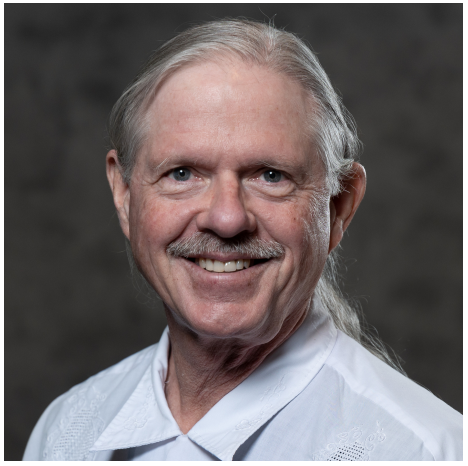
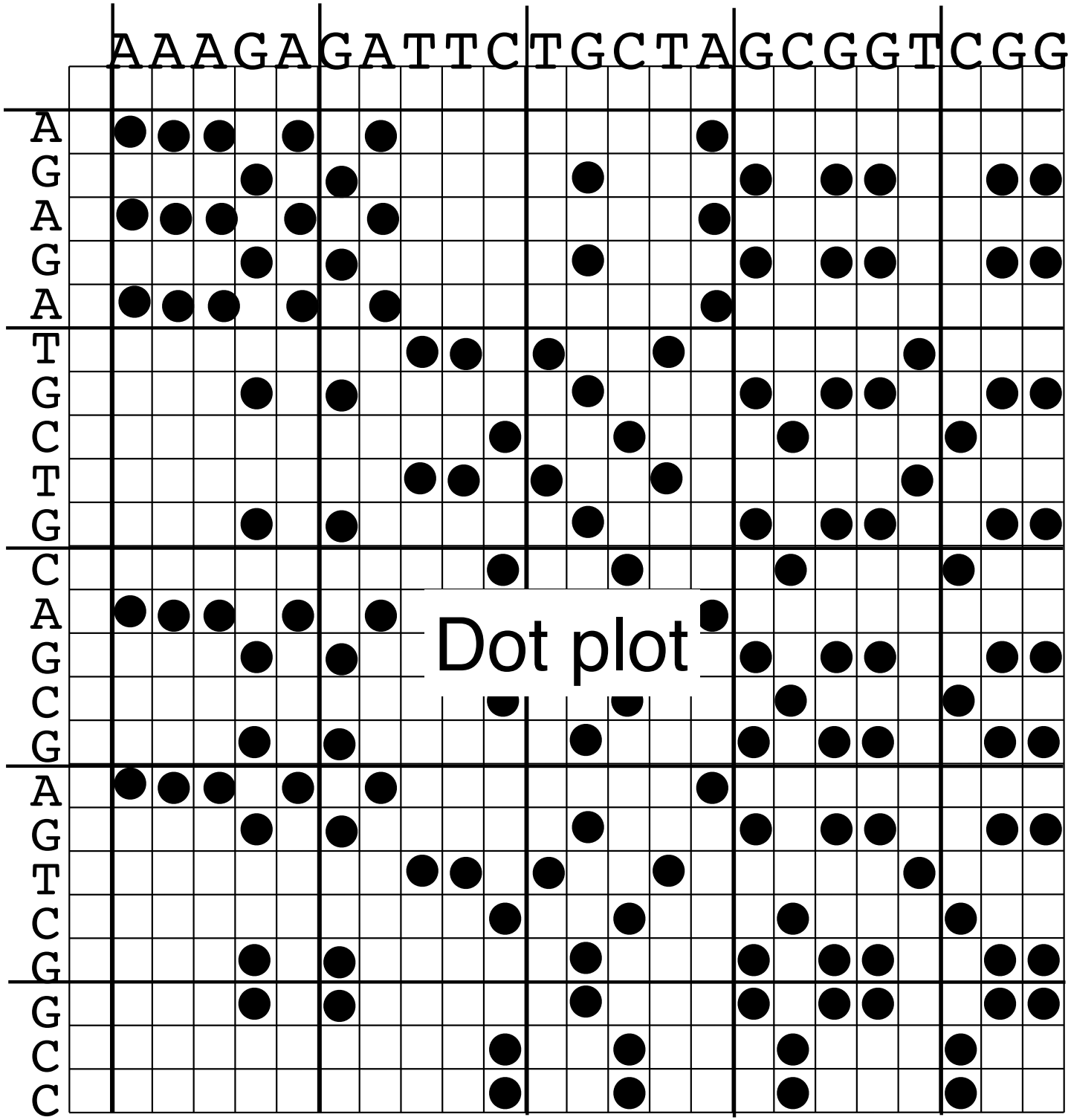
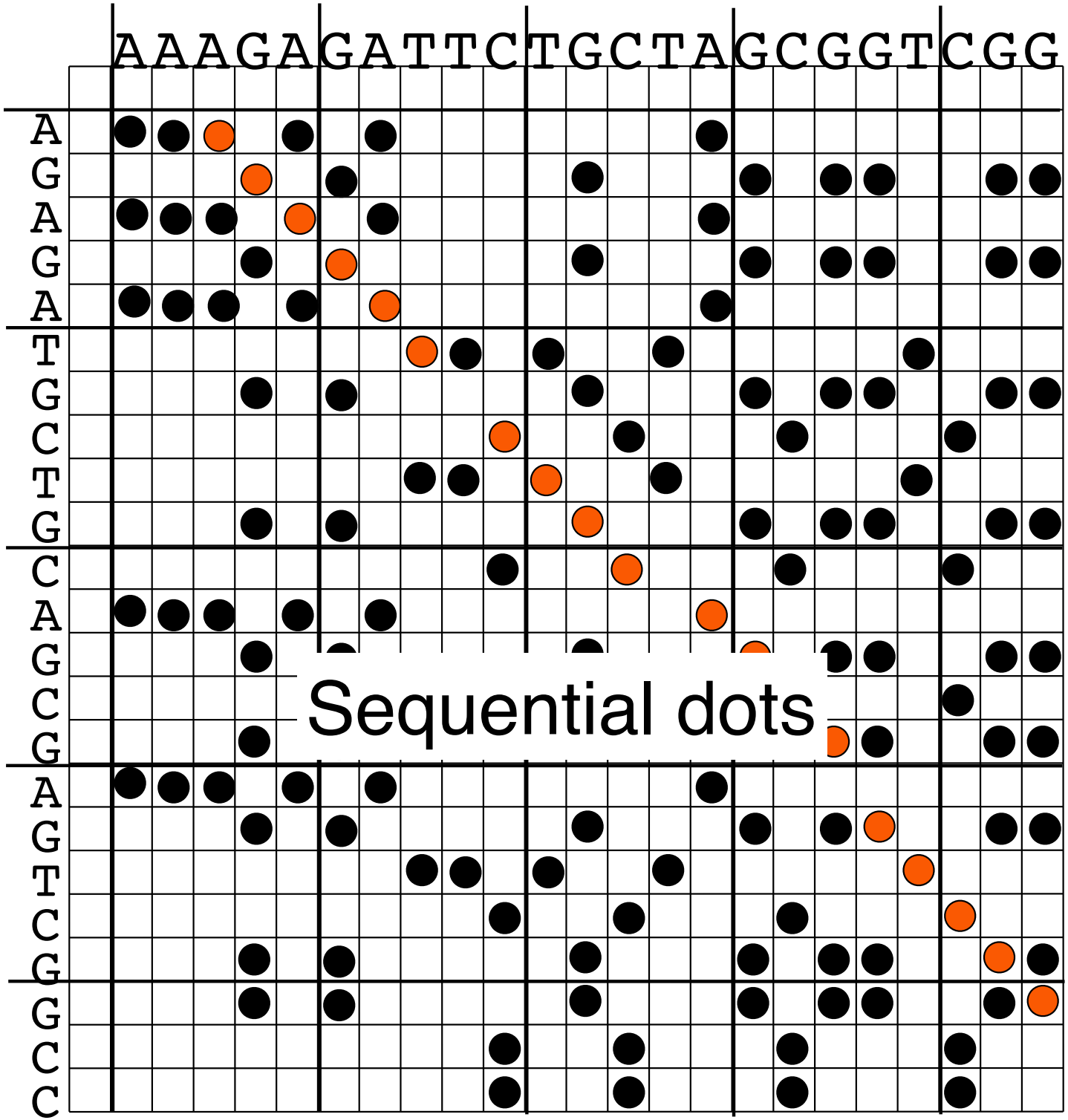
Prof. Bystroff talks about BIOINFORMATICS

