

Proteins: counts as genetics ok?

Bridging the Bench-Machine Learning Gap

Dr. Emily A. Beck

Dr. Jake Searcy

Learning Objectives

Learn the basics of protein structures and why ML is needed

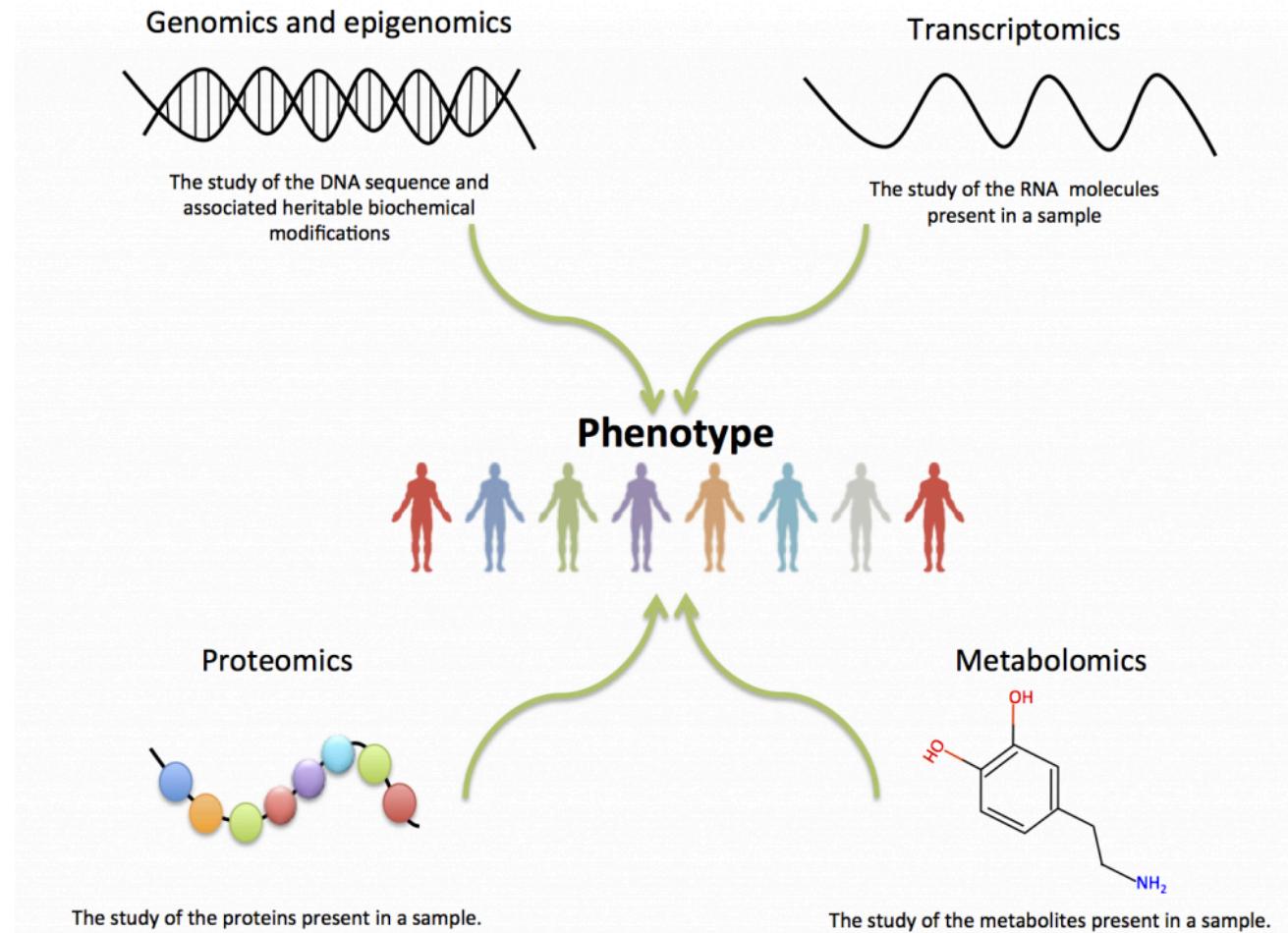
Understand basic structures of protein sequencing files

Learn about the cool things you can do with single amino acid sequences

Learn about cool things you can do with multiple alignments of amino acid seq

After the break we will learn to make a multiple alignment file via command line, run software assessing amino acid changes for deleterious effects on function, predict structure changes in our proteins

Many questions in genomics are trying to get at biological significance of mutations



We have spent a lot of time focusing on SNPs in DNA/RNA.

How those impact amino acids and protein structure is a fundamental question in biology.

There is a suite of questions biologists and biochemists have struggled to answer

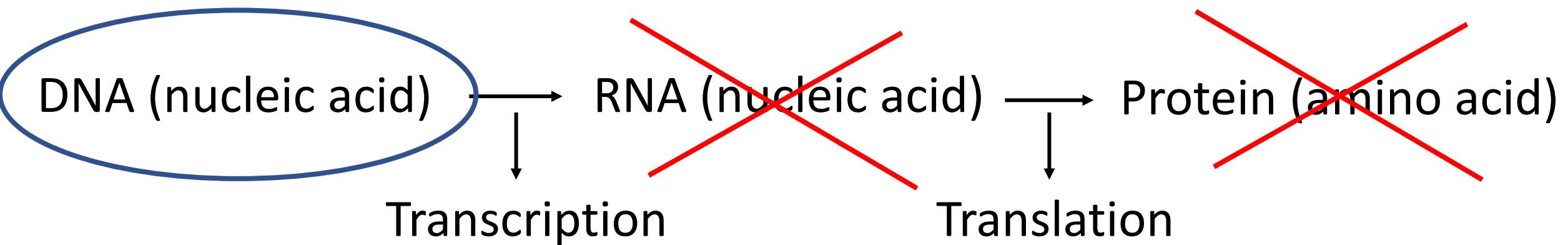
Does this mutation cause a change in the amino acid sequence of a protein?

How different are the amino acid changes?

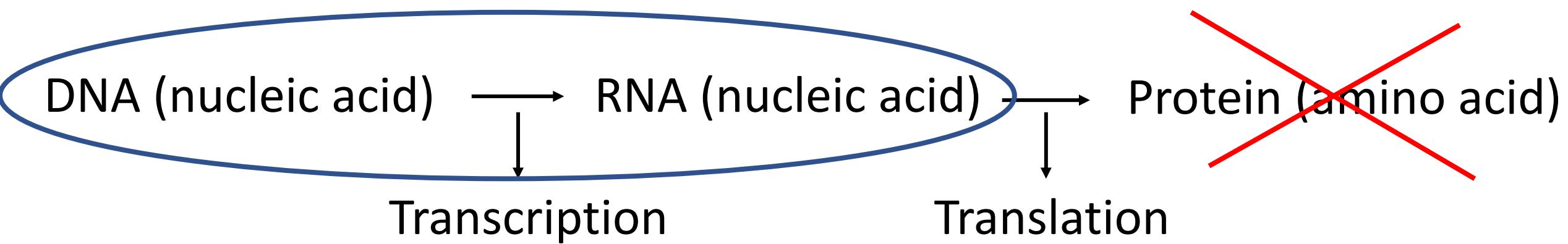
Do the amino acid changes impact the function of my protein?

Do the amino acid changes impact the structure of my protein?

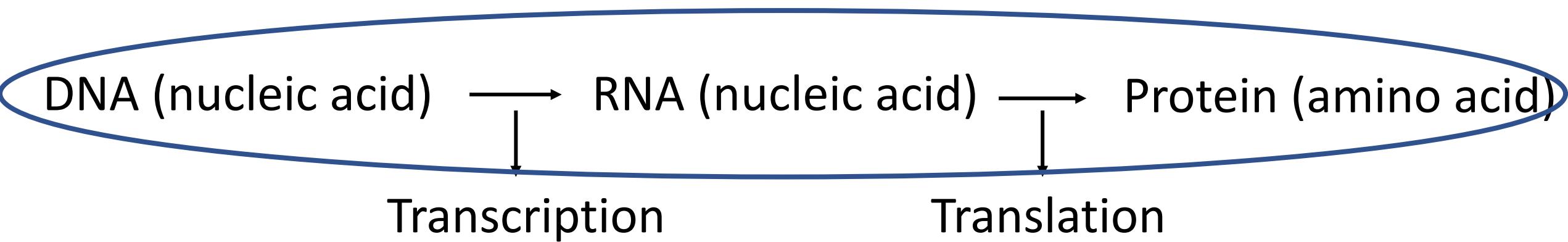
Synonymous or Silent mutations can occur in noncoding regions and have no impact on protein sequence



Synonymous changes can also occur within protein coding regions but do not result in an amino acid change

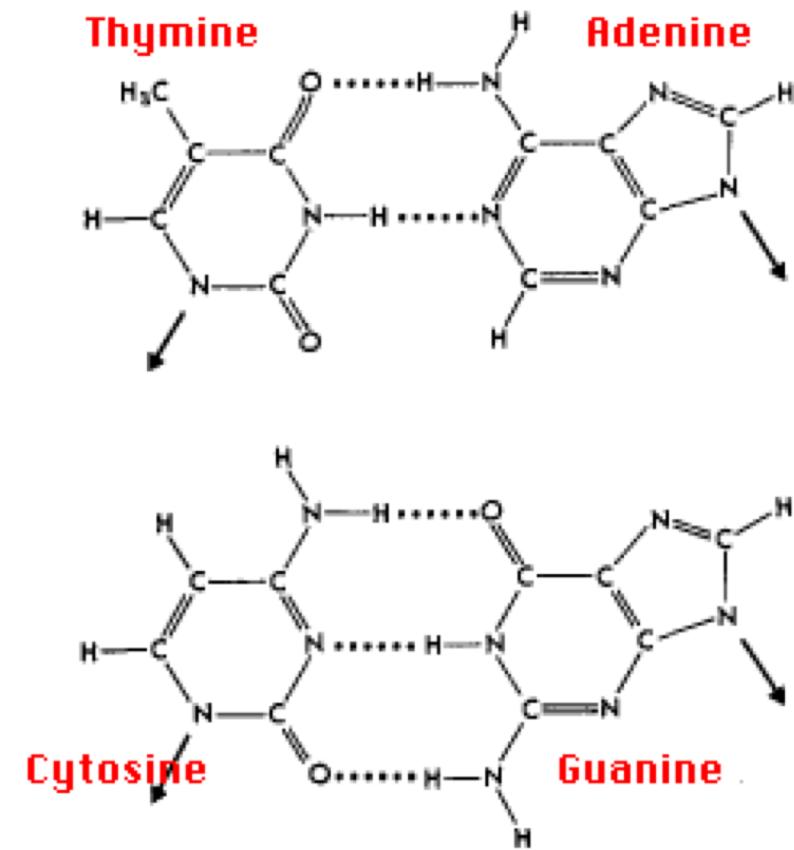
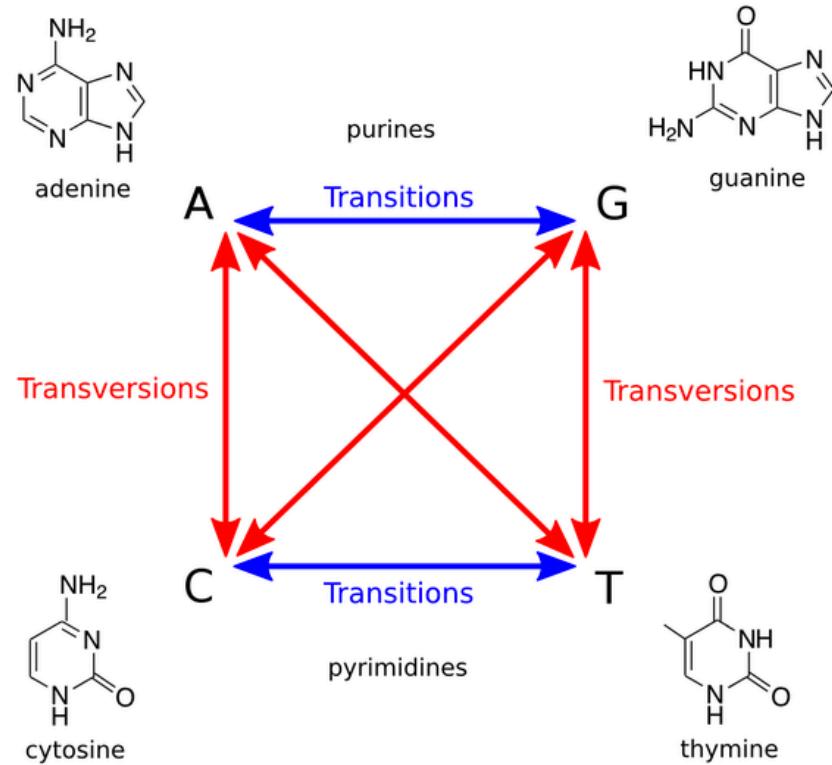


Nonsynonymous changes (replacement changes) persist through translation and a different amino acid is used



