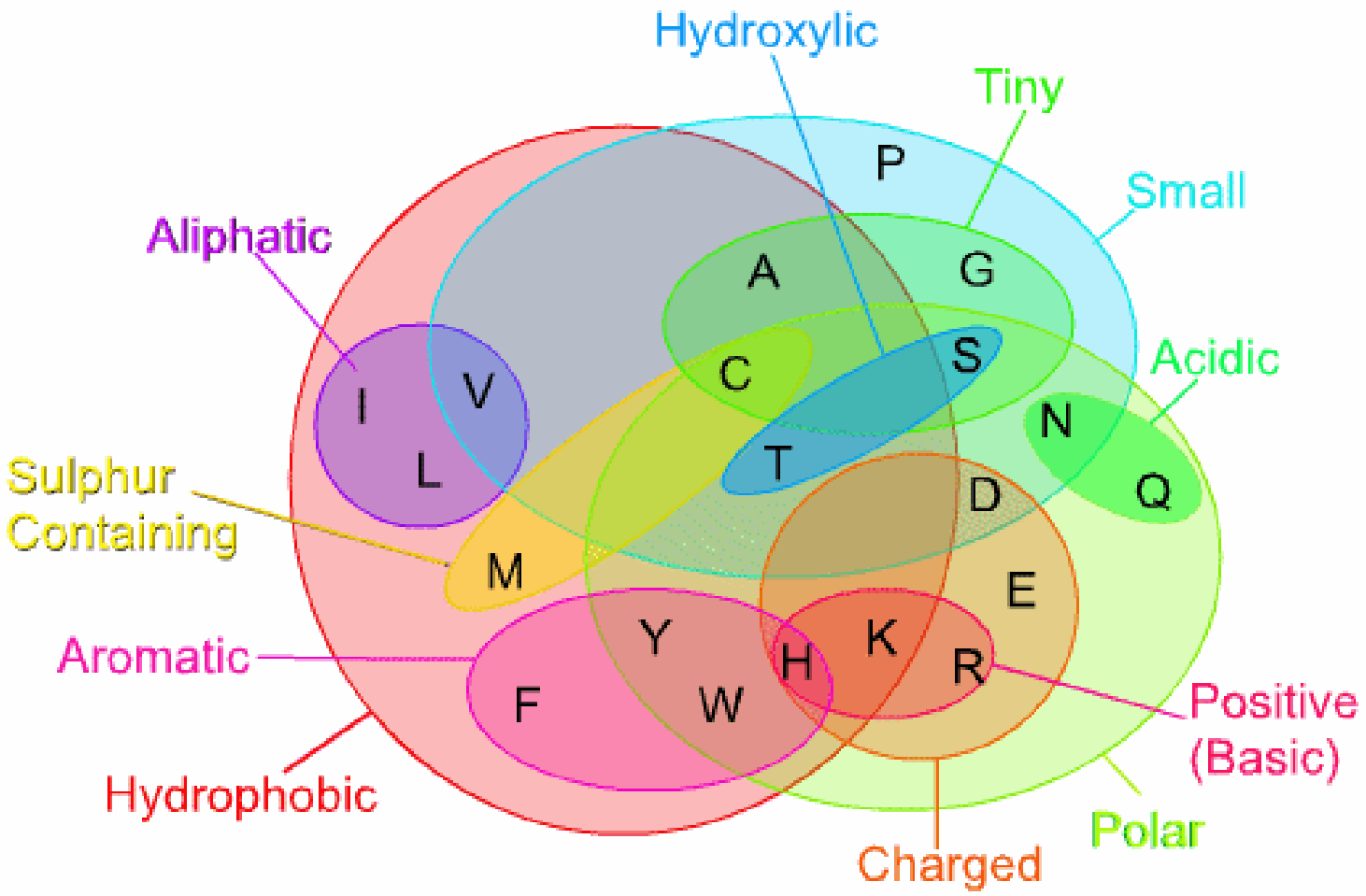
Sequence based mutation analysis

Carina Demel
Eva Reisinger

Sequence based mutation Analysis

- Amino Acid Properties
- Substitution Matrices
- SNPs
- Polyphen
- SIFT
- SNAP
- Comparison

Amino Acid Properties



Amino Acids

A alanine (ala)
R arginine (arg)
N asparagine (asn)
D aspartic acid (asp)
C cysteine (cys)
Q glutamine (gln)
E glutamic acid (glu)
G glycine (gly)
H histidine (his)
I isoleucine (ile)
L leucine (leu)
K lysine (lys)
M metioneine (met)
F phenyalanine (phe)
P proline (pro)
S serine (ser)
T threonine (thr)
W trytophan (trp)
Y tyrosine (tyr)

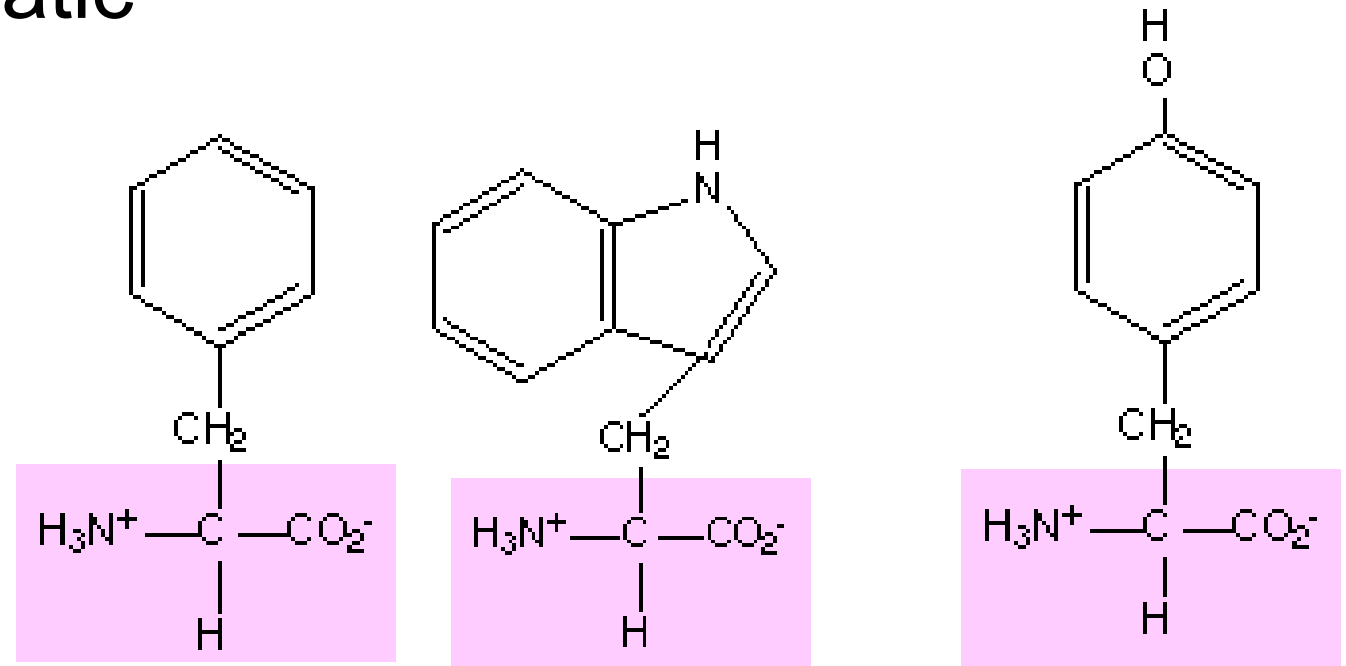
http://www.weightlossandnutritionsecrets.com/wp-content/uploads/2010/07/Amino-Acids-Chart.gif

Amino Acid Properties

- Hydrophobic
  - Aromatic
  - Aliphatic
  - Sulphur containing
- Polar
  - Hydroxylic
  - Charged
  - Acidic
  - Positive (basic)