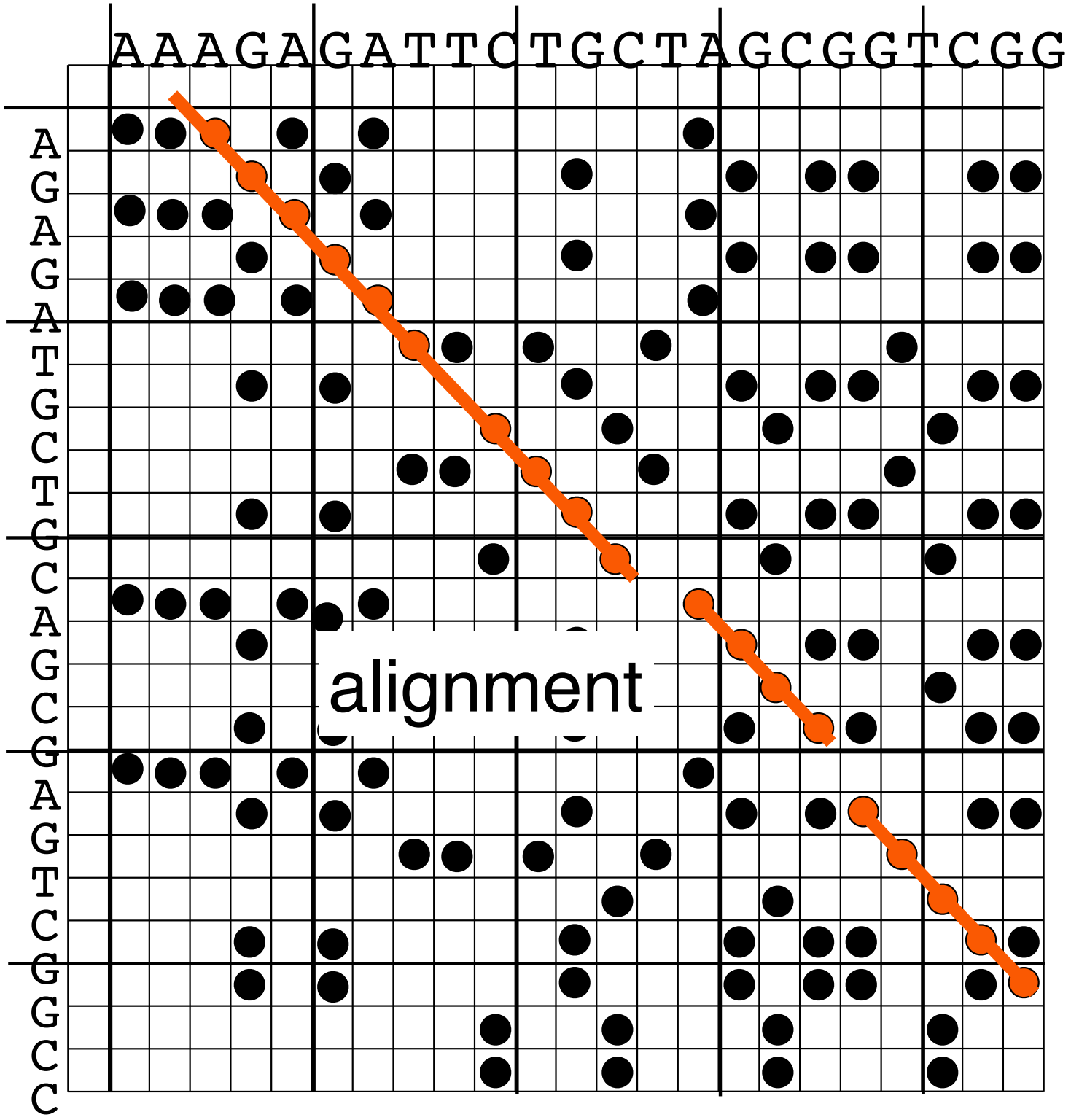


hi

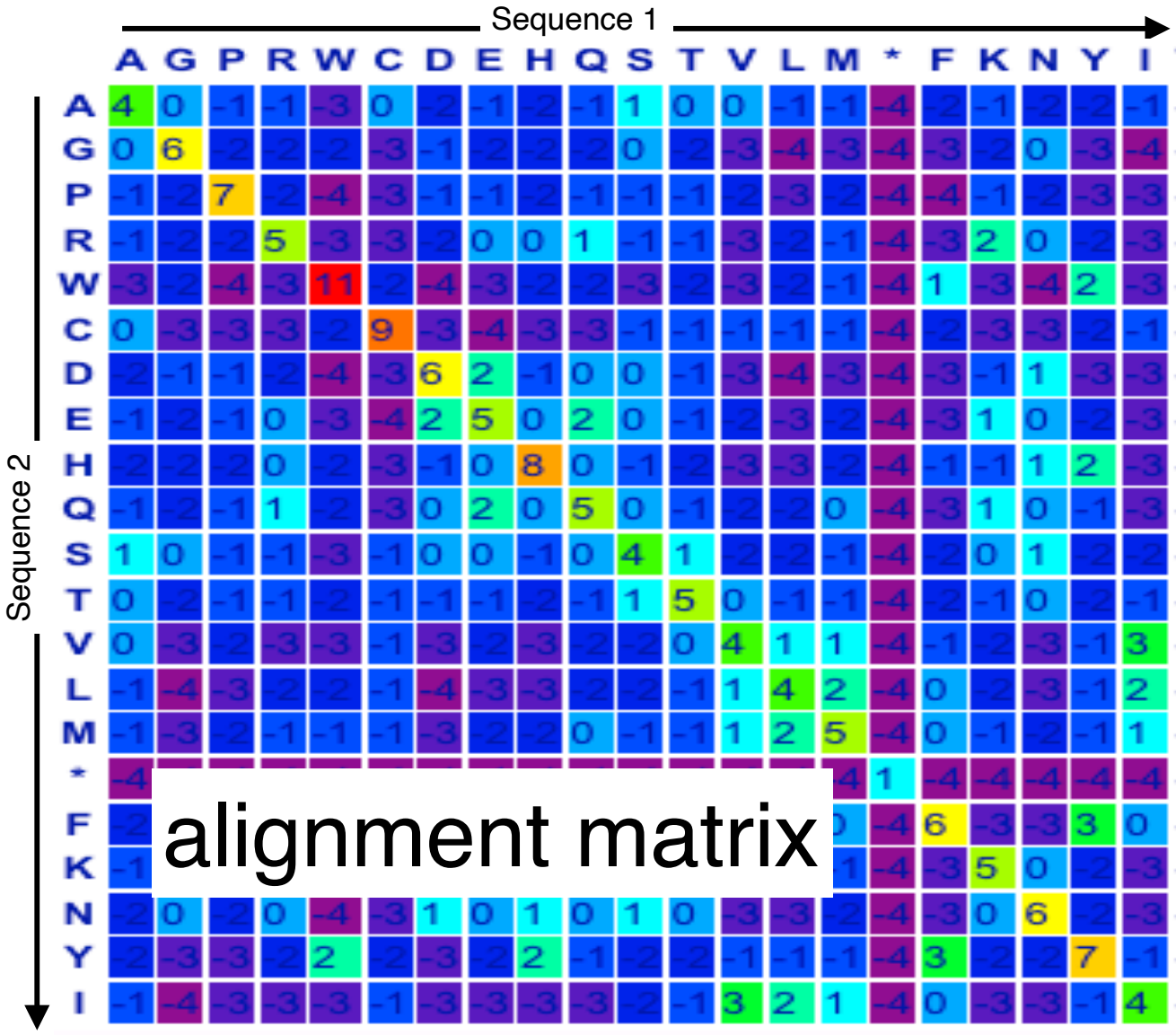
- **Sequence database searching**
- Phylogenetic Trees
- Protein Structure


AAAGAGATTCTGCTAGCGGTCGG
AGAGATGCTAGCGGTCGG
TGCTGCTAGCGGTCGG
CAGCGCGCTAGCGGTCGG
AGTCGGCGTAGCGGTCGG
GCCC
Dot plot

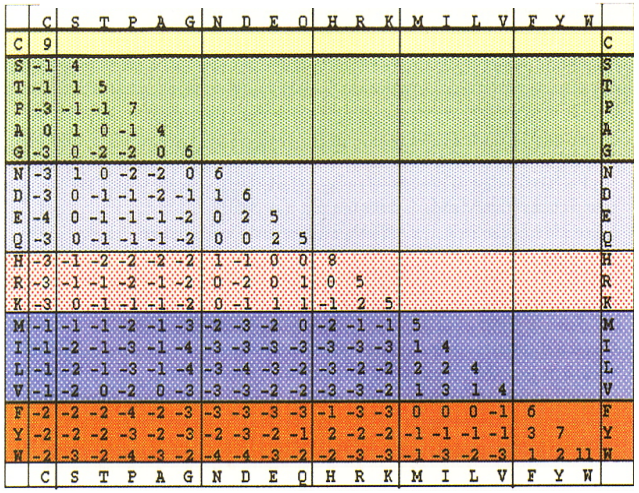

AAAGAGATTCTGCTAGCGGTCGG
A
G
A
G
A
T
G
C
T
G
C
A
G
C
G
A
G
T
C
G
G
C
C
Sequential dots


AAAGAGATTCTGCTAGCGGTCGG
AGAGATGCTCCAGCGAGTCGGCC
alignment

Protein sequence alignment uses a "substitution" score instead of dots.

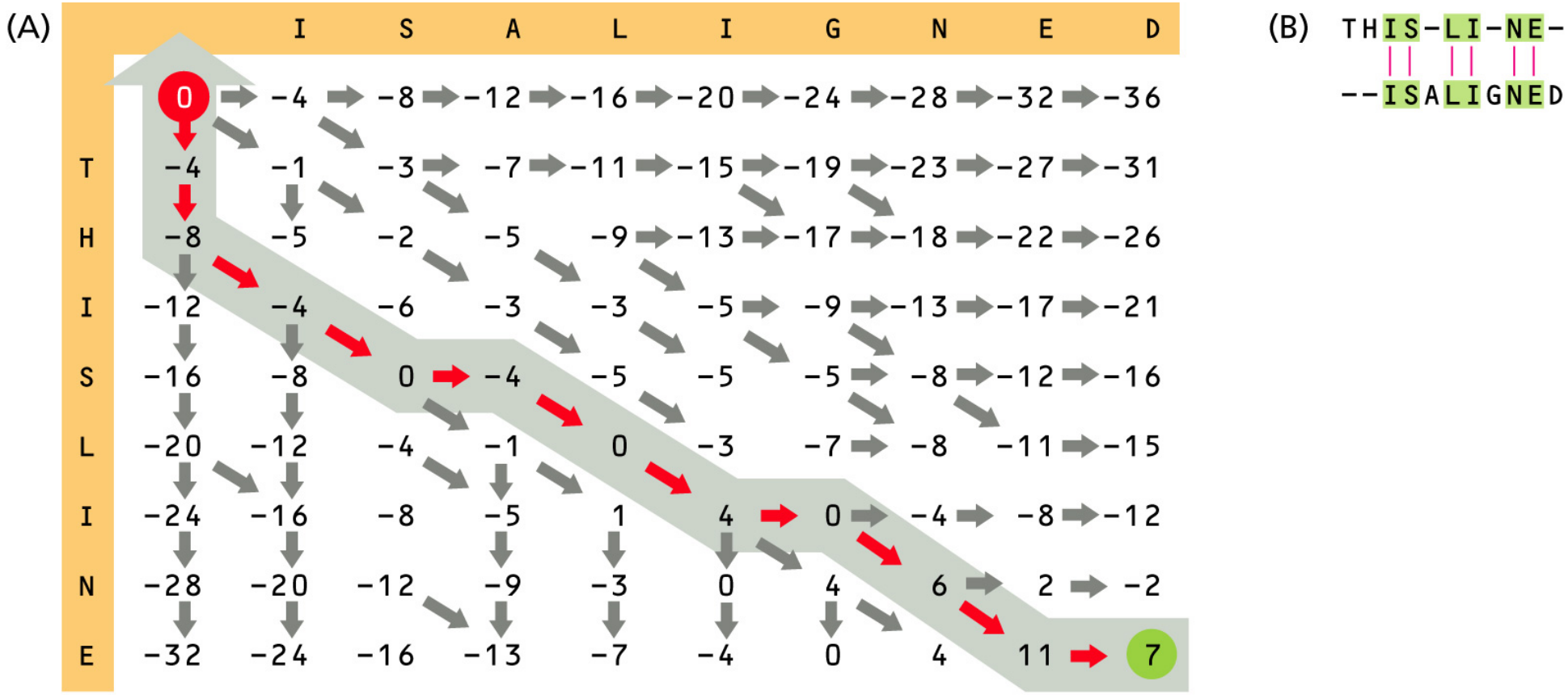


substitution table



lookup table for scores

The best sequential pathway through the substitution scores is the alignment



"dynamic programming" algorithm.

