Protein Evolution
Lecture 2

Lucy Colwell

ICTS Winter School on Quantitative Systems Biology

Dec 13 2017

How to extract useful information from protein sequences?

Each protein is constructed as a specific chain or string of amino acids. There are 20 different amino acids which are used with roughly equal frequencies across all proteins.

In a specific protein of interest, such as hemoglobin, the order of amino acids is highly important and contains all the information necessary to produce the folded, functional molecule.

1	2	3	4	5	6	7	8 …….
M	V	L	S	P	A	D	K ……

We would like to find a probability model for the sequence of amino acids that corresponds to each protein of interest.

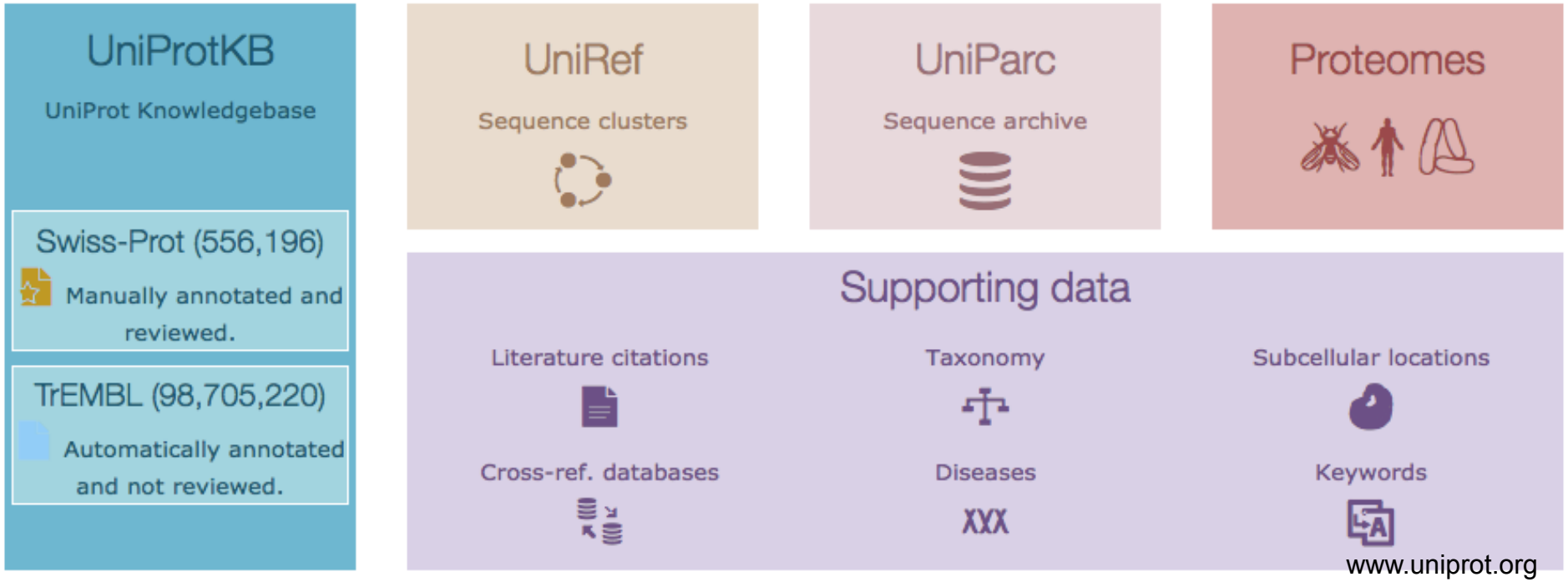
$P(A_1,\ldots,A_L)$ = Probability that a sequence produces a folded, functional hemoglobin molecule.

Sequence Similarity

$P(A_1,\ldots,A_L)$ = Probability that a sequence produces a folded, functional hemoglobin molecule.

To proceed, we need data – if we can find lots of examples of hemoglobin sequences, then we can treat this as an inverse problem, and look for a model that reproduces the statistics of our observed data.

How can we find this data? Over the last 50 years there has been an explosion of technological innovation, which has enabled us to determine the sequences of huge numbers of biological organisms, and thus of numerous proteins.



Protein evolution

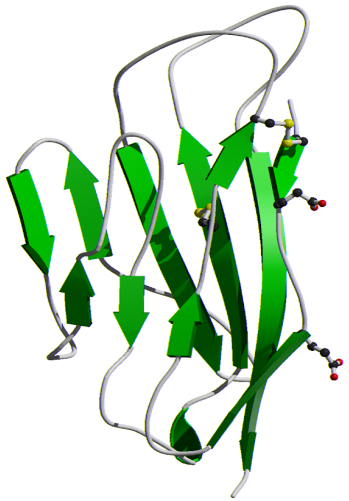
Families of homologous proteins

- Similar/dissimilar sequence
- Common 3D structure
- Common function

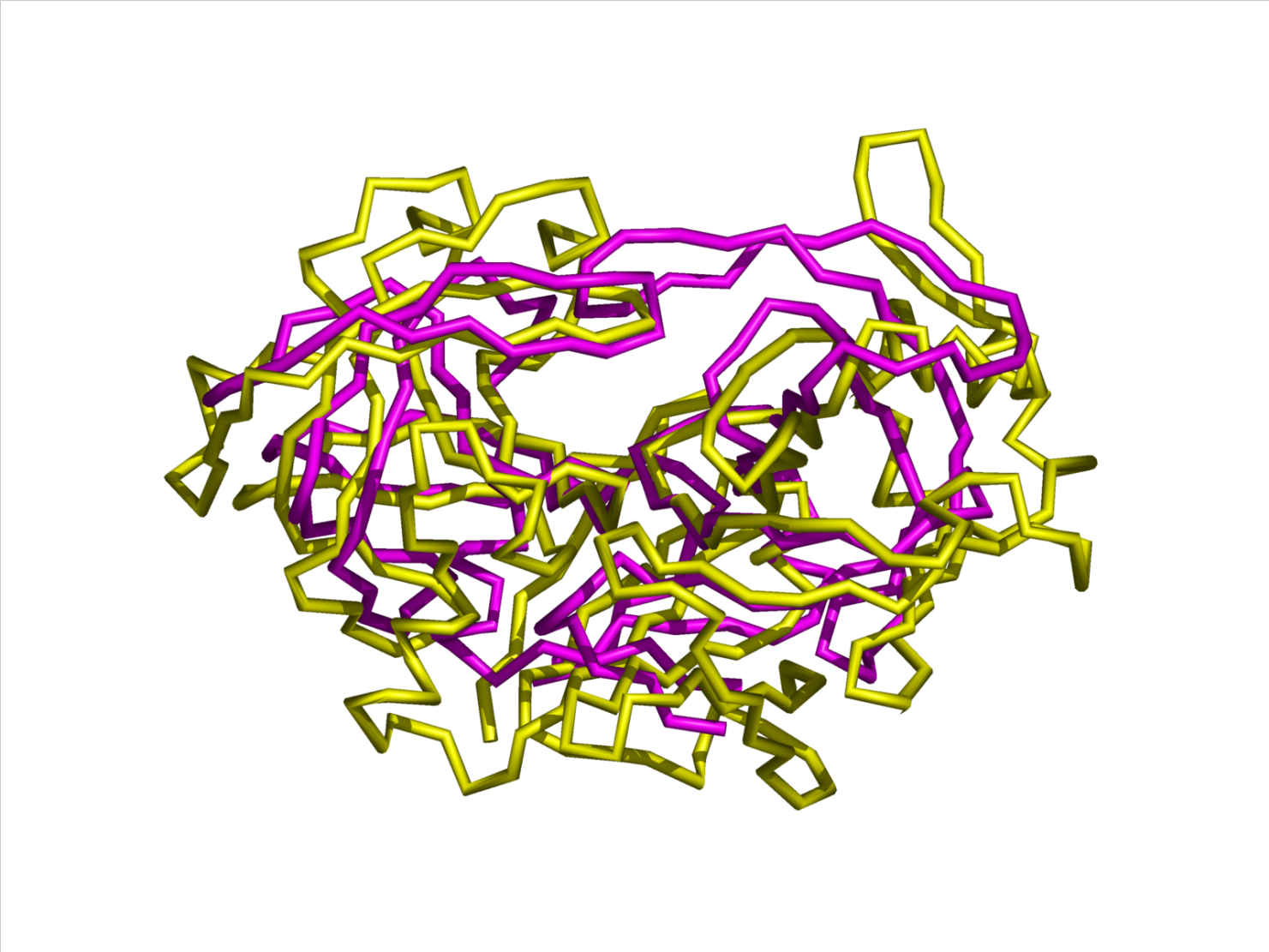
Protein → Known family

VCVEVPSETEA...

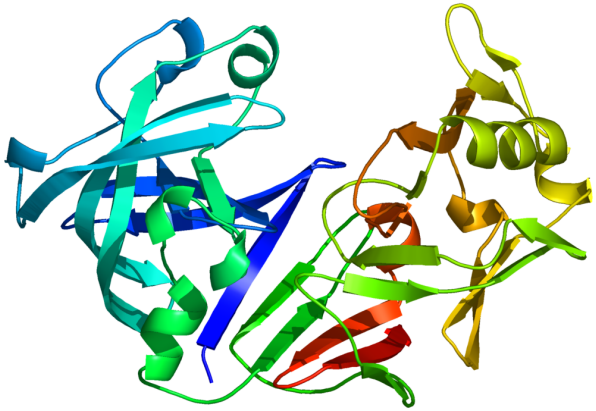
3D structure, function



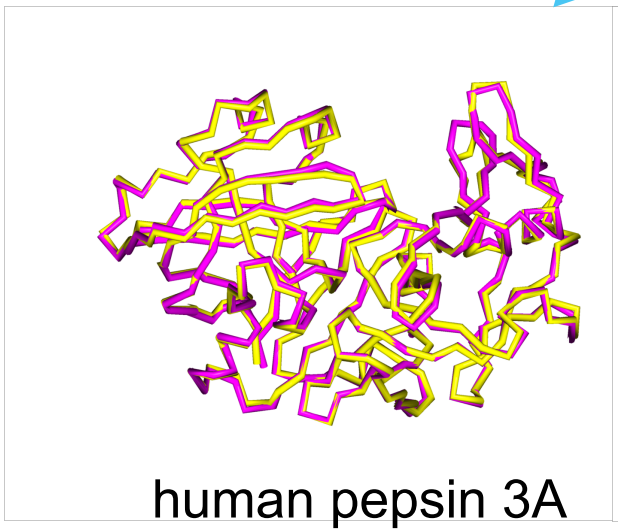
Similar or dissimilar?



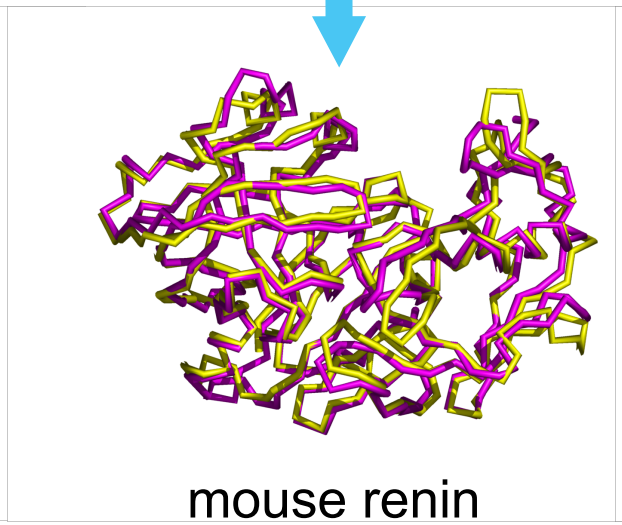
To compare proteins, we need to be able to measure similarity between their sequence and their 3D structures.



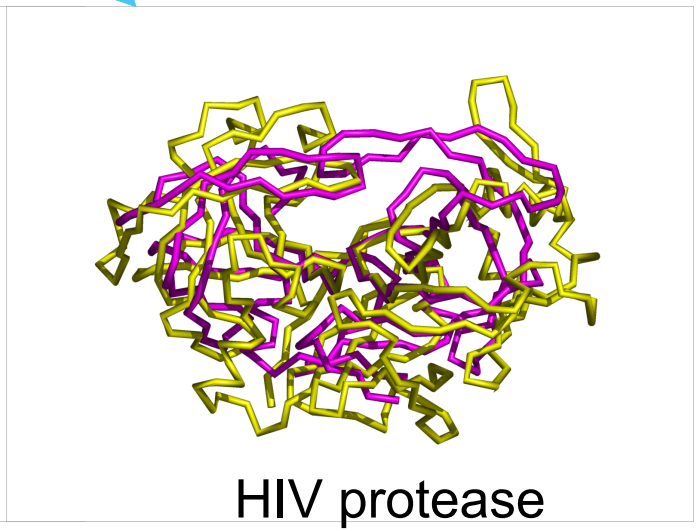
porcine pepsin



human pepsin 3A
(PID 86%, RMSD 0.8 Å)



mouse renin
(PID 43%, RMSD 1.6 Å)



HIV protease
(PID 13%, RMSD 2.6 Å)

PID = percentage sequence identity, RMSD = root mean squared deviation

Pairwise sequence alignment

To group sequences into families, we need to be able to answer the question: are two sequences related? This is the most basic sequence analysis task that we will come across. We approach this by first aligning the sequences, and then asking whether the alignment suggests that the sequences are related, or could have occurred by chance.

Pairwise sequence alignment

To group sequences into families, we need to be able to answer the question: are two sequences related? This is the most basic sequence analysis task that we will come across. We approach this by first aligning the sequences, and then asking whether the alignment suggests that the sequences are related, or could have occurred by chance.

Consider this alignment between parts of the human haemoglobin alpha and beta chains:

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
            G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN   GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
```

Here identical positions are shown in the middle with letters, while similar positions are shown with plus signs. To define 'similar' we use a substitution matrix – more on this soon.

Note there are many positions at which the two residues are identical, and many others that we call 'functionally conservative' – for example the D-E pair towards the end (both are negatively charged amino acids).

Pairwise sequence alignment

To group sequences into families, we need to be able to answer the question: are two sequences related? This is the most basic sequence analysis task that we will come across. We approach this by first aligning the sequences, and then asking whether the alignment suggests that the sequences are related, or could have occurred by chance.

Consider this alignment between parts of the human haemoglobin alpha and beta chains:

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
            G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN   GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
```

Here identical positions are shown in the middle with letters, while similar positions are shown with plus signs. To define 'similar' we use a substitution matrix – more on this soon.

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
            ++ ++++H+ KV   + +A  ++          +L+ L+++H+ K
LGB2_LUPLU  NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG
```

In contrast the alignment with leghemoglobin from yellow lupin is much weaker, although these two proteins are evolutionarily related, and have the same 3D structures.

Pairwise sequence alignment

To group sequences into families, we need to be able to answer the question: are two sequences related? This is the most basic sequence analysis task that we will come across. We approach this by first aligning the sequences, and then asking whether the alignment suggests that the sequences are related, or could have occurred by chance.

Consider this alignment between parts of the human haemoglobin alpha and beta chains:

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
            G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN   GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
```

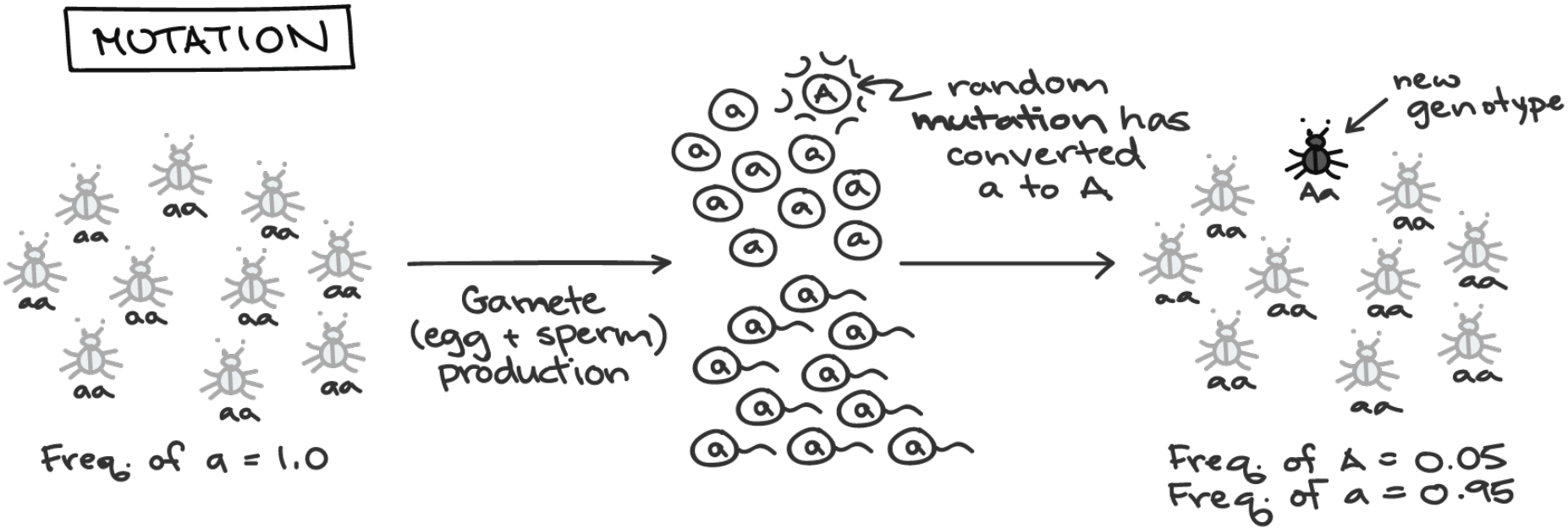
Here identical positions are shown in the middle with letters, while similar positions are shown with plus signs. To define 'similar' we use a substitution matrix – more on this soon.

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
            GS+ + G +   +D L  ++ H+ D+  A +AL D    ++AH+
F11G11.2    GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

Finally this alignment with a nematode glutathione S-transferase homologue has a similar number of identities and similarities, but the structure and function are completely different.

Scoring models

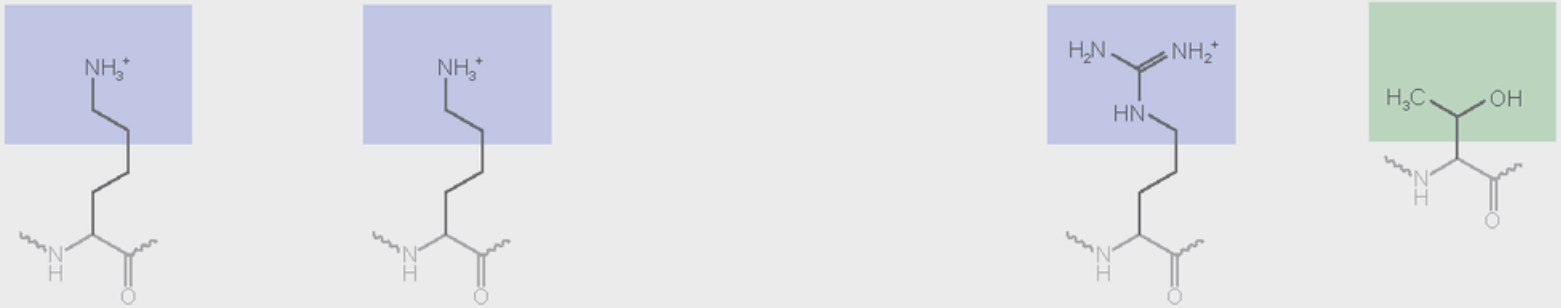
The purpose of a scoring algorithm is to assess the evidence that two sequences have diverged from a common ancestor by a process of mutation and selection.



Scoring models

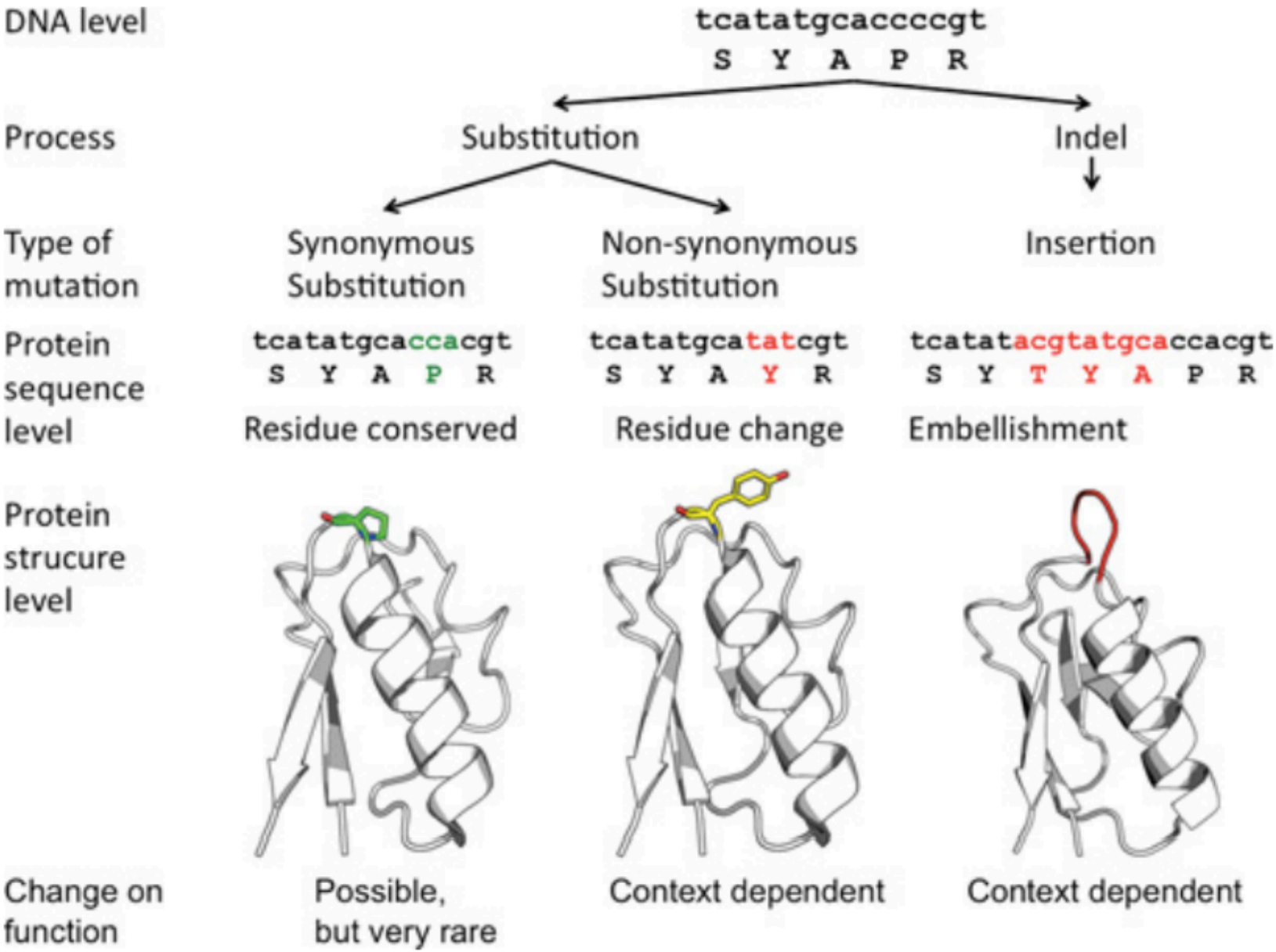
The purpose of a scoring algorithm is to assess the evidence that two sequences have diverged from a common ancestor by a process of mutation and selection.

	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	Lys	Lys	STOP	Arg	Thr



basic
polar

Scoring models

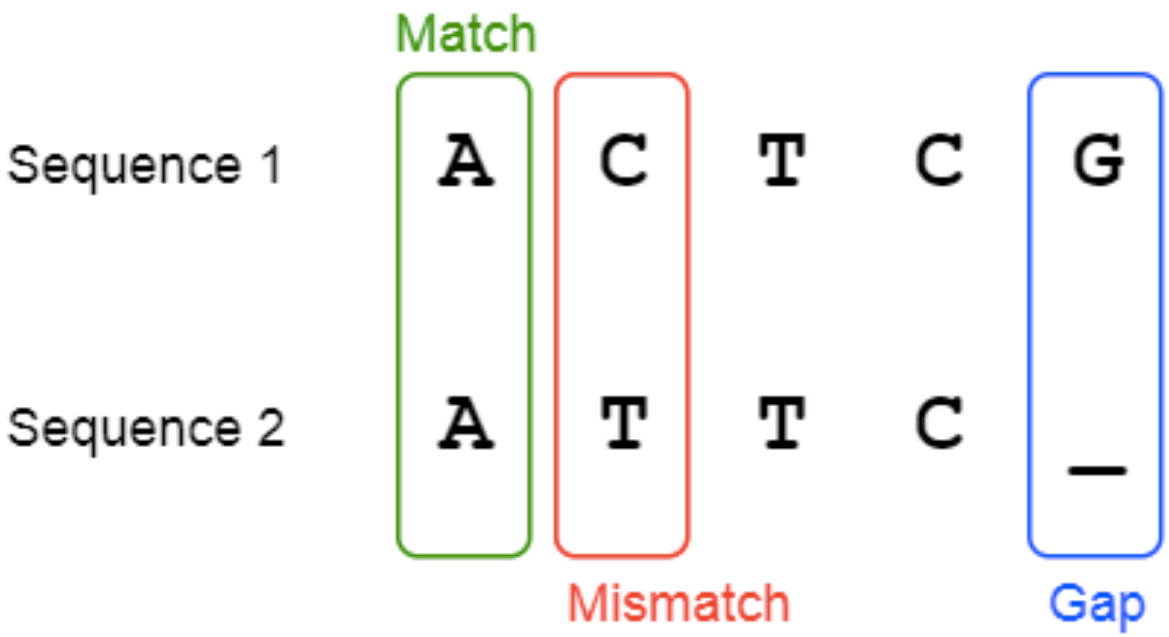


Scoring models

Total score of an alignment will contain a term for each aligned pair of residues, plus terms for each gap introduced. This will correspond to the logarithm of the relative likelihood that the sequences are related, compared to being unrelated.

Scoring models

Total score of an alignment will contain a term for each aligned pair of residues, plus terms for each gap introduced. This will correspond to the logarithm of the relative likelihood that the sequences are related, compared to being unrelated.



For example, S(A, A) = S(T, T) = S(C, C) = +5
S(C, T) = -3
S(gap) = ? -8?

Substitution Matrices

We need score terms for each aligned residue pair. The intuition for proteins alluded to in these slides could yield 210 scoring terms for all possible pairs of amino acids, but it is useful to have a guiding theory for what the scores mean.

Substitution Matrices

We need score terms for each aligned residue pair. The intuition for proteins alluded to in these slides could yield 210 scoring terms for all possible pairs of amino acids, but it is useful to have a guiding theory for what the scores mean.

Consider two protein sequences *x* and *y,* we want to compare the random model:

$$x_1, x_2, x_3, x_4, x_5$$
$$y_1, y_2, y_3, y_4, y_5$$

$$P(x, y|R) = \prod_i q_{x_i} \prod_j q_{y_j}$$

with the match model:

$$P(x, y|M) = \prod_i p_{x_i y_i}$$

Substitution Matrices

We need score terms for each aligned residue pair. The intuition for proteins alluded to in these slides could yield 210 scoring terms for all possible pairs of amino acids, but it is useful to have a guiding theory for what the scores mean.

Consider two protein sequences *x* and *y,* we want to compare the random model:

x_1, x_2, x_3, x_4, x_5
y_1, y_2, y_3, y_4, y_5

$$P(x,y|R) = \prod_i q_{x_i} \prod_j q_{y_j}$$

with the match model:

$$P(x,y|M) = \prod_i p_{x_i y_i}$$

So we want

$$\frac{P(x,y|M)}{P(x,y|R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \prod_i q_{y_i}} = \prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}$$

This leads to the score $S = \sum_i s(x_i, y_i)$, where $s(a,b) = \log\left(\frac{p_{ab}}{q_a q_b}\right)$ is the log likelihood ratio that (a,b) is an aligned or match pair, vs occurred at random.

Substitution Matrices – BLOSUM 50

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**5**	−2	−1	−2	−1	−1	−1	0	−2	−1	−2	−1	−1	−3	−1	1	0	−3	−2	0
R	−2	**7**	−1	−2	−4	1	0	−3	0	−4	−3	3	−2	−3	−3	−1	−1	−3	−1	−3
N	−1	−1	**7**	2	−2	0	0	0	1	−3	−4	0	−2	−4	−2	1	0	−4	−2	−3
D	−2	−2	2	**8**	−4	0	2	−1	−1	−4	−4	−1	−4	−5	−1	0	−1	−5	−3	−4
C	−1	−4	−2	−4	**13**	−3	−3	−3	−3	−2	−2	−3	−2	−2	−4	−1	−1	−5	−3	−1
Q	−1	1	0	0	−3	**7**	2	−2	1	−3	−2	2	0	−4	−1	0	−1	−1	−1	−3
E	−1	0	0	2	−3	2	**6**	−3	0	−4	−3	1	−2	−3	−1	−1	−1	−3	−2	−3
G	0	−3	0	−1	−3	−2	−3	**8**	−2	−4	−4	−2	−3	−4	−2	0	−2	−3	−3	−4
H	−2	0	1	−1	−3	1	0	−2	**10**	−4	−3	0	−1	−1	−2	−1	−2	−3	2	−4
I	−1	−4	−3	−4	−2	−3	−4	−4	−4	**5**	2	−3	2	0	−3	−3	−1	−3	−1	4
L	−2	−3	−4	−4	−2	−2	−3	−4	−3	2	**5**	−3	3	1	−4	−3	−1	−2	−1	1
K	−1	3	0	−1	−3	2	1	−2	0	−3	−3	**6**	−2	−4	−1	0	−1	−3	−2	−3
M	−1	−2	−2	−4	−2	0	−2	−3	−1	2	3	−2	**7**	0	−3	−2	−1	−1	0	1
F	−3	−3	−4	−5	−2	−4	−3	−4	−1	0	1	−4	0	**8**	−4	−3	−2	1	4	−1
P	−1	−3	−2	−1	−4	−1	−1	−2	−2	−3	−4	−1	−3	−4	**10**	−1	−1	−4	−3	−3
S	1	−1	1	0	−1	0	−1	0	−1	−3	−3	0	−2	−3	−1	**5**	2	−4	−2	−2
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	2	**5**	−3	−2	0
W	−3	−3	−4	−5	−5	−1	−3	−3	−3	−3	−2	−3	−1	1	−4	−4	−3	**15**	2	−3
Y	−2	−1	−2	−3	−3	−1	−2	−3	2	−1	−1	−2	0	4	−3	−2	−2	2	**8**	−1
V	0	−3	−3	−4	−1	−3	−3	−4	−4	4	1	−3	1	−1	−3	−2	0	−3	−1	**5**

Henikoff, Steven, and Jorja G. Henikoff. "Amino acid substitution matrices from protein blocks." PNAS 1992.

Alignment algorithms

To construct the BLOSUM matrix a set of aligned, ungapped regions from protein families were assembled and clustered such that two sequences go in the same cluster if their percentage of identical residues exceeds 50%.

Alignment algorithms

To construct the BLOSUM matrix a set of aligned, ungapped regions from protein families were assembled and clustered such that two sequences go in the same cluster if their percentage of identical residues exceeds 50%.

Once we have a scoring model, we then need a way to find the optimal alignment between two sequences. If we allow gaps, then there are

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \simeq \frac{2^{2n}}{\sqrt{\pi n}}$$

possible global alignments between two sequences of length *n*. This means it is not computationally feasible to enumerate all of these, even for moderate values of *n*.

Alignment algorithms

To construct the BLOSUM matrix a set of aligned, ungapped regions from protein families were assembled and clustered such that two sequences go in the same cluster if their percentage of identical residues exceeds 50%.

Once we have a scoring model, we then need a way to find the optimal alignment between two sequences. If we allow gaps, then there are

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \simeq \frac{2^{2n}}{\sqrt{\pi n}}$$

possible global alignments between two sequences of length *n*. This means it is not computationally feasible to enumerate all of these, even for moderate values of *n*.

Luckily, dynamic programming algorithms provide a way of efficiently finding the optimal alignment. The idea is to start from an existing smaller alignment, and evaluate all possible next moves.

Global alignment: Needleman-Wunsch algorithm – align whole sequences.

Local alignment: Smith-Waterman algorithm – align subsequences.

Durbin et al. Biological Sequence Analysis, CUP 1998

Next – multiple sequence alignment

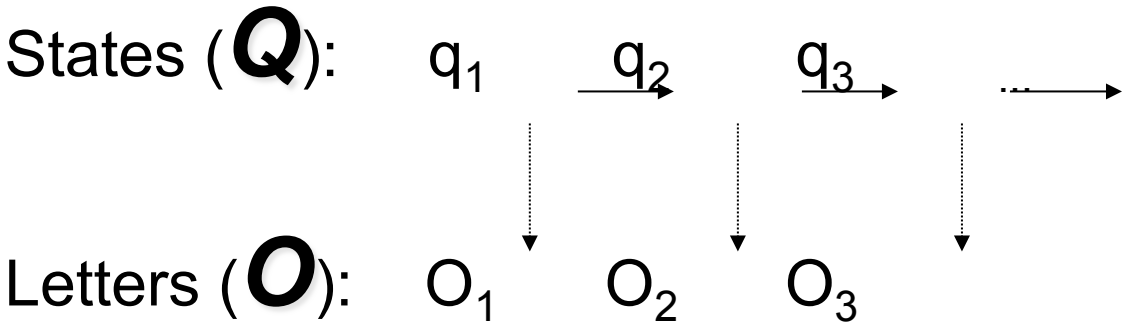
```
Helix                 AAAAAAAAAAAAAAAA   BBBBBBBBBBBBBBBBCCCCCCCCCCC
HBA_HUMAN  ---------VLSPADKTNVKAAWGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF
HBB_HUMAN  --------VHLTPEEKSAVTALWGKV----NVDEVGGEALGRLLVVYPWTQRFFESF
MYG_PHYCA  ---------VLSEGEWQLVLHVWAKVEA--DVAGHGQDILIRLFKSHPETLEKFDRF
GLB3_CHITP ----------LSADQISTVQASFDKVKG------DPVGILYAVFKADPSIMAKFTQF
GLB5_PETMA PIVDTGSVAPLSAAEKTKIRSAWAPVYS--TYETSGVDILVKFFTSTPAAQEFFPKF
LGB2_LUPLU --------GALTESQAALVKSSWEEFNA--NIPKHTHRFFILVLEIAPAAKDLFS-F
GLB1_GLYDI ---------GLSAAQRQVIAATWKDIAGADNGAGVGKDCLIKFLSAHPQMAAVFG-F
Consensus            Ls....  v a W kv .  .    g . L.. f . P .   F  F

Helix          DDDDDDDEEEEEEEEEEEEEEEEEEE            FFFFFFFFFFF
HBA_HUMAN  -DLS-----HGSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL-
HBB_HUMAN  GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL---D--NLKGTFATLSELHCDKL-
MYG_PHYCA  KHLKTEAEMKASEDLKKHGVTVLTALGAILKK----K-GHHEAELKPLAQSHATKH-
GLB3_CHITP AG-KDLESIKGTAPFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG-
GLB5_PETMA KGLTTADQLKKSADVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF-
LGB2_LUPLU LK-GTSEVPQNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-
GLB1_GLYDI SG----AS---DPGVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVGVRHKGYGN
Consensus   .  t    .. . v..Hg kv. a   a...l   d   . a l. l   H  .

Helix       FFGGGGGGGGGGGGGGGGG     HHHHHHHHHHHHHHHHHHHHHHHHHH
HBA_HUMAN  -RVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
HBB_HUMAN  -HVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
MYG_PHYCA  -KIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
GLB3_CHITP --VTHDQLNNFRAGFVSYMKAHT--DFA-GAEAAWGATLDTFFGMIFSKM-------
GLB5_PETMA -QVDPQYFKVLAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
LGB2_LUPLU --VADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
GLB1_GLYDI KHIKAQYFEPLGASLLSAMEHRIGGKMNAAAKDAWAAAYADISGALISGLQS-----
Consensus    v.   f  l . ..  ....    f   . aa. k. .     l sky
```

Hidden Markov Models

A **Hidden Markov Model** (HMM) is a discrete-time finite-state Markov chain coupled with a sequence of letters emitted when the Markov chain visits its states.



The sequence ***O*** of emitted letters is called "the observed sequence" because we often know it while <u>not</u> knowing the state sequence ***Q***, which we call"hidden".

Used extensively throughout computational biology for 'labeling' data – and in particular for classifying and aligning protein sequence data.

What is a hidden Markov model?

Sean R Eddy

NATURE BIOTECHNOLOGY VOLUME 22 NUMBER 10 OCTOBER 2004

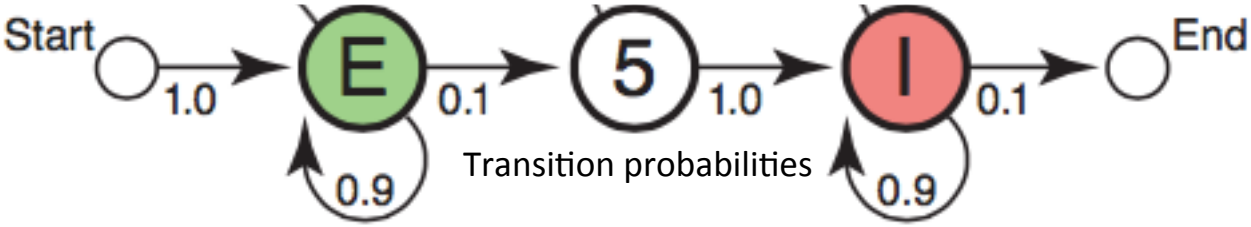
As a simple example, imagine the following 5′ splice-site recognition problem. Assume we are given a DNA sequence that begins in an exon, contains one 5′ splice site and ends in an intron. The problem is to identify where the switch from exon to intron occurred—where the 5′ splice site (5′SS) is.

What is a hidden Markov model?

Sean R Eddy

NATURE BIOTECHNOLOGY VOLUME 22 NUMBER 10 OCTOBER 2004

As a simple example, imagine the following 5′ splice-site recognition problem. Assume we are given a DNA sequence that begins in an exon, contains one 5′ splice site and ends in an intron. The problem is to identify where the switch from exon to intron occurred—where the 5′ splice site (5′SS) is.



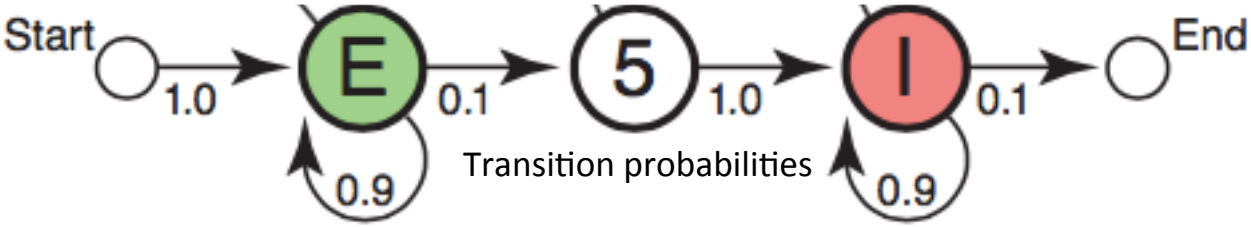
HMM with three states – exon, 5′SS and intron, and the probabilities of moving between these states.

What is a hidden Markov model?

Sean R Eddy

NATURE BIOTECHNOLOGY VOLUME 22 NUMBER 10 OCTOBER 2004

As a simple example, imagine the following 5′ splice-site recognition problem. Assume we are given a DNA sequence that begins in an exon, contains one 5′ splice site and ends in an intron. The problem is to identify where the switch from exon to intron occurred—where the 5′ splice site (5′SS) is.



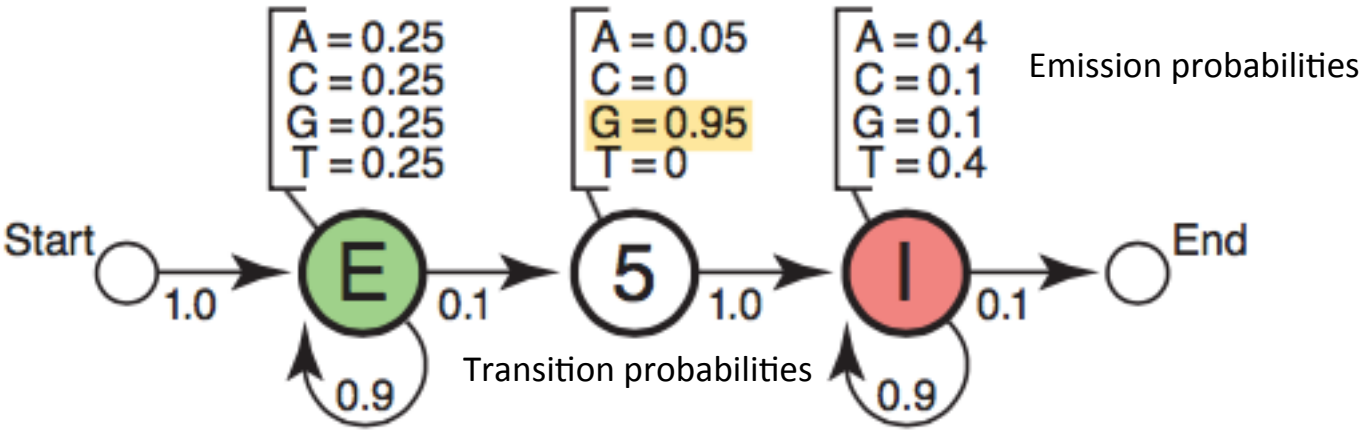
HMM with three states – exon, 5′SS and intron, and the probabilities of moving between these states. To estimate which state each element of the DNA sequence is in, we need information about the statistics of exon, splice site and intron sequences – these are the emission probabilities.