Amino Acid Properties

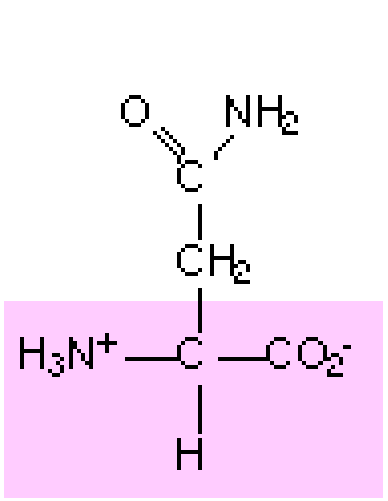
- Aromatic



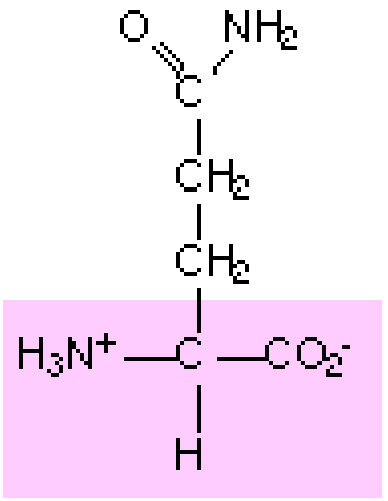
Phe **Trp** **Tyr**

Amino Acid Properties

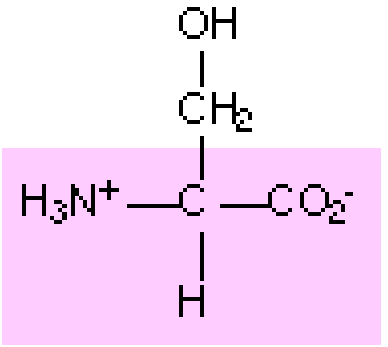
- Polar/ Uncharged:



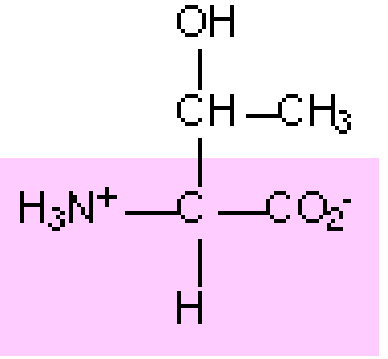
Asn



Gln



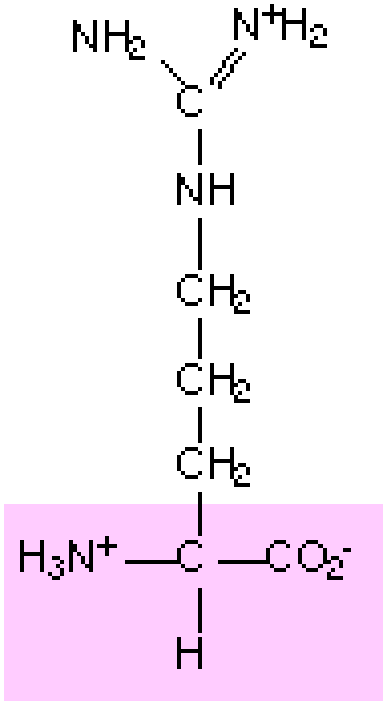
Ser



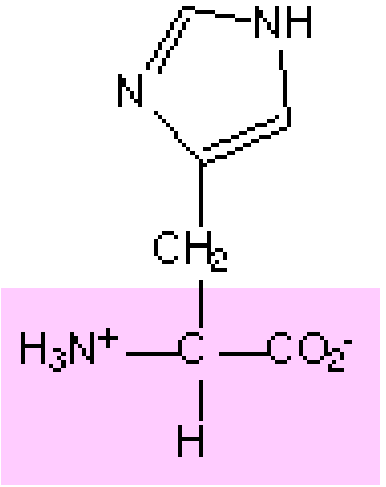
Thr

Amino Acid Properties

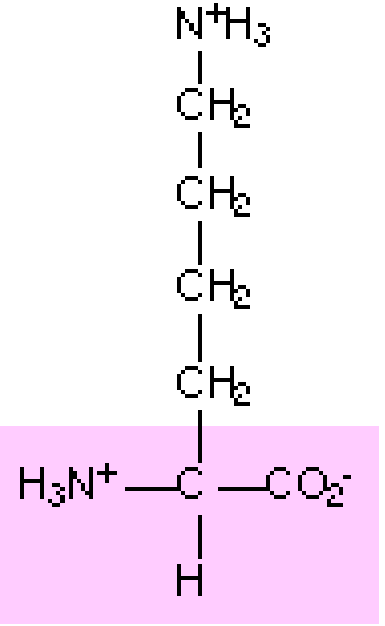
- Positively Charged



Arg



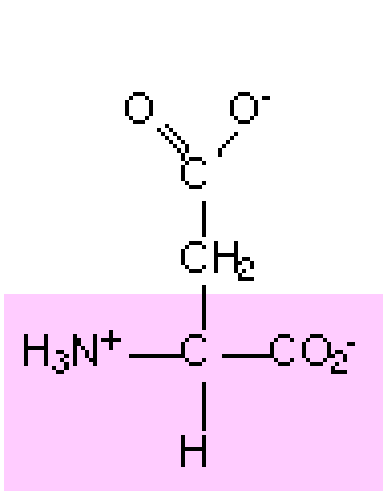
His



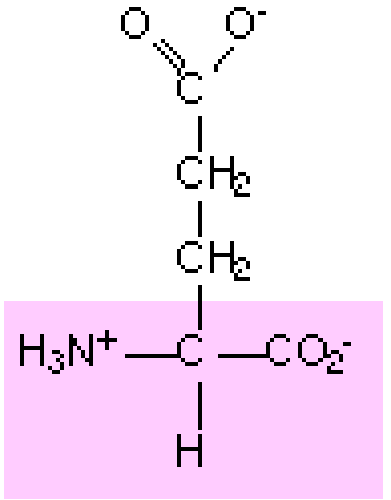
Lys

Amino Acid Properties

- Negatively Charged



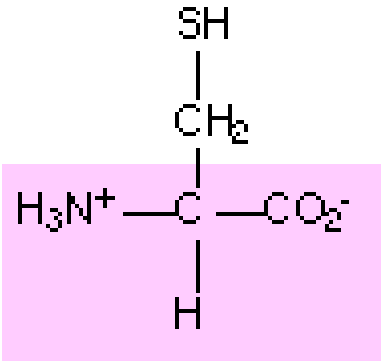
Asp



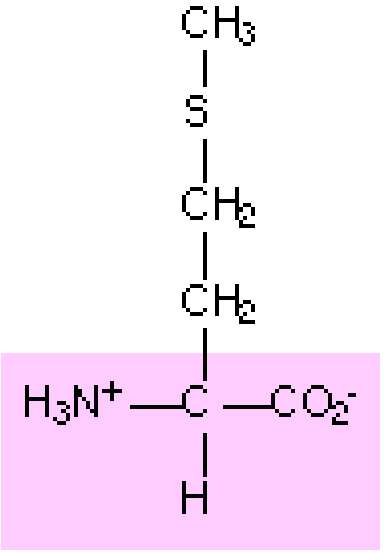
Glu

Amino Acid Properties

- Sulfur Containing



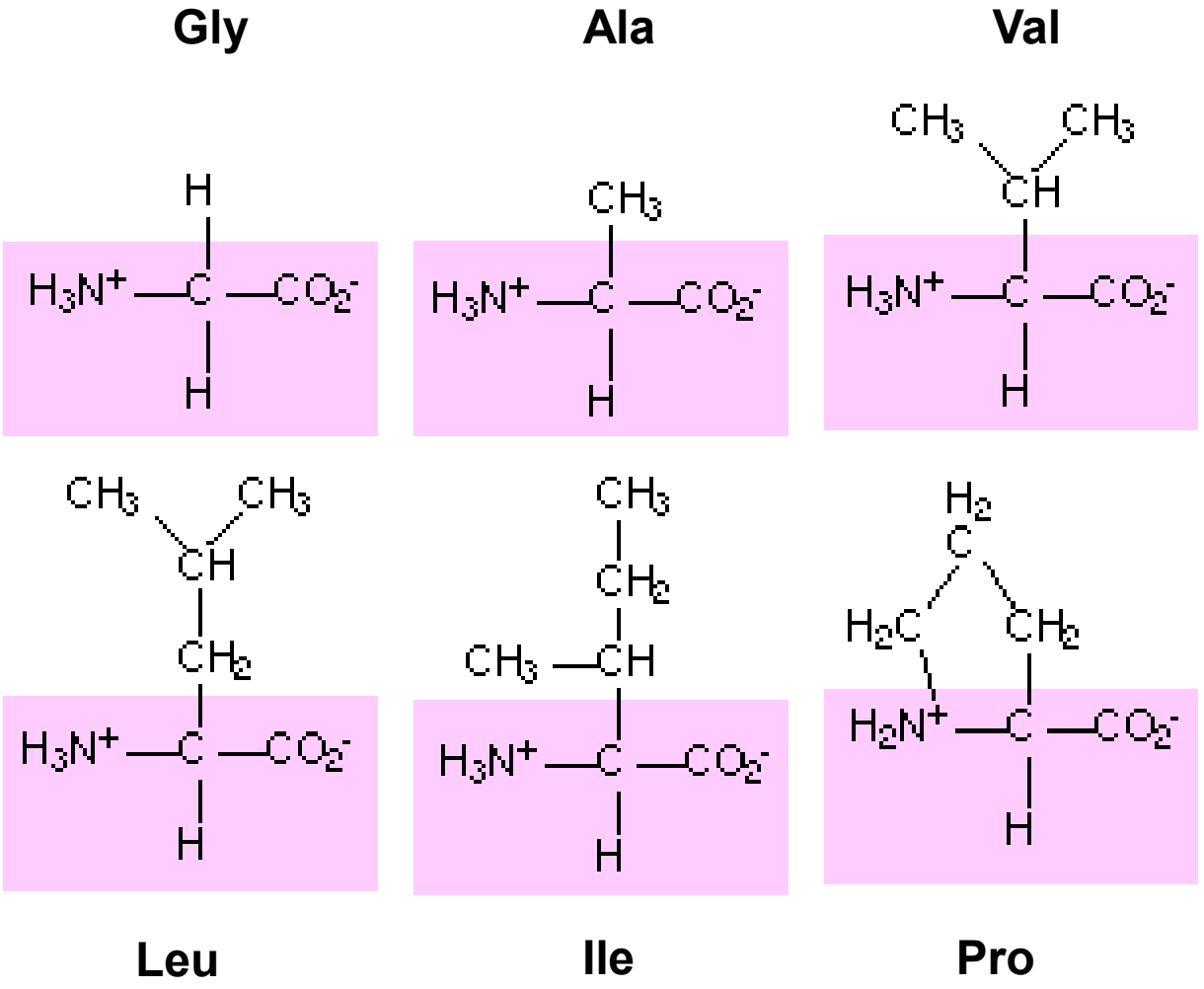
Cys



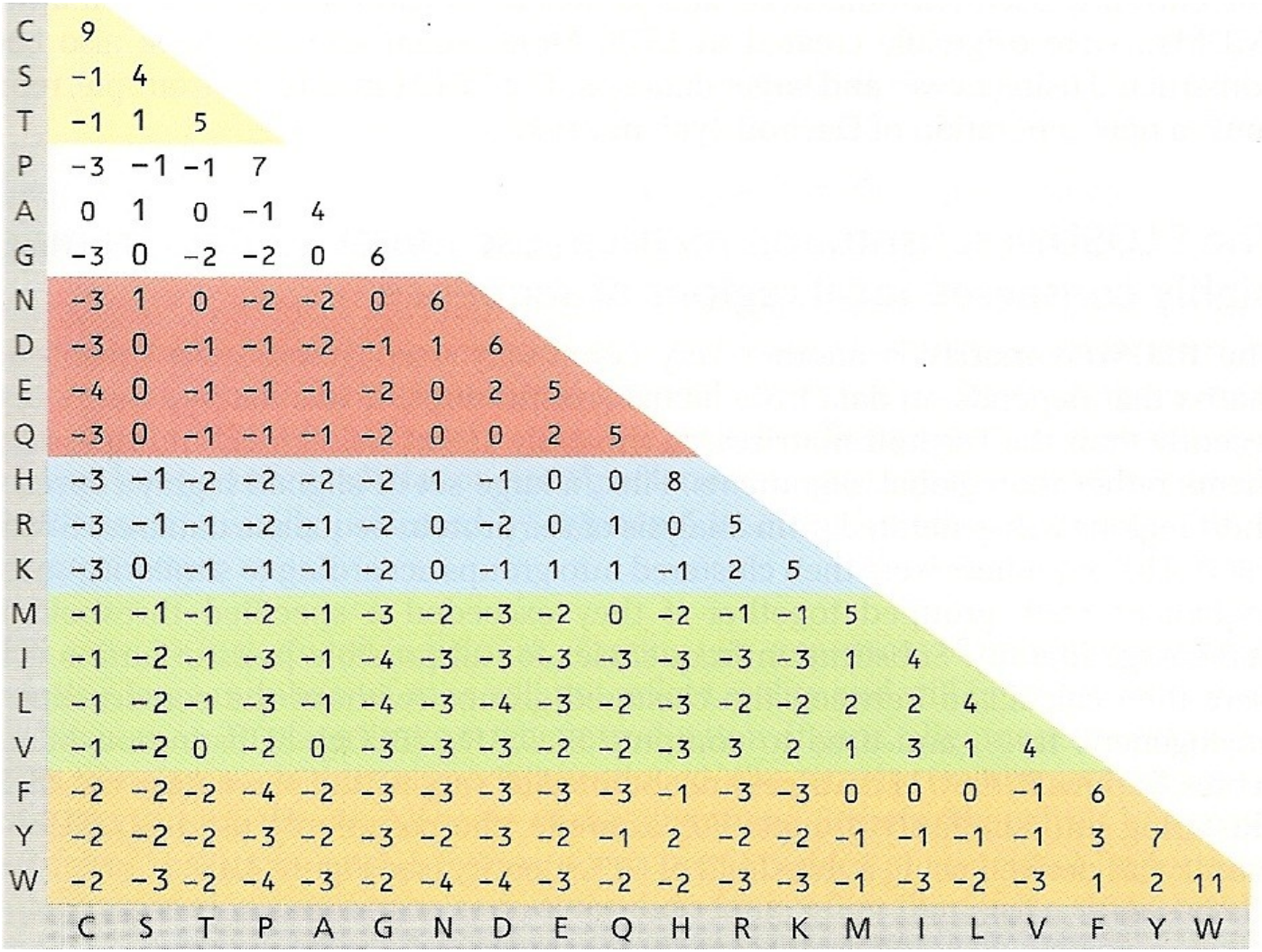
Met

Amino Acid Properties

- Aliphatic
  - Tiny:
    G, A
  - Small:
    V, P



Substitution matrices



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	3	2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

BLOSUM 62

Substitution Matrices

- Reflect relative rate in which one AA changes to another AA during evolution
- Often symmetric
- Based on:
  - Chemical properties of AA
  - Empirical data (PAM, BLOSUM)

