



JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

Protein structure prediction

Miguel Andrade

Faculty of Biology,

Institute of Organismic and Molecular Evolution

Johannes Gutenberg University

Mainz, Germany

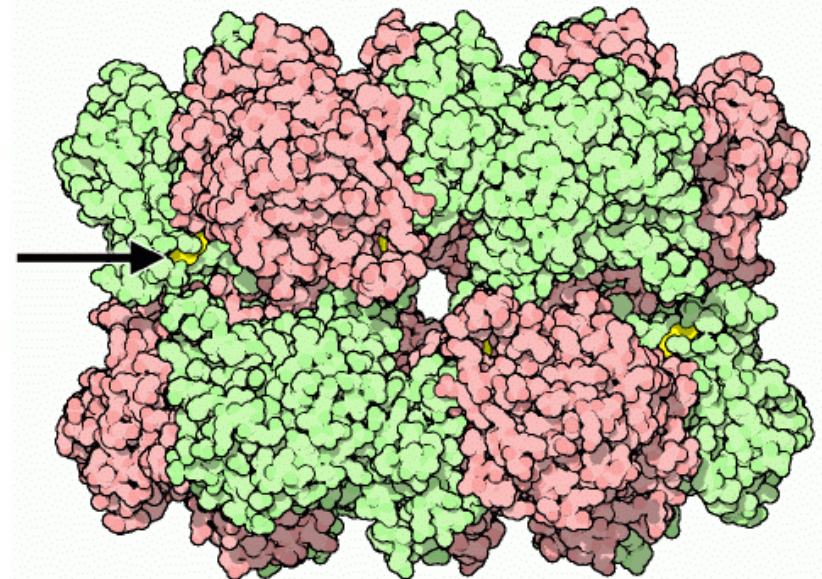
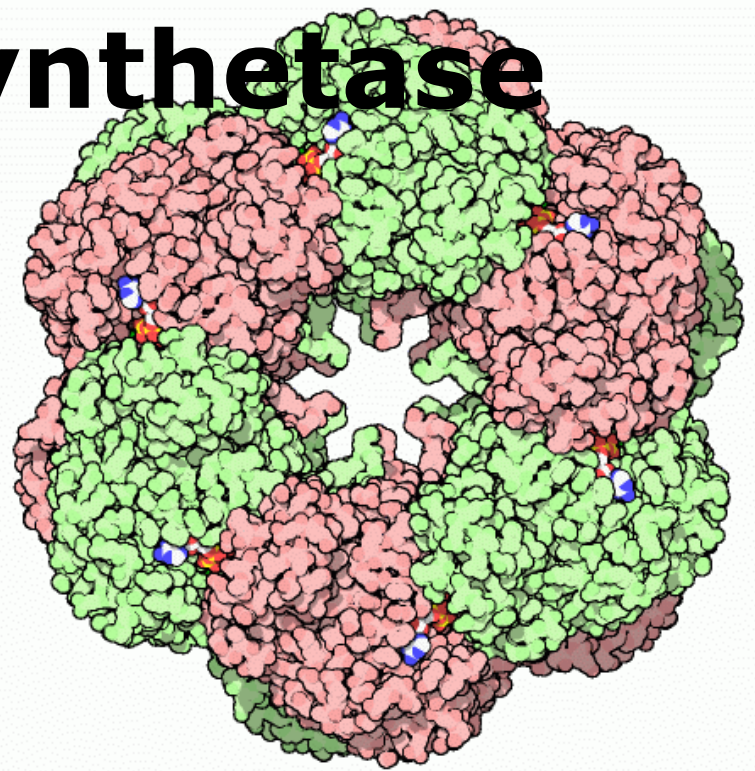
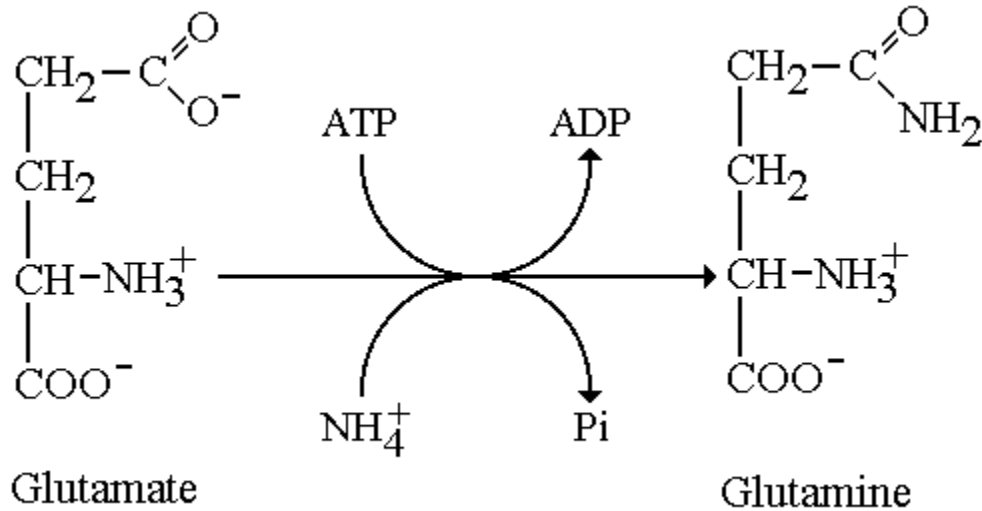
andrade@uni-mainz.de

Mount Everest

A photograph of Mount Everest, the highest mountain in the world, covered in snow and set against a clear blue sky. The mountain's peak is sharp and pointed, with snow clinging to its ridges and crevasses. The foreground shows snow-covered slopes and rocky outcrops.

Age: 60M years

Glutamine synthetase

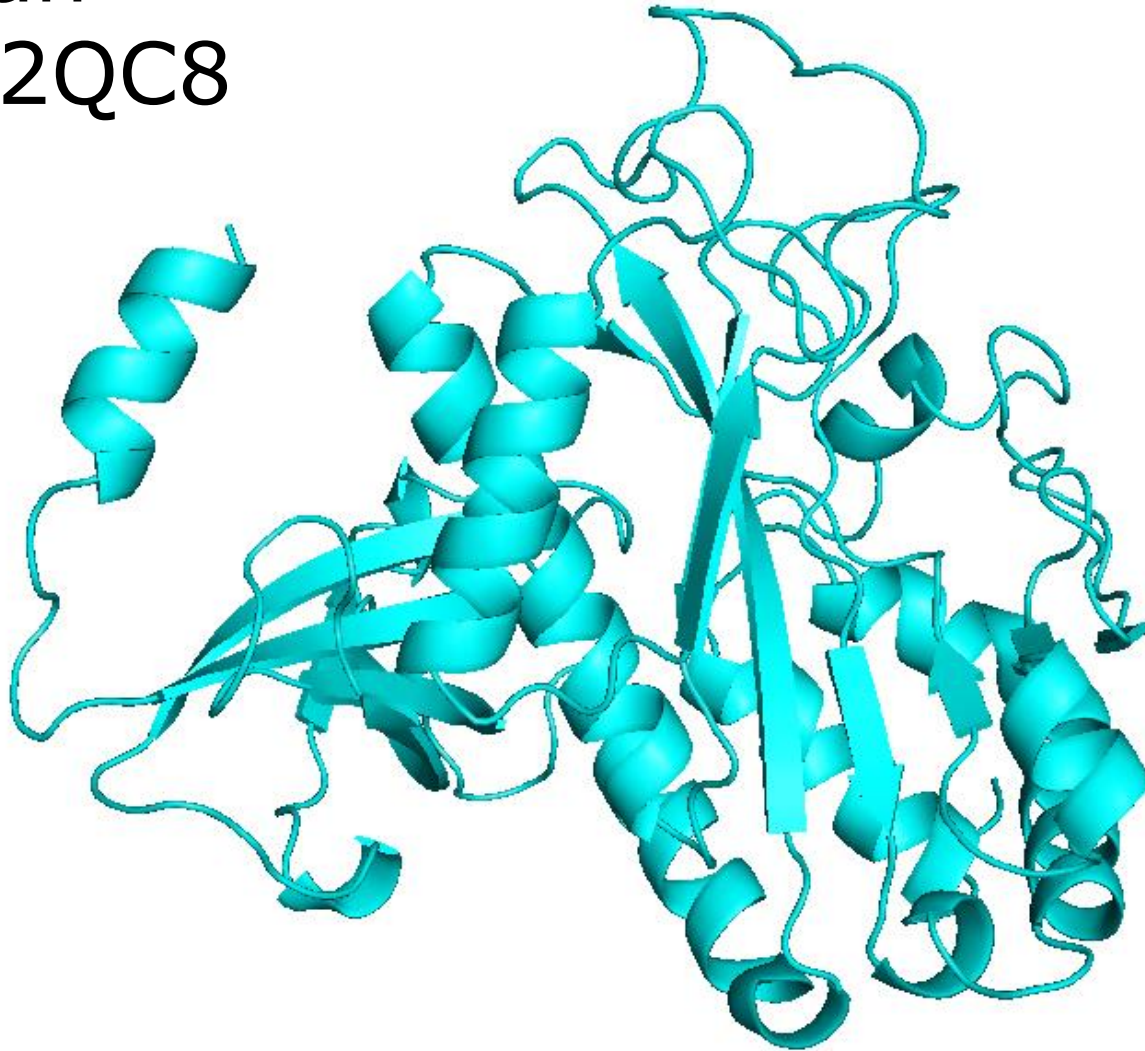


Age: +3500M years

Glutamine synthetase

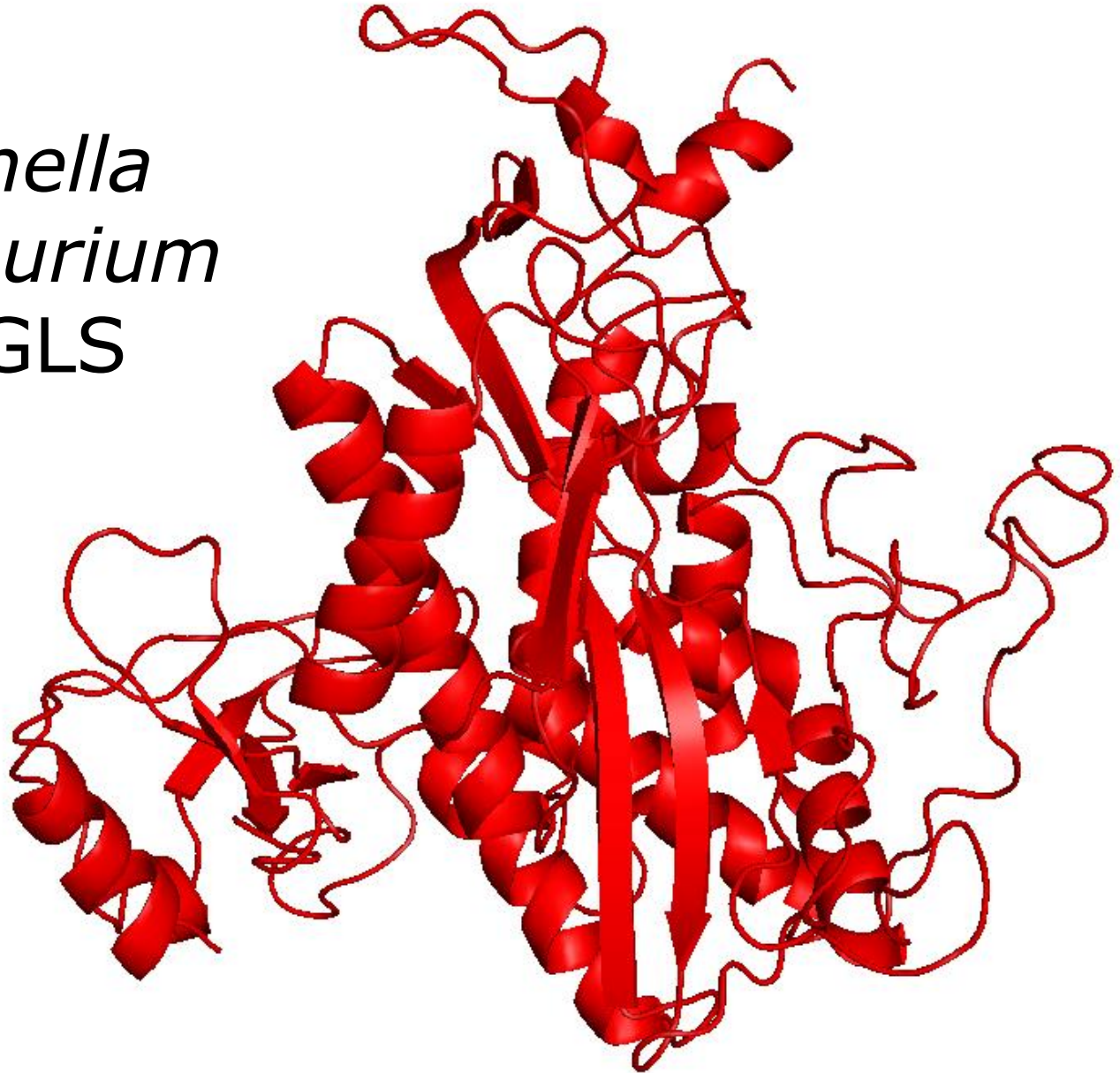
Human

PDB:2QC8

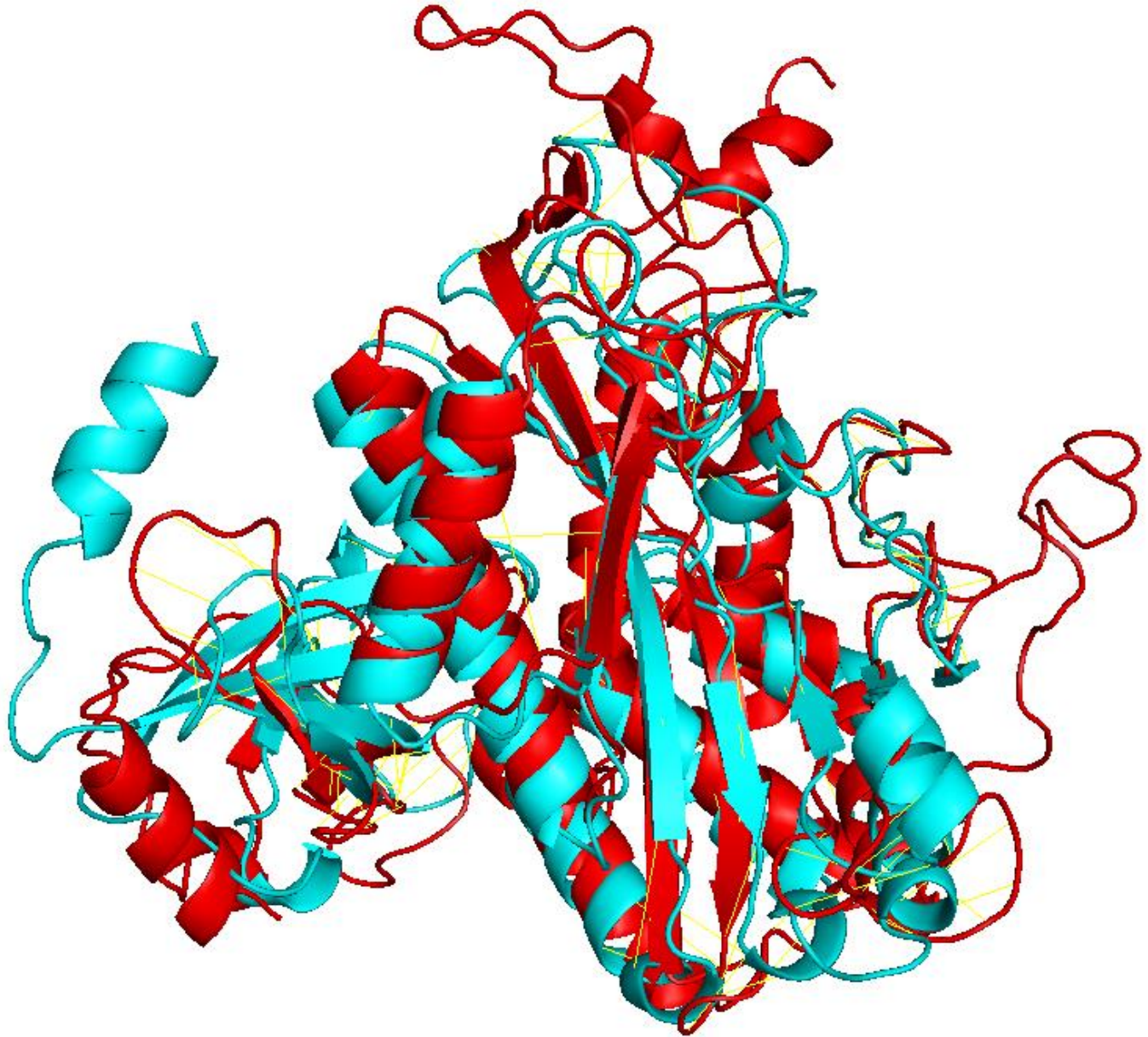


Glutamine synthetase

Salmonella
typhimurium
PDB:2GLS



Glutamine synthetase



Time line

Earth: 4.6 By

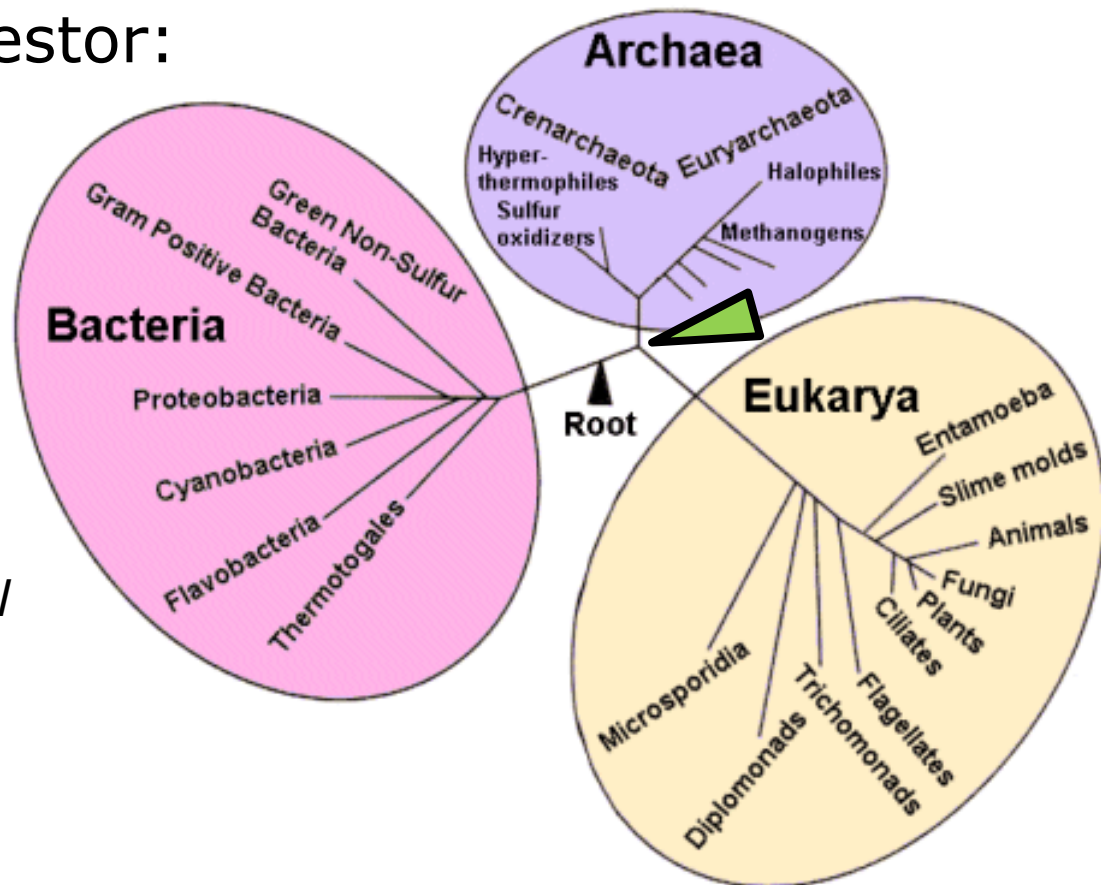
Origin of life: 3.9 By – 3.5 By

Last Common Ancestor:
3.5 – 3.8 By

Glansdorff & Labedan
(2008) *Biology Direct*

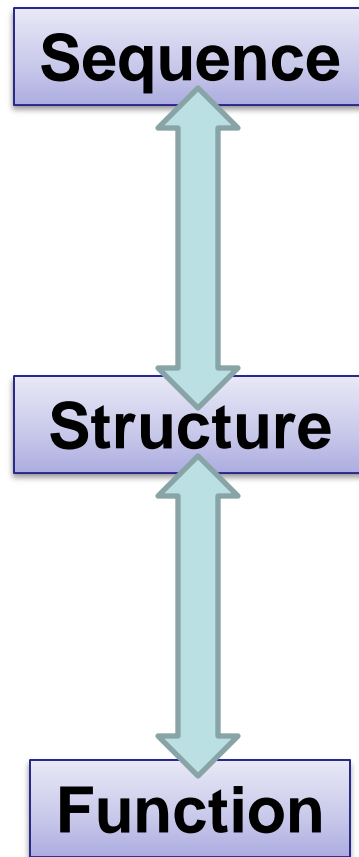
4.29 By

Sheridan *et al.* (2003)
Geomicrobiology Journal



Sequence and function

Evolutionary constraints



MTQDELKKAVGWAALQYVQ

PG

LG

EK

DA

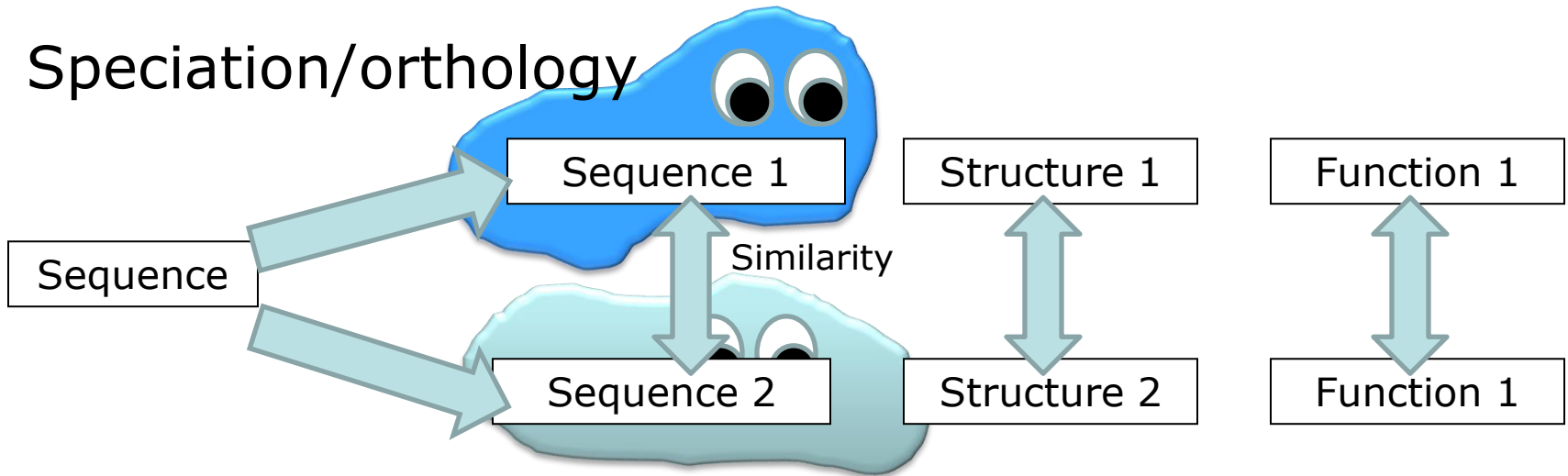
ST



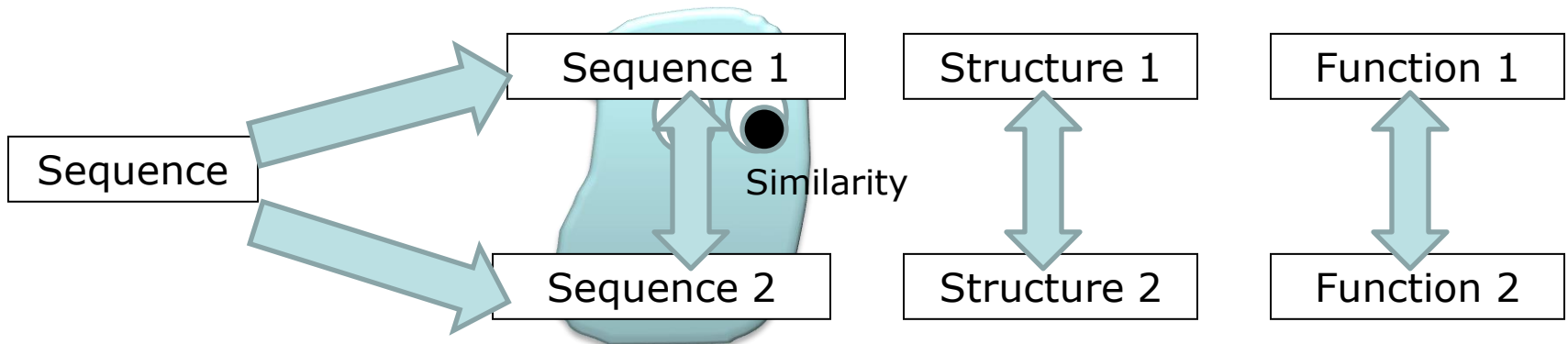
Sequence and function

Evolutionary constraints

Speciation/orthology



Gene duplication/paralogy



Sequence pairwise alignment

```
>gs_human gi|74271837|ref|NP_001028216.1| glutamine synthetase [Homo sapiens]
MTTSASSHLNKGIKQVYMSLPQGEKVQAMYIWIDGTGEGLRCKTRTLTLDSEPKCVEELPEWNFDSSTLQS
EGSNSDMYLVPAAMFRDPFRKDPNKLVLCEVFKYNRRPAETNLRHTCKRIMDMVSNQHPWFGMEQEYTLM
GTDGHPFGWPSNGFPGPQGPYYCGVGADRAYGRDIVEAHYRACLYAGVKIAGTNAEVMPAQWFEFQIGPCE
GISMGDHLWVARFILHRVCEDFGVIAFTDPKPIPGNWNGAGCHTNFSTKAMREENGLKYIEEAIEKLSKR
HQYHIRAYDPKGGLDNARRLTGFHETSNINDFSAGVANRSASIRIPRTVGOEKKGYFEDRRPSANCDPFS
VTEALIRTCLLNETGDEPFQYKN
```

```
>gs_salmonella gi|16767272|ref|NP_462887.1| glutamine synthetase [Salmonella
enterica subsp. enterica serovar Typhimurium str. LT2]
MSAEHVLTMLNEHEVKFVDLRFTDTKGKEQHVTIPAHQVNAEFFEKGKMGFDGSSIGGWKGINESDMVLMP
DASTAVIDPFFADSTLIIRCDILEPGTLQGYDRDPRSIakraedyLRATGIADTVLFGPEPEFFLFDDIR
FGASISGSHVAIDDIEGAWNSSTKYEGGNKGHRPGVKGGYFPVPPVDSAQDIRSEMCLVMEQMGLVVEAH
HHEVATAGQNEVATRENTMTKKADEIQIYKYVVHNVHRFGKTATFMPKPMFGDNGSGMHCHMSLAKNGT
NLFSGDKYAGLSEQALYYIGGVIKHAKAINALANPTTNSYKRLVPGYEAPVMLAYSARNRSASIRIPVVA
SPKARRIEVRFDPANPYLCFAALLMAGLDGIKNKIHPGEAMDKNLYDLPPEEAKEIPQVAGSLEEALN
ALDLDFLKAAGGVFTDEAIDAYIALRREEDDRVRMTPHPVEFELYYSV
```

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 FDGSSIGGWKGINESDMVLMPDASTAVIDPFFADSTLIIRCDILEPGTLOGYDRDPRSIA 109

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

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

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

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

Query 298 RRLTGFHETSNINDFSAGVANRSASIRIPRTVQEKKGYPEDRRPSANCDPFSVTEALIR 357
          +RL  +E  +  +SA  NRSASIRIP V  K  E R P  +P+  AL+
Sbjct 321 KRLVPGYEAPVMLAYSAA--RNRSASIRIP-VVASPKARRIEVRFDPDPAANPYLCFAALLM 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]
MTTSASSHLNKGIKQVYMSLPQGEKVQAMYIWIWIDGTGEGLRCKTRTLTLDSEPKCVEELPEWNFDSSTLQS
EGSNSDMYLVPAAMFRDPFRKDPNKLVLCEVFKNRRPAETNLRHTCKRIMDMVSNQHPWFGMEQEYTLM
GTDGHPFPGWPSNGFPGPQGPYYCGVGADRAYGRDIVEAHYRACLYAGVKIAGTNAEVMPAQWEFQIGPCE
GISMGDHLWVARFILHRVCEDFGVIAFTDPKPIPGNWNWAGACHTNFSSTKAMREENGLKYIEEAIEKLSKR
HQYHIRAYDPKGGLDNARRLTGFHETSNINDFSAGVANRSASIRIPRTVVGQEKKGYPFEDRRPSANCDPFS
VTEALIRTCLLNETGDEPFQYKN
```

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

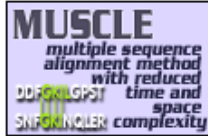
```
MPTRNLEIEPADLWRILKASGIKYVKFIIVDINGAPRSEIVPIDMAKDLFIDGMPFDASSIPSYSTVTKS
DFVAYVDPRAVYVEYWQDGKQVADVFTMVSDIADKPSPLDPRRVLNDALEQARSKGYEFLMGVEVEFFVIK
EDGGKPVFADPGIYFDGWNVTVQSQFMKELITAIADAGINYTKTHHEVAPSQYEVNIGATDPLRLADQIV
YFKIMAKDIARKYGLVATFMPKPFVWVNGSGAHTHISVWKGDKNLFQSSTGKITEECGYAISAILSNARA
LSSFVAPLVNSYKRLVPHYEAPTRIVWGYANRSAMIRIPQYKMRINRIEYRHPDPSMNPYLAFTAIKTM
IRGLEEKKEPPPTTEEVAYELANALET PATLEDTLKELSKSFLATEL PSELVNAYIKIKQNEWEDYLTNV
GPWEKTWNIITQWEYNKYLVTA
```

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

```
MSAEHVLTMLNEHEVKFVDLRFTDTKGKEQHVTIPAHQVNAEFFEKGKMGFDGSSIGWKGINESDMVLMP
DASTAVIDPFFADSTLIIRCDILEPGTLQGYDRDPRSIakraedyLRATGIADTVLFGPEPEFFLFDDIR
FGASISGSHVAIDDIAGAWNSSTKYEGGNKGHRPGVKGGYFPVPPVDSAQDIRSEMCLVMEQMGLVVEAH
HHEVATAGQNEVATRFNTMTKKADEIQIYKYVVHNVHRFGKTATFMPKPMFGDNGSGMHCHMSLAKNGT
NLFSGDKYAGLSEQALYYIGGVIKHAKAINALANPTTNSYKRLVPGYEAPVMLAYSARNRSASIRIPVVA
SPKARRIEVRFDPAAANPYLCFAALLMAGLDGKIKNIHPGEAMDKNLYDLPPEEAKEIPQVAGSLEEALN
ALDLDREFLKAGGVFTDEAIDAYIALRREEDDRVRMT'PHPVEFELYYSV
```


```
>gs_yeast gi|330443748|ref|NP_015360.2| Gln1p [Saccharomyces cerevisiae S288c]
```

```
MAEASIEKTQILQKYLELDQRGRIIAEYVWIDGTGNLRSKGRTLKKRITSIDQLPEWNFDSSTNQAPGH
DSDIYLPVAYYPDPFRRGDNIIVLAACYNNDGTPNKFNRHEAAKLFAAHKDEEIIWFGLEQEYTLFDMY
DDVYGWPKGGYPAPQGPYYCGVGAGKVYARDMIEAHYRACLYAGLEISGINAEVMPQSQWEFQVGPCTGID
MGDQLWMARYFLHRVAEEFSGIKISFHPKPLKGDWNGAGCHTNVSTKEMRQPGGMKYIEQAIEKLSKRHAE
HIKLYGSDNDMRLTGRHETASMTAFSSGVANRGSIRIPRSVAKEGYGYFEDRRPASNIDPYLVTGIMCE
TVCGAIDNADMTKEFERESS
```



- [Help](#)
- [MUSCLE website](#)
- [Jalview](#)
- [Programmatic Access](#)
- [Download](#)

- [Related Applications](#)
 - [Pairwise Sequence Alignment](#)
 - [Multiple Sequence Alignment](#)
 - [Phylogeny](#)

MUSCLE related literature 

Search for MUSCLE related literature in Medline... [more](#)

EBI > Tools > Multiple Sequence Alignment > MUSCLE

MUSCLE - 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.

Internet Explorer users: If button presses (including copy/paste operations) don't appear to work please try enabling Compatibility View.

Use this tool

STEP 1 - Enter your input sequences

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

Or upload a file: No file chosen

STEP 2 - Set your Parameters

OUTPUT FORMAT:

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(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://www.ebi.ac.uk/Tools/msa/muscle/>

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

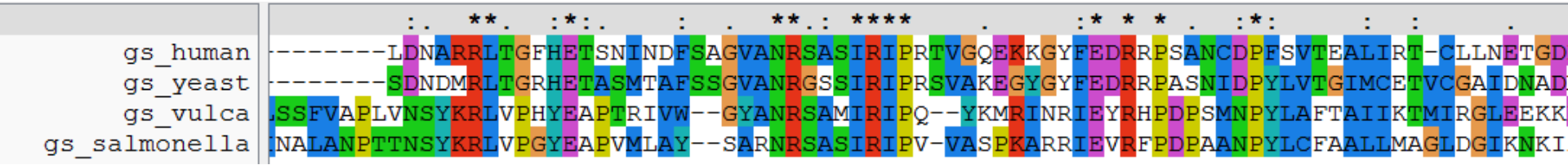
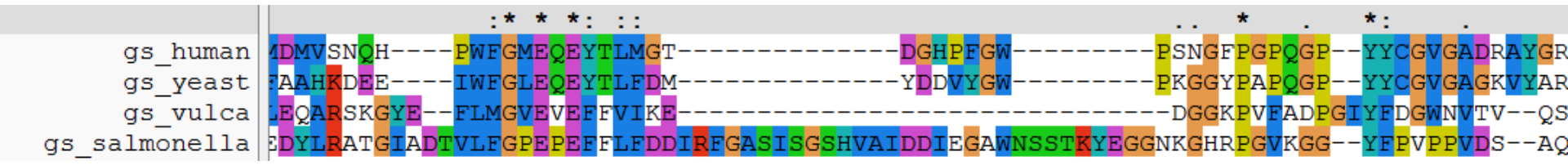
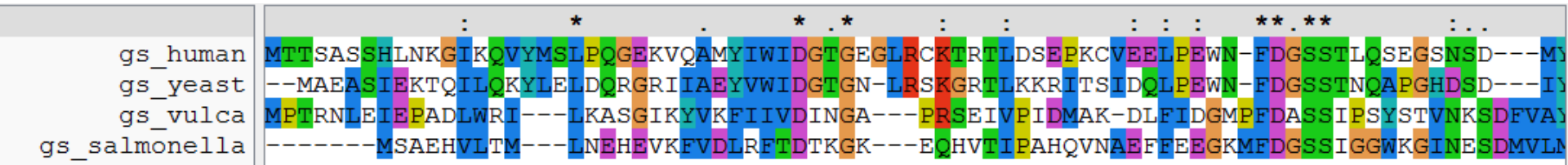
-----FQYKN-----

>gs_yeast gi|330443748|ref|NP_015360.2| Gln1p [Saccharomyces cerevisiae S288c]
--MAEASIEKTQILQKYLELDQRGRIIAEYVWIDGTGN-LRSKRTLKKRITSIDQLPEW
N-FDGSSTNQAPGHSD---IYLKPVAYYDPFRRGDNIVVLAACYNNDGTPN-KFNHRH
EAAKLFAAHKDEE----IWFGLEQEYTLFDM-----YDDVYGW-----
-PKGYPAPQGP--YYCGVGAGKVYARDMIEAHYRACLYAGLEISGINAEVMPA-QWEFQ
VGPCTGIDMGDQLWMARYFLHRVAEEFGIKISFHPKPLKGDWNGAGCHTNVSTKEMRQPG
GMKYIEQAIEKLSKRHAHEHIKLYG-----SDNDMRLTGRHETASMTAFSSGV
ANRGSSIRIPRSVAKEGYGYFEDRRPASNIDPYLVGTGIMCETVCGAIDNADMT-----

-----KEFERESS-----

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

>gs_salmonella gi|16767272|ref|NP_462887.1| glutamine synthetase [Salmonella enterica]
-----MSAEHVLTM---LNEHEVKFVDLRFTDTKGK---EQHVTIPAHQVNAEFFEEG
KMFDFGSSIGGWKGINESDMVLMPPDASTAVIDPFFADSTLIIRCDILEPGTLQGYDRDPRS
IAKRAEDYL RATGIADTVLFGPEPEFFLFDDIRFGASISGSHVAIDDI EGAWNSSTKYEG
GNKGHRPGVKGG--YFPVPPVDS--AQDIRSEMCLVMEQMGLVVEAHHHEVATAGQNEVA
TRFNTMTKKADEIQIYKYVVHNVVHRFGKTATFMPKPMFGD-NGSGMHCHMS---LAKNG
TNLFSGDKYAGLSEQALYYIGGVIKHAKAINALANPTTNSYKRLVPGYEAPVMLAY--SA
RNRASIRIPV-VASPKARRIEVRFDPANPYLCFAALLMAGLDGIKNIHPGEAMDKN