What is a hidden Markov model?

Sean R Eddy

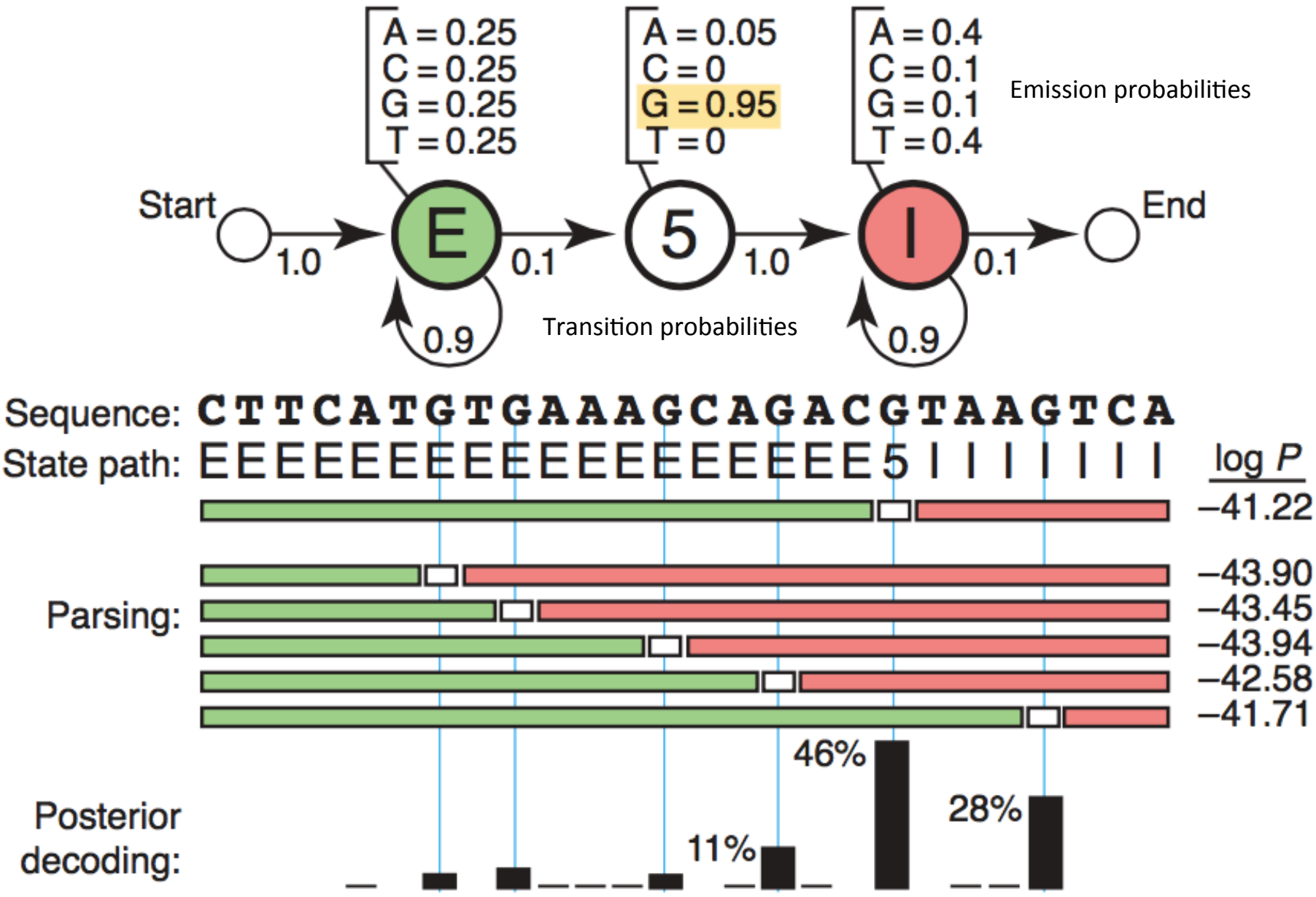
NATURE BIOTECHNOLOGY VOLUME 22 NUMBER 10 OCTOBER 2004

As a simple example, imagine the following 5′ splice-site recognition problem. Assume we are given a DNA sequence that begins in an exon, contains one 5′ splice site and ends in an intron. The problem is to identify where the switch from exon to intron occurred—where the 5′ splice site (5′SS) is.



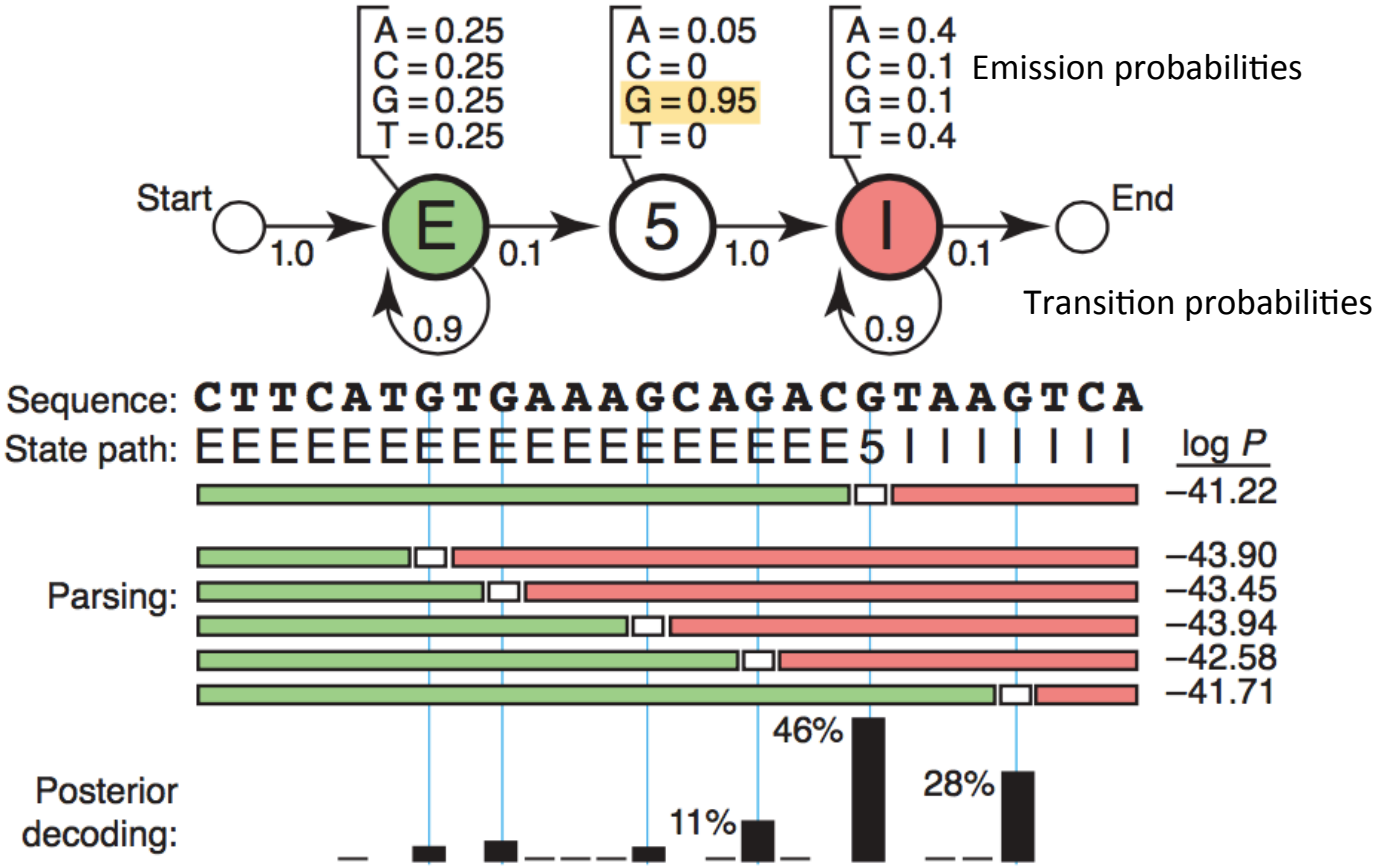
HMM with three states – exon, 5′SS and intron, and the probabilities of moving between these states. To estimate which state each element of the DNA sequence is in, we need information about the statistics of exon, splice site and intron sequences – these are the emission probabilities.

Eddy, Nature Biotech, 2004



A toy HMM for 5′ splice site recognition.

Eddy, Nature Biotech, 2004



The probability P(S,π|HMM,θ) that an HMM with parameters θ generates a state path π and an observed sequence S is the product of all the emission probabilities and transition probabilities that were used.

Here for the 26 nucleotide sequence, there are 27 transitions and 26 emissions – the log of the product of all 53 probabilities yields log P(S,π|HMM,θ) = –41.22.

Eddy, Nature Biotech, 2004

EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam 31.0 (March 2017, 16712 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM ENTRY	View Pfam annotation and alignments
VIEW A CLAN	See groups of related entries
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	enter any accession or ID [Go] [Example]

Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.

Or view the help pages for more information

pfam.xfam.org

```
GLB1_PARCH/17-123          QKKIVRKTWHQ.LMR......NKTSFVTDLFIRIFAYDPAAQNKFPQM.AGMS.ASQLRSSRQMQAHAIRVSSIMSEYIE
HMP_ECOLI/6-103            TIATVKATIPL.LVE......TGPKLTAHFYDRMFTHNPELKEIFN.M...........SNQRNGDQREALFNAIAAYAS
HMP_ECOLI/6-103 (SS)       HHHHHHHHHHH.HHT......THHHHHHHHHHHHHHH-GGGGGTS-.-...........-TT--SHHHHHHHHHHHHHHG
BAHG_VITST/6-103           TINIIKATVPV.LKE......HGVTITTTFYKNLFAKHPEVRPLFD.M.......GRQESLEQPKALAMTVLAAAQNIEN
BAHG_VITST/6-103 (SS)      HHHHHHHHHHH.HHH......THHHHHHHHHHHHHHH-GGGGGGG-.-........-SS-TTHHHTTHHHHHHHHHHTTTG
GLB1_CALSO/5-105           DIKNVQDTWGK.LYD.....QWDAVHASKFYNKLFKDSEDISEAFVKA........GTGSGIAMKRQALVFGAILQEFVA
GLB2_CALSO/5-108           DIAAVQTSWRR.CYCS...WDNE..DGLKFYQTLFDSNSKIRHAFESA........GATNDTEMEKQANLFGLMMTQFID
A0A0E8VH55_MYCTX/42-136    DALRVLQNAFK.L........DDPELVRRFYAHWFALDASVRDLFP.P..........DMGAQRAAFGQALHWVYGELVA
GLB1B_ANATR/17-122         QKDLLRLSWGV.LSV......DMEGTGLMLMANLFKTSSAARTKFARL.GDVS...AGKDNSKLRGHSITLMYALQNFID
GLB_BUSCA/6-116            QKTALKESWKV.LGAD...GPTMMKNGSLLFGLLFKTYPDTKKHFKHF..DDATFAAMDTTGVGKAHGVAVFSGLGSMIC
GLB_CERRH/6-116            SKSALASSWKT.LAKD...AATIQNNGATLFSLLFKQFPDTRNYFTHF.GNMS.DAEMKTTGVGKAHSMAVFAGIGSMID
HBP2_CASGL/13-122          QEALVVKSWSA.MKP......NAGELGLKFFLKIFEIAPSAQKLFS.FLKD..SNVPLERNPKLKSHAMSVFLMTCESAV
HBP1_CASGL/7-116           QEALLKQSWEV.LKQ......NIPAHSLRLFALIIEAAPESKYVFS.FLKD..SNEIPENNPKLKAHAAVIFKTICESAT
LGB1_LUPLU/8-116           QVALVKSSFEE.FNA......NIPKNTHRFFTLVLEIAPGAKDLFS.FLKGSS..EVPQNNPDLQAHAGKVFKLTYEAAI
LGB1_MEDSA/7-112           QEALVNSSWEA.FKQ......NLPRYSVFFYTVVLEKAPAAKGLFS.F...LKNSAEVQDSPQLQAHAEKVFGLVRDSAV
GLB1_GLYDI/6-111           QRQVIAATWKD.IAGA....DNGAGVGKDCLIKFLSAHPQMAAVFG.F.SG.......ASDPGVAALGAKVLAQIGVAVS
GLB1_GLYDI/6-111 (SS)      HHHHHHHHHHH.HHTT....TTTHHHHHHHHHHHHHH-CCHHHHHT.-.SS.......TT-HHHHHHHHHHHHHHHHHHH
GLBP1_GLYDI/7-115          QVAALKASWPE.VSAG....DGGAQLGLEMFTKYFHENPQMMFIFG.YSGR.T..EALKHSSKLQHHGKVIIDQIGKAVA
GLB_APLJU/6-113            DAGLLAQSWAP.VFA......NSDANGASFLVALFTQFPESANFFNDF.KG.KSLADIQASPKLRDVSSRIFARLNEFVS
GLB73_CHITH/23-129         EASLVQSSWKA.VSH..........NEVDILAAVFAAYPDIQAKFPQF.AG.KDLASIKDTGAFATHATRIVSFLSEVIA
GLB6_CHITH/22-128          QADLVKKTWST.VKF..........NEVDILYAVFKAYPDIMAKFPQF.AG.KDLDSIKDSAAFATHATRIVSFLSEVIS
GLB3_CHITH/20-120          QISTVQASFDK.VKG..........DPVGILYAVFKADPSIMAKFTQF.AG.KDLESIKGTAPFEIHANRIVGFFSKIIG
GLB3_CHITH/20-120 (SS)     HHHHHHHHHHT.TTT..........-HHHHHHHHHHH-HHHHTT-TTT.TT.S-HHHHTTSHHHHHHHHHHHHHHHHHHH
GLB1_CHITH/21-126          QIAAAKASWNT.VKN..........NQVDILYAVFKANPDIQTAFSQF.AG.KDLDSIKGTPDFSKHAGRVVGLFSEVMD
GLB10_CHITH/11-117         EVEQVQATWKA.VSH..........DEVEILYTVFKAHPDIMAKFPKF.AG.KDLEAIKDTADFAVHASRIIGFFGEYVT
GLB8_CHITH/9-115           QLALFKSSWNT.VKH..........NEVDILYAVFKANPDIQAKFPQF.AG.KDLDSIKDSADFAVHSGRIVGFFSEVIG
GLB2_CHITH/22-128          EASLVRGSWAQ.VKH..........SEVDILYYIFKANPDIMAKFPQF.AG.KDLETLKGTGQFATHAGRIVGFVSEIVA
HBF1_URECA/7-113           QIKAIQDHWFLNIKG......CLQAAADSIFFKYLTAYPGDLAFFHKF.SS.VPLYGLRSNPAYKAQTLTVINYLDKVVD
HBF1_URECA/7-113 (SS)      HHHHHHHHHHHTGG......GHHHHHHHHHHHHHHSGGGGGG-TTT.TT.S-GGGGGG-HHHHHHHHHHHHHHHHHHH
GLBT_CHITH/13-116          QVAAVKGDWEK.IKG..........SGVEILYFFLNKFPGNFPMFKKL.GN..DLAAAKGTAEFKDQADKIIAFLQGVIE
MYG_CYPCA/3-108            DAELVLKCWGG.VEA......DFEGTGGEVLTRLFKQHPETQKLFPKF.VG.IASNELAGNAAVKAHGATVLKKLGELLK
MYG_ALLMI/7-113            EWKHVLDIWTK.VES......KLPEHGHEVIIRLLQEHPETQERFEKF.KHMKTADEMKSSEKMKQHGNTVFTALGNILK
MYG_GALGA/3-108            DWDKVNSVWSA.MEA......NITAVGQNILLRLFEQYPESQSYFPKL.KN.KSLGELKDTADIKAQADTVLKALGNIVK
MYG_HETPO/3-108            EWEHVNKVWAV.VEP......DIPAVGLAILLRLFKEHKETKDLFPKF.KE.IPVQQLGNNEDLRKHGVTVLRALGNILK
HBAM_LITCT/6-102           EKSAVASLWEK.IAP......QTNKLGAESMERLFKNHPETKSFFSRF..DIS.....PGSQDLLTHGGKIFGALGEAIK
HBAD_ERYML/6-106           DRRLLQASVGK.LGC......RLEDIGADALNRLLITFPQSKTYFSHF..NLS.....PGSKDIIHQGEKVGKALDSALK
HBAD_PASMO/6-106           DKKLIQQIWGK.LGG......AEEEIGADALWRMFHSYPSTKTYFPHF..DLS.....QGSDQIRGHGKKVVAALSNAIK
HBAT_HORSE/7-107           DRATVRALWKK.MGS......NVGVYATEALERMFLGFPSTTTYFLHL..DLS.....LGSTQVKAHGQKVADALTLAVE
HBA1_BOSMU/6-106           DKGNVKAAWGK.VGG......HAAEYGAEALERMFLSFPTTKTYFPHF..DLS.....QGSAQVKGHGAKVAAALTKAVE
HBA1_IGUIG/6-106           DKNHIRAIWGH.VDN......NPEAFGVEALTRLFLAYPATKTYFAHF..DLN.....PGSAQIKAHGKKVVDALTQAVN
HBA1_PLEWA/7-107           DKHNVKAIWDH.VKG......HEEAIGAEALYRMFCCMPTTRIYFP.A.KDLS.....ERSSYLHSHGKKVVGALTNAVA
HBA1_XENBO/7-107           DKKHIKAIMPS.IAA......HGDKFGGEALYRMFLVNPKTKTYFPTF..DFH.....HNSKQISAHGKKVVDALNEASN
HBAZ_CAPHI/7-107           ERTIILSLWSK.IST......QADVIGTETLERLFSCYPQAKTYFPHF..DLH.....SGSAQLRAHGSKVVAAVGDAVK
HBA3_PLEWA/7-107           EKALVVGLCGK.ISG......HCDALGGEALDRLFASFGQTRTYFSHF..DLS.....PGSADVKRHGGKVLSAIGEAAK
HBA_CATCL/6-107            DKADVKIAWAK.ISP......RADEIGAEALGRMLTVYPQTKTYFAHW.ADLS.....PGSGPVKHGKKVIMGAIGDAVT
HBB_HETPO/7-106            ELHEITTTWKS.I........DKHSLGAKALARMFIVYPWTTRYFGNL.KEFT.....ACSYGVKEHAKKVTGALGVAVT
```

Family: *Globin* (PF00042)

66 architectures | 4699 sequences | 12 interactions | 1761 species | 2557 structures

Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures

Jump to...
enter ID/acc Go

Summary: Globin

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

Wikipedia: Globin | Pfam | InterPro

This is the Wikipedia entry entitled "Globin". More...

Globin

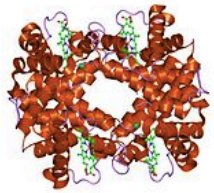
Edit Wikipedia article

Not to be confused with globulin or globular protein.

The **globins** are a superfamily of heme-containing globular proteins, involved in binding and/or transporting oxygen. These proteins all incorporate the globin fold, a series of eight alpha helical segments. Two prominent members include myoglobin and hemoglobin. Both of these proteins reversibly bind oxygen via a heme prosthetic group. They are widely distributed in many organisms.[2]

Contents [hide]

Globin family



the Structure of deoxyhemoglobin Rothschild 37 beta Trp----Arg: a mutation that creates an intersubunit chloride-binding site.[1]

Identifiers	
Symbol	Globin
Pfam	PF00042
Pfam clan	CL0090
InterPro	IPR000971
PROSITE	PS01033
SCOP	1hba
SUPERFAMILY	1hba
CDD	cd01067

Structure

Globin superfamily members share a common three-dimensional fold.[3] This 'globin fold' typically consists of eight alpha helices, although some proteins have additional helix extensions at their termini.[4] Since the globin fold contains only helices, it is classified as an all-alpha protein fold.

The globin fold is found in its namesake globin families as well as in phycocyanins. The globin fold was thus the first protein fold discovered (myoglobin was the first protein whose structure was solved).

Helix packing

The eight helices of the globin fold core share significant nonlocal structure, unlike other structural motifs in which amino acids close to each other in primary sequence are also close in space. The helices pack together at an average angle of about 50 degrees, significantly steeper than other helical packings such as the helix bundle. The exact angle of helix packing depends on the sequence of the protein, because packing is mediated by the sterics and hydrophobic interactions of the amino acid side chains near the helix interfaces.

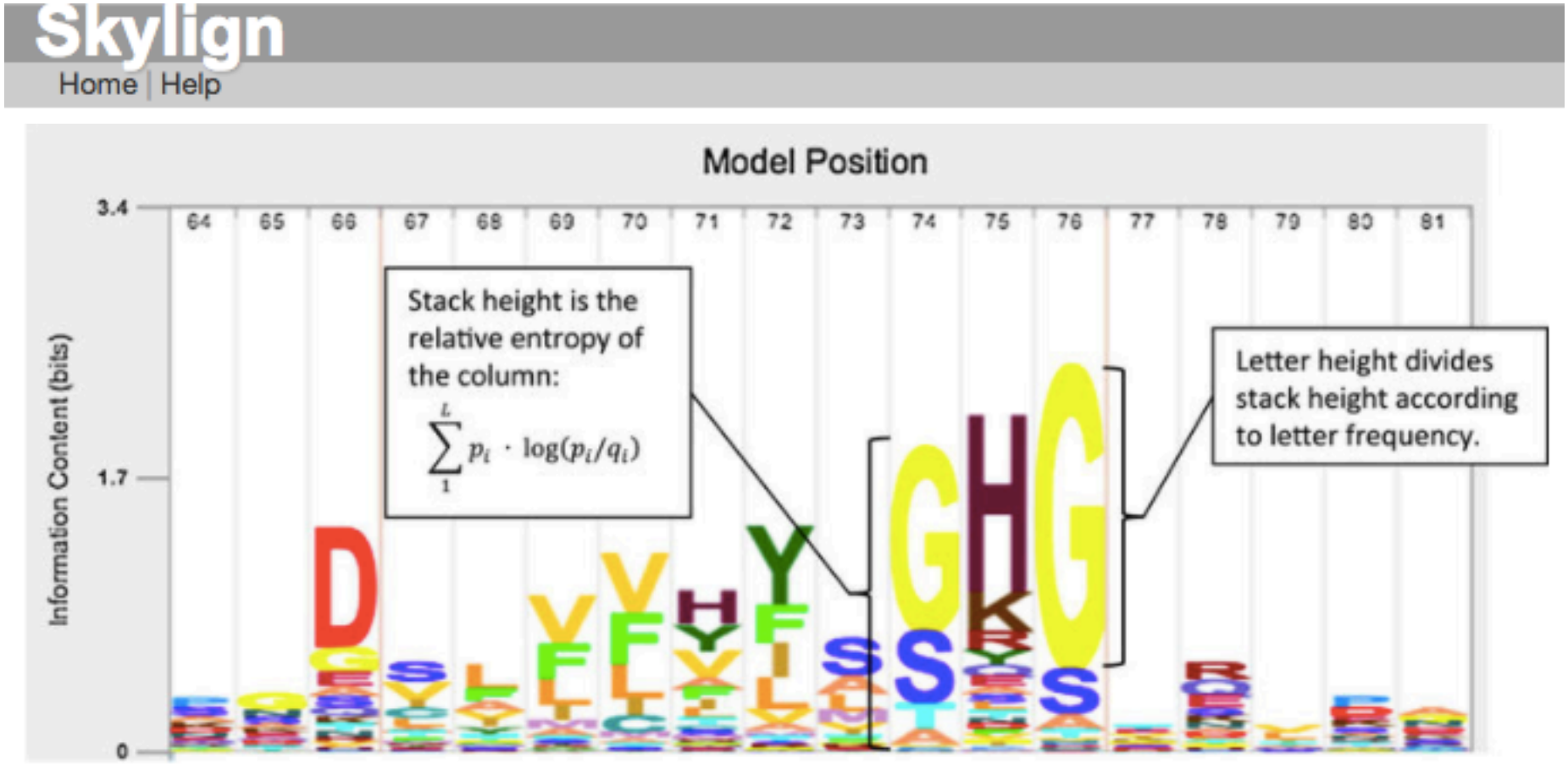


Figure 1

Example profile logo. This logo shows positions 64 to 81 of the Peptidase_C14 profile HMM from Pfam (PF00656, Pfam 27.0), produced using Skylign. The profile HMM was constructed using *hmmbuild* (default parameters) from HMMER 3.1 on the Pfam seed alignment. One of the active sites of this Caspase domain is found at position 75. This site is invariant in active peptidases, but not in this profile HMM. This is the result of two forces: (1) the Pfam alignment includes non-peptidase homologs, which do not contain a Histidine at this position, and (2) HMMER intentionally drives down the information content per position (using an approach called entropy weighting [12]) to increase sensitivity to remote homologs.

BMC Bioinformatics 2014 **15**:7

https://doi.org/10.1186/1471-2105-15-7

EMBL-EBI

Pfam

keyword search Go

Family: *Globin* (PF00042)

66 architectures | 4699 sequences | 12 interactions | 1761 species | 2557 structures

Format an alignment

	Seed (73)	Full (4699)	Representative proteomes				UniProt (14977)	NCBI (21099)	Meta (34)
			RP15 (1041)	RP35 (2260)	RP55 (4063)	RP75 (5877)			
Alignment:	◉	○	○	○	○	○	○	○	○

Format: Selex

Order: ◉ Tree ○ Alphabetical

Sequence: ◉ Inserts lower case ○ All upper case

Gaps: Gaps as "." or "-" (mixed)

Download/view: ◉ Download ○ View

Generate

20 July 1973, Volume 181, Number 4096

SCIENCE

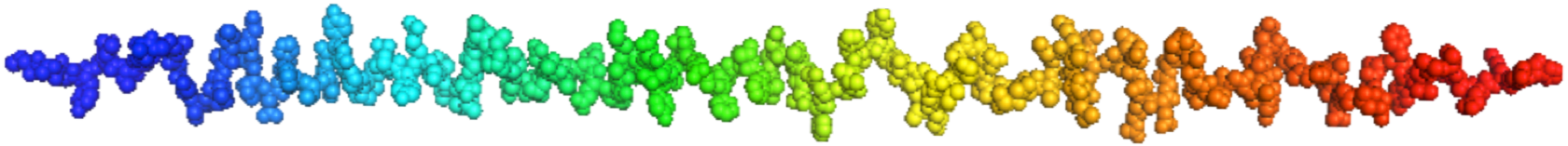
Principles that Govern the Folding of Protein Chains

Christian B. Anfinsen

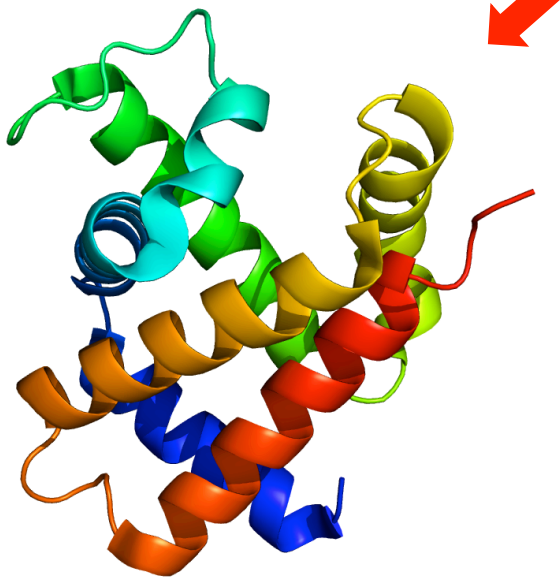
Empirical considerations of the large amount of data now available on correlations between sequence and three-dimensional structure (*48*), together with an increasing sophistication in the theoretical treatment of the energetics of polypeptide chain folding (*49*) are beginning to make more realistic the idea of the a priori prediction of protein conformation. It is certain that major advances in the understanding of cellular organization, and of the causes and control of abnormalities in such organization, will occur when we can predict, in advance, the three-dimensional phenotypic consequences of a genetic message.

Anfinsen CB (1973). "Principles that govern the folding of protein chains". Science 181 (4096): 223–230.

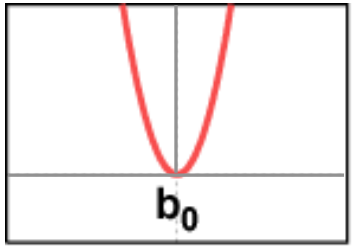
What makes this problem hard?



Polymer self-assembles into unique 3D structure

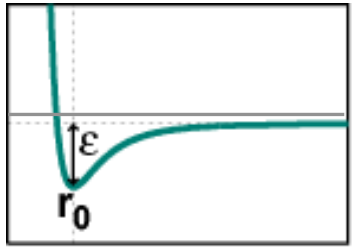


Potential function for MD



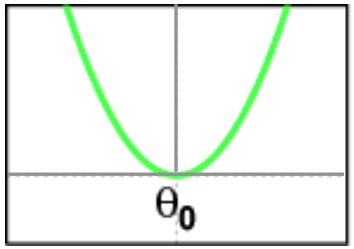
Bond

$$\sum_{i}^{bonds} K_{b,i}(b_i - b_{0,i})^2$$



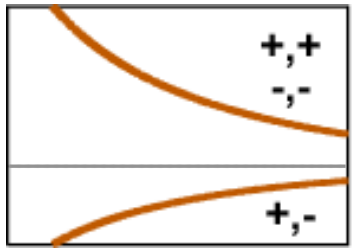
van der Waals

$$\sum_{pairs\cdot i,j} [\varepsilon_{ij}\left(\frac{r_{0,ij}}{r_{ij}}\right)^{12} - 2\varepsilon_{ij}\left(\frac{r_{0,ij}}{r_{ij}}\right)^{6}]$$



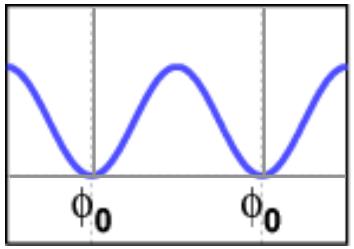
Angle

$$\sum_{i}^{\substack{bond\\angles}} K_{\theta,i}(\theta_i - \theta_{0,i})^2$$



Electrostatic

$$332 \sum_{pairs\cdot i,j} \left(\frac{q_i q_j}{r_{ij}}\right)$$



Dihedral

$$\sum_{i}^{\substack{torsion\\angles}} K_{\phi,i}\{1 - \cos[n_i(\phi_i - \phi_{0,i})]\}$$

$$F = -\frac{\partial U}{\partial x}$$

$$F = ma$$

$$a = \frac{v_3 - v_2}{\partial t}$$

$$v = \frac{x_3 - x_2}{\partial t}$$

$$E = U + K$$

$$\partial t = 2 \text{ fs}$$

Evaluate forces and perform integration for every atom

Each picosecond of simulation time requires 500 iterations of cycle

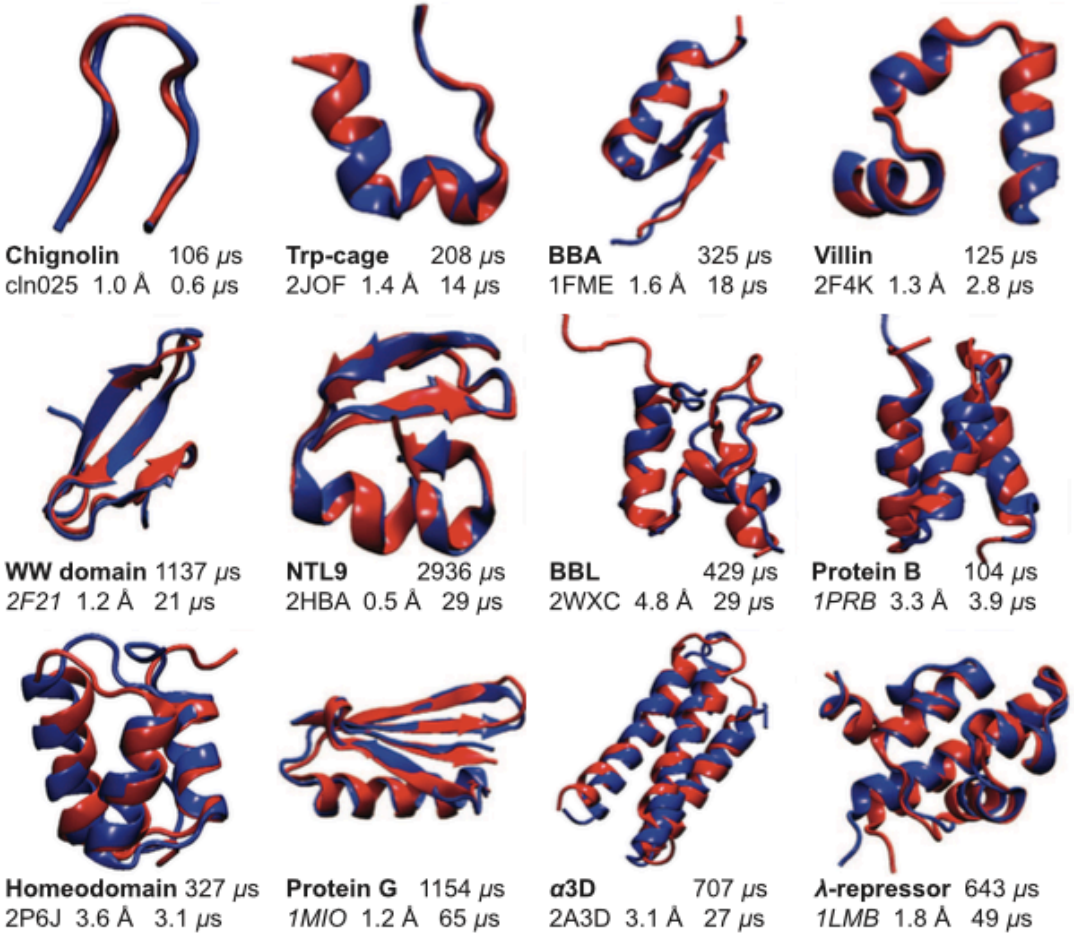
E.g. w/ 50,000 atoms, each ps (10^{-12} s) involves 25,000,000 evaluations

How Fast-Folding Proteins Fold

Kresten Lindorff-Larsen,[1*†] Stefano Piana,[1*†] Ron O. Dror,[1] David E. Shaw[1,2†]

An outstanding challenge in the field of molecular biology has been to understand the process by which proteins fold into their characteristic three-dimensional structures. Here, we report the results of atomic-level molecular dynamics simulations, over periods ranging between 100 μs and 1 ms, that reveal a set of common principles underlying the folding of 12 structurally diverse

Our research group has developed a specialized supercomputer, called Anton, which greatly accelerates the execution of atomistic molecular dynamics (MD) simulations. In addition, we recently modified the CHARMM force field in an effort to make it more easily transferable among different protein classes.



Lindorff-Larsen, Kresten, et al. "How fast-folding proteins fold." Science 334.6055 (2011): 517-520.

Rosetta – an algorithm that introduced assembly of sequence similar fragments was a major step forward in computational approaches to the protein folding problem.

Swiftly followed by 'folding at home' and the

Crystal structure of a monomeric retroviral protease solved by protein folding game players

Firas Khatib[1], Frank DiMaio[1], Foldit Contenders Group, Foldit Void Crushers Group, Seth Cooper[2], Maciej Kazmierczyk[3], Miroslaw Gilski[3,4], Szymon Krzywda[3], Helena Zabranska[5], Iva Pichova[5], James Thompson[1], Zoran Popović[2], Mariusz Jaskolski[3,4] & David Baker[1,6]

Following the failure of a wide range of attempts to solve the crystal structure of M-PMV retroviral protease by molecular replacement, we challenged players of the protein folding game Foldit to produce accurate models of the protein. Remarkably, Foldit players were able to generate models of sufficient quality for successful molecular replacement and subsequent structure determination. The refined structure provides new insights for the design of antiretroviral drugs.

Khatib, Firas, et al. Nature structural & molecular biology 18.10 (2011): 1175-1177.

WHAT IF...

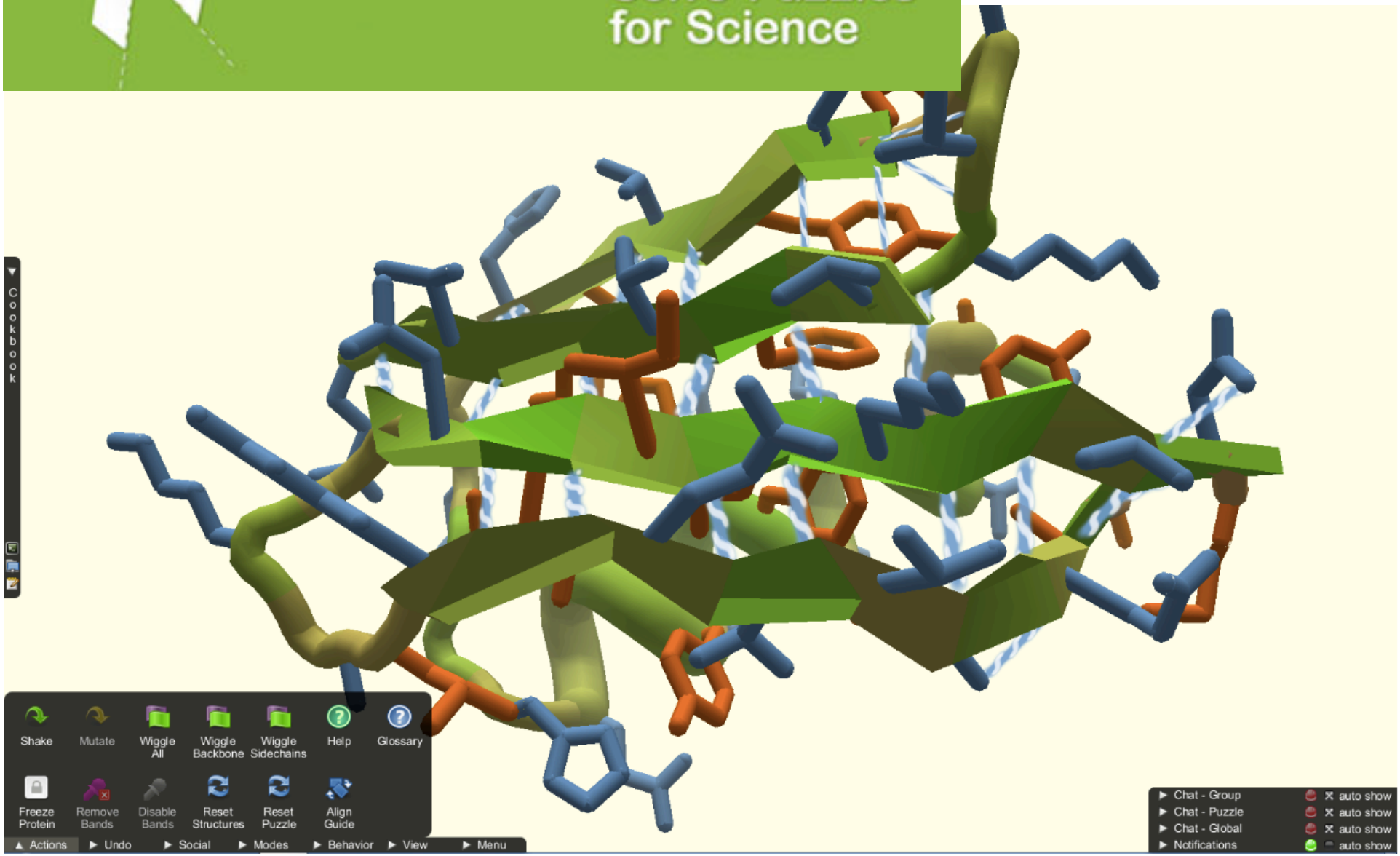
you could help find a cure?

PLAY VIDEO

Help Stanford University scientists studying Alzheimer's, Huntington's, Parkinson's, and many cancers by simply running a piece of software on your computer.

foldit

Solve Puzzles for Science



How to extract useful information from protein sequences?

We can think of protein sequences from different species as samples from a probability distribution…..

Species

```
                                  20        30        40        50        60        70        80        90        100
Racoon                            VEEVGGEALGRLLVVYPWTQRFFESFGDLSSADAIMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Ring-tailed coati                 VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAIMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Red fox                           VDEVGGEALGRLLIVYPWTQRFFDSFGDLSTPDAVMGNAKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Weddell seal                      VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPNAIMSNPKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDQLHVDPEI
Harbor seal                       VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSADAIMGNPKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Eurasion badger                   VDEVGGEALGRLLVVYPWTQRYFDSFGDLSTPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Red panda                         VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Ferret                            VDEVGGETLGRLLVVYPWTQRFFDSFGDLSSPDAVMSNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Fur seal                          VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPEI
River otter                       VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Beach marten                      VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
European mink                     VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Ratel                             VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Colobus monkey                    VDAVGGEALGRLLVVYPWTQRFFESFGDLSSPAAVMGNPKVKAHGKKVLGAFSDGLAHLDSLKGTFSQLSELHCDKLHVDPEI
Lowland gorilla                   VDAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFSQLSELHCDKLHVDPEI
Human                             VDAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFSQLSELHCDKLHVDPEI
Spider monkey                     VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPAAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
Brown spider monkey               VDEVGGEALGRLLVVYPWTQRFFESFGALSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
Tamarin                           VDEVGGEALGRLLVVYPWTQRFFESFGALSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
Moustached tamarin                VDEVGGEALGRLLVVYPWTQRFFESFGALSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
Capuchin                          VDEVGGEALGRLLVVYPWTQRFFDSFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLTHLDNLKGTFAQLSELHCDKLHVDPEI
White Capuchin                    VDEVGGEALGRLLVVYPWTQRFFDSFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLTHLDNLKGTFAQLSELHCDKLHVDPEI
white sapajou                     VDEVGGEALGRLLVVYPWTQRFFDSFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLTHLDNLKGTFAQLSELHCDKLHVDPEI
squrrel monkey                    VEDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
red colobus                       VDEVGGEALGRLLVVYPWTQRFFDSFGDLSTADAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
Collared titi                     VXEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMSNXKVKAHGKKVLGAFSDGLAHLDNLKSTFAQLSELHCDKLHVDPEI
Common marmoset                   VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLTHLDNLKGTFAHLSELHCDKLHVDPEI
Howler monkey                     VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMHNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
Black spider monkey               VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
Goeldi's marmoset                 VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMNNPKVKAHGKKVLGAFSDGLTHLDNLKGTFAQLSELHCDKLHVDPEI
Brown tamarin                     VEEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCNKLHVDPEI
black and red tamarin             VEEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMNNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
moustached tamarin                VEEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMNNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
papio sphinx                      VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPEI
olive baboon                      VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPAAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPEI
gelada baboon                     VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPAAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPEI
green monkey                      VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
sootey mangabey                   VDEVGGEALGRLLVVYPWTQRFFESFGDLSNPDAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPEI
crab eating macaque               VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPEI
rhesus macaque                    VDEVGGEALGRLLLVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPEI
western colobus monkey            VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
```

We want to use this data to parameterize a probability model for the protein sequence.

$P(A_1,\ldots,A_L)$ = Probability that a sequence produces a folded, functional hemoglobin molecule.

First order models

$P(A_1,\ldots,A_L) =$ Probability that a sequence produces a folded, functional hemoglobin molecule.

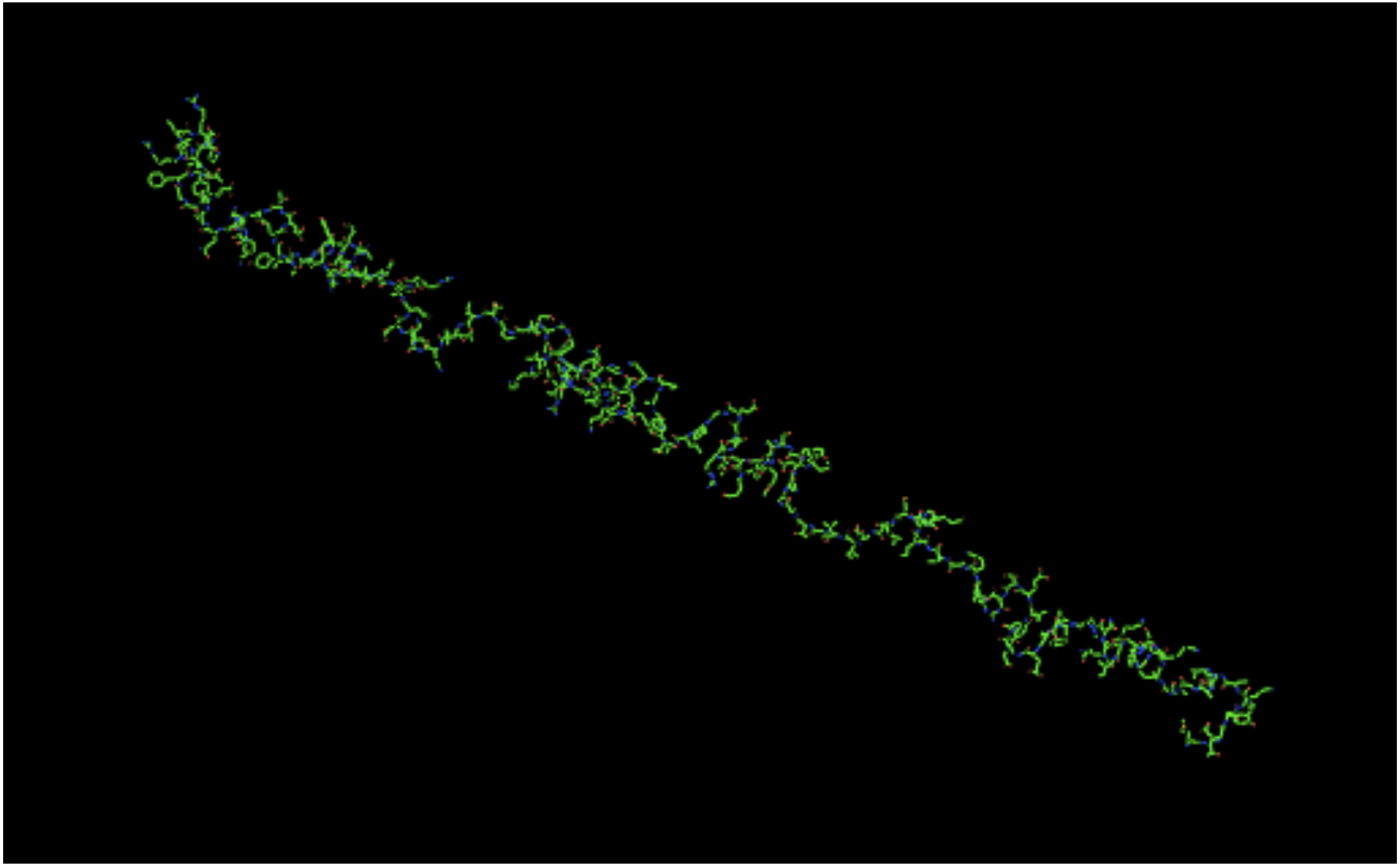
To start, we might ask that the model matches the distribution of amino acids seen at each sequence position in the data (the set of available sequences).

$$P_i(A_i) = \sum_{\{A_k | k \neq i\}} P(A_1,\ldots,A_L) = f_i(A_i)$$

Hidden Markov Models are often used to model protein sequence evolution. Different sequence positions (variables) are assumed to evolve independently of one another – there are no interactions.

Does this model tell us anything about protein structure and function?

Those residues close in sequence will be close in 3D structure….



but their conservation level doesn’t provide much information about the 3D structure.

First order models

$P(A_1,\ldots,A_L) =$ Probability that a sequence produces a folded, functional hemoglobin molecule.

To start, we might ask that the model matches the distribution of amino acids seen at each sequence position in the data (the set of available sequences).

$$P_i(A_i) = \sum_{\{A_k | k \neq i\}} P(A_1,\ldots,A_L) = f_i(A_i)$$

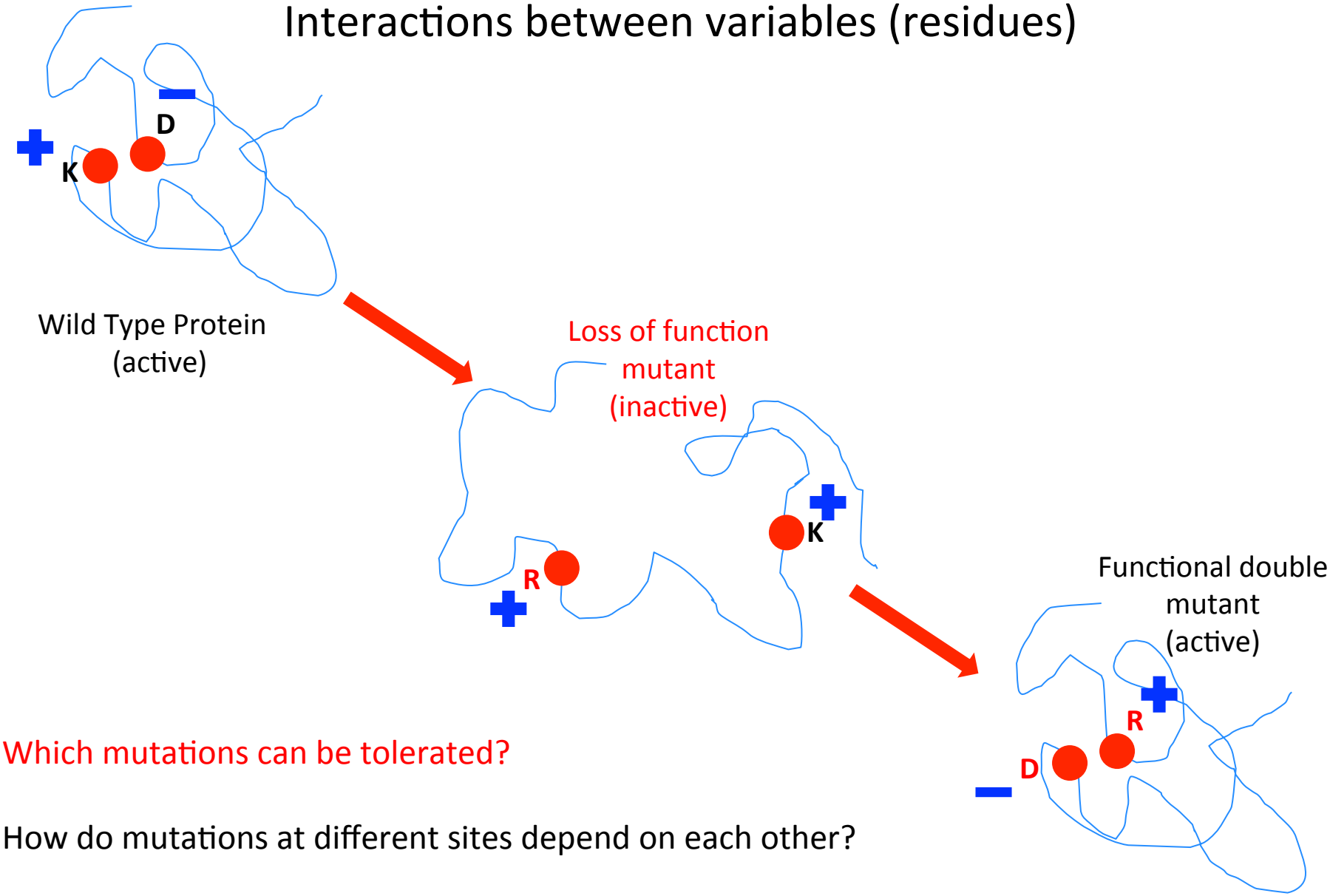
Hidden Markov Models are often used to model protein sequence evolution. Different sequence positions (variables) are assumed to evolve independently of one another – there are no interactions.

Does it tell us anything about the 3D protein structure…….NOT MUCH

How can we get to 3D protein structure and/or function?

INTERACTIONS!!!

Interactions between variables (residues)

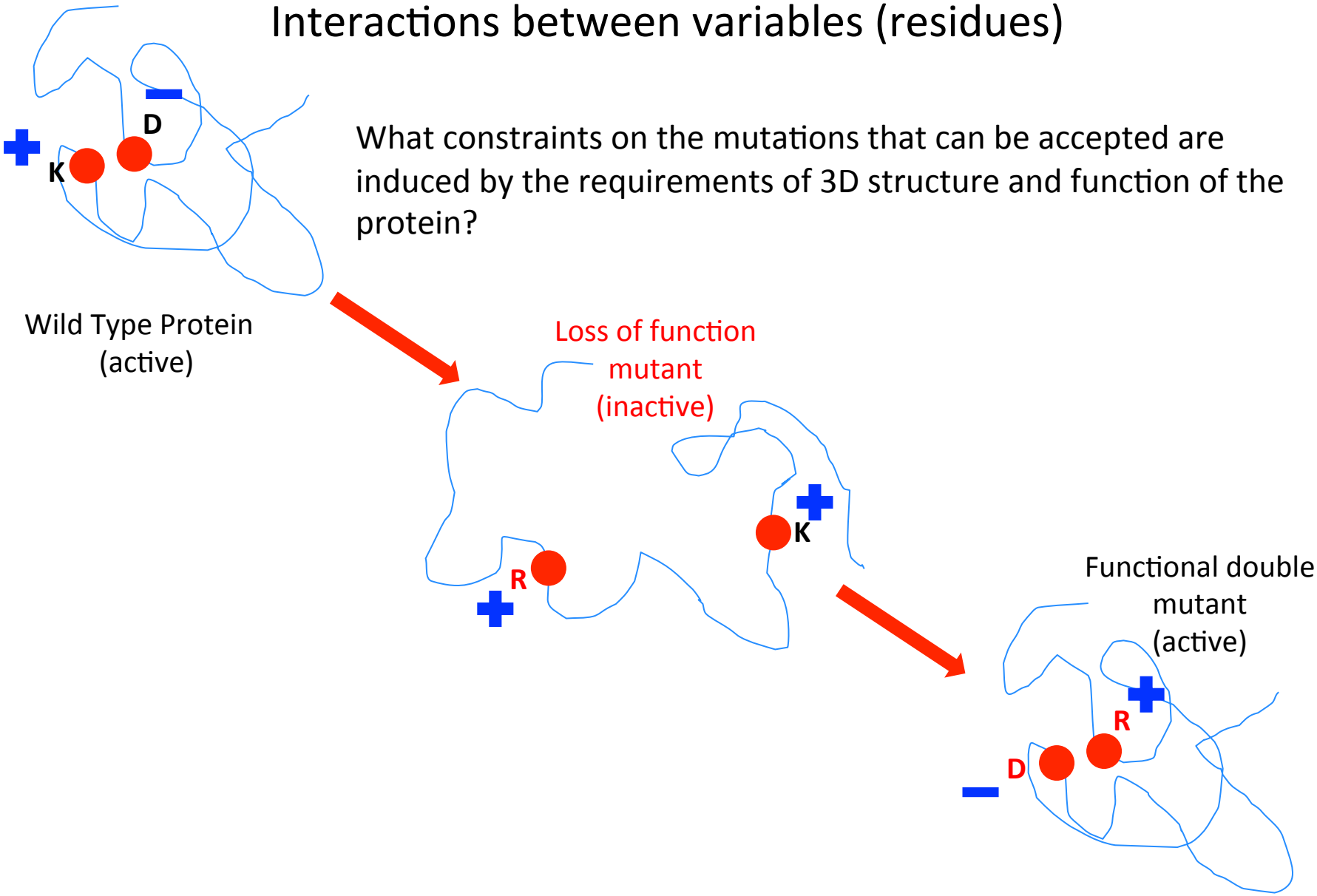


Which mutations can be tolerated?

How do mutations at different sites depend on each other?

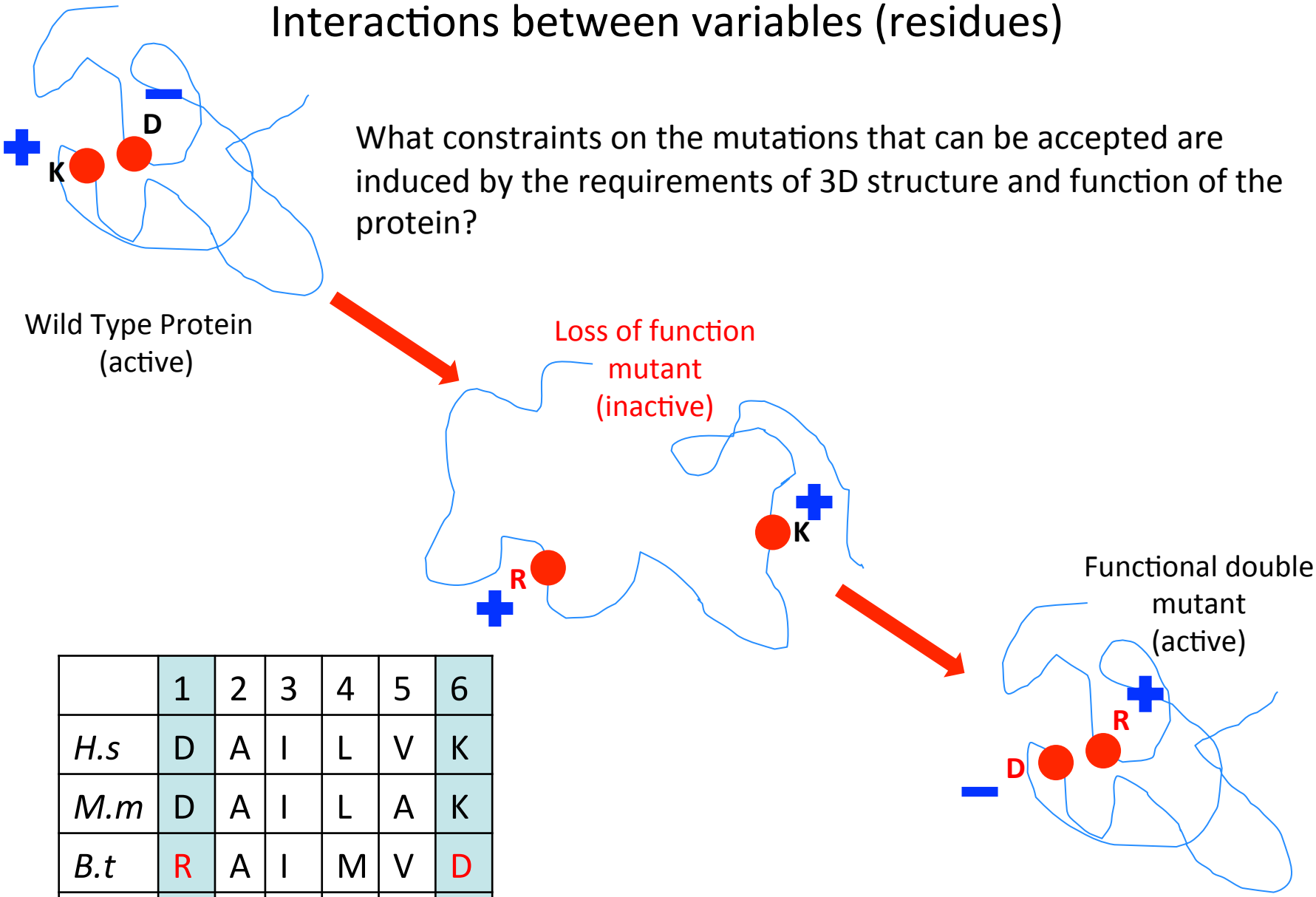
Interactions between variables (residues)

What constraints on the mutations that can be accepted are induced by the requirements of 3D structure and function of the protein?



Interactions between variables (residues)

What constraints on the mutations that can be accepted are induced by the requirements of 3D structure and function of the protein?



	1	2	3	4	5	6
H.s	D	A	I	L	V	K
M.m	D	A	I	L	A	K
B.t	R	A	I	M	V	D
G.g	R	V	I	L	V	D

correlation

Interactions lead to covariance of amino acids

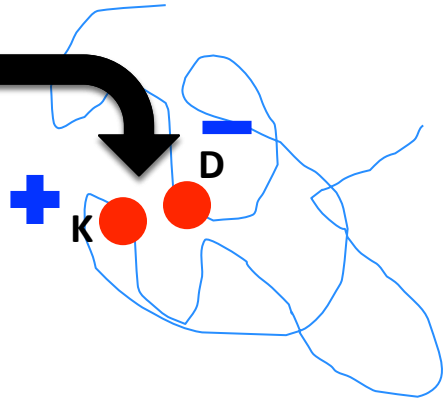
```
                                  20        30        40        50        60        70        80        90       100
Racoon                            VEEVGGEALGRLLVVYPWTQRFFESFGDLSSADAIMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPE
Ring-tailed coati                 VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAIMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPE
Red fox                           VDEVGGEALGRLLIVYPWTQRFFDSFGDLSTPDAVMGNAKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPE
Weddell seal                      VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPNAIMSNPKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDQLHVDPE
Harbor seal                       VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSADAIMGNPKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPE
Eurasion badger                   VDEVGGEALGRLLVVYPWTQRYFDSFGDLSTPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPE
Red panda                         VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPE
Ferret                            VDEVGGETLGRLLVVYPWTQRFFDSFGDLSSPDAVMSNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPE
Fur seal                          VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPE
River otter                       VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPE
Beach marten                      VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPE
European mink                     VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPE
Ratel                             VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPE
Colobus monkey                    VDAVGGEALGRLLVVYPWTQRFFESFGDLSSPAAVMGNPKVKAHGKKVLGAFSDGLAHLDSLKGTFSQLSELHCDKLHVDPE
Lowland gorilla                   VDAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFSQLSELHCDKLHVDPE
Human                             VDAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFSQLSELHCDKLHVDPE
Spider monkey                     VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPAAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPE
Brown spider monkey               VDEVGGEALGRLLVVYPWTQRFFESFGALSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPE
Tamarin                           VDEVGGEALGRLLVVYPWTQRFFESFGALSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPE
Moustached tamarin                VDEVGGEALGRLLVVYPWTQRFFESFGALSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPE
Capuchin                          VDEVGGEALGRLLVVYPWTQRFFDSFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLTHLDNLKGTFAQLSELHCDKLHVDPE
White Capuchin                    VDEVGGEALGRLLVVYPWTQRFFDSFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLTHLDNLKGTFAQLSELHCDKLHVDPE
white sapajou                     VDEVGGEALGRLLVVYPWTQRFFDSFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLTHLDNLKGTFAQLSELHCDKLHVDPE
squrrel monkey                    VEDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPE
red colobus                       VDEVGGEALGRLLVVYPWTQRFFDSFGDLSTADAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPE
Collared titi                     VXEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMSNXKVKAHGKKVLGAFSDGLAHLDNLKSTFAQLSELHCDKLHVDPE
Common marmoset                   VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLTHLDNLKGTFAHLSELHCDKLHVDPE
Howler monkey                     VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMHNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPE
Black spider monkey               VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPE
Goeldi's marmoset                 VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMNNPKVKAHGKKVLGAFSDGLTHLDNLKGTFAQLSELHCDKLHVDPE
Brown tamarin                     VEEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCNKLHVDPE
black and red tamarin             VEEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMNNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPE
moustached tamarin                VEEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMNNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPE
papio sphinx                      VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPE
olive baboon                      VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPAAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPE
gelada baboon                     VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPAAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPE
green monkey                      VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPE
sootey mangabey                   VDEVGGEALGRLLVVYPWTQRFFESFGDLSNPDAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPE
crab eating macaque               VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPE
rhesus macaque                    VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPE
western colobus monkey            VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPE
```

Exploit correlation structure of protein sequences

What constraints on the mutations that can be accepted are induced by the requirements of 3D structure and function of the protein?

	1	2	3	4	5	6
H.s	D	A	I	L	V	K
M.m	D	A	I	L	A	K
B.t	R	A	I	M	V	D
G.g	R	V	I	L	V	D

correlated



Wild Type Protein
(active)

Can we **invert correlations** in the sequence data to provide information about **the 3D protein structure?**

Burger, Lukas, and Erik Van Nimwegen. PLoS computational biology 6.1 (2010): e1000633.
Marks*, Debora S., Colwell*, Lucy J. et al. PloS one 6.12 (2011): e28766,

Measure pair correlations in the sequence alignment

	1	2	3	4	5	6
H.s	K	D	I	L	V	D
M.m	K	D	I	L	V	D
S.c	D	D	I	K	V	H
S.p	D	E	I	L	V	H

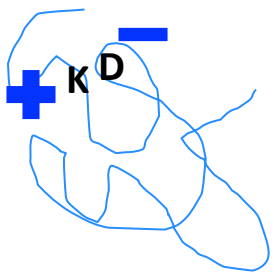
	K	D	I	L	V	D
K	X					I
D		X				
I			X			
L				X		
V					X	
D	I					X

?

Compute (for example) the mutual information for each pair of columns

$$MI_{ij} = \sum_{A_i, A_j = 1}^{q} f_{ij}(A_i, A_j) \ln \left(\frac{f_{ij}(A_i, A_j)}{f_i(A_i) f_j(A_j)} \right)$$

Wild Type Protein



Measure pair correlations in the sequence alignment

	1	2	3	4	5	6
H.s	K	D	I	L	V	D
M.m	K	D	I	L	V	D
S.c	D	D	I	K	V	H
S.p	D	E	I	L	V	H

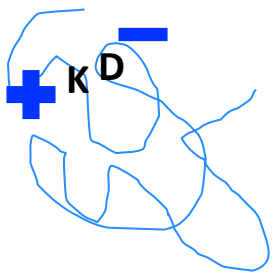
	K	D	I	L	V	D
K	X					I
D		X				
I			X			
L				X		
V					X	
D	I					X

Compute (for example) the mutual information for each pair of columns

$$MI_{ij} = \sum_{A_i, A_j = 1}^{q} f_{ij}(A_i, A_j) \ln \left(\frac{f_{ij}(A_i, A_j)}{f_i(A_i) f_j(A_j)} \right)$$

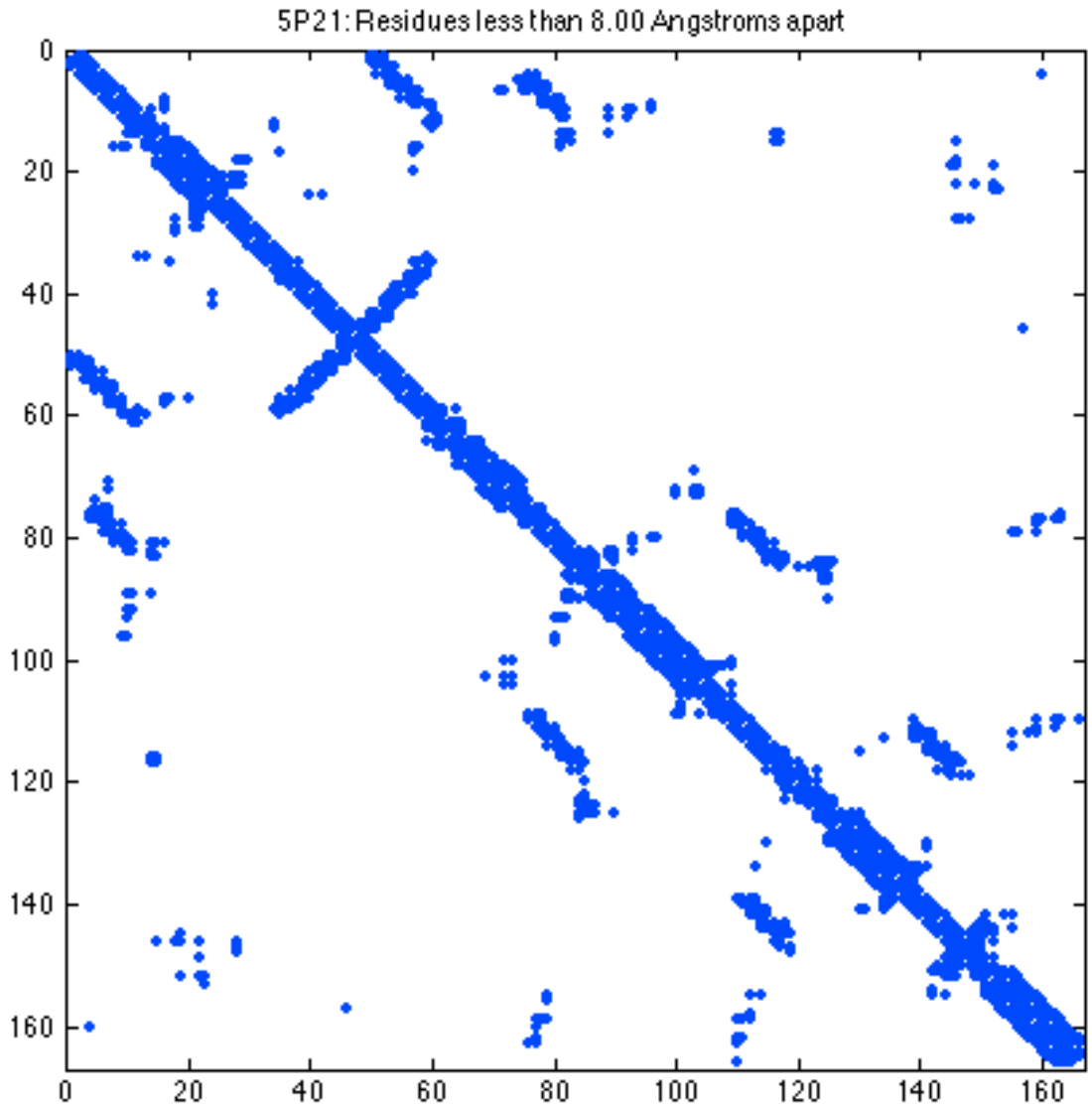
?

Wild Type Protein



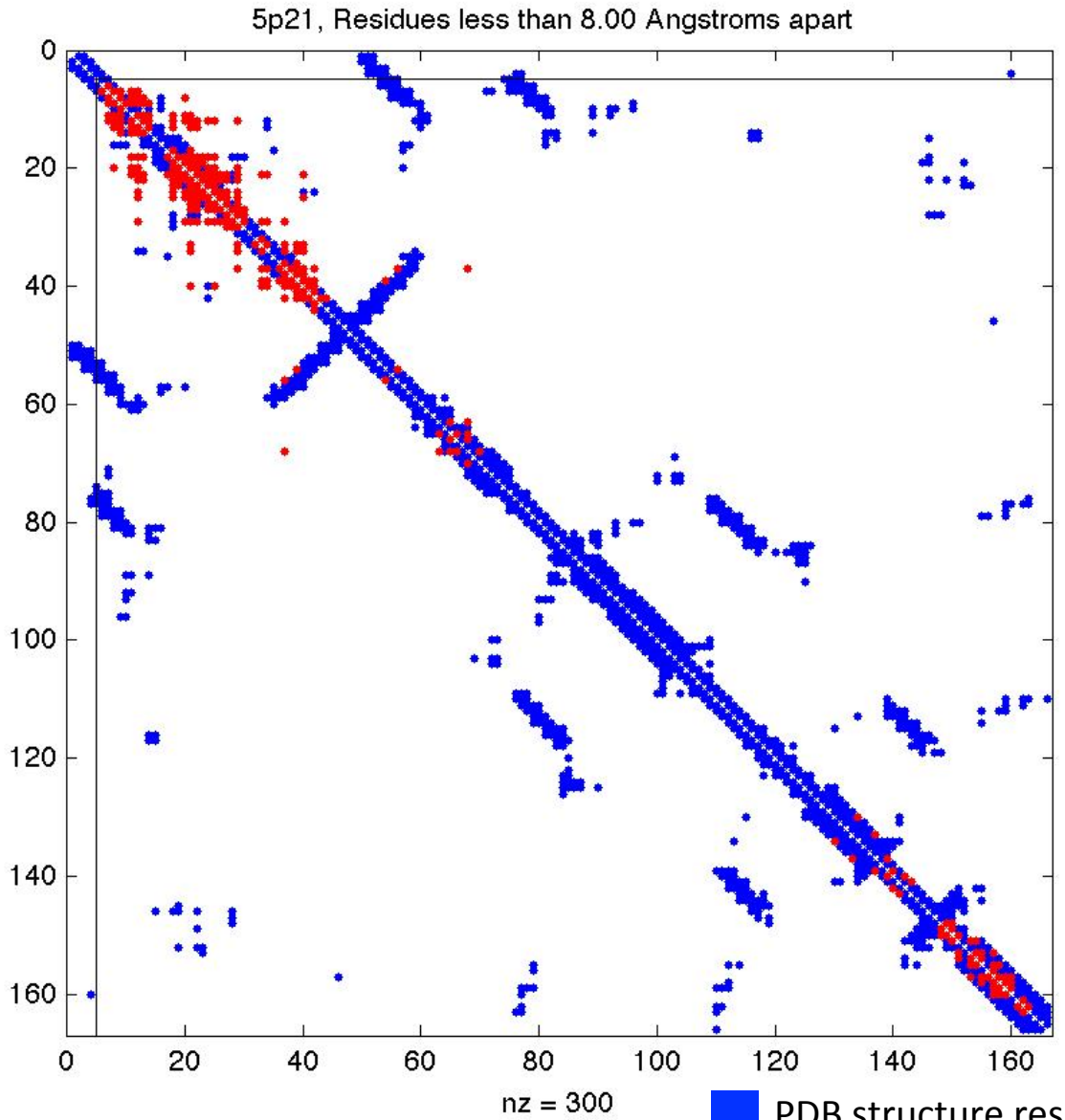
Are the most correlated residue pairs close in tertiary protein structure?

Testing: Are highly correlated pairs close in structure?



PDB structure residue contacts

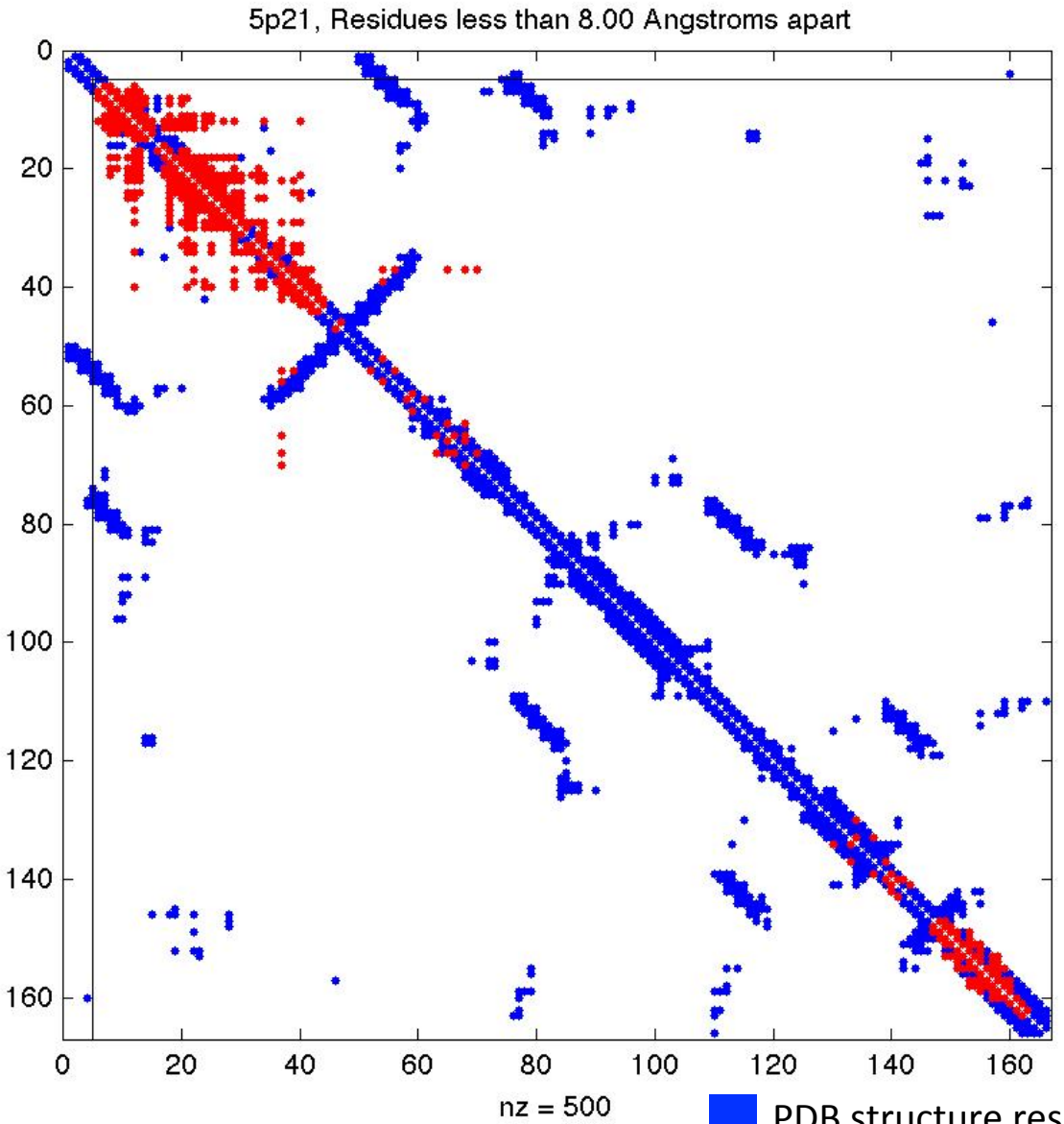
Testing MI: Are high-scoring pairs close in structure?



PDB structure residue contacts

residue pairs with 150 highest MI values

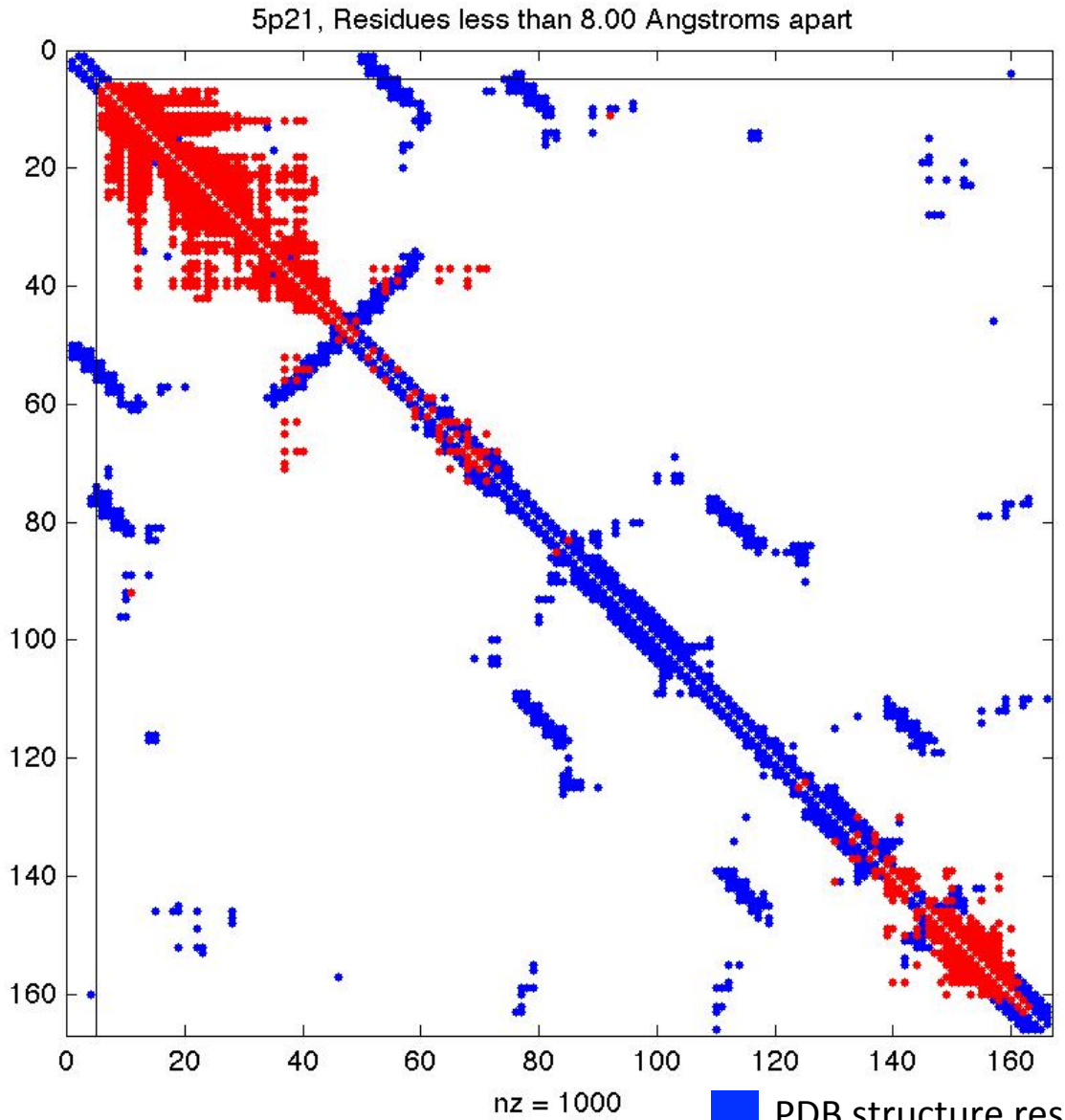
Testing MI: Are high-scoring pairs close in structure?



PDB structure residue contacts

residue pairs with 250 highest MI values

Testing MI: Are high-scoring pairs close in structure?

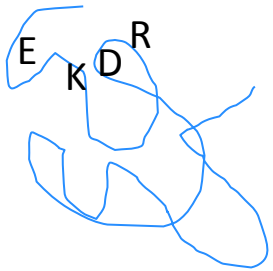


PDB structure residue contacts

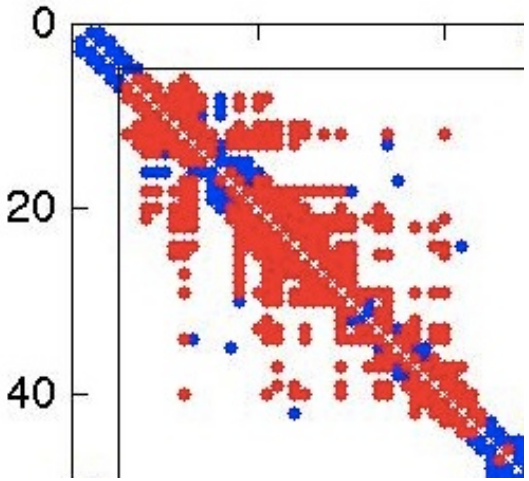
residue pairs with 500 highest MI values

Why does this fail? Construct and analyze a model

Wild Type Protein



K ⟷ D



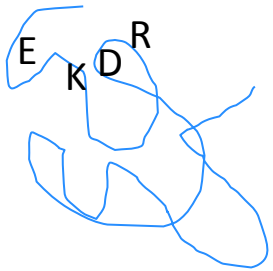
	K	D	E	R
K	X	I		
D	I	X		
E			X	
R				X

PDB structure residue contacts

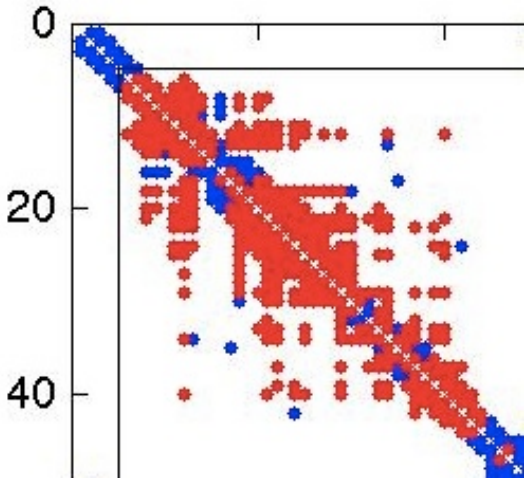
residue pairs with 100 highest MI values

Construct and analyze a model

Wild Type Protein



E ⟷ K ⟷ D ⟷ R



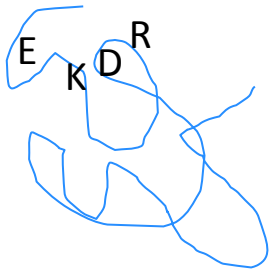
	K	D	E	R
K	X	I	I	
D	I	X		I
E	I		X	
R		I		X

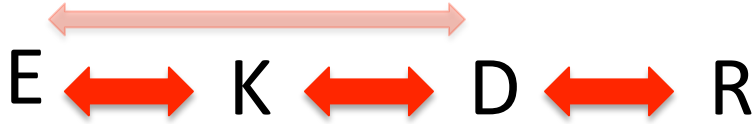
PDB structure residue contacts

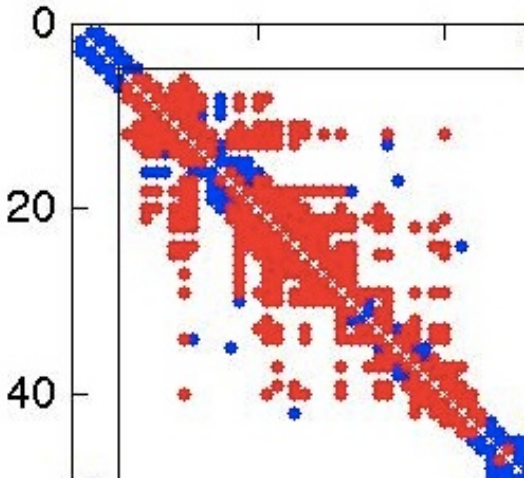
residue pairs with 100 highest MI values

Construct and analyze a model

Wild Type Protein







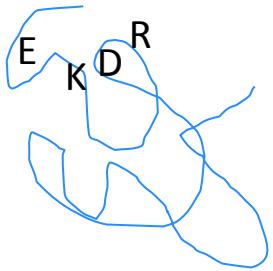
	K	D	E	R
K	X	I	I	
D	I	X	C	I
E	I	C	X	
R		I		X

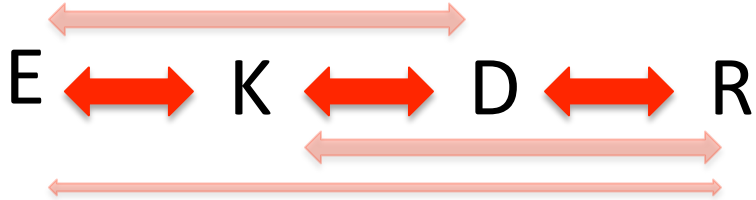
PDB structure residue contacts

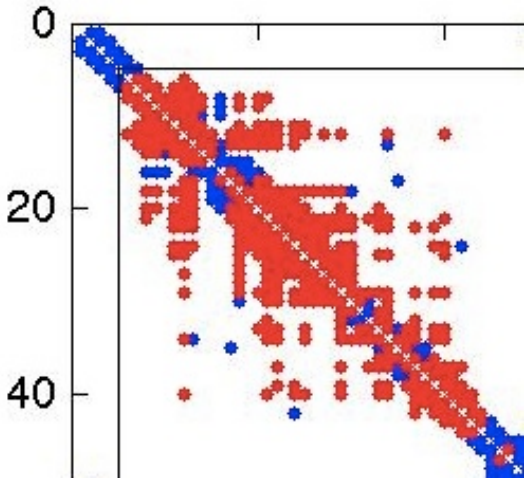
residue pairs with 100 highest MI values

Construct and analyze a model

Wild Type Protein







	K	D	E	R
K	X	I	I	C
D	I	X	C	I
E	I	C	X	C
R	C	I	C	X

PDB structure residue contacts

residue pairs with 100 highest MI values

Probability model

$P(A_1,\ldots,A_L)$ = Probability a sequence is part of protein family

$$P_i(A_i) = \sum_{\{A_k | k \neq i\}} P(A_1,\ldots,A_L) = f_i(A_i)$$

$$P_{ij}(A_i, A_j) = \sum_{\{A_k | k \neq i, j\}} P(A_1,\ldots,A_L) = f_{ij}(A_i, A_j)$$

Satisfy these constraints and choose the maximum entropy or least constrained model:

$$S = -\sum_{\{A_i | i=1,\ldots,L\}} P(A_1,\ldots,A_L) \ln P(A_1,\ldots,A_L)$$

Probability model

$P(A_1,\ldots,A_L)$ = Probability a sequence is part of protein family

$$= \frac{1}{Z}\exp\left\{-\sum_i h_i A_i - \sum_{(i,j)} e_{ij}\left(A_i, A_j\right)\right\}$$

Probability model

$P(A_1,\ldots,A_L)$ = Probability a sequence is part of protein family

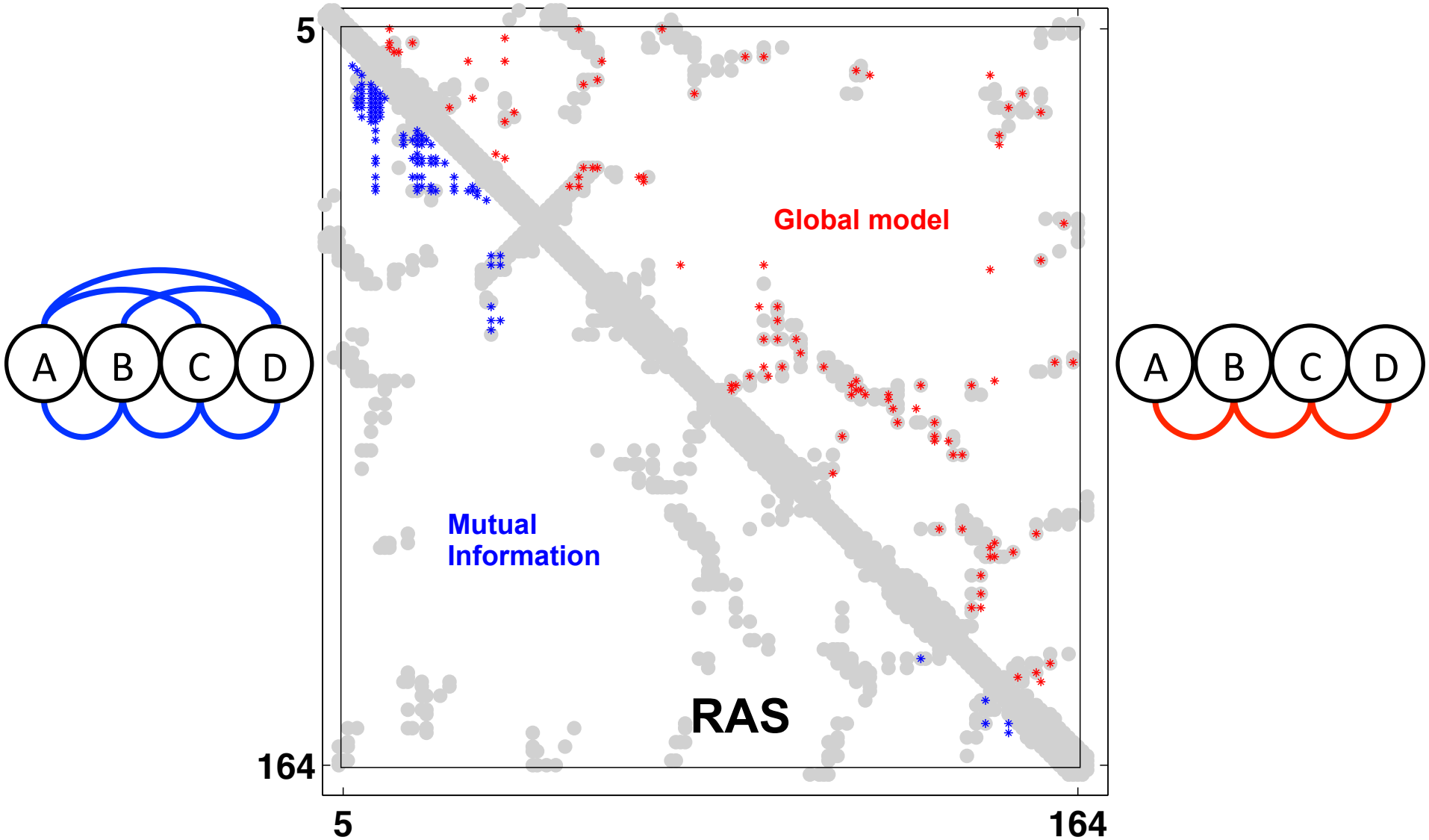
$$= \frac{1}{Z}\exp\left\{-\sum_i h_i A_i - \sum_{(i,j)} e_{ij}\left(A_i, A_j\right)\right\}$$

naïve mean field approximation, assume small couplings

$$e_{ij}\left(A_i, A_j\right) = \left(C^{-1}\right)_{ij}\left(A_i, A_j\right)$$

Sequences to structure:

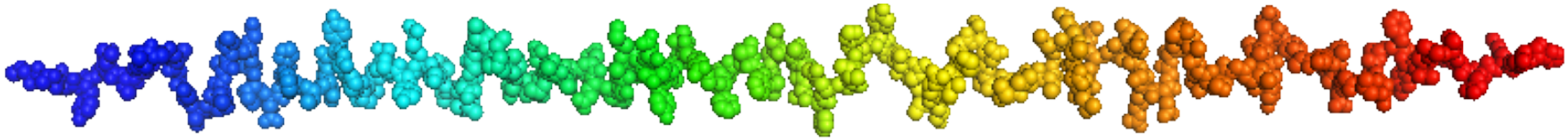
Solution: A global probability model – high scoring pairs are close in structure!



Marks*, Colwell* *et al*, PloS One 2011

Sequences to structure:

Is this enough information to fold the protein?



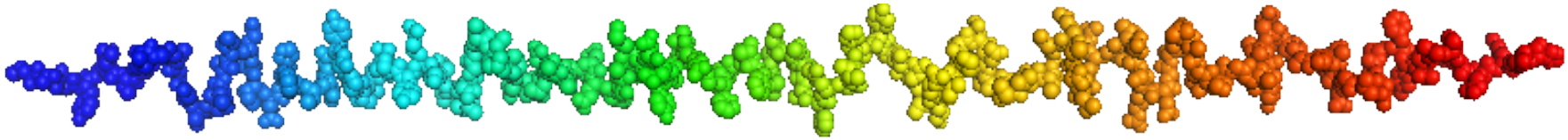
Start with an extended polypeptide, add predicted secondary structure

Our hypothesis is the high scoring pairs will be close in the structure – so constrain the distance between the residues in each pair.

Marks*, Colwell* *et al*, PloS One 2011

Sequences to structure:

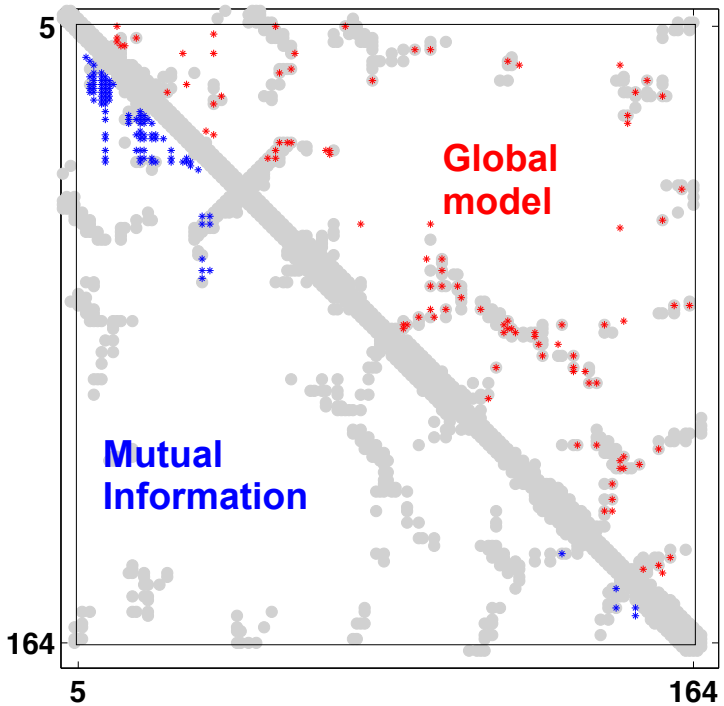
Is this enough information to fold the protein?

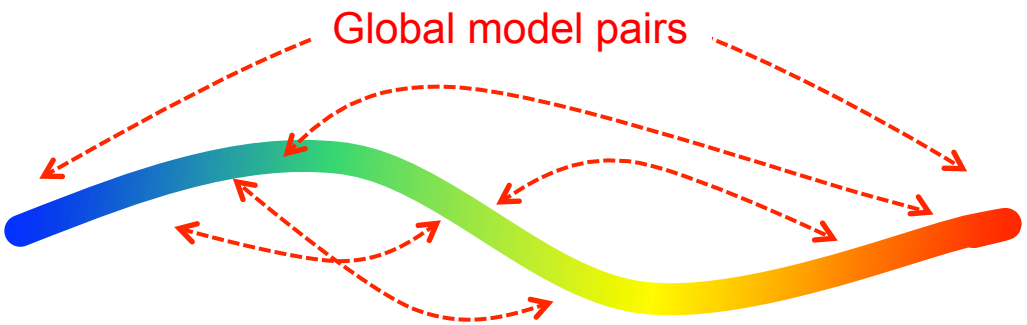


Start with an extended polypeptide, add predicted secondary structure

Our hypothesis is the high scoring pairs will be close in the structure – so constrain the distance between the residues in each pair.

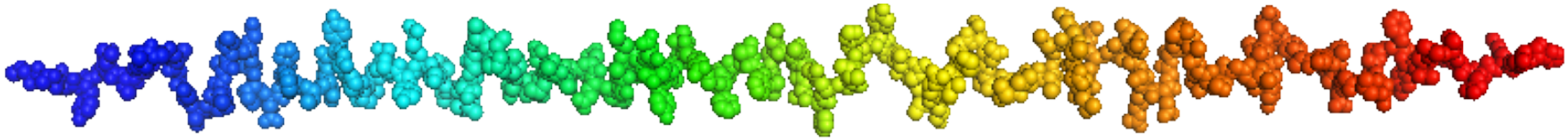
This massively reduces the space of possible 3D conformations of the protein





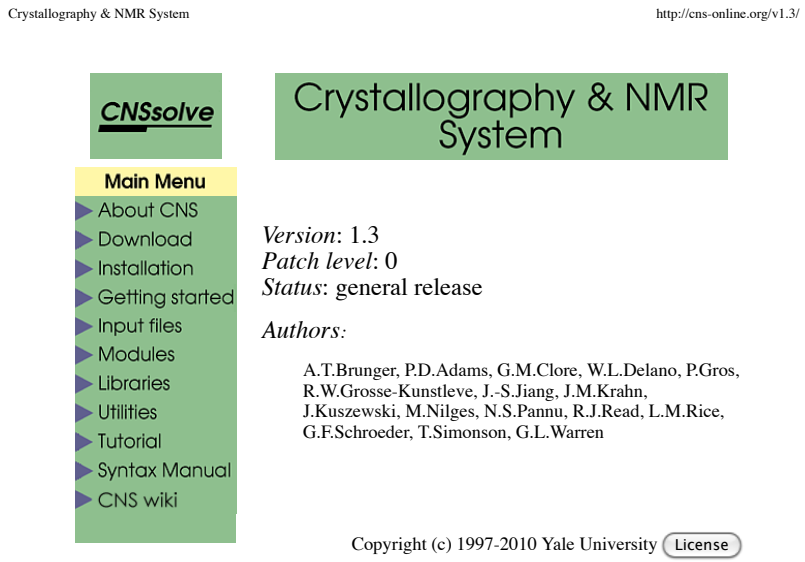
Marks*, Colwell* *et al*, PloS One 2011

Folding the protein



Start with an extended polypeptide

Our hypothesis is the high scoring pairs will be close in the structure – so we constrain the distance between the two residues in each pair.



Use these distance constraints together with predicted secondary structure in the standard distance geometry and simulated annealing protocol from CNS to generate structures

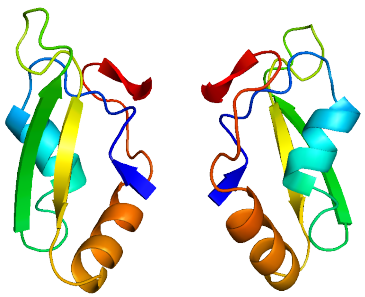
Marks*, Colwell* *et al*, PloS One 2011

Sequences to structure:

Correlations in the mutation patterns of amino acids within a protein used to predict tertiary structure.

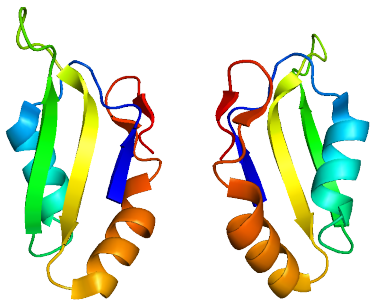
An RNA binding domain: 2.9Å C-alpha rmsd over 67 residues

predicted
blind top ranked



ELAV4_HUMAN

observed
crystal structure



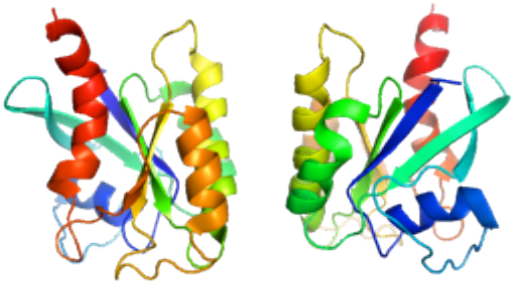
1G2E.pdb

Sequences to structure:

Correlations in the mutation patterns of amino acids within a protein used to predict tertiary structure.

RAS: 3.5Å C-alpha rmsd over 161 residues

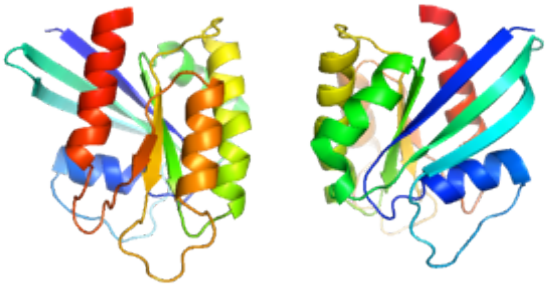
predicted
blind top ranked



RASH_HUMAN

180°

observed
crystal structure



5P21.pdb

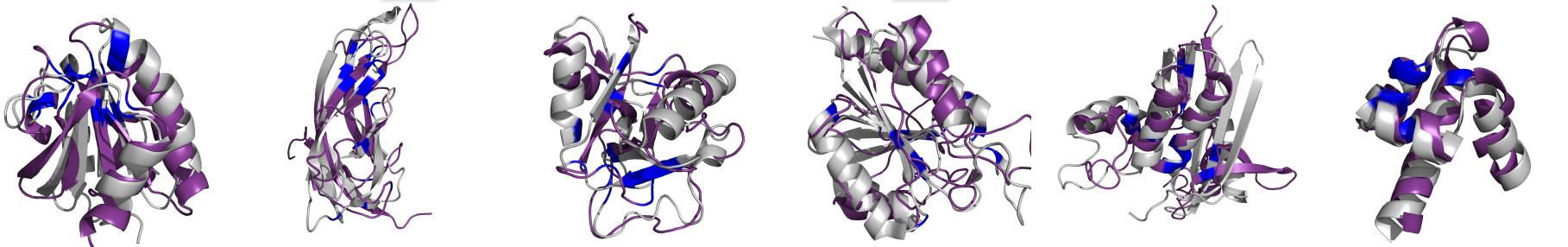
180°

Beta strands positioned well enough to predict correct registration with the exception of the beta one strand, which was very recently shown to exist in both the conformations that we find.

Comparison of predicted and observed structures

Crystal structure*

Predicted structure



Thiredoxin: 3.5 Å, 97 res

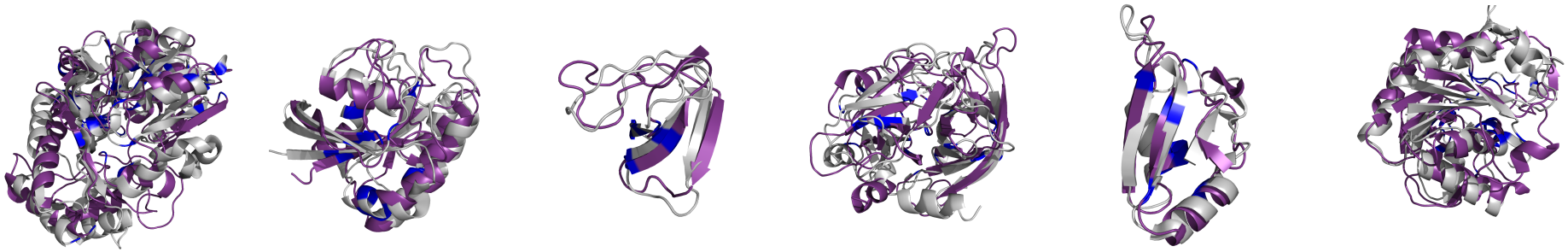
Cadherin: 3.8 Å, 88 residues

Lectin: 4.0 Å 100 res

MOEA_ECOLI: 4.0 Å 128 res

RnaseH: 3.5 Å 114 res

Q9KZ96_STRCO: 1.46 Å 49 res



O85354_CAUCR: 3.74 Å, 278 res

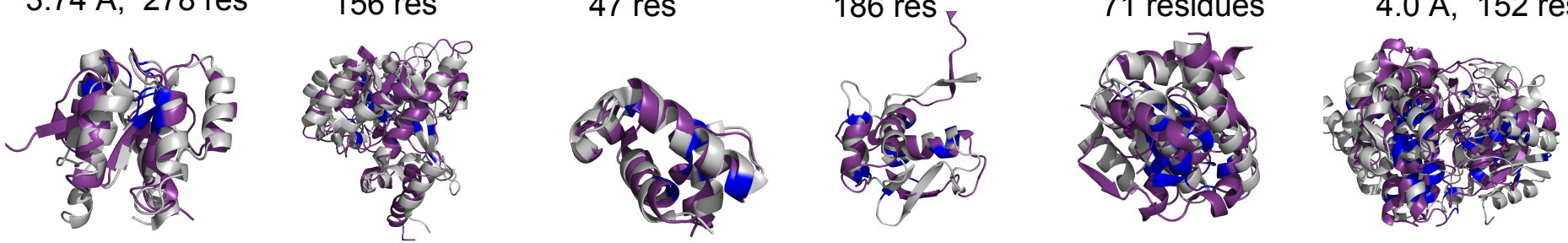
Ras: 3.2 Å, 156 res

SH3: 3.6 Å, 47 res

Trypsin: 4.27 Å, 186 res

RRM_1: 3.16 Å, 71 residues

ISPD_ECOLI: 4.0 Å, 152 res



CheY: 2.98 Å, 107 residues

Q3SJE6_THIDA: 3.6 Å, 206 res

GERE_BACSU: 1.8 Å, 49 res

OmpR: 3.9 Å 62 res

GMHA_VIBCH : 4.0 Å, 133 res

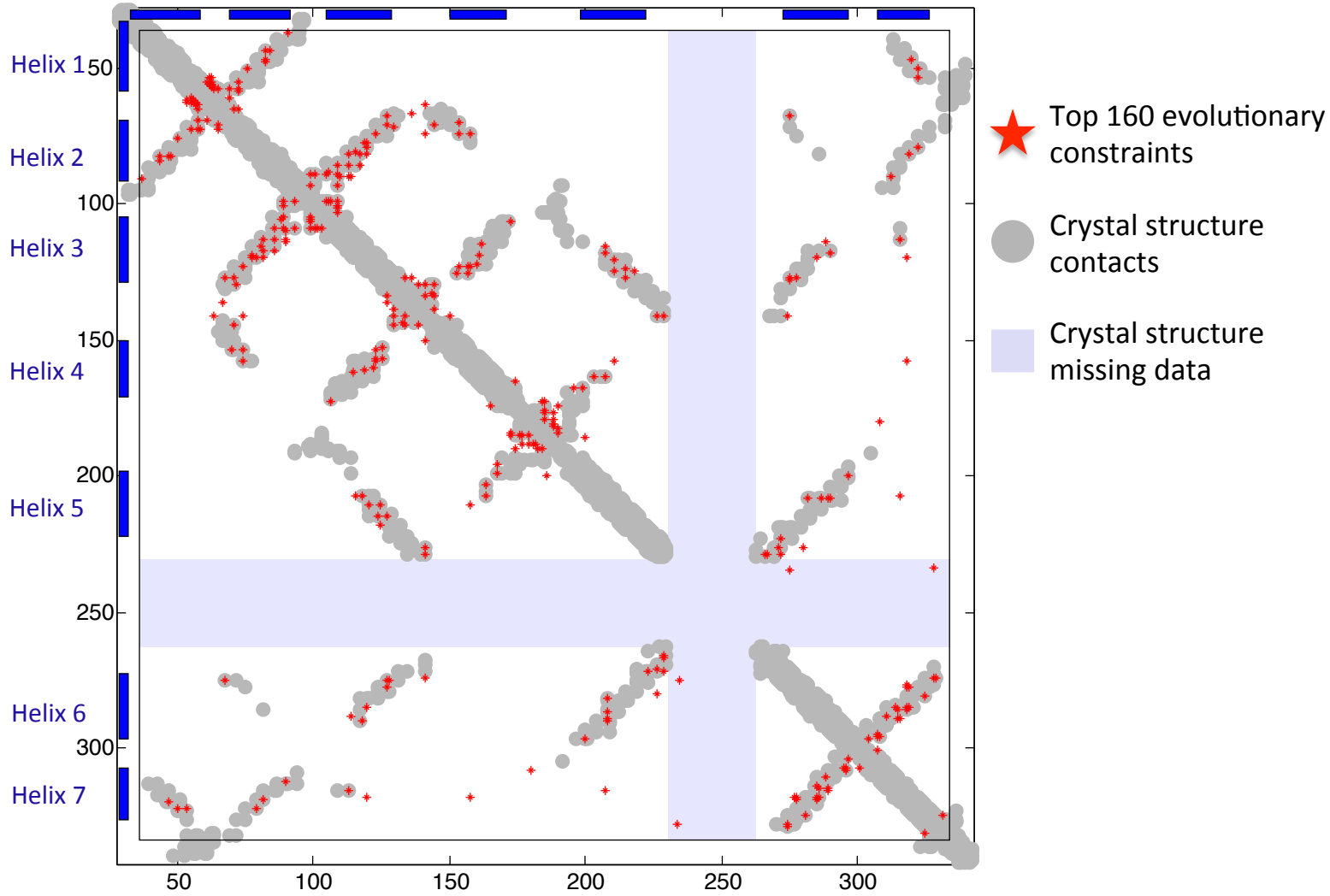
Q7WRJ3_KLEPN 4.5 Å, 232 res

Challenge : transmembrane proteins

Hard experimentally
No high throughput
Drug targets - > 40%

De novo prediction

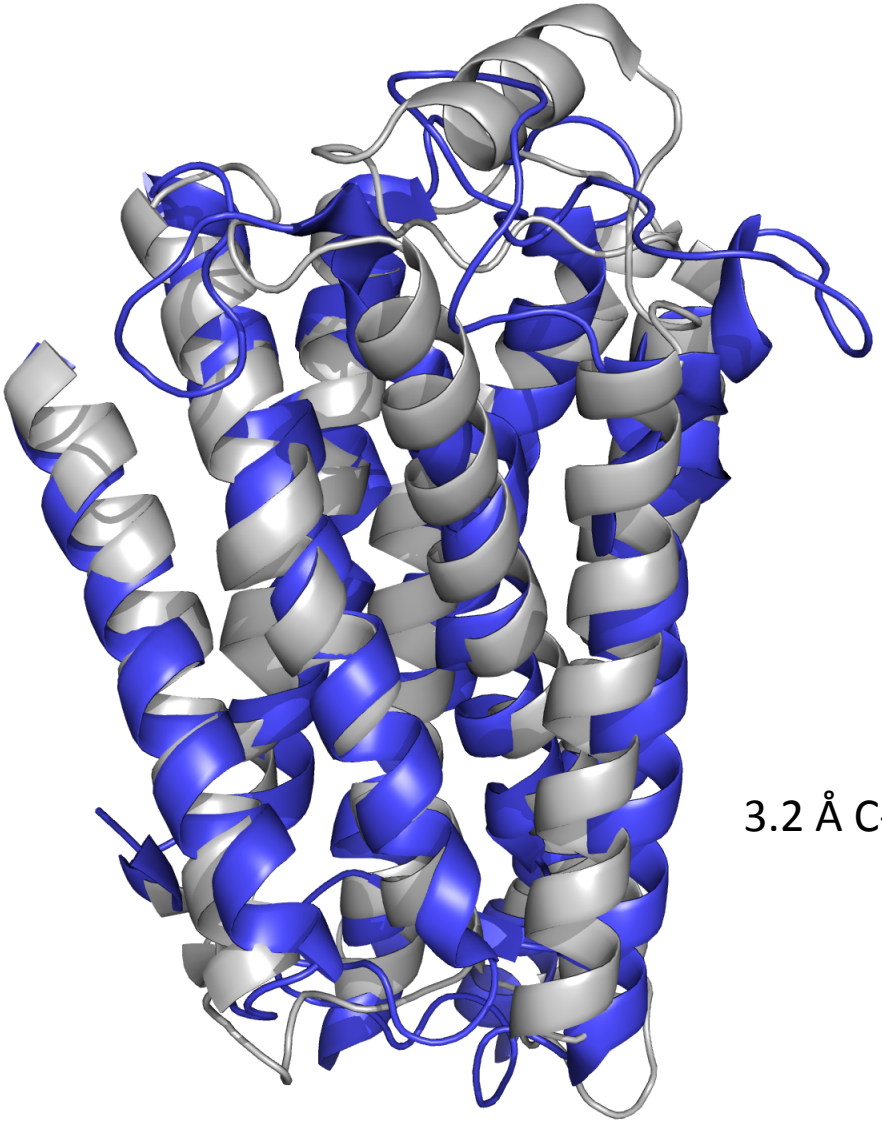
β2 adrenergic receptor : evolutionary constraints



The β2 adrenergic receptor, a seven transmembrane helix GPCR family member.

Hopf, Colwell *et al*, Cell 2012

Blind prediction for β2 adrenergic receptor



Crystal structure*

Predicted structure

3.2 Å C-α rmsd over 200 residues

*Crystal Structure: Rasmussen SG, DeVree BT, Zou Y, Kruse AC, Chung KY, Kobilka TS, Thian FS, Chae PS, Pardon E, Calinski D, Mathiesen JM, Shah ST, Lyons JA, Caffrey M, Gellman SH, Steyaert J, Skiniotis G, Weis WI, Sunahara RK, Kobilka BK. Nature. 2011 Jul 19.

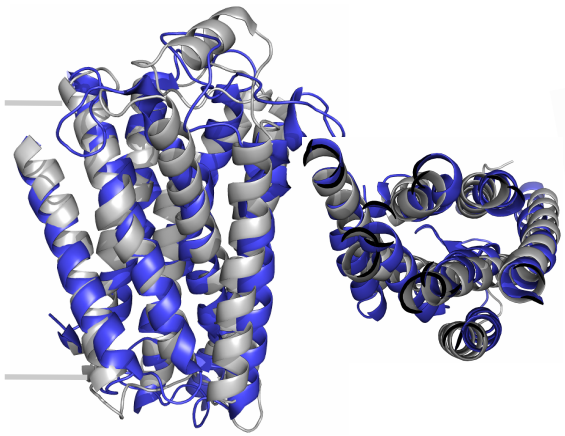
Hopf, Colwell *et al*, Cell 2012

Sequences to structure:

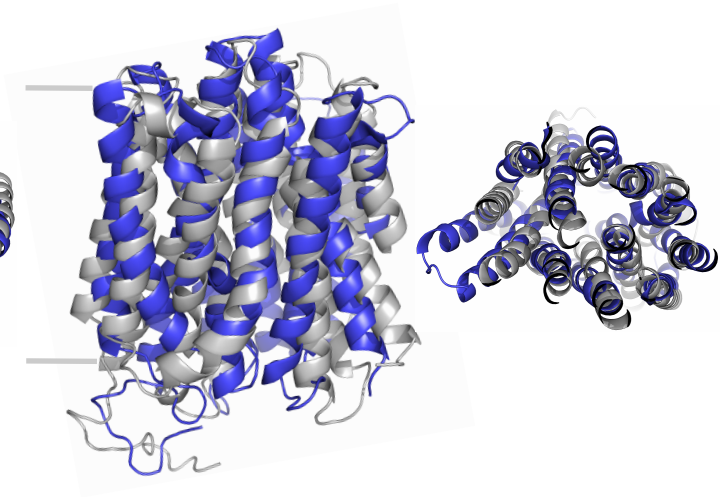
Blind structure prediction.

observed predicted

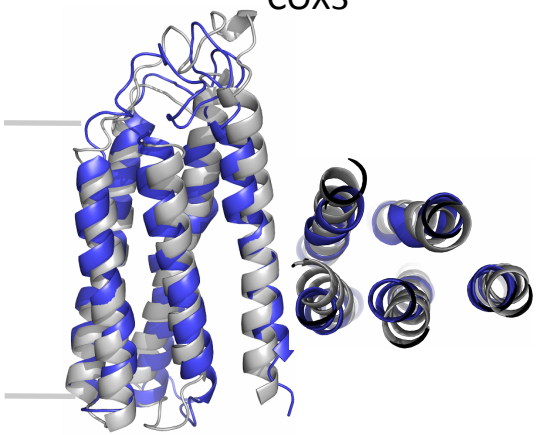
β2 adrenergic receptor



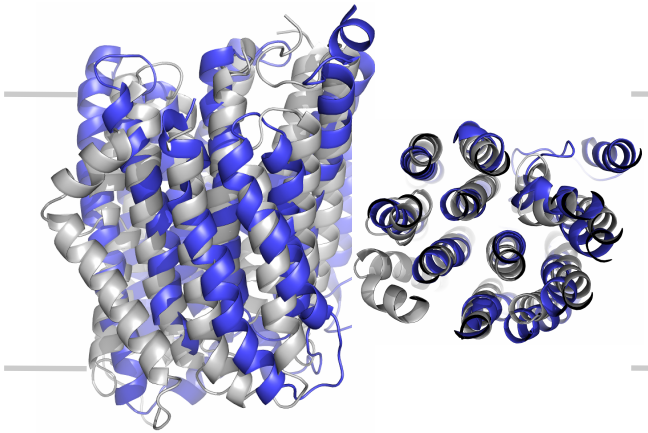
G-3-P transporter GlpT



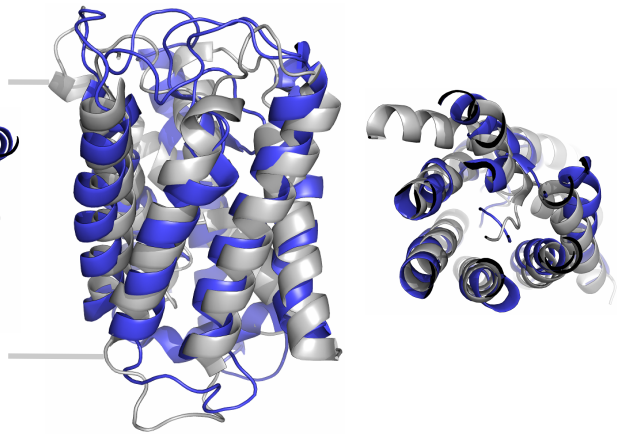
COX3



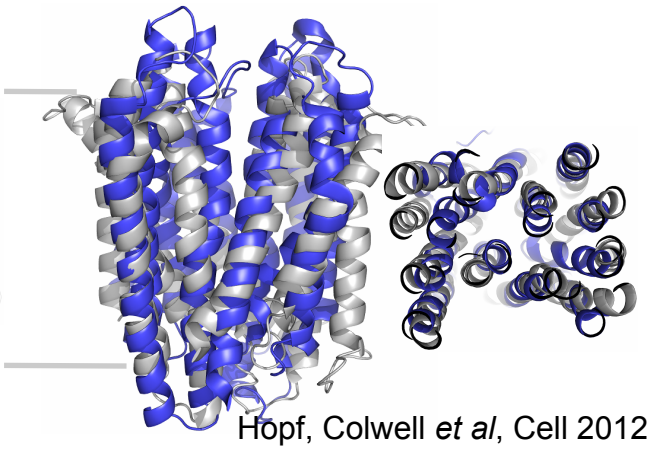
NADH-quinone oxidoreductase subunit N 1



Aquaporin-0



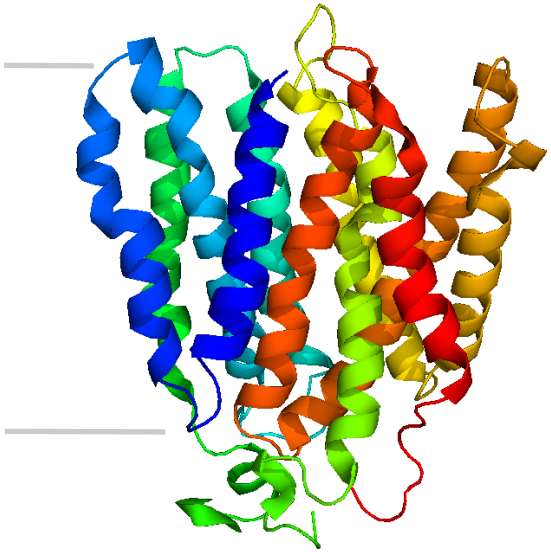
Multidrug resistance protein norM



Hopf, Colwell *et al*, Cell 2012

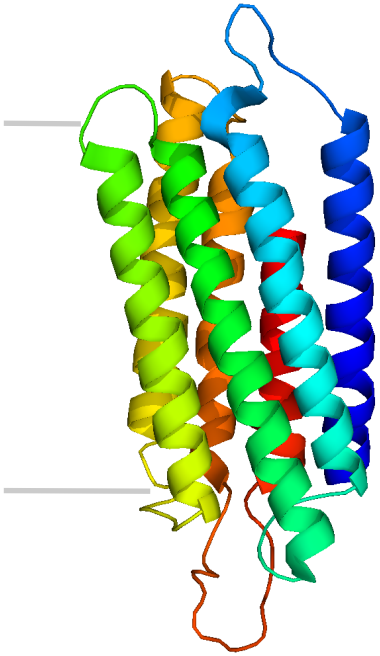
11 medically important membrane proteins of unknown structure predicted

OCTN1



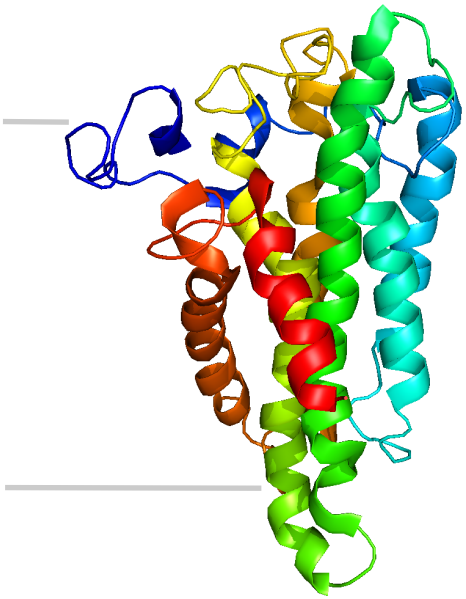
Crohn's disease,
rheumatoid arthritis

Adiponectin receptor 1



diabetes, obesity,
cancer

MT-ND1



LHON, MELAS,
Alzheimer, Parkinson