PAM

- Point Accepted Mutation- Matrix
- Margret Dayhoff, 1978
- Created by observing differences in closely related proteins
- Based on 71 protein families with 85% identical amino acids
- → based on an evolutionary model
- Symmetric ($P(A \rightarrow B)=P(B \rightarrow A)$)

PAM

- Assumptions:
  - Mutations are independent
  - Substitutions are independent of their position in the sequence
- PAM1: rates for substitution if 1% of AA has changed → 99% similarity
- PAM2 = PAM1 * PAM1
- PAM250: 250 mutations have been fixed per 100 residues: sequence similarity 20%

PAM1

- Builing a 1-PAM matrix *M*:
  - List of accepted mutations
  - Probability of occurence p_a for each amino acid *a* (in a large, sufficiently varied sequence set) $\sum_a p_a = 1$
- Assumptions:
  - Accepted mutations are undirected: $a \leftrightarrow b$
  - Mutations are immediate: $a \rightarrow b$ and not $a \rightarrow c \rightarrow b$

PAM1

- From the list of accepted point mutations compute f_{ab}, the frequency of mutations $a \leftrightarrow b$
- Undirected mutations: $f_{ab} = f_{ba}$
- Total number of mutations of *a:* $f_a = \sum_{b \neq a} f_{ab}$
- Total number of aa occurances involved in mutations:

$$f = \sum_a f_a$$

PAM1

- Relative mutability of a = probability that the given aa a will change in the evolutionary period of interest

$$m_a = \frac{number\ of\ changes}{number\ of\ occurances} = \frac{f_a}{p_a}$$

Aligned Sequences	A	D	A
	A	D	B

Amino Acids	A	B	D
Observed changes (f_a)	1	1	0
Frequency of occurrence (total) (p_a)	3	1	2
Relative Mutability	0.33	1	0

$$m_a = \frac{f_A}{p_A} = \frac{1}{3} = 0.33$$

- Scaled to number of replacements of aa a per 100 residues in the alignments

$$m_A = \frac{f_A}{100\ f\ p_A}$$

Relative mutability

Relative Mutabilities of the Amino Acids[a]

More mutable					Less mutable
Asn	134		His	66	
Ser	120		Arg	65	
Asp	106		Lys	56	
Glu	102		Pro	56	
Ala	100		Gly	49	
Thr	97		Tyr	41	
Ile	96		Phe	41	
Met	94		Leu	40	
Gln	93		Cys	20	
Val	74		Trp	18	

[a] The value for Ala has been arbitrarily set at 100.

- Most mutable: Asn, Ser, Asp, Glu
- Least mutable: Cys and Trp

PAM1

- Probability of a remaining unchanged: $M_{aa} = 1 - m_a$
- $M_{ab} = P(a \to b) = P(a \to b | a\ changed) * P(a\ changed) = \frac{f_{ab}}{f_a} m_a$
- Normalizing of M_{aa} and M_{ab} → Transitionmatrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**9867**	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	1	**9913**	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	4	1	**9822**	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	6	0	42	**9859**	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	1	1	0	0	**9973**	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	3	9	4	5	0	**9876**	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	10	0	7	56	0	35	**9865**	4	2	3	1	4	1	0	3	4	2	0	1	2
G	21	1	12	11	1	3	7	**9935**	1	0	1	2	1	1	3	21	3	0	0	5
H	1	8	18	3	1	20	1	0	**9912**	0	1	1	0	2	3	1	1	1	4	1
I	2	2	3	1	2	1	2	0	0	**9872**	9	2	12	7	0	1	7	0	1	33
L	3	1	3	0	0	6	1	1	4	22	**9947**	2	45	13	3	1	3	4	2	15

PAM1

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	3																		
T	-3	2	4																	
P	-3	1	-1	6																
A	-3	1	1	1	3															
G	-5	1	-1	-2	1	5														
N	-5	1	0	-2	0	0	4													
D	-7	0	-1	-2	0	0	2	5												
E	-7	-1	-2	-1	0	-1	1	3	5											
Q	-7	-2	-2	0	-1	-3	0	1	2	6										
H	-4	-2	-3	-1	-3	-4	2	0	-1	3	7									
R	-4	-1	-2	-1	-3	-4	-1	-3	-3	1	1	6								
K	-7	-1	-1	-2	-2	-3	1	-1	-1	0	-2	2	5							
M	-6	-2	-1	-3	-2	-4	-3	-4	-4	-1	-4	-1	0	8						
I	-3	-2	0	-3	-1	-4	-2	-3	-3	-3	-4	-2	-2	1	6					
L	-7	-4	-3	-3	-3	-5	-4	-5	-4	-2	-3	-4	-4	3	1	5				
V	-2	-2	0	-2	0	-2	-3	-3	-3	-3	-3	-3	-4	1	3	1	5			
F	-6	-3	-4	-5	-4	-5	-4	-7	-6	-6	-2	-4	-6	-1	0	0	-3	8		
Y	-1	-3	-3	-6	-4	-6	-2	-5	-4	-5	-1	-6	-6	-4	-2	-3	-3	4	8	
W	-8	-2	-6	-7	-7	-8	-5	-8	-8	-6	-5	1	-5	-7	-7	-5	-8	-1	-1	12

- Derive transition probabilites for larger amounts of evolution (M^k) via matrix muliplication

PAMk = PAMk-1 * PAM1

- Scoring matrix: *logarithm of odds (lod)* matrix *S*

$$S_k ab = 10 \log_{10} \frac{M^k_{ab}}{p_b}$$

BLOSUM

- BLOcks Substitution Matrix
- Henikoff & Henikoff 1992
- Based on ungapped local alignments = blocks
- Block ≈ conserved region of a protein family
- BLOSUM62 based on blocks with 62% sequence identity

BLOSUM

1. Deriving a frequency table from blocks
   - Count all possible pairs in each column of each block → f_{ab}
2. Calculate a Logarithm of Odds (lod) Matrix
   - Obs probability of occurance for each pair *a, b* → $q_{ab} = f_{ab} / \sum_{a=1}^{20} \sum_{b=1}^{a} f_{ab}$
   - Exp probability of *a*th aa in an a, *b* pair is:

$$p_a = q_{aa} + \frac{\sum_{b \neq a} q_{ab}}{2}$$

...A...
...A...
...A...
...A...
...A...
...A...
...S...
...A...
...A...
...A...

$$f_{AA} = 8 + 7 + \ldots 1 = 36$$
$$f_{AS} = 9$$

$$q_{AA} = \frac{36}{45} = 0.8$$
$$q_{AS} = \frac{9}{45} = 0.2$$

$$p_A = \frac{36 + \frac{9}{2}}{45} = 0.9$$
$$p_S = \frac{(\frac{9}{2})}{45} = 0.1$$

BLOSUM

...A...
...A...
...A...
...A...
...A...
...A...
...S...
...A...
...A...
...A...

- Exp probability of occurance for each *a,b* pair: $e_{ab} = p_a * p_b \; for \; a = b$

  $e_{ab} = p_a * p_b + p_b * p_a \; for \; a \neq b$

$e_{AA} = 0.9 * 0.9 = 0.81$

$e_{AS} = 2 * (0.9 * 0.1)$

$e_{AA} = 0.1 * 0.1 = 0.01$

- Odds ratio matrix

  each entry: q_{ab} / e_{ab}

- Lod ratio: $s_{ab} = \log_2(q_{ab} / e_{ab})$ $\longrightarrow$ multiplied by scaling factor, rounded $\longrightarrow$ Scoring matrix

  observed frequencies as expected: $s_{ab} = 0$

  less than expected: $s_{ab} < 0$

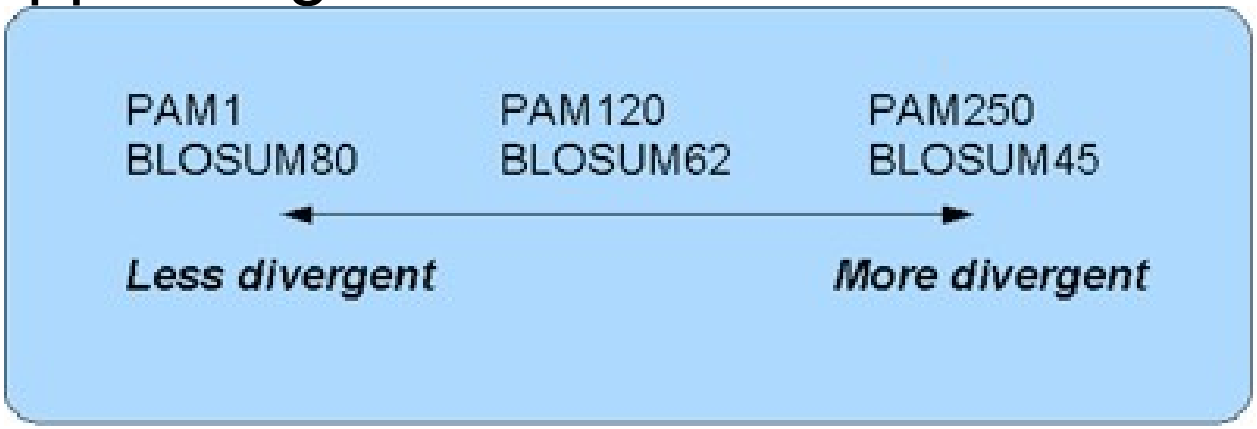
  more than expected: $s_{ab} > 0$

PAM vs BLOSUM

- PAM: basted on explicit evolutionary model,

  BLOSUM: based on protein families
- PAM: based on mutations observed in a global alignment, with highly conserved and mutable regions.

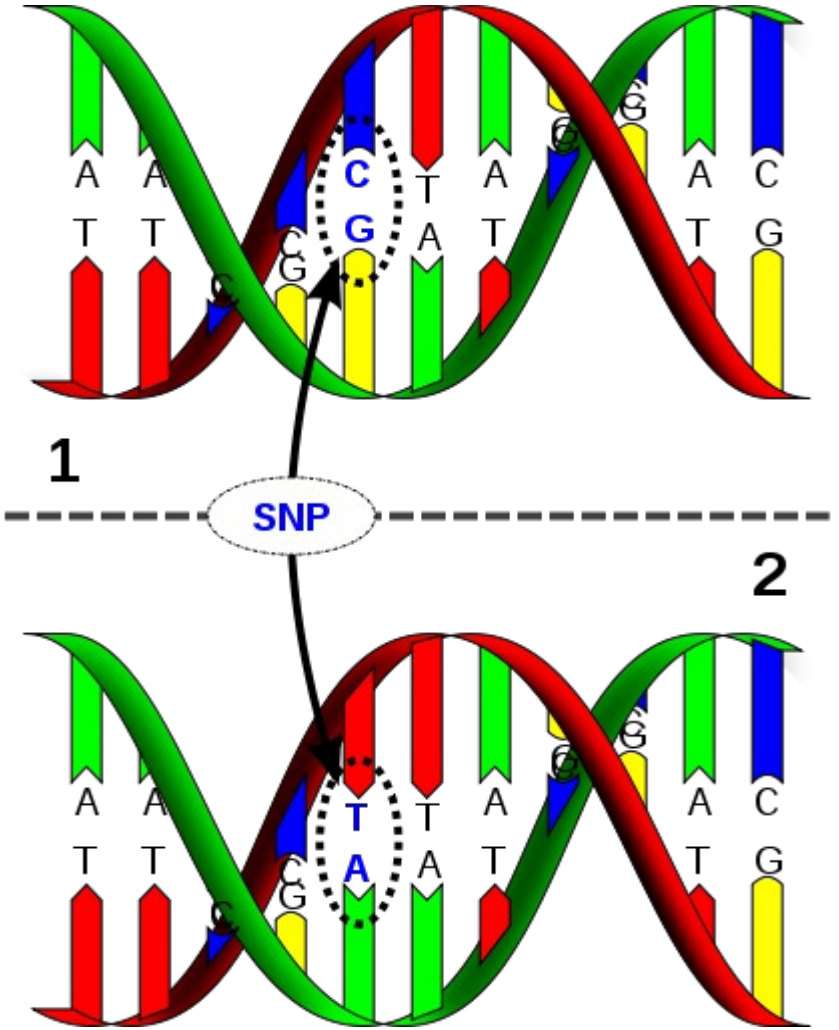
  BLOSUM: based on highly conserved regions in local ungapped alignments
-



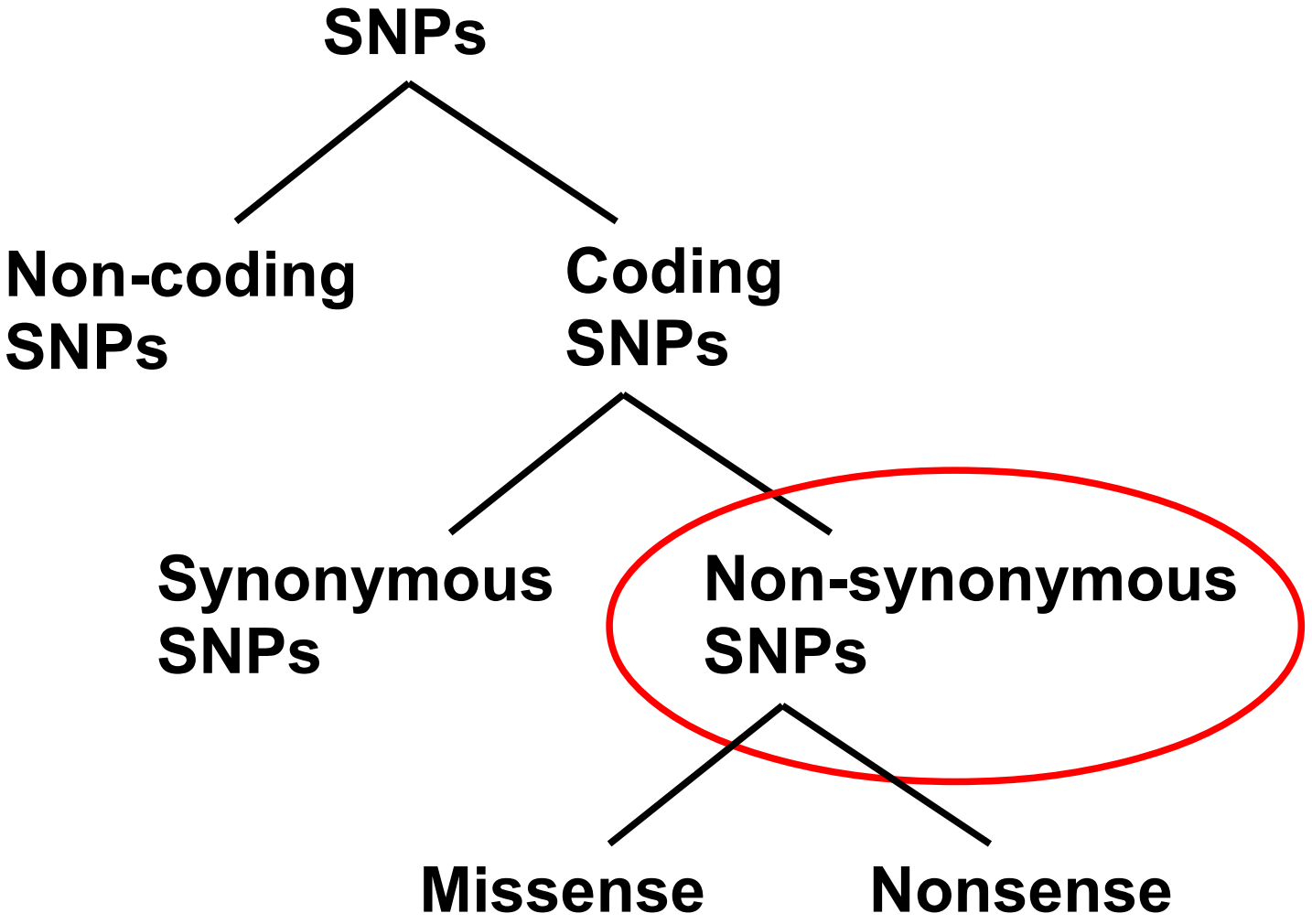
SNPs

- **S**ingle **N**ucleotide **P**olymorphism
- DNA sequence variation
  - between members of a biological species or
  - paired chromosomes in an individual
- Located in coding regions, non-coding regions and intergenic regions

SNPs



SNPs


SNPs
Non-coding SNPs
Coding SNPs
Synonymous SNPs
Non-synonymous SNPs
Missense
Nonsense

SNP databases

- DbSNP
- SNPedia
- OMIM
  - SNPs → diseases
- Human Gene Mutation Database
  - SNPs, human inherited diseases → gene mutations

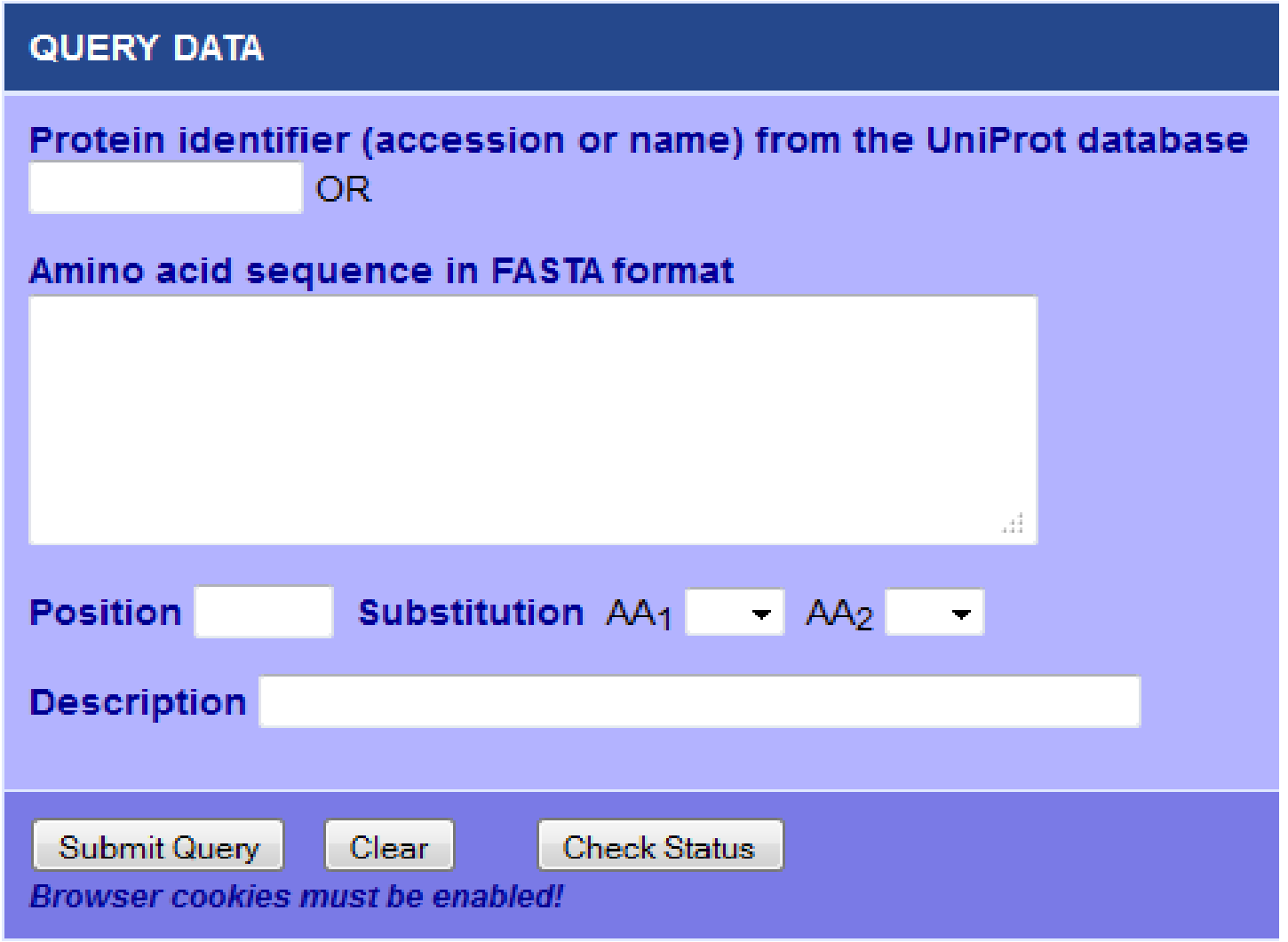
PolyPhen

- Background
- Input
- Procedure
- Prediction
- Output
- PolyPhen2

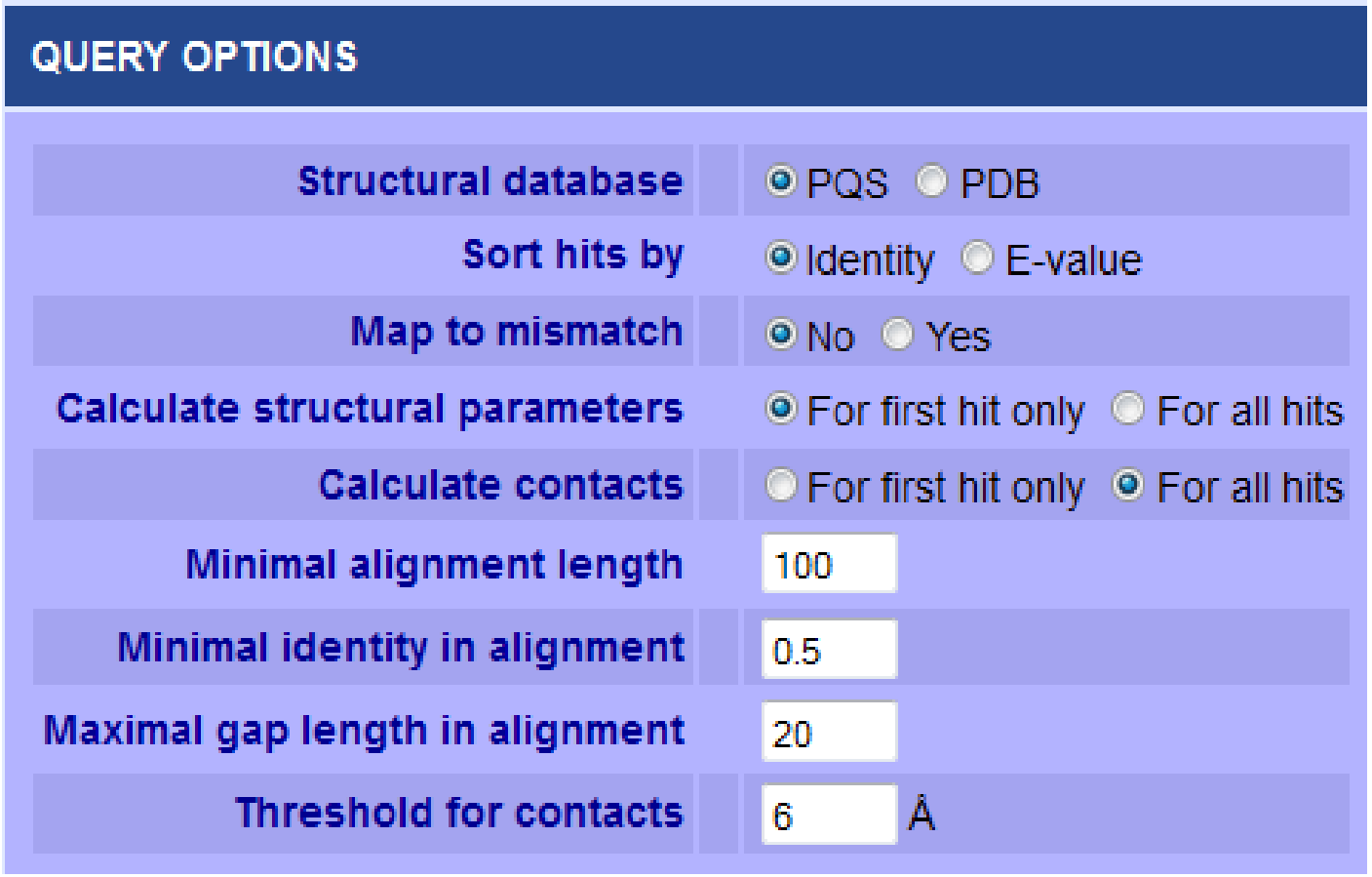
Background

- PolyPhen (=*Poly*morphism *Phen*otyping)
- Automated tool
- Predicts possible impact of an amino acid substitution on the structure and function of a human protein
- Uses straightforward empirical rules which are applied on the information characterizing the substitution

Input


QUERY DATA
Protein identifier (accession or name) from the UniProt database
OR
Amino acid sequence in FASTA format
Position
Substitution
AA1
AA2
Description
Submit Query
Clear
Check Status
Browser cookies must be enabled!

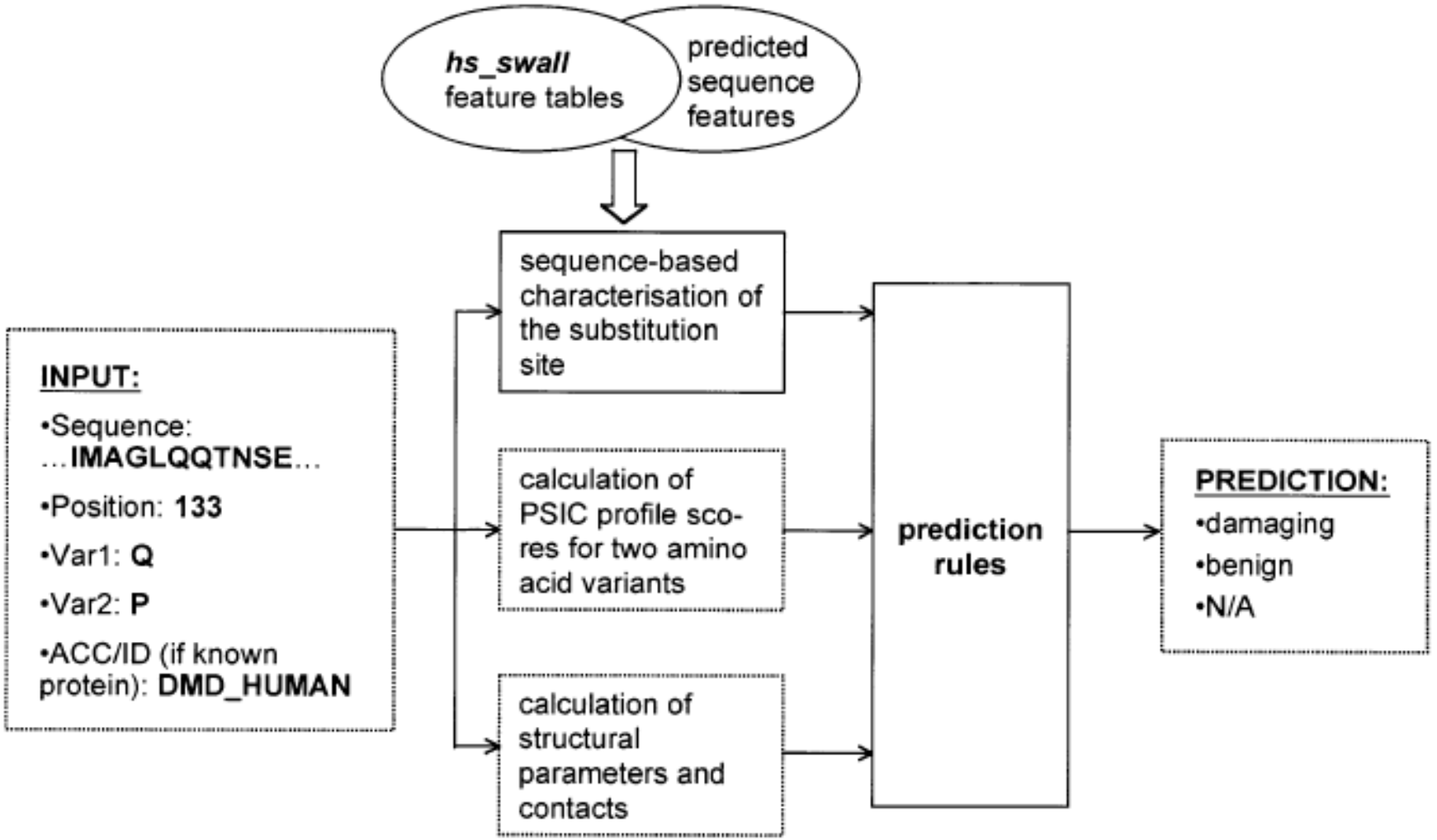
Input


QUERY OPTIONS
Structural database
PQS
PDB
Sort hits by
Identity
E-value
Map to mismatch
No
Yes
Calculate structural parameters
For first hit only
For all hits
Calculate contacts
For first hit only
For all hits
Minimal alignment length
100
Minimal identity in alignment
0.5
Maximal gap length in alignment
20
Threshold for contacts
6
Å

Procedure

- Sequence-based characterisation of the substitution site
- Calculation of PSIC profile scores for two amino acid variants
- Calculation of structural parameters and contacts

Procedure


hs_swall
feature tables
predicted
sequence
features
INPUT:
•Sequence:
…IMAGLQQTNSE…
•Position: 133
•Var1: Q
•Var2: P
•ACC/ID (if known
protein): DMD_HUMAN
sequence-based
characterisation of
the substitution
site
calculation of
PSIC profile sco-
res for two amino
acid variants
calculation of
structural
parameters and
contacts
prediction
rules
PREDICTION:
•damaging
•benign
•N/A

Sequence-based characterisation of the substitution site

- Search query protein in **hs_swall** *(*human proteins subset of UniProt)
- Checks where the substitution is located (different annotations in hs_swall)
- Substitution in an annotated or transmembrane region?