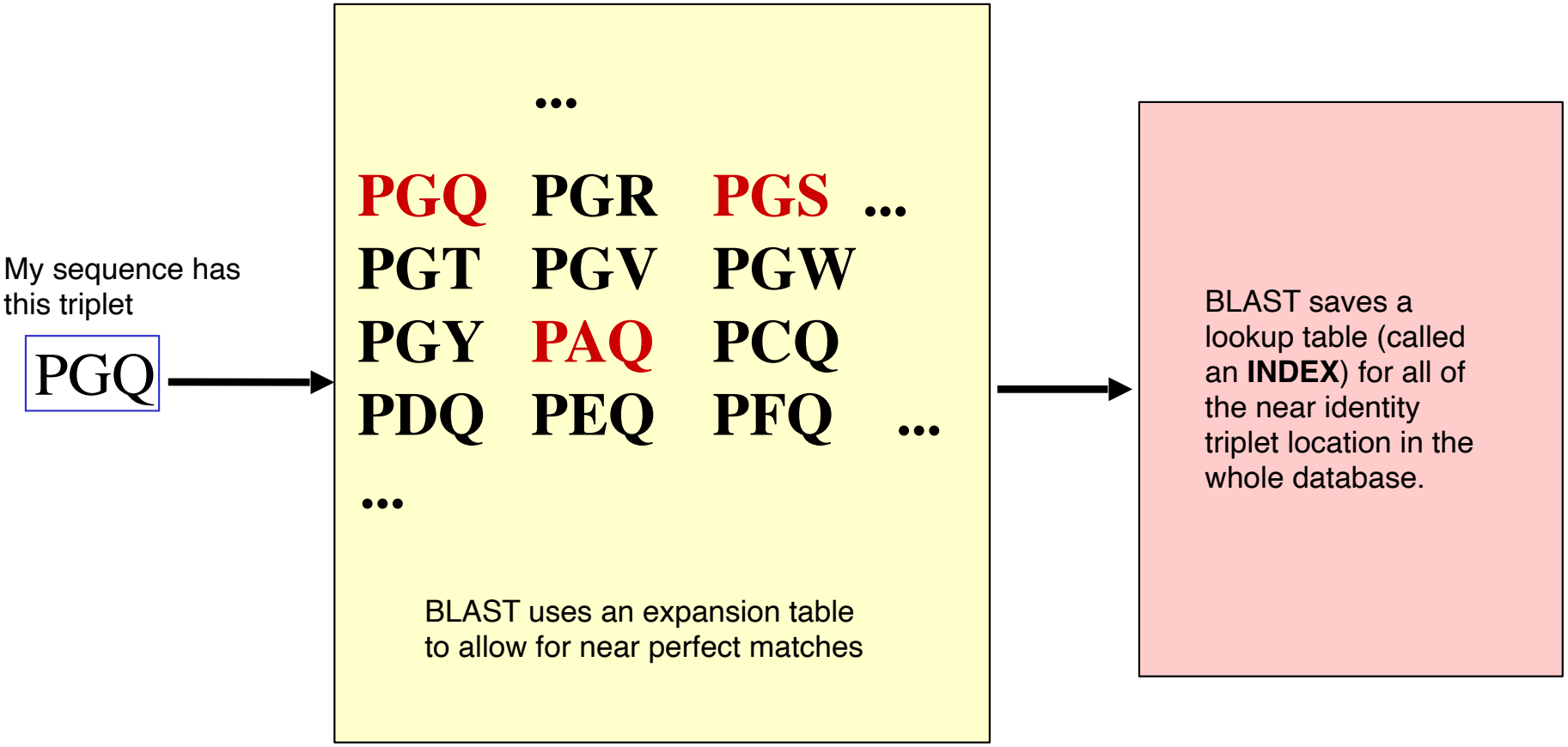
BLAST searches millions of sequences

GenBank contains *over 162 million sequences!!*

The score for each should be the *optimal* alignment score. Even if we can do 1 per millisecond, it would take 45 hours to do one search. BLAST usually finishes in under a minute.

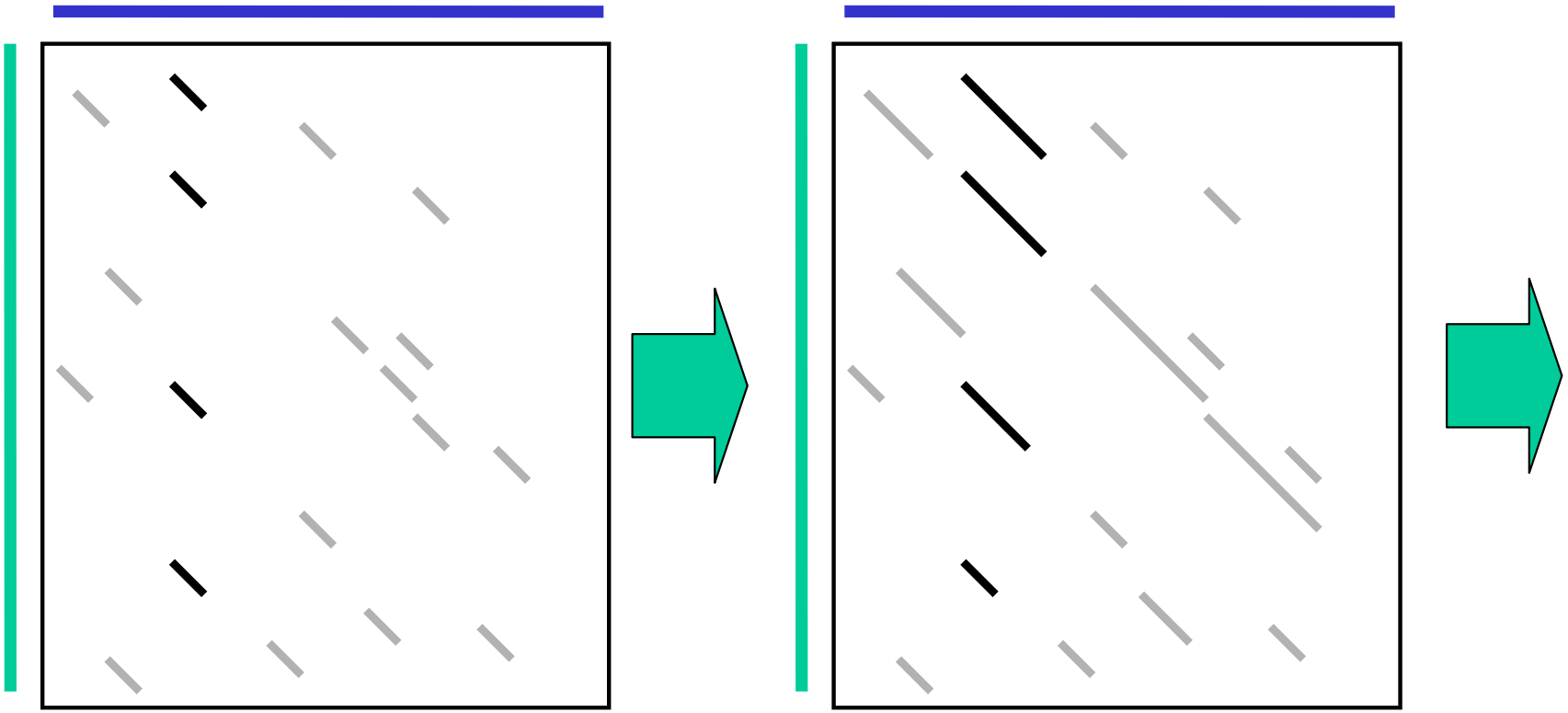
How does BLAST do it so fast?

BLAST precalculates all triplet hits in the database.



This is all done when BLAST is set up, before any searches are carried out.

