Protein Physics 2016 Lecture 12, March 1

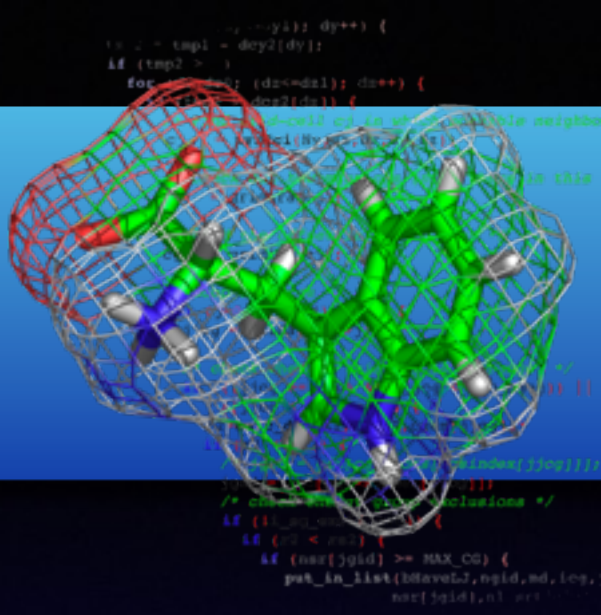
The Bioinformatics Approach to Proteins

Magnus Andersson

magnus.andersson@scilifelab.se

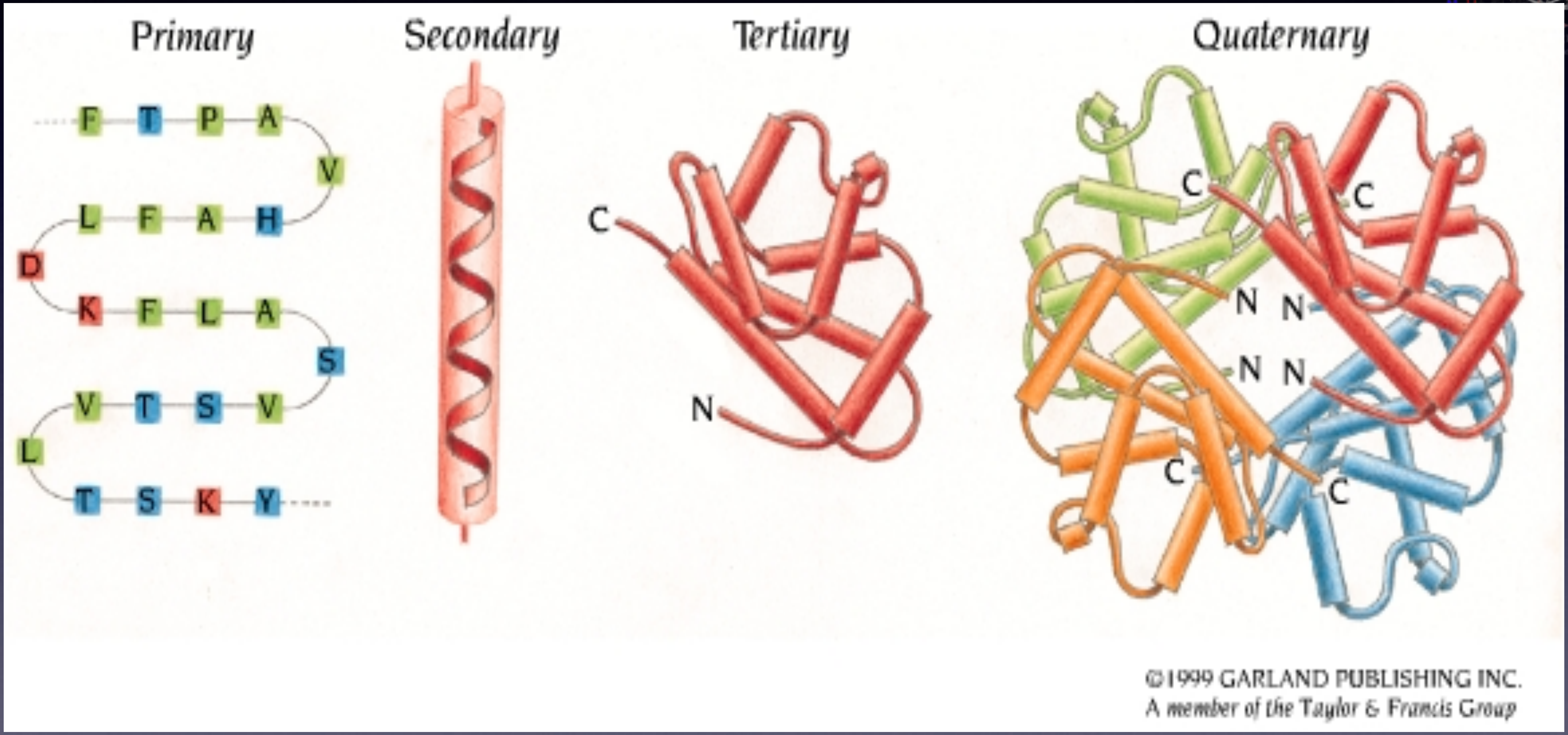
Theoretical & Computational Biophysics

Bioinformatics



- Genomes, genes & evolution
- Large scale databases
- Sequence comparison, finding genes
- Sequence - structure - function
- Evolution vs. laws of nature
    - *Computer science vs. chemistry/physics?*

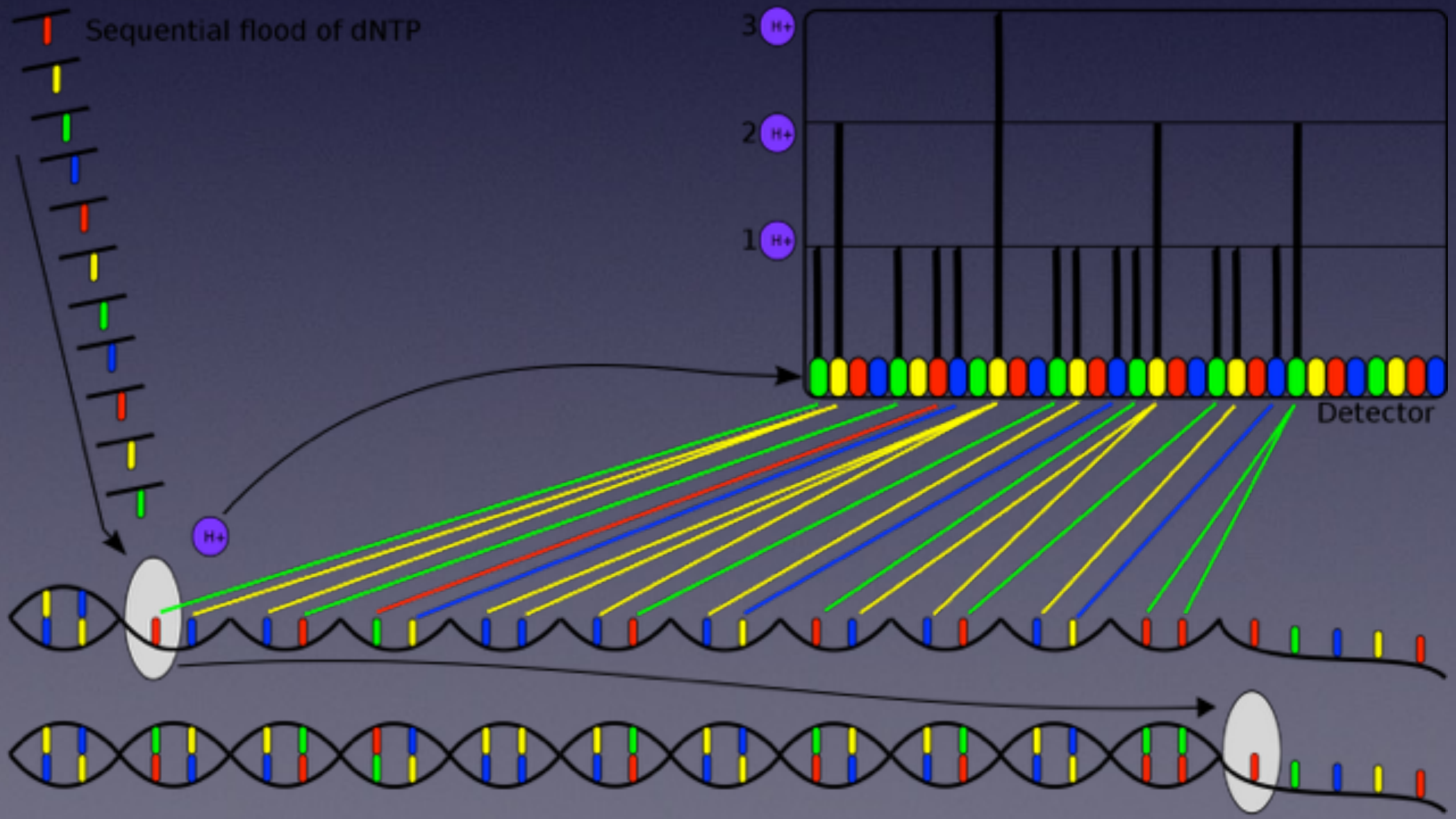
Intellectual & practical problems



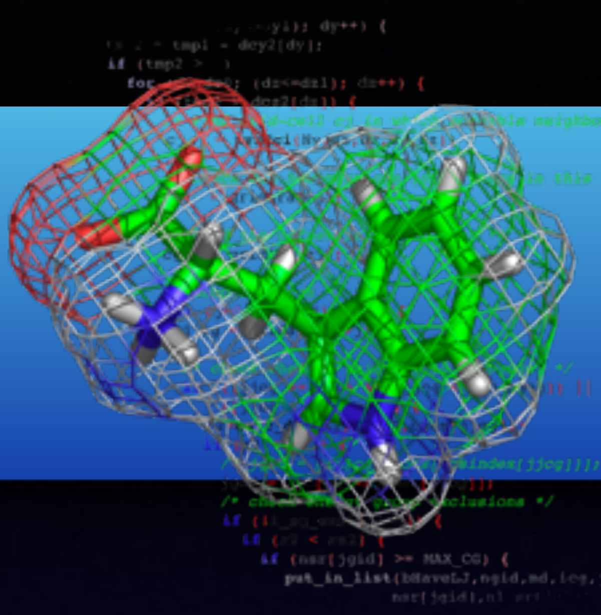
©1999 GARLAND PUBLISHING INC.
A member of the Taylor & Francis Group

*It is interesting to understand **how** structure forms, but it would also be worth a lot if we could just **predict** the final structure!*

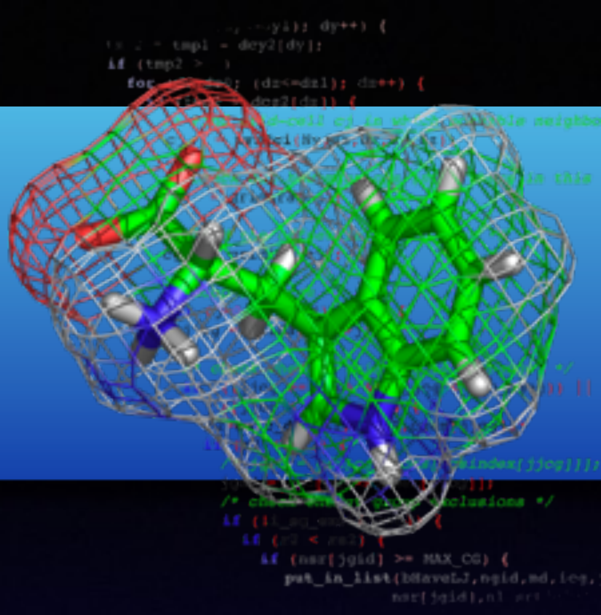
DNA sequencing



DNA vs protein

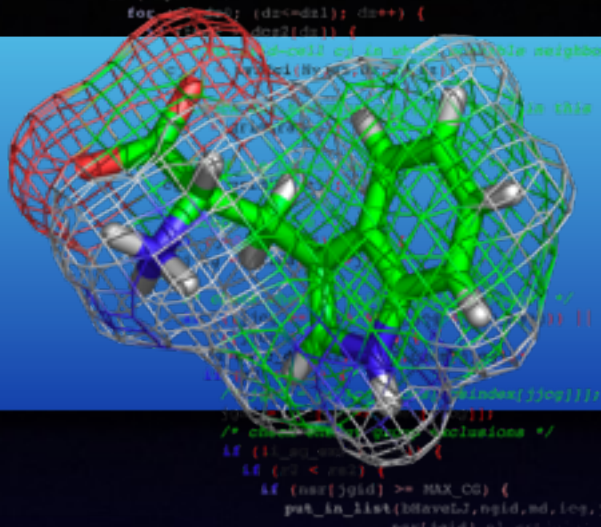


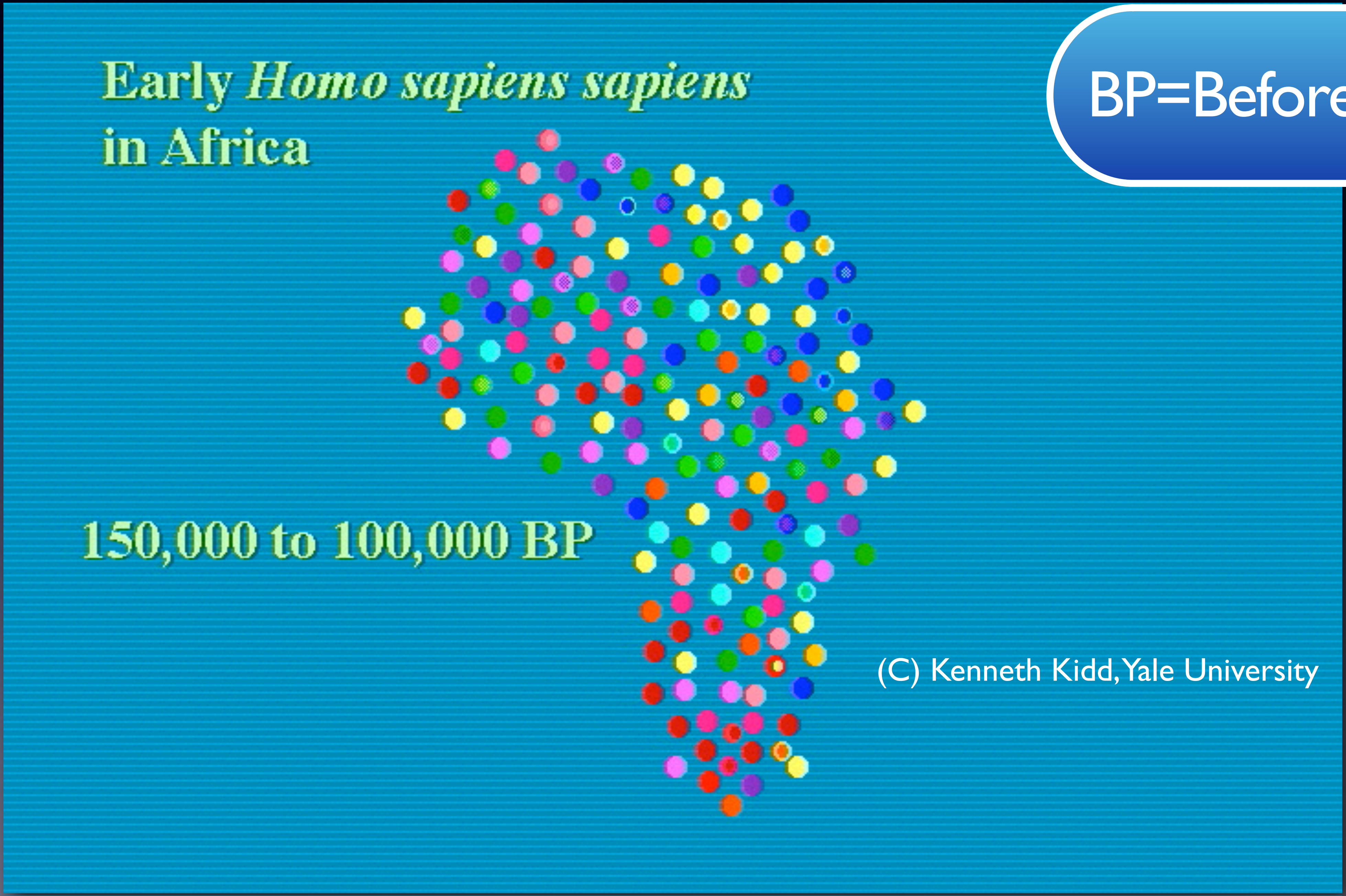
- 1.2% protein-coding DNA in human
  - ORF: Open Reading Frame
  - ATG ... ... TAA
- 20,000-25,000 genes in human
- How do we find & study similarities?



Examples

Human evolution





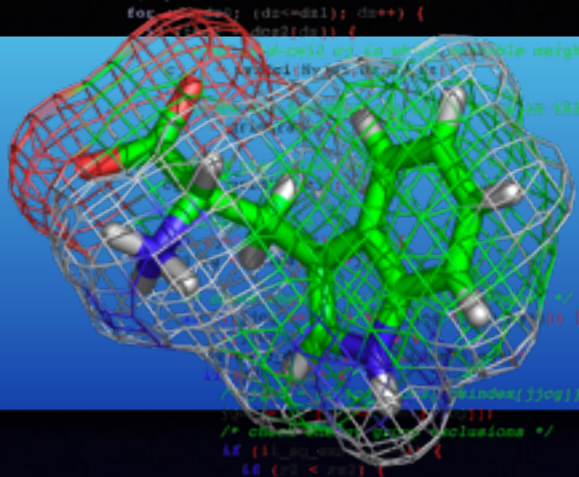
Early *Homo sapiens sapiens* in Africa

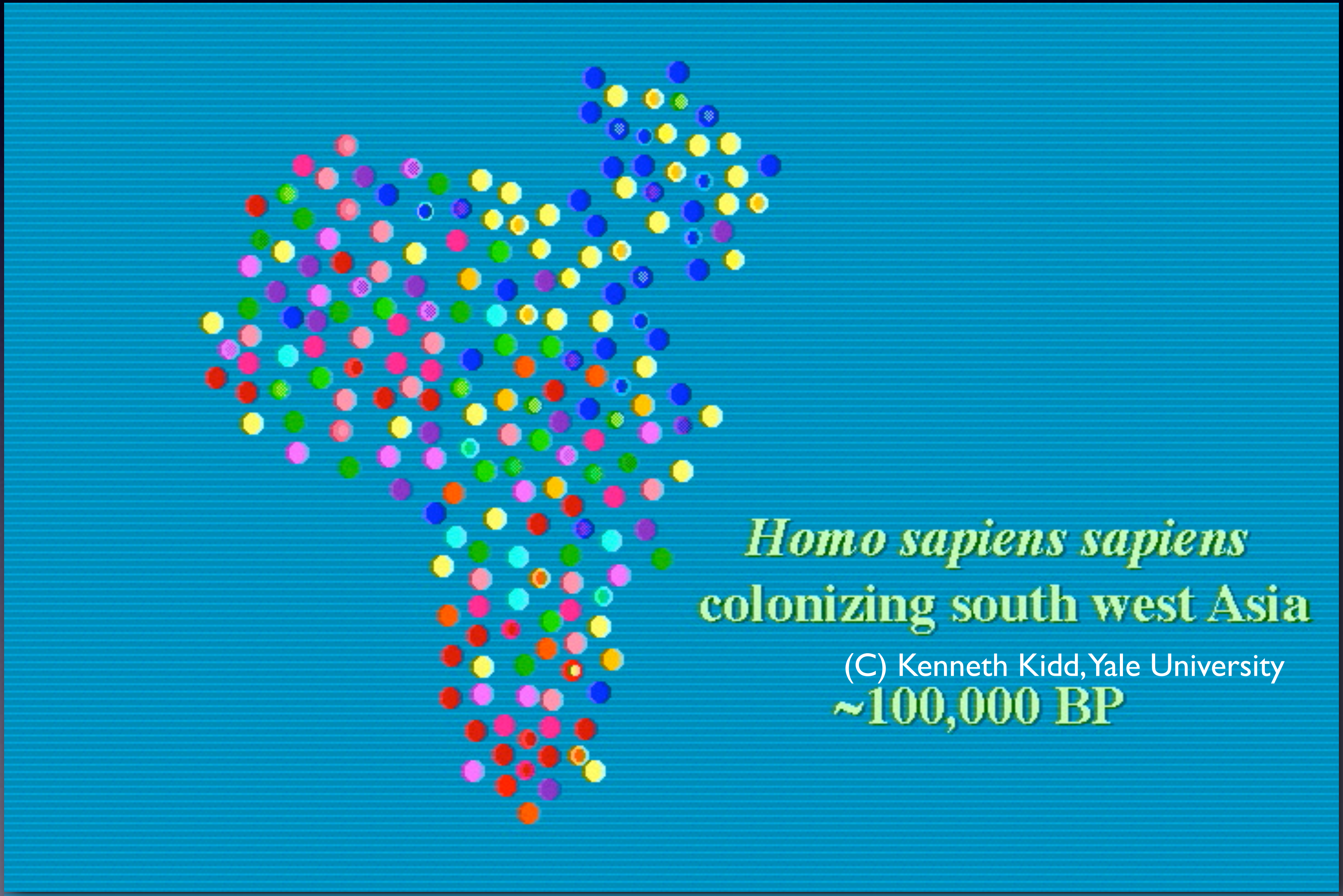
150,000 to 100,000 BP

(C) Kenneth Kidd, Yale University

BP=Before Present

Human evolution



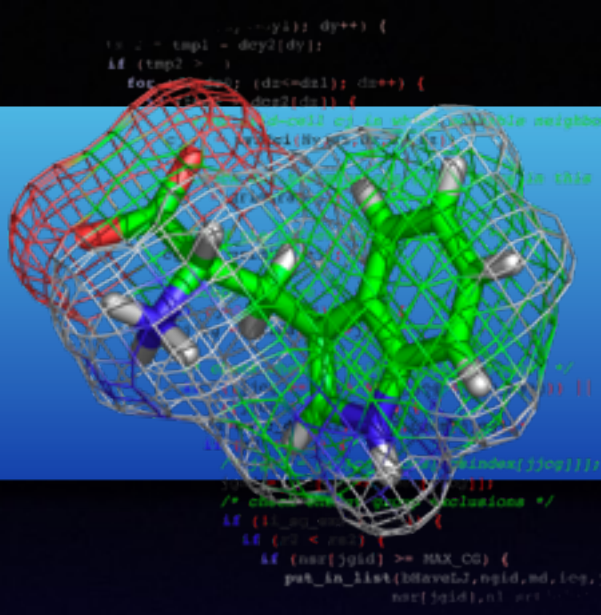


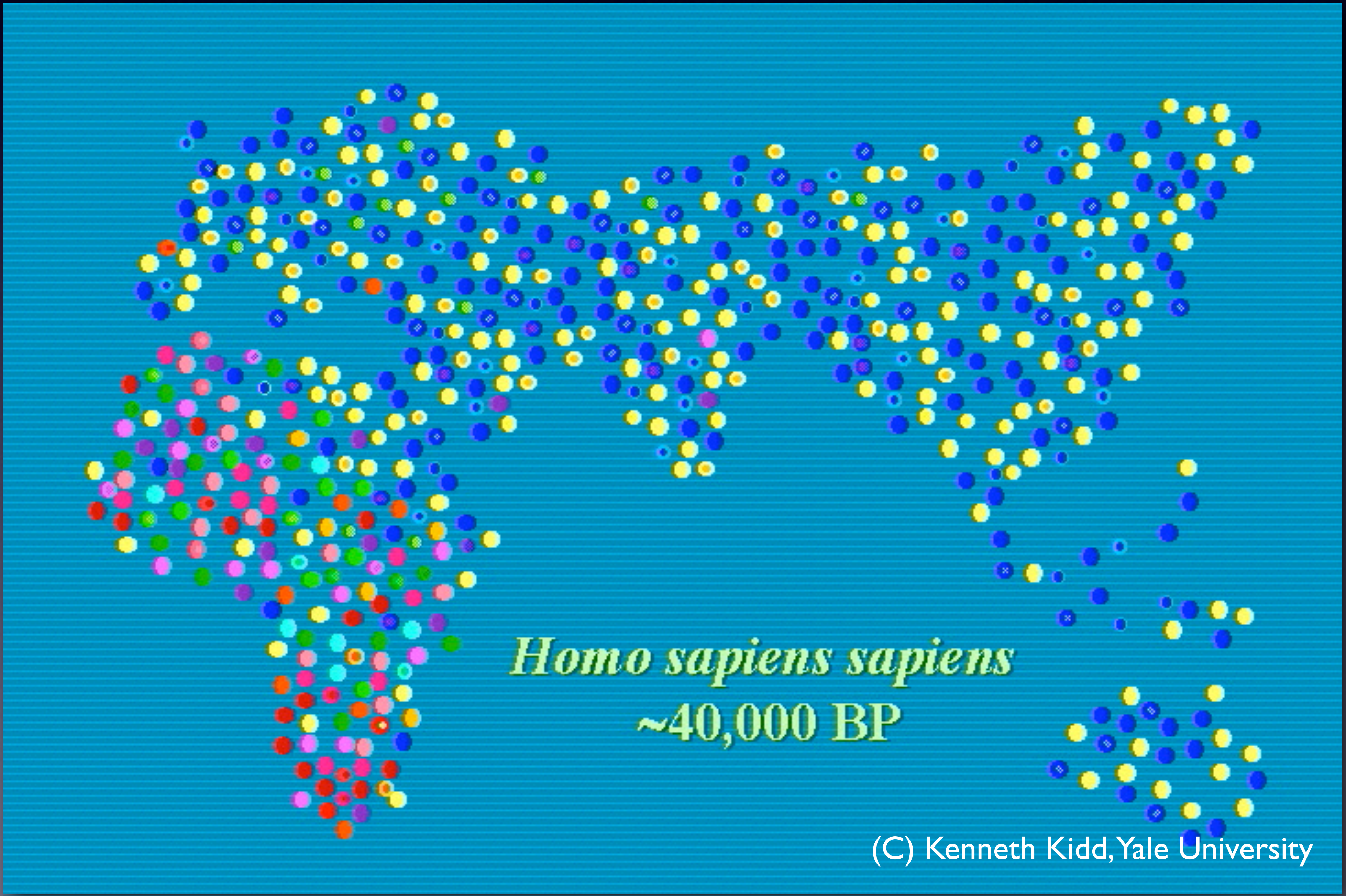
Homo sapiens sapiens
colonizing south west Asia

(C) Kenneth Kidd, Yale University

~100,000 BP

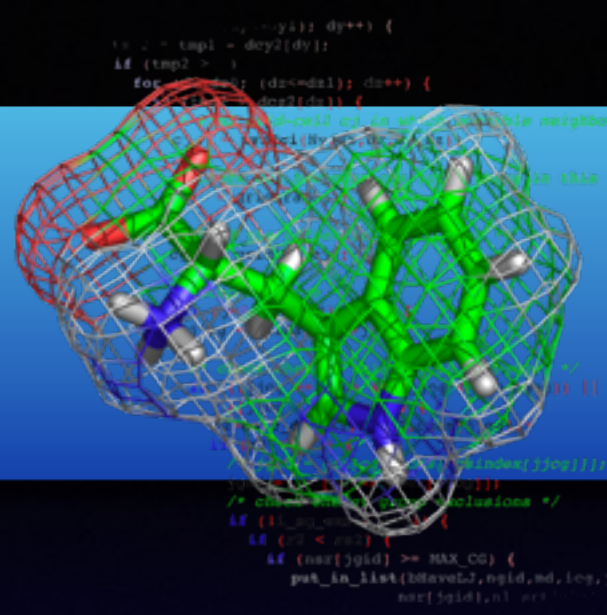
Human evolution





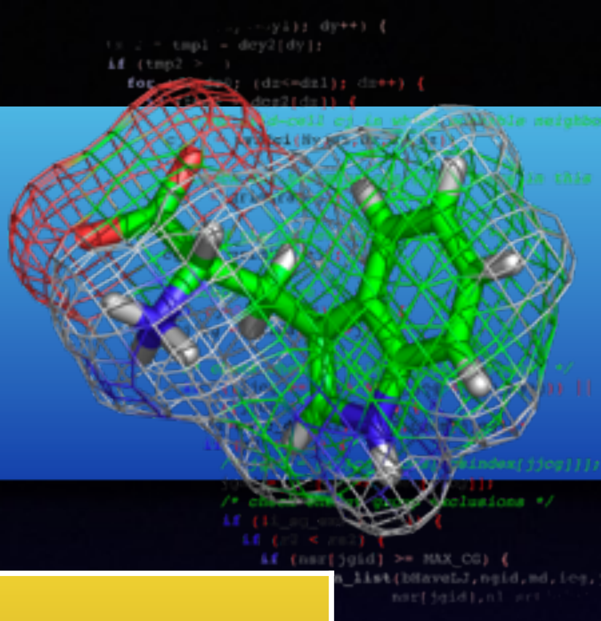
(C) Kenneth Kidd, Yale University

BRCA genes



- BRCA1/BRCA2 (=BReast CAncer)
- Some DNA mutations in these mean 85% risk of developing breast cancer
- New efficient genetic tests for screening
    - Frequent mamograms if positive
    - Possibly preventive breast removal

Nucleotides determine the amino acid sequence



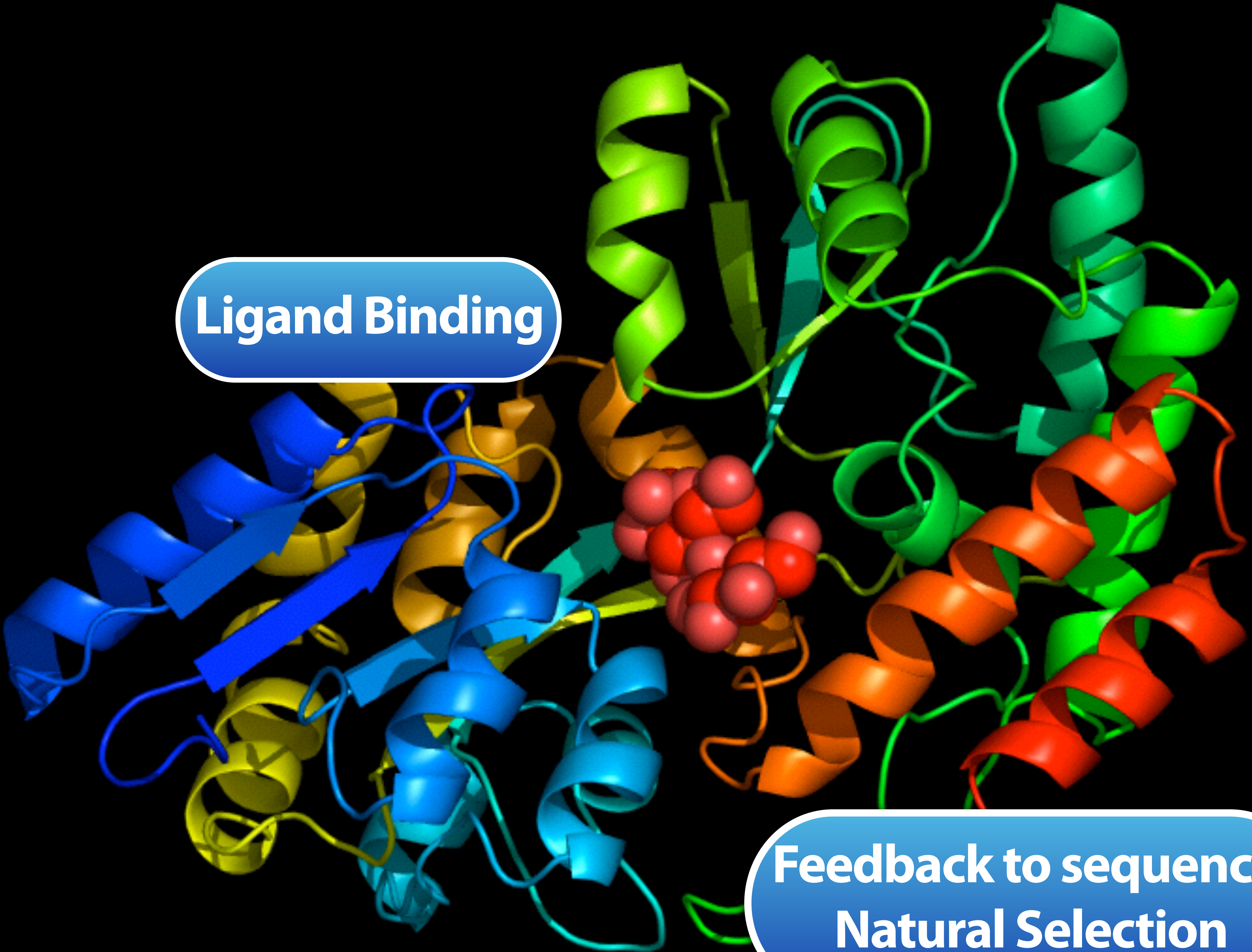
1

	T	C	A	G	
T	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	T C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	T C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	T C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	T C A G

2

3

```
  1 KIEEGKLVIW INGDKGYNGL AEVGKKFEKD TGIKVTVEHP

 41 DKLEEKFPQV AATGDGPDII FWAHDRFGGY AQSGLLAEIT

 81 PDKAFQDKLY PFTWDAVRYN GKLIAYPIAV EALSLIYNKD

121 LLPNPPKTWE EIPALDKELK AKGKSALMFN LQEPYFTWPL

161 IAADGGYAFK YENGKYDIKD VGVDNAGAKA GLTFLVDLIK

201 NKHMNADTDY SIAEAAFNKG ETAMTINGPW AWSNIDTSKV

241 NYGVTVLPTF KGQPSKPFVG VLSAGINAAS PNKELAKEFL

301 ENYLLTDEGL EAVNKDKPLG AVALKSYEEE LAKDPRIAAT

341 MENAQKGEIM PNIPQMSAFW YAVRTAVINA ASGRQTVDEA

361 LKDAQTRITK
```



Ligand Binding

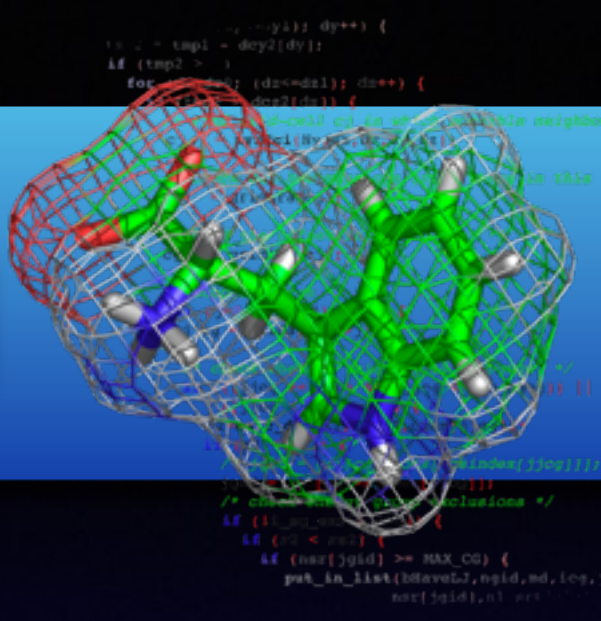
Feedback to sequence: Natural Selection

Sequence

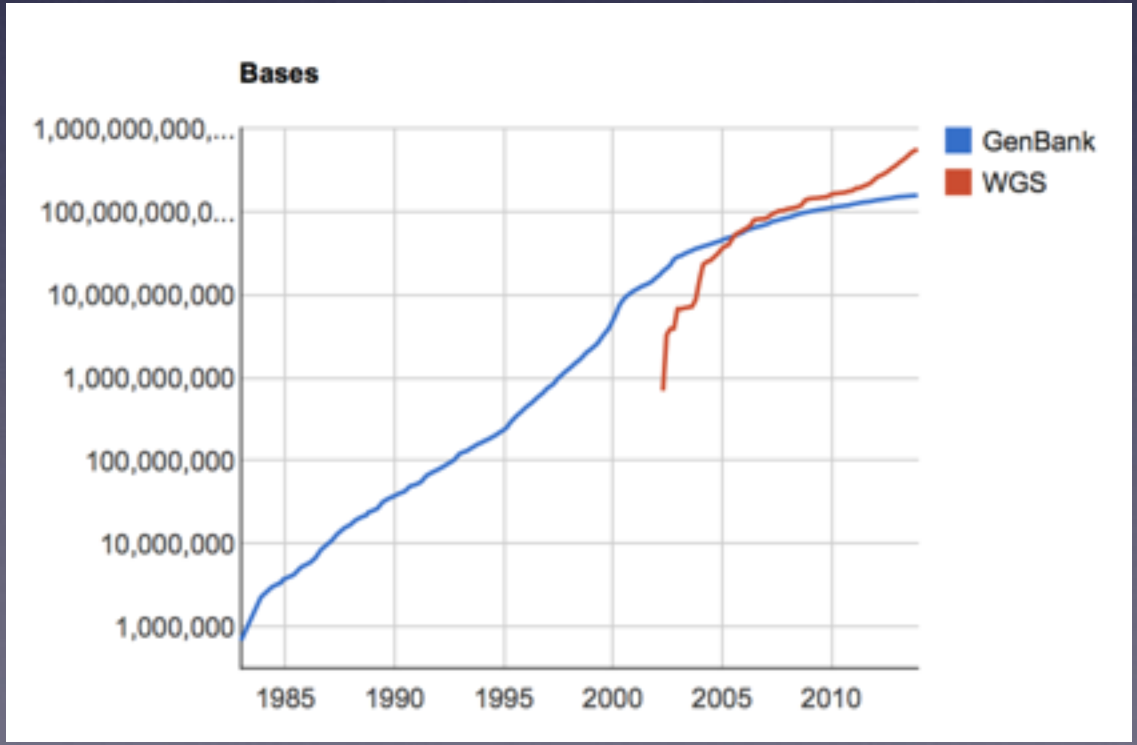
Structure

Function

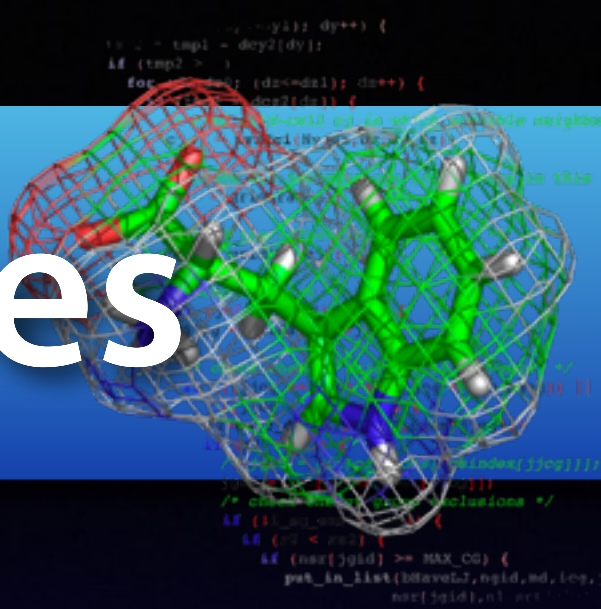
Genome Sequencing



- In total 184,938,063,614 DNA bases from 179,295,769 different sequence records (Dec 2014)
- 12,367 genomes sequenced completely (Jan 9, 2014)
- Over 20,000 partially complete
  - 436 metagenomic studies
- www.genomesonline.org

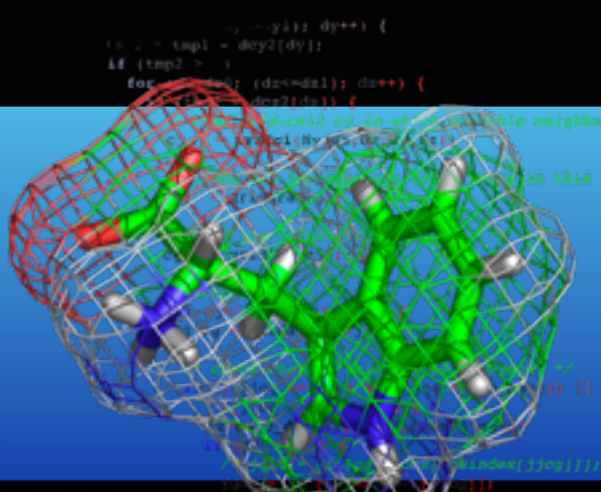


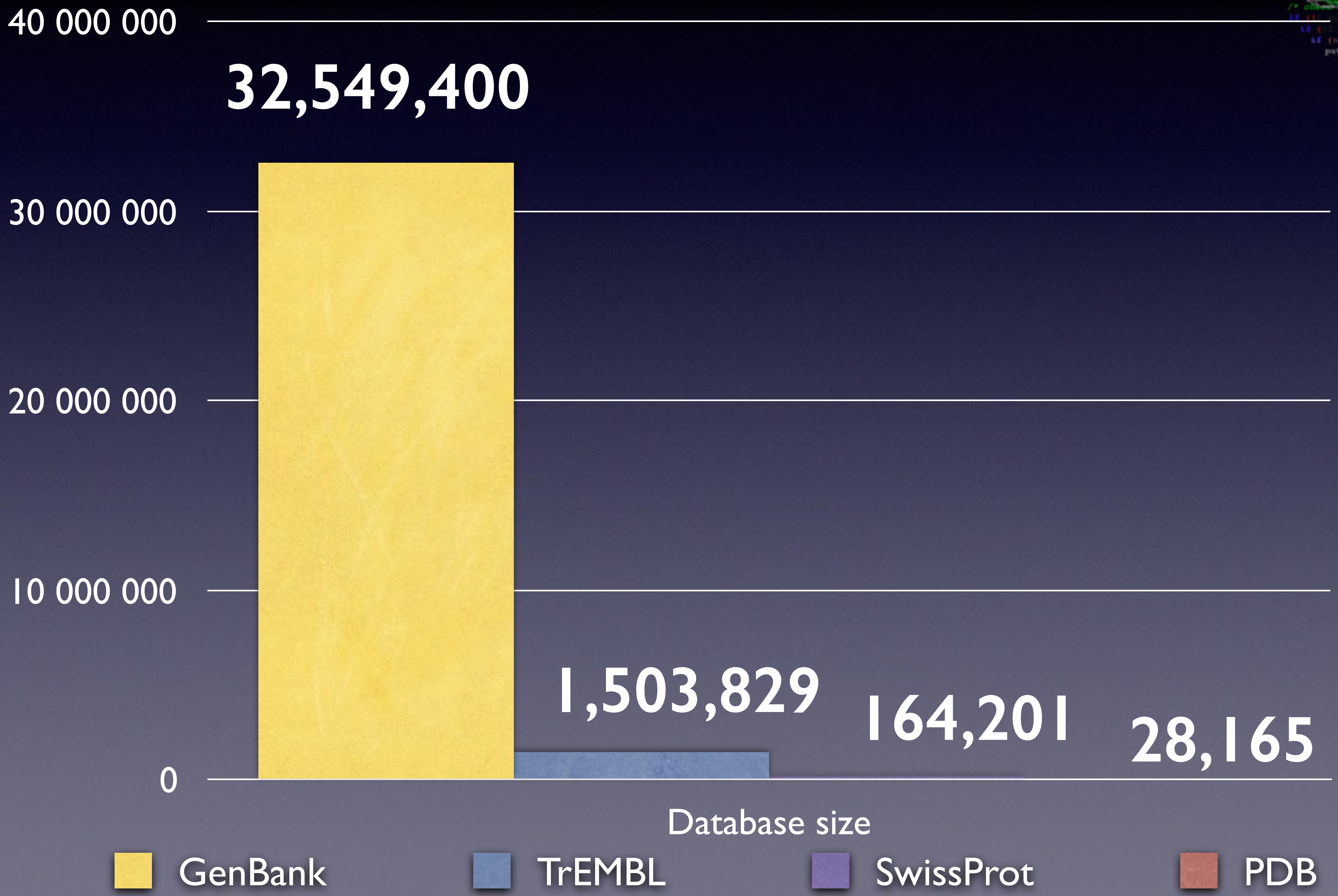
Some Public Databases



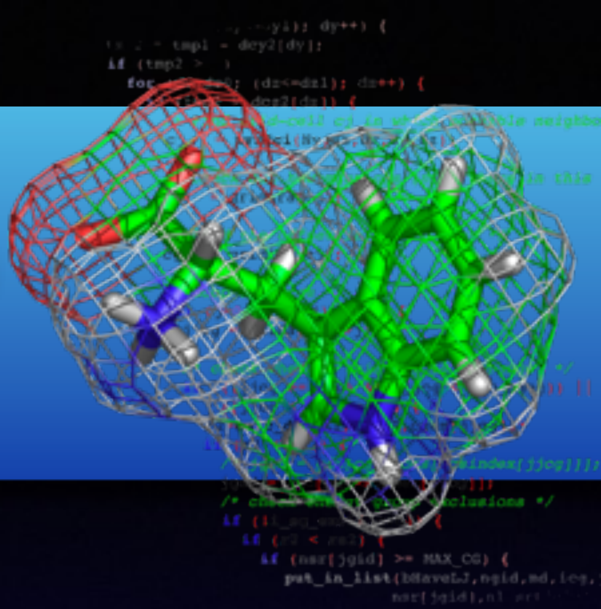
- GenBank (NCBI) - genome sequences
  - Huge, but lots of junk
- SwissProt/TrEMBL - Annotated seqs.
  - Genes known to code for proteins
- Protein Data Bank (PDB)
  - Coordinates of 3D protein structures

Old data from 2007, but to show relative size:



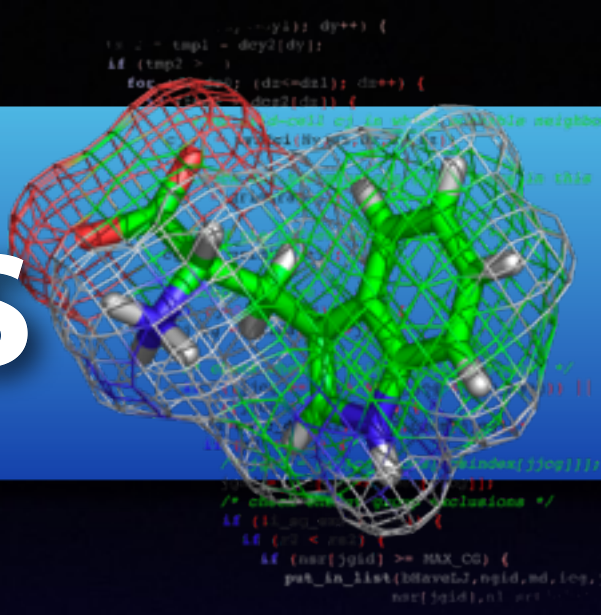


Sequence Similarity



- Natural selection:
  - Random mutation/insertion/deletion
  - Survival of the fittest
- Evolution from older ancestors
- Proteins (genes) from a common ancestor are called *Homologs*

Paralogs / Orthologs



- ***Paralogs:*** Homologous proteins that perform different (but related) functions in the same organism
- ***Orthologs:*** Homologous proteins that perform the same (or very similar) function in different organisms

Myoglobin from 9 species

Are these paralogs or orthologs?

```
MYHU   ..MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPE..
MYCZ   ...GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPE..
MYMQV  ...GLSDGEWQLVLNIWGKVEADIPSHGQEVLISLFKGHPE..
MYOY   ...GLSDAEWQLVLNVWGKVEADIPGHGQDVLIRLFKGHPE..
MYFXBE ...GLSDGEWQIVLNIWGKVETDLAGHGQEVLIRLFKNHPE..
MYDG   ...GLSDGEWQIVLNIWGKVETDLAGHGQEVLIRLFKNHPE..
MYWHL  ...GLSDGEWQLVLNVWGKVEADLAGHGQDILIRLFKGHPE..
MYPN   ...GLNDQEWQQVLTMWGKVESDLAGHGHAVLMRLFKSHPE..
MYTUY  .......ADFDAVLKCWGPVEADYTTMGGLVLTRLFKEHPE..
Consensus GLSDGewQL  N   K  A   GH QEv IR   G
```

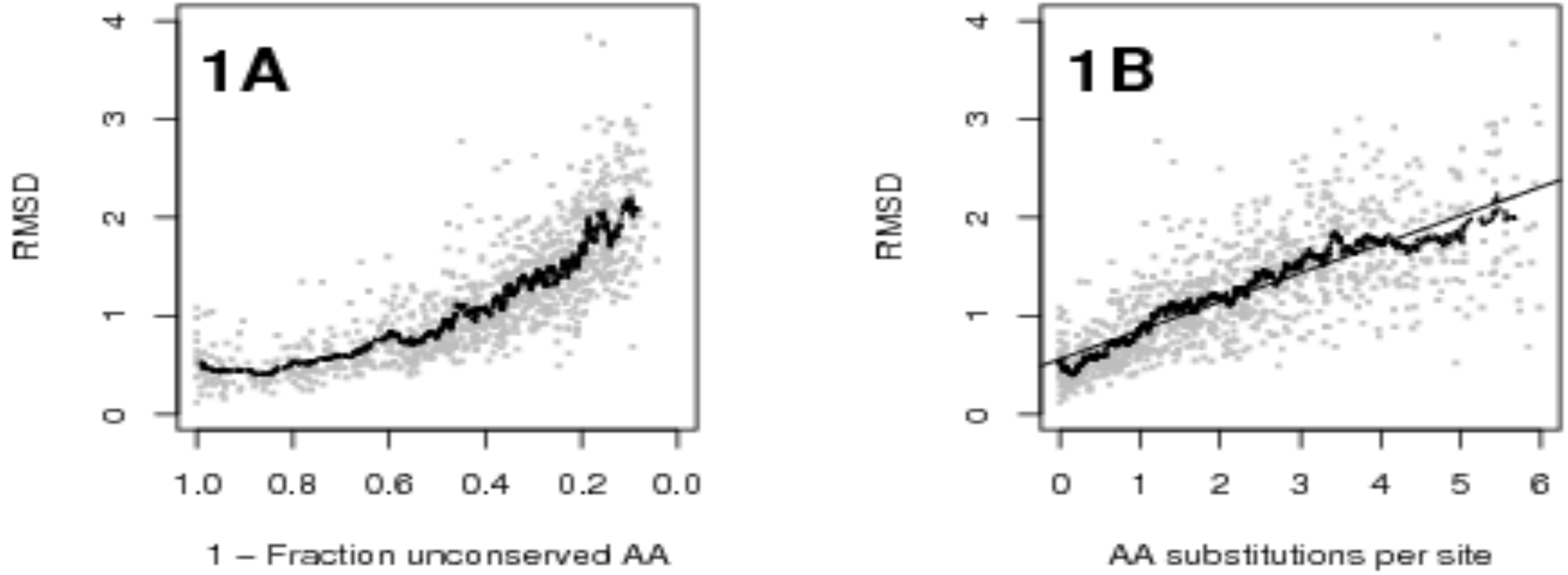
Structure distance: RMSD

- Defined almost like a standard deviation

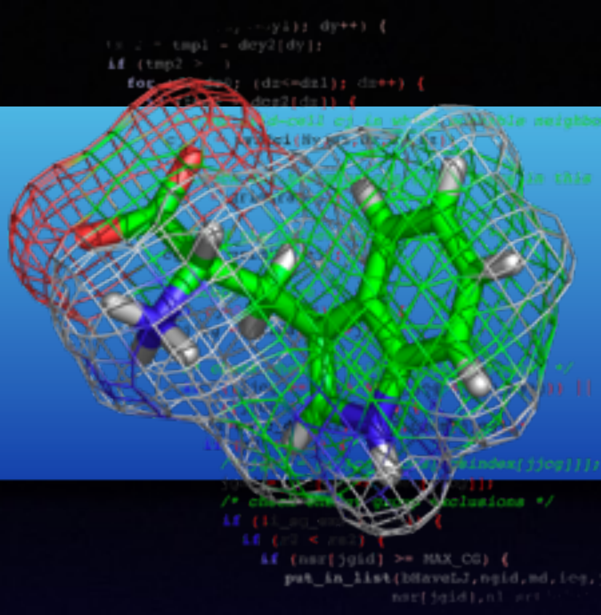
$$\sum_{i=1}^{n}\sqrt{\frac{(x_a-x_b)^2+(y_a-y_b)^2+(z_a-z_b)^2}{n}}$$

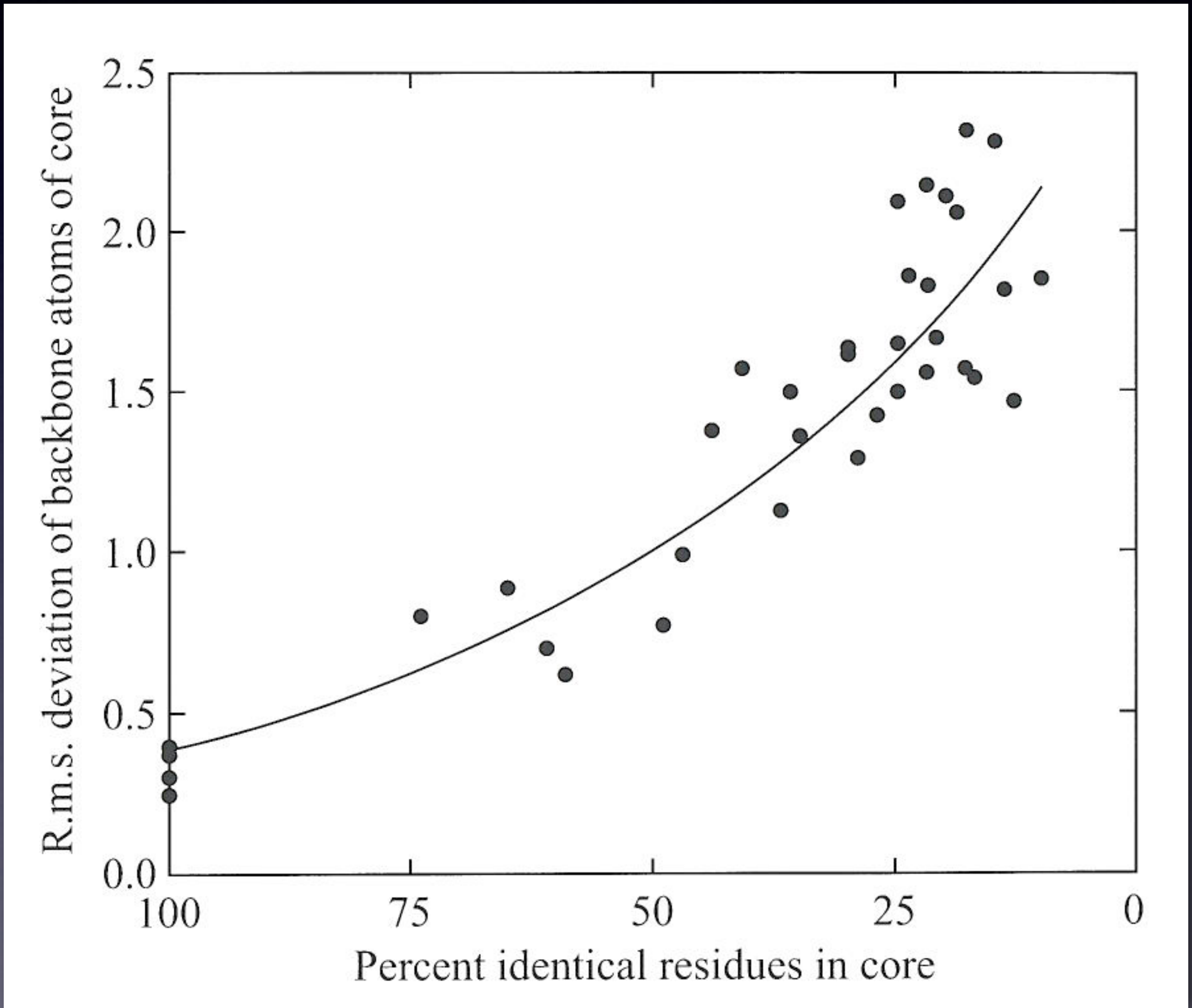
- Average displacement of atoms
- X-ray: 0.2 Å NMR: 1-2 Å
- Homology models: 1-3 Å

Structural change depends on evolutionary distance!



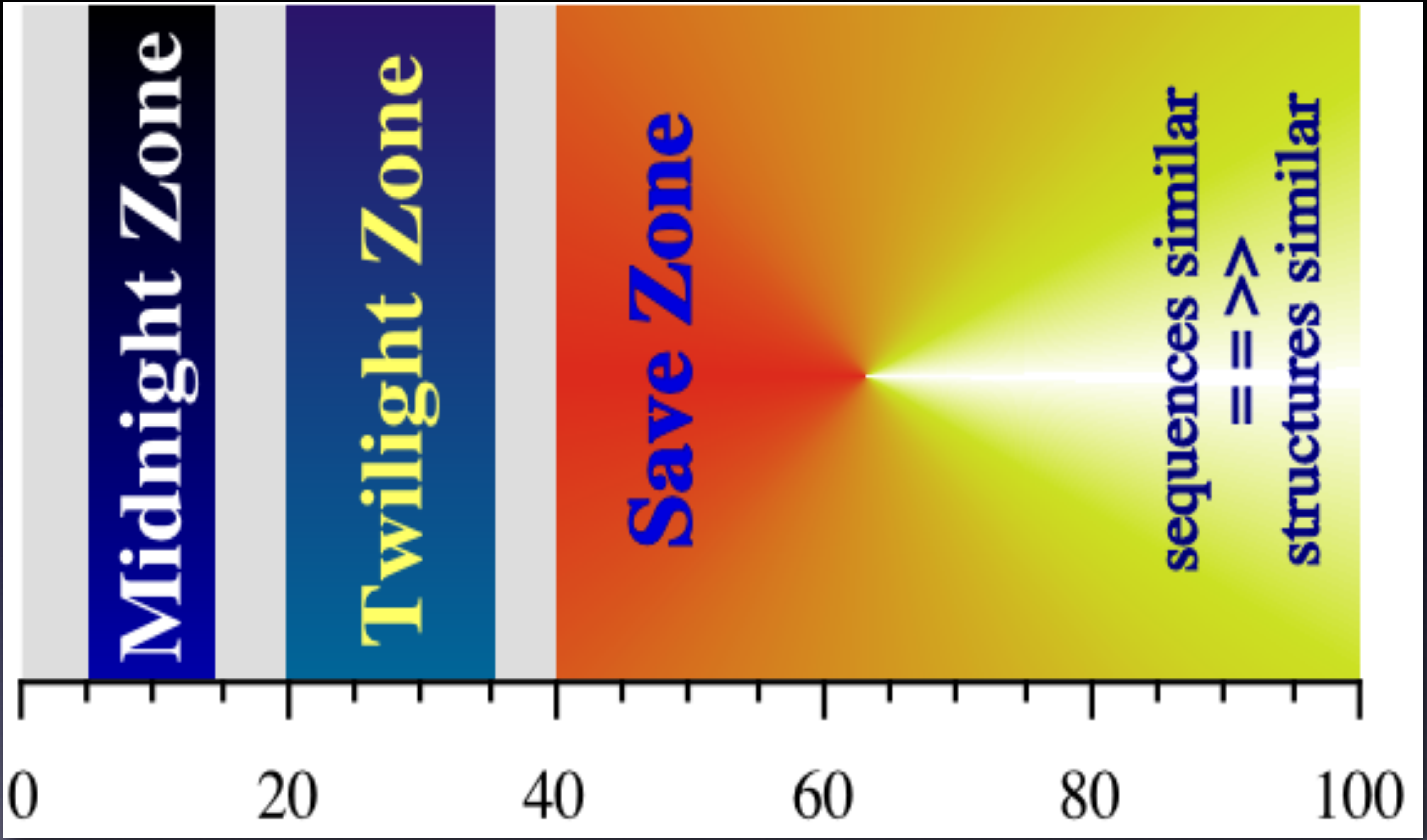
Homology is useful for structure prediction




R.m.s. deviation of backbone atoms of core
2.5
2.0
1.5
1.0
0.5
0.0
100
75
50
25
0
Percent identical residues in core

If we know the structure of a homologous protein, we might be able to build a model based on this relative!

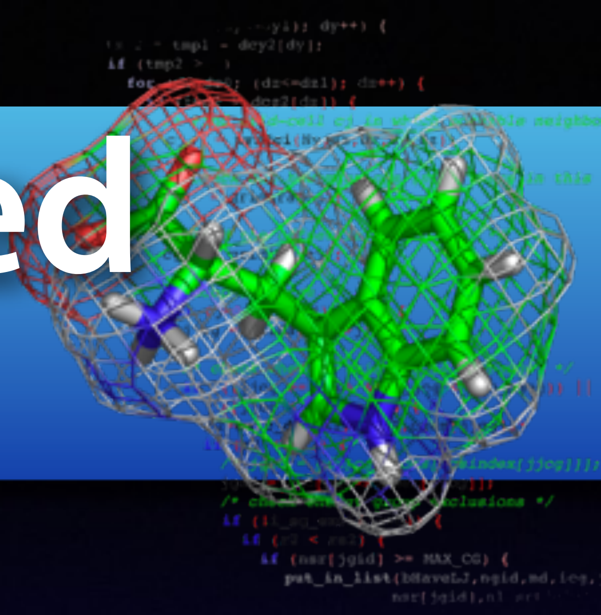
Impossible *Hard* *Easy*



Sequence identity

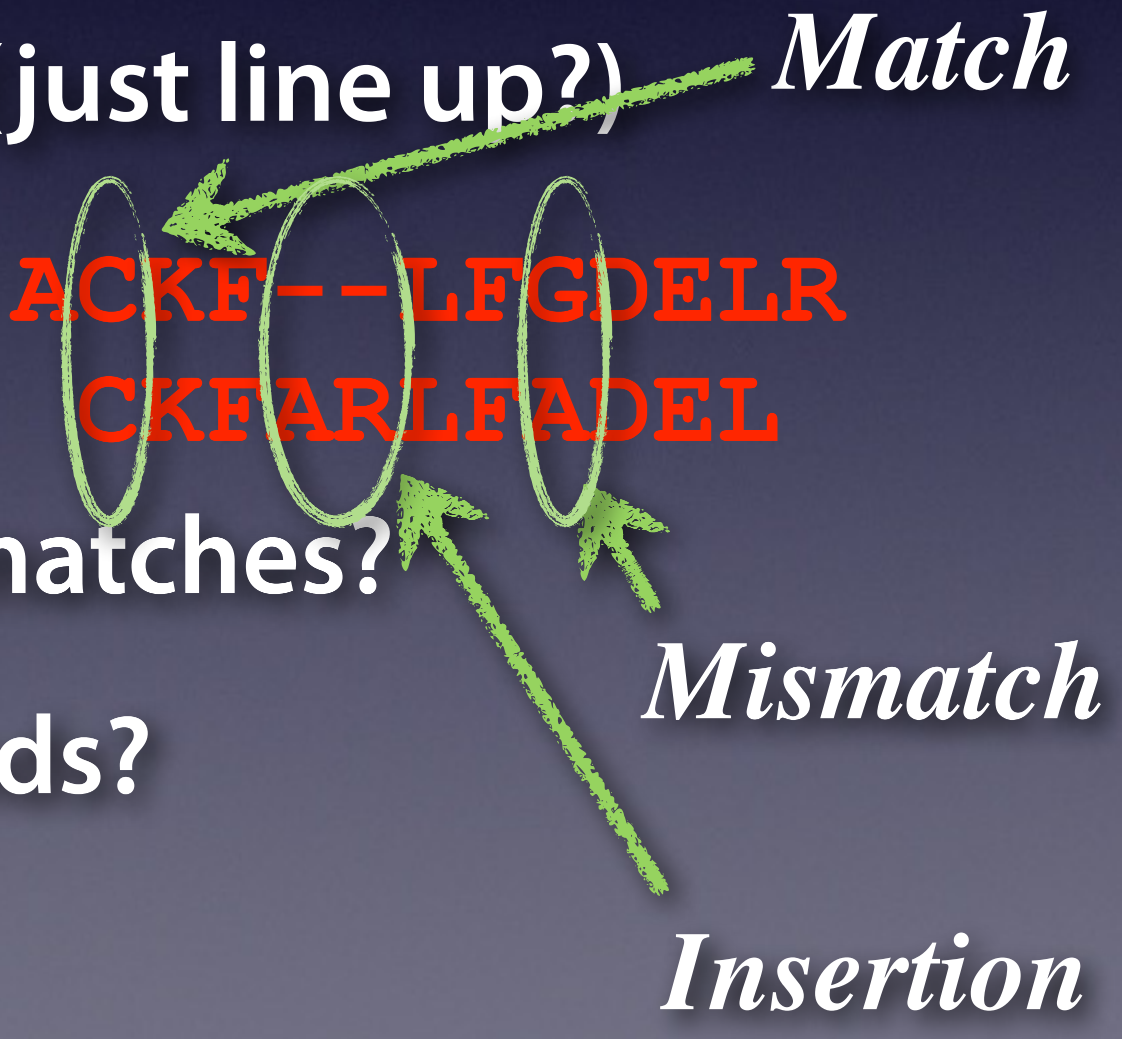
But: Proteins are either homologs or not - the question is only when we can detect it! (You can't be 50% siblings)

Homology can be detected from sequence similarity



- How do we locate & assess similarities?
- Alignment of sequences (just line up?)

```
ACKFLFGDELR
CKFARLFADEL
```



```
ACKF--LFGDELR
  CKFARLFADEL
```

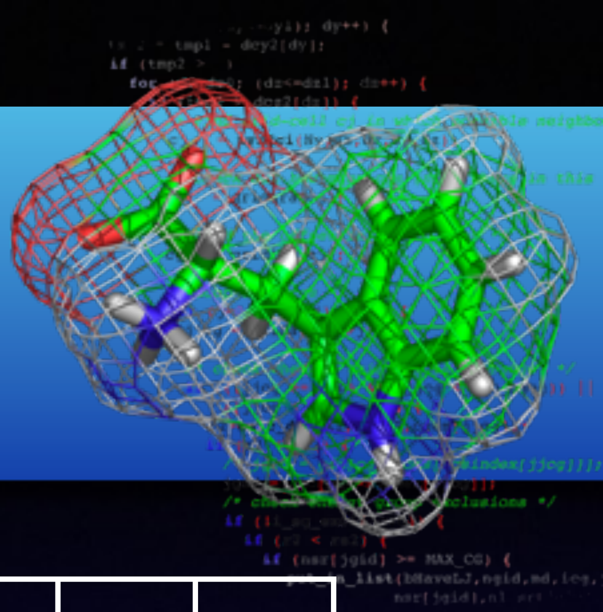
Match

Mismatch

Insertion

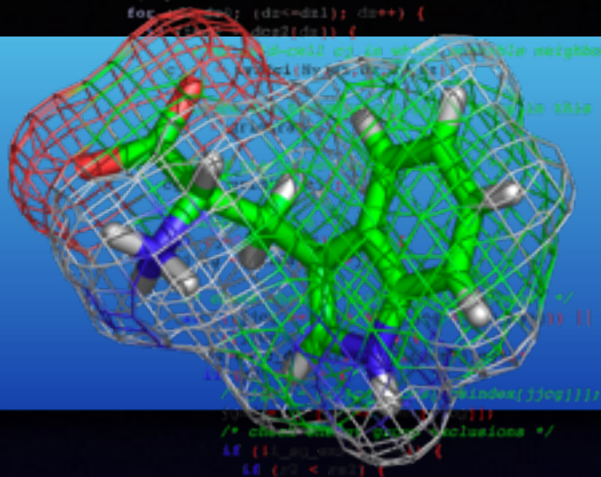
- What do we do with mismatches?
- Insertions? Deletions? Ends?

A Simple Dot Plot



	A	C	K	F	L	F	G	D	E	L	R
C		🟥									
K			🟥								🟥
F				🟥		🟥					
A	🟥										
R											🟥
L					🟥					🟥	
F				🟥		🟥					
G							🟥				
D								🟥			
E									🟥		
L					🟥					🟥	

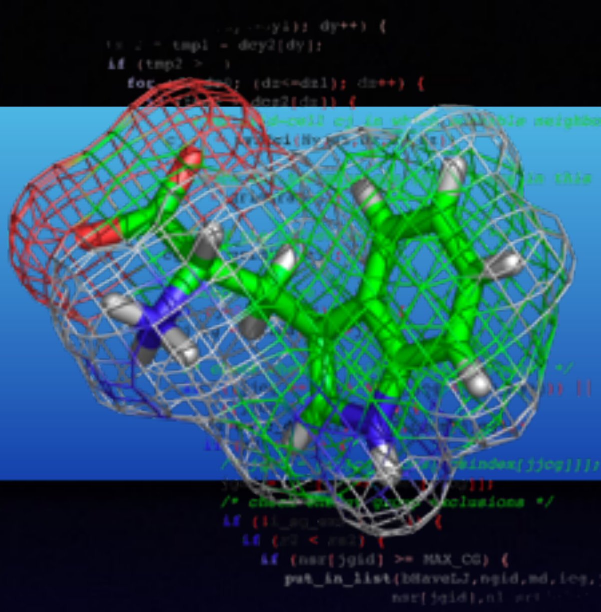
Filtered Dot Plot



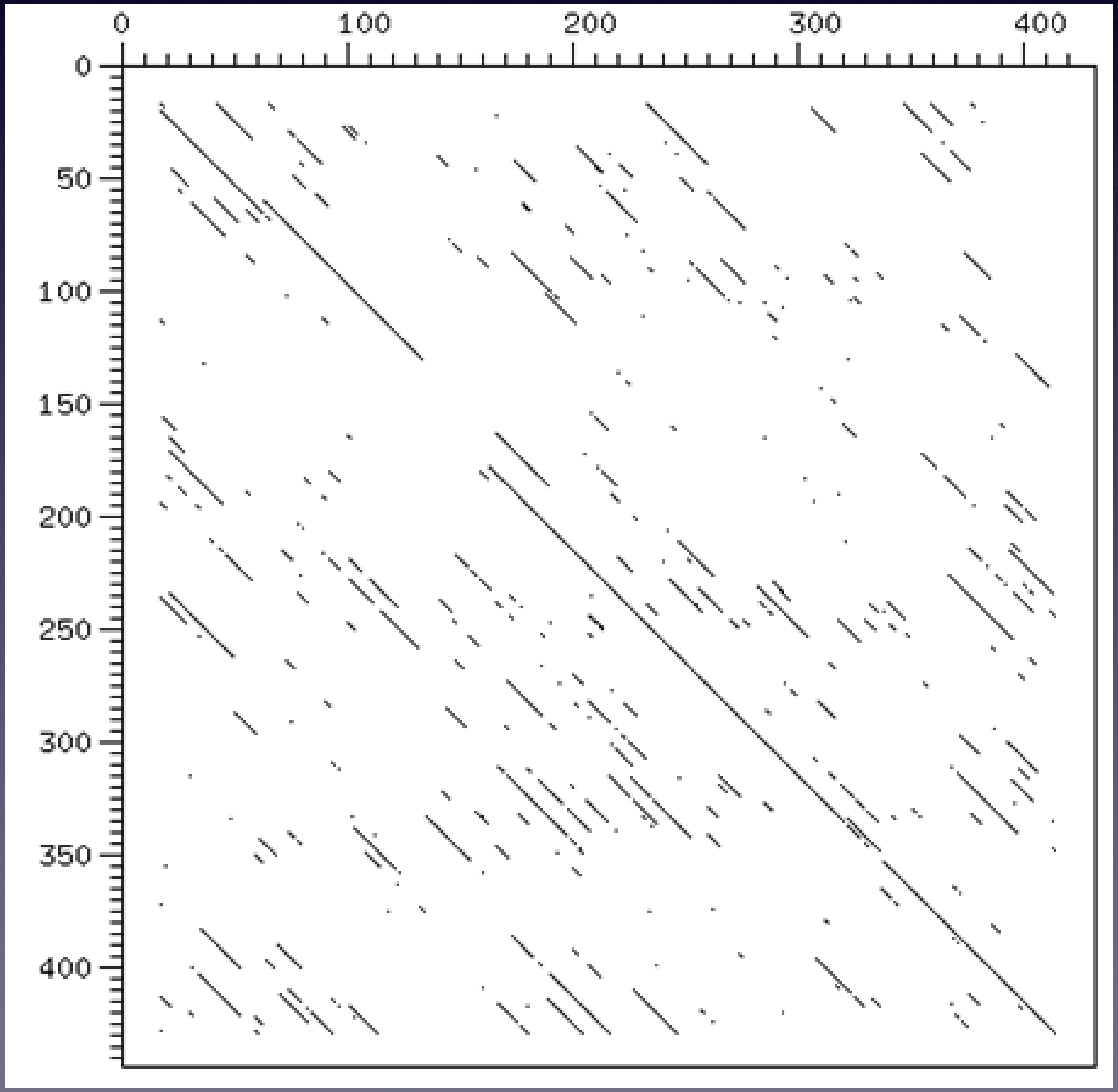
Remove all hits shorter than three positions

	A	C	K	F	L	F	G	D	E	L	R
C		🟥									
K			🟥								
F				🟥							
A											
R											
L					🟥						
F						🟥					
G							🟥				
D								🟥			
E									🟥		
L										🟥	

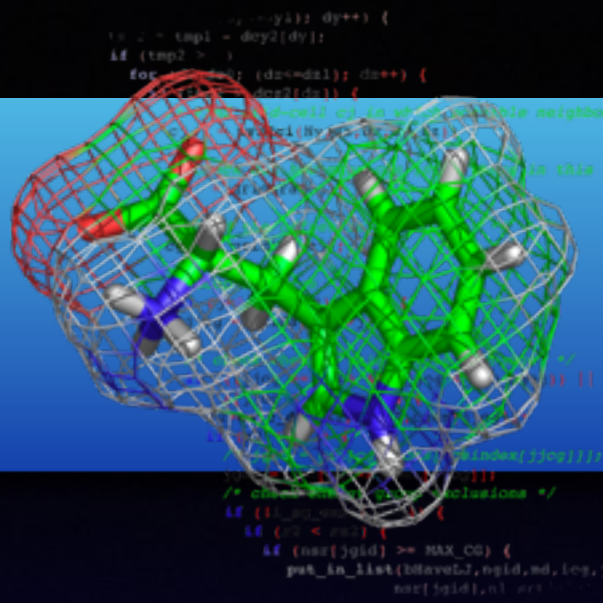
Realistic Dot Plot



- Hemoglobin
  α chain vs. β chain
- Lots of false hits
- Hard to quantify

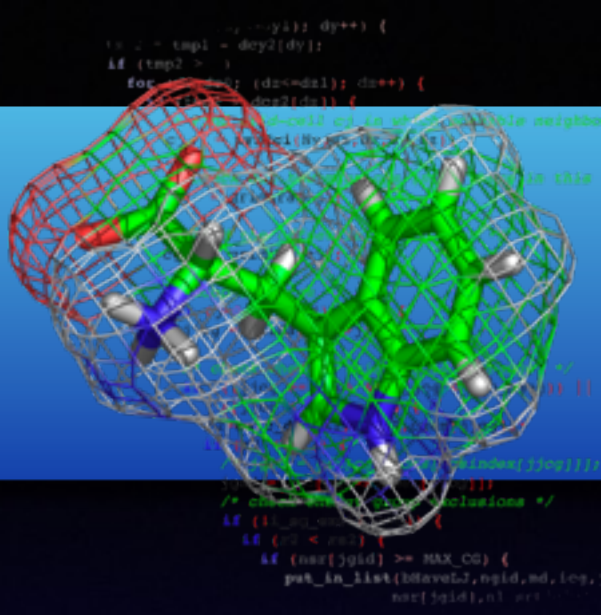


Quantify Similarity



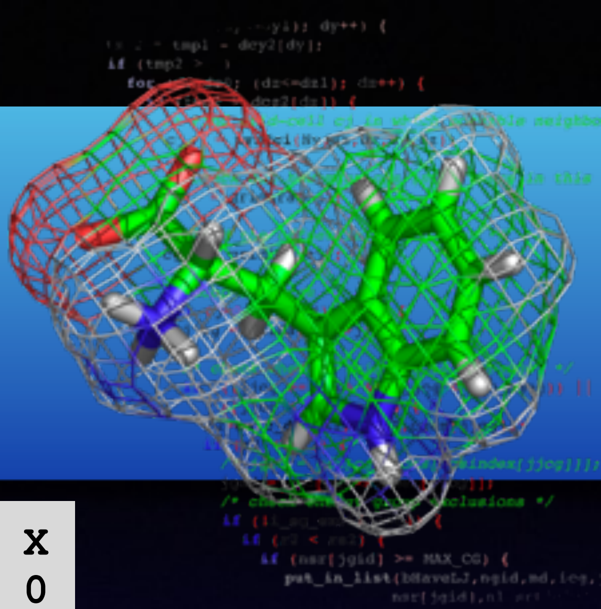
- What do we mean by "similar"?
  - Must it cover the whole sequence?
  - Do we allow gaps?
- Any way of pairing residues/gaps in the sequences is called an *alignment*
  - Good alignments maximize similarity without adding too many gaps

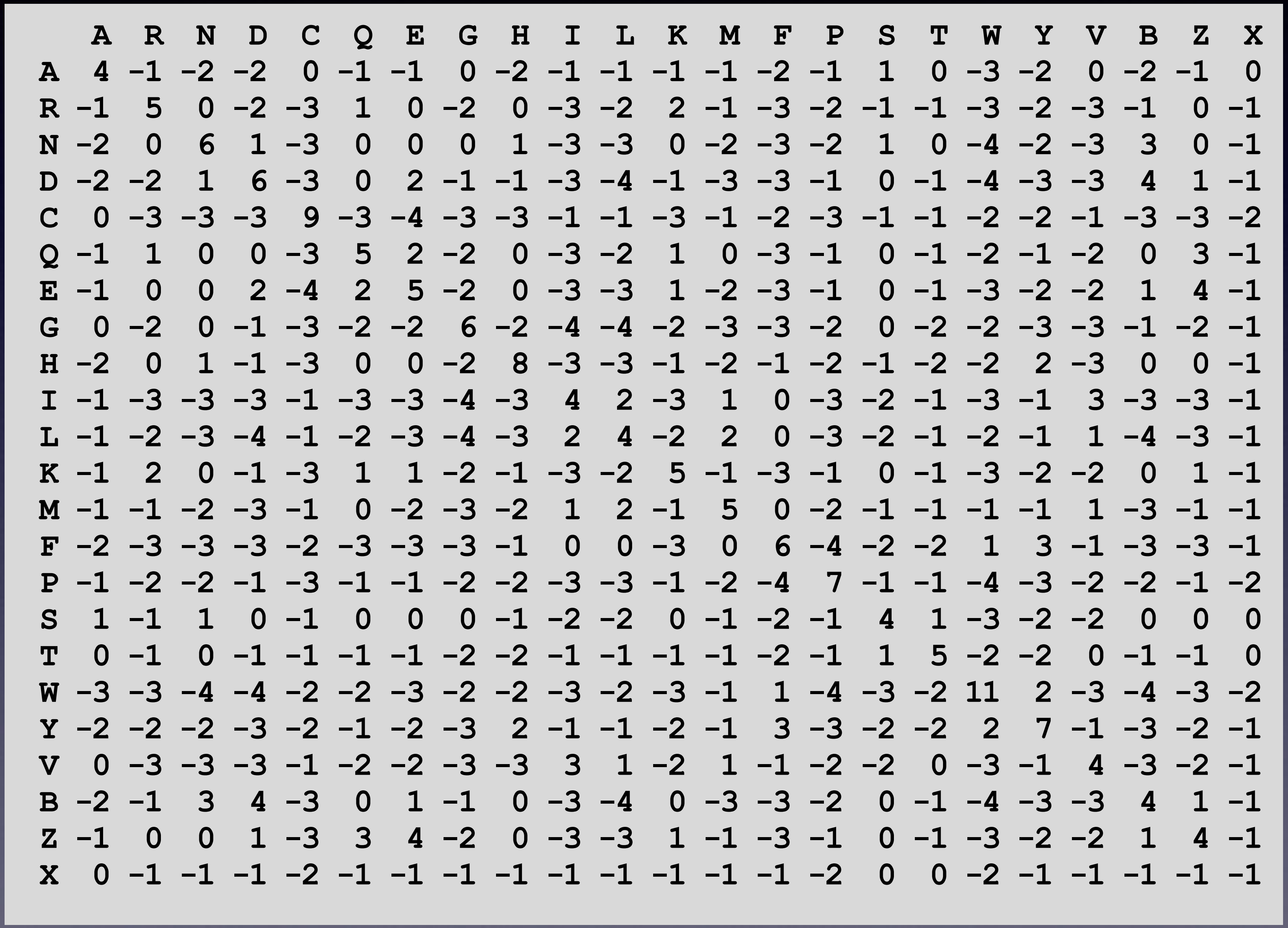
Similarity Measures



- Amino acid substitution scores
  - Conserved amino acids (very good)
  - Similar amino acids (OK)
  - Neutral
  - Significantly different (very bad)
- Substitution scores: 20*20 matrix
- Example matrices: PAM250, BLOSUM62

BLOSUM62

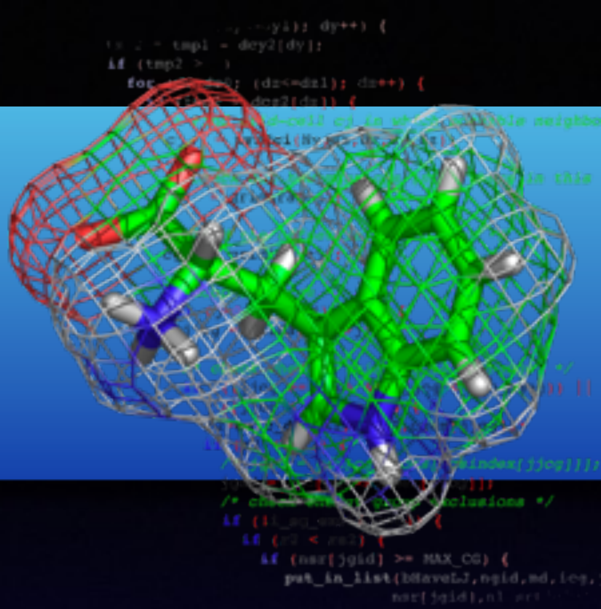




	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1
C	0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

B=D or N (Asp or Asn)   Z=E or Q
(Glu or Gln)   X=any amino acid

Alignment Scoring



- We *could* define any scoring we want
- Use a simple setup for two examples: Match=3, Mismatch=-1, Gap=-2

1

```
DEFYWLKKPAGTSVQND
 |||| |      ||||
EEFYWKKPAGTSAVQND
```

Score: 19

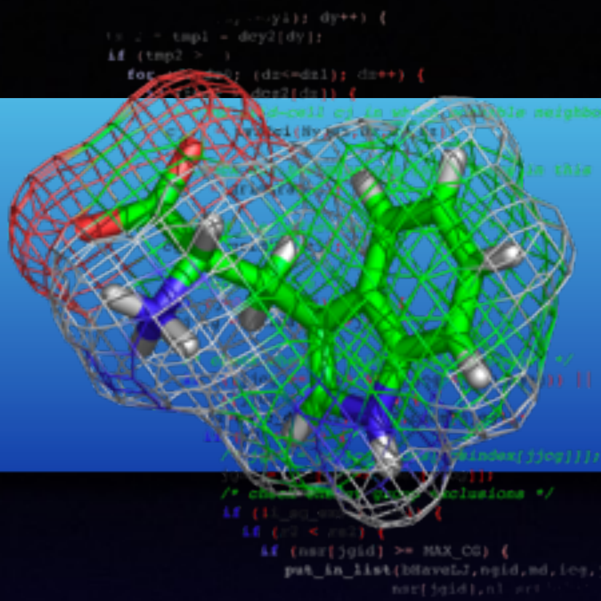
2

```
DEFYWLKKPAGTS-VQND
 |||| ||||||| ||||
EEFYW-KKPAGTSAVQND
```

Better!

Score: 40

Similarity better than identity for alignments!



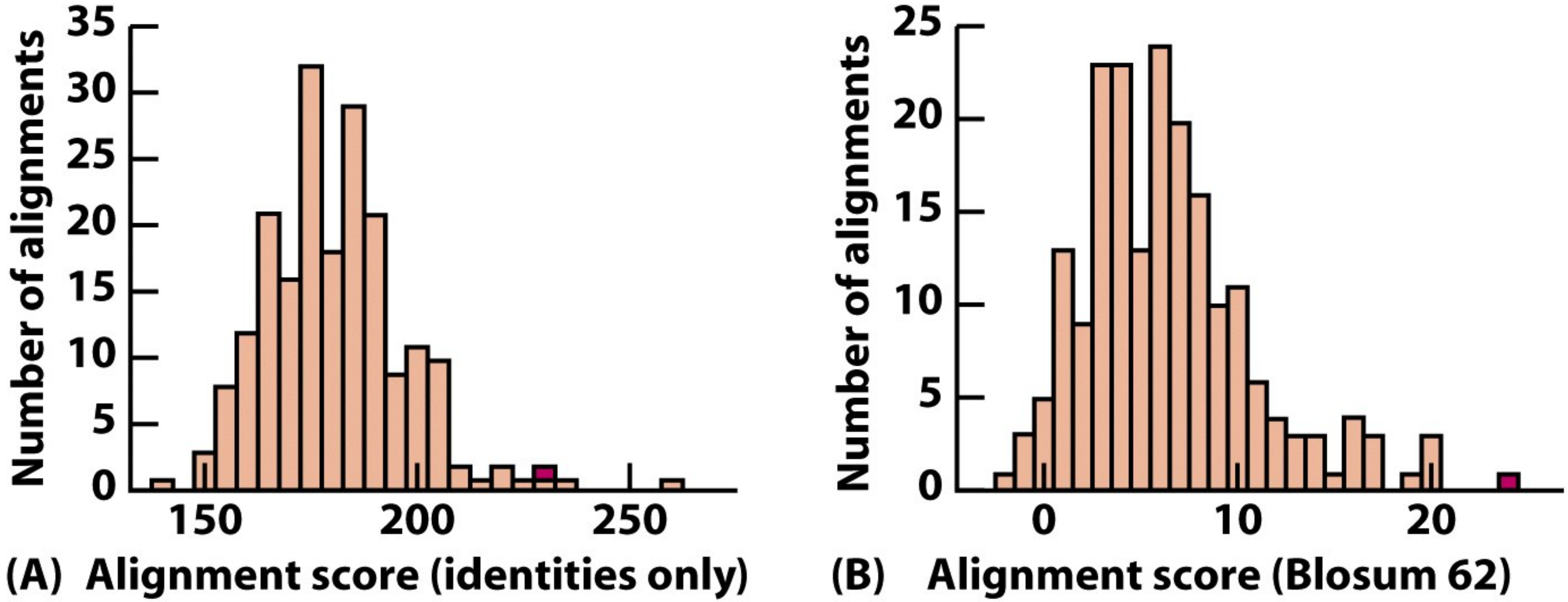
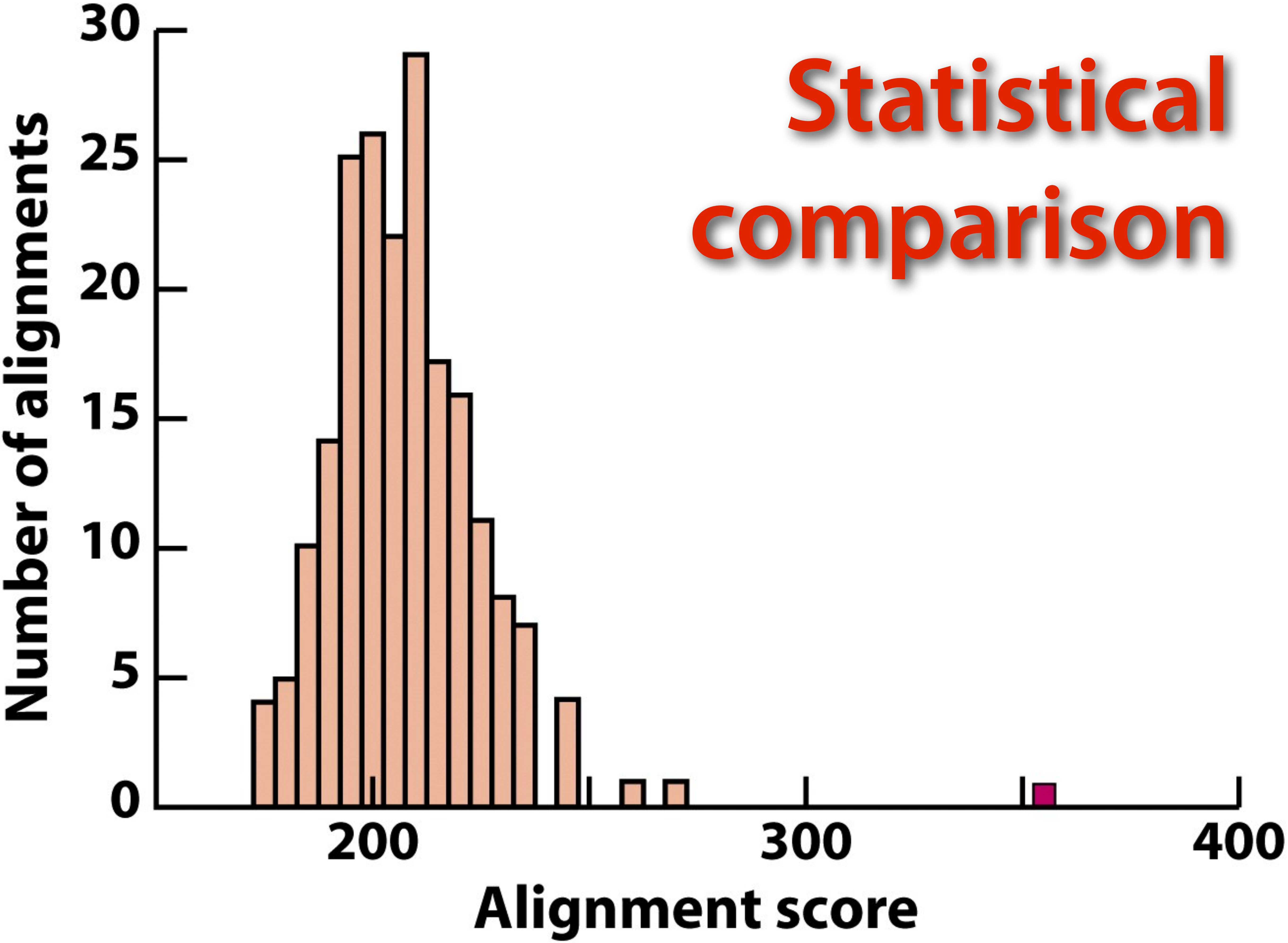
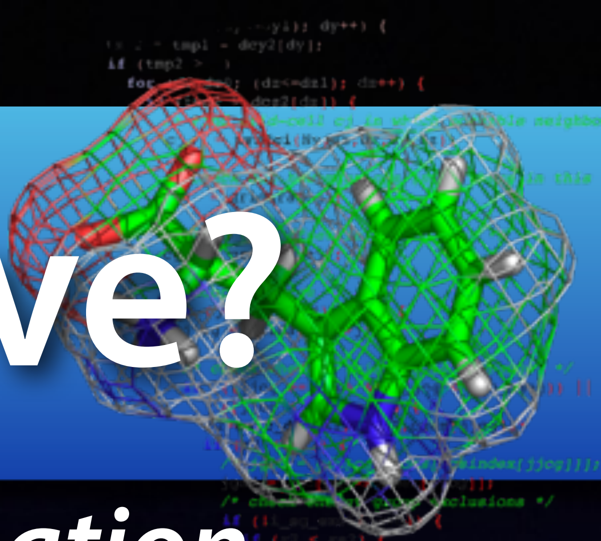


Figure 6-11
Biochemistry, Sixth Edition

Statistical comparison



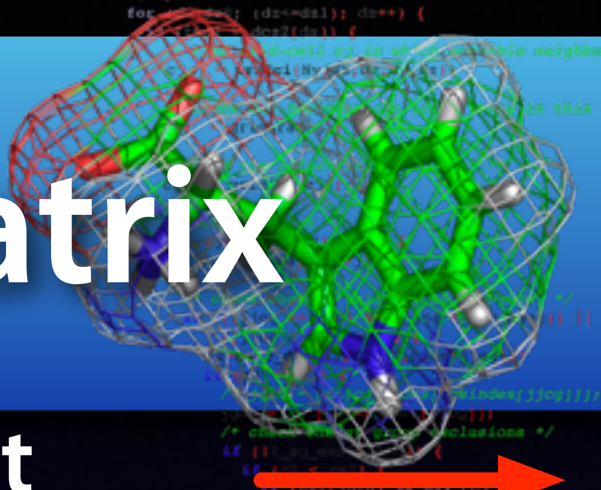
How can we improve?



- The key here was *evolutionary information*
- Can you find and use more such data?

```
MYHU    ..MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPE..
MYCZ    ...GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPE..
MYMQV   ...GLSDGEWQLVLNIWGKVEADIPSHGQEVLISLFKGHPE..
MYOY    ...GLSDAEWQLVLNVWGKVEADIPGHGQDVLIRLFKGHPE..
MYFXBE  ...GLSDGEWQIVLNIWGKVETDLAGHGQEVLIRLFKNHPE..
MYDG    ...GLSDGEWQIVLNIWGKVETDLAGHGQEVLIRLFKNHPE..
MYWHL   ...GLSDGEWQLVLNVWGKVEADLAGHGQDILIRLFKGHPE..
MYPN    ...GLNDQEWQQVLTMWGKVESDLAGHGHAVLMRLFKSHPE..
MYTUY   .......ADFDAVLKCWGPVEADYTTMGGLVLTRLFKEHPE..
Consensus  GLSDGewQL  N   K  A   GH QEv IR   G
```

Position-Specific Scoring Matrix

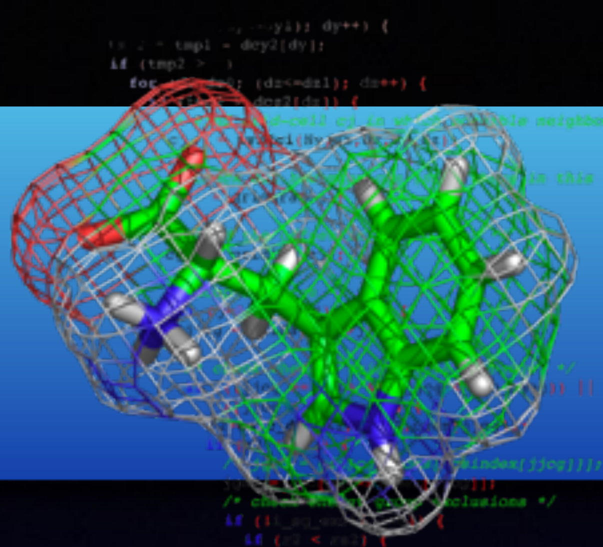


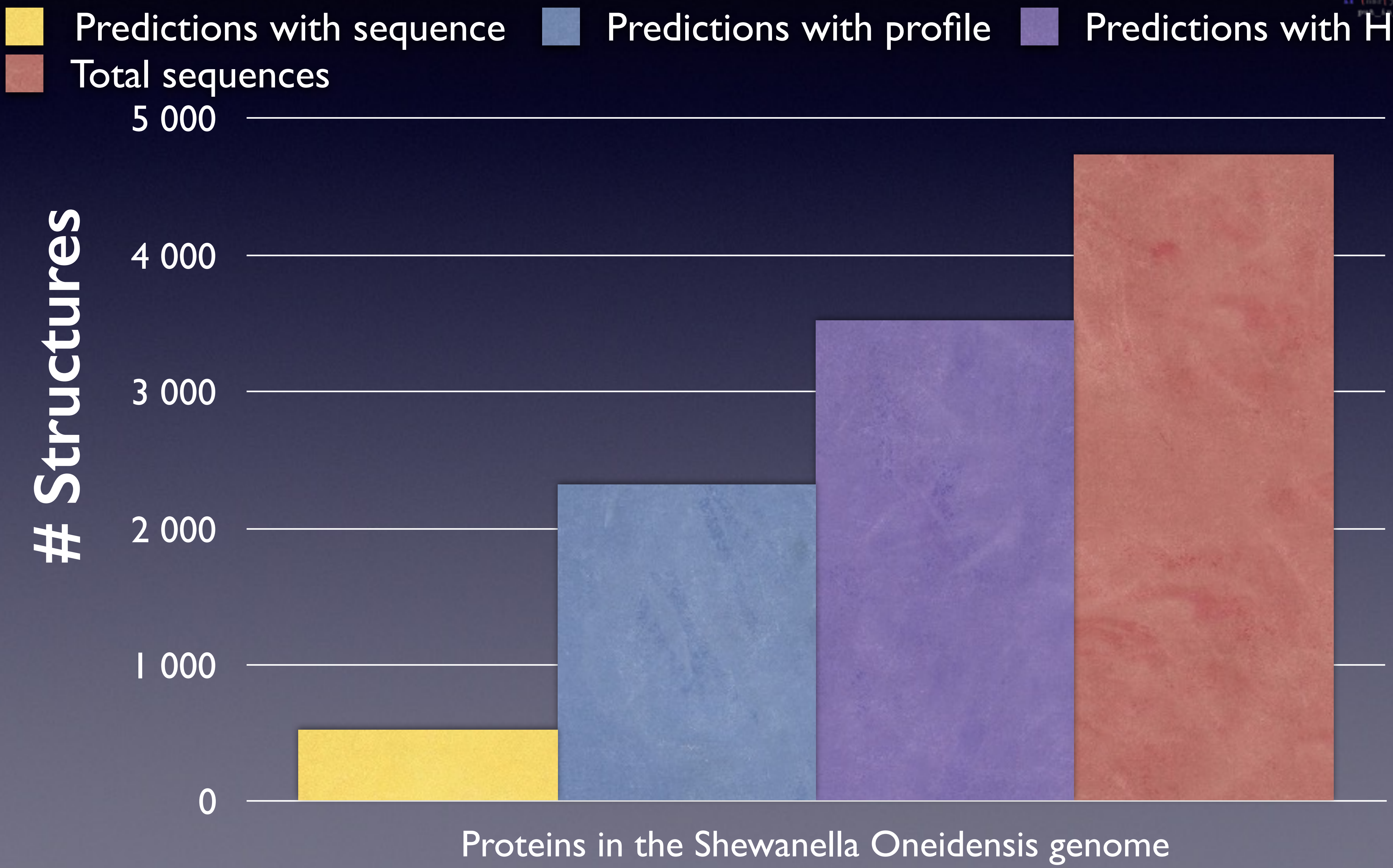
Position in our multiple sequence alignment

Amino acids

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
A	-3	3	0	1	2	1	0	1	-4	2	0	-1	2	-1	-3	-4	2	2	1	1	2
C	4	2	-4	0	1	2	1	6	2	-1	1	2	4	-4	3	-2	-1	4	0	6	4
D	0	-4	-4	-2	-4	1	-2	3	2	3	3	-3	5	2	1	1	3	5	-2	3	5
E	-2	3	2	-3	1	0	-4	2	-3	2	-1	3	-1	0	3	0	2	-1	-3	2	-1
F	1	-2	0	-5	1	0	3	-2	3	3	3	0	-2	2	1	1	3	-2	-5	-2	-2
G	3	3	3	1	0	-2	0	-1	-1	4	-4	-4	7	0	2	0	4	7	1	-1	7
H	0	0	-2	-3	4	-4	-4	5	2	5	4	0	-6	2	1	1	5	-6	-3	5	-6
I	-4	1	0	1	2	0	4	3	-1	2	0	-1	4	1	3	2	2	4	1	3	4
K	3	-4	-1	3	0	0	3	2	-4	-4	-5	3	3	-2	-1	1	-4	3	3	2	3
L	3	1	-4	0	4	0	4	1	3	-1	0	-4	-3	-5	4	1	-1	-3	0	1	-3
M	2	-1	1	2	2	1	-1	3	3	3	2	-1	0	2	0	4	3	0	2	3	0
N	1	2	2	-3	-3	2	1	2	0	-1	-4	-2	0	-3	3	1	-1	0	-3	2	0
P	-4	-5	2	3	3	-2	3	3	0	2	4	1	-1	-2	3	-2	2	-1	3	3	-1
Q	3	2	5	4	0	3	0	5	5	-4	5	0	2	0	-1	-1	-4	2	4	5	2
R	1	-4	1	-2	-4	0	0	4	-1	3	1	-2	0	2	4	5	3	0	-2	4	0
S	2	0	0	2	-2	4	2	1	-1	0	4	-4	-1	-4	2	4	0	-1	2	1	-1
T	1	1	-4	4	0	1	-1	3	1	-4	2	-2	0	-1	-1	-4	-4	0	4	3	0
V	-2	4	2	2	1	5	1	-4	-1	1	1	2	2	0	2	3	1	2	2	-4	2
W	0	5	-4	0	3	3	3	2	3	2	4	3	-1	1	2	-4	2	-1	0	2	-1
Y	5	-2	1	-2	2	4	-1	2	1	3	3	0	-2	-1	1	3	3	-2	-2	2	-2

Search sensitivity

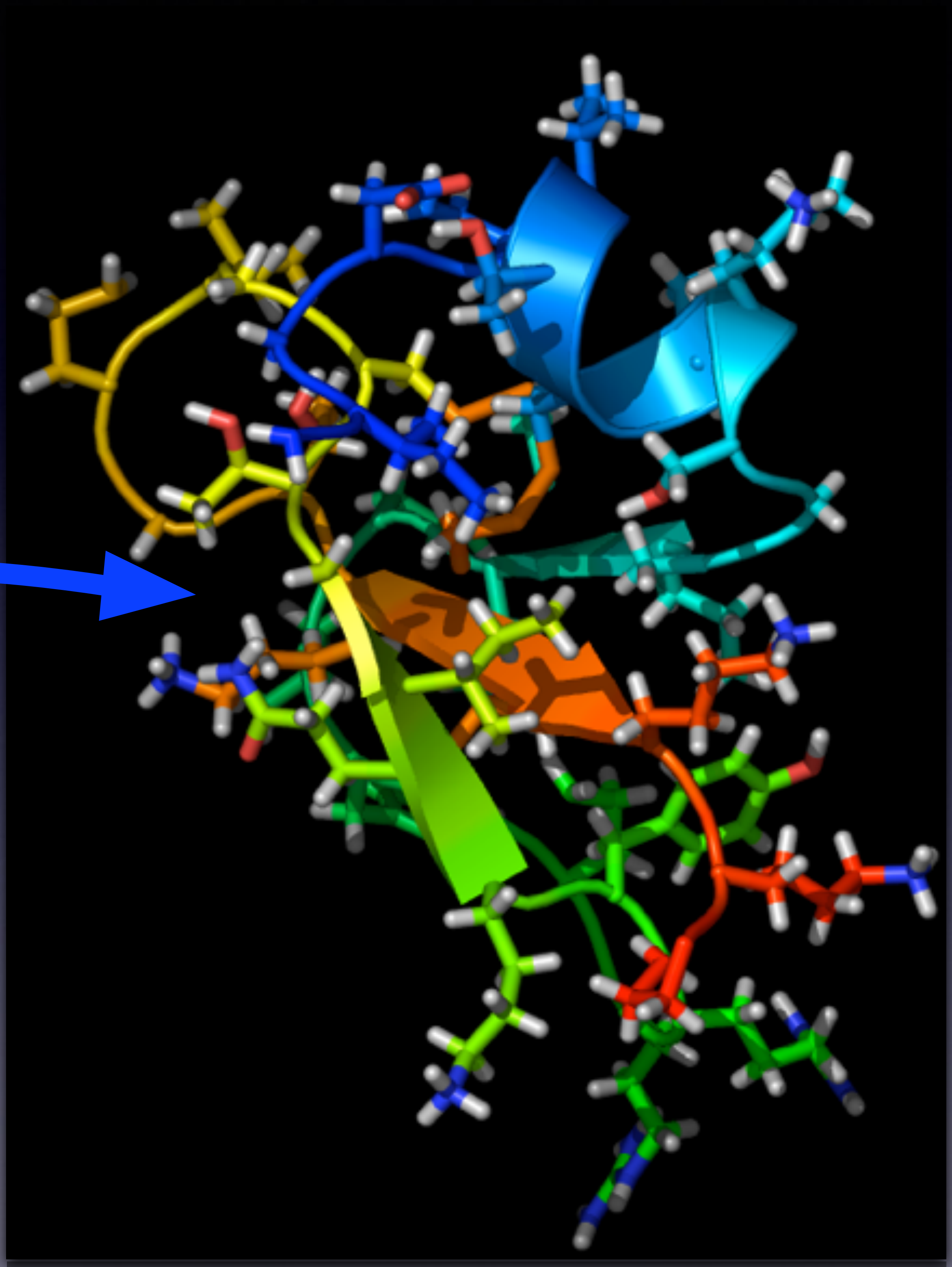




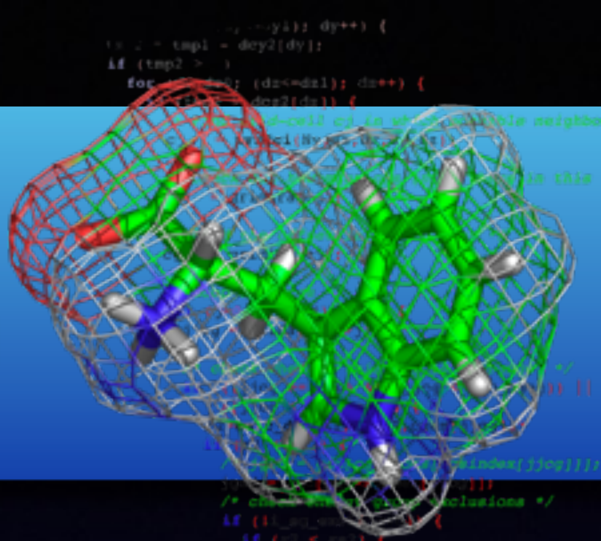
Protein Structure Classification & Prediction

Not quite trivial...

```
KIEEGKLVIW INGDKGYNGL
AEVGKKFEKD TGIKVTVEHP
DKLEEKFPQV AATGDGPDII
FWAHDRFGGY AQSGLLAEIT
PDKAFQDKLY PFTWDAVRYN
GKLIAYPIAV EALSLIYNKD
LLPNPPKTWE EIPALDKELK
AKGKSALMFN LQEPYFTWPL
IAADGGYAFK YENGKYDIKD
VGVDNAGAKA GLTFLVDLIK
NKHMNADTDY SIAEAAFNKG
ETAMTINGPW AWSNIDTSKV
NYGVTVLPTF KGQPSKPFVG
VLSAGINAAS PNKELAKEFL
ENYLLTDEGL EAVNKDKPLG
AVALKSYEEE LAKDPRIAAT
MENAQKGEIM PNIPQMSAFW
YAVRTAVINA ASGRQTVDEA
```

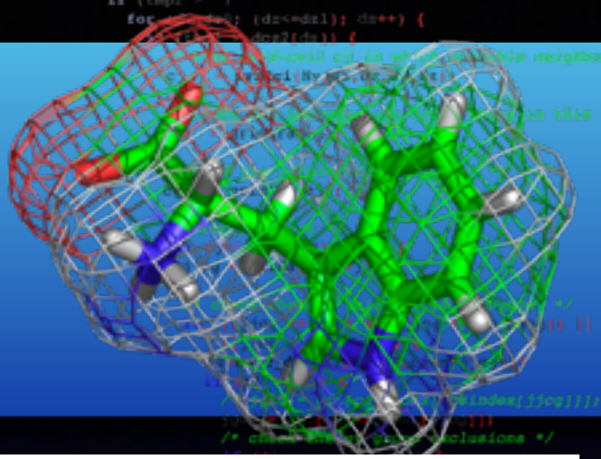


Structure prediction

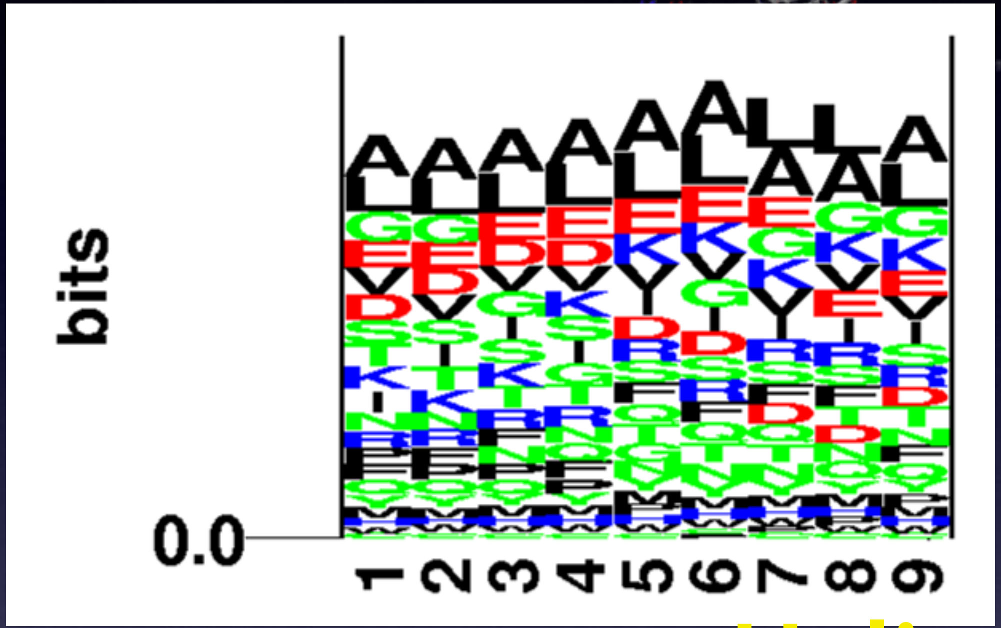


Method	Knowledge	Approach	Difficulty	Useful?
Secondary structure prediction	Sequence-structure statistics	Predict helix, strand, or coil for each residue	Medium	Sometimes (membrane prots.)
Homology modeling	Homologs of known structure	Identify sequence homologs, copy 3D coords and modify	Fairly easy	Quite reliable with high identity. Use in drug design.
Fold recognition	Proteins of known structure	Assemble parts from (several) proteins - often not homologs	Medium to hard	More of a long shot, but models are often correct
Ab initio	Physics and general biology statistics	Simulate folding, or generate lots of structures and pick the best one	Extremely hard	Does not yet work reliably. Too hard?

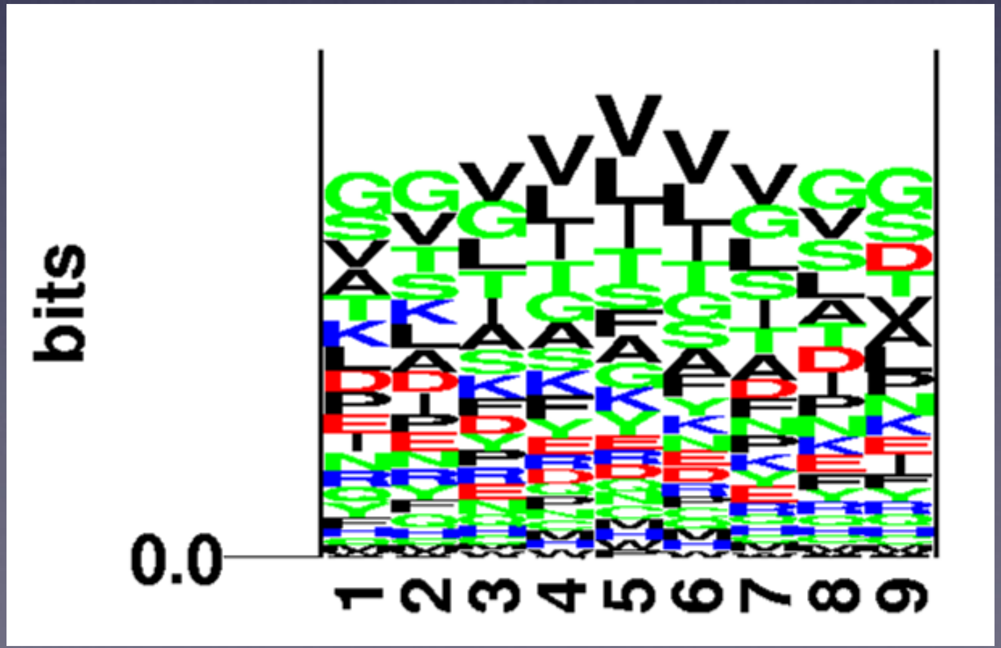
Secondary structure



- Hydrophobicity patterns in helices/strands
- AA Preferences for helix/strand/coil
- Best methods reach ~80% accuracy
- Special case: Predicting transmembrane helices and their in/out topology!

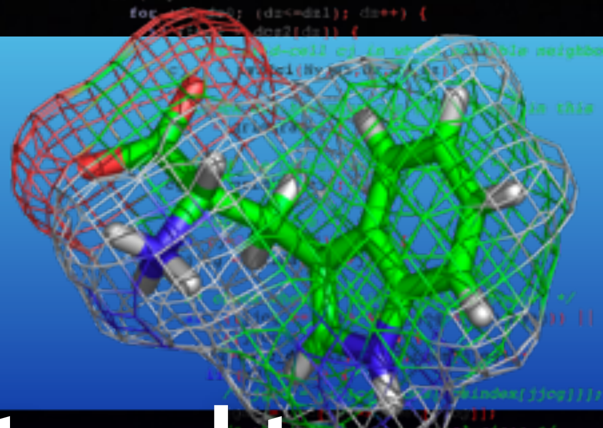


α-Helix

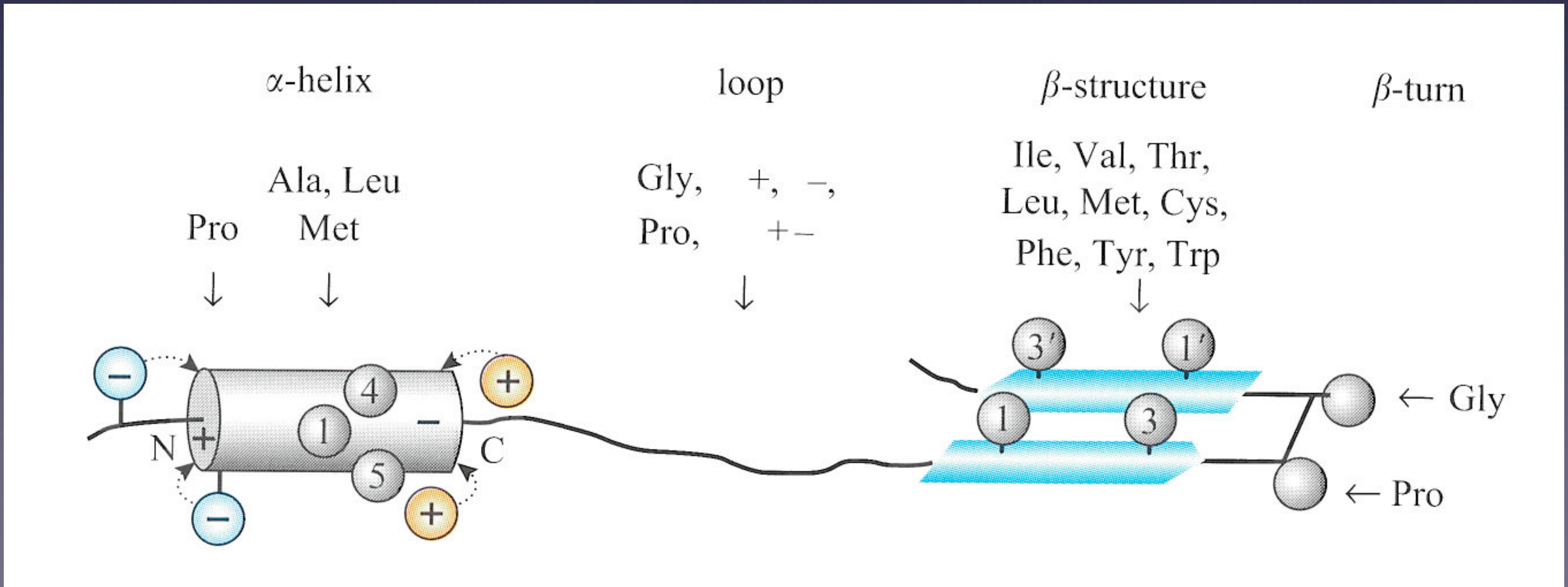


β-Strand

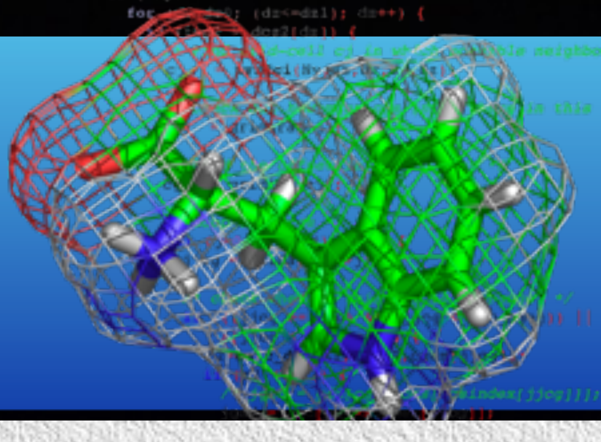
Chou-Fasman

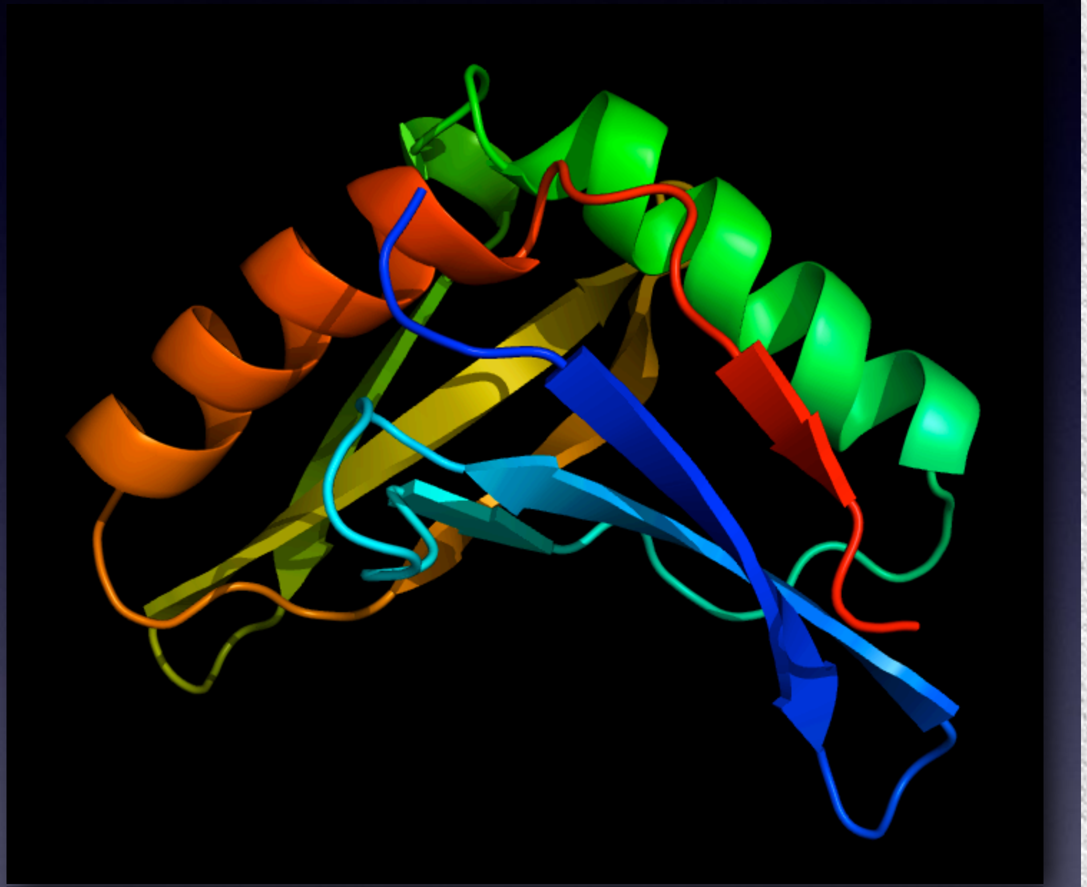


- Determine the probability of helix, sheet and turn for each residue based on available structures
- Single unfavorable residues can occur
- But the rolling average properties of amino acids should be a useful predictor

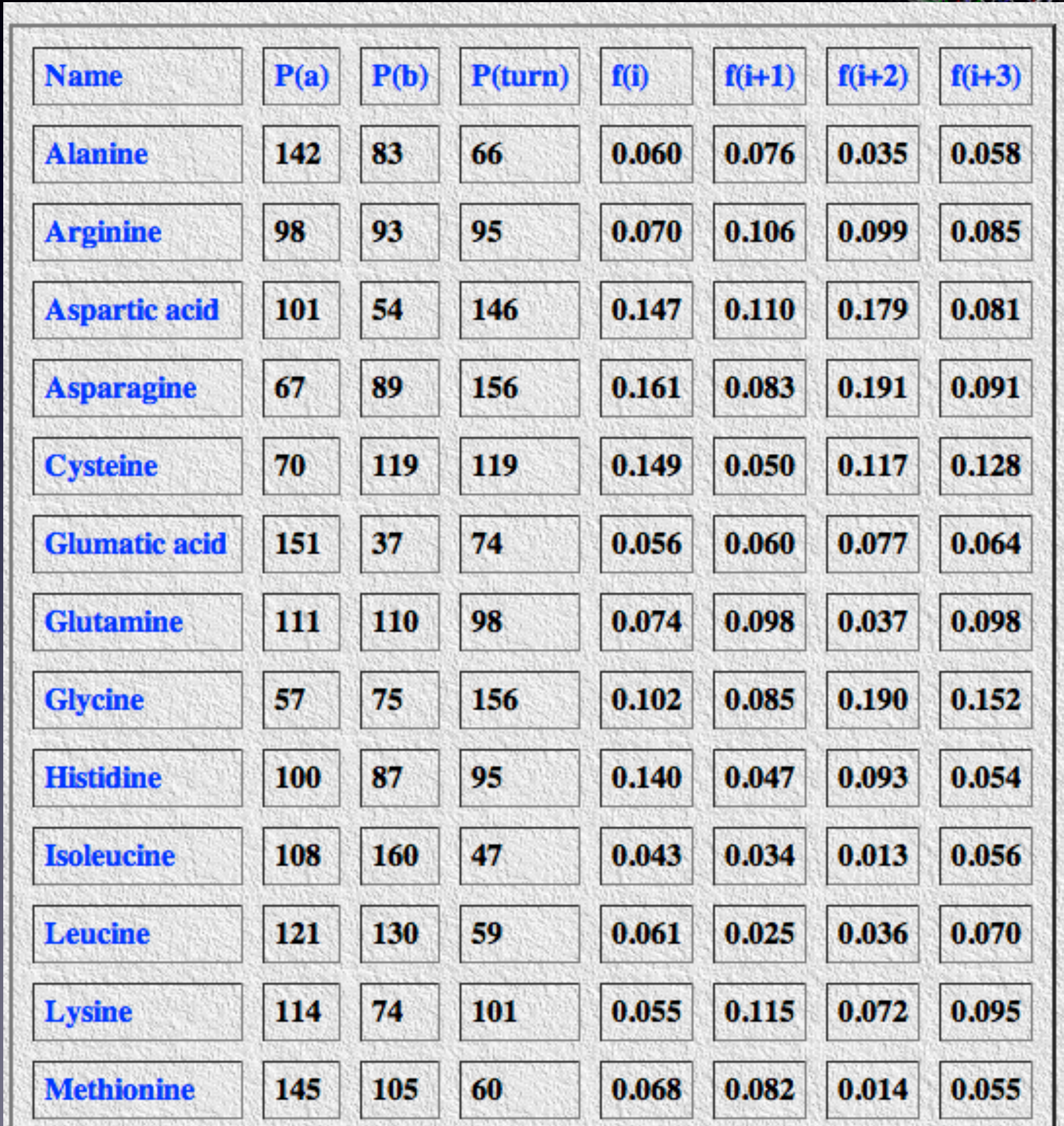


Chou-Fasman data



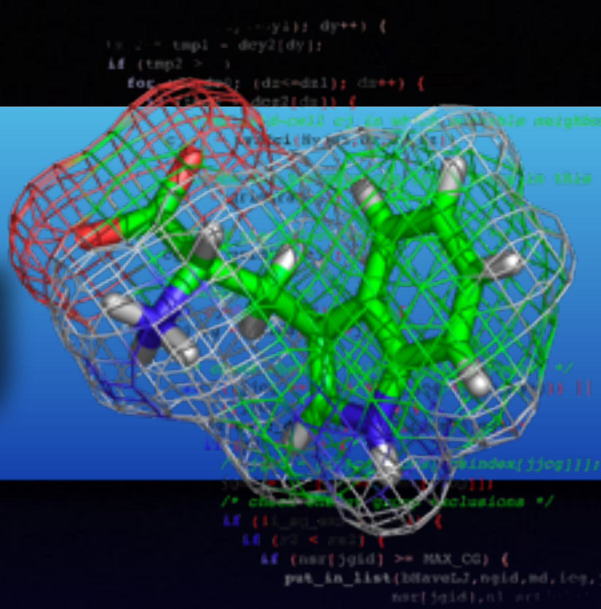


"Propensity" rather than probability, but it contains the same information



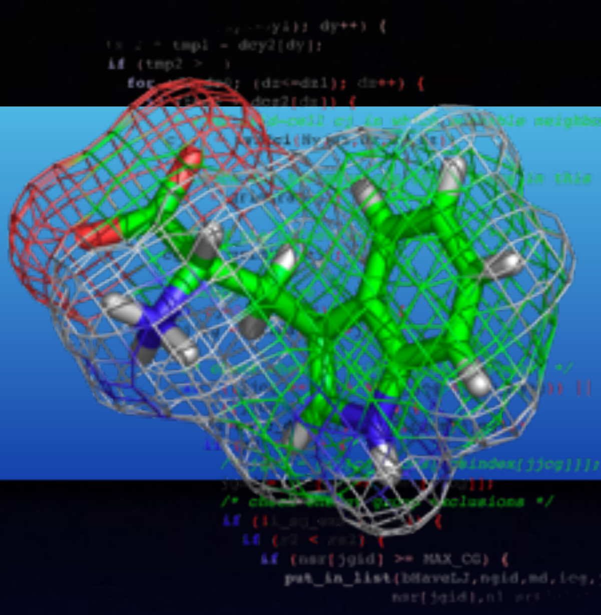
Name	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	142	83	66	0.060	0.076	0.035	0.058
Arginine	98	93	95	0.070	0.106	0.099	0.085
Aspartic acid	101	54	146	0.147	0.110	0.179	0.081
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Cysteine	70	119	119	0.149	0.050	0.117	0.128
Glumatic acid	151	37	74	0.056	0.060	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.190	0.152
Histidine	100	87	95	0.140	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.070
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055

Homology modeling

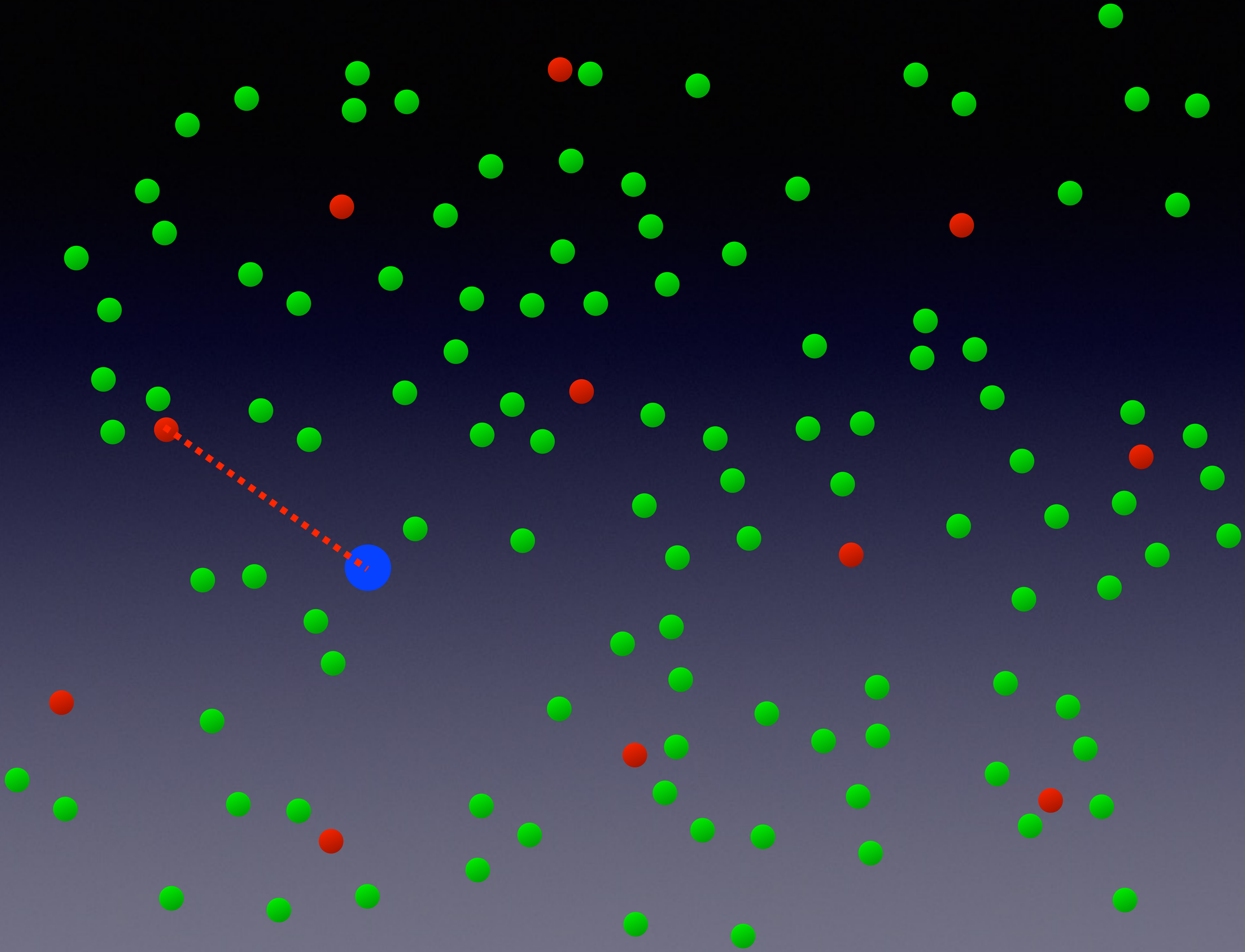


- Protein structures are stable
  - Small sequence changes usually only lead to small variations in 3D structure
- Insertions/deletions usually occur in loop regions, not in helices or sheets
- Sequence matching methods are very good at finding homologs
- Ideally you only need to rebuild side chains

Model Quality?



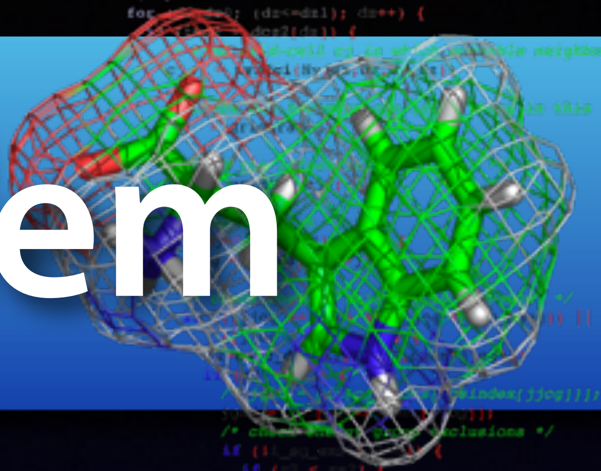
- Depends on modeling distance
- 95% identical residues: perfect model
- 20% identical residues: questionable
- Structural Genomics
  - Reducing modeling distances by determining more 3D crystal structures

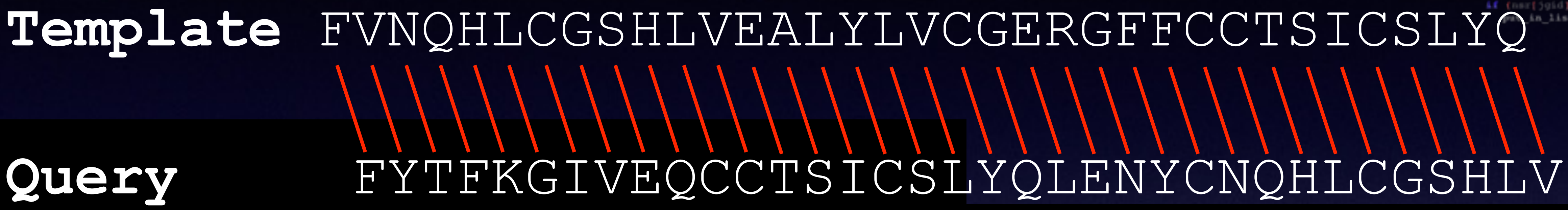


We only need experimental structures for a set of representative folds to create reasonable models for 90% of proteins

Goal of the Structural Genomics Project is 100,000 new structures

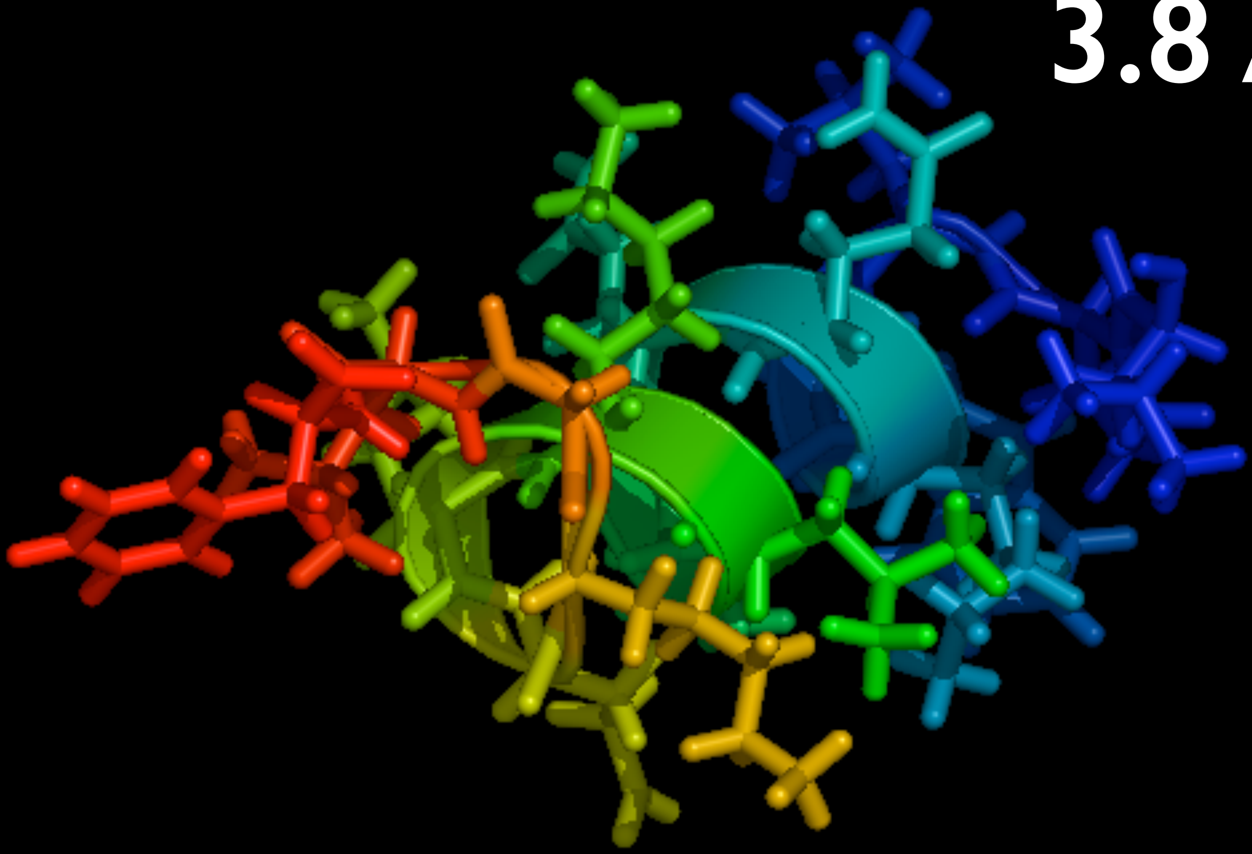
The Alignment Problem

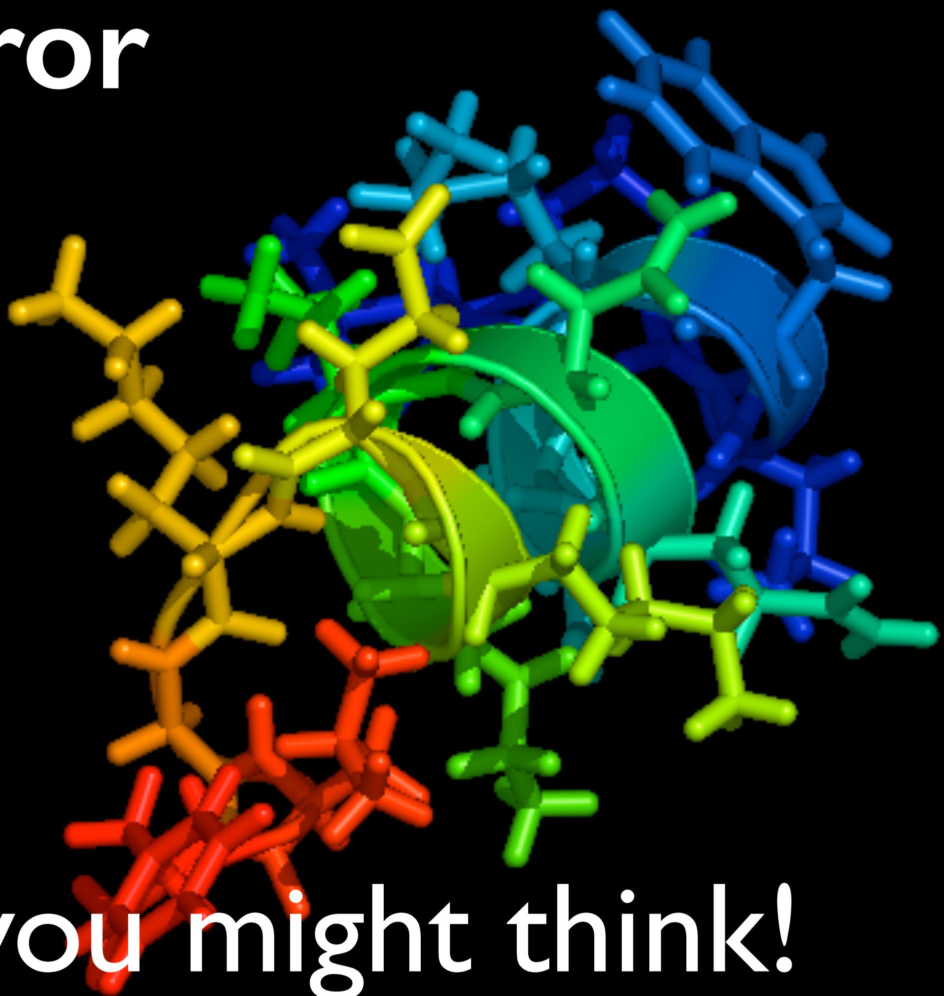




```
Template FVNQHLCGSHLVEALYLVCGERGFFCCTSICSLYQ

Query      FYTFKGIVEQCCTSICSLYQLENYCNQHLCGSHLV
```

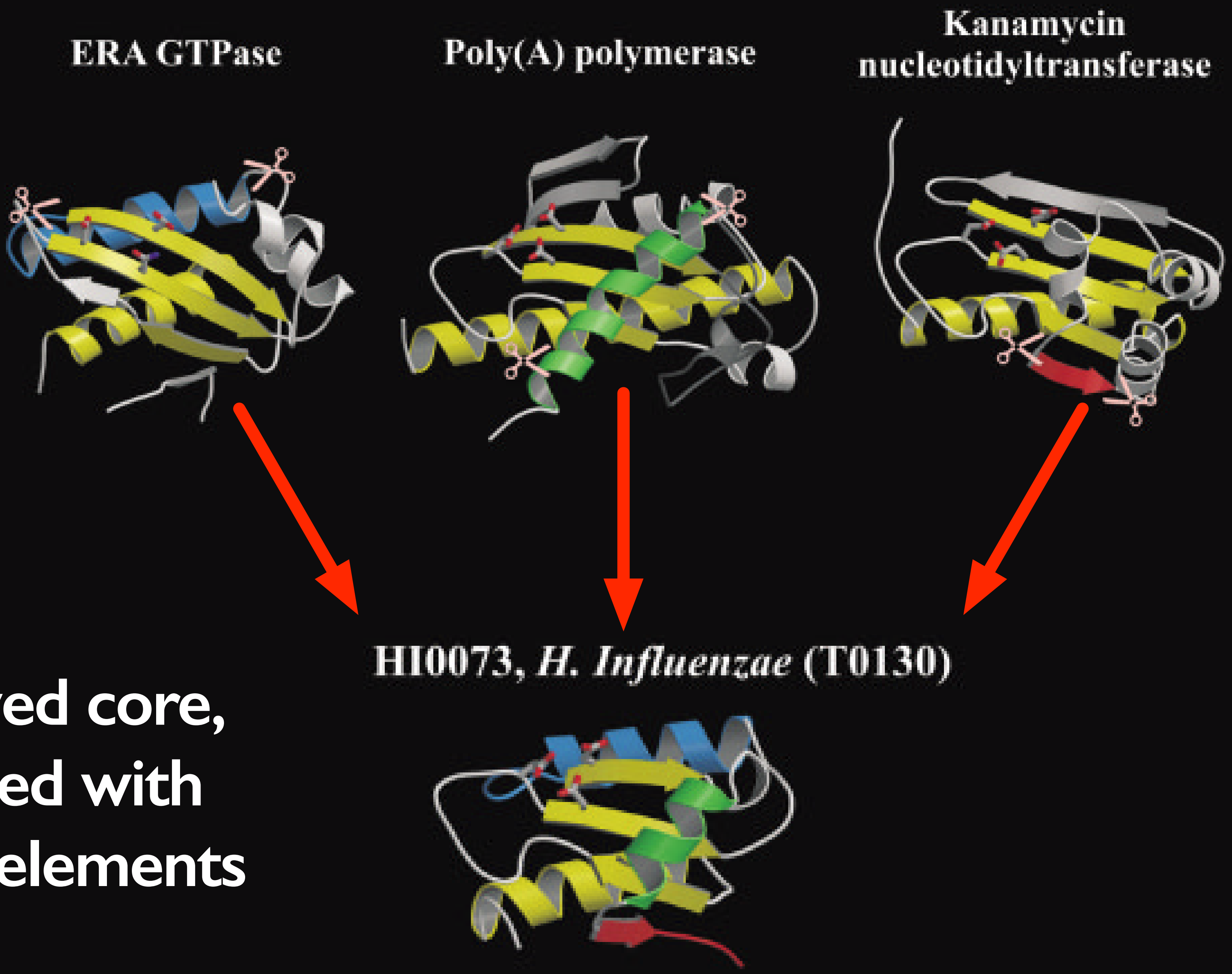
3.8 Å error





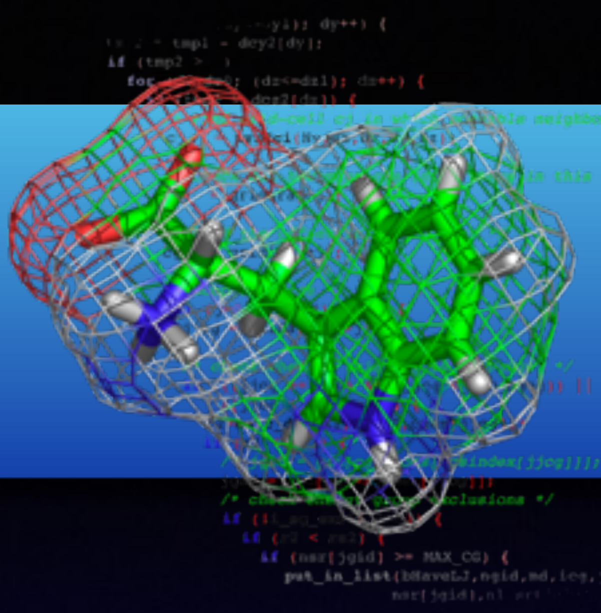
Prediction is harder than you might think!

Multiple Templates


ERA GTPase
Poly(A) polymerase
Kanamycin nucleotidyltransferase
HI0073, *H. Influenzae* (T0130)

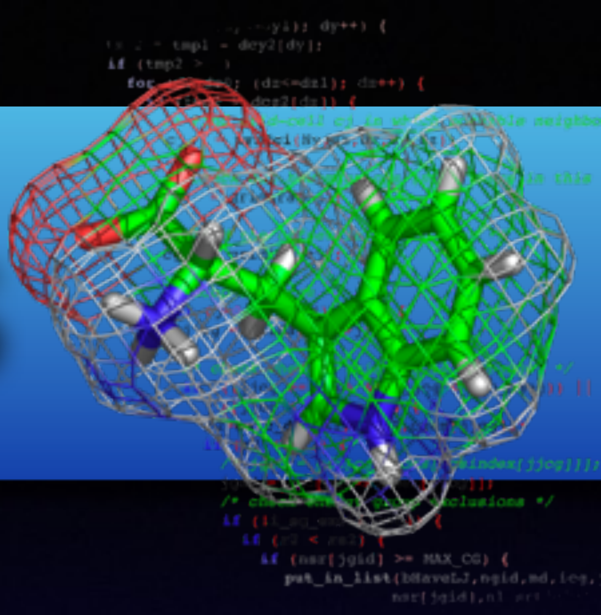
Conserved core, combined with different elements

Ab Initio prediction



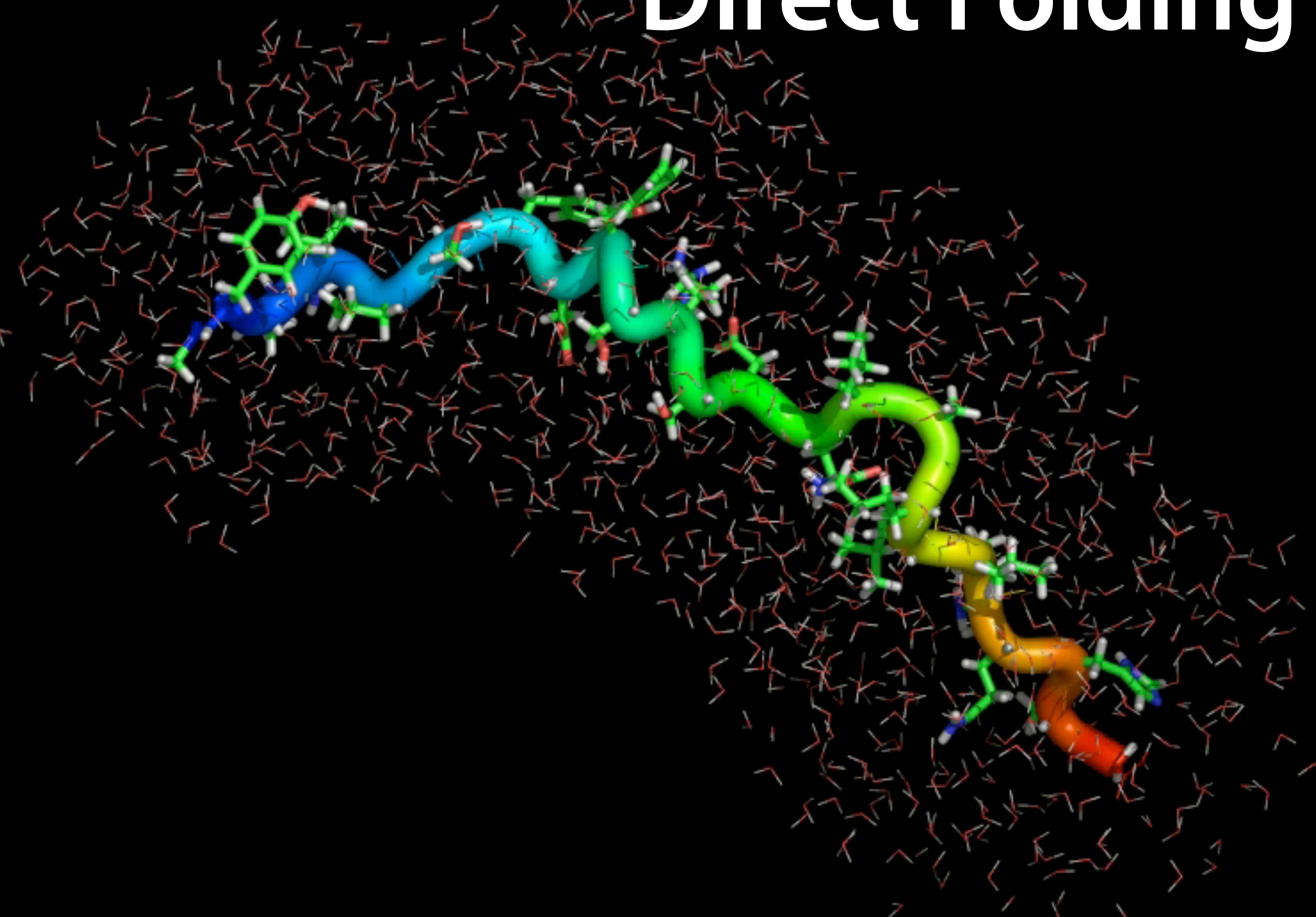
- Consider a 100 residue protein
- Assume there are 10 conformations/aa
- 10^100 stuctures to test
- Levintal's paradox: It would take the age of the universe to test everything
- In practice it must be a guided process
- But how do you do it in a computer?

Possible approaches

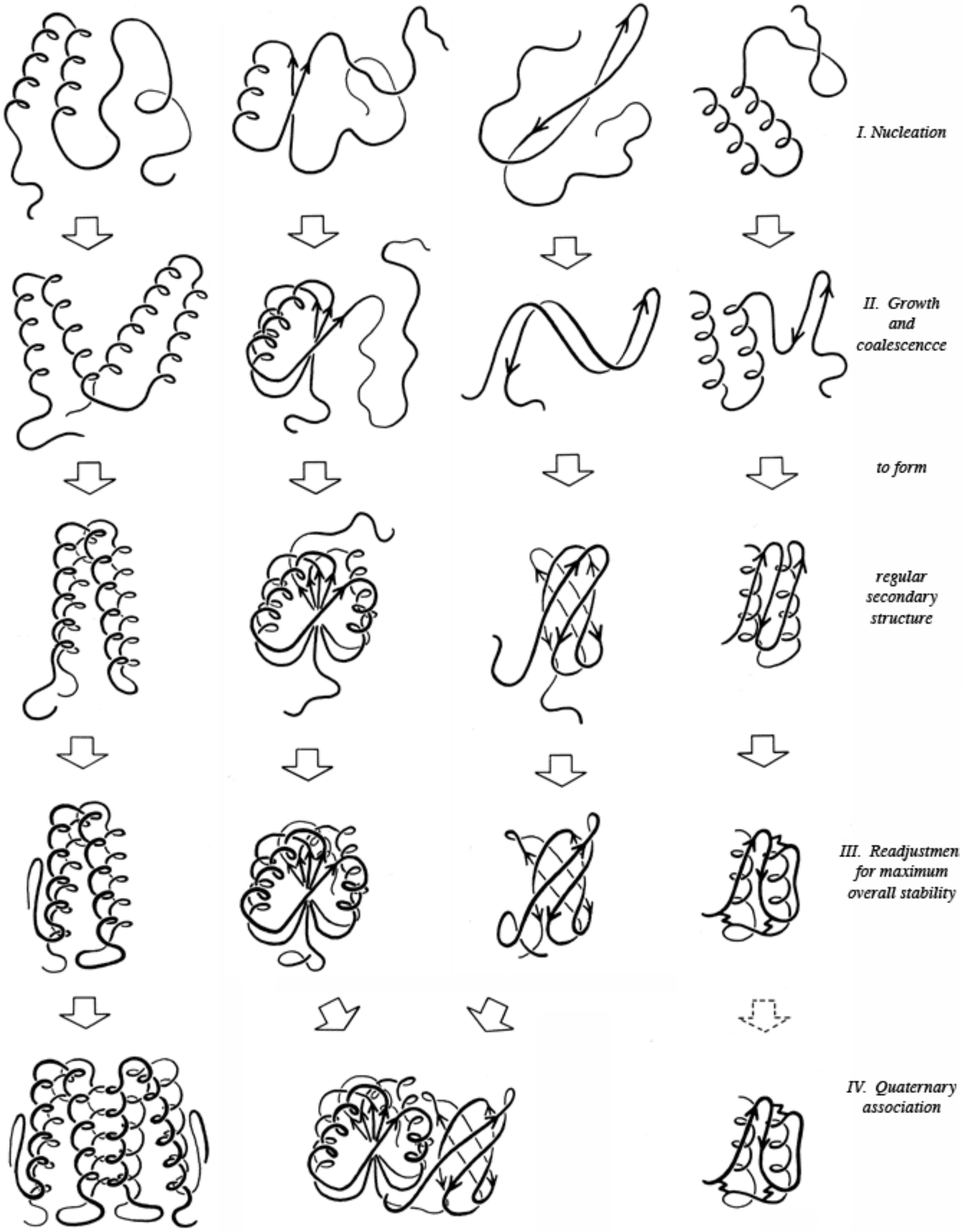


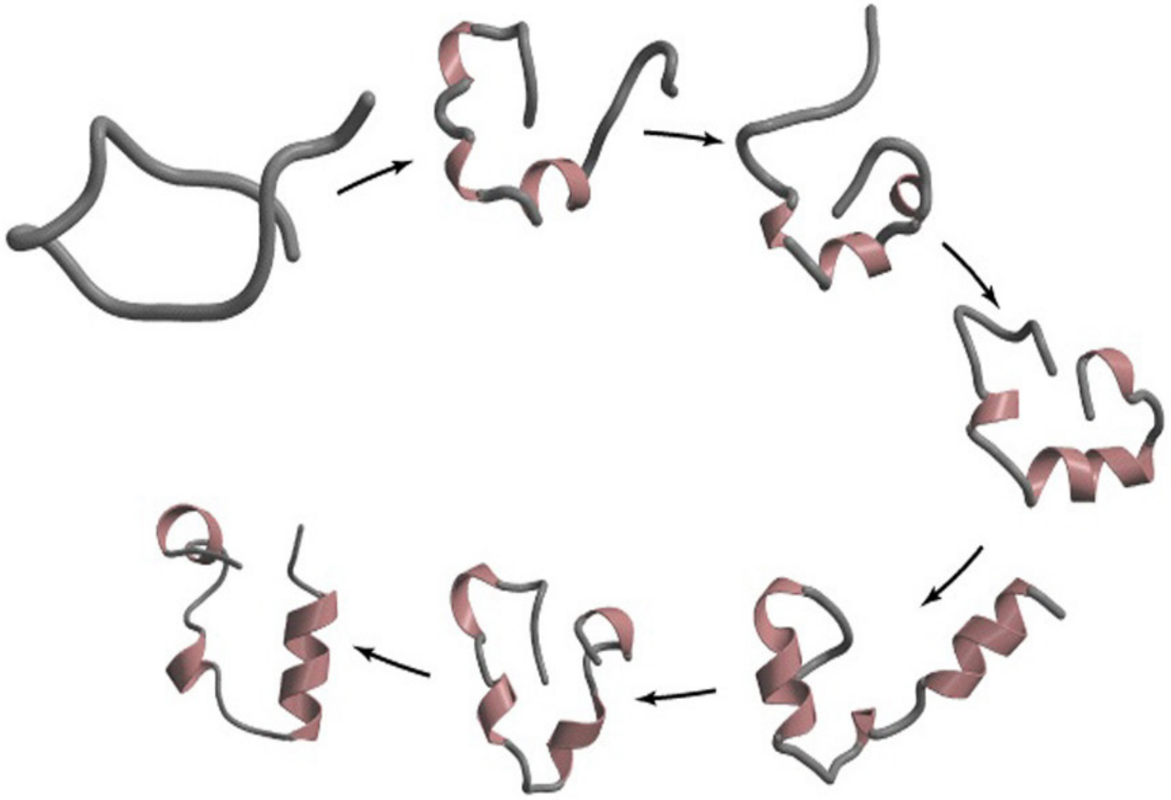
- Brute force physical simulation
  - Would provide both the path & goal
  - Even supercomputers are usually too slow
- Smarter *ab initio* algorithms
  - The path is usually NOT the goal
  - Create test structures & find the best
  - Fragment assembly: ROSETTA (Baker)

Direct Folding

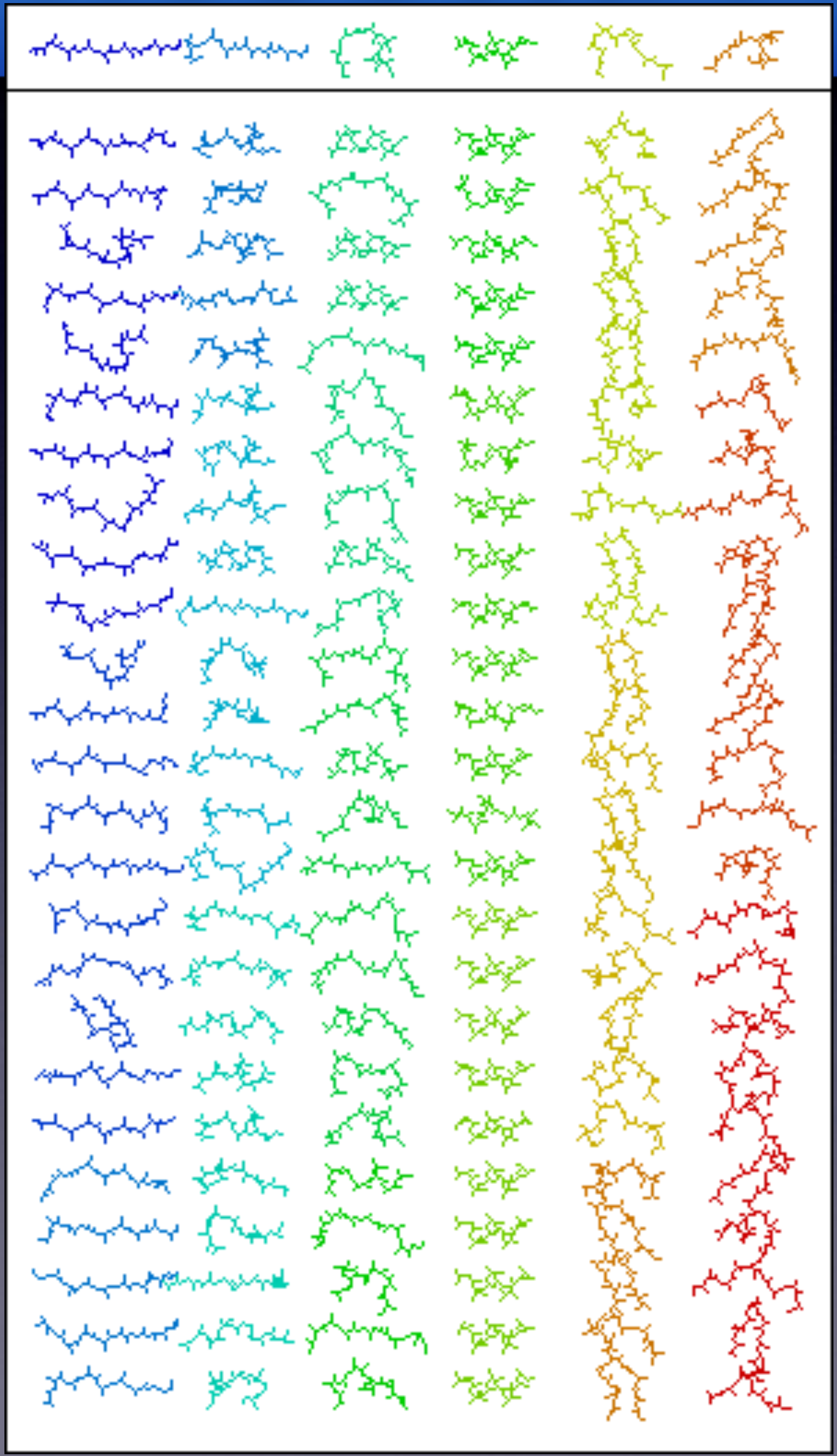


The Rosetta idea


I. Nucleation
II. Growth and coalescencce
to form
regular secondary structure
III. Readjustmen for maximum overall stability
IV. Quaternary association



Rosetta Fragment libraries



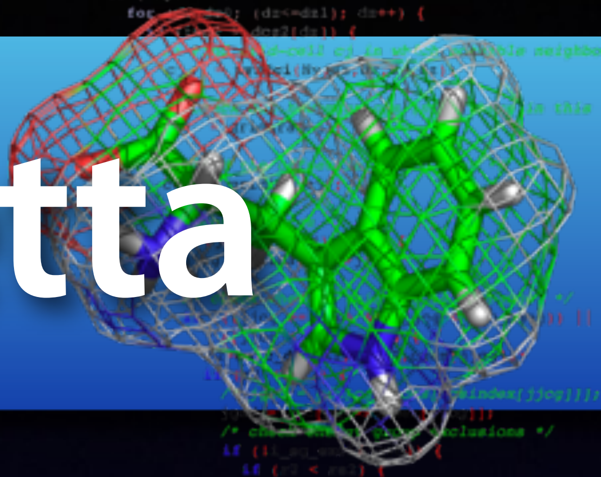
25-200 fragments for each 3 and 9 residue sequence window

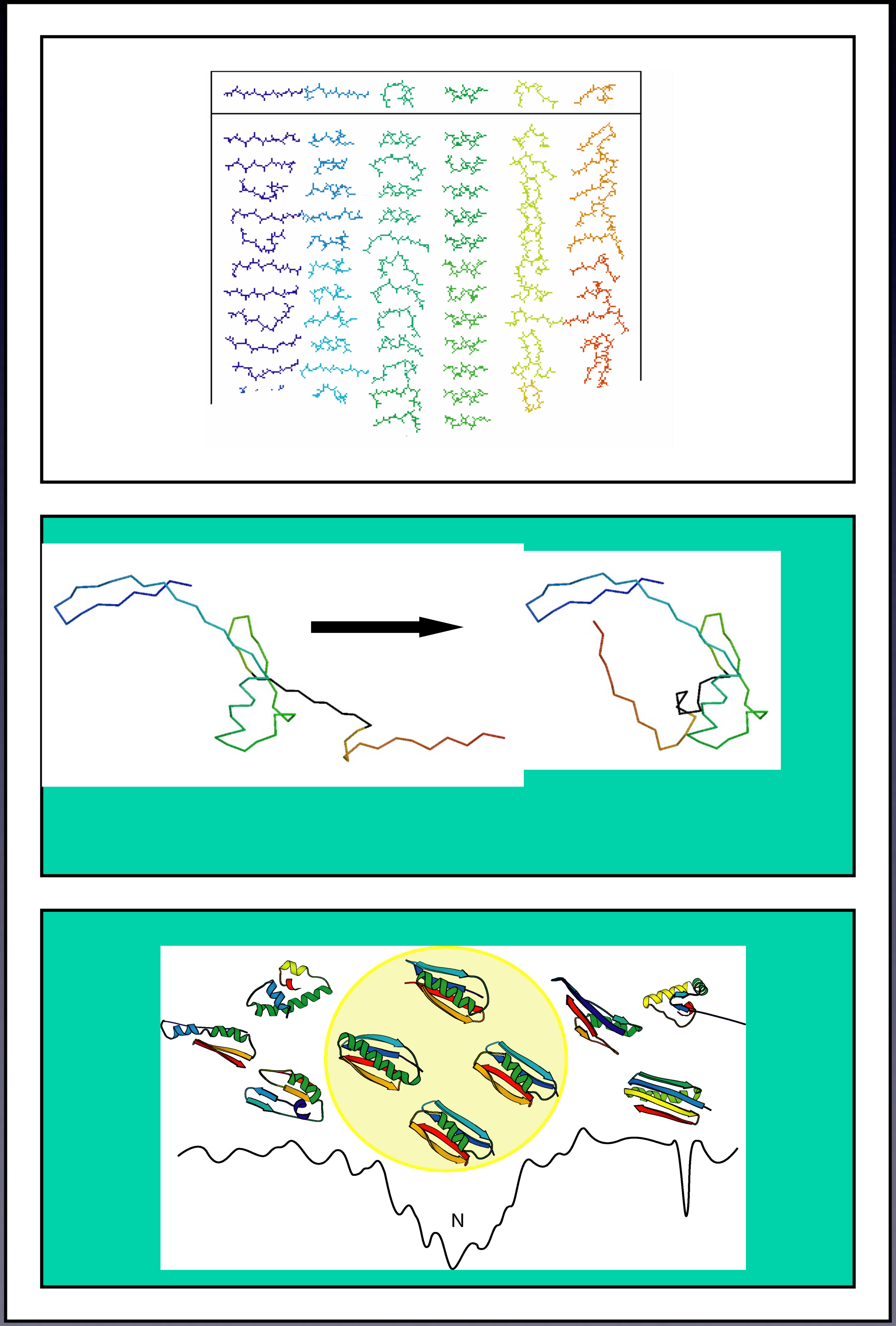
Selected from known structures

Better than 2.5Å resolution

< 50% sequence identity

Prediction with Rosetta

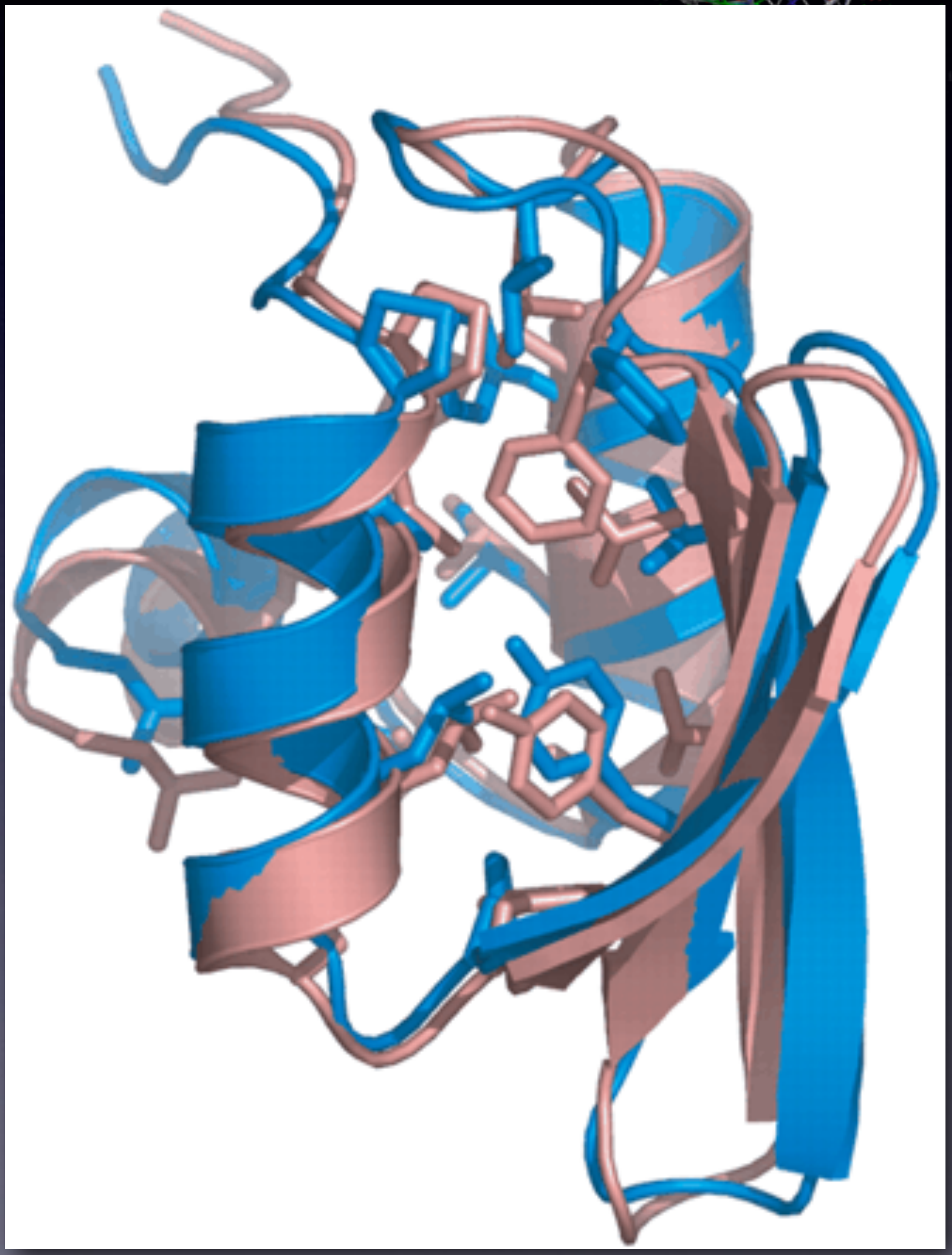




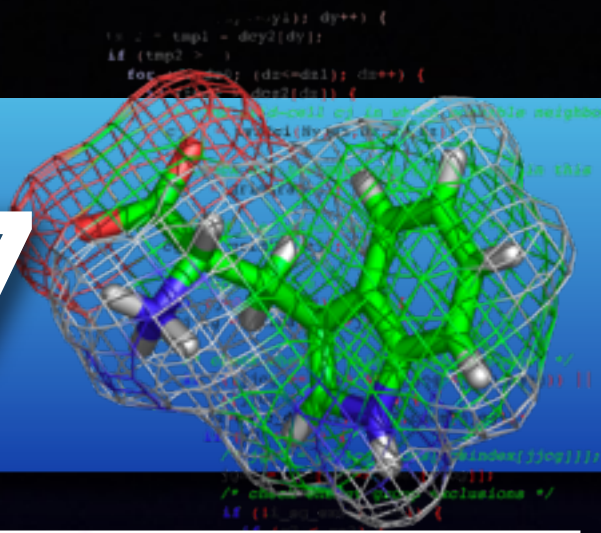
- Select fragments with good local properties
- Assemble into protein-like folds (lots of them)
- Use physics-based energy functions to try and select the best one

Rosetta Successes

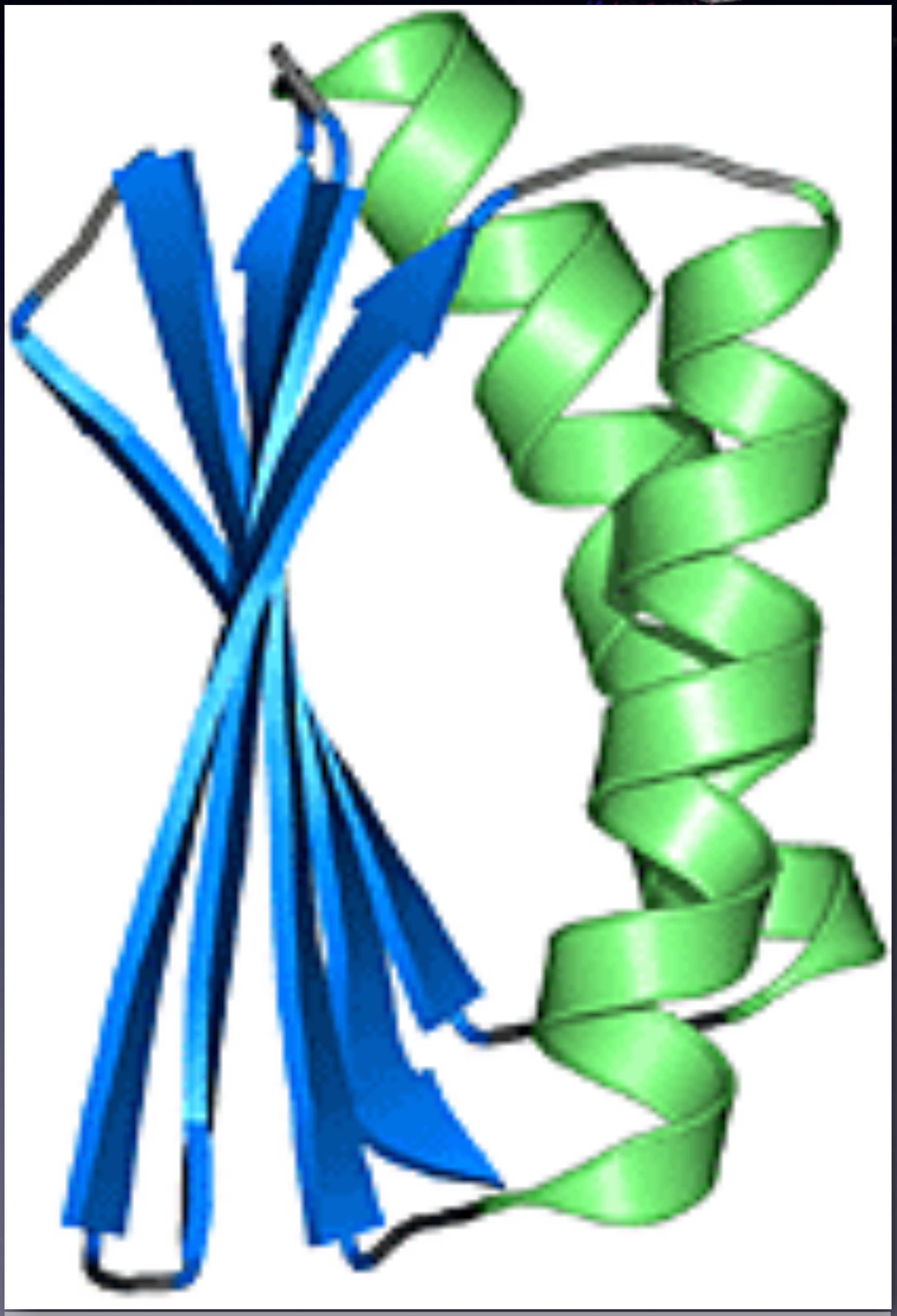
- Refinement: Make small moves in torsion angles
- Rebuild sidechains
- Minimize energy
- Repeat refinement, etc.
- Bradley, Science 2005:
  5 of 16 structures predicted to within 1.5Å resolution!



Rosetta Design: TOP7



- Can you design a completely new fold not seen in nature?
- Iterate design & refinement
- Extremely stable structure
- Determined structure in experiments to confirm: Less than 1.2Å difference!



INFORMATION + PHYSICS = LIFE

DNA Sequence → RNA Sequence → Protein Sequence → Folded Protein

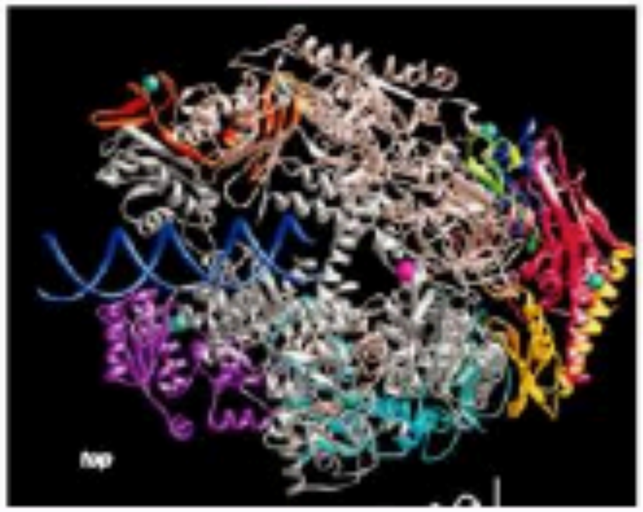
- *in silico*

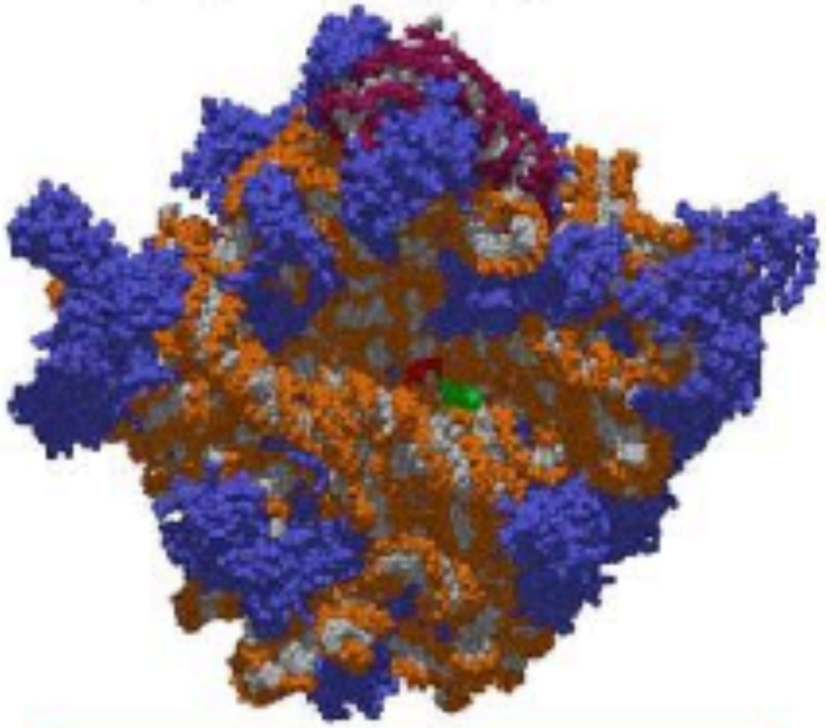
Easy: Change T to U

Easy: Triplet Code

Hard: Folding is many body simulation

- *in vivo*





Easy: Folding is free by laws of physics

Hard: Transcription Polymerase

Hard: Translation Ribosome

Michael Levitt

©Michael Levitt 01