  → Uses **PHAT** (transmembrane-specific substitution matrix) score to evaluate possible functional effect of a nsSNP

Calculation of PSIC profile scores for two amino acid variants

- BLAST to identify homologues of the input
- Retain hits that have
  - Sequence identity of 30- 94%
  - Length of Alignment > 50
- PSIC software (**P**osition-**S**pecific **I**ndependent **C**ounts) to calculate a *profile matrix*

Calculation of PSIC profile scores for two amino acid variants

- Computes difference between profile scores in the polymorphic position
- Big difference
  → Substitution is yet rarely or never observed
- Shows the number of aligned sequences at the query position
  → Used to assess the reliability of profile score calculations

Calculation of structural parameters and contacts

- Mapping of the substitution site to known protein **3D structures** with BLAST
- **Structural parameters** like secondary structure, phi-psi dihedral angles...
  → Some are taken from DSSP
- **Contacts**
  → May reveal the role of a residue for the protein function

Prediction

- Bases on 4 main sources:
  - Sequence annotation describing the substitution position
  - Sequence prediction
  - Multiple alignment
  - Structure

Prediction

- Probably damaging
  → With high confidence supposed to affect protein function or structure
- Possibly damaging
  → Supposed to affect protein function or structure
- Benign
  → Most likely lacking any phenotypic effect
- Unknown
  → No prediction possible

Output

- 3 main sections
  - Query → mostly resembling the input
  - Prediction
  - Details
    - Sequence features of the substitution site
    - PSIC profile scores for two amino acid variants
    - Structural parameters and contacts

Output

Query

Acc number	Position	AA_1	AA_2	Description
21040341	176	C	Y	.1| hemochromatosis protein isoform 3 precursor; hereditary haemochromatosis protein [Homo sapiens]

Prediction

This variant is predicted to be probably damaging

Prediction	Available data	Prediction basis	Substitution effect	Prediction data
probably damaging	FT alignment	alignment	N/A	PSIC score difference: 2.943

Details

PSIC PROFILE SCORES FOR TWO AMINO ACID VARIANTS

Score1	Score2	|Score1-Score2|	Observations	Diagnostics	Multiple alignment around substitution position
+2.415	-0.528	**2.943**	9	precomputed	Sequences: all Flanks: 25 Show alignment

MAPPING OF THE SUBSTITUTION SITE TO KNOWN PROTEIN 3D STRUCTURES

Database	Initial number of structures	Number of structures
PQS	709	0

PolyPhen-2

- Comparison with PolyPhen
  - Similarities
    - Input
    - Procedure
  - Differences
    - Prediction
      - Uses Naïve Bayes classifier to predit the functional significance of a substitution
      - 2 datasets were used to train and test prediction models
        - HumDiv
        - HumVar

SIFT

- Background
- Input
- Procedure
- Output

Background

- **S**orts **i**ntolerant **f**rom **t**olerant amino acid substitutions (=SIFT)
- Sequence homology-based
- Predicts whether an amino acid substitution in a protein will have a phenotypic effect
- Based on the premise that protein evolution is correlated with protein function

Input

- NCBI GI/ RefSeq ID
  → Predictions are based on pre-computed BLAST searches and are returned within a minute
- Protein sequence
  → FASTA format
  → The entire SIFT procedure will be executed and results will be returned to you
  → Slow

Input

- Group of related sequences
  → FASTA format
  → Skip the first two steps of SIFT
  → Fast

- Multiple alignment
  → CLUSTAL, MSF, or FASTA format
  → The first three steps are skipped
  → Very fast

Procedure

- Get related sequences

  → With PSI-BLAST

- Choose closely related sequences
  - Build sequence groups with >90% identity
  - Make consensus sequences for each group
  - Find conserved reagions between query seq. and consensus seq. with MOTIF
  - Extract conserved regions from sequences aligned by PSI-BLAST

Procedure

- Build a checkpoint file with the conserved regions of the query seq and the consensus seqs.
- Search in the consensus seq the remaining conserved regions with PSI-BLAST
- Add the top hit to the alignment in the checkpoint file → calculate the conservation score
- Rebuild the checkpoint file

→ Until: - The score is under a threshold

- The score decreases

Procedure

- Obtain alignment
  - From the initial PSI-BLAST search results
- Calculate probabilities
  - For each amino acid to be at this specific postion

Output

- SIFT Predictions for Substitutions
  - SIFT Score
  - Median Info
  - Seqs at Position
- Genome Tool Output
- Single Protein Output
  - A table of probabilities
  - Predictions for each position

SIFT Predictions for Substitutions

- SIFT Score
  - Score <= 0.05 → damaging
  - Score > 0.05 → tolerated
- Median Info
  - Measure for the diversity of the sequences used for prediction
- Seqs at Position
  - Number of sequences that have an amino acid at the position of prediction

Single Protein Output

- A table of probabilities

pos	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
9I	0.75	0.71	0.12	0.39	0.68	0.35	0.36	0.30	0.81	1.00	0.87	0.24	0.42	0.28	0.54	0.76	0.58	0.58	0.94	0.02	0.39

- Predictions for each position

Predict Not Tolerated	Position	Seq Rep	Predict Tolerated
cwdfmiyvgpshnalte	7Q	0.95	K Q R

nonpolar, uncharged polar, basic, acidic.

SNAP

- Screening for non-acceptable polymorphisms
- Neural network based
- Prediction about the functionality of a mutated protein

SNAP

- Input: protein sequence & substitutions (XposY)
  - *Y: Substitute all residues in sequence by Y (scan)
  - Pos*: substitute the residue in position pos by all other residues
  - X*Y: subsitute all residues X in sequence by Z

SNAP

- Functional/structural annotations beneficial
- Derived in silico protein information
  - Evolutionary information (residue conservation within sequence families)
  - Protein structure (secondary structure, solvent accessibility)
  - ...

SNAP

- Classifies all nsSNPs in all proteins into
  - Neutral (no effect)
  - non-neutral (effect on function)

- Provides Reliability Index (level of confidence of a particular prediction)

SNAP

Input sequence

```
MVNSTHRGMHTSLHLWNRSSYRLHSNASESLGKGYSDGGCYEQLFVSPEVFVTLGVISLL
ENILVIVAIAKNKNLHSPMYFFICSLAVADMLVSVSNGSETIVITLLNSTDTDAQSFTVN
IDNVIDSVICSSLLASICSLLSIAVDRYFTIFYALQYHNIMTVKRVGIIISCIWAACTVS
GILFIIYSDSSAVIICLITMFFTMLALMASLYVHMFLMARLHIKRIAVLPGTGAIRQGAN
MKGAITLTILIGVFVVCWAPFFLHLIFYISCPQNPYCVCFMSHFNLYLILIMCNSIIDPL
IYALRSQELRKTFKEIICCYPLGGLCDLSSRY
```

Input substitutions

```
R7H, S30F, E100A
```

```
Result of SNAP prediction
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Yana Bromberg & Burkhard Rost
NAR (2007)
```

```
# Query  : dict_h19775

nsSNP       Prediction   Reliability   Expected Accuracy
---------  ----------  ------------  -----------------
R7H          Neutral          5                89%
S30F        Non-neutral       4                82%
E100A       Non-neutral       3                78%
```

Comparison

- Test on a subset of the PMD database (Sub-PMD)

Sub-PMD data set		
SNAP	SIFT	PolyPhen
69.5 ± 0.4	70.6 ± 0.5	67.8 ± 0.5
79.9 ± 0.4	73.6 ± 0.5	73.5 ± 0.5
48.3 ± 0.6	55.9 ± 0.6	48.9 ± 0.7
41.0 ± 1.1	36.0 ± 1.0	29.7 ± 1.1
70.9 ± 0.4	NR	NR

- MCC (Matthew's correlation coefficient) → quality
- ROC AUC → Performanz

References/Sources

- Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. (1978). "A model of evolutionary change in proteins". Atlas of Protein Sequence and Structure 5 (3): 345–352
- A. Churbanov, PAM matrix for BLAST algorithm, 2002
- S. Henikoff and J. G. Henikoff, Amino acid substitution matrices from protein blocks. Proc Nat Acad Sci 89(22):10915–10919, 1992
- Yana Bromberg and Burkhard Rost, SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Research, 2007, Vol. 35, No. 11 3823-3835
- http://genetics.bwh.harvard.edu/pph/
- http://genetics.bwh.harvard.edu/pph/snps_server_and_survey.pdf
- http://swift.cmbi.ru.nl/gv/dssp/
- http://genetics.bwh.harvard.edu/pph2/
- http://sift.jcvi.org/
- http://sift.jcvi.org/www/SIFT_help.html#SIFT_OUTPUT_SUBST
- http://www.cs.tau.ac.il/~ruppin/mud.pdf