LYDLPPEEAKEIPQVAGSLEEALNALDLDFEFLKAGGVFTDEAIDAYIALRREEDDRVRM
TPHP-----VEFELYYSV-



ClustalW, JalView

Determination of protein structure

X-ray crystallography (179K in PDB)

- need crystals

Nuclear Magnetic Resonance (NMR)
(14K)

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

Electron microscopy (17K)

- Recently resolution improving a lot!

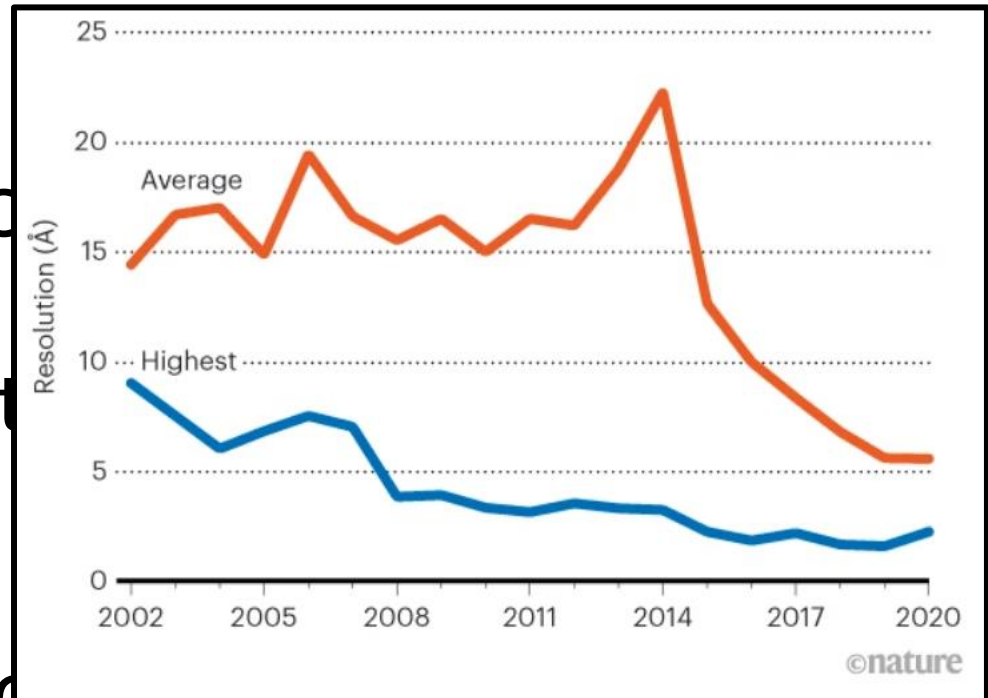
Determination of protein structure

X-ray crystallography (179K in PDB)

- need crystals

Nuclear Magnetic Resonance (14K)

- proteins in solution
- lower size limit



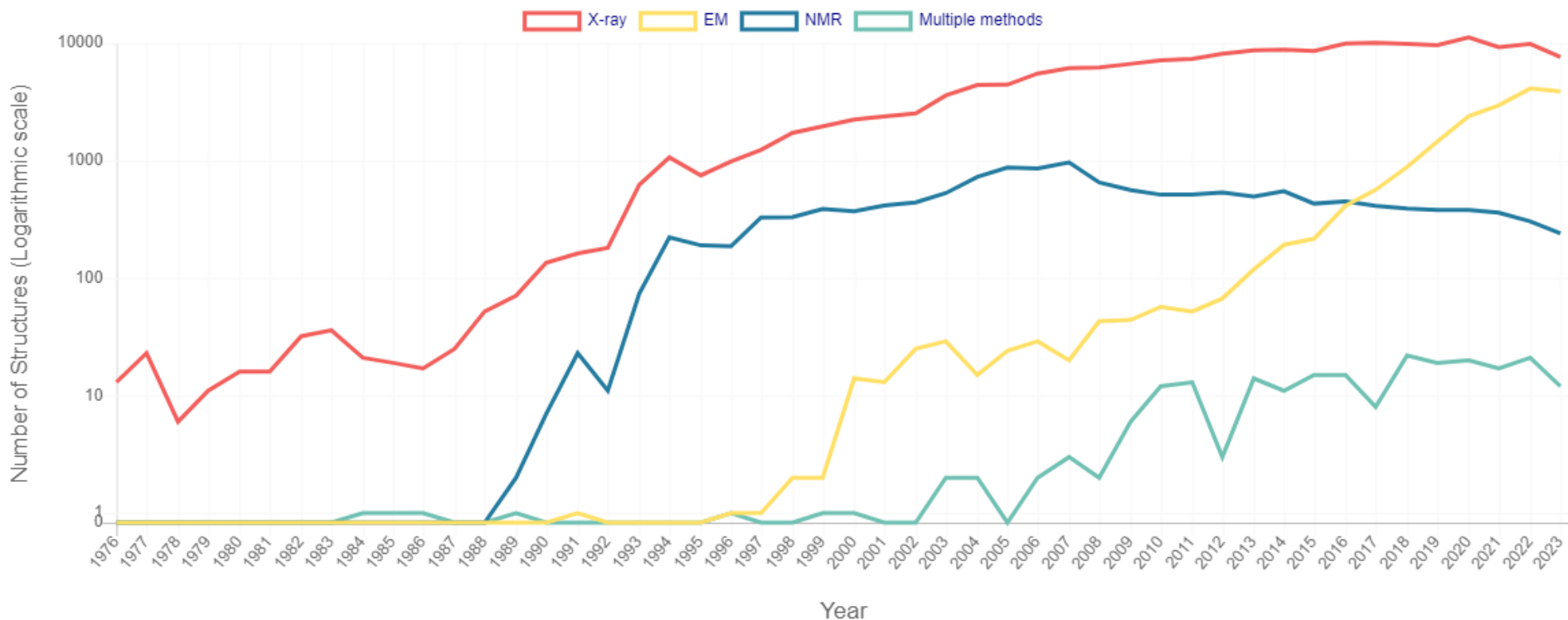
Electron microscopy (17K)

- Recently resolution improving a lot!

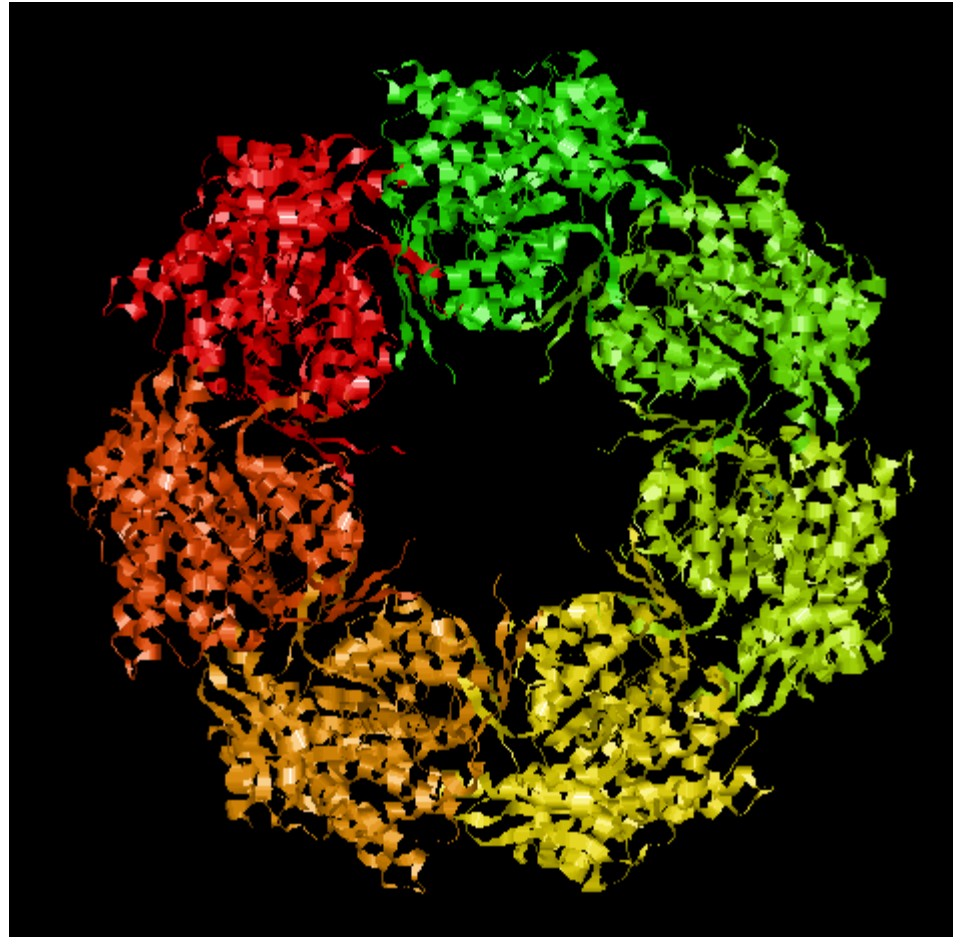
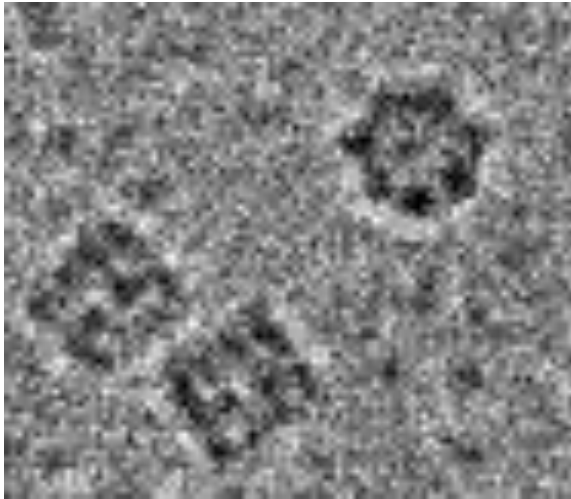
Determination of protein structure

X-ray crystallography (179K in PDB)

Number of Released PDB Structures per Year

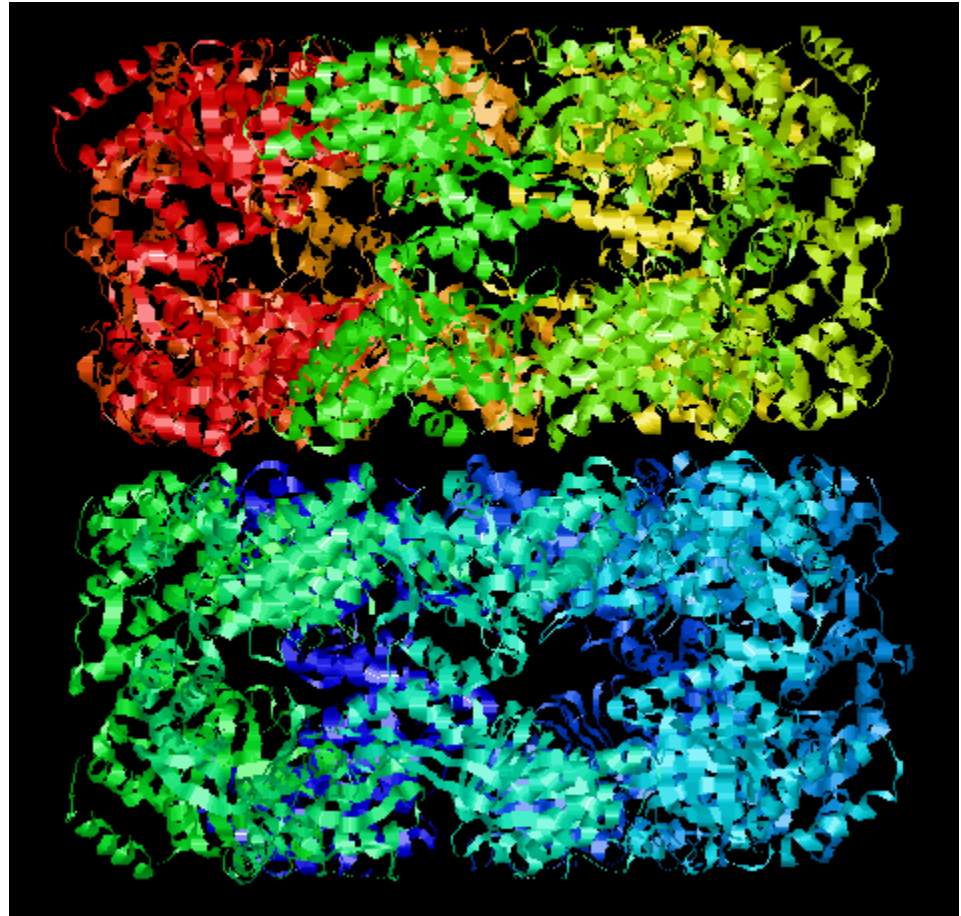
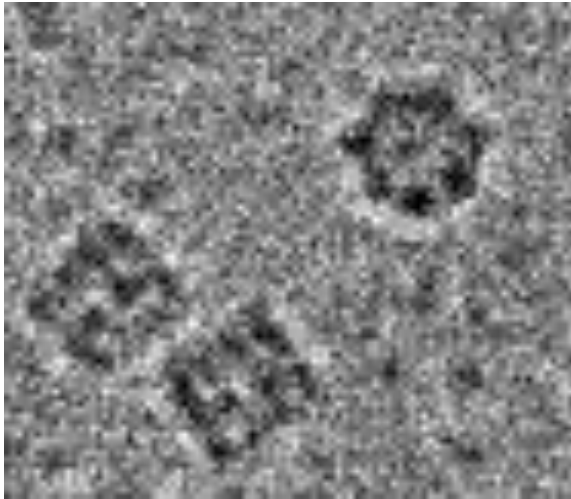


Determination of protein structure



resolution 2.4 Å

Determination of protein structure



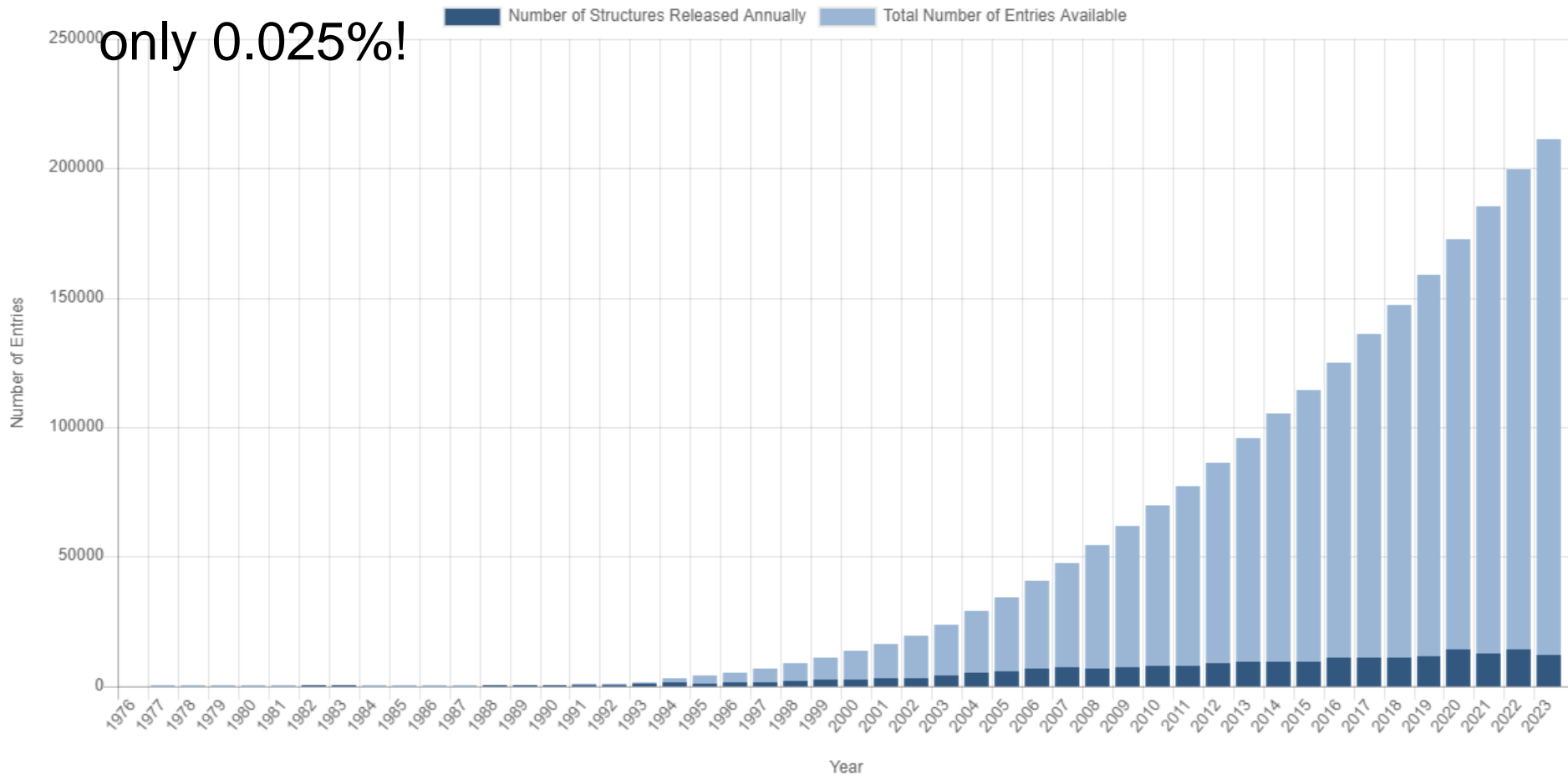
resolution 2.4 Å

Structural genomics

Currently: 211K protein 3D structures
from around 62K sequences in UniProt (how do I know?)

252M sequences in UniProt

only 0.025%!

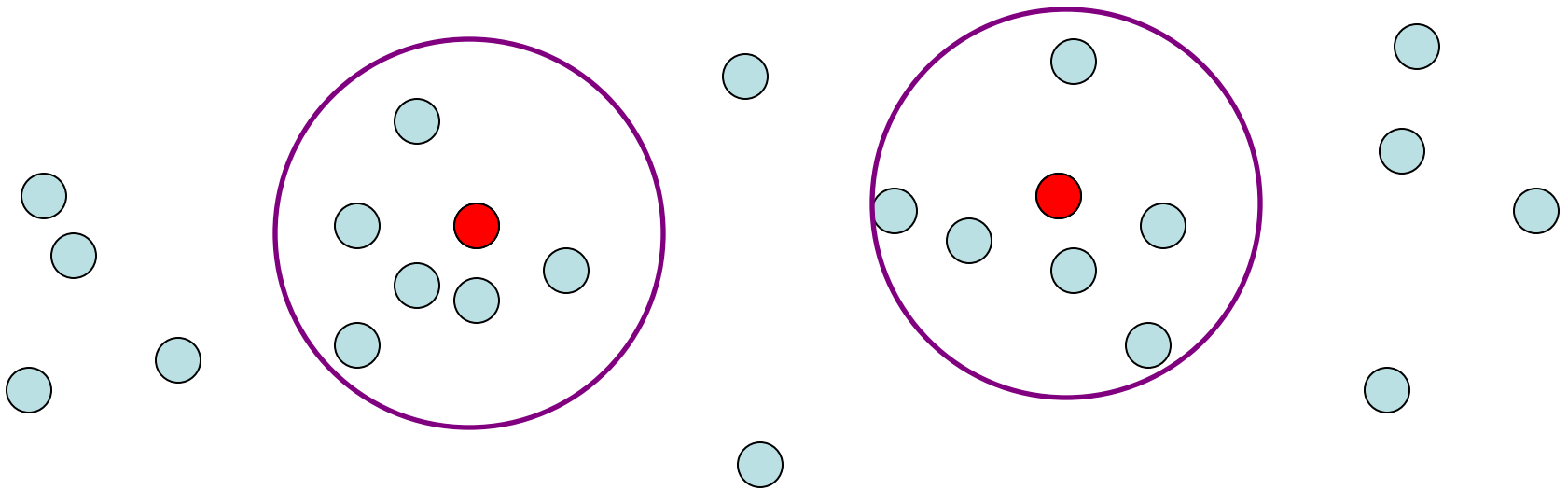


Structural genomics

Currently: 187K protein 3D structures
from around 55K sequences in UniProt (how do I know?)

226M sequences in UniProt

only 0.024%!



50% sequences covered (25% in 1995)

Protein structure prediction

Ab initio

Explore conformational space

Limit the number of atoms

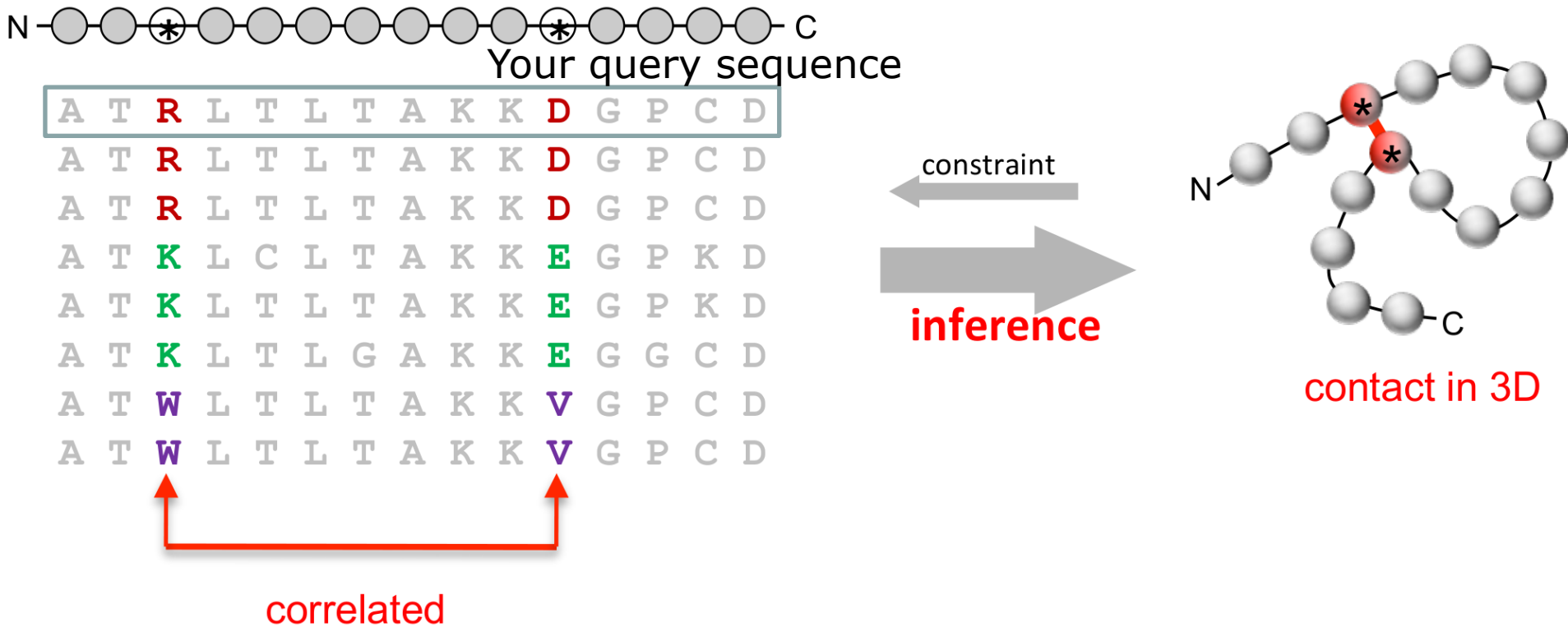
Break the problem into fragments of sequence

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

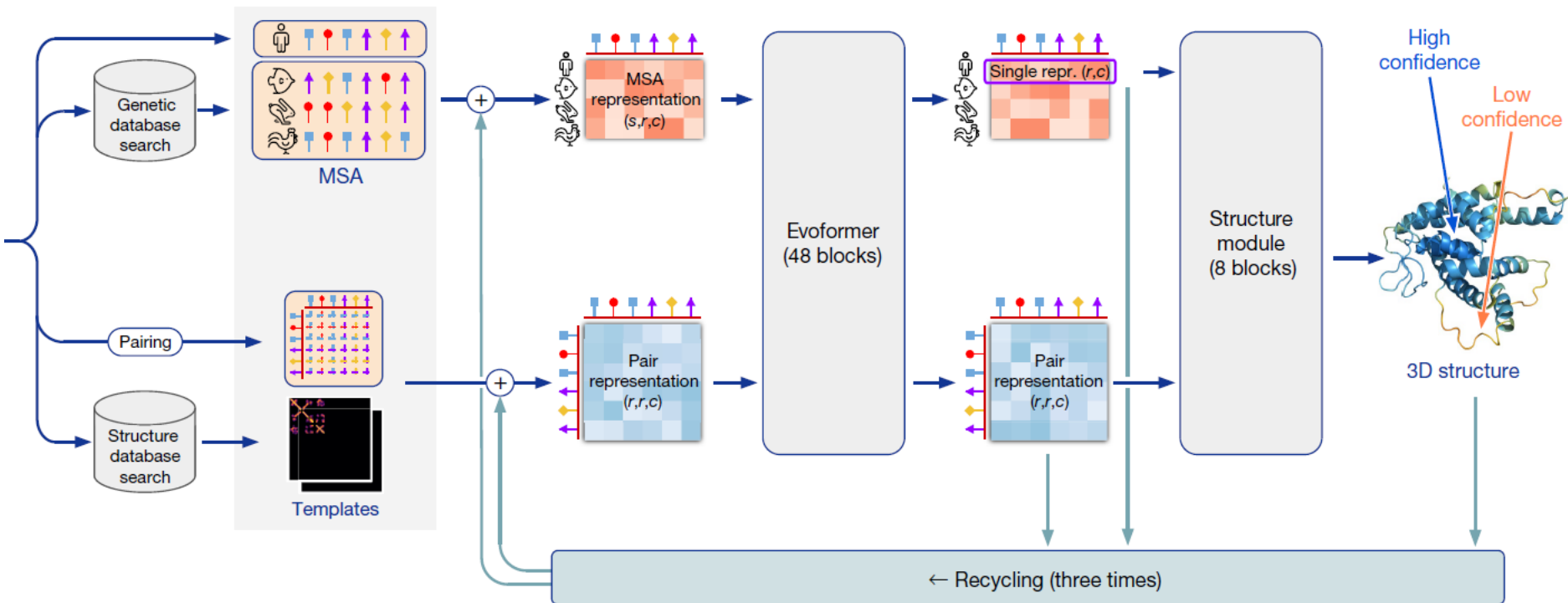
AlphaFold: DeepMind, Google



Demis Hassabis



Input sequence



Jumper et al (2021) Nature

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. Structureⁱ)

Status | Reviewed - A

Model Confidence:

Very high (pLDDT > 90)

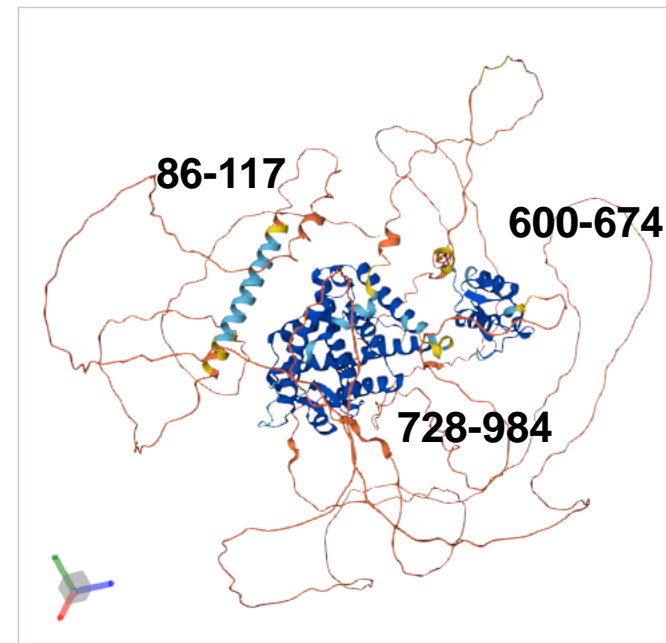
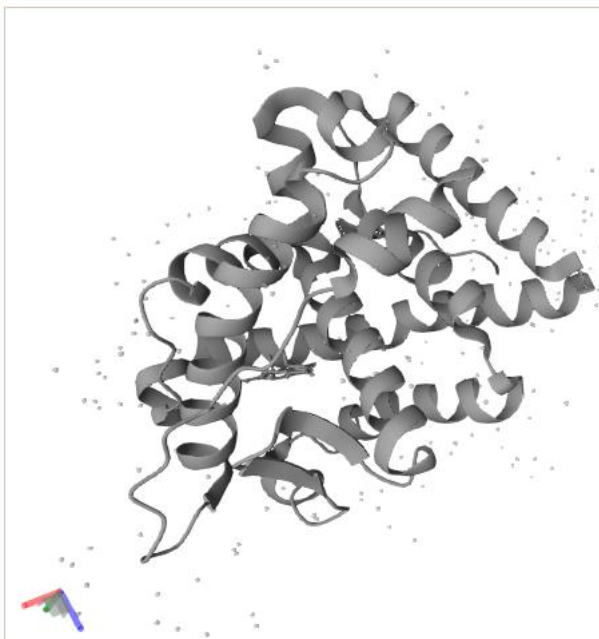
Confident (90 > pLDDT > 70)

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/>

(limited to model organisms)

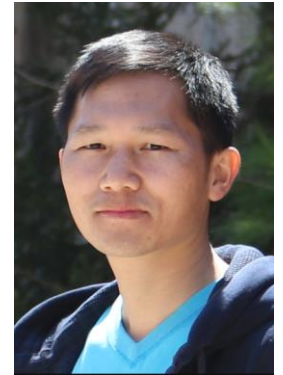
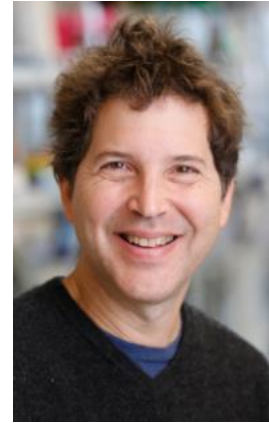
Colab notebook (simplified version / limited server / takes hours)

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/>

Run online at

<https://yanglab.nankai.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*