



JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

Protein Structure Prediction

Miguel Andrade

Faculty of Biology,

Institute of Organismic and Molecular Evolution

Johannes Gutenberg University

Mainz, Germany

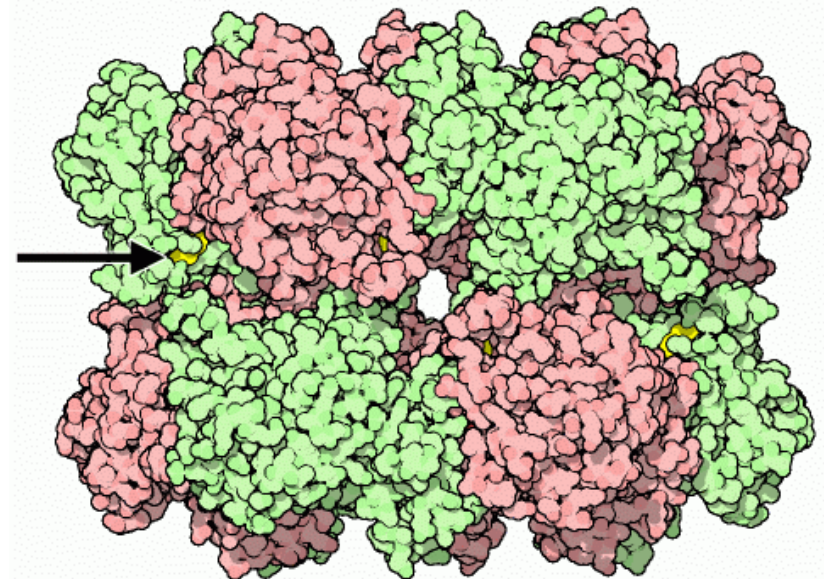
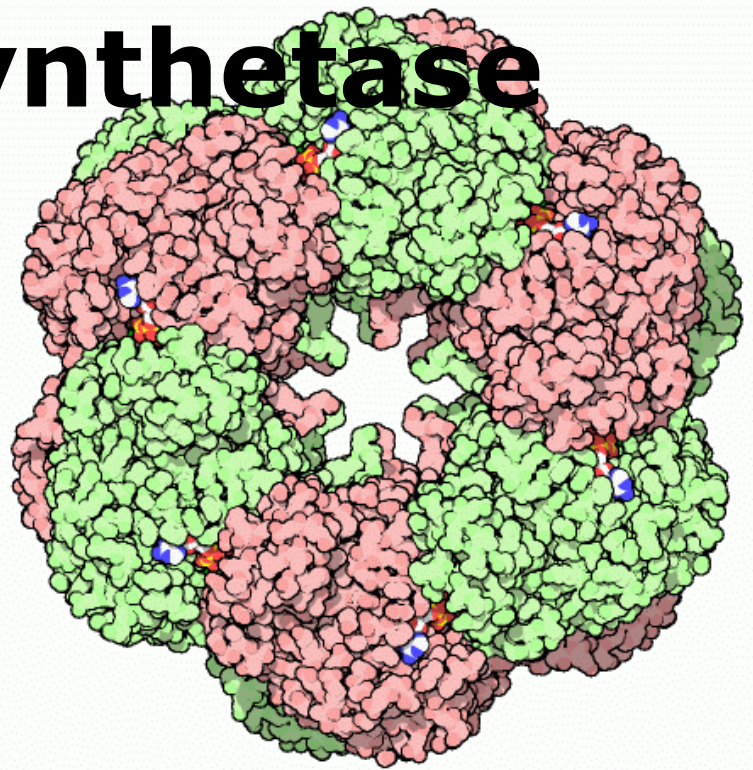
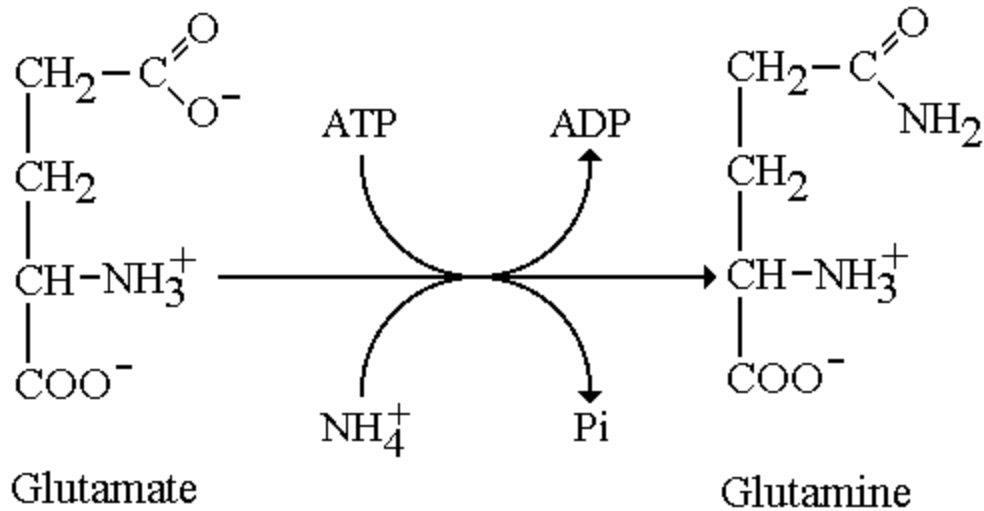
andrade@uni-mainz.de

Mount Everest

Age: 60M years



Glutamine synthetase

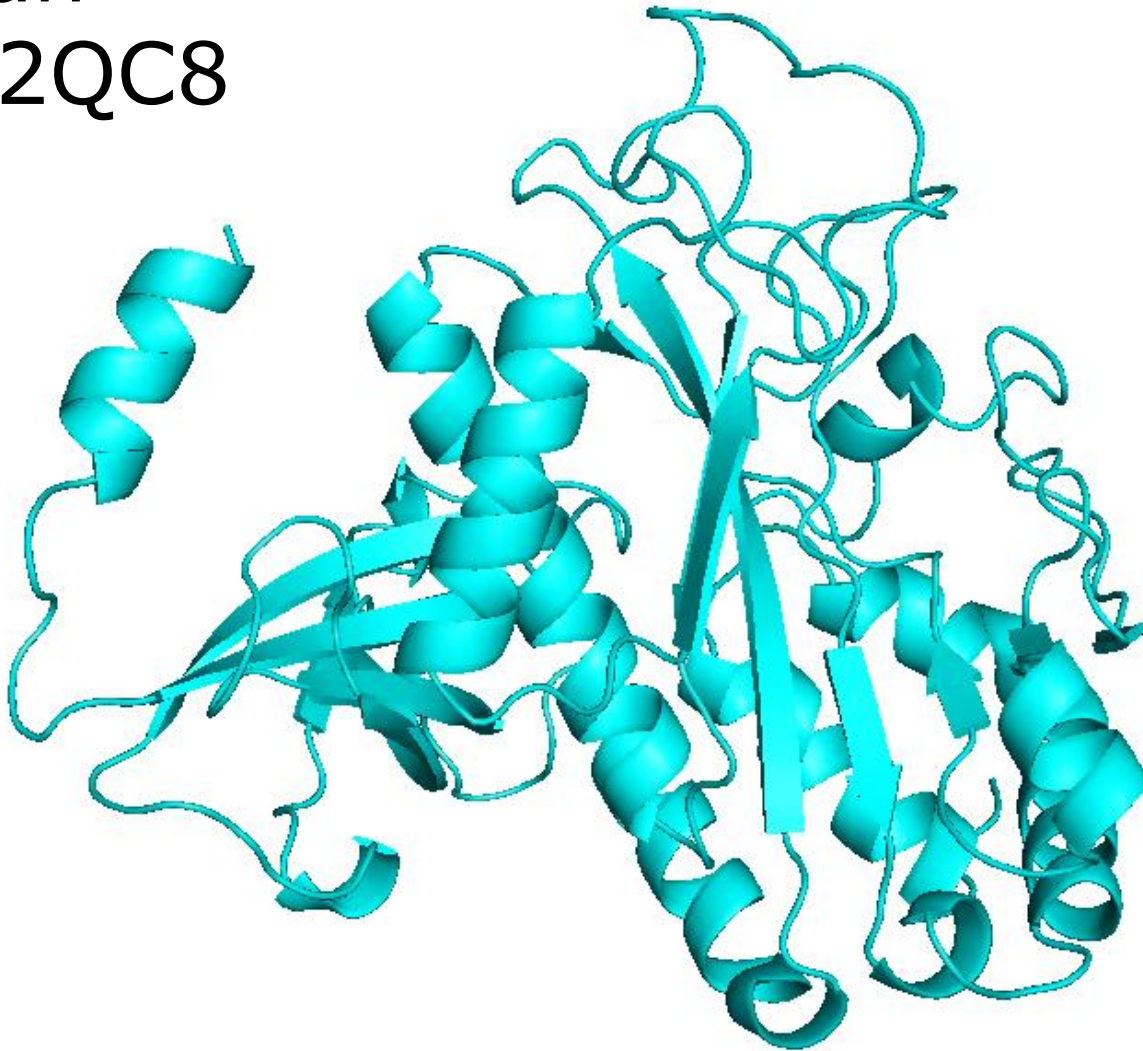


Age: +3500M years

Glutamine synthetase

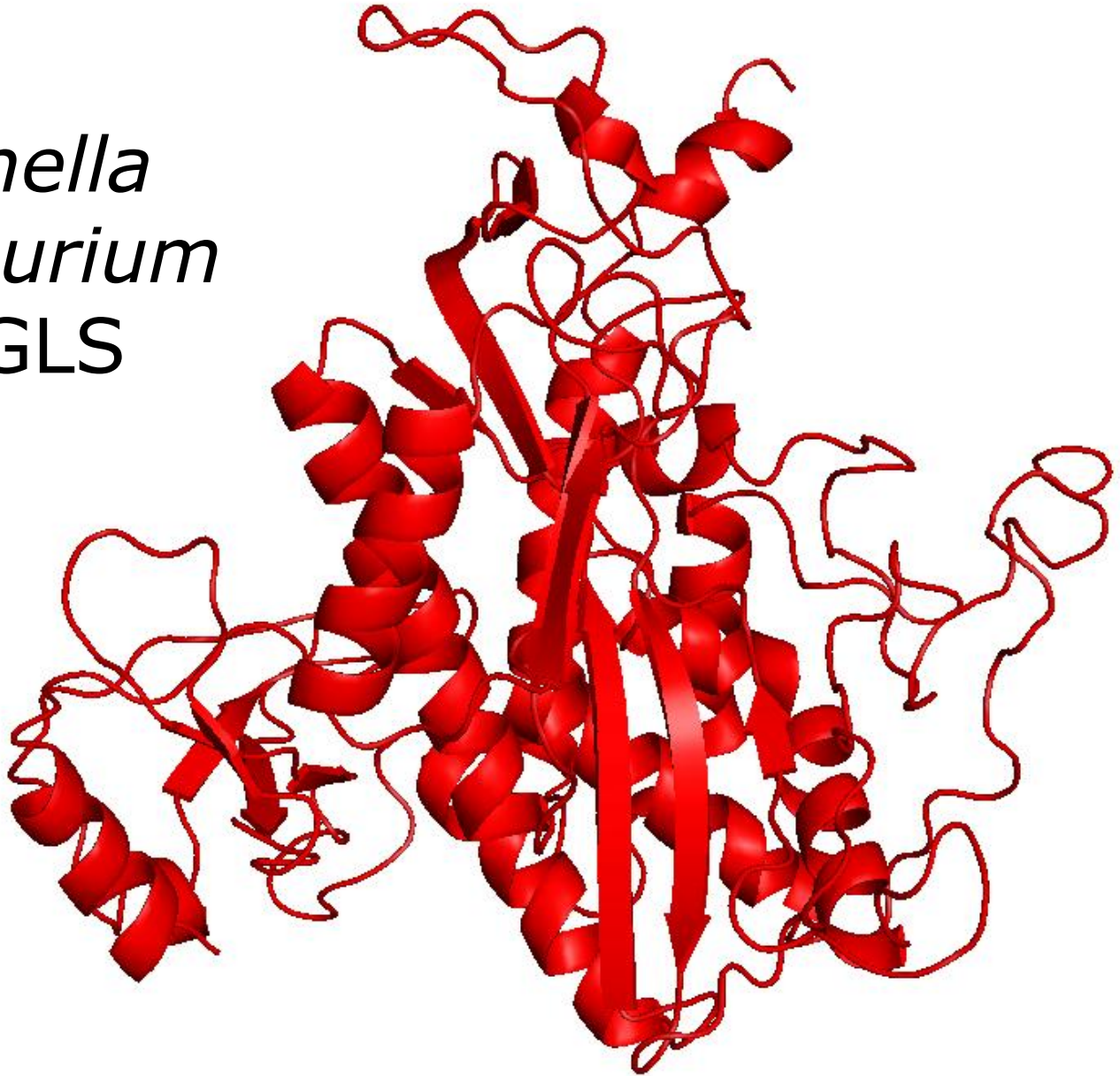
Human

PDB:2QC8

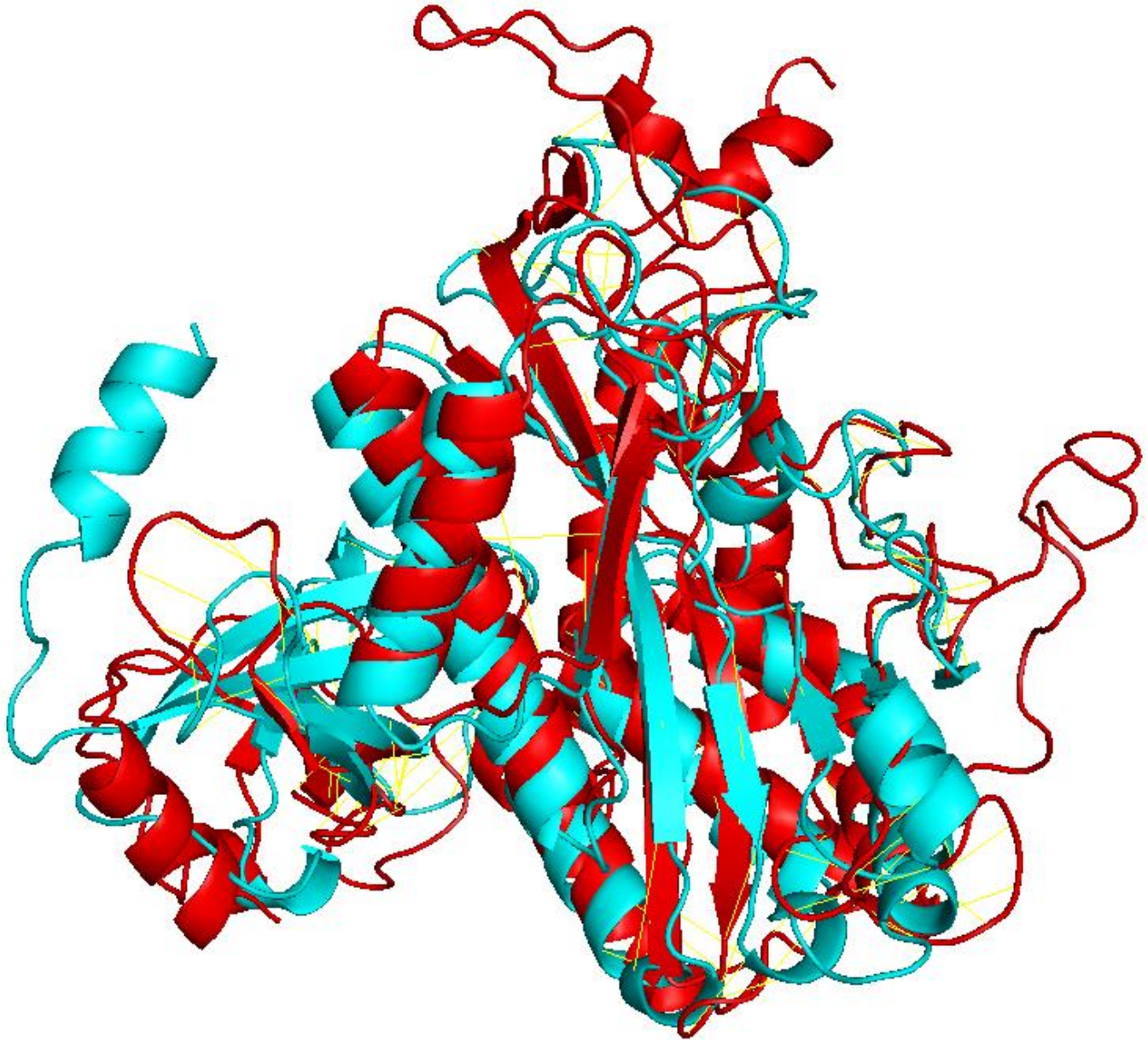


Glutamine synthetase

Salmonella
typhimurium
PDB:2GLS



Glutamine synthetase



Time line

Earth: 4.6 By

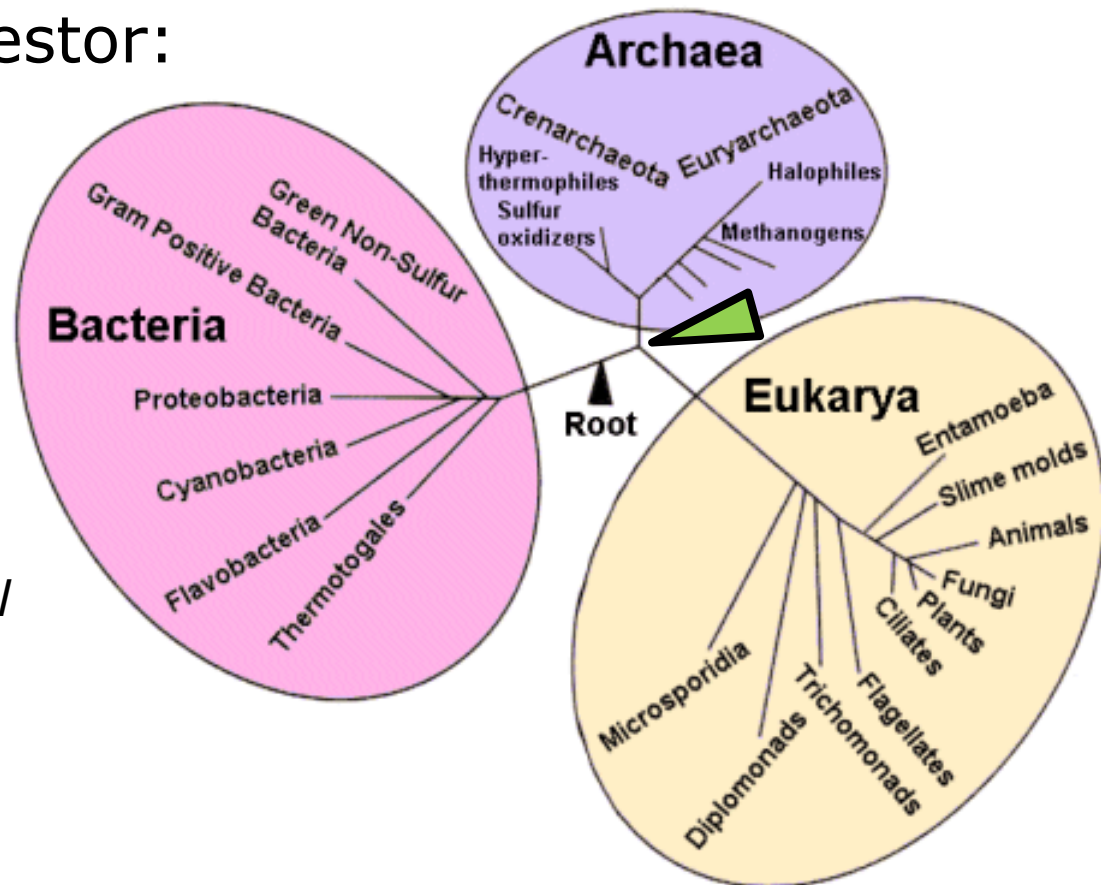
Origin of life: 3.9 By – 3.5 By

Last Common Ancestor:
3.5 – 3.8 By

Glansdorff & Labedan
(2008) *Biology Direct*

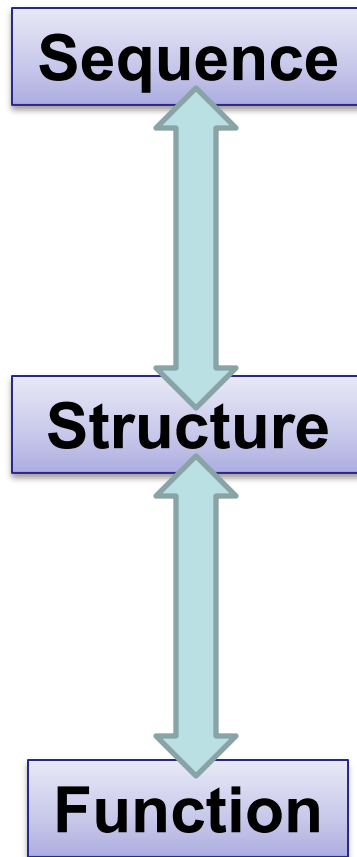
4.29 By

Sheridan *et al.* (2003)
Geomicrobiology Journal



Sequence and function

Evolutionary constraints



MTQDELKKAVGWAALQYVQ

P
L
E

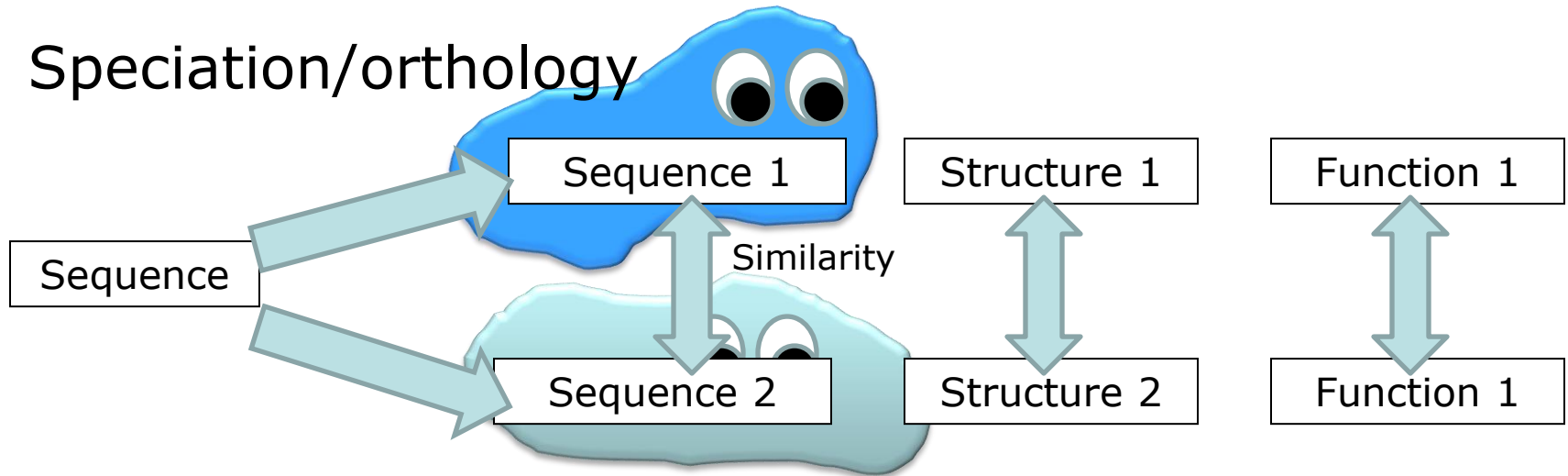
A
T



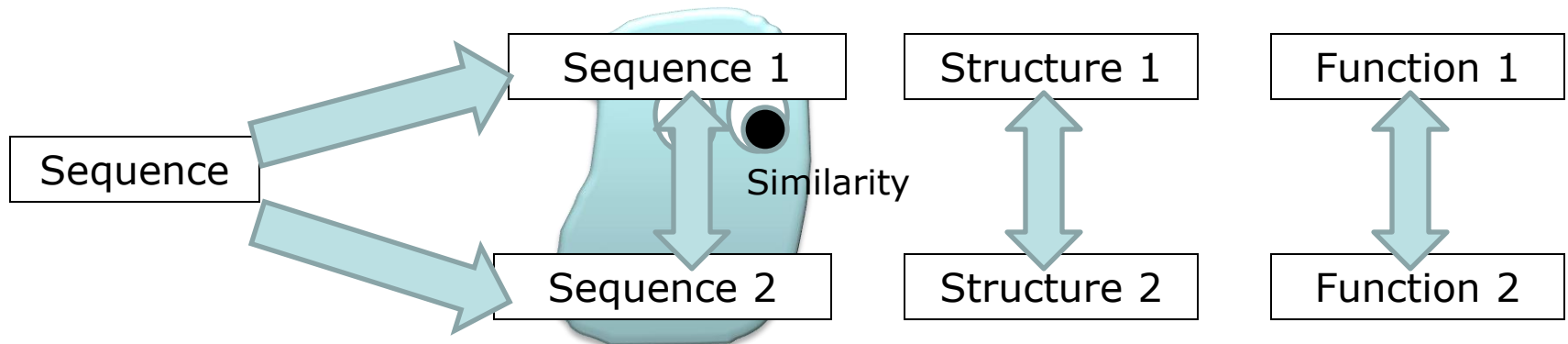
Sequence and function

Evolutionary constraints

Speciation/orthology



Gene duplication/paralogy



Sequence pairwise alignment

```
>gs_human gi|74271837|ref|NP_001028216.1| glutamine synthetase [Homo sapiens]  