Take a quick step back to DNA

There are two classes of DNA mutations **Transitions** and **Transversions**



Transversions are more likely to impact amino acid sequences

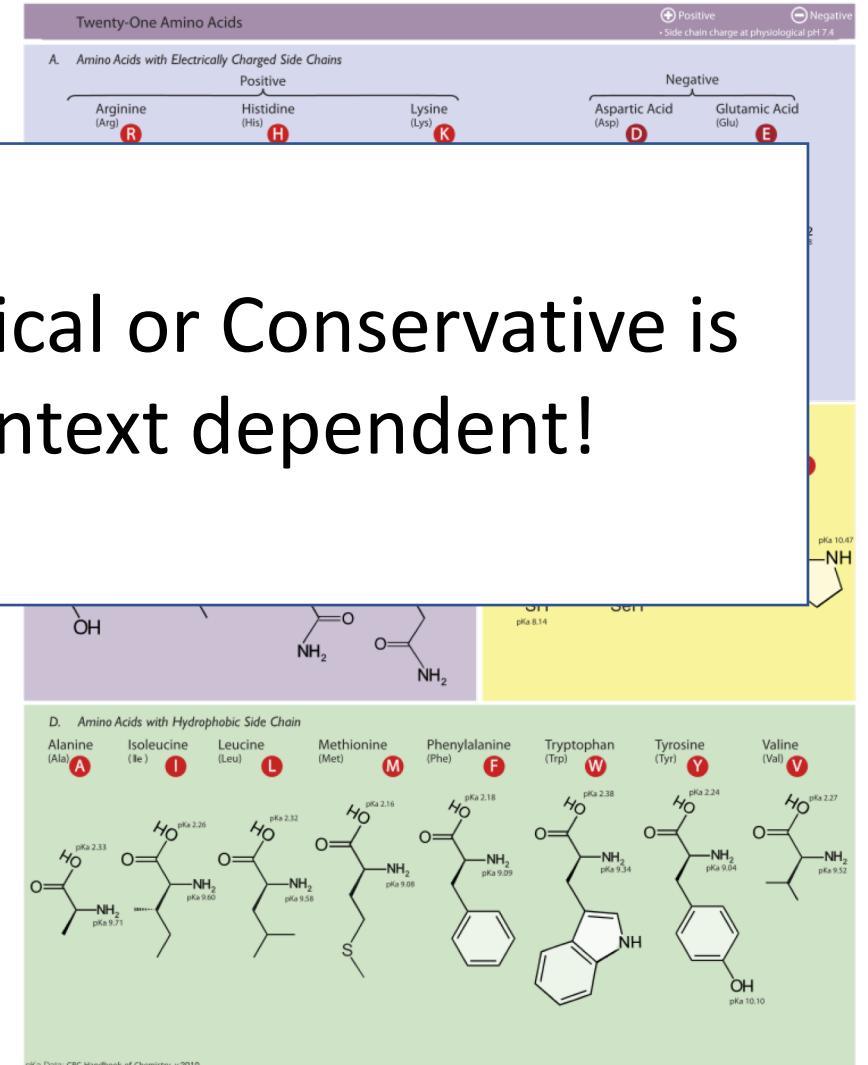
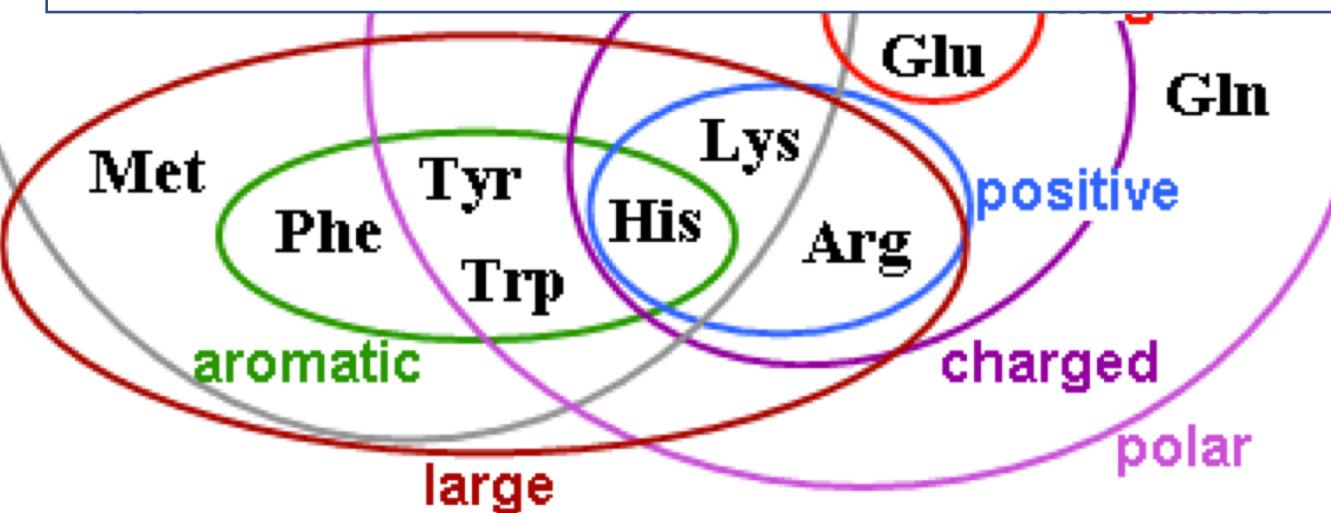
GENETIC CODE TABLE

		SECOND LETTER					
		U	C	A	G		
FIRST LETTER	U	UUU UUC UUA UUG } Phe	UCU UCC UCA UCG } Ser	UAU UAC } Tyr	UGU UGC } Cys	U C	A G
	C	CUU CUC CUA CUG } Leu	CCU CCC CCA CCG } Pro	CAU CAC CAA CAG } His	CGU CGC CGA CGG } Arg	U C	A G
	A	AUU AUC AUA AUG } Ile	ACU ACC ACA ACG } Thr	AAU AAC AAA AAG } Asn	AGU AGC AGA AGG } Ser	U C	A G
	G	GUU GUC GUA GUG } Val	GCU GCC GCA GCG } Ala	GAU GAC GAA GAG } Asp	GGU GGC GGA GGG } Gly	U C	A G
THIRD LETTER							

Within Nonsynonymous changes there are two groups Radical and Conservative changes.



Whether something is classified as Radical or Conservative is difficult to determine. Many are context dependent!

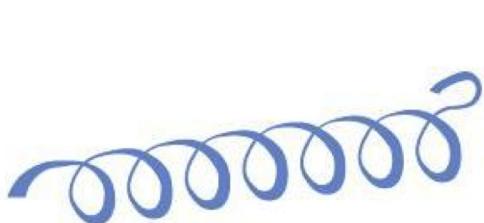


There are a lot of software packages to help you assess the impacts of nonsynonymous changes.

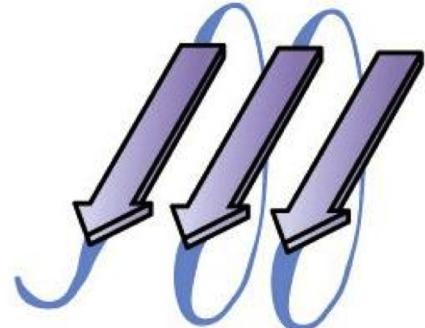
Primary structure: Changes in amino acid sequences

Secondary structure: Changes in amino acid sequences

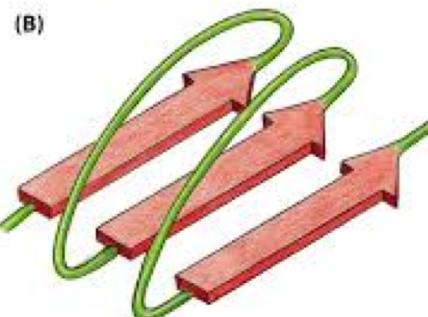
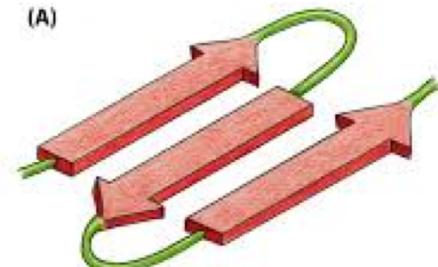
Secondary structure



α helix



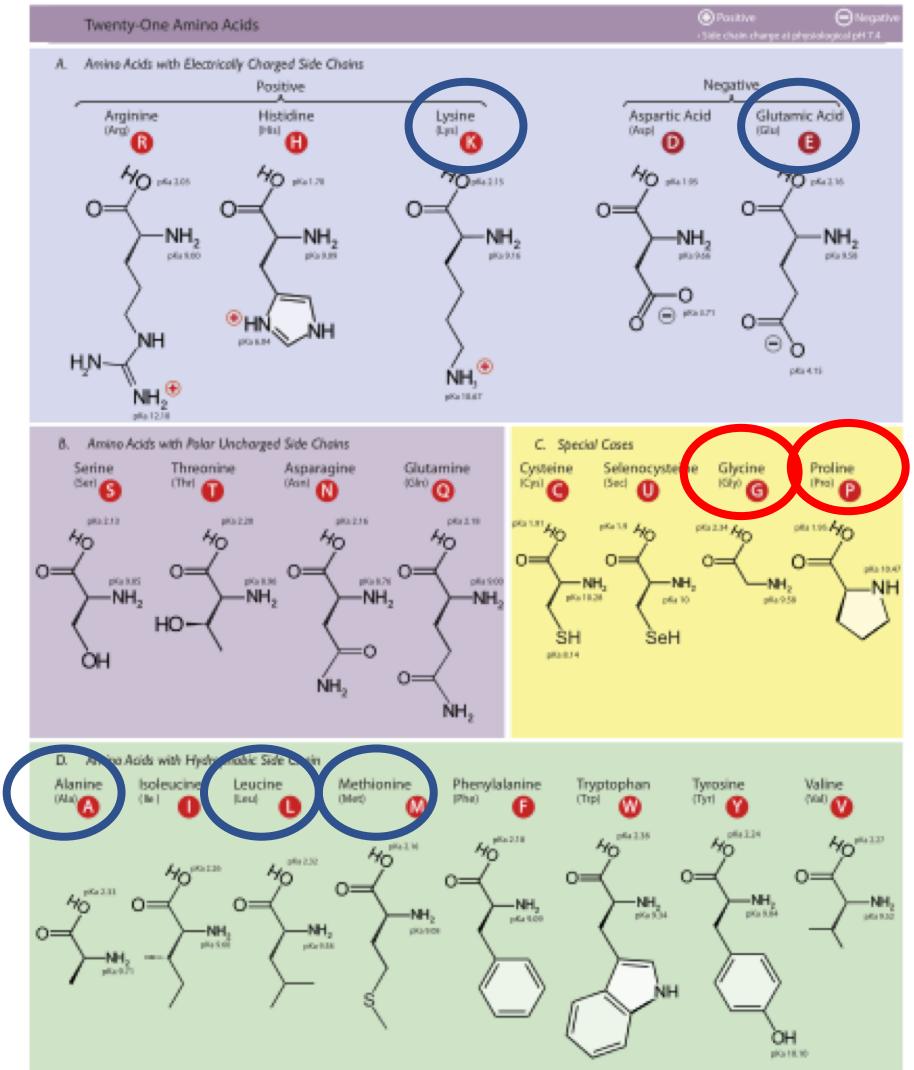
β sheet



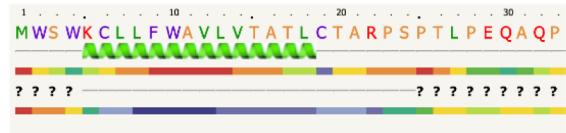
Some amino acids have properties that favor alpha helices or Beta pleated sheets

“MALEK” Amino acids readily form Alpha helices

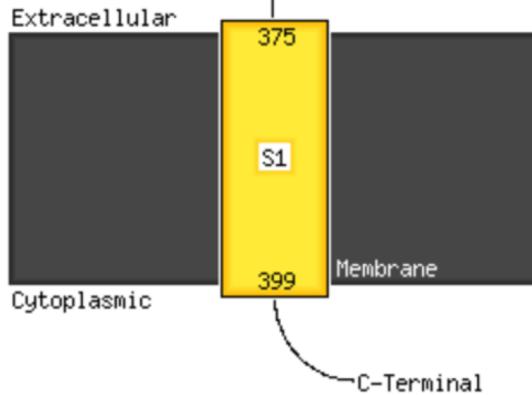
Glycine and Proline DO NOT work well in alpha helices



Using Secondary structure predictions you can also assess impacts on alpha helices, transmembrane domains, beta sheets etc.



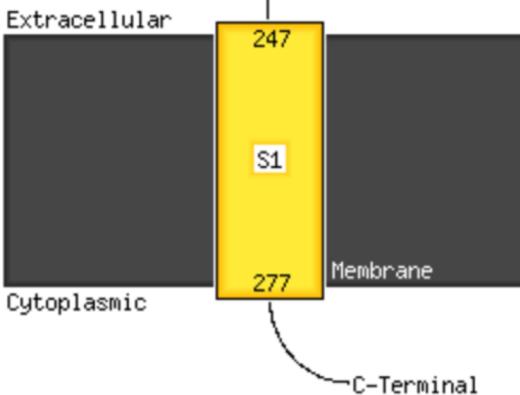
1-33 Signal peptide
N-Terminal



Humans



N-Terminal



Sea dragons

You can also compare sequences to assess impact of mutations.

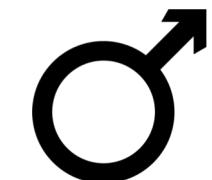
Example: A SNP that determines sex



HHTRANSMEMBRANEDOMAINSEQUENCERR

HHTRANSMEMBRANEDOMAINSEQUENCERR

HHTRANSMEMBRANEDOMAINSEQUENCERR



YHTRANSMEMBRANEDOMAINSEQUENCELR

YHTRANSMEMBRANEDOMAINSEQUENCELR

YHTRANSMEMBRANEDOMAINSEQUENCELR

Two Nonsynonymous changes fixed between the sexes flanking a transmembrane domain

-In Both cases you have repeated amino acids HH & RR

-in males there is a change: HH -> YH and RR -> LR

Both are changes from a positively charged aa to a hydrophobic aa

→ PROVEAN Tools

PROVEAN Protein

PROVEAN Protein Batch

Human

Mouse

PROVEAN Genome Variants

Human

Mouse

→ About

→ FAQ

→ News

→ Download

→ Help

→ Contact Us

→ Related Links

PROVEAN (**P**rotein **V**ariation **E**ffect **A**nalyzer) is a software tool which predicts whether an amino acid substitution or indel has an impact on the biological function of a protein.

PROVEAN is useful for filtering sequence variants to identify nonsynonymous or indel variants that are predicted to be functionally important.

The performance of PROVEAN is comparable to popular tools such as SIFT or PolyPhen-2 [1]. [Read more](#).

A fast computation approach to obtain pairwise sequence alignment scores enabled the generation of precomputed PROVEAN predictions for 20 single AA substitutions and a single AA deletion at every amino acid position of all protein sequences in human and mouse [2].

This work is funded by the National Institutes of Health [grant number 5R01HG004701-04].

References:

1. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE* 7(10): e46688.
2. Choi Y (2012) A Fast Computation of Pairwise Sequence Alignment Scores Between a Protein and a Set of Single-Locus Variants of Another Protein. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine (BCB '12)*. ACM, New York, NY, USA, 414-417.
(* This is the author's version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution. The definitive version was published in ACM BCB '12. <http://doi.acm.org/10.1145/2382936.2382989>)
3. Choi Y, Chan AP (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 31(16): 2745-2747.

There are also online tools specifically designed to assess impacts on TM domains

DeepTMHMM

A Deep Learning Model for Transmembrane Topology Prediction and Classification

Protein structure prediction using deep learning methods have seen several advancements within the last years. In this project, we investigate deep learning for prediction of the membrane topology of transmembrane proteins. Transmembrane proteins are relevant for drug development since they make up more than 50% of all human drug targets .

DeepTMHMM is currently the most complete and best-performing method for the prediction of the topology of both alpha-helical and beta-barrel transmembrane proteins. The model encodes the primary amino acid sequence by a pre-trained language model and decodes the topology by a state space model to produce topology and type predictions at unprecedented accuracy. DeepTMHMM makes it possible to scan full proteomes in order to detect both classes of transmembrane proteins, and we anticipate our method to be very valuable for the research community.

Note: If you want to run DeepTMHMM on more than 10000 sequences per run, then you can run it locally by following the instructions below.

Running DeepTHMHH From the Terminal

You can run DeepTMHMM directly from the terminal through the `pybiolib` package:

To perform tasks comparing between sequences
you need a high quality multiple alignment

Up until now when we have talked about “alignments” we have been referring to reads aligned to a reference

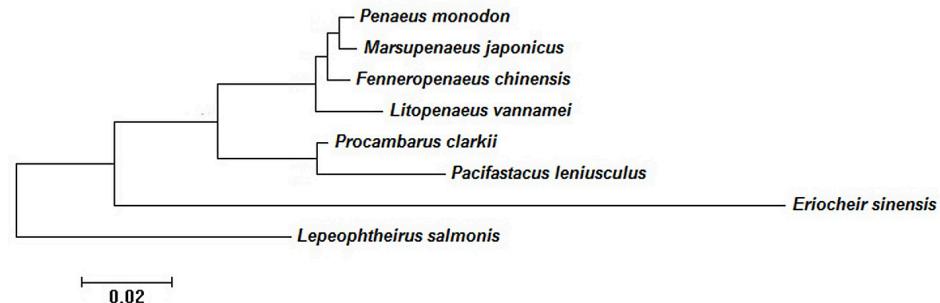
A

Penaeus monodon
Marsupenaeus japonicus
Fenneropenaeus chinensis
Litopenaeus vannamei
Procambarus clarkii
Pacifastacus leniusculus
Eriocheir sinensis
Lepeophtheirus salmonis

Penaeus monodon
Marsupenaeus japonicus
Fenneropenaeus chinensis
Litopenaeus vannamei
Procambarus clarkii
Pacifastacus leniusculus
Eriocheir sinensis
Leneophterheirus salmonis

Penaeus monodon
Marsupenaeus japonicus
Fenneropenaeus chinensis
Litopenaeus vannamei
Procambarus clarkii
Pacifastacus leniusculus
Eriocheir sinensis
Lernanthrus sinicus
Salmo salar

B



We are now focusing on
multiple alignments from
Fasta formats which can be
used for lots of biological
inferences

Multiple alignments can be in nucleotide or amino acid formats.

Bioedit

Geneious

I use Geneious for most things, and even find the free version (with most features locked) helpful for visualization

- visualize bam files
- call consensus
- annotate sequences
- design primers
- translate/reverse/complement/align/edit

Multiple alignments can also be done on the command line for example with MAFFT

MAFFT

Section: Mafft Manual (1)
Updated: 2007-06-09
[Index](#) [Return to Main Contents](#)

NAME

mafft - Multiple alignment program for amino acid or nucleotide sequences

SYNOPSIS

```
mafft [options] input [> output]
linsi input [> output]
ginsi input [> output]
einsi input [> output]
fftensi input [> output]
fftnsi input [> output]
nwns input [> output]
nwnsi input [> output]
mafft-profile group1 group2 [> output]

input, group1 and group2 must be in FASTA format.
```

DESCRIPTION

MAFFT is a multiple sequence alignment program for unix-like operating systems. It offers a range of multiple alignment methods.

Accuracy-oriented methods:

```
*L-INS-i (probably most accurate; recommended for <200 sequences; iterative refinement method incorporating local pairwise alignment information):
mafft --localpair --maxiterate 1000 input [> output]
linsi input [> output]

*G-INS-i (suitable for sequences of similar lengths; recommended for <200 sequences; iterative refinement method incorporating global pairwise alignment information):
mafft --globalpair --maxiterate 1000 input [> output]
ginsi input [> output]

*E-INS-i (suitable for sequences containing large unalignable regions; recommended for <200 sequences):
mafft --ep 0 --genafpair --maxiterate 1000 input [> output]
einsi input [> output]

For E-INS-i, the --ep 0 option is recommended to allow large gaps.
```

The protein folding problem:

We have come a long way in predicting impacts on secondary structures but how changes impact folding and tertiary structure is still a massive problem

“Maybe more than any other area of biology, Machine Learning is needed to solve the protein folding problem”

Enter AlphaFold2

Colab Notebook: AlphaFold2.ipynb

File Edit View Insert Runtime Tools Help

Share Sign In

+ Code + Text Copy to Drive

Connect Editing

ColabFold: AlphaFold2 using MMseqs2

{x}

Easy to use protein structure and complex prediction using [AlphaFold2](#) and [AlphaFold2-multimer](#). Sequence alignments/templates are generated through [MMseqs2](#) and [HHsearch](#). For more details, see [bottom](#) of the notebook, checkout the [ColabFold GitHub](#) and read our manuscript. Old versions: [v1.0](#), [v1.1](#), [v1.2](#), [v1.3](#)

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: Making protein folding accessible to all. *Nature Methods*, 2022.



Input protein sequence(s), then hit Runtime -> Run all

query_sequence: PIAQIHLGRSDEQKETLIREVSEAIRSLDALTSVRVIITEMAKGHFGIGGELASK

- Use : to specify inter-protein chainbreaks for **modeling complexes** (supports homo- and hetero-oligomers). For example PI...SK:PI...SK for a homodimer

jobname: test

use_amber:

template_mode: none

- "none" = no template information is used, "pdb70" = detect templates in pdb70, "custom" - upload and search own templates (PDB or mmCIF format, see [notes below](#))

Show code

MSA options (custom MSA upload, single sequence, pairing mode)

msa_mode: MMseqs2 (UniRef+Environmental)

pair_mode: unpaired+paired

- "unpaired+paired" = pair sequences from same species + unpaired MSA, "unpaired" = separate MSA for each chain, "paired" - only use paired sequences.

Show code

Advanced settings

model_type: auto

Keep in mind that proteins are dynamic and changes in the environment can impact structure

This adds a layer of complexity to the protein folding problem

After our break we will work through using prediction software for assessing impacts of nonsynonymous changes on protein structure and function

- (1) Phyre2
- (2) PROVEAN
- (3) AlphaFold2