

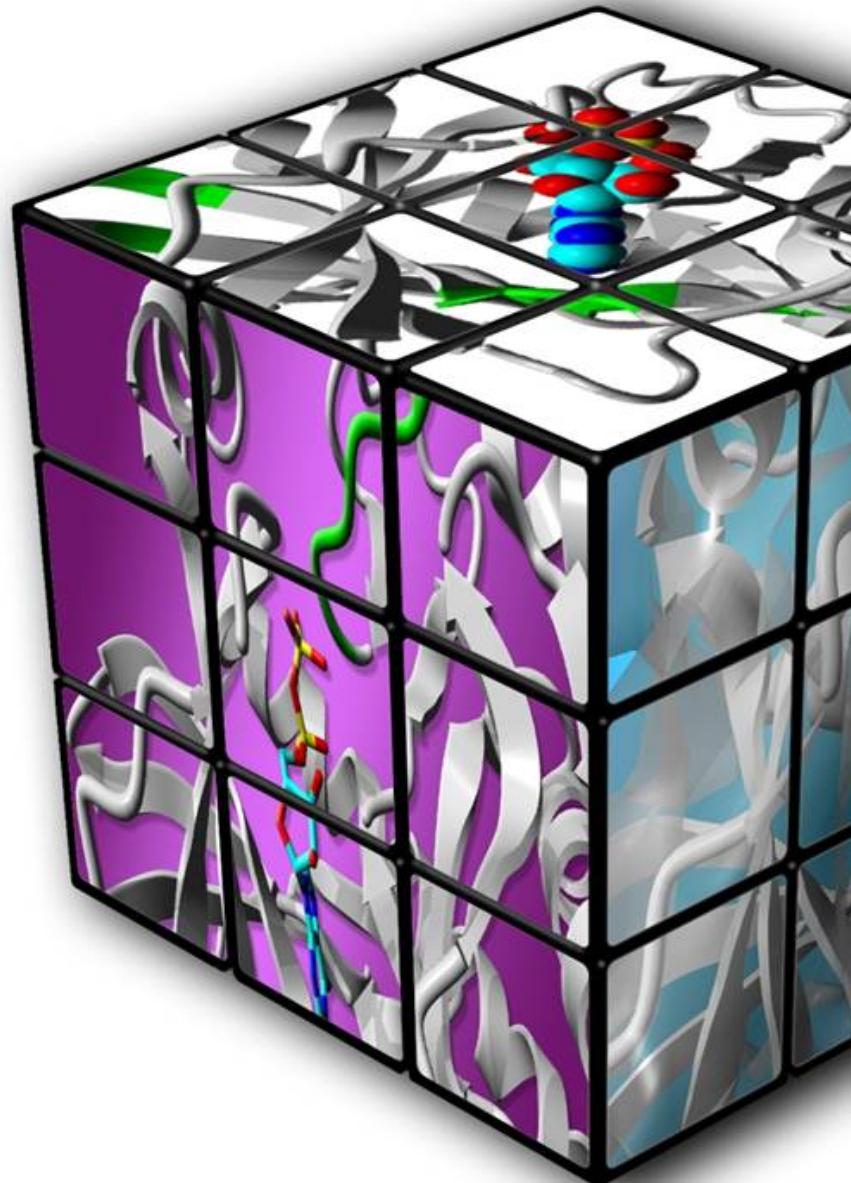
Variant Effect Prediction

course 2019

Mutational effects on protein level

Hanka Venselaar
CMBI - Radboudumc

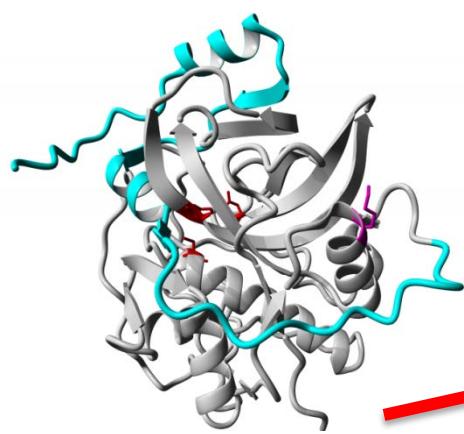
Hanka.Venselaar@radboudumc.nl



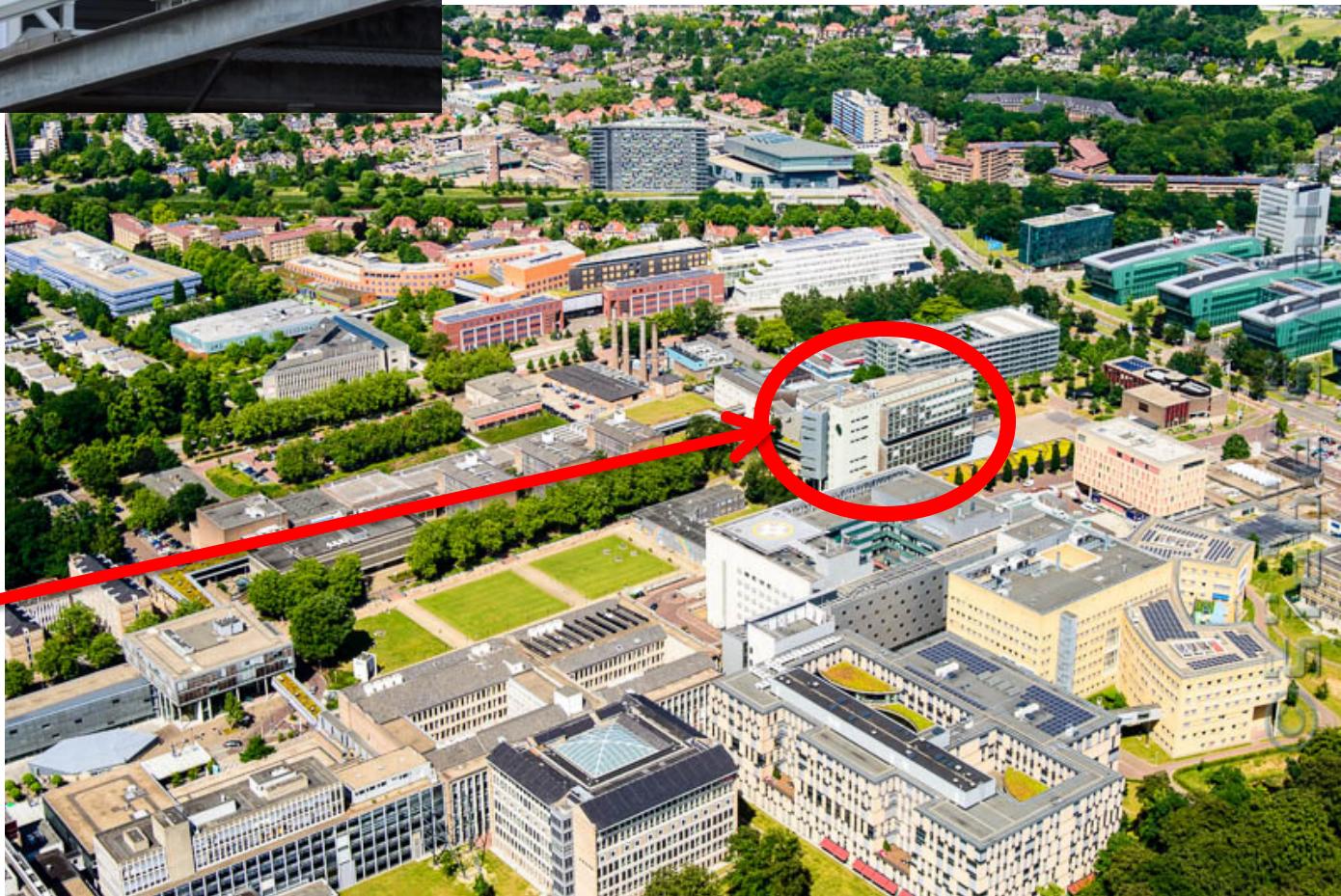


Radboud Institute for Molecular Life Sciences

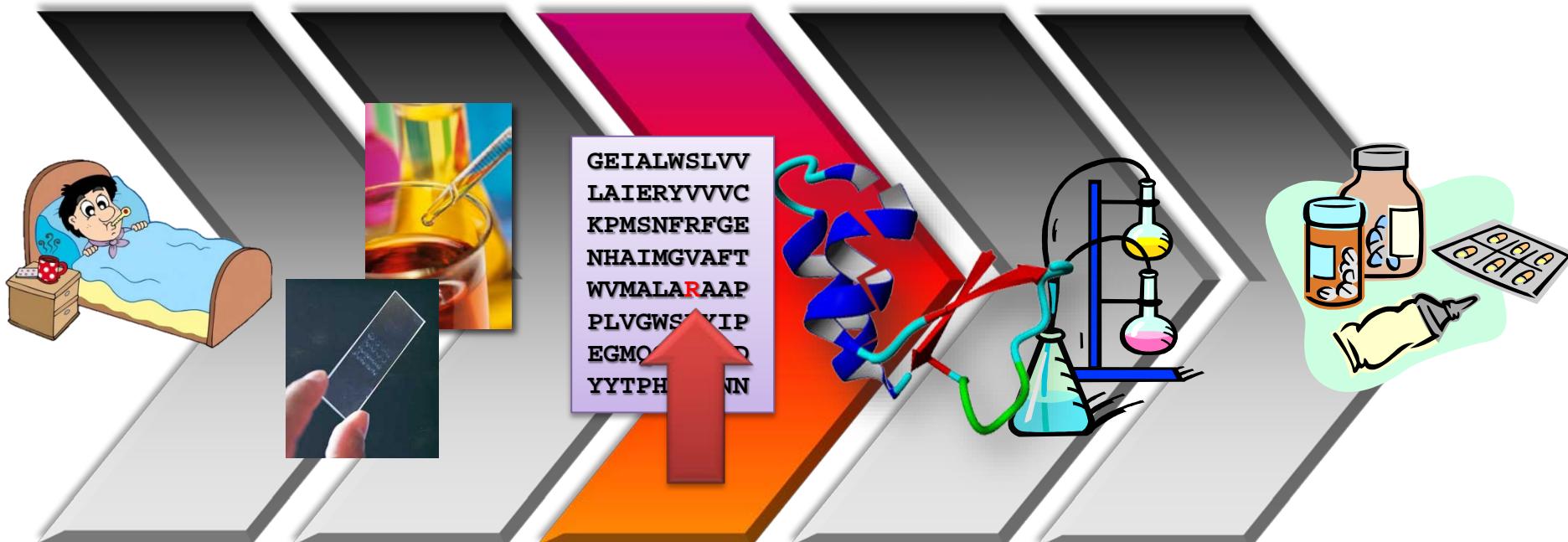
“Today’s Molecules for Tomorrow’s Medicine”



Centre for
Molecular and
Biomolecular
Informatics (CMBI)



Understanding (and improving) life...



Patient in hospital

Experiments

Bioinformatics

New experiments

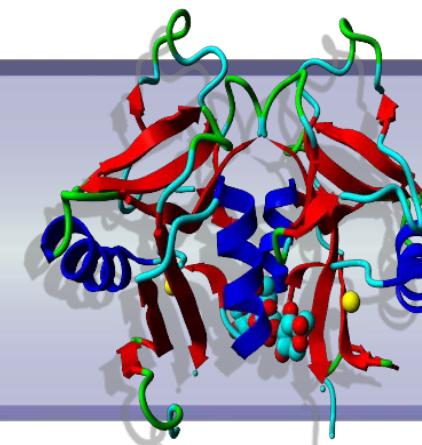
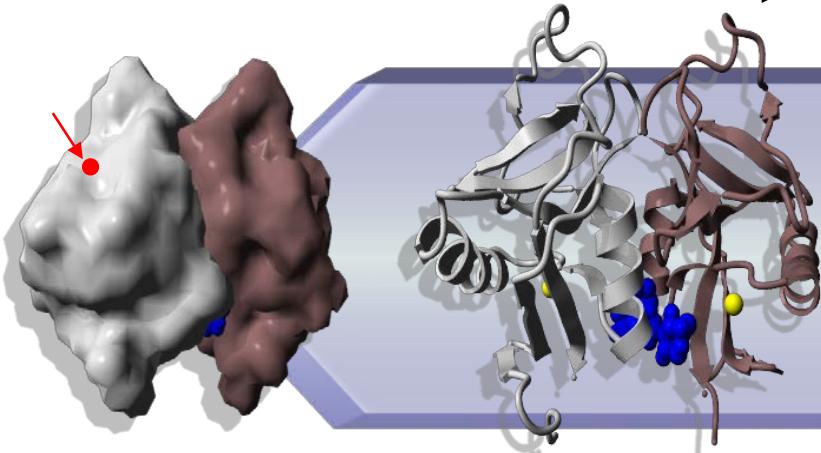
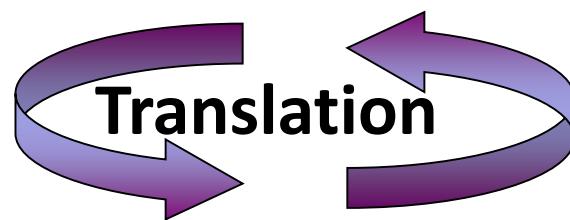
Drug development

Structural bioinformatics plays a central role...

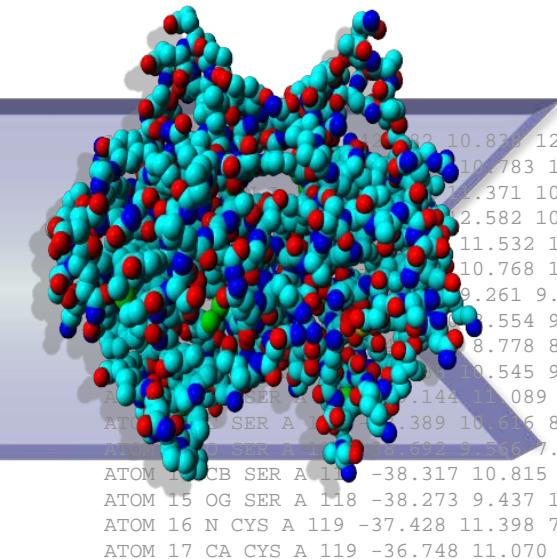
Understanding life: from Lab to PC to lab



Lab



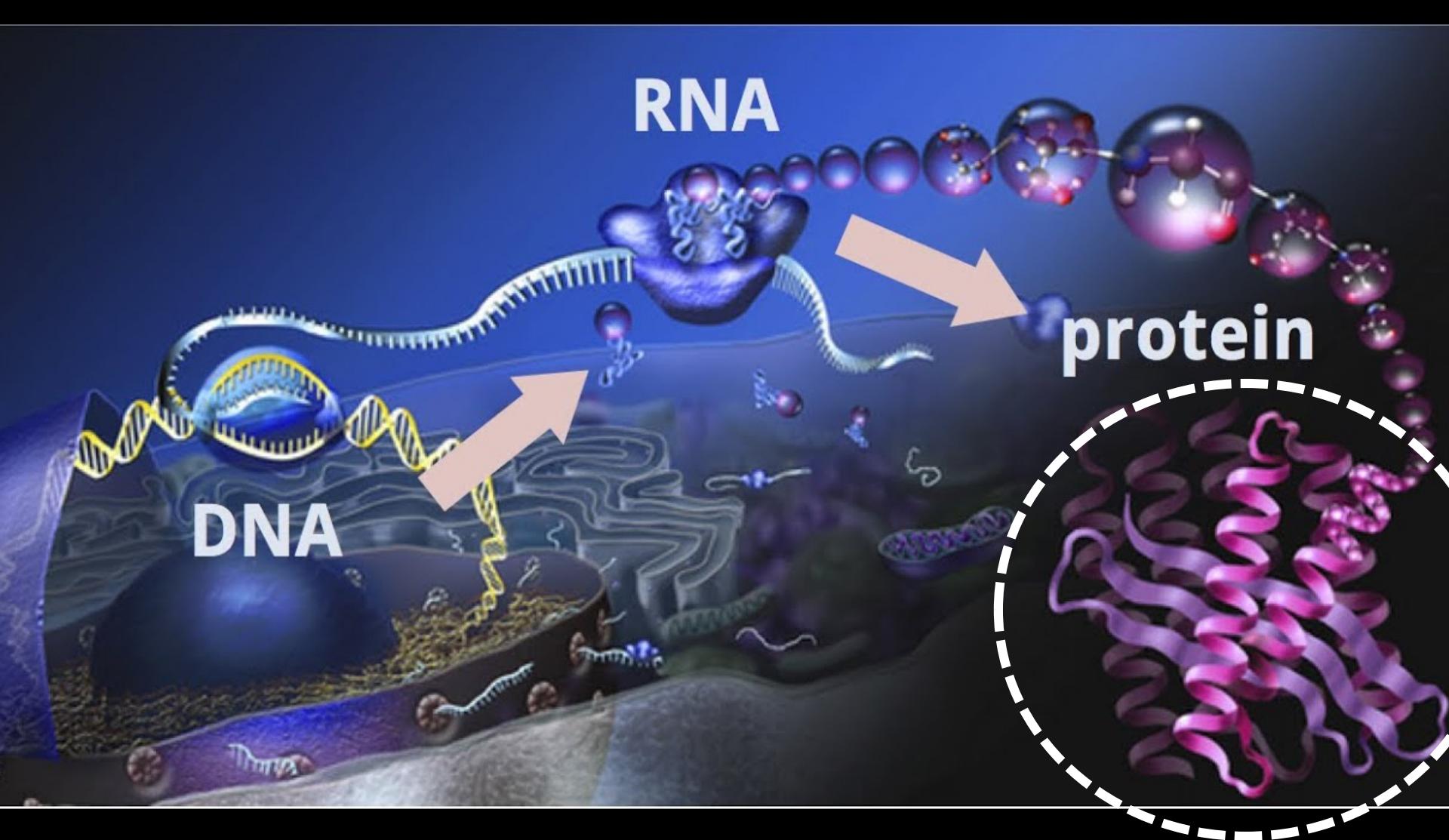
Bioinformatics



KKIALSDARSMKHALREIKIIRRL
DHDNIVKVYEVLGPKGTDLQGELF
KFSVAYIVQEYMETDLARLLEQGT
LAEEHAKLFMYQLLRGLKYIHSAN
VLHRDLPANIFISTEDLVL**K**IGDF
GLARIVDQHYSHKGYLSEGLVTKW
YRSPRLLLSPNNYTKAIDMWAAAGC
ILAEMLTGRMLFAGAHELEQMQLL
ETIPVIREEDKDELLRVMPSFVSS

?

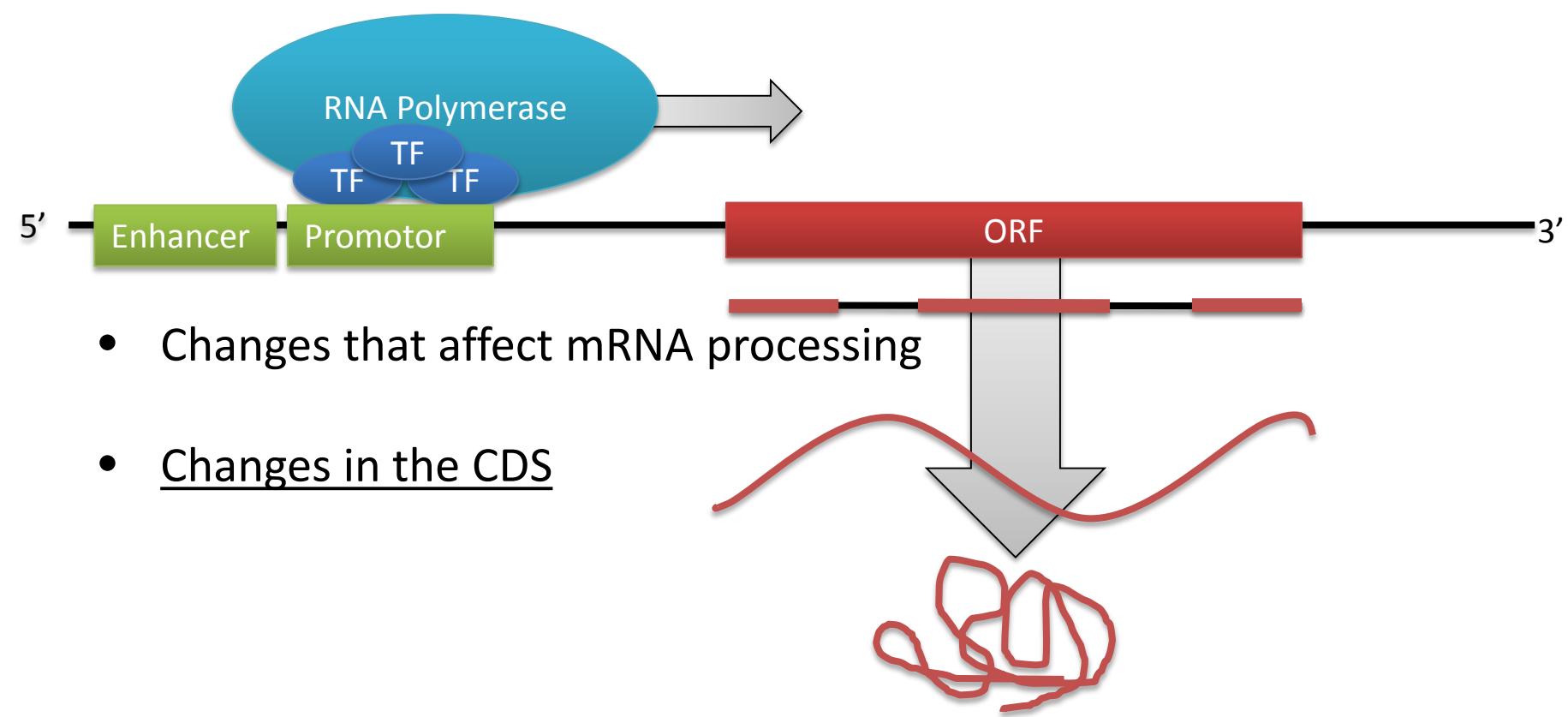
So much can go wrong...



...but let's focus on the level of proteins

DNA variations that affect a protein

- Changes in regulation (f.e. enhancer/promotor sites)
- Changes in splice sites
- Changes in the flanking regions
- Changes that affect translation speed (could be in ORF)



Mutations in coding sequence affect the protein' structure

- Silent, same amino acid → no effect on structure
- Nonsense, stop codons → Usually clear effect on structure, truncation
- Insertions and deletions → Effect depends on location
- Frameshifts → Usually clear effect on structure
- **Missense → Effects vary, easiest for automatic analysis**

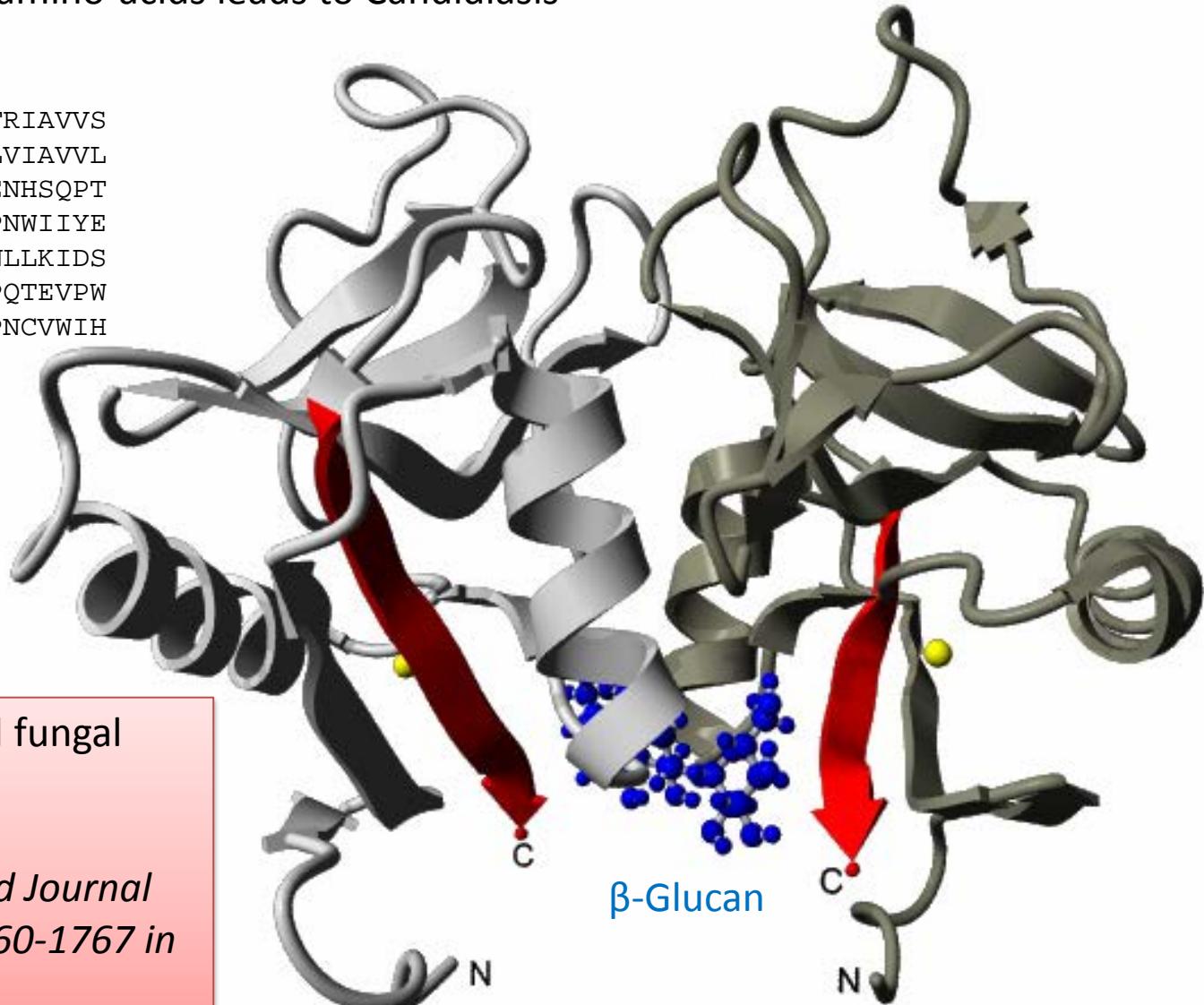
Example: Truncated protein

Normal function of Dectin is to recognize parts of the fungal cell-wall

Mutation: Deletion of 10 amino-acids leads to Candidiasis

>Dectin_1_Isoform_a

```
MEYHPDLENLDEDGYTQLHFDSQSNTRIAVVS  
EKGSCAASPPWRLIAVILGILCLVILVIAVVL  
GTMAIWRSNSGSNTLENGYFLSRNKENHSQPT  
QSSLEDSVTPTKAVKTTGVLLSSPCPPNIIYE  
KSCYLFMSMSLNSWDGSKRQCWQLGSNLLKIDS  
SNELGFIVKQVSSQPDNSFWIGLSRPQTEVPW  
LWEDGSTFSSNLFQIRTTATQENPSPNCVWIH  
VSVIYDQLCSVPSYSICEKKFSM
```



Candidiasis = uninhibited fungal infections (C. Albicans)

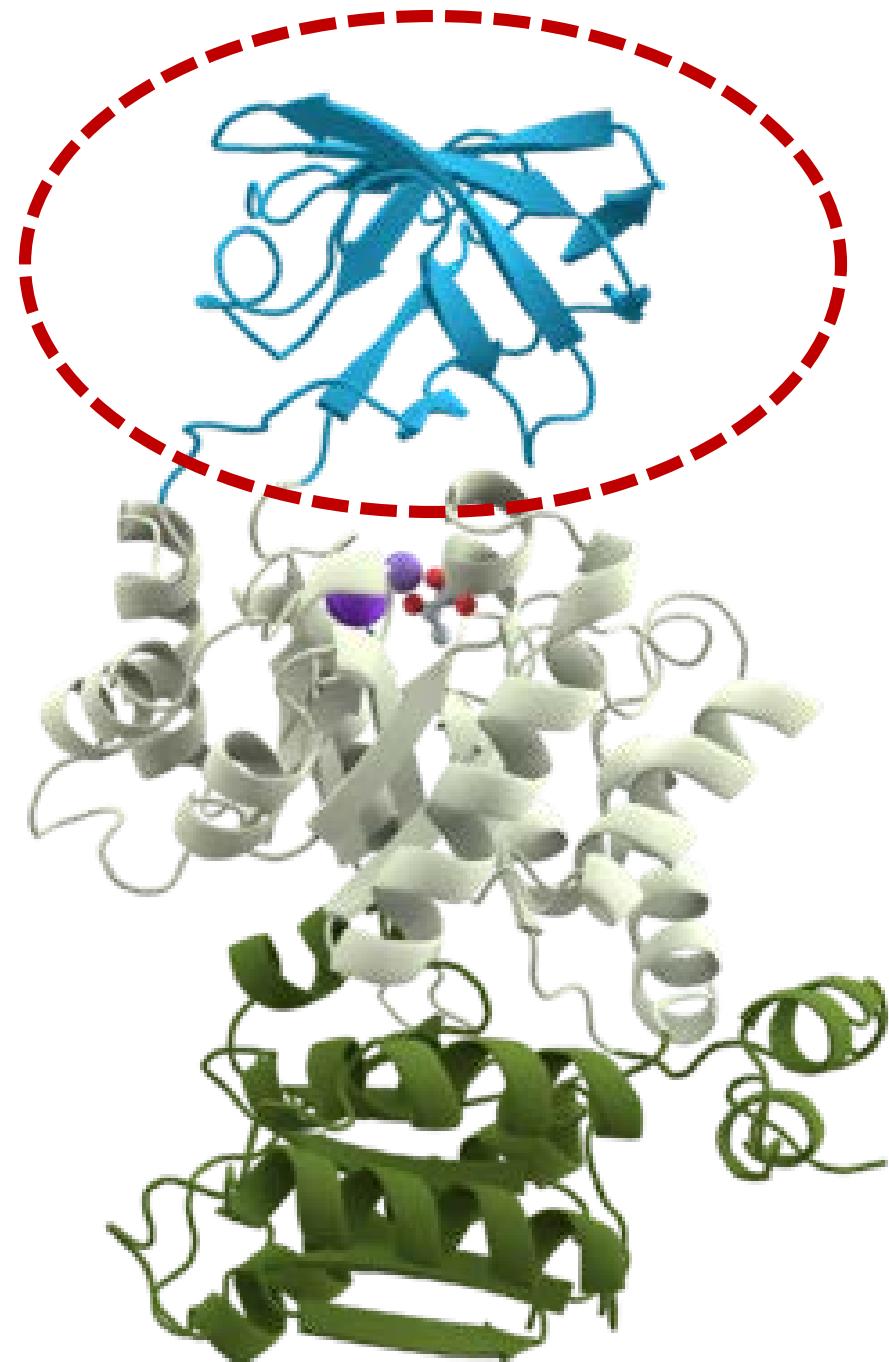
Published in New England Journal of Medicine 361 (18), 1760-1767 in 2009

Side note:

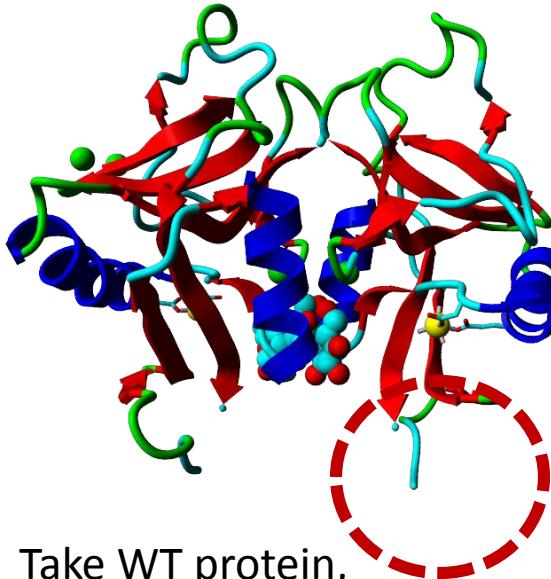
Protein domains are often stable structural units with a particular function that can also exist on their own.

Deletion of a domain might result in a stable truncated protein that disturbs the function of the protein that is still produced by the healthy allele.

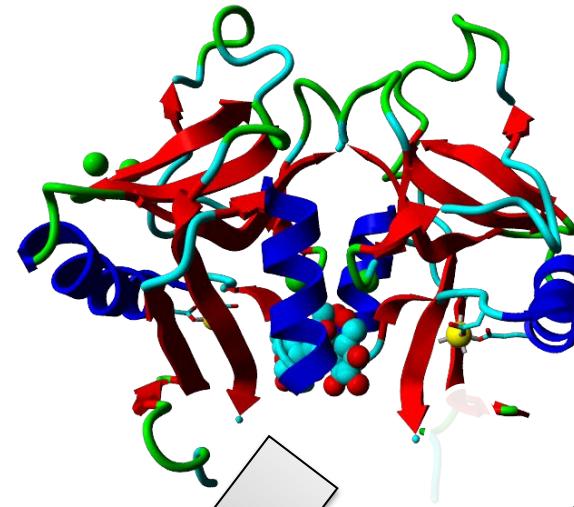
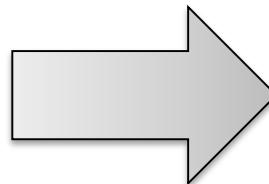
→ Dominant negative effect



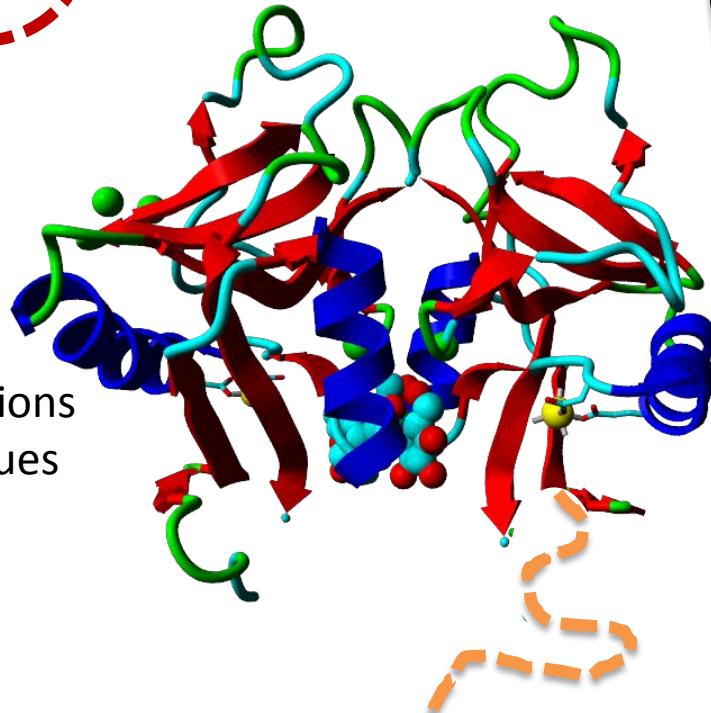
Insertions/Deletions/Frame shifts are difficult to predict



Take WT protein,
Identify residues
to change



Make this
change, in this
case a deletion



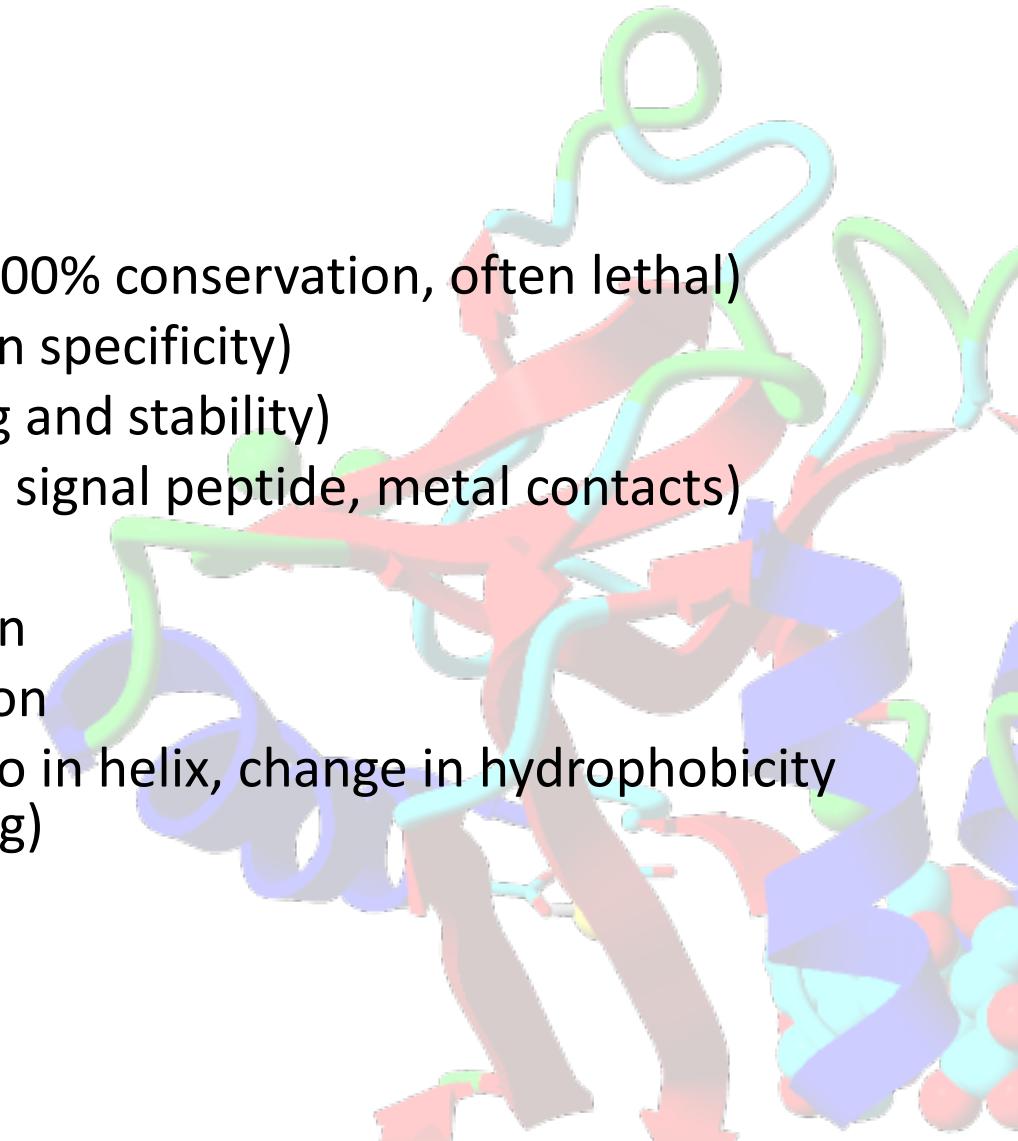
Predict positions
of new residues
from scratch

To do this you need a
protein structure and a
template for the new
residues
Smaller changes are
easier!

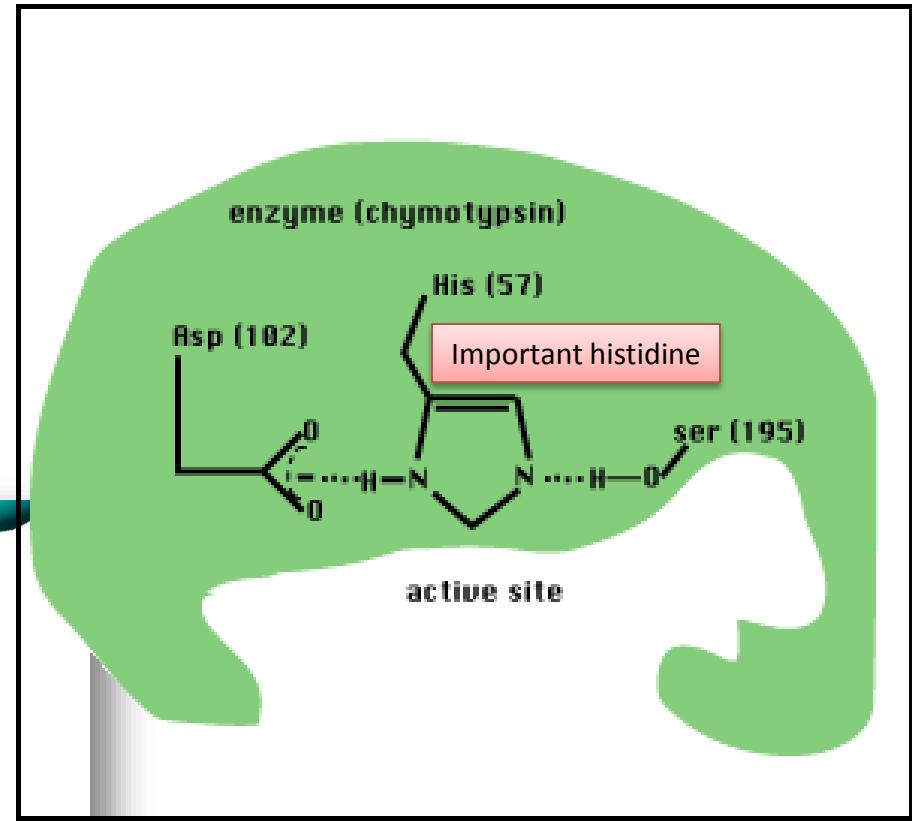
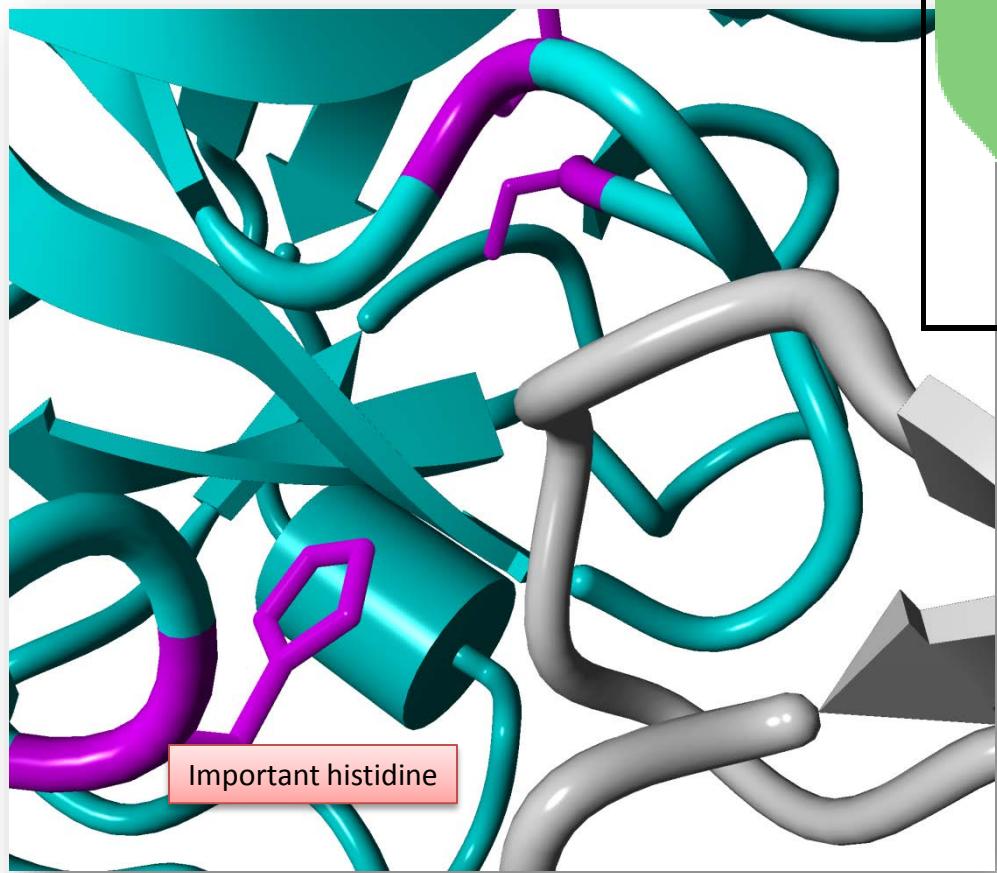
Examples of missense mutations on protein level

Mutations can occur in/on the:

- Active site (often in Uniprot, 100% conservation, often lethal)
- Binding site/pocket (changes in specificity)
- Core of protein (affects folding and stability)
- Functional site (like PTM sites, signal peptide, metal contacts)
- Surface in dimerization sites
- Surface of a membrane protein
- Surface without known function
- Very specific mutations (X->Pro in helix, change in hydrophobicity pattern in strand, helix-capping)



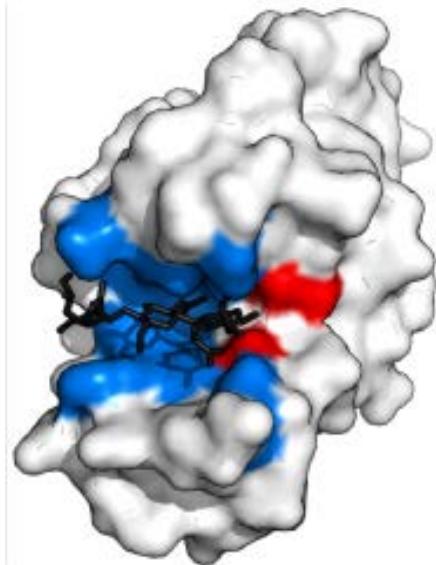
Active site mutations are Damaging



Schematic

Real life

Binding site mutations



ACTIVE SITE

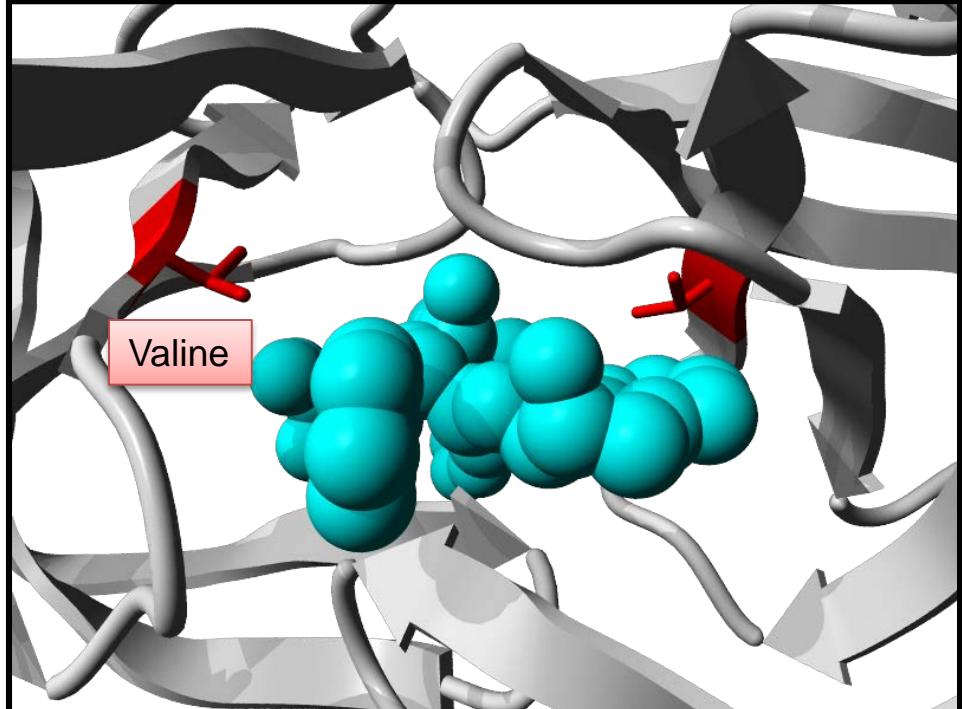
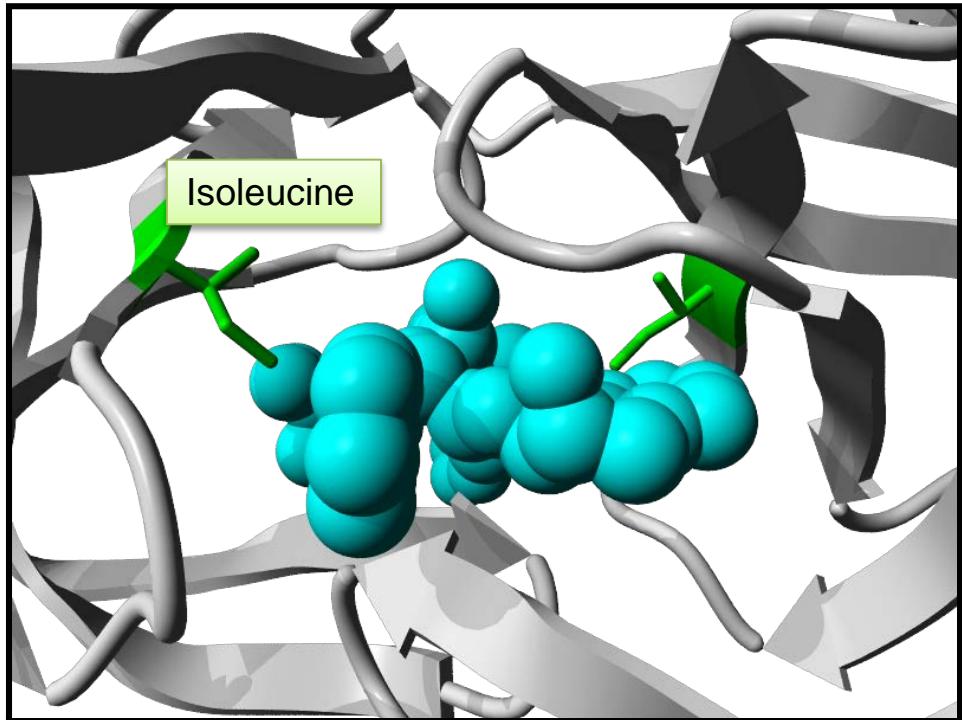
BINDING SITES

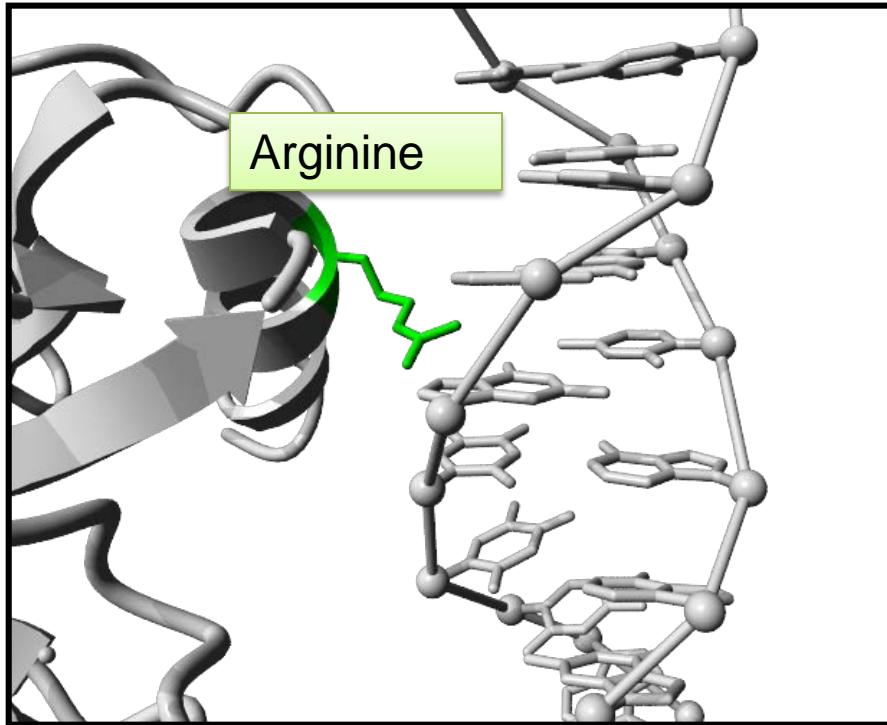
Bind and orient
substrate(s)

CATALYTIC SITE

Reduce chemical
activation energy

Different shaped binding
site will affect substrate
binding



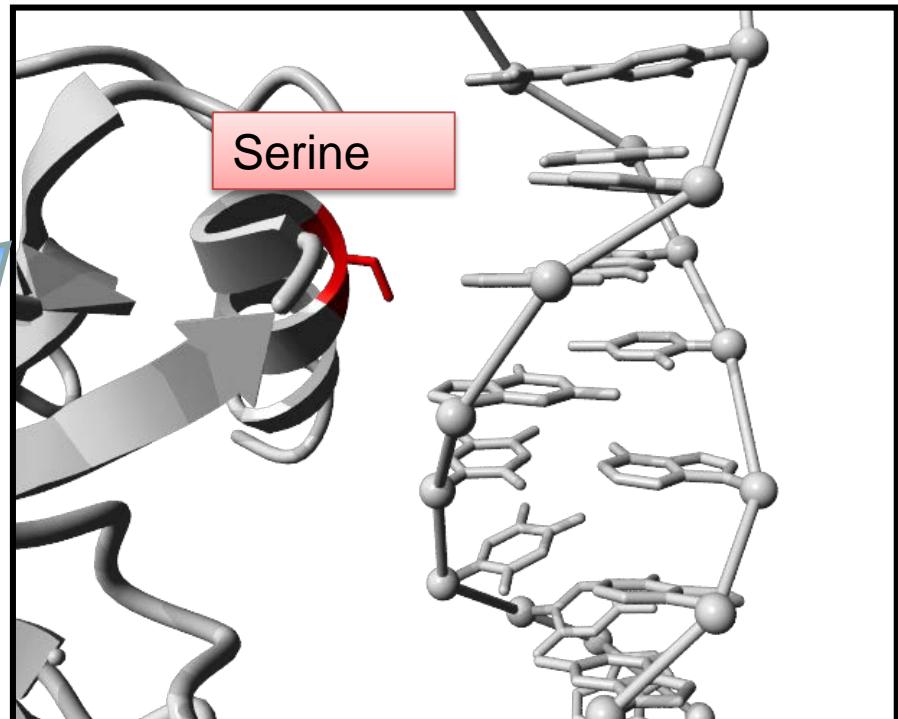
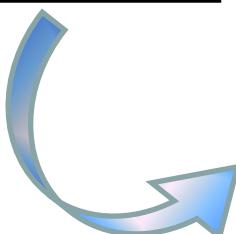


Mutations affecting any type of interaction

For example: interaction with DNA

Other interactions that can be disturbed:

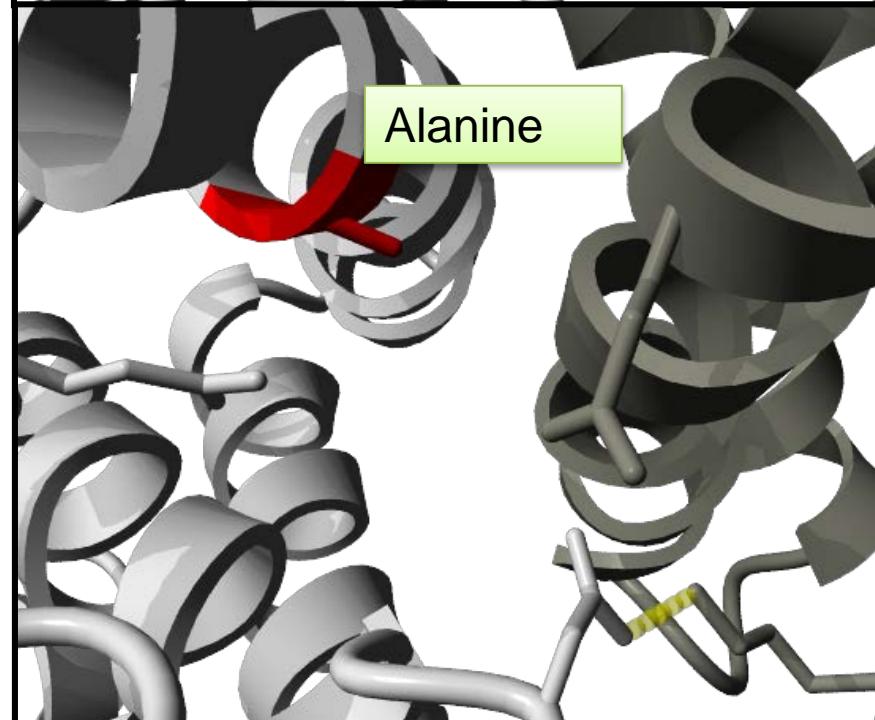
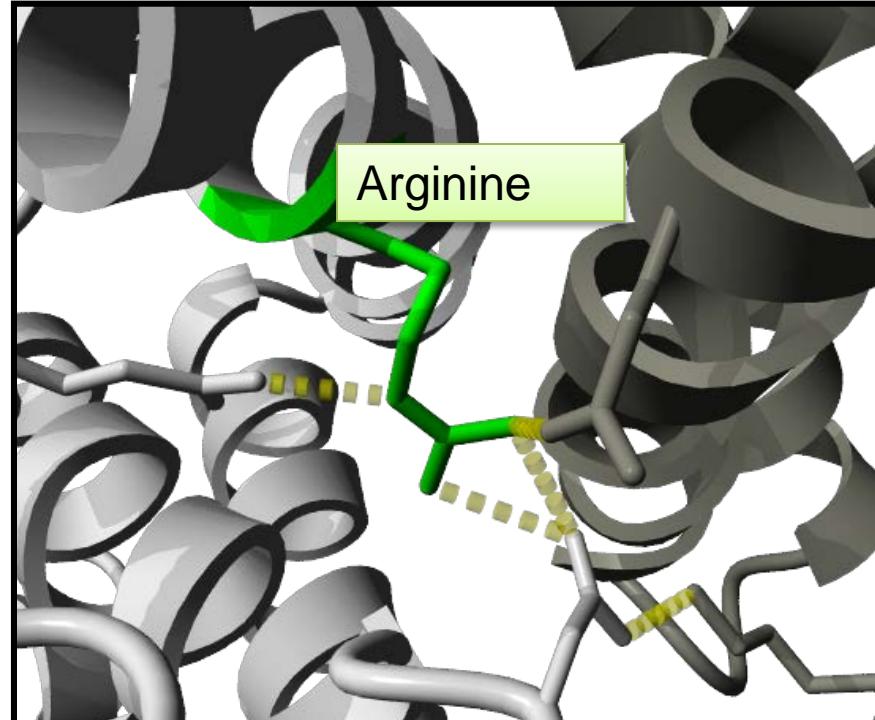
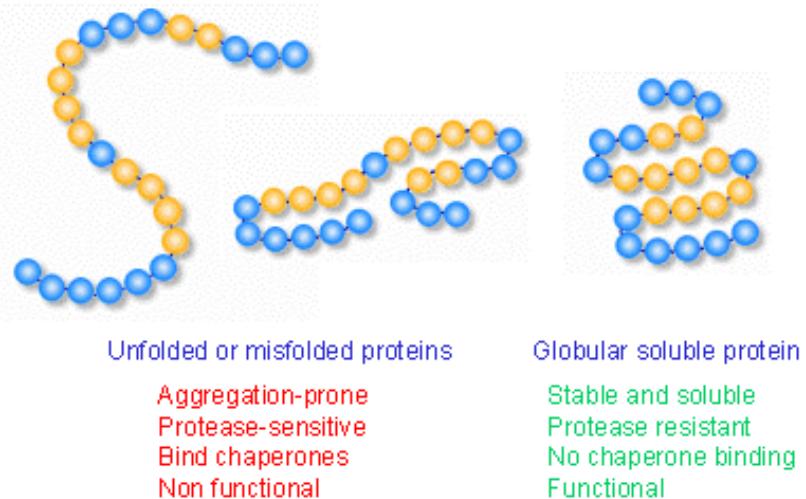
- metals
- co-factors
- lipids
- other proteins
(stable/transient)
- etc



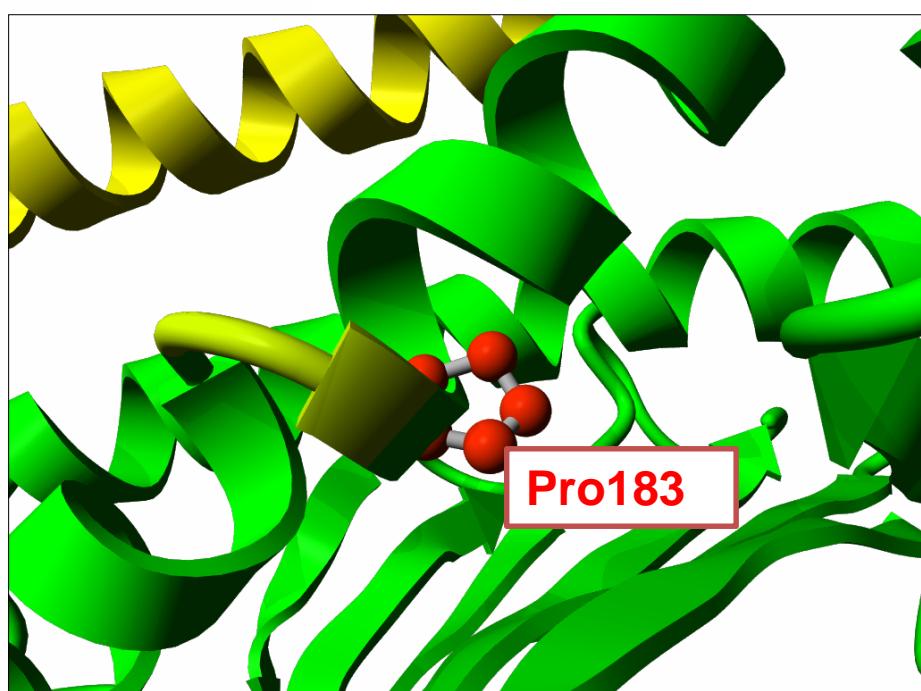
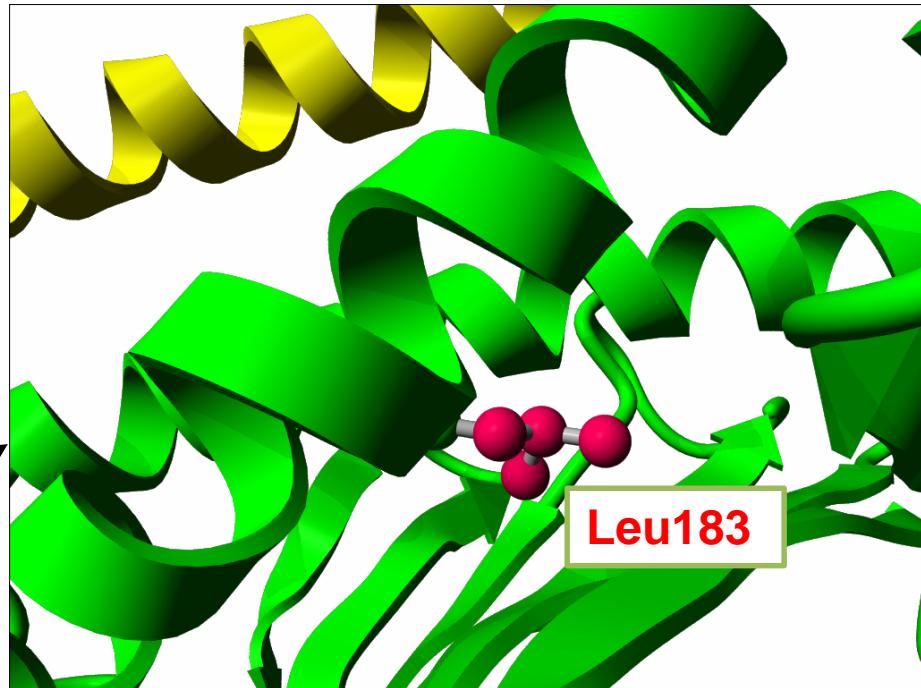
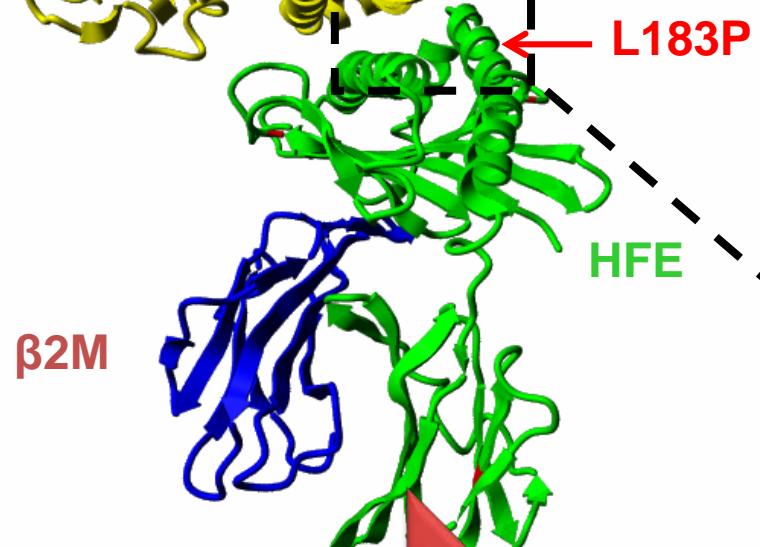
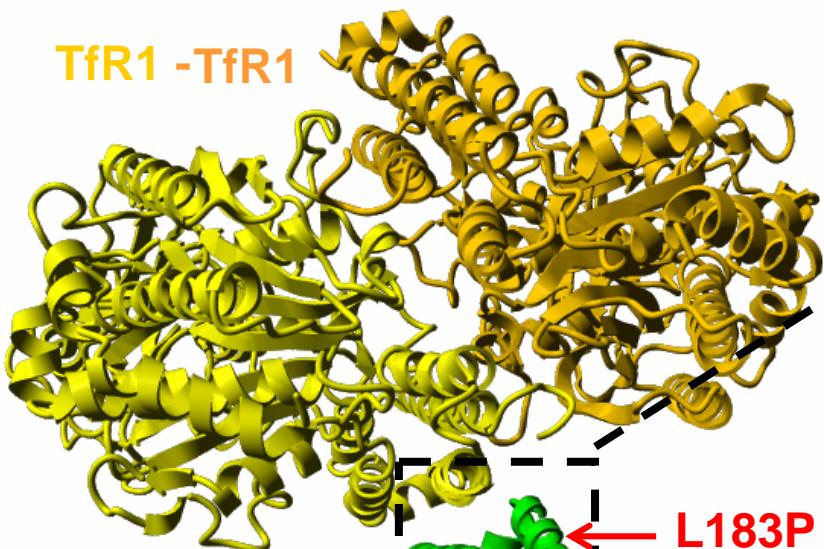
Mutations in the core of the protein

The core should be correctly folded and stable, otherwise the protein will unfold.

Delicate balance between folded and unfolded



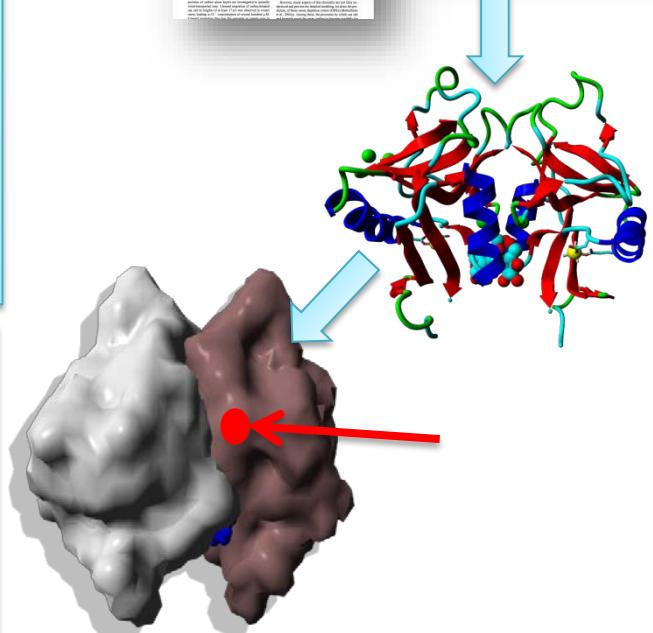
Very specific mutation,
anything to Proline in a helix

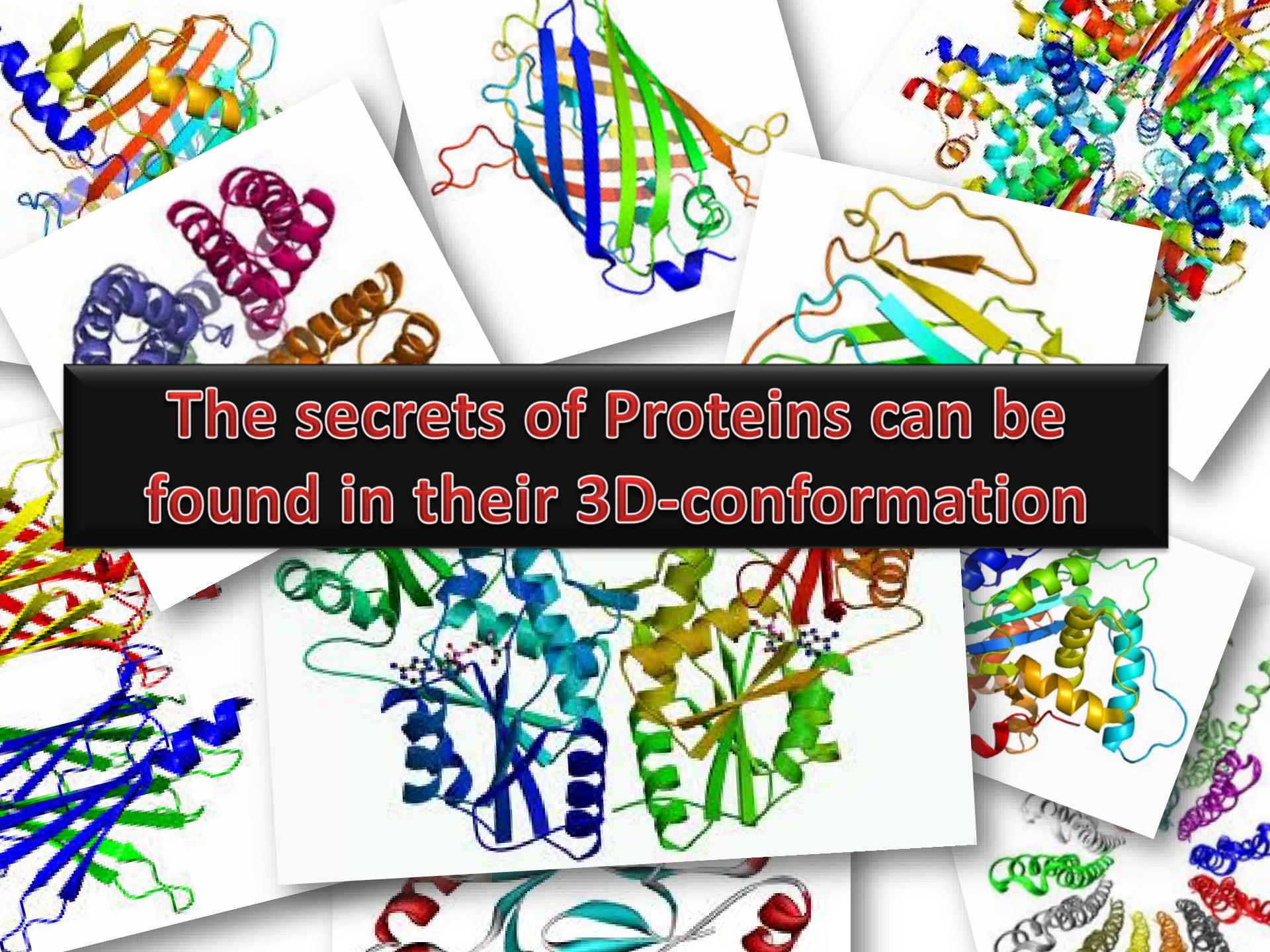


Missense mutations: How to analyze their effect on the protein's structure?

- Step 1: Find the sequence of your protein
 - Step 2: Find information about the function of the residues of interest (for example in Uniprot, from databases or literature, and maybe from your own experiments)
 - Step 3: Map this information where possible on the available proteins structure(s)
 - Step 4: Analyze what would happen to the function of the residue when it mutates into the new residue

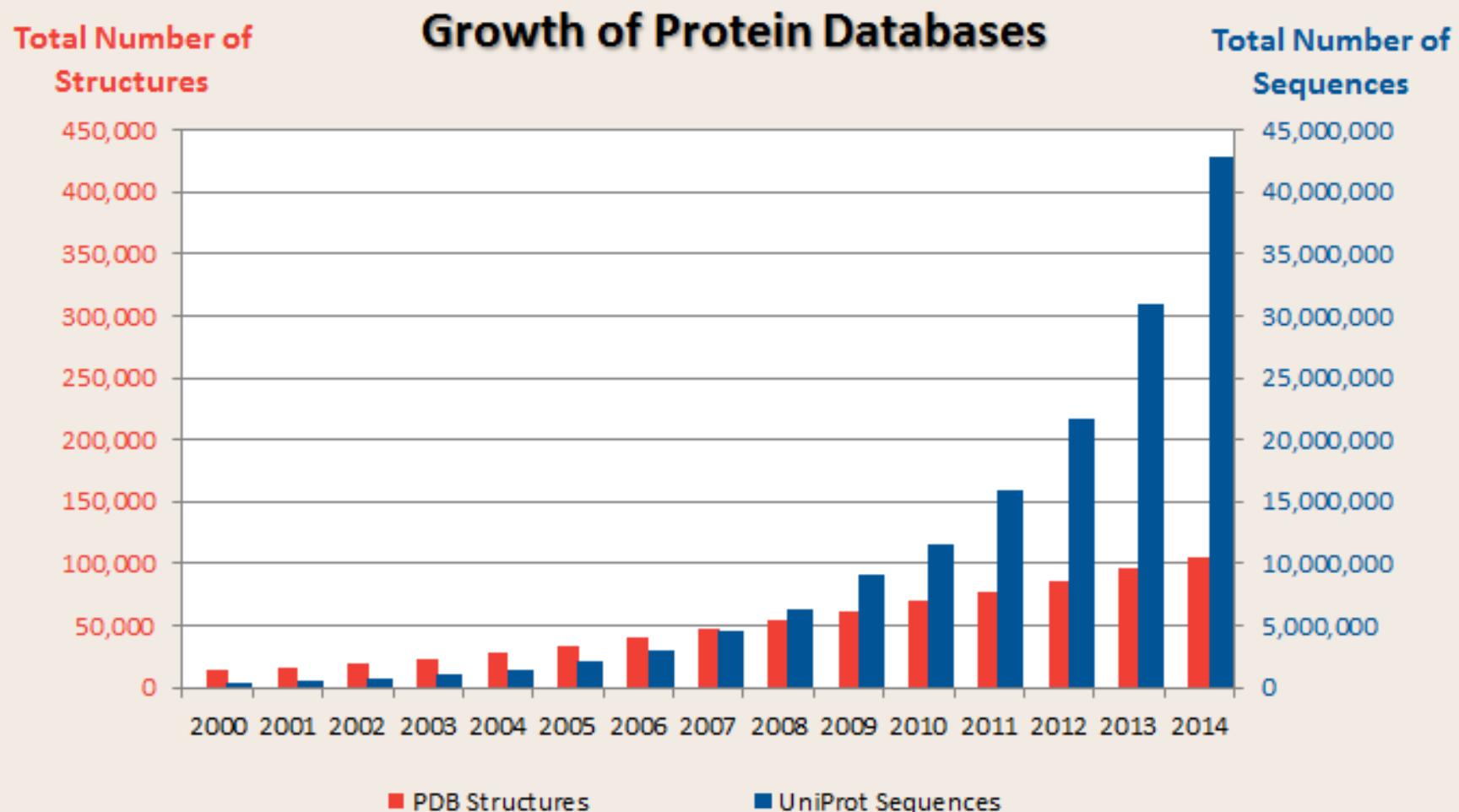
For this, you will need **information** about the structure and function of the protein and **combine** this with whatever is known about the wildtype and mutant residue.





The secrets of Proteins can be
found in their 3D-conformation

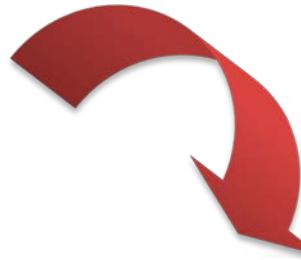
**Protein structures contain a wealth of information...
...there is just 1 problem....**



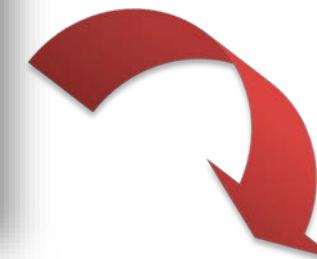
How to find your 3D-structure of interest?

```
>RELB_HUMAN
MLRSGPASGPSVPTGRAMPSRRVA
RPPAAPELGA LGSPDLSSSLA V S
RSTDELEI IIDEYIKENGFG LDGGP
GEGLPRLVSRGAASLSTVTLGPVA
PPATPPPWCPLGRLVSPAPGP GP
QPHLVITEQPKQRGMRF RYECRSA
GSILGESSTEASKTLPAIELRDCG
GLREVEVTACLVWKDWPHRVH PHS
LVGKDCTDGICRVRLRPHVRHSFN
NLGIQCVRKKEIEAAIERKIQLGI
DPYNAGSLKNHQEVDMNVVRICFQ
ASYRDQQQMRRMDPVLYDKKST
NTSELRICRINKESGPCTGG EELY
LLCDKVQKEDISVVFSRASWEGRA
DFSQADVHRQIA
```

1: Take the Amino acid sequence



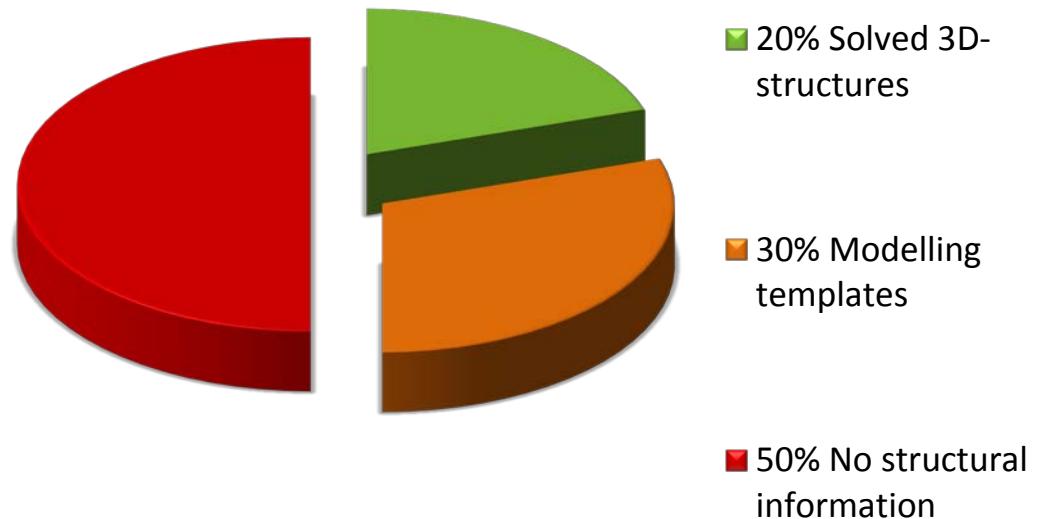
2: BLAST against the PDB



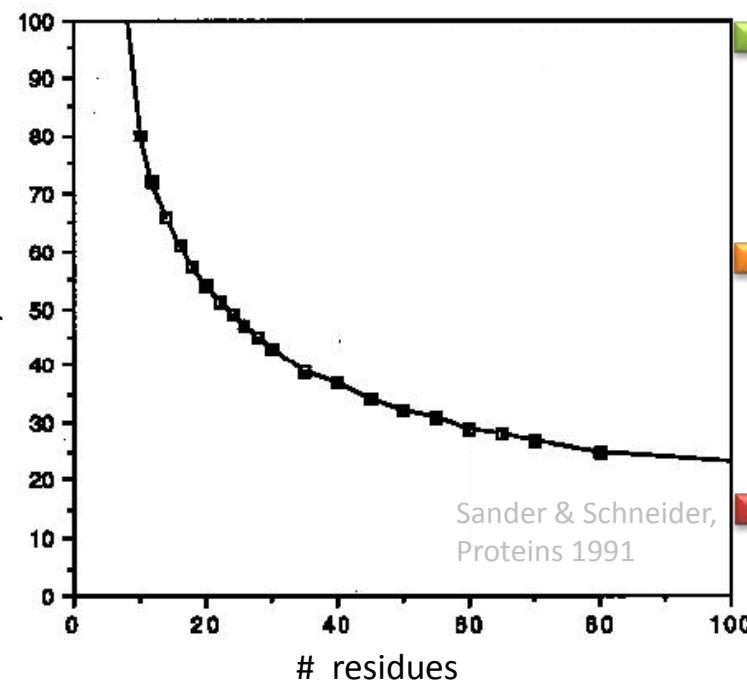
3: Analyze the results, you have 3 options:

3D-structures for human proteins

PDB BLAST
=
3 possible
results



Sander & Schneider plot



100% BLAST hit, PDB-file = structure

Enough sequence identity to build model

No (homologous) structure, use sequence-based predictions

Option 1:
100% PDB hit
⇒ Structure exists
⇒ Download and
study it

RCSB PDB PROTEIN DATA BANK

ATOM	3	C	GLY	A	1	46.692	49.648	101.284	0.0
ATOM	4	O	GLY	A	1	46.895	50.222	102.381	0.0
ATOM	5	N	SER	A	2	47.283	48.516	100.951	1.0
ATOM	6	CA	SER	A	2	46.277	47.866	101.761	1.0
ATOM	7	C	SER	A	2	49.212	47.031	100.845	1.0
ATOM	8	O	SER	A	2	49.060	47.195	99.630	1.0
ATOM	9	CB	SER	A	2	47.438	47.091	102.800	1.0
ATOM	10	OG	SER	A	2	46.276	46.356	102.404	1.0
ATOM	11	N	HIS	A	3	50.147	46.186	101.370	1.0
ATOM	12	CA	HIS	A	3	51.129	45.389	100.609	1.0
ATOM	13	C	HIS	A	3	50.953	43.905	100.849	1.0
ATOM	14	O	HIS	A	3	50.530	43.595	101.950	1.0
ATOM	15	CB	HIS	A	3	52.555	45.574	100.990	1.0
ATOM	16	CG	HIS	A	3	52.940	47.090	100.611	1.0
ATOM	17	ND1	HIS	A	3	53.371	47.470	99.422	1.0
ATOM	18	CD2	HIS	A	3	52.956	48.175	101.433	1.0
ATOM	19	CE1	HIS	A	3	53.676	48.730	99.476	1.0

PDB file

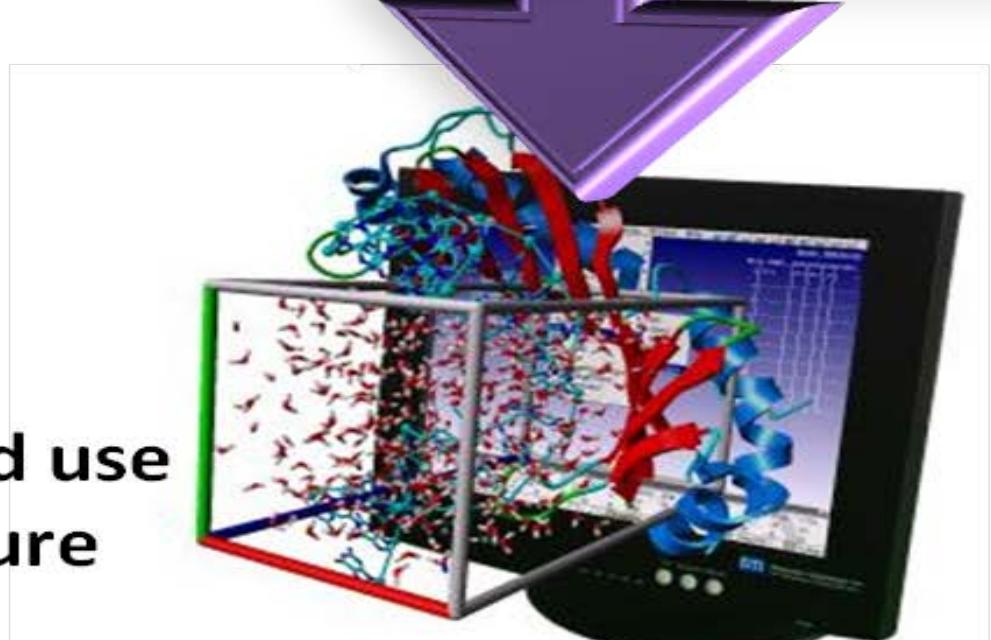


But also: PyMol, Jmol, VMD, Chimera

Visualization and use of the 3D structure

More on :

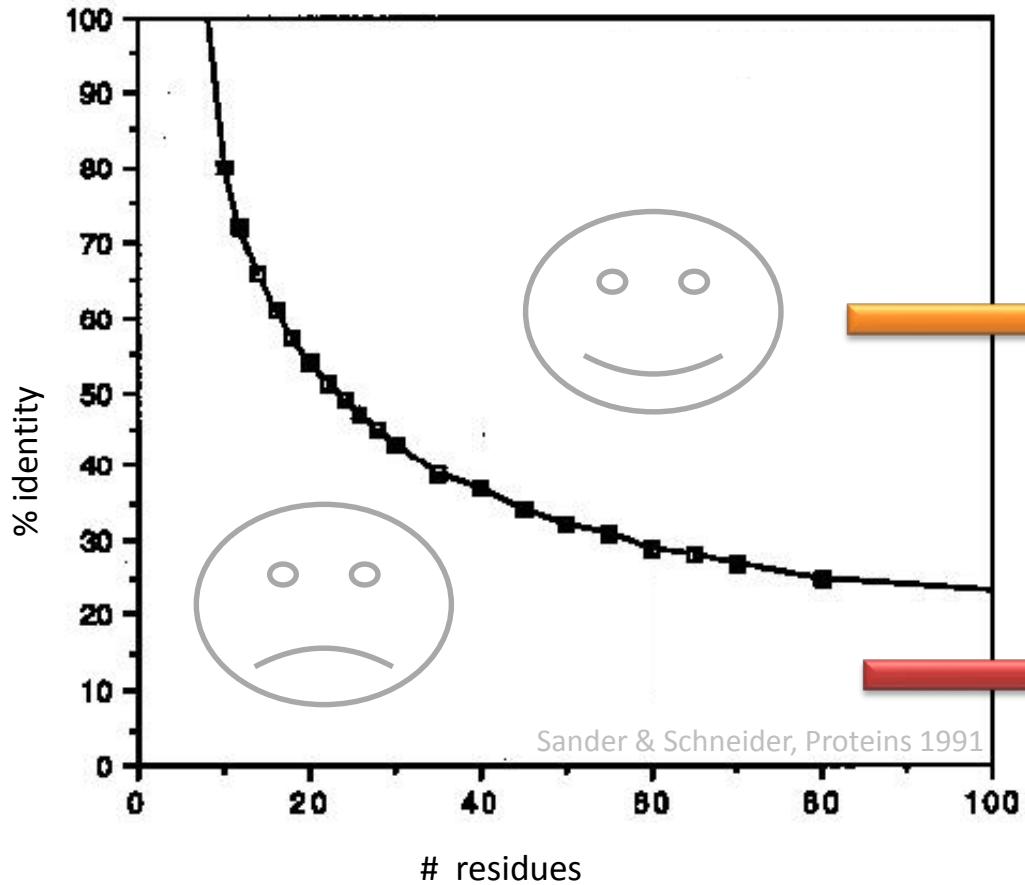
https://www.rcsb.org/pdb/static.do?p=software/software_links/molecular_graphics.html



Option 2:

A structure with % identity between 30-99% exists
⇒ you found a homolog
⇒ Build a homology model and study it

Sander & Schneider plot



Safe homology modelling zone,
quality will depend on % identity
in your alignment

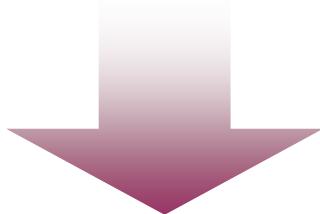
No-go area

Homology modeling in short...

Prediction of structure based upon a highly similar structure

2 basic assumptions:

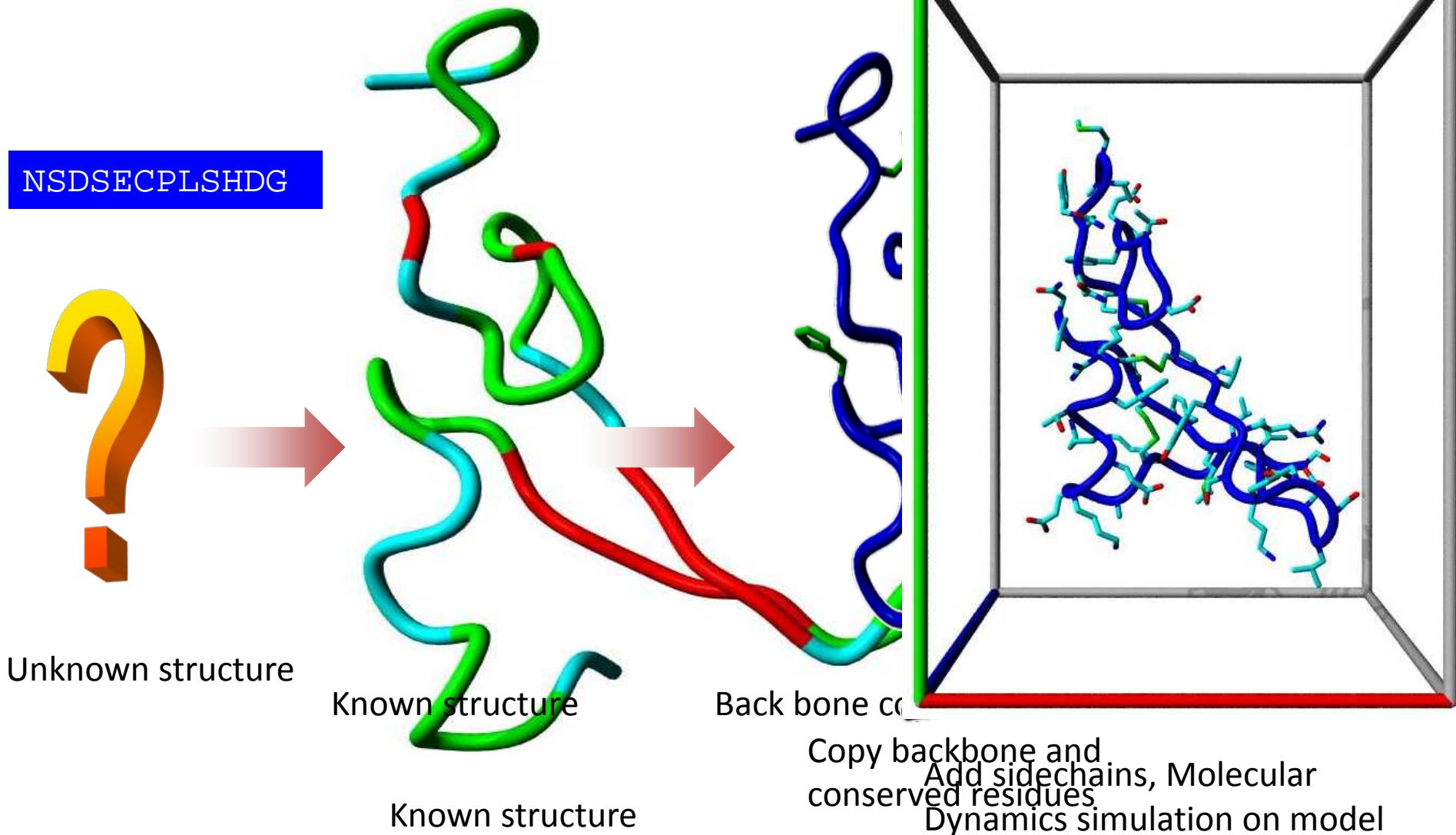
- Structure defines function
- During evolution structures are more conserved than sequence



Use one structure to predict another

Homology modeling in short...

Prediction of structure based upon a highl...



KGIAGPGLPD

PRAQV

LVVTI

TVLLR

RAFSYVL

LTTLDHWSS

PVKGQILMRL

VLELGTYCGY

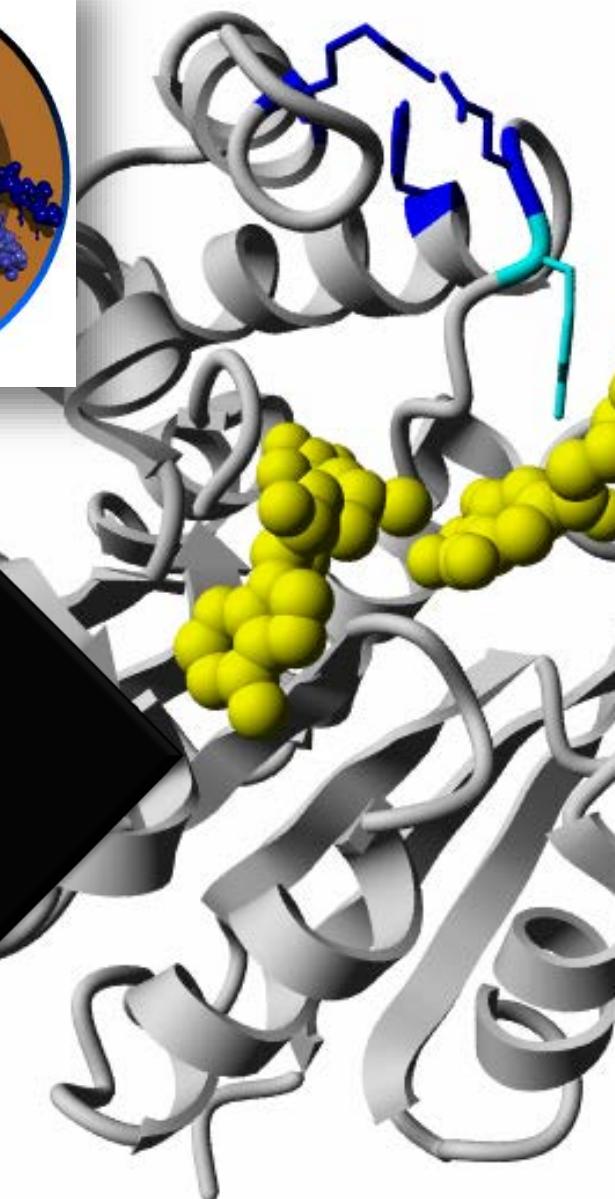
LPPGGRLLT

VAEKLIRAG

IVGSSEDVIP



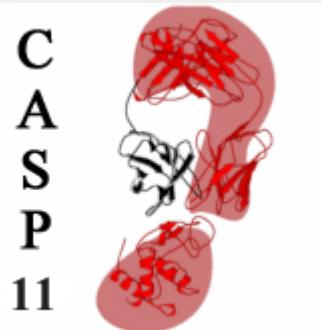
Magic
happens
here..



Other modelling
programs/tools/packages



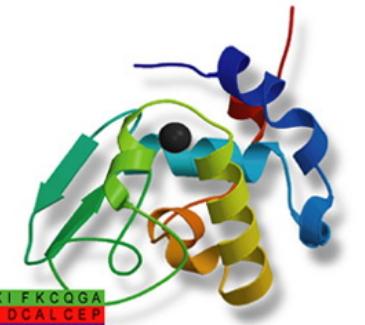
FoldX



Modeller

Program for Comparative Protein
Structure Modelling by Satisfaction
of Spatial Restraints

A I L V G S M P R R D G M E R K D O L L K A N V K I F K C Q G A
V E V C P Y D C F Y E G P N P L V I H P D E C I D C A L C E P
G A C K P E C R V N I Q G S - - Y A I D A D S C I D C G S
Q - - I A C G A C K P E C P V N I Q G S - - Y A I D A D S



 ROBETTA
Full-chain Protein Structure Prediction Server

BETA

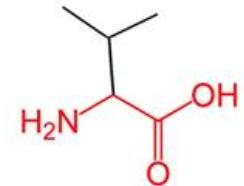
Phyre²



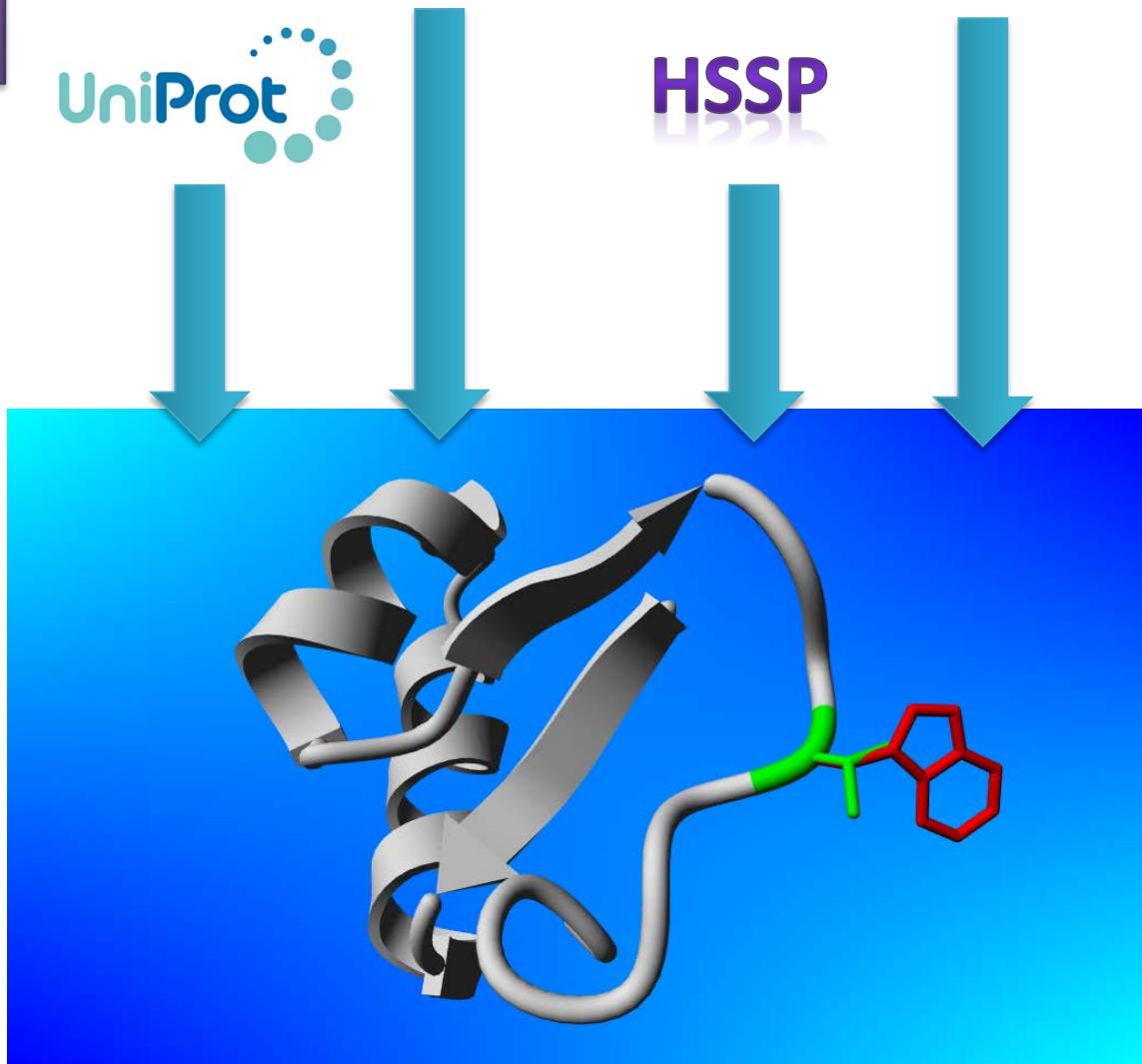
I-TASSER
Protein Structure & Function Predictions

Use the 3D-structure/model to map information from literature, Swissprot, conservation scores, your own knowledge to analyse the protein

Homology models can be used as a “normal” PDB structure, just keep the quality in mind

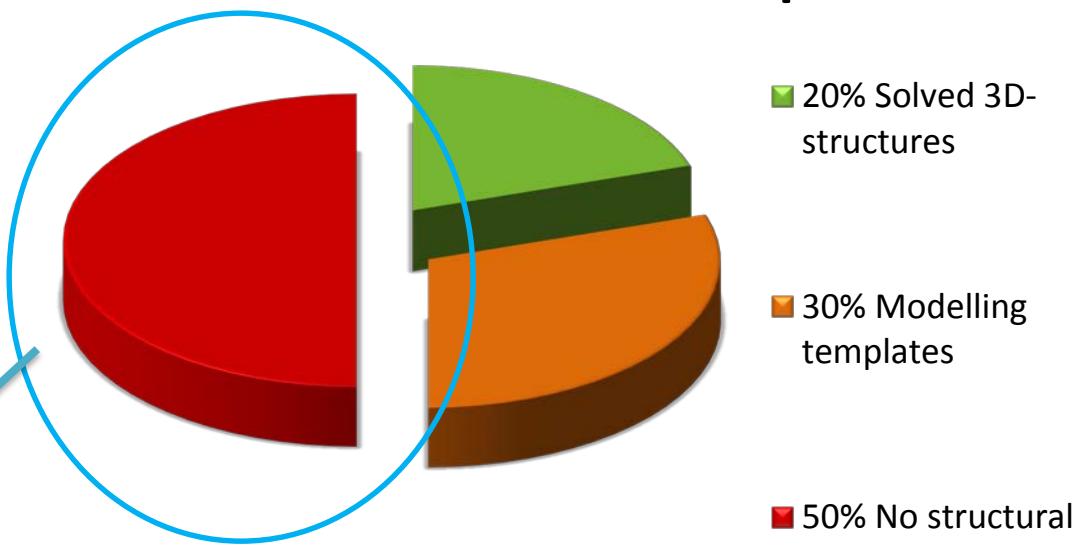


HSSP



Option 3:
No structure or
possible modelling
template available
⇒ No 3D-information
⇒ Use predictions

3D-structures for human proteins



What can we predict based on sequence alone?

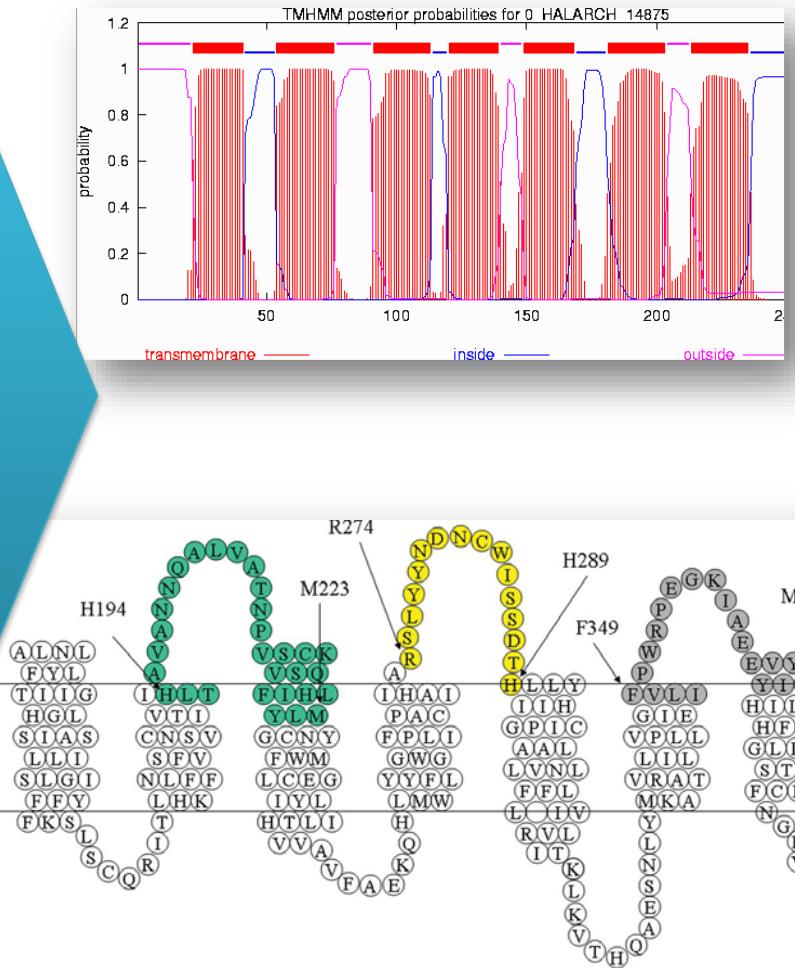
- Transmembrane regions
- Secondary Structure
- Accessibility
- Domains
- Motifs such as signal peptides/localisation signals
- Conservation (using Multiple Sequences)
- Disordered regions

One example: Transmembrane predictions

```
>RELB_HUMAN
MLRSGPASGPSVPTGRAMPSRRVA
RPPAAPELGALGSPDLSSLSLAVS
RSTDELEIIDEYIKENGFLGDGGP
GEGLPRLVSRGAASLSTVTLGPVA
PPATPPPWCPLGLRVSPAPGPGP
QPHLVITEQPQKQRGMRFRYECRSA
GSILGESSTEASKTLPAIELRDCG
GLREVEVTACLVWKDWPHRVHPHS
LVGKDCTDGICRVRRLRPHVRHSFN
NLGIQCVRKKEIEAAIERKIQQLGI
DPYNAGSLKNHQEVDMNVVRICFQ
ASYRDQQGQMRRMDPVLVYDKKST
NTSELRICRINKESGPCTGGEELY
LLCDKVQKEDISVVFSRASWEGRA
DFSQADVHRQIA
```

Choose one:

- TMHMM
- PsiPred
- Tpred
- DAS
- Phobius
- OCTOPUS
- PREDTMR
- And many more

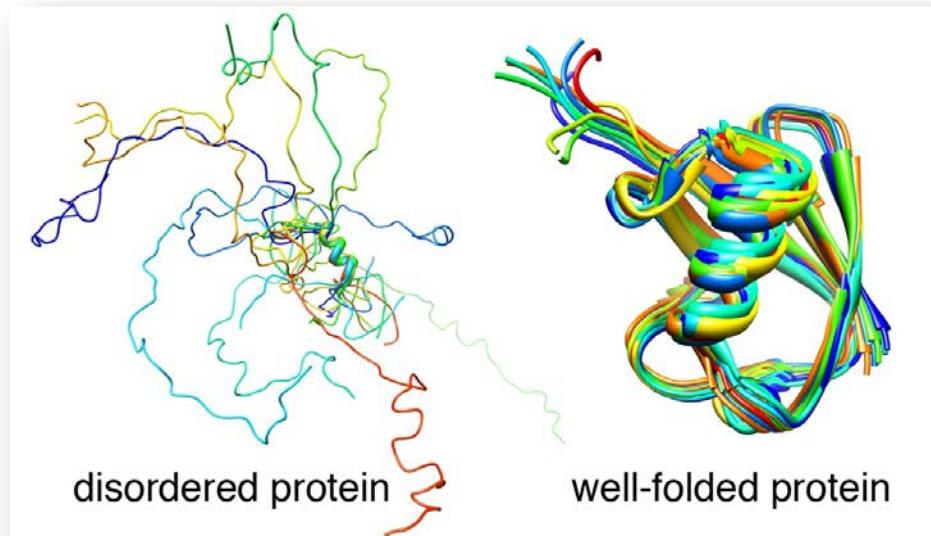


List partly from: https://molbiotools.ca/Protein_secondary_structure.htm

Another example: Disorder prediction

It has been estimated that at least 50% of eukaryotic proteins possess at least one long (>40-amino-acid) loop, while this fraction is lot lower in prokaryotes and Archaea.

From: Supratim Choudhuri,
in [Bioinformatics for Beginners](#), 2014



KMAD About Methods Download Why Examples API ▾

Job status: SUCCESS

Disorder prediction

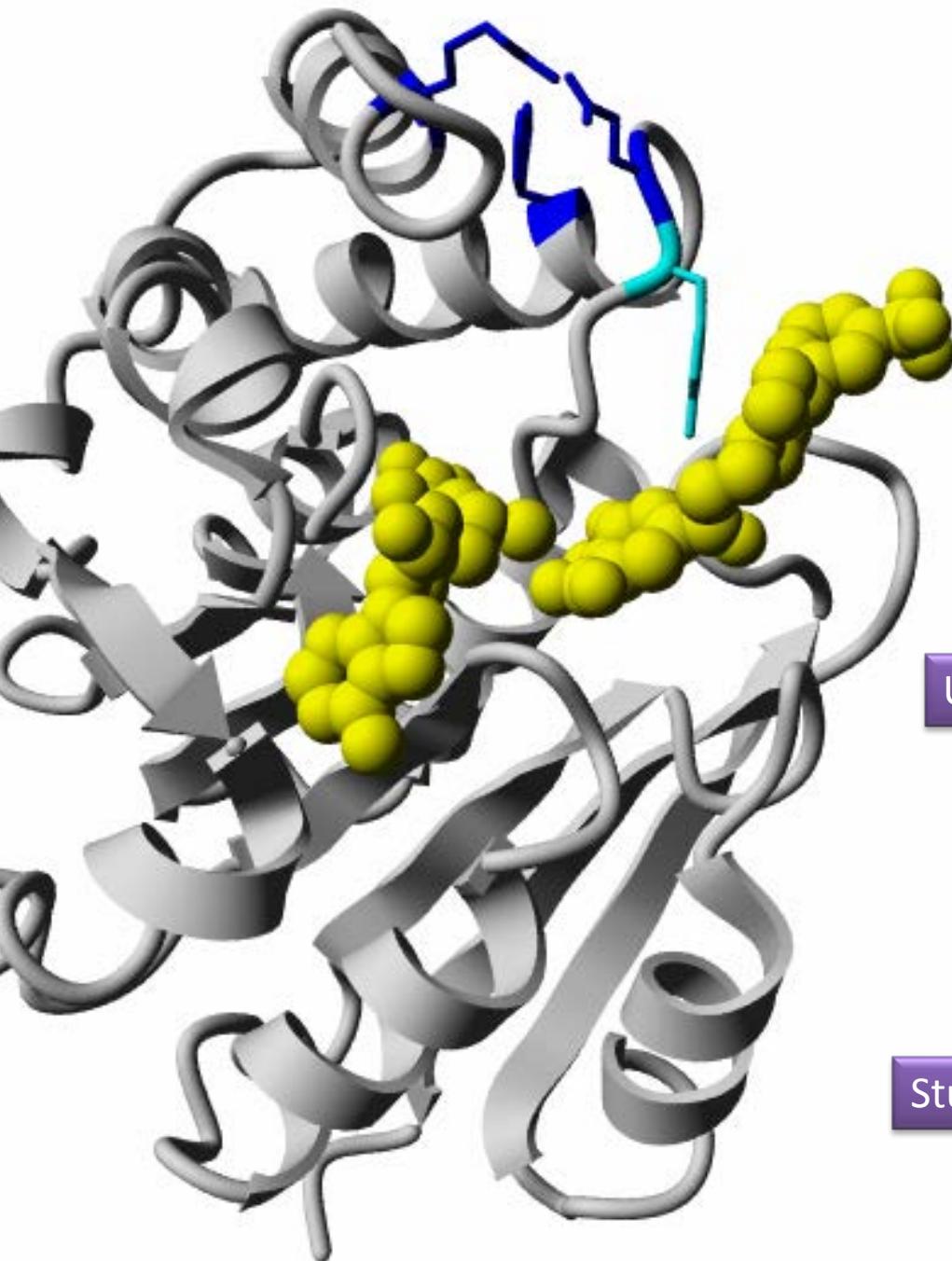
A - disordered A - ambiguous disorder prediction A - structured

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100	105	110	115	120	125		
filtered	MVLSPADKTNVKA	A	WGKVGA	HAGEYGA	EALERMF	L	FPTTKTYFPHFDL	SHGS	AQVK	GKGKK	VADAL	TNAVA	HVDMPN	ALS	SDLH	AHK	L	RVD	PVNFKL	L	SHCL	L	VTLAAH	L	P	PAEFTPAVHASLDKFLASVSTVLT	SKYR
consensus	MVLSPADKTNVKA	A	WGKVGA	HAGEYGA	EALERMF	L	FPTTKTYFPHFDL	SHGS	AQVK	GKGKK	VADAL	TNAVA	HVDMPN	ALS	SDLH	AHK	L	RVD	PVNFKL	L	SHCL	L	VTLAAH	L	P	PAEFTPAVHASLDKFLASVSTVLT	SKYR
globplot	MVLSPADKTNVKA	A	WGKVGA	HAGEYGA	EALERMF	L	FPTTKTYFPHFDL	SHGS	AQVK	GKGKK	VADAL	TNAVA	HVDMPN	ALS	SDLH	AHK	L	RVD	PVNFKL	L	SHCL	L	VTLAAH	L	P	PAEFTPAVHASLDKFLASVSTVLT	SKYR
psipred	MVLSPADKTNVKA	A	WGKVGA	HAGEYGA	EALERMF	L	FPTTKTYFPHFDL	SHGS	AQVK	GKGKK	VADAL	TNAVA	HVDMPN	ALS	SDLH	AHK	L	RVD	PVNFKL	L	SHCL	L	VTLAAH	L	P	PAEFTPAVHASLDKFLASVSTVLT	SKYR

Please cite: KMAD: Knowledge Based Multiple Sequence Alignment for Intrinsically Disordered Proteins DOI: 10.1093/bioinformatics/btv663

[Download](#)

Predictions tools:
-KMAD
-Disopred
-PsiPred
-Globplot
-PreDisorder
etc



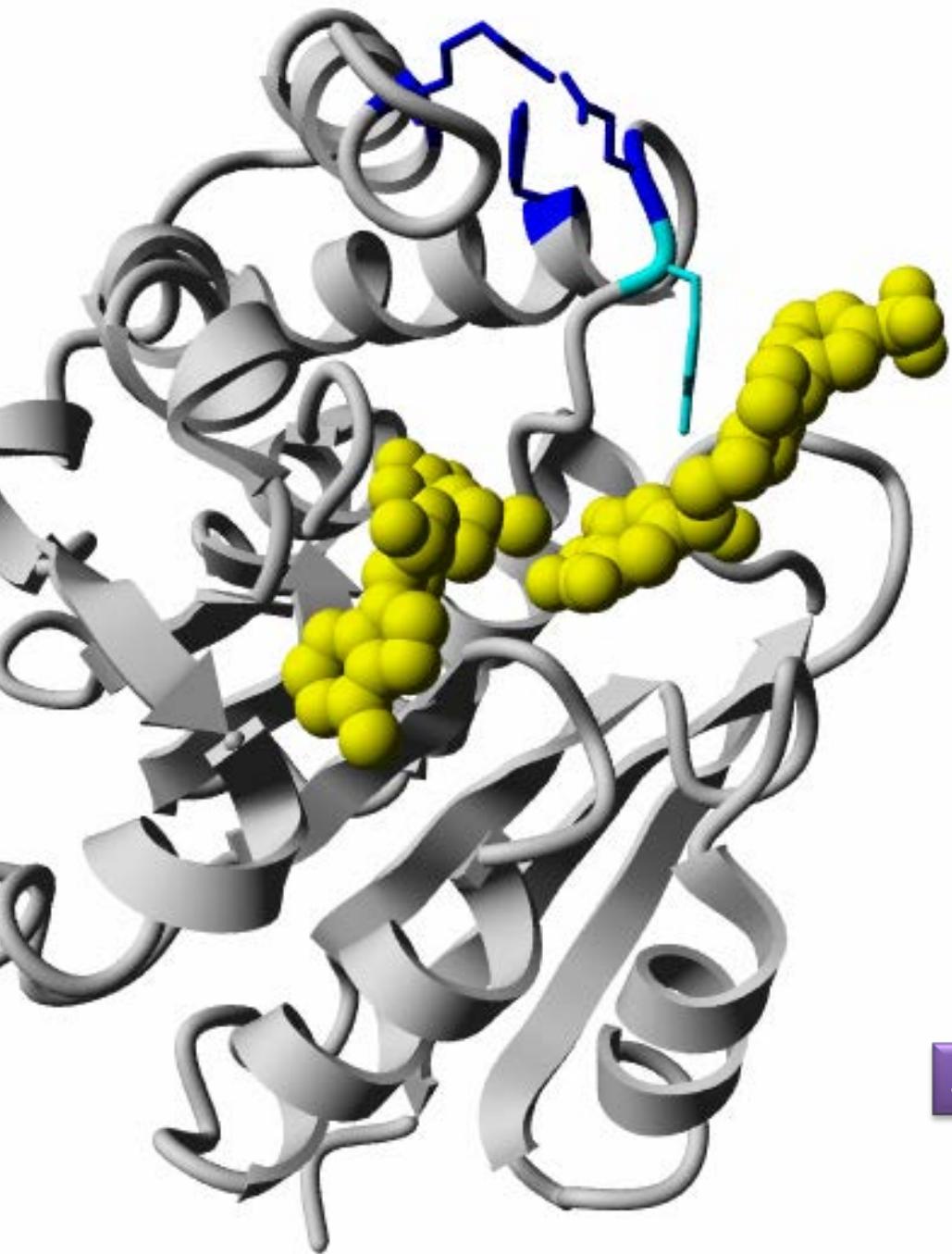
Using existing and predicted protein information we can:

Understand protein function

Understand protein localisation

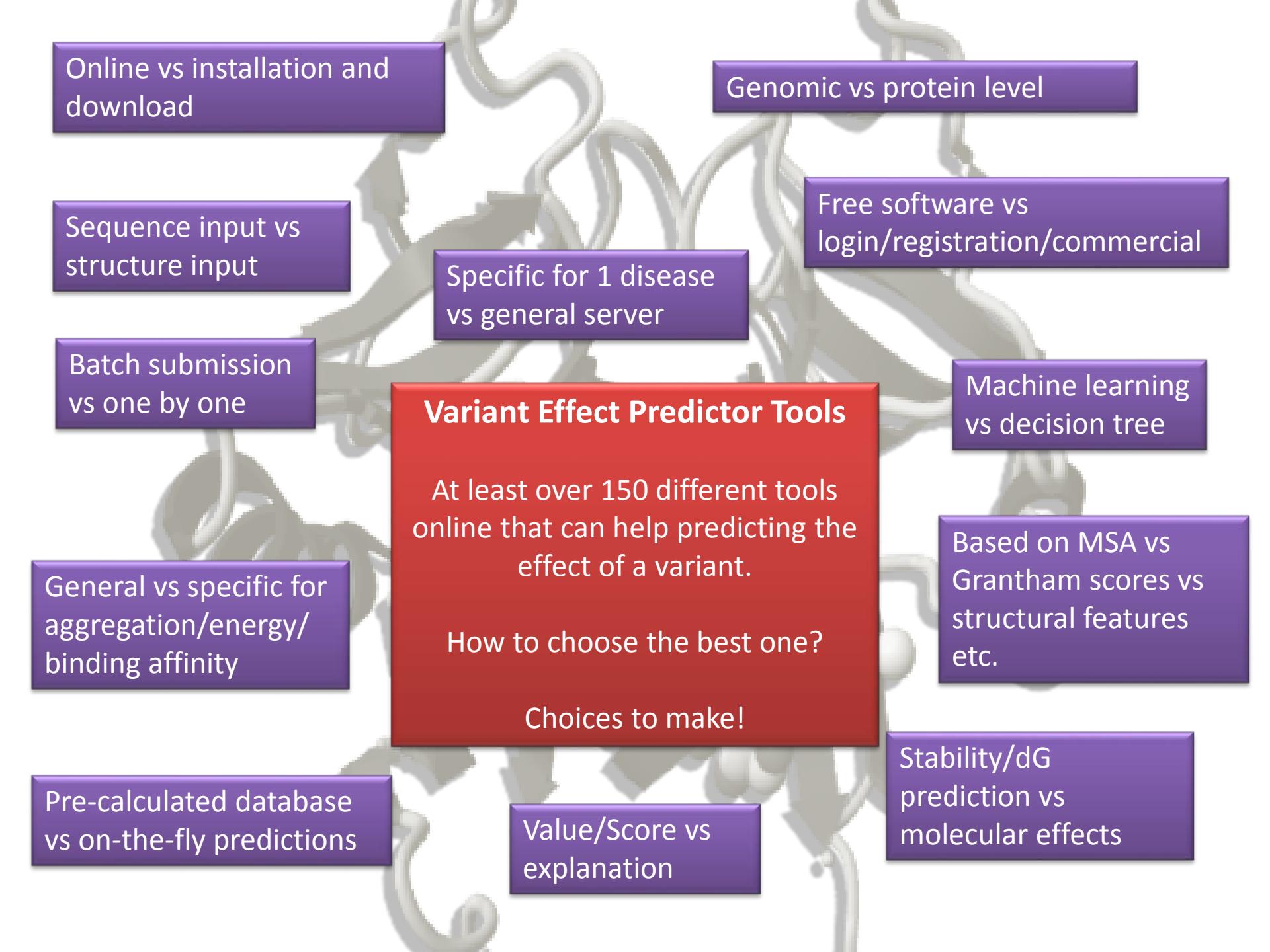
Design intelligent experiments

Study the effect of mutations



Variant Effect Prediction

Study the effect of mutations



Online vs installation and download

Sequence input vs structure input

Batch submission vs one by one

General vs specific for aggregation/energy/binding affinity

Pre-calculated database vs on-the-fly predictions

Genomic vs protein level

Free software vs login/registration/commercial

Machine learning vs decision tree

Based on MSA vs Grantham scores vs structural features etc.

Stability/dG prediction vs molecular effects

Variant Effect Predictor Tools

At least over 150 different tools online that can help predicting the effect of a variant.

How to choose the best one?

Choices to make!

Value/Score vs explanation

Choices: Input, Method and Output

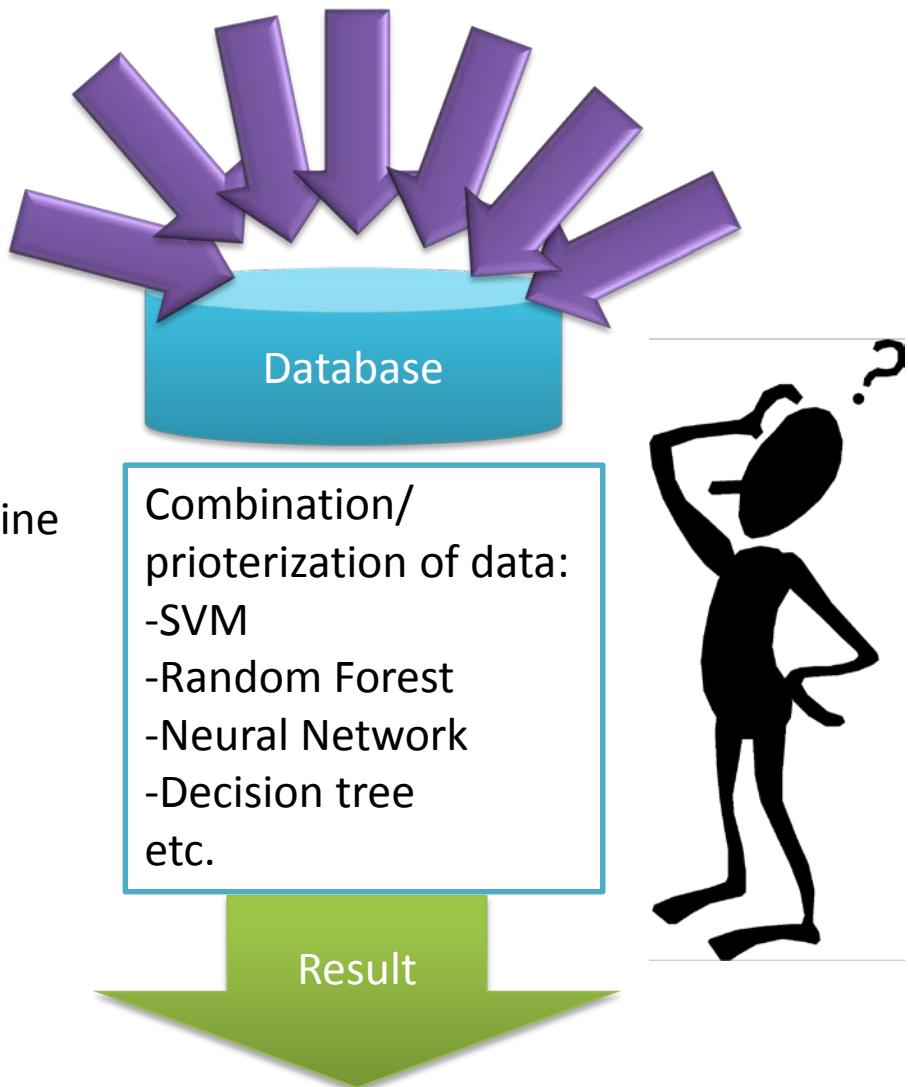
Input: What type of data do you have?

Method: Which data sources have been used??

Methods: Which method was used to combine the data and to draw a conclusion from the data?

...(and how was this method tested?)

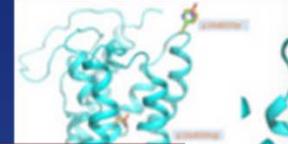
Output: What are the results?



Choices: Input type – Protein level vs Genomic level

KD4v - Comprehensible Knowledge Discovery System For Missense Variants

HOME | INTERPRETABLE RULES | WEB SERVICES | DOV

KD4v asks for Uniprot accession code, position and mutation

Prediction for new human missense variant

Uniprot Accession number
P20020
(SM2PH human protein, e.g P20711)

Position
100

Wild type residue
Gln (Q)

Mutant residue
Ile (I)

Knowledge base
For all human proteins

Prediction | **Reset**

Taster mutation t@sting

Gene HGNC gene symbol, NCBI Gene ID, Ensembl gene ID [show available transcripts](#) [clear input](#)

Transcript Ensembl transcript ID

Position / snippet refers to coding sequence (ORF) transcript (cDNA sequence) gene

Alteration **all types by sequence**

enter a few bases around your alteration

Format:

ACTGTC[A/T] GTGTF A substituted by T
ACTGTC[AG/T] GTGTF AG substituted by T
ACTGTC[ACGT/] GTGTF ACGT deleted
ACTGTC[-/AA] GTGTF AA inserted

single base exchange by position

enter position and new base

insertion or deletion by position

enter positions of ...last wild type base before alteration
...first wild type base after alteration and the inserted bases (if applicable)

MutationTaster asks for ORF/cDNA or gene, special format to indicate mutant

- single query
- query_chromosome
- QueryEngine
- MutationDistill
- RegulationSpot
- other applications
- slides ESHG20

PolyPhen-2 prediction of functional effects of human nsSNPs

[Home](#)[About](#)[Help](#)[Downloads](#)[Batch query](#)[WHESS.db](#)

Phenotyping v2) is a tool which predicts possible impact of an amino acid substitution on the structural and comparative considerations. Please, use the form below to submit your query.

Query Data

Protein or SNP identifier

Protein sequence
in FASTA format

```
>gnl|sprot|RELB_HUMAN Transcription factor RelB  
MLRSGPASGPSVPTGRAMPSRRVARPPAAPEL GALGSPDLSSL SLAVSR  
TDELEIIDEYIKENGFG LDGGQPGPGEGLPRLVSRAASLSTVTLGPVP  
PATPPPWGCPCLGRLVSPAPGPGPQPHLVITEQPKQRGMFRYCEGRSAG  
SILGESSTEASKTLPAIELRDCGGLREVEVTACLVWKDWPHRVHPHSLVG
```

Position

250

Substitution

AA ₁	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
AA ₂	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V

Query description

Click to select **Gly** (Glycine)

[Submit Query](#) [Clear](#) [Check Status](#)

Choices: Input type

PolyPhen asks for
sequence, position
and mutant

Select a Residue to Mutate

10 20 30 40 50 60 70 80 90 100 110

MLRSGPASGPSVPTGRAMPSRRVARPPAAPEL GALGSPDLSSL SLAVSR TDELEIIDEYIKENGFG LDGGQPGPGEGLPRLVSRAASLSTVTLGPVP APPATPPPWGCP

Position: 67

0 130 140 150 160 170 180 190 200 210

GPGPQPHLVITEQPKQRGMFRYCEGRSAGS SILGESSTEASKTLPAIELRDCGGLREVEVTACLVWKDWPHRVHPHSLVGKDCTDGICRVRLRPHVS

HOPE asks for the
sequence and allows
you to indicate the
mutation

PREV

NEXT

Choices: Input type – PDB files

✓ Predict Mutant Stability from Existing PDB structures

PDB ID (4 letters) [View Structure from PDB](#)

Experimental Method Thermal Denaturants

Stability Prediction: one amino acid all amino acids *

[Reset values](#)

[Predict Stability](#)



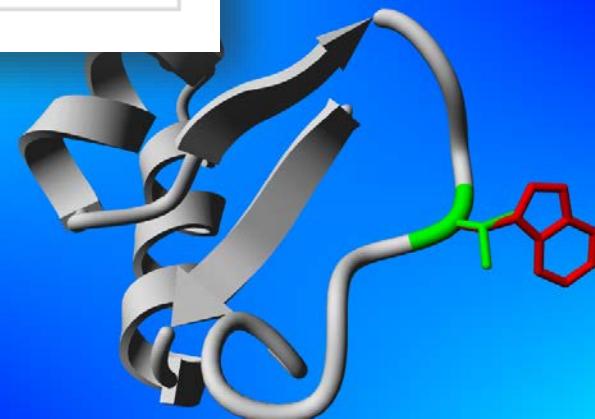
✓ Predict Mutant Stability from Custom Protein Structures

Choose a Protein Structure File (in PDB file format)

[Choose File](#) No file chosen

[Reset values](#) [Upload](#)

CUPSAT needs a
PDB file. You can
upload your own
model



Choices: Input type – extra info needed?

Multiple sequence alignment in FASTA format:*

>sp|P04637|P53_HUMAN - Homo sapiens (Human).
SVP-SQKT-YQGSYGFRLGFL
>sp|P10360|P53_CHICKEN - Gallus gallus (Chicken).
VVPSTEDYGGDFDFRVGVFV

Please make sure that in your alignment '-' correspond to gaps and 'X' to unknown amino-acids because the results will be different.

(click here for more details)

Select your file to upload: No file chosen

OR

Paste your alignment:

OR

Select one of our library alignments:

--Select the gene--

(click here for more details)

-----Select the depth-----

Substitutions list:*

S96T
S96P
S96A

(click here for more details)

Select your file to upload: No file chosen

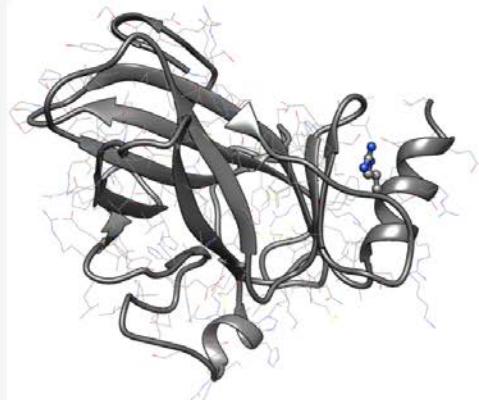
OR

Paste your list of substitutions:

Align-GVD asks for a simple substitution list on protein level, but you need to provide your own MSA (or choose one of the 20 pre-calculated ones)

E-mail (optional):

Choices: Input type – batch mode (for a single protein)



Run example

Disclaimer

No PDB files will be retained on the system after being uploaded by the user.

SDM allows single mutations but also a file with a list of mutations

Single mutation

Description

Wild-type protein structure - PDB format (Ex.: [2OCJ](#))

Choose File No file chosen

OR

Provide a 4-letter PDB code:

(Example: 2OCJ)

Mutation (Example: R282W)

Mutation chain (Example: A)

Predict reverse mutation

Run SDM

Mutation list

Description

Wild-type protein structure - PDB format (Ex.: [2OCJ](#))

Choose File No file chosen

OR

Provide a 4-letter PDB code:

(Example: 2OCJ)

Mutation list file Format (Ex.: [2OCJ_mut_list](#))

Choose File No file chosen

Predict reverse mutation

Run SDM

Choices: Input type – Genome level, VCF files

The screenshot shows the CRAVAT web application interface. It consists of three main sections: 1. Input, 2. Analysis, and 3. Results.

1 Input: This section allows users to upload a VCF file or paste its contents. It includes a text area for variants and a file upload field. A red callout box highlights the feature: "CRAVAT allows both file upload and copy/paste of the VCF file".

2 Analysis: This section lets users choose analysis programs. Options include VEST-4, CHASM-3.1, and GeneCard and PubMed annotation.

3 Results: This section provides options to send the analysis report via email and includes a checkbox for including text reports.

Header: CRAVAT 5.2.1 - Now in GRCh38!
Still using GRCh37/hg19? ?
Follow @CravatMupitTool

[Help](#) | [Release notes](#) | [How to Cite](#) | [Contact](#) | [Log-in](#) >

SUBMIT
What will I get?

Choices: Input type – VCF files and more...

wANNOVAR

Home Tutorial Example Related projects ▾

WGLAB Recent Updates

Basic Information

Email

Sample Identifier

Input File

or Paste Variant Calls

I agree to the [Terms of Use](#). Please note that commercial users need to obtain a [license](#).

Disease/Phenotype

Enter Disease or Phenotype Terms

wANNOVAR allows variant calls, but also needs more parameters

[10/19/2017] The detailed amino acid changes for indels are now included in the output (through -polish argument in table_annovar). The server also handles duplicated entries (multiple identical variants) in the input file correctly.

Parameter Settings

Result duration	<input type="text" value="15 days"/>	<input type="button" value="🔍"/>
Reference Genome	<input type="text" value="hg19"/>	<input type="button" value="🔍"/>
Input Format	<input type="text" value="VCF"/>	<input type="button" value="🔍"/>
Gene Definition	<input type="text" value="RefSeq Gene"/>	<input type="button" value="🔍"/>
Individual analysis	<input type="text" value="Individual analysis"/>	<input type="button" value="🔍"/>
Disease Model	<input type="text" value="rare recessive Mendelian disease"/>	<input type="button" value="🔍"/>

Choices: Input type – VCF files and more...

Phen-Gen

[HOME](#) [ABOUT](#) [INSTRUCTIONS](#) [DEMO](#) [RESULTS](#)

Phen-Gen online

Important note: We only accept 1 family per run. To evaluate unrelated

INPUT FILES [?]

Variants:* No file chosen

In [VCF](#). Sample VCF

Phenotypes:* No file chosen

In a text file with 1 [HPO](#) term per line. [Sample phenotype file](#)

Pedigree:* No file chosen

In [PED](#) format. [Sample PED](#)

RUN PARAMETERS [?]

Disease inheritance pattern:

- Recessive
- Dominant

Type of prediction:

- Coding
- Genomic

Stringency:

Only digits (0-9)

Your e-

To get n

Phen-Gen even needs a pedigree
file + parameters

[Remove association test](#)

Phenotype

one line each, 1 for control, 2 for case, . for missing

And I-Fish allows an
association test and
Pedigree test-file

Or [upload](#) from file

Model

general

model wish to test at each SNP in ass

[Remove co-segregation test](#)

Pedigree

1	101	0	0	1	1
1	102	0	0	2	2
1	201	0	0	1	1
1	202	101	102	2	2
1	203	0	0	1	2
1	204	101	102	2	1
1	205	101	102	1	1
1	206	101	102	1	1

Or [upload](#) from file

Allele
Frequency

0.0001

disease allele frequency used in co-segregation

Penetrance

AA

Aa

aa

Choices: Input type

The tool you use will depend on the type of input you have.

- Input files/data: Genome locations / Accession codes / Gene names / Protein structures
- How many variants do you have?
- Do you want to install your own software or use an online server?

Choices: Input, Method and Output

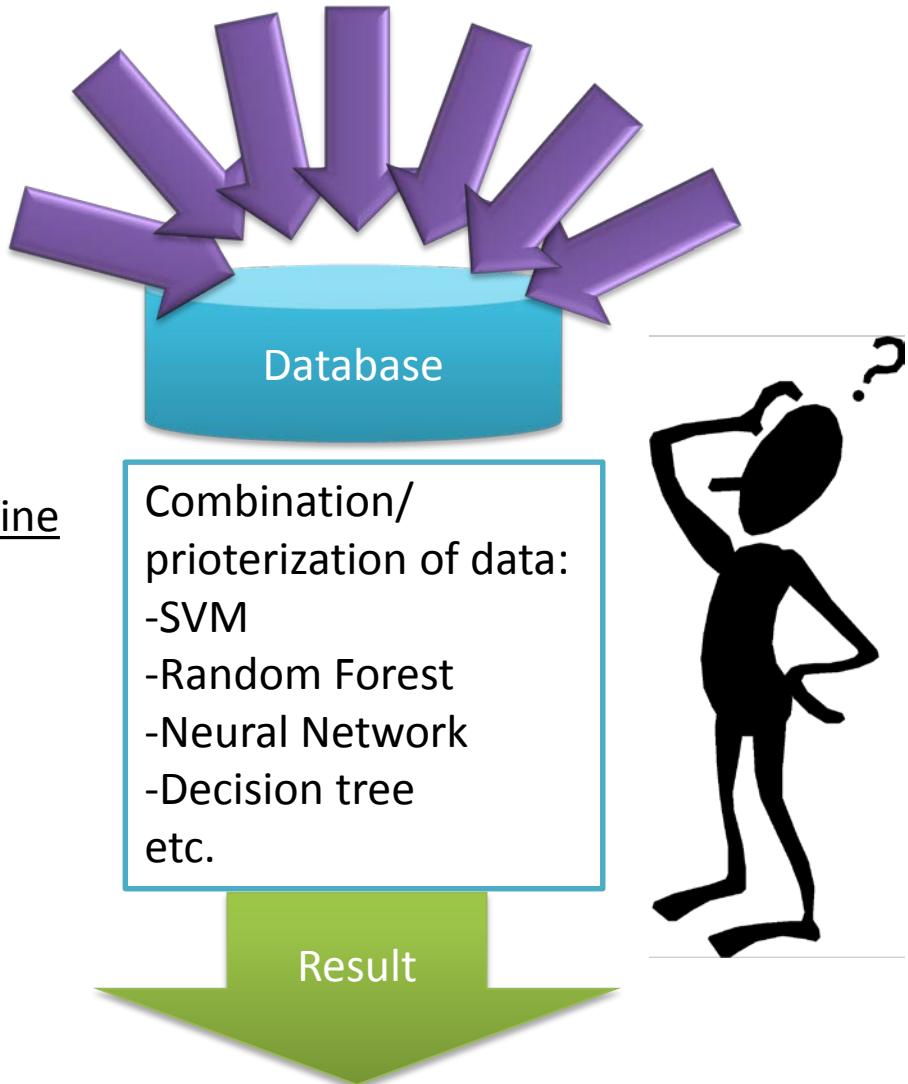
Input: What type of data do you have?

Method: Which data sources have been used??

Methods: Which method was used to combine the data and to draw a conclusion from the data?

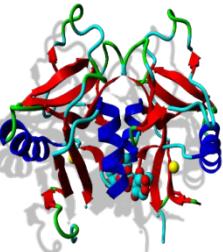
...(and how was this method tested?)

Output: What are the results?

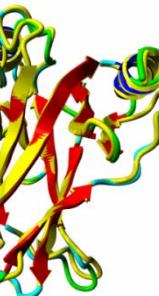


Possible Information sources for VEP tools:

3D Structures



Homology Models



Disorder prediction



Annotated SwissProt/
UniProt features



GO-terms



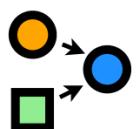
Domain Information



(Variant) Databases



Information specific for
protein/disease

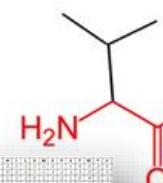
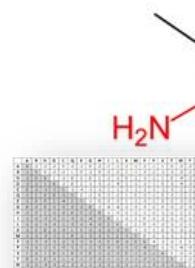


Sequence-based predictions

Position 12	Julia HD	Struct
R	S	K
I	T	L
V	N	A
A	G	M
S	C	H
T	D	F
P	E	R
Y	W	S
W	S	P
F	Y	T

(P, Y or W/JuSaPi7)

Conservation scores from
multiple sequence alignments

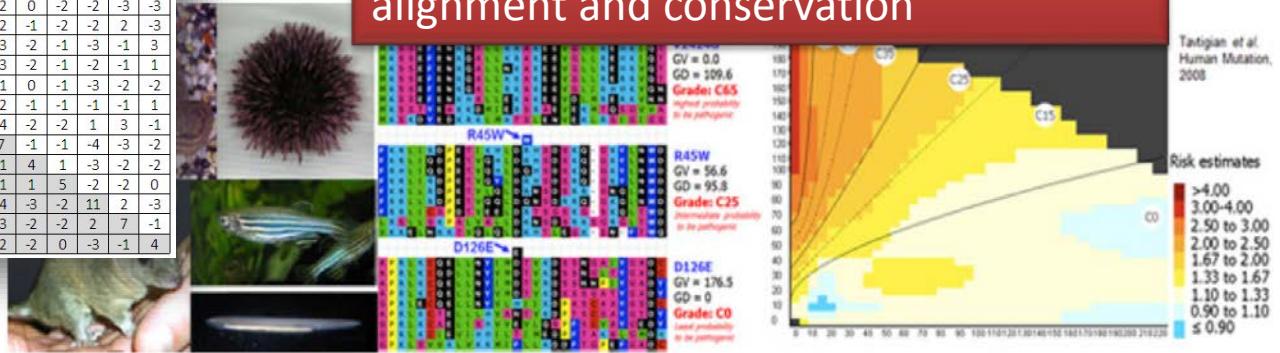


Amino Acid properties/
Grantham Scores

Choices: Methods - What type information was used and how?

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

Easiest option:
Grantham scores, Multiple sequence alignment and conservation



ALIGN GVGD

Align-GVGD is a freely available, web-based program that combines the biophysical characteristics of amino acids and protein multiple sequence alignments to predict where missense substitutions in genes of interest fall in a spectrum from enriched deleterious to enriched neutral. Align-GVGD is an extension of the original Grantham difference to multiple sequence alignments and true simultaneous multiple

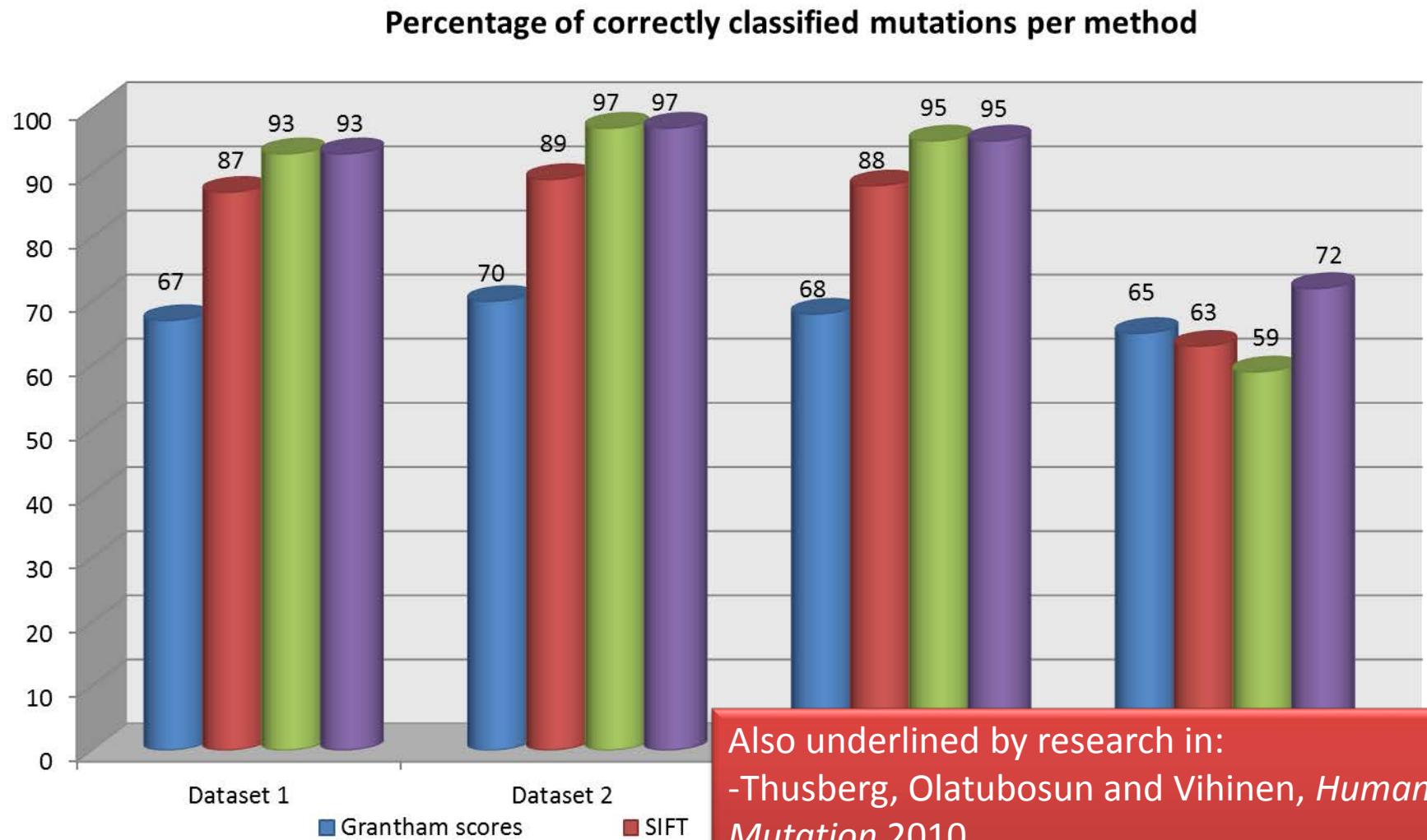


Sorting Intolerant From Tolerant

Advantage: Fast methods, always possible

Disadvantage: not the best results, information is missing

A small experiment shows that adding information to simple Grantham scores and conservation is beneficial for the prediction

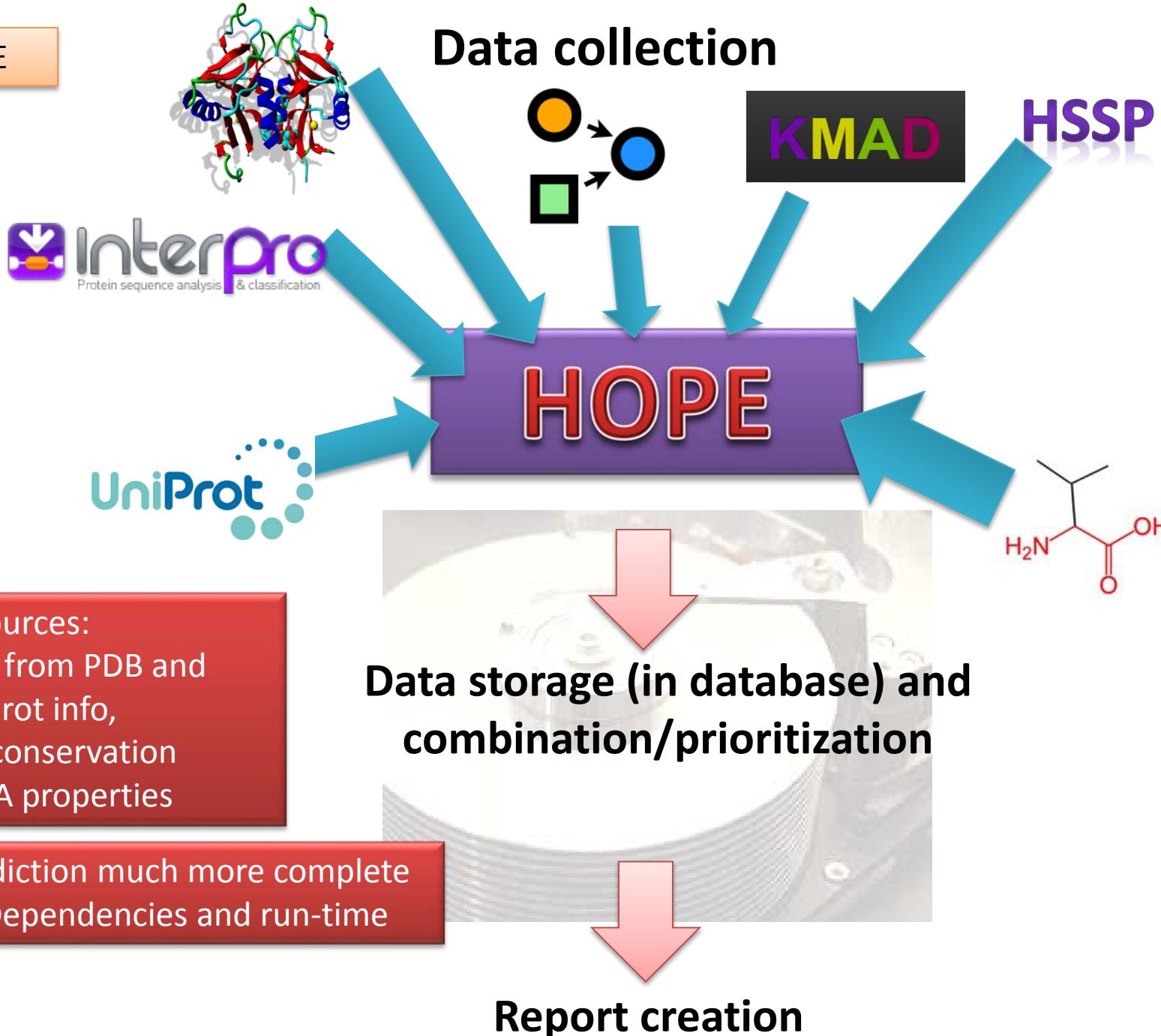


Also underlined by research in:
-Thusberg, Olatubosun and Vihinen, *Human Mutation* 2010
-Saunders and Baker, *JMB* 2002
-Bromberg and Rost, *NAR* 2007

Source: [BMC Bioinformatics](#). 2013 Dec 4;14:352.

Status quo of annotation of human disease variants

Example: HOPE



Choices: Methods – Testing / Benchmarking



All human proteins

	Description	VarMod: modelling the functional effects of non-synonymous variants.	Set of rules
1	We used Human Protein Variation Server. We analyzed missense variants split into 2 pathogenicity polymorphism and neutral polymorphism. Author information	Pappalardo M ¹ , Wass MN ² . Abstract Unravelling the genotype-phenotype relationship in humans remains a challenging task in genomics studies. Recent advances in sequencing technologies mean there are now	View Download (in Prolog code)

We have developed a sequence conservation-based artificial neural network predictor called NetDiseaseSNP which classifies nsSNPs as disease-causing or neutral. Our method uses the excellent alignment generation algorithm of SIFT to identify related sequences and a combination of 31 features assessing sequence conservation and the predicted surface accessibility to produce a single score which can be used to rank nsSNPs based on their potential to cause disease. NetDiseaseSNP classifies successfully disease-causing and neutral mutations. In addition, we show that NetDiseaseSNP discriminates cancer driver and passenger mutations satisfactorily. Our method outperforms other state-of-the-art methods on several disease/neutral datasets as well as on cancer driver/passenger mutation datasets and can thus be used to pinpoint and prioritize plausible disease candidates among nsSNPs for further investigation. NetDiseaseSNP is publicly available as an online tool as well as a web service: <http://www.cbs.dtu.dk/services/NetDiseaseSNP>.

that functional nsSNVs are enriched at protein-protein interfaces and protein-ligand binding sites and uses these characteristics to make predictions. In benchmarking on a set of nearly 3000 nsSNVs VarMod performance is comparable to an existing state of the art method. The VarMod web server provides extensive resources to investigate the sequence and structural features associated with the predictions including visualisation of protein models and complexes via an interactive JSmol molecular viewer. VarMod is available for use at <http://www.wasslab.org/varmod>.

ant dataset. INPS performs very well in predicting the effect of non-synonymous variants using a rule-based method mCSM. When

0.27 0.46 0.79

BER MCC AUC

Auto-Mute (RF, 400 trees, 10-fold CV, 1790 nsSNPs)	0.76	0.30	0.38	0.79
Bao and Cui (2005) (RF, 1000 trees, 10-fold CV, 4013 nsSNPs)	0.77	0.29	0.32	~ 0.75
Bao and Cui (2005) (SVM, 10-fold CV, 4013 nsSNPs)	0.76	0.32	0.27	~ 0.73

Choices: Methods – Test-sets

mCSM-AB

mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures

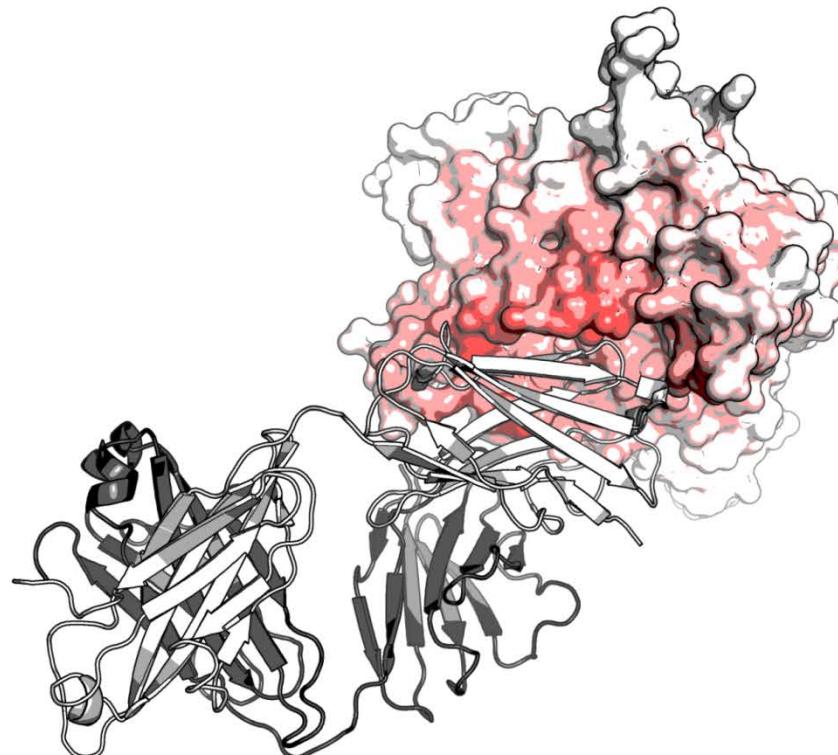
Douglas E. V. Pires & David B. Ascher

Nucleic Acids Research, v. 44 (W1), p. W469-W473, 2016.  

Abstract

Summary: Computational methods have traditionally struggled to predict the effect of mutations in antibody-antigen complexes on binding affinity. This has limited their usefulness during antibody engineering and development, and their ability to predict biologically relevant escape mutations. Here we demonstrate that graph-based signatures can be used to accurately predict the effect of mutations on antibody binding affinity. We show that mCSM-AB performs better than comparable methods that have been previously used for antibody engineering.

This mCSM-AB was created and tested for antibody-antigens complexes. Open door: other complexes might result in bad predictions



Choices: Methods

A few conclusion about testing/benchmarking

- Each tool seems to outperform other ones, but this depends on the test sets and the other tools
- Tools were created with different purposes in mind
- Check article / manual on website
- The Machine Learning method (SV/RF/etc.) doesn't really matter
- But: adding structure information does seem to improve the results

Choices: Input, Method and Output

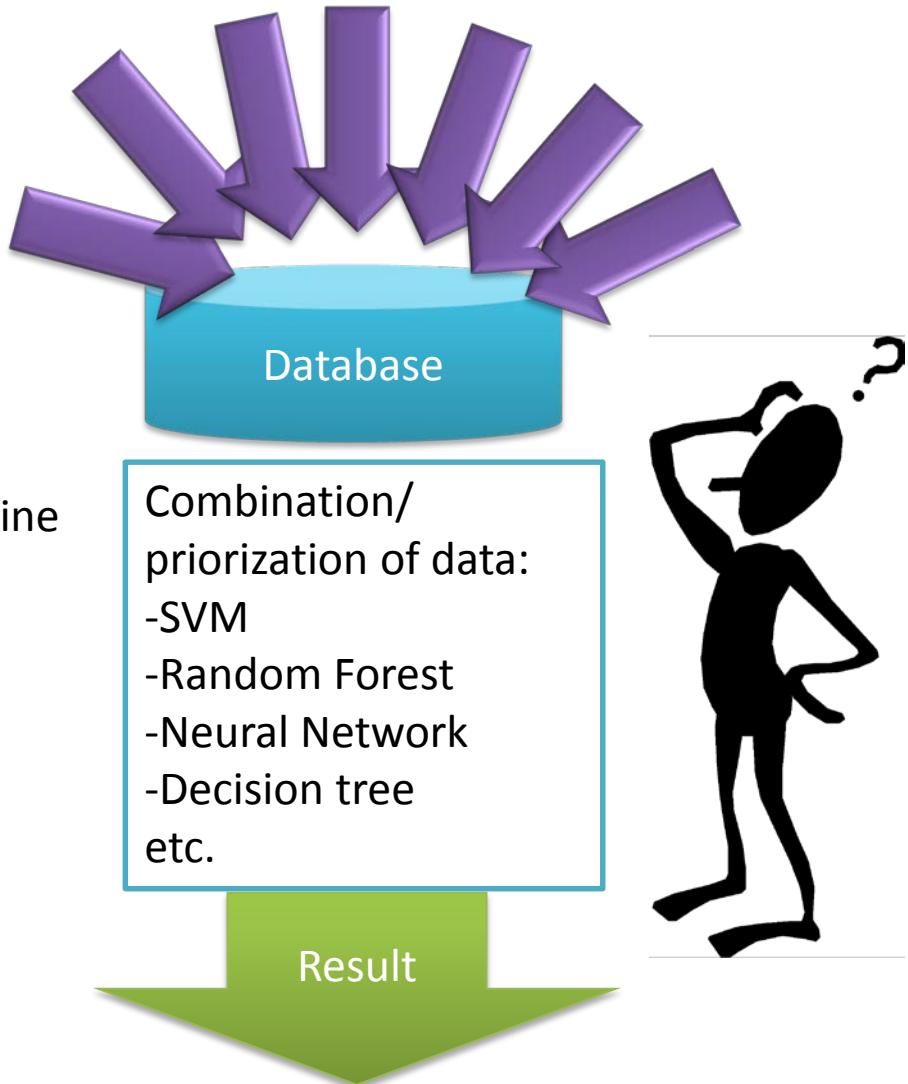
Input: What type of data do you have?

Method: Which data sources have been used??

Methods: Which method was used to combine the data and to draw a conclusion from the data?

...(and how was this method tested?)

Output: What are the results?



Choices: The Results

Wide variety in results-pages:

- Single value vs extensive test
- Pictures, graphs vs text
- Values for dG, scores vs a single word (disease/benign)
- Positive scores / negative scores
- Different ranges like 0-1 or 10-100
- Result by email vs online (or both)
- One result vs result list
- 3D structure
- Explanation of the effect

Position: 125
Original Amino Acid: P
Substitute Amino Acid: W

Prediction Results:

1. Predicted both value and sign of energy change using SVM and sequence information only.

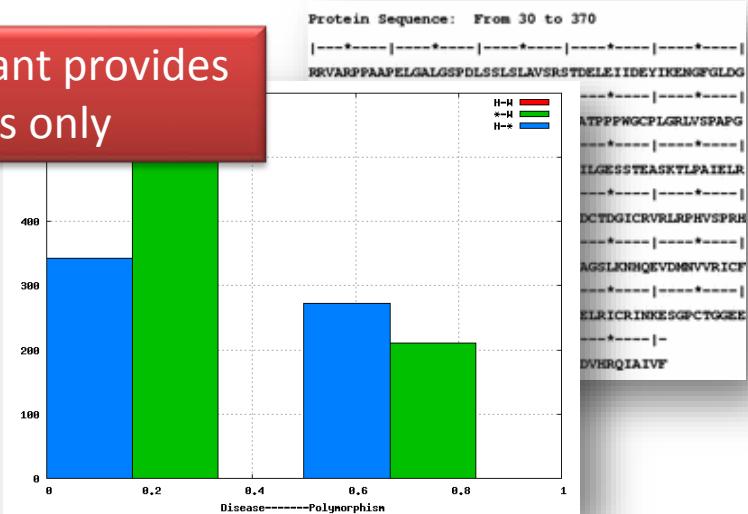
Detail delta G = -0.64383457 (DECREASE stability)

2. Prediction of the sign (direction) of energy change using SVM and sequence information only.

Method 1: Support Vector Machine, use sequence information only.
Effect: DECREASE the stability of protein structure.
Confidence Score: -0.7518157

Method 2: Neural Network, use sequence information only.
Effect: DECREASE the stability of protein structure.

I-Mutant provides Graphs only



ELASPIC: Energy scores only

Protein	Mutation	Type	Template	Seq iden	Align score	Model	$\Delta G_{wt}^{(Fold)}$	$\Delta G_{mut}^{(Fold)}$	$\Delta \Delta G$
Q9BUL8	E94A	Core	3I8i_B	1.000	1.000	-1.399	8.02506	8.26325	0.198

MuPro provides scores for stability change

1. Predicted both value and sign of energy change using SVM and sequence information only.

Detail delta G = -0.64383457 (DECREASE stability)

2. Prediction of the sign (direction) of energy change using SVM and sequence information only.

Method 1: Support Vector Machine, use sequence information only.
Effect: DECREASE the stability of protein structure.
Confidence Score: -0.7518157

Method 2: Neural Network, use sequence information only.
Effect: DECREASE the stability of protein structure.

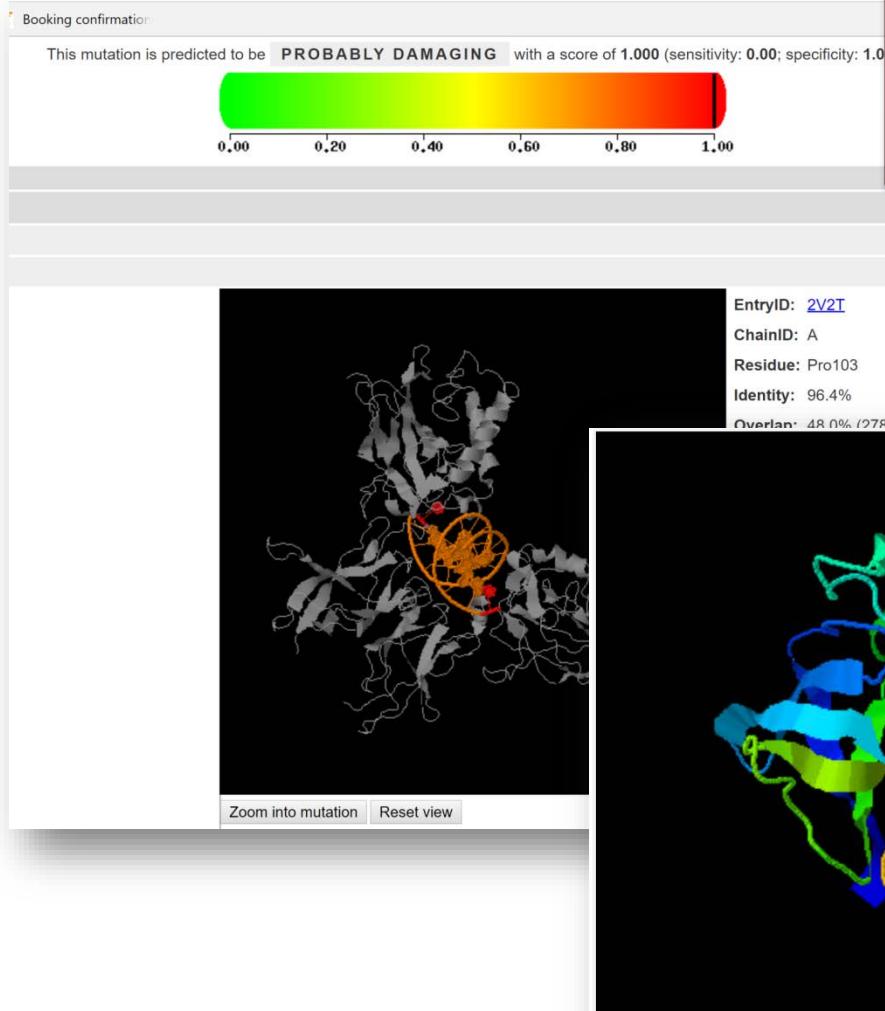
SNPS&GO provides a classification Disease vs Neutral



JSmol

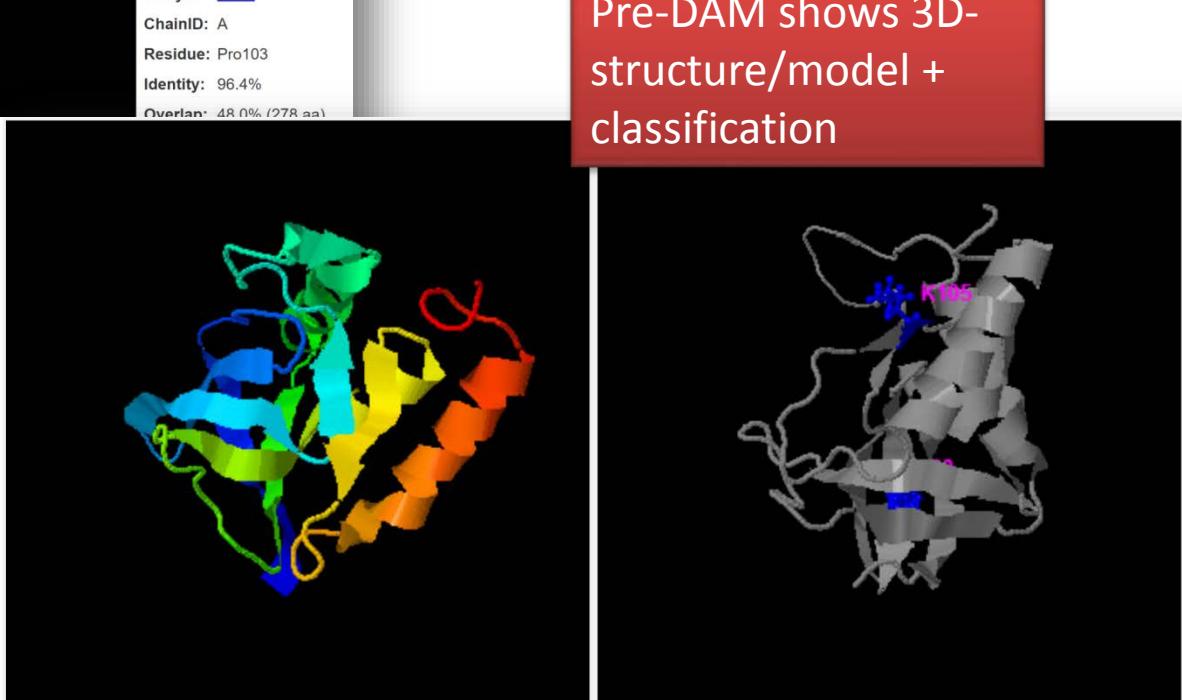
Mutation	Prediction	RI	Probability	Method
C26Q	Disease	8	0.887	PANTHER: F[C]=80% F[Q]=0% S3Ds&G
	Disease	9	0.960	
C83R	Disease	10	0.977	PANTHER: F[C]=94% F[R]=0% S3Ds&G
	Disease	9	0.970	
S124L	Neutral	8	0.106	PANTHER: F[S]=10% F[L]=15% S3Ds&G
	Neutral	5	0.231	

Choices: The Results



PolyPhen provides a classification + sometimes a 3D visualisation

Pre-DAM shows 3D-structure/model + classification



Position	Wild-type	mutant type	prediction
20	L	A	neutral
105	K	M	disease

Choices: The Results: More information = better understanding

UniProt Entry: P12883

Mutant	Structures	Effects	JSON
Ser 242 -> Glu	6	Binding (2) HBonds (2) Impact (6) Interface (2)	[JSON]

Hover over the 'Effects' names for an explanation. Number of residues involved in each effect.

Native residue was involved in a specific HBond or packing interaction with another protein chain or ligand.

SAAP provides structural features and explanation

Site prediction results in protein gnl|sprot|RELB_HUMAN Transcription factor RelB

MutPred provides a score but also a hypothesis based on features

Mutation	Probability of deleterious mutation	MOLECULAR MECHANISM		DISRUPTED		Top 5 features	Exact PTM Match
		Actionable Hypotheses	Confident Hypotheses	Very Confident Hypotheses	Very Confident Hypotheses		
P125W	0.530	Loss of disorder (P = 0.0269) Loss of glycosylation at P125 (P = 0.0415)				Loss of disorder (P = 0.0269) Loss of glycosylation at P125 (P = 0.0415) Gain of MoRF binding (P = 0.057) Loss of phosphorylation at T130 (P = 0.1205) Loss of loop (P = 0.2897)	

Mutation Assesor provides functional impact scores and a few features

Read [how it works](#) document for more details.

Protein	AA	Score	Impact	Pos.	Incl. Impact	FI score	Uniprot	Refseq	MSA height	Codon start position	Func. region	Protein bind.site	DNA/RNA bind.site	small.mol bind.site
1 EGFR_HUMAN G719S	G719S	EGFR	msa	pdb	medium	3.245	EGFR_HUMAN	NP_005219	1416	isoforms	hg19:chr7:55241707		1	STU W3
2 EGFR_HUMAN G724S	G724S	EGFR	msa	pdb	medium	2.245	EGFR_HUMAN	NP_005219	1416	isoforms	hg19:chr7:55241722			ANP 111
3 EGFR_HUMAN E734K	E734K	EGFR	msa	pdb	low	0.97	EGFR_HUMAN	NP_005219	1416	isoforms	hg19:chr7:55242430			
4 EGFR_HUMAN L747F	L747F	EGFR	msa	pdb	low	1.655	EGFR_HUMAN	NP_005219	1416	isoforms	hg19:chr7:55242469		1	W2R
5 EGFR_HUMAN R748P	R748P	EGFR	msa	pdb	low	1.255	EGFR_HUMAN	NP_005219	1416	isoforms	hg19:chr7:55242472		1	VX6 STU
6 EGFR_HUMAN Q787R	Q787R	EGFR	msa	pdb	neutral	0.355	EGFR_HUMAN	NP_005219	1416	isoforms	hg19:chr7:55249061			

CRAVAT provides and extensive output file on both protein/gene level, also nice summary graphs

16 15

Job Info: rkim_20170731_225423

Filter: CHASM p-value: 1, VEST p-value: 2, Allele frequency: 1, Hide synonymous checked

Columns: Variant Info, Structure, Variant Impact, CHASM, VEST, Disease Association, Population Stats, Study, Mutation, NDEX

Summary: Job Info, Variant, Noncoding, Error

Variant Info: UID: TR10:1_NA00001, Gene: KRAS

Pathogenicity impact: VEST, CHASM

Cancer driver impact: CHASM

Disease Association: Driver gene class: Oncogene, TARGET: Cetuximab, PubMed gene hits: None, ClinVar Clin. Sig.: None, ClinVar Diseases: None, ClinVar Dis. Ref.: None, CGC Role: Oncogene, CGC Inheritance: somatic, CGC Somatic Tumors: pancreatic, colo..., CGC Germline Tumors: None, COSMIC variant hits: None

Population Stats: 1000 Genomes, ESP6500, ExAC

MuPIT: 9

Protein Diagram: Ras, My Mutations, TCGA Mutations: All tissues

Sequence Ontologies

Mutation Summary: 17 mutations, 1 Coding, 16 Non-Coding

Cancer Genome Landscapes: 1 Oncogenes, 8 Other genes, Vogelstein

Job Info: rkim_20170731_225423, Input File: manual_input.txt, Date Run: 2017-08-01, Processing Time: 00:01:45, Analyses: CHASM, VEST, CHASM Tissue: Other, Number Input lines: 17, Number Errors: 0, Number Variants: 17, Number Genes: 4, Number Samples: 3, Number Non-Coding: 0

Summary: Job Info, Gene, Variant, Noncoding, Error

Chrom	Position	Strand	Ref. base(s)	Alt. base(s)	Sample ID	HUGO symbol	Sequence ontology	Protein sequence change	CHASM cancer driver p-value (missense)	VEST pathogenicity p-value (non-silent)	Occurrences in COSMIC	PubMed articles	Mappability Warning
chr12	25245321	+	GTA	-	NA00...	KRAS	Inframe del	IQ21K		0	0	0	
chr12	25245321	+	GTA	-	NA00...	KRAS	Inframe del	IQ21K		0	0	0	
chr12	25245321	+	GA	-	NA00...	KRAS	Frameshift del	IQ21IANSE...		0.0093	0	0	
chr12	25245321	+	GA	-	NA00...	KRAS	Frameshift del	IQ21IANSE...		0.0093	0	0	
chr12	25245322	TA	-	TA	NA00...	KRAS	Frameshift del	I21TANSES...		0.0093	0	0	

17 out of 17 variants, Export, Show Full Detail

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

HOPE, an automatic web server that is easy to use
for (bio)medical scientists & Provides structural
explanation of point mutations

Not a classifier, focus on explanation

The screenshot shows the HOPE web interface. At the top, there's a purple header bar with links for 'HOPE', 'Create report', 'Support', 'About', and 'API'. Below the header, a large white box contains the text 'Welcome to HOPE!'. It describes HOPE as an easy-to-use web service that analyses the structural effect of a point mutation in a protein sequence. A 'TRY HOPE' button is visible at the bottom left of this box.

Data collection &
combination is
hidden for user

The screenshot shows the HOPE web interface. At the top, there's a purple header bar with links for 'HOPE', 'Create report', 'Support', 'About', and 'API'. Below the header, a white box contains the text 'Method'. It explains that HOPE can build a model of a protein of interest based on a homologous structure, using Yasara & WHAT IF Twinset. Structural information was collected from WHAT IF Web services, the UniProt database, and DAS-servers. As a modelling template, HOPE identified PDB: 2F1Z. More information about the protein of interest can be found in UniProt entry: UBP48_HUMAN. A 'method page' link is provided for more information. To the right, there's a detailed ribbon presentation of a protein structure, colored by element (alpha-helix in blue, beta-strand in red, turn in green, 3/10 helix in yellow, random coil in cyan). Other molecules in the complex are shown in grey. Below the ribbon, there's a section titled 'Amino Acids' which discusses the mutation of a methionine into a valine at position 415, showing the chemical structures of the original and mutant residues.

Method

The exact 3D-structure of your protein of interest is unknown. However, HOPE is able to build a model of your protein of interest based on a homologous structure. The model will be built using the Yasara & WHAT IF Twinset. Structural information was collected using information from WHAT IF Web services, the UniProt database, and a series of DAS-servers. As a possible modelling template, HOPE identified PDB: 2F1Z. More information about your protein of interest can be found in UniProt entry: UBP48_HUMAN. See the [method page](#) for more information.

Amino Acids

You are interested in the mutation of a methionine into a valine at position 415. The figure below shows the schematic structures of the original (left) and the mutant backbone, which is the same for each amino acid, is colored red. The side chain, u colored black.

CC(C(=O)O)N Mutates into CC(N)O

The figure below shows the schematic structures of the original (left) and the mutant backbone, which is the same for each amino acid, is colored red. The side chain, u colored black.

Overview of the protein in ribbon-presentation. The protein is coloured by element; α -helix=blue, β -strand = red, turn=green, 3/10 helix=yellow and random coil=cyan. Other molecules in the complex are coloured grey when present.

Movies

URL: www.cmbi.ru.nl/hope

Method	Result
Grantham-score	101
PolyPhen-2	Probably damaging
SIFT	Not tolerated
SNP&GO	Disease 10
SNAP	Non-neutral 93%
SNPs3D	-2,35
MutPred	0.918
nsSNPAnalyzer	Disease
Panther	-6,95
PHD-SNP	Disease 9

HOPE report R287W

Contacts

The wildtype residue forms a hydrogen bond with the Leucine on position 271 and with the Tyrosine on position 237. The size difference between wild-type and mutant residue makes that the new residue is not in the correct position to make the same hydrogenbond as the original wild-type residue did. The difference in hydrophobicity will affect hydrogenbond formation. The difference in charge will disturb the ionic interaction made by the original, wild-type residue.

Conservation

Only this residue type was found at this position. Mutation of a 100% conserved residue is usually damaging for the protein.

Domains

The mutated residue is buried in a domain that is important for binding of other molecules. The differences between the wild-type and mutant residue might disturb the core structure of this domain and thereby affect the binding properties.

Amino acid properties

There is a difference in charge between the wild-type and mutant amino acid. The charge of the buried wild-type residue is lost by this mutation. The wild-type and mutant amino acids differ in size. The mutant residue is bigger than the wild-type residue. The wild-type residue was buried in the core of the protein. The mutant residue is bigger and probably will not fit.


HOPE provides
understanding and insight
in the molecular effect of
your mutation

Automatic approaches seem necessary

Nature Genetics Editorial – Jan 2012

“We are more likely to send to review a more complete analysis with strong mechanistic insight explaining how variants in a gene result in altered phenotypes.”

(See: *Nature Genetics* volume44, page1 (2012))

American College of Medical Genetics & Association of Medical Pathologists (2015)

Guideline for variant classification

“If all of the in silico programs tested agree on the prediction, then this evidence can be counted as supporting. If in silico predictions disagree, however, then this evidence should not be used in classifying a variant.”

How to make sense out of this?

Method	Result
Grantham-score	89
PolyPhen-2	Probably damaging
SIFT	Not tolerated
SNP&GO	Disease 10
SNAP	Non-neutral 70%
SNPs3D	-1.43
MutPred	0.82
nsSNPAnalyzer	Neutral
Panther	-2.93
PHD-SNP	Disease 2

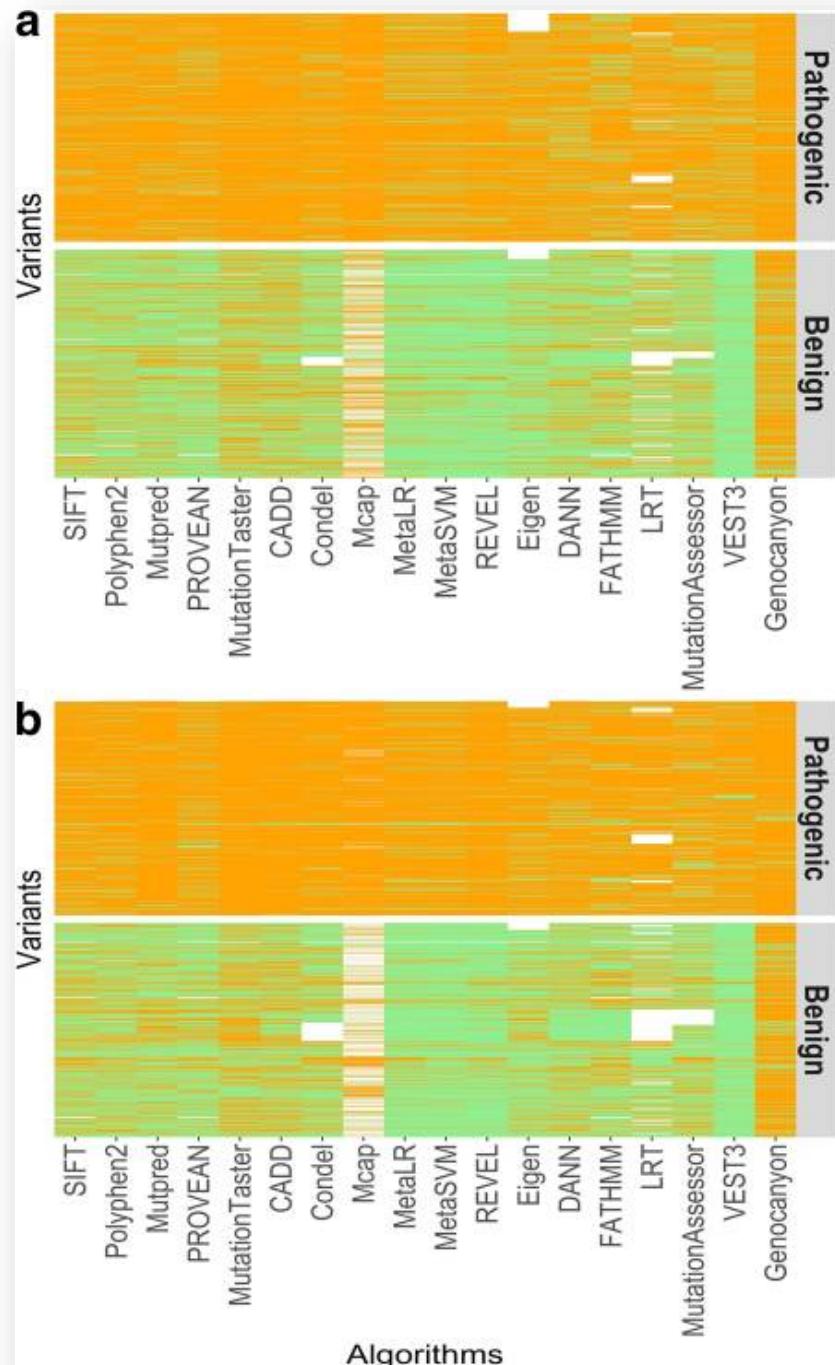
Evaluation of *in silico* algorithms for use with ACMG/AMP clinical variant interpretation guidelines

R. Ghosh *et al.*

~15.000 variations from ClinVar tested on 18 different tools. Very low concordance of ~40% for pathogenic variants and ~3% for benign variants.

Predictions from five commonly used algorithms (Polyphen, SIFT, CADD, PROVEAN and MutationTaster) resulted in higher, 79% for pathogenic variants and 33% for benign variants

No combination of tools led to 100% correctly predicted variants



One last note: Meta-tools are the new thing...

REVEL: Rare Exome Variant Ensemble Learner

[About](#) [Downloads](#) [Contact](#)

SNP Function Prediction (FuncI)

Query by :

Chromosome Position

Chromosome:

1

Start position(bp):

End position(bp):

SNP Functional Predictions:

- nsSNP
- Splicing Regulation
- Stop Codon
- Polyphen Prediction
- SNPs3D Prediction
- TFBS Prediction
- miRNA Binding Site Prediction
- Regulatory Potential Score
- Conservation Score
- Nearby Genes

About REVEL

REVEL is an ensemble method for predicting the pathogenicity of missense variants based on a combination of individual tools: MutPred, FATHMM v2.3, VEST 3.0, PolyPhen-2, SIFT, PROVEAN, MutationAssessor, MutationTaster, SiPhy, phyloP, and phastCons. REVEL was trained using recently discovered pathogenic and rare neutral missense variants, excluding those previously used to train its constituent tools. The REVEL score for an individual missense variant ranges from 0 to 1, with higher scores reflecting greater likelihood that the variant is disease-causing. When applied to two independent datasets, REVEL had the best overall performance ($p < 10^{-12}$) compared with any individual tool and seven ensemble methods: MetaLR, KGGSeq, Condel, CADD, DANN, and Eigen. Importantly, REVEL also had the best performance for distinguishing pathogenic from rare neutral variants with allele frequencies $< 0.5\%$. Compared with other ensemble methods, the area under the receiver operating characteristic curve (AUC) for REVEL was 0.046–0.182 higher in an independent test set of 1953 disease variants and 123,935 putatively neutral exome sequencing variants, and 0.027–0.143 higher in an independent test set of 1953 pathogenic and 2406 benign variants recently reported in ClinVar. We provide precomputed REVEL scores for all human missense variants to facilitate the identification of pathogenic variants in the sea of rare variants discovered in large-scale studies.



PREDICTSNP¹

Consensus classifier for prediction of disease related amino acid mutations

Meta-**SNP**



The solution: Consensus predictors?

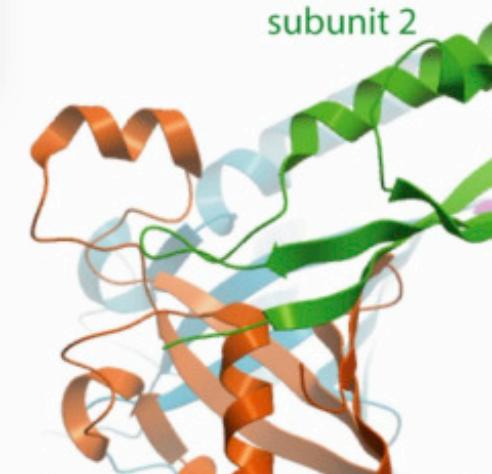
From Gene Volume 680, 5 January 2019,

A review study: Computational techniques for expecting the impact of non-synonymous single nucleotide variants in human diseases

M.Hassan et al.

Computational tools that are available on the website as web servers can provide better predictions of the effect of SNVs if used combined rather than alone.

→ Still for full mechanistic understanding, a 3D analysis is necessary



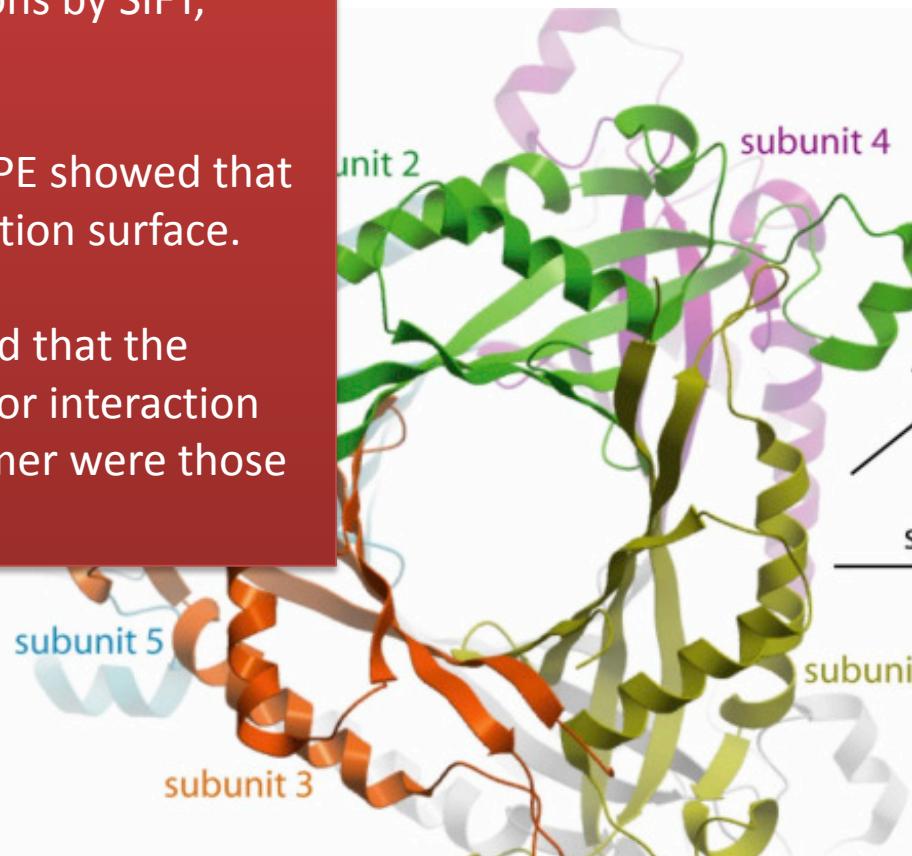
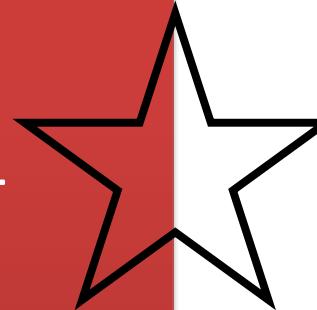
Role of protein structure in variant annotation: structural insight of mutations causing 6-pyruvoyl- tetrahydropterin synthase deficiency

J. Muniz et al.

Comparison of predictions for 95 ClinVar variations by SIFT,
PolyPhen and HOPE. Only 51% concordance.

20 Variants were falsely classified as benign, HOPE showed that
these mutations occurred on the putative interaction surface.

Information from the protein structure suggested that the
variants affecting amino acid residues required for interaction
between monomeric subunits of the PTPS hexamer were those
misclassified as benign by *in silico* algorithms.



Choices: The Results

Some conclusions about the way VEP tools present their results

- Wide variety in ways to present the results (of course related to the type of data that is used to create the results)
- Output often in terms of Disease/Non-disease, Damaging / Benign, Not-tolerated/Tolerated, scale 10-0, scale 0-100, Non-neutral / Neutral, etc.
→ Good for classification, but not for usable for deeper understanding
- HOPE is the only server that provides insight instead of scores
- Although a few servers nowadays do use and sometimes explain some of the changes in the structure.

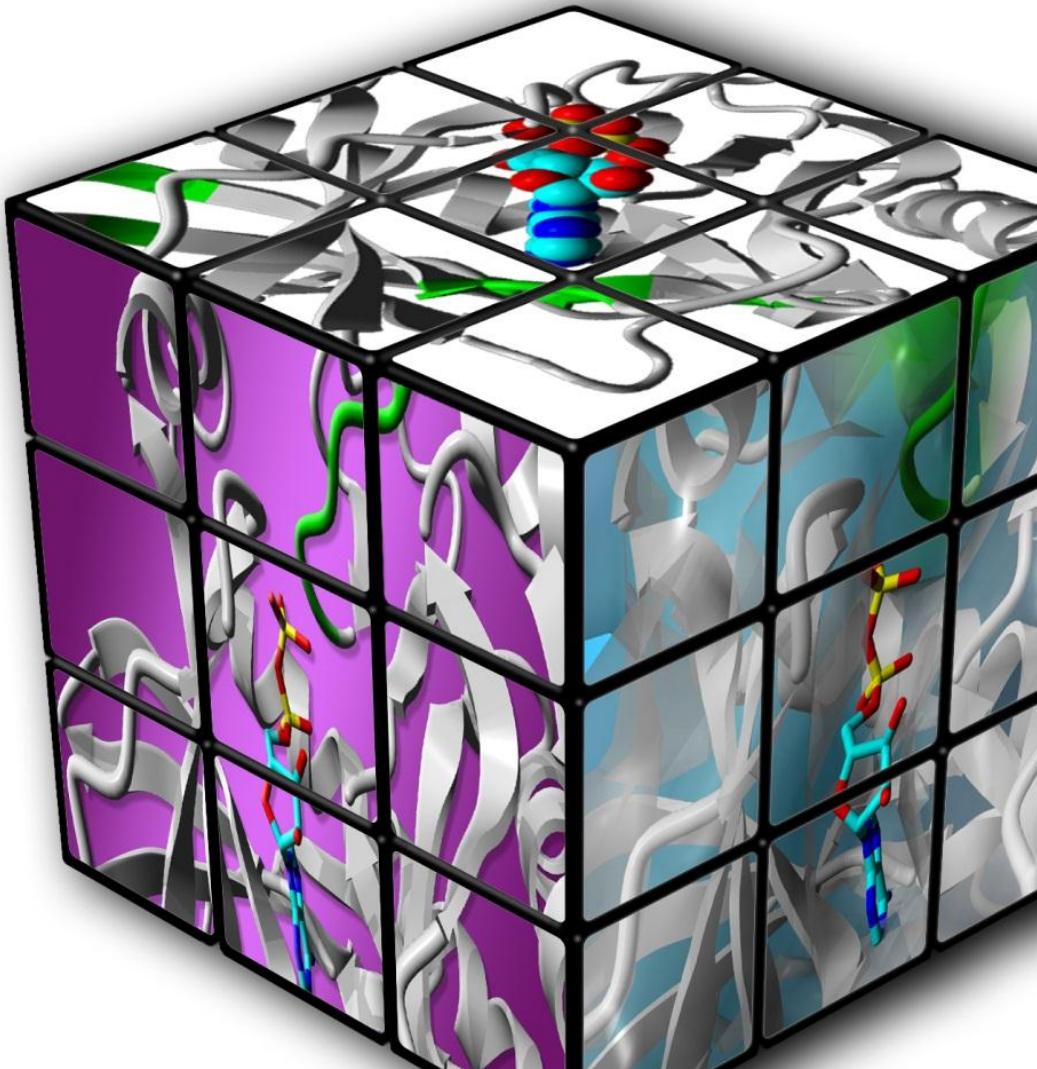
My suggestion for your analysis: Choose a tool that suits your data and needs, use a tool like HOPE to gain more insight in the one mutation of interest

Thank you...

Hanka Venselaar

CMBI / RadboudUMC

**International Post-graduate Course
Variant effect prediction
April 12th 2018
Hanka.Venselaar@radboudumc.nl**



Interesting literature

[Genome Med.](#) 2017 Dec 18;9(1):113. doi: 10.1186/s13073-017-0509-y.

Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: a proposed framework.

[Brief Bioinform.](#) 2014 Mar; 15(2): 256–278.

Published online 2013 Jan 21. doi: [10.1093/bib/bbs086](https://doi.org/10.1093/bib/bbs086)

PMCID: PMC3956068 , PMID: [23341494](#)

A survey of tools for variant analysis of next-generation genome sequencing data

[Genome Med.](#) 2014; 6(10): 87.

Published online 2014 Oct 22. doi: [10.1186/s13073-014-0087-1](https://doi.org/10.1186/s13073-014-0087-1)

PMCID: PMC4254438, PMID: [25473426](#)

Computational approaches to interpreting genomic sequence variation

[J Mol Biol.](#) 2013 Nov 1; 425(21): 4047–4063.

Published online 2013 Aug 17. doi: [10.1016/j.jmb.2013.08.008](https://doi.org/10.1016/j.jmb.2013.08.008)

PMCID: PMC3807015, PMID: [23962656](#)

Towards Precision Medicine: Advances in Computational Approaches for the Analysis of Human Variants

[Brief Bioinform.](#) 2013 Jul;14(4):448-59. doi: 10.1093/bib/bbt013. Epub 2013 Mar 15.

Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools.

[Hum Mutat.](#) 2016 Jun;37(6):579-97. doi: 10.1002/humu.22987. Epub 2016 Apr 15.

Variation Interpretation Predictors: Principles, Types, Performance, and Choice.

- HOPE
- PolyPhen
- AutoMute
- CupSat
- DUET
- ELASPIC
- I-Mutant
- INPS-3D
- KD4v
- PreDAM
- SAPPpred
- SNPeffect
- SusPect
- MutationTaster
- MutPred
- nsSNPAnalyzer
- Panther-PSEP
- PdD-SNP
- PMUT
- PONP-2
- SIFT/PROVEAN
- SNAP / SNAP 2
- SNP&GO
- FatHMM
- NetDiseaseSNP
- VARMOD
- stSNP

Servers that predict **mutation effects on protein level**. (all work with a sequence or sequence ID and mutation)

Meta predictors

- CONDEL
- META-SNP
- REVEL
- PredictSNP
- FuncPred

Protein stability change /energy change predictors

- CUPSAT
- DUET
- I-Mutant
- INPS-3D
- mCSM
- MuPro
- SAAFEC
- SDM
- Auto-MUTE

- Binding Affinity Predictors**
- BeatMusic
 - BindProfX
 - mCMS-AB / mCSM-lig
 - MutaBind
 - SAAMBE
 - StructMan

Lists of servers that are freely available online without registration. These lists do not pretend to be complete!

Servers that predict **mutation effect on genomic level**

- MutationTaster
- VEP
- PhenGen
- TransFic
- CRAVAT

Protein Aggregation Predictors

- Aggrescan
- Fold Amyloid
- PASTA2
- ProA