MTTSASSHLNKGIKQVYMSLPQGEKVQAMYIWIDGTGEGLRCKTRTLDSEPKCVEELPEWNFDGSSTLQS  
EGSNSDMYLVPAAMFRDPFRKDPNKLVLCEVFKYNRRPAETNLRHTCKRIMDMVSNQHPWFGMEQEYTLM  
GTDGHPFGWPSNGFPGPQGPYYCGVGADRAYGRDIVEAHYRACLYAGVKIAGTNAEVMMPAQWEFQIGPCE  
GISMGDHLWVARFILHRVCEDFGVIAITFDPKPIPGNWNAGCHTNFSTKAMREENGLKYIEEAIEKLSKR  
HQYHIRAYDPKGGLDNARRLTGFHETSNINDFSAGVANRSASIRIPRTVGQEKKGYFEDRRPSANCDPFS  
VTEALIRTCLLNETGDEPFQYKN
```

```
>gs_salmonella gi|16767272|ref|NP_462887.1| glutamine synthetase [Salmonella  
enterica subsp. enterica serovar Typhimurium str. LT2]  
MSAEHVLTMLNEHEVKFVDLRFTDTKGKEQHVTIPAHQVNAEFFEEGKMFDGSSIGGWKGINESDMVLMP  
DASTAVIDPFFADSTLIIRCDILEPGTLQGYDRDPRSIakraEDYLRATGIADTVLFGPEPEFFLFDDIR  
FGASISGSHVAIDDIEGAWNSSTKYEGGNKGHRPGVKGGYFPVPPVDSAQDIRSEMCLVMEQMGLVVEAH  
HHEVATAGQNEVATRFNTMTKKADEIQIYKYVVHNVAHFRFGKTATFMPKPMFGDNGSGMHCHMSLAKNGT  
NLFSGDKYAGLSEQALYYIGGVIKHAKAINALANPTTNSYKRLVPGYEAPVMLAYSARNRSASIRIPVVA  
SPKARRIEVRFPDPAANPYLCFAALLMAGLDGIKNKIHPGEAMDKNLYDLPPEEAKEIPQVAGSLEEALN  
ALDLDREFLKAGGVFTDEAIDAYIALRREEDDRVRMTPHPVEFELYYSV
```

Sequence pairwise alignment

BLAST (Altschul et al, 1990)

>lcl|39919 unnamed protein product
Length=469

Score = 70.5 bits (171), Expect = 1e-17, Method: Compositional matrix adjust.
Identities = 102/363 (28%), Positives = 138/363 (38%), Gaps = 96/363 (26%)

```
Query 62  FDGSSTLQSEGSN-SDMYLVPAA--MFRDPFRKDPNKLVLCEVFK-----YNRRP---- 108
          FDGSS   +G N SDM L+P A      DPF D   ++ C++ +      Y+R P
Sbjct 50  FDGSSIGGWKGINESDMVLMPDASTAVIDPFFADSTLIIRCDILEPGTLQGYDRDPRSIA 109

Query 109 --AETNLRHTCKRIMDMVSNQHPWFGMEQEYTLMGTDGHPFGWPSNGF----- 154
          AE LR T   I D V      FG E E+ L   D   FG   +G
Sbjct 110 KRAEDYLRTG--IADTV-----LFGPEPEFFLF--DDIRFGASISGSHVAIDIEGAWN 160

Query 155 -----PGPQGPYYCGVGADRAYGRDI-----VEAHYRACLYAG 187
          PG +G Y+      D A   +DI              VEAH+      AG
Sbjct 161 SSTKYEGGNKGHRPGVKGGYFPVPPVDSA--QDIRSEMCLVMEQMGLVVEAHHHEVATAG 218

Query 188 VKIAGTNAEVMPAQWEFQIGPCEGISMGDHLWVARFILHRVCEDFGVIATFDPKPIPG-N 246
          T     M   +              D + + ++++H V   FG   ATF PKP+ G N
Sbjct 219 QNEVATRFNTMTKK-----ADEIQIYKYVVHNVAHREFGKTATFMPKPMFGDN 265

Query 247 WNGAGCHTNFSTKAMREENGLKYIEEAIEKLSKRHQYHIRAYDPKGGLDNA----- 297
          +G CH + +      +G KY      LS++   Y+I              NA
Sbjct 266 GSGMHCHMSLAKNGTNLFSGDKY-----AGLSEQALYYIGGVIKHAKAINALANPTTNSY 320

Query 298 RRLTGFHETSNINDFSAGVANRSASIRIPRTVQGEKKGYFEDRRPSANCDPFSVTEALIR 357
          +RL   +E   +   +SA   NRSASIRIP V   K     E R P   +P+   AL+
Sbjct 321 KRLVPGYEAPVMLAYSAA--RNRSASIRIP-VVASPKARRIEVRFPDPAANPYLCFAALLM 377

Query 358 TCL 360
          L
Sbjct 378 AGL 380
```

Multiple sequence alignment

```
>gs_human gi|74271837|ref|NP_001028216.1| glutamine synthetase [Homo sapiens]
MTTSASSHLNKGIKQVYMSLPQGEKVQAMYIWIWIDGTGEGLRCKTRTLDSEPKCVEELPEWNFDGSSTLQS
EGSNSDMYLVPAAMFRDPFRKDPNKLVLCEVFYKYNRRPAETNLRHTCKRIMDMVSNQHPWFGMEQEYTLM
GTDGHPFGWPSNGFPGPQGPYYCGVGADRAYGRDIVEAHYRACLYAGVKIAGTNAEVMPAQWEFQIGPCE
GISMGDHLWVARFILHRVCEDFGVIAITFDPKPIPGNWNWAGCHTNFSTKAMREENGLKYIEEAIEKLSKR
HQYHIRAYDPKGGLDNARRLTGFHETSNINDFSAGVANRSASIRIPRTVGQEKKGyFEDRRPSANCDPFS
VTEALIRTCLLNETGDEPFQYKN
```

```
>gs_vulca gi|307594850|ref|YP_003901167.1| glutamine synthetase [Vulcanisaeta
distributa DSM 14429]
```

```
MPTRNLEIEPADLWRILKASGIKYVKFIIVDINGAPRSEIVPIDMAKDLFIDGMPFDASSIPSYSTVNKS
DFVAYVDPRAVYVEYWQDGKVADVFTMVSDIADKPSPLDPRRVLNDALEQARSKGYEFLMGVEVEFFVIK
EDGGKPVFADPGIYFDGWNVTVQSQFMKELITAIADAGINYTKTHHEVAPSQYEVNIGATDPLRLADQIV
YFKIMAKDIARKYGLVATFMPKPFWGVNGSGAHTHISVWKDGKNLFQSSTGKITEECGYAISAILSNARA
LSSFVAPLVNSYKRLVPHYEAPTRIVWGYANRSAMIRIPQYKMRINRIEYRHPDPSMNPYLAFTAIKTM
IRGLEEKKEPPPPTEEVAYELANALET PATLEDTLKELSKSFLATELPSELVNAYIKIKQNEWEDYLTNV
GPWEKTWNIITQWEYNKYLVTA
```

```
>gs_salmonella gi|16767272|ref|NP_462887.1| glutamine synthetase [Salmonella
enterica subsp. enterica serovar Typhimurium str. LT2]
```

```
MSAEHVLTMLNEHEVKFVDLRFTDTKGKEQHVTIPAHQVNAEFFEEGKMGFDGSSIGGWKGINESDMVLMP
DASTAVIDPFFADSTLIIRCDILEPGTLQGYDRDPRSIakraEDYLratGIADTVLFGPEPEFFLFDDIR
FGASISGSHVAIDDIEGAwnsSTKYEGGNKGHRPGVKGGYFPVPPVDSAQDIRSEMCLVMEQMGLVVEAH
HHEVATAGQNEVATRFNTMTKKADEIQIYKYVVHNVahRFgKTATfMPKPMFGDNGSGMHCHMSLAKNGT
NLFSGDKYAGLSEQALYYIGGVikHAKAINALANPTTNSYKRLVPGYEAPVMLAYSARNRSASIRIPVVA
SPKARRIEVRFPDPAANPYLCFAALLMAGLDGIKNKIHPGEAMDKNLYDLPPEEAKEIPQVAGSLEEALN
ALDLDREFLkAGGVFTDEAIDAYIALRREEDDRVRMTPHPVEFELYYSV
```

```
>gs_yeast gi|330443748|ref|NP_015360.2| Glnlp [Saccharomyces cerevisiae S288c]
```

```
MAEASIEKTQILQKYLELDQRGRIIAEYVWIDGTGNLRSKGRTLKKRITSIDQLPEWNFDGSSTNQAPGH
DSDIYLKPVAYYPDPFRRGDNIVVLAACYNNDGTPNKFNRHEAAKLFAAHKDEEIWFGLEQEYTLFDMY
DDVYGWPKGGYPAPQGPYYCGVGAGKVYARDMIEAHYRACLYAGLEISGINAEVMPsQWEFQVGPCTGID
MGDQLWMARYFLHRVAEEFGIKISFHPKPLKGDWNGAGCHTNVSTKEMRQPGGMKYIEQAIEKLSKRHAe
HIKLYGSDNDMRLTGRHETASMTAFSSGVANRGSSIRIPRSVAKEGYGYFEDRRPASNIDPYLVTGIMCE
TVCGAIDNADMTKEFERESS
```

Multiple Sequence Alignment

MUSCLE stands for **M**ultiple **S**equence **C**omparison by **L**og-**E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.

Important note: This tool can align up to 500 sequences or a maximum file size of 1 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of sequences in any supported format:

Or upload a file: Keine ausgewählt

[Use a example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

STEP 2 - Set your Parameters

OUTPUT FORMAT:

ClustalW

The default settings will fulfill the needs of most users.

(Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

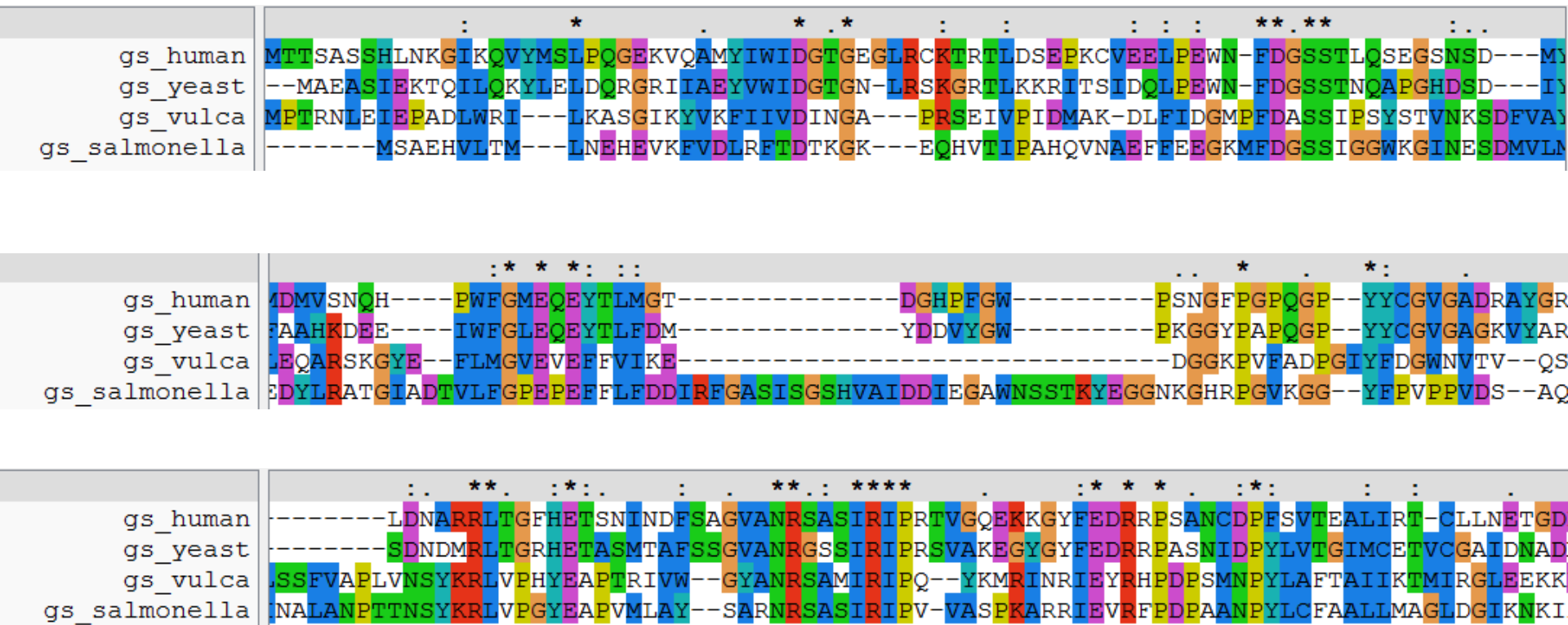
<http://ebi.ac.uk/Tools/msa/muscle/>

>gs_human gi|74271837|ref|NP_001028216.1| glutamine synthetase [Homo sapiens]
MTTSASSHLNKGIKQVYMSLPQGEKVQAMYIWIDGTGEGLRCKTRTLTLDSEPKCVEELPEW
N-FDGSSTLQSEGSNSD---MYLVPAAMFRDPFRKDPNKLVLCEVFKYNRRPA-ETNLRH
TCKRIMDMVSNQH----PWFGEQEYTLTGMT-----DGHFPGW-----
-PSNGFPGPQGP--YYCGVGADRAYGRDIVEAHYRACLYAGVKIAGTNAEVMMPA-QWEFQ
IGPCEGISMGDHLWVARFILHRVCEDFGVIATFDPKPIPGNWNAGCHTNFSTKAMREEN
GLKYIEEAIEKLSKRHQYHIRAYDPKGG-----LDNARRLTGFHETSININDFSAGV
ANRSASIRIPRTVQGEKKGYFEDRRPSANCDPFSVTEALIRT-CLLNETGDEP-----
-----FQYKN-----

>gs_yeast gi|330443748|ref|NP_015360.2| Glnlp [Saccharomyces cerevisiae S288c]
--MAEASIEKTQILQKYLELDQRGRIIAEYVWIDGTGN-LRSKGRTLKKRITSIDQLPEW
N-FDGSSTNQAPGHDS---IYLKPVAYYPDPFRRGDNIVVLAACYNNDGTPN-KFNHRH
EAAKLFAAHKDEE----IWFGLEQEYTLFDM-----YDDVYGW-----
-PKGYPAPQGP--YYCGVGAGKVYARDMIEAHYRACLYAGLEISGINAEVMPS-QWEFQ
VGPCTGIDMGDQLWMARYFLHRVAEEFGIKISFHPKPLKGDWNGAGCHTNVSTKEMRQPG
GMKYIEQAIEKLSKRHAEHKLYG-----SDNDMRLTGRHETASMTAFSSGV
ANRGSSIRIPRSVAKEGYGYFEDRRPASNIDPYLVTGIMCETVCGAIDNADMT-----
-----KEFERESS--

>gs_vulca gi|307594850|ref|YP_003901167.1| glutamine synthetase [Vulcanisaeta distributa DSM 14429]
MPTRNLEIEPADLWRI---LKASGIKYVKFIIVDINGA---PRSEIVPIDMAK-DLFIDG
MPFDASSIPSYSTVNKSDVFAYVDPRAVYVEYWQDGKVADVFTMVSDIADKPS-PLDPRR
VLNDALEQARSKGYE--FLMGVEVEFFVIKE-----
--DGGKPVFADPGIYFDGWNVTV--QSQFMKELITAIADAGINYTKTHHEVAPS-QYEVN
IGATDPLRLADQIVYFKIMAKDIARKYGLVATFMPKPFWGV-NGSGAHTHIS---VWKDG
KNLF-QSSTGKITEECGYAISAILSNNARLSSSFVAPLVNSYKRLVPHYEAPTRIVW--GY
ANRSAMIRIPQ--YKMRINRIEYRHPDPSMNPYLAFTAIKTMIRGLEEKKEPPPTTEEV
AYELA--NALETP---ATLEDTLK--ELSKSFLATE--LPSELVNAYIKIKQNEWEDYLT
NVGPWEKTWNIITQWEYNKYLVTA

>gs_salmonella gi|16767272|ref|NP_462887.1| glutamine synthetase [Salmonella enterica]
-----MSAEHVLTM---LNEHEVKFVDLRFTDTKGK---EQHVTIPAHQVNAEFFEEG
KMFDGSSIGGWKGINESDMVLMPPDASTAVIDPFFADSTLIIRCDILEPGTLQGYDRDPRS
IAKRAEDYLRTATGIADTVLFGPEPEFFLFDDIRFGASISGSHVAIDDIAGAWNSSTKYEG
GNKGHRPGVKGG--YFPVPPVDS--AQDIRSEMCLVMEQMGLVVEAHHHEVATAGQNEVA
TRFNTMTKKADEIQIYKYVVHNVHRFGKTATFMPKPMFGD-NGSGMHCHMS---LAKNG
TNLFSGDKYAGLSEQALYYIGGVIKHAKAINALANPTTNSYKRLVPGYEAPVMLAY--SA
RNRSASIRIPV-VASPKARRIEVRFPDPAANPYLCFAALLMAGLDGIKNKIHPGEAMDKN
LYDLPPEEAKEIPQVAGSLEEALNALDLDFEFLKAGGVFTDEAIDAYIALRREEDDRVRM
TPHP-----VEFELYYSV-



ClustalW, JalView

Determination of protein structure

X-ray crystallography (200K in PDB)

- need crystals

Electron microscopy (30K)

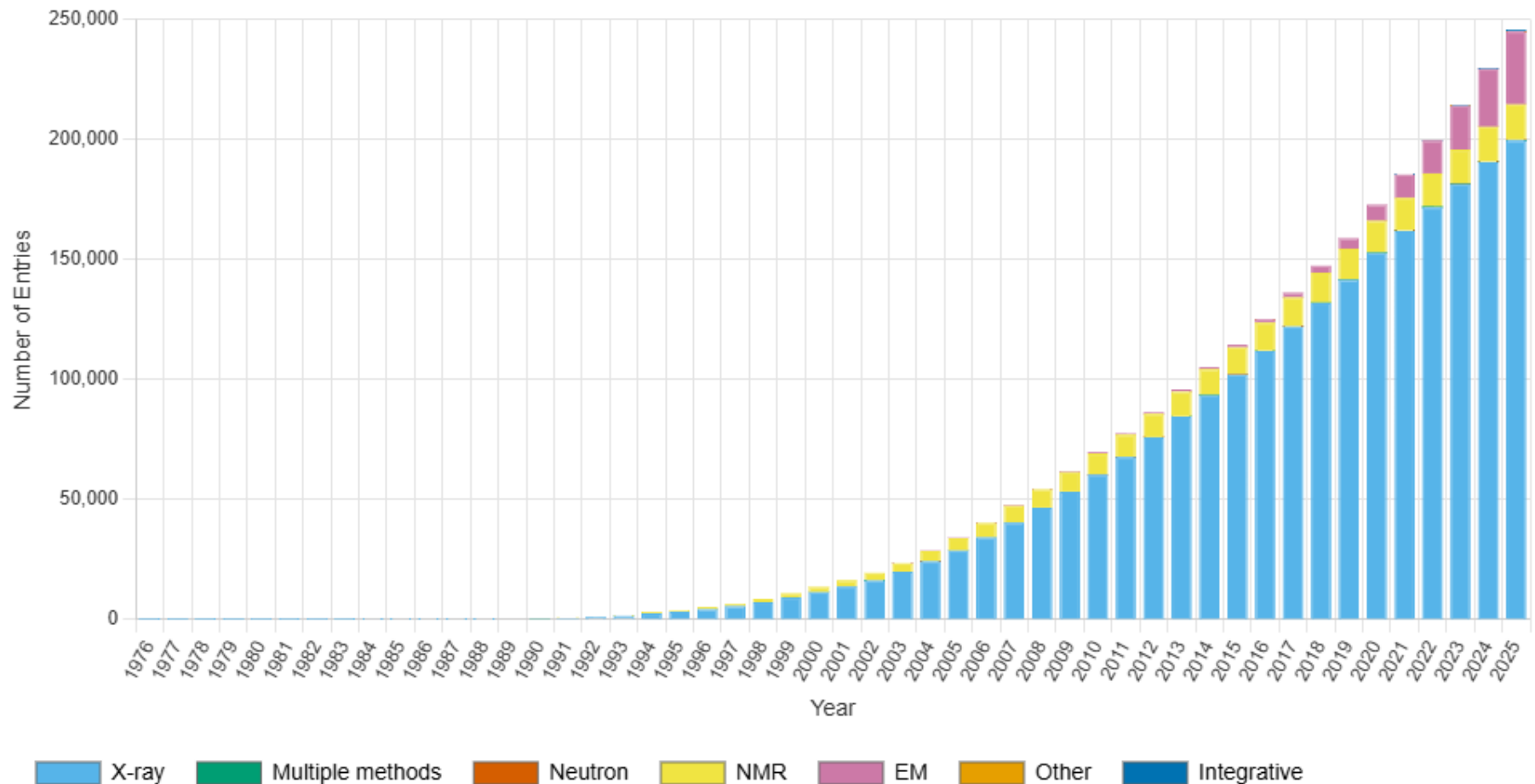
- Was low resolution ($>5\text{\AA}$)

Nuclear Magnetic Resonance (NMR)
(14K)

- proteins in solution
- lower size limit (600 aa)

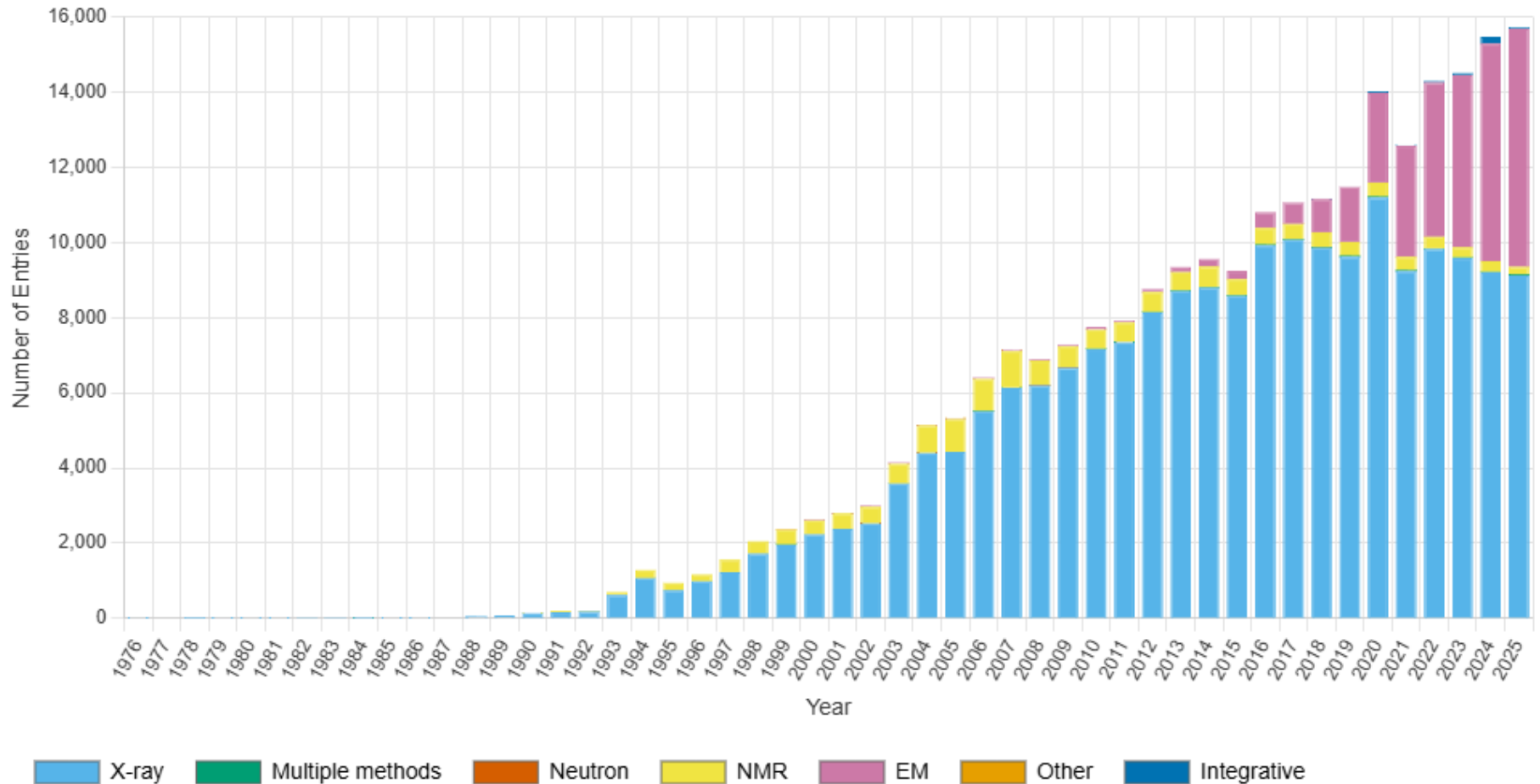
Determination of protein structure

PDB Data Growth by Experimental Method



Determination of protein structure

PDB Data Growth by Experimental Method

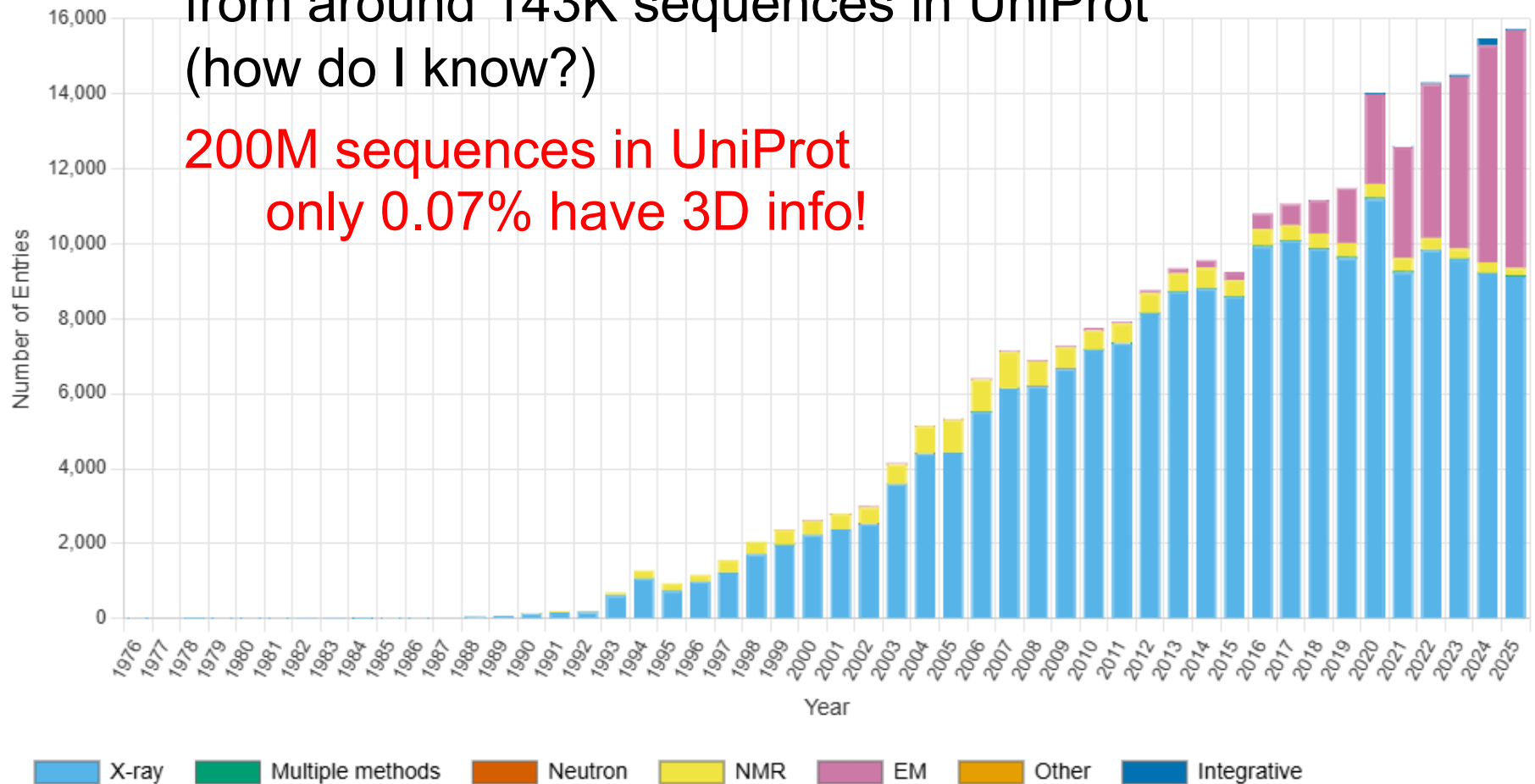


Determination of protein structure

PDB Data G

Currently: 241K protein 3D structures
from around 143K sequences in UniProt
(how do I know?)

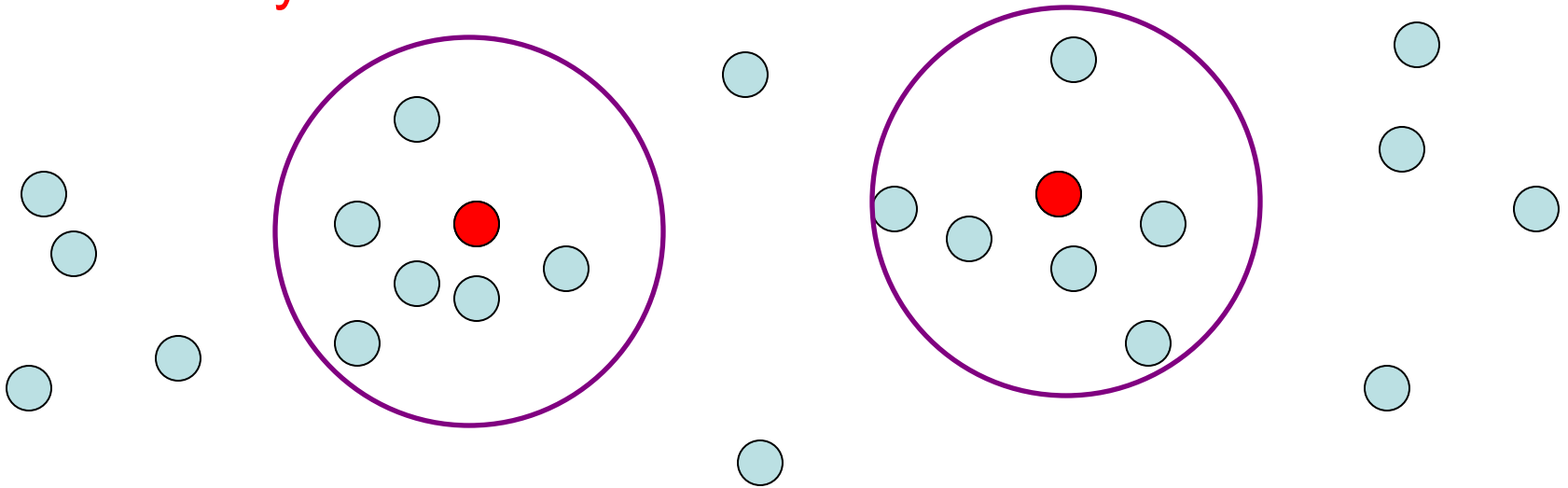
200M sequences in UniProt
only 0.07% have 3D info!



Determination of protein structure

Currently: 241K protein 3D structures
from around 143K sequences in UniProt
(how do I know?)

200M sequences in UniProt
only 0.07% have 3D info!



50% sequences covered (25% in 1995)

Protein structure prediction

Ab initio

Explore conformational space

Limit the number of atoms

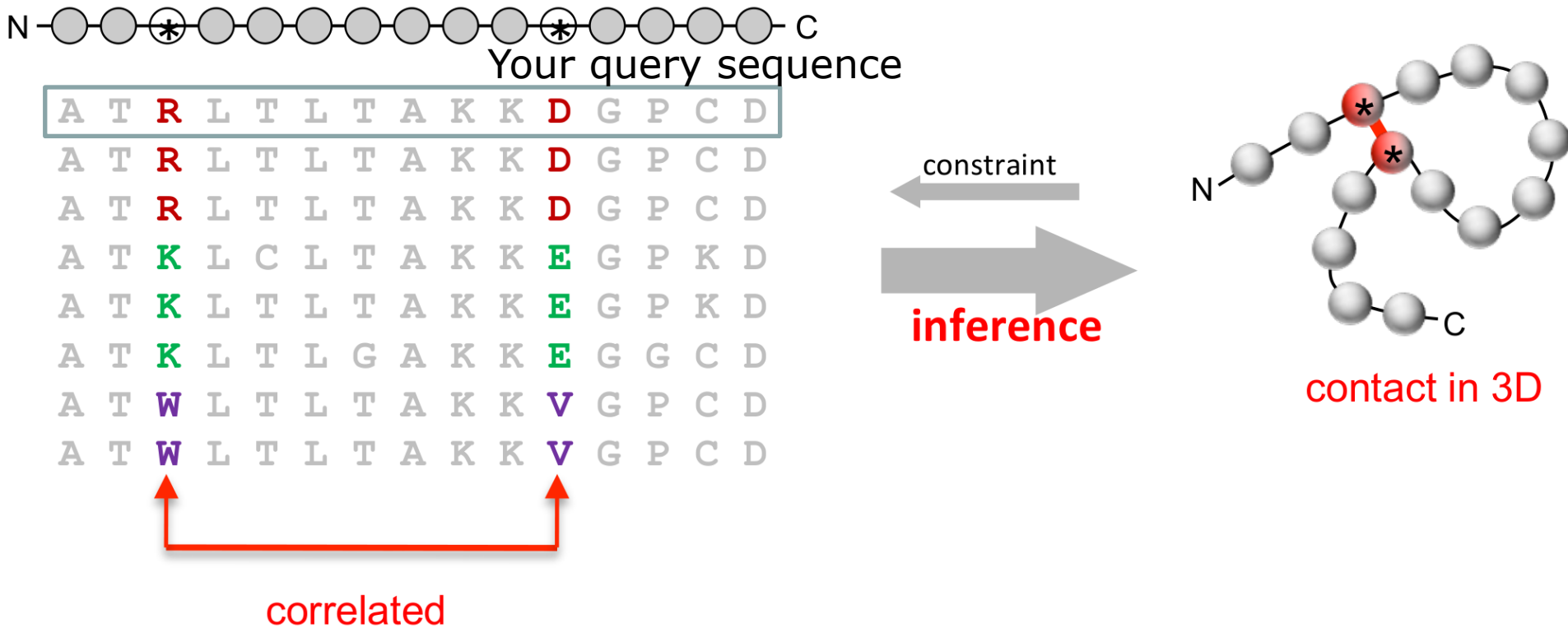
Break the problem into fragments of sequence

Optimize hydrophobic residue burial and pairing of beta-strands

Limited success...

Protein structure prediction

Correlated mutations



https://commons.wikimedia.org/wiki/File:Correlated_mutation.png

Protein structure prediction Combined

Homology to solved structures

Correlated sequence variation in homologs

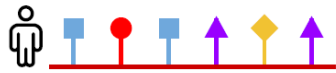
Generation of a structure following physical constraints

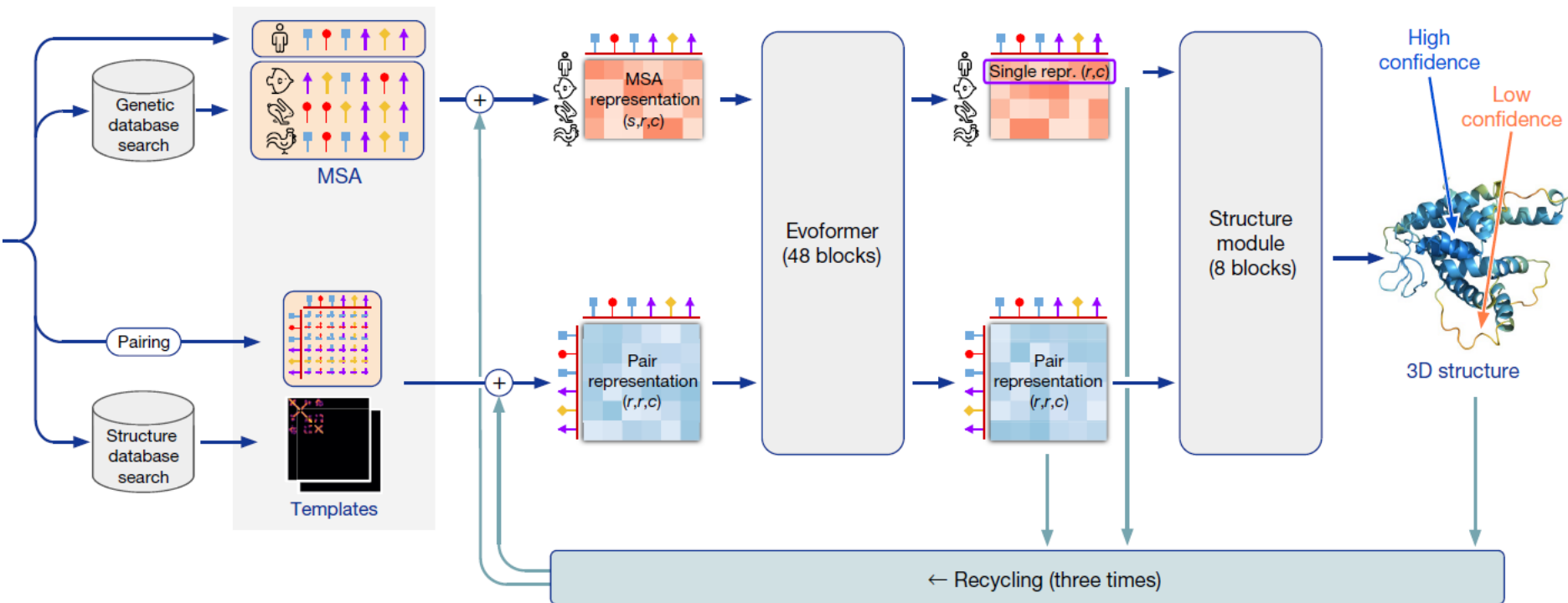
Protein structure prediction

AlphaFold: DeepMind, Google

Demis
Hassabis



A diagram showing an input sequence as a horizontal red line with six colored arrows (blue, red, blue, purple, yellow, purple) pointing upwards from it.
Input sequence



Jumper et al (2021) Nature

Nobel prize in Chemistry 2024

Protein structure prediction

AlphaFold: DeepMind, Google

UniProtKB - P08235 (MCR_HUMAN)

Display

[Help video](#)

[BLAST](#)

[Align](#)

[Format](#)

[Add to basket](#)

[History](#)

Entry

[Publications](#)

[Feature viewer](#)

[Feature table](#)

Protein | **Mineralocorticoid receptor**

Gene | **NR3C2**

Organism | *Homo sapiens* (H. sapiens)

Status | [Reviewed](#) - A

Structureⁱ

Model Confidence:

Very high (pLDDT > 90)

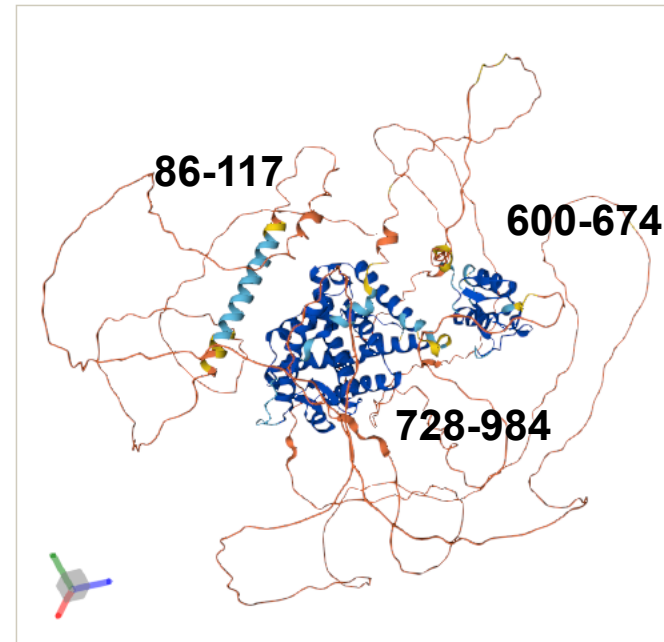
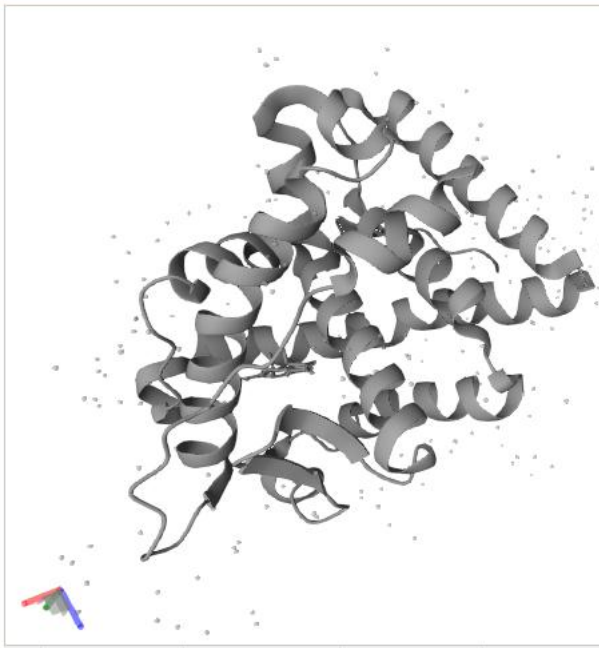
Confident (90 > pLDDT > 70)

Low (70 > pLDDT > 50)

Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions with low pLDDT may be unstructured in isolation.

Structureⁱ



AlphaFold	AF-P08235-F1	Predicted			1-984
-----------	--------------	-----------	--	--	-------

PDB	1Y9R	X-ray	1.96 Å	A/B	731-984
-----	------	-------	--------	-----	---------

Protein structure prediction

AlphaFold: DeepMind, Google

Precomputed models:

UniProt

<https://alphafold.ebi.ac.uk/>

(200M models)

AlphaFold server

<https://alphafoldserver.com/>

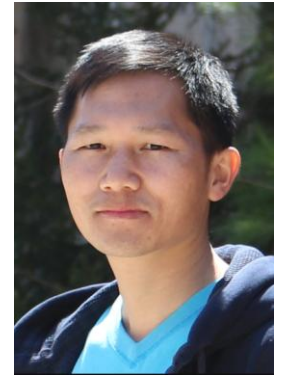
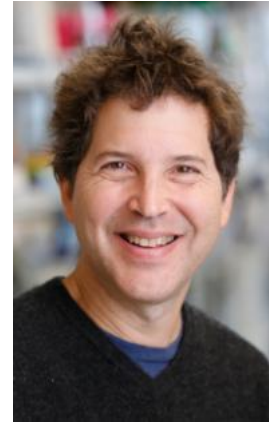
Complexes, PTMs, RNA, DNA, ligands,
ions (30 models per day)

Source code (Needs 3 Tb disk space)

Protein structure prediction

trRosetta: David Baker & Jianyi Yang

Needs large multiple sequence alignments to predict contacts



Predictions available for all PFAM domains

Example:

<https://www.ebi.ac.uk/interpro/entry/pfam/PF07887/rosettafold/>

Runs online at

<https://yanglab.qd.sdu.edu.cn/trRosetta/>

Du et al (2021) *Nature Protocols*

Protein structure prediction

C-I-Tasser: Yang Zhang



Run online at
<https://zhanggroup.org/C-I-TASSER/>

Zheng et al (2021) *Cell Reports Methods*

Protein structure prediction

AlphaFold Multimer

Predicts protein complexes

Evans et al (2022) *bioRxiv*

AlphaFold 3

Structure of proteins (with PTMs),
ligands, RNA/DNA

Abramson et al (2024) *Nature*

[nature](#) > [articles](#) > article

Article | [Open access](#) | Published: 08 May 2024

**Accurate
interaction**

Update November 11, 2024: As of November 2024, we have released [AlphaFold 3 model code and weights](#) for academic use to help advance research. Learn more about [AlphaFold tools](#).

Protein structure prediction

AlphaFold 3

Structure of proteins (with PTMs),
ligands, RNA/DNA

Abramson et al (2024) *Nature*

AlphaFold Server allows you to model a structure consisting of many biological molecules

[Learn more](#) ▾

⬆ Upload JSON

↻ Clear



Entity type
Protein ▾

Copies
1

TPTPTIQEDG	SPALIAKRAS	VTESCNIGYA	STNGGTTGGK	GGATTTVSTL
AQFTKAAESS	GKLNIVVKGK	ISGGAKVRVQ	SDKTIIGQKG	SELVGTGLYI
NKVKNVIVRN	MKISKVKDSN	GDAIGIQASK	NVWVDHCDLS	SDLKSGKDYI
DGLLDITHGS	DWVTVSNTFL	HDHFKASLIG	HTDSNAKEDK	GKLHVTYANN
YWYNVNSRNP	SVRFGTVHIY	NNYYLEVGS	AVNTRMGAQV	RVESTVFDKS
TKNGIISVDS	KEKGYATVGD	ISWGSSTNTA	PKGTLGSSNI	PYSYNLYGKN
NVKARVYGTA	GQTLGFAAAS	FLEQKLISEE	DLNSAVDHHH	HHH

PTMs 22T: Alpha-D-mannose 44T: Alpha-D-mannose 45T: Alpha-D-mannose 46T: Alpha-D-mannose 48S: Alpha-D-mannose
54T: Alpha-D-mannose



Entity type
Ion ▾

Copies
1

Ca²⁺