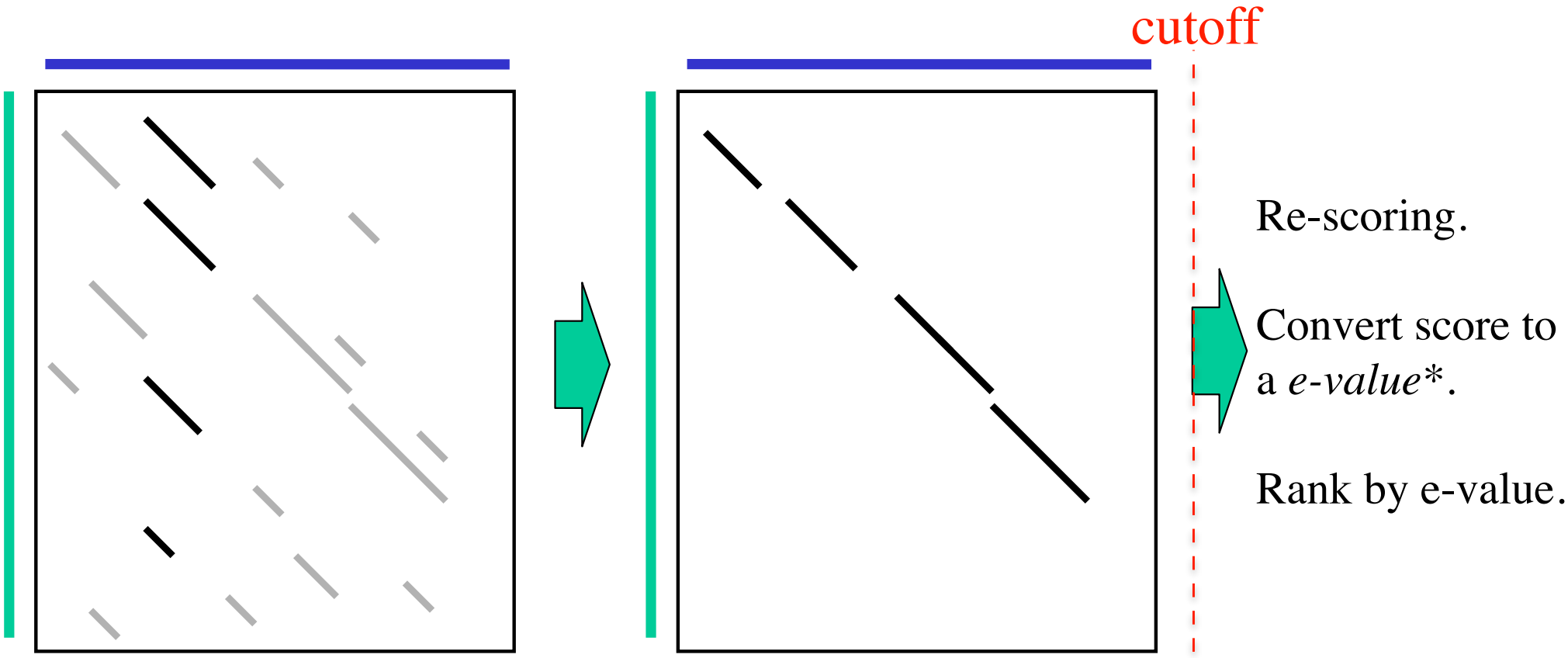
BLAST finds diagonal arrangements of triplet hits



triplet hits in one database protein

Hits are joined by extension

BLAST scores only the best hits (saves time)



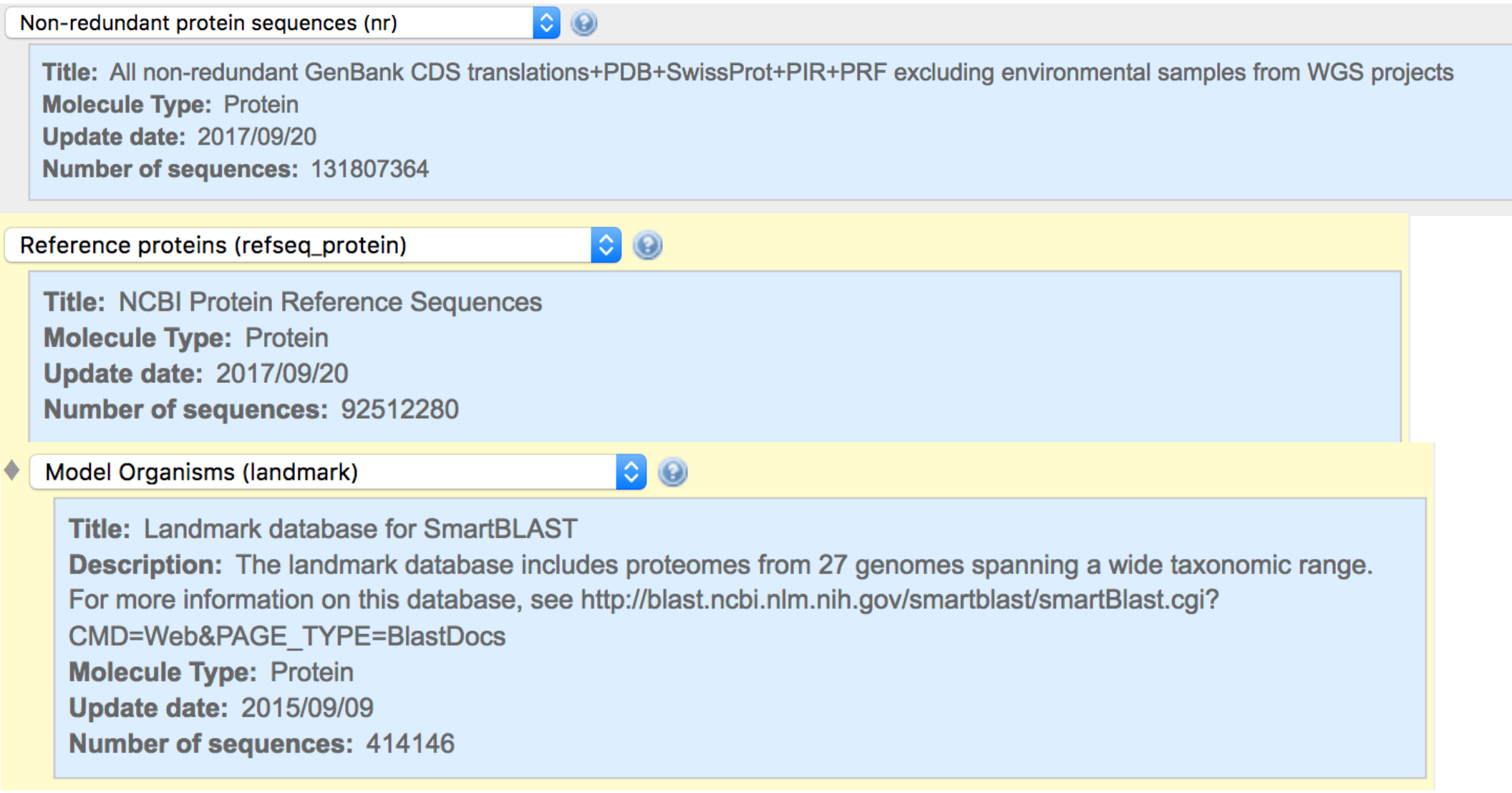
BLAST connects the diagonals (FASTA algorithm)

This protein is given a score, and we save it for later only if the score passes a cutoff.

*later...

Protein Databases available for BLAST search

Go to BLAST search page (i.e. blastp) , select a database to search and then select ? to learn a little about that database.



Protein Databases available for BLAST search

On BLAST search page, select.a database to search and then select ? to learn a little about that database.

UniProtKB/Swiss-Prot(swissprot)

Title: Non-redundant UniProtKB/SwissProt sequences.
Molecule Type: Protein
Update date: 2017/09/20

Patented protein sequences(pat)

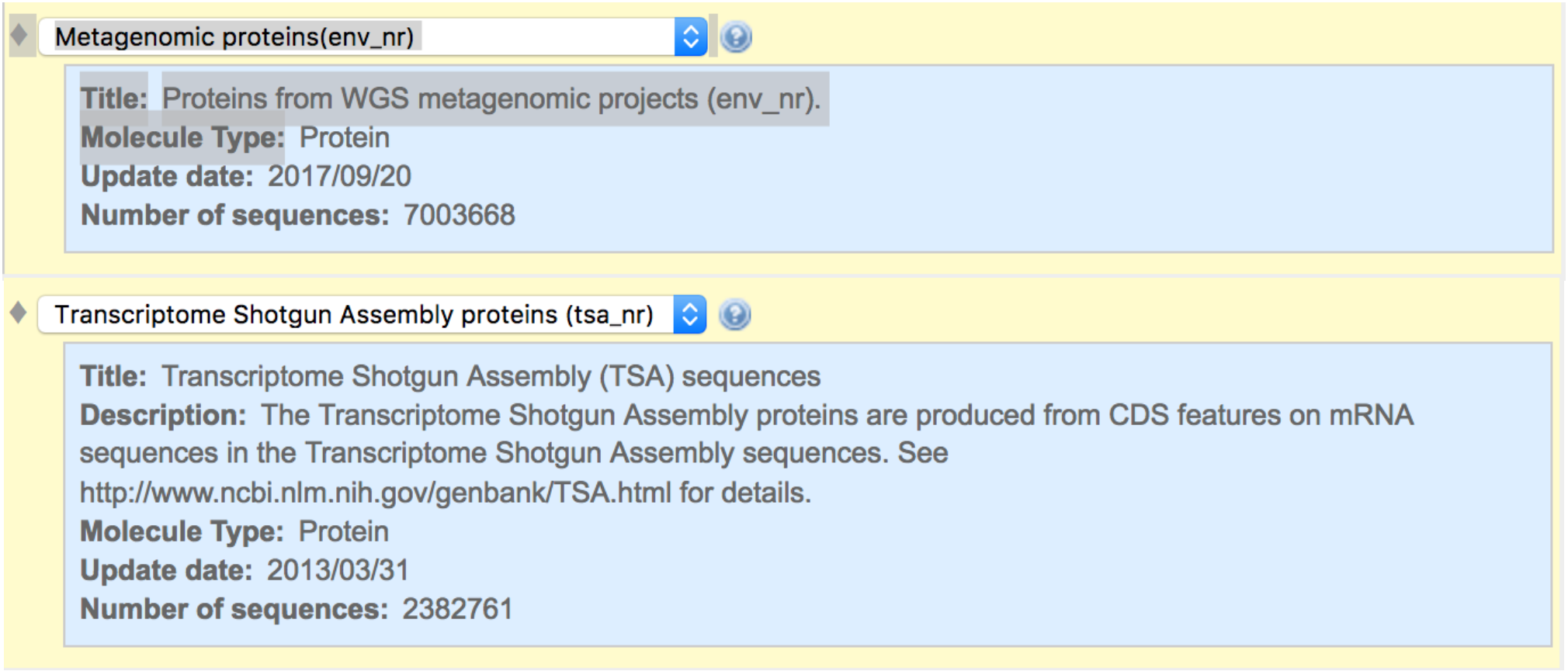
Title: Protein sequences derived from the Patent division of GenBank
Molecule Type: Protein
Update date: 2017/09/20
Number of sequences: 2054527

Protein Data Bank proteins(pdb)

Title: PDB protein database
Description: This database consists of sequences from the Protein Data Bank (PDB), which contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies.
Molecule Type: Protein
Update date: 2017/09/20
Number of sequences: 93500

Protein Databases available for BLAST search

On BLAST search page, select.a database to search and then select ? to learn a little about that database.



forms of BLAST

BLAST	query	database
blastn	nucleotide	nucleotide
blastp	protein	protein
tblastn	protein	translated DNA
blastx	translated DNA	protein
tblastx	translated DNA	translated DNA
psi-blast	protein, profile	protein
phi-blast	pattern	protein

IUPAC nuc
A
C
G
T (or U)
R
Y
S
W
K
M
B
D
H
V
N
. or -

How significant is that?

Please give me a number for...

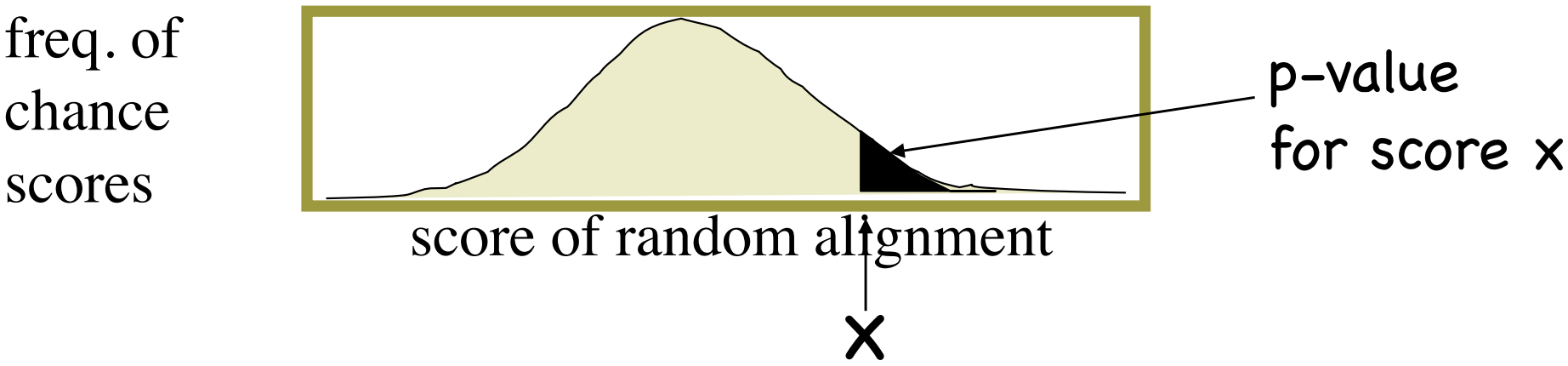
...how likely the data would not have been the result of chance,...

...as opposed to...

...a specific inference.

p-value

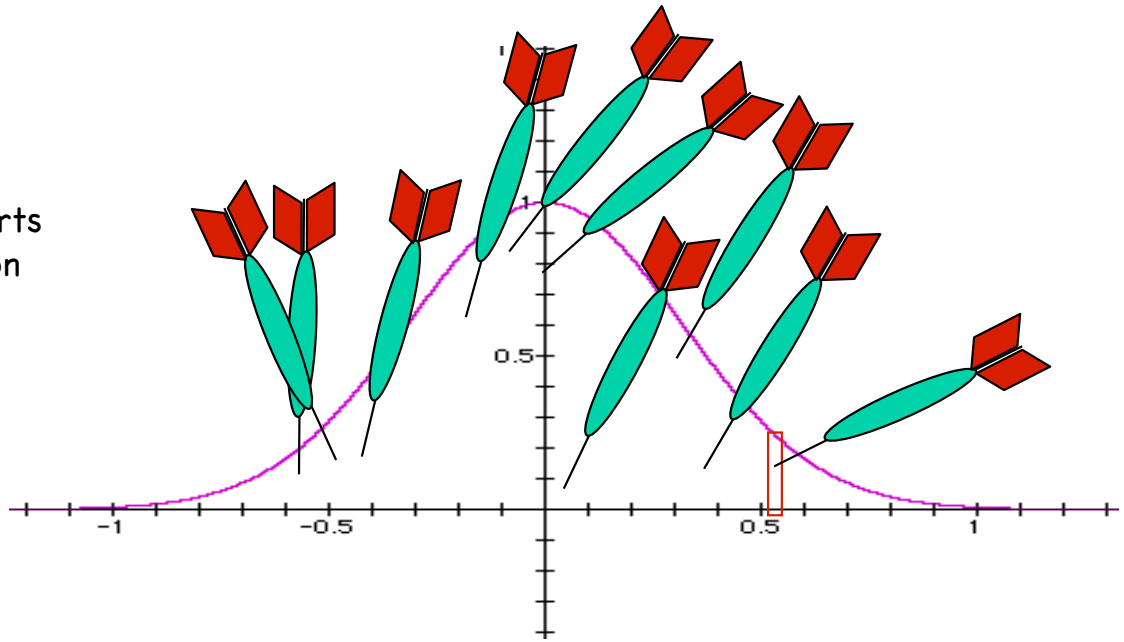
The p-value for score x is the likelihood that a chance score is greater than x.



Thus, the p-value is the probability that the result is "by chance."

p-value is the significance of one (1) alignment score.
e-value is the significance of the highest score out of many tries.

Throwing **n** random darts at a normal distribution

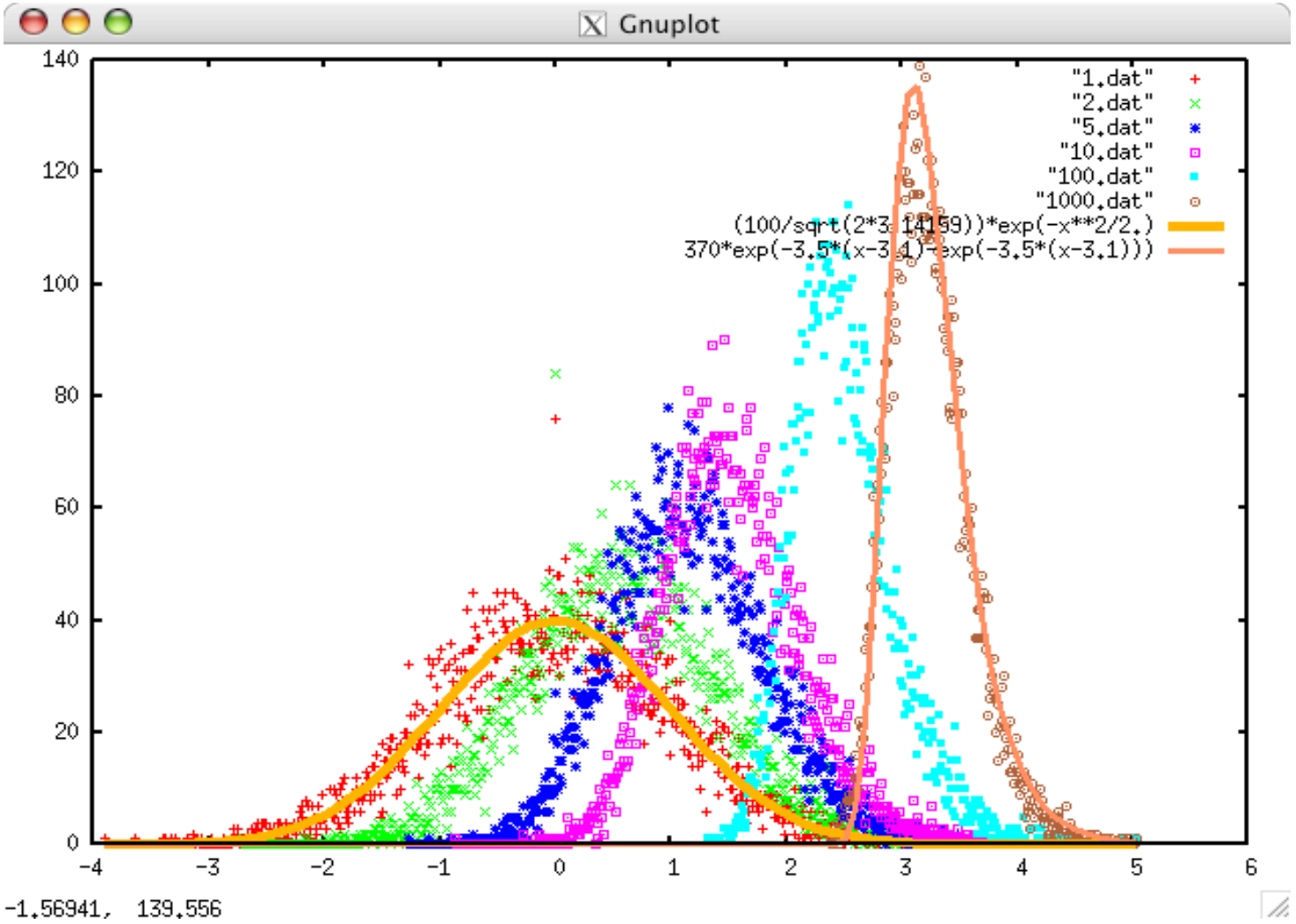


Searching a database of 162 million sequences for one hit is like trying 162 million times to get one good high score.

So, the *number of times you will see that **high score** by chance* is the p-value times 162 million.

e-value = p-value * 162,000,000 (GenBank search)

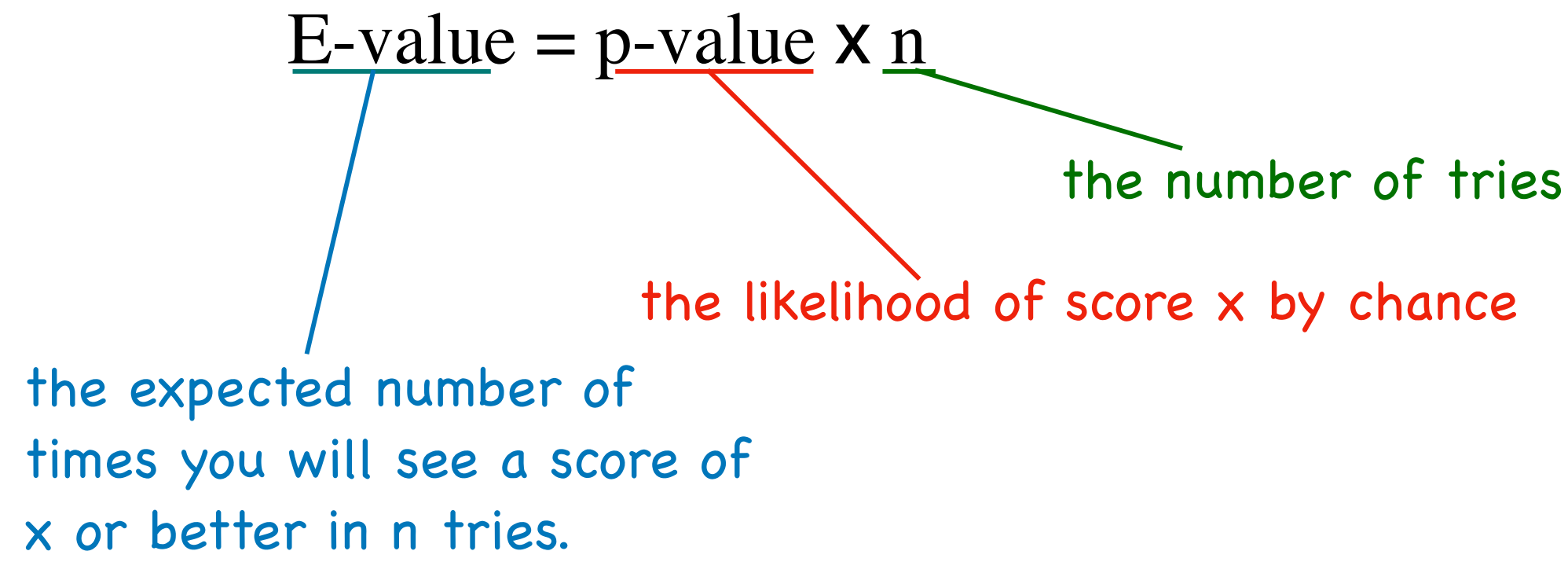
The Extreme Value Distribution



Histograms of Random Gaussian scores, best of N tries.
N=1,2,5,10,100,1000
EVD is the "null" model for database hits.

e-value

A better metric of significance.



Pop-quiz

BLAST HIT.....	e-value
1. annotation	3.0
2. annotation	3.0
3. annotation	3.0
4. annotation	3.0
5. annotation	3.0
6. annotation	3.0
7. annotation	3.0
8. annotation	3.0
9. annotation	3.0
10. annotation	3.0

How many of the above 10 hits are the expected to be by chance?

Pop-quiz

BLAST HIT.....	e-value
1. annotation	1.0
2. annotation	2.0
3. annotation	3.0
4. annotation	4.0
5. annotation	5.0
6. annotation	6.0
7. annotation	7.0
8. annotation	8.0
9. annotation	9.0
10. annotation	10.0

How many of the above 10 hits are the expected to be by chance?

Pop-quiz

BLAST HIT.....	e-value
1. annotation	0.0
2. annotation	0.01
3. annotation	0.01
4. annotation	0.01
5. annotation	0.02
6. annotation	0.02
7. annotation	0.02
8. annotation	0.02
9. annotation	0.02
10. annotation	10.0

How many of the above 10 hits are the expected to be by chance?

Biology Grad Core course: Discussion Topic

Merck Smith-Kline was the author on a study of Trioxx, an anti-inflammatory drug used to treat arthritis, for which it was know to be effective. The study followed over 500 long-time Trioxx users and an equal number of control subjects who had never used the drug. Dr. Smith-Kline was looking for correlations between the use of Trioxx and the incidence of any disease other than arthritis, in any demographic group. He noted in the study that Tunisian Americans, in the age range from 45-55, male or female, and who had been a vegetarian for more than 6 months at any time in their lives, had a "strong negative correlation" between the use of Trioxx and the incidence of restless leg syndrome (RLS), and began touting Trioxx as an effective anti-RLS drug.

The numbers were as follows:

Total Tunisian American vegetarians age 45-55 : 32
Total Tunisian American vegetarians age 45-55 Trioxx users : 16
Total Tunisian American vegetarians age 45-55 Trioxx non-users : 16
Total Tunisian American vegetarians age 45-55 who have RLS : 4
Total Tunisian American vegetarians age 45-55 who do not have RLS : 28

	Trioxx non-users	Trioxx users
RLS	4	0
no RLS	12	16

Dr. Smith-Kline correctly calculated the correlation between Trioxx and RLS as follows:
Corr =, where $u_i = 1$ if subject *i* is a user, and 0 otherwise, r_i is 1 if the subject has RLS and 0 otherwise.

$$\frac{\sum \left(u_i - \langle u \rangle\right)\left(r_i - \langle r \rangle\right)}{\sqrt{\sum \left(u_i - \langle u \rangle\right)^2 \sum \left(r_i - \langle r \rangle\right)^2}}$$

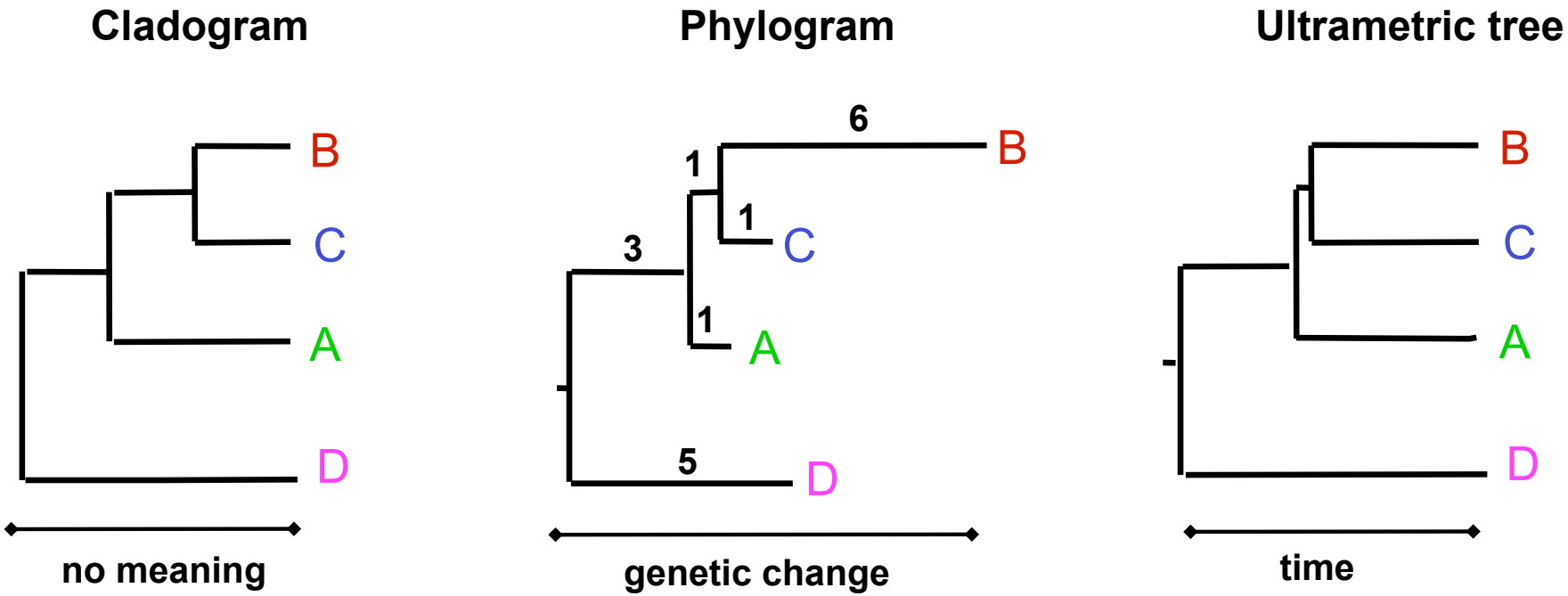
The sums were carried out over all 32 subjects in the subset, and the resulting correlation was -0.378. This confidence level was cited as 99%, since the p-value for this correlation was 0.01, The sample size of 32 and the uneven distribution of subjects with RLS were taken into account.

The data itself was collected correctly and the calculations were correct, both for the correlation and its confidence. Yet Merck Smith-Kline did something dishonest in this study. What was it and what specific question would you ask him to reveal his dishonesty?

Bioinformatics

- Sequence database searching
- **Phylogenetic Trees**
- Protein Structure

Evolutionary time



(D:5,(A:1,(C:1,B:6):1):3)

parenthesis (Newick) notation has both labels and distances.

Multiple Sequence Alignment

A multiple sequence alignment is made using many pairwise sequence alignments

Construct a distance-based tree

	A	B	C	D	E	F
A		97				32
B			77		55	
C				90		40
D					61	
E						33
F						

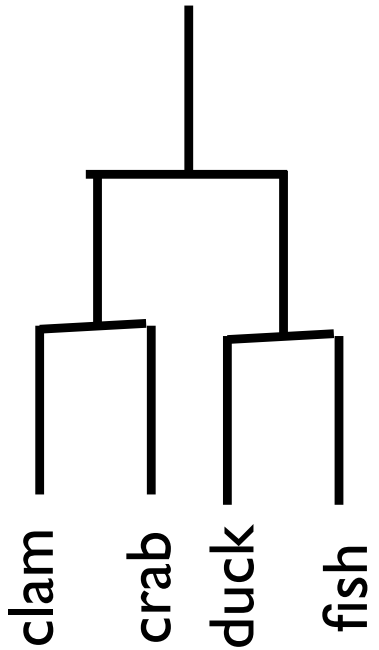
distances

A
B
C
D
E
F

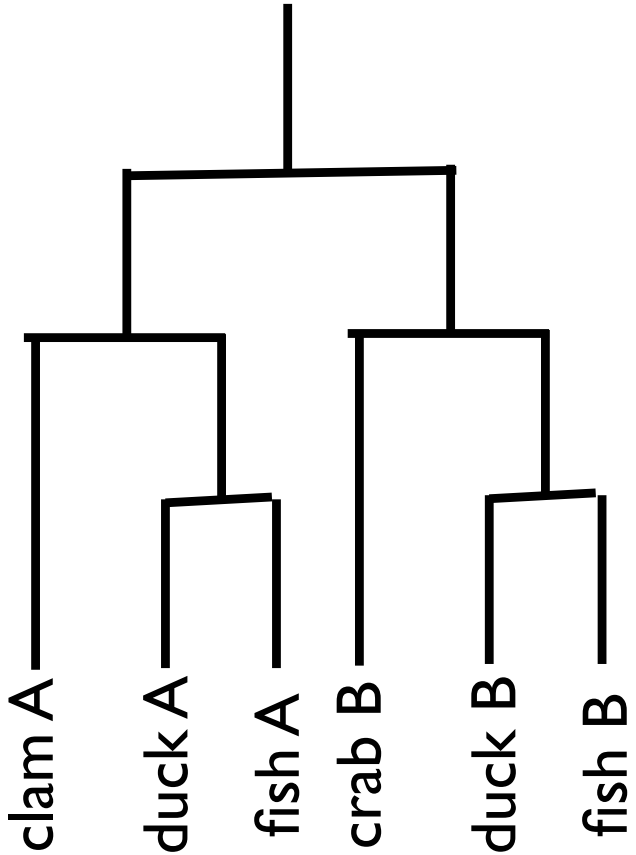
Draw tree here

Sequence homology trees are complicated by paralogy

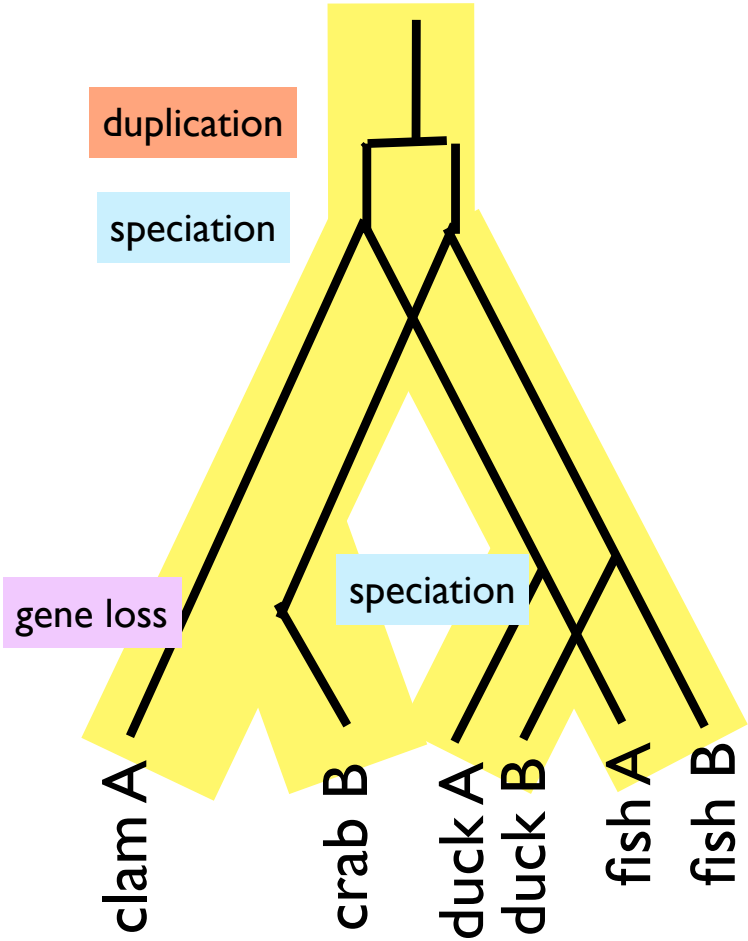
Orthologs: homologs originating from a speciation event
Paralogs: homologs originating from a gene duplication event.



True Species tree



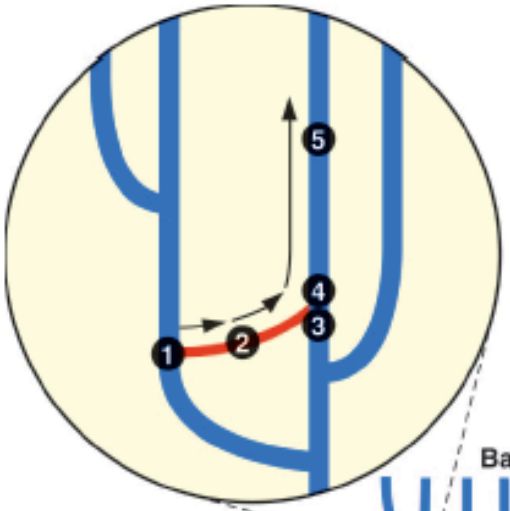
Sequence tree !!

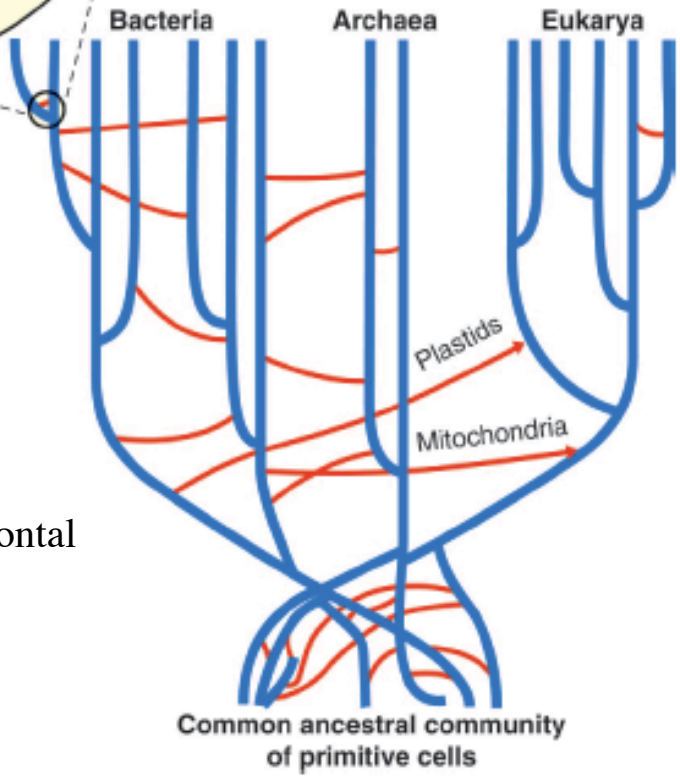


reconciled trees

Life is not strictly a tree --

horizontal gene transfer





BF Smets, T Barkay (2005) "Horizontal gene transfer: perspectives at a crossroads of scientific disciplines" Nature Reviews Microbiology.

Discrete Steps Needed for Stability of Gene Transfer

Stably incorporating horizontally transferred genes into a recipient genome involves five distinct steps (Fig. 1). **1.** First, a particular segment of DNA or RNA is prepared for transfer from the donor strain through one of several processes, including excision and circularization of conjugative transposons, initiation of conjugal plasmid transfer by synthesis of a mating pair-formation protein complex, or packaging of nucleic acids into phage virions. **2.** Next, the segment is transferred either by conjugation, which requires contact between the donor and recipient cells, or by transformation and transduction without direct contact. **3.** During the third step, genetic material enters the recipient cell, where cell exclusion may abort the transfer. **4.** Otherwise, during the fourth step, the incoming gene is integrated into the recipient genome by legitimate or sitespecific recombination or by plasmid circularization and complementary strand synthesis. Barriers to transfer during this step come from restriction modification systems, failure to integrate and replicate within the new host genome, and incompatibility with resident plasmids. **5.** In the final step, transferred genes are replicated as part of the recipient genome and transmitted to daughter cells in stable fashion over successive generations. Researchers from different disciplines tend to focus on specific stages within this five-step sequence. Thus, evolutionary biologists who examine microbial genomes for evidence of past transfers tend to look at HGTs from the perspective of step five. Molecular biologists are more likely to examine the details of the transfer events, while microbial ecologists look more broadly when they describe the magnitude and diversity of the mobile gene pool, sometimes called the mobilome.

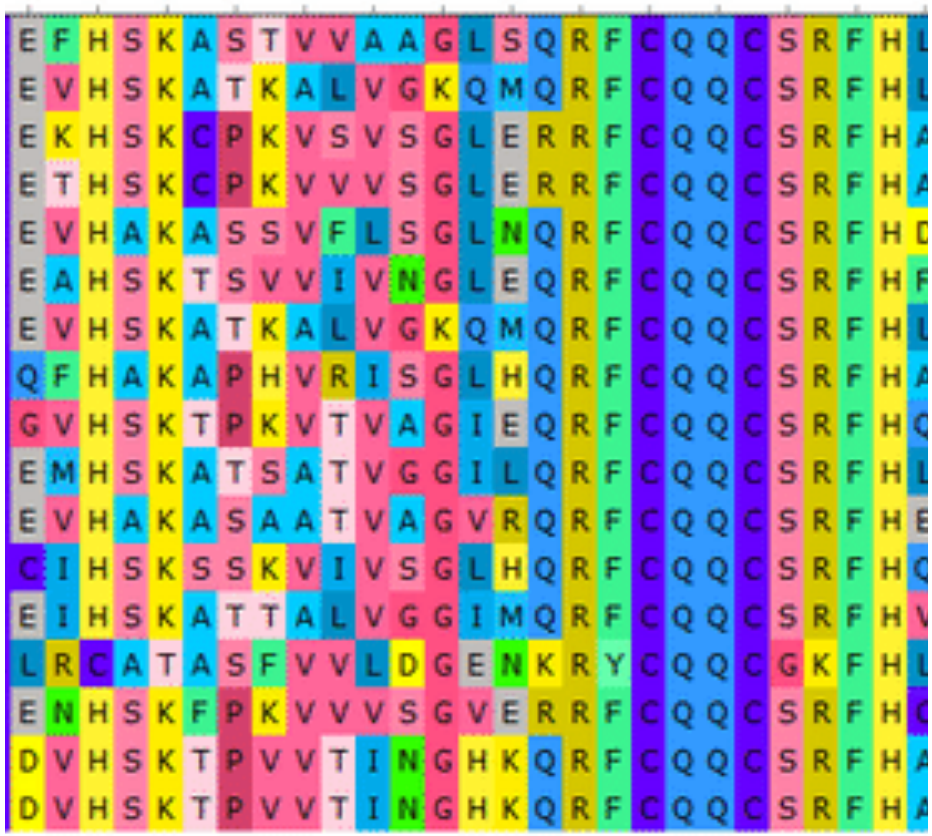
Use orthologs to make trees

- To make the right inferences about evolution, make sure your phylogenetic tree is composed of orthologs

How do you know it's an ortholog?

1. It has the same function in both species.
2. It has about the same number of differences across species as other orthologs.
3. Often you can't know.

Functional inference from multiple sequence alignments

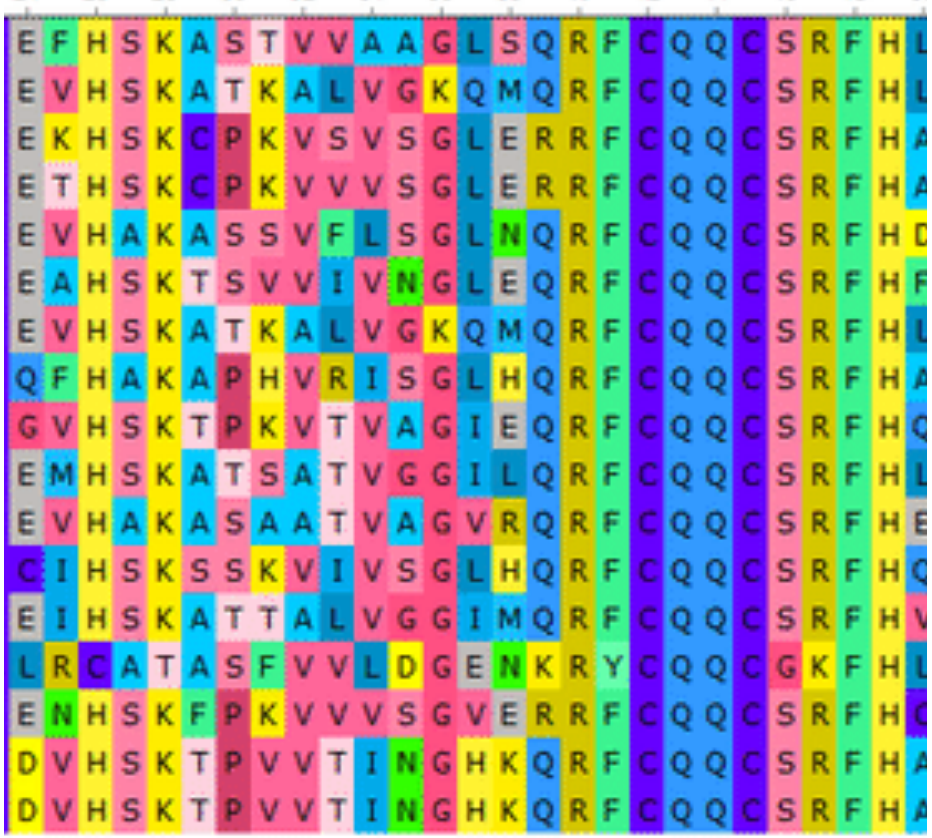


Not conserved

Conserved

folding

function



Not conserved

Conserved

species differences

folding

stability

kinetics

function

enzyme activity

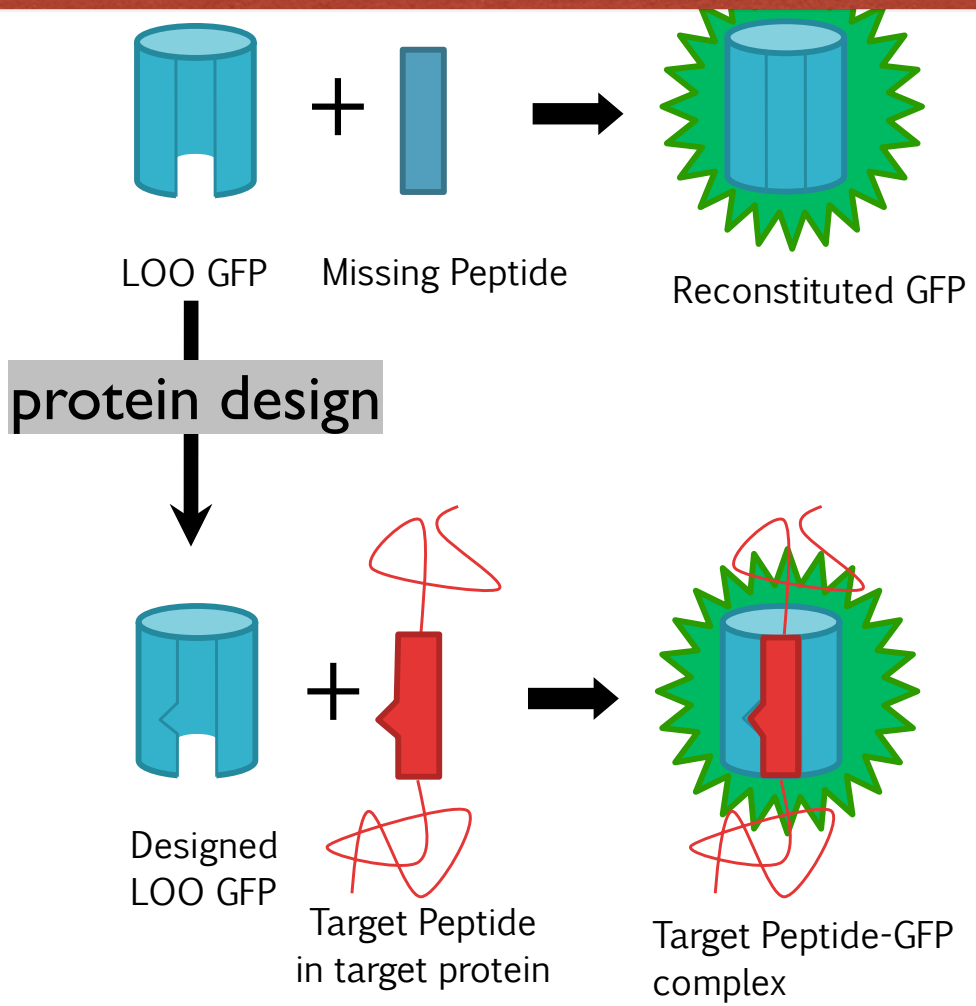
binding

post-translational modification

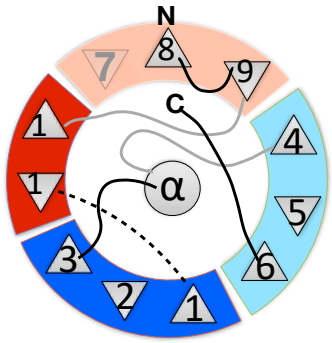
Next time:

- Visit rcsb.org
- Try visualizing a protein.
- Locate a residue that is conserved across all species in a BLAST search.
- Locate one that is conserved *except in one species*. What might be its function?

Computationally designed LOO-GFP biosensors



Crone, D. E., Huang, Y. M., Pitman, D. J., Schenkelberg, C., Fraser, K., Macari, S., & Bystroff, C. (2013). GFP-based biosensors. In *State of the Art in Biosensors-General A*



LOO7-HA

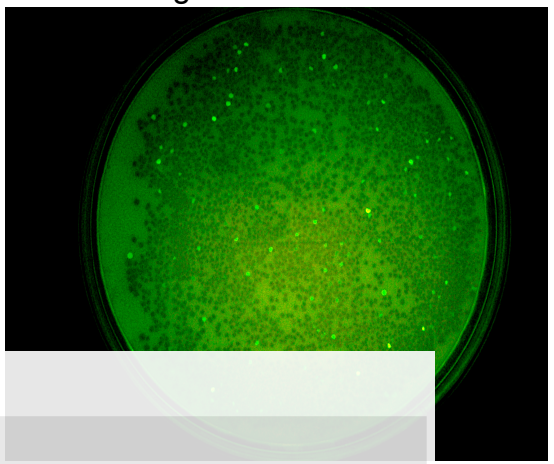
1.... Strand 7 changed to H1-antigen SSHEVSLGVSSA

2.... Dead-end elimination design algorithm applied to surrounding positions.

3.... Gene library made by *degenerate codon*

4.... Library was co-expressed with HA peptide. Colonies picked.

Overnight colonies of LOO7-HA library



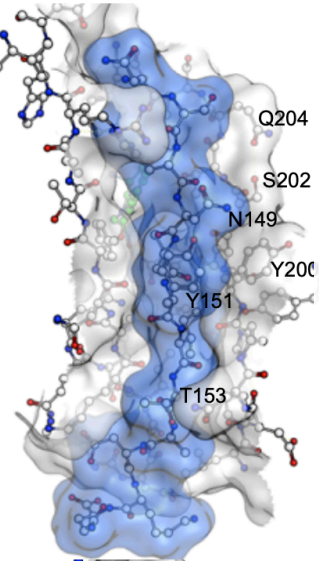


Table 1. Results of Design and Selection for Designable Positions of LOO7-GFP[a]

		DEEdesign				
position	wild-type residue	input	output	degenerate codon	library	LOO7-HA4
83	F	AFILMVW	FW	TKS	FLCW	**W**
84	F	AFILMVW	FM	WTK	MILF	F
161	I	AFILMVW	ILV	VTA	ILV	I
163	A	AFILMVW	I	ATC	I	**I**
164	N	NST	NST	AVY	NST	**T**
165	F	AFILMVW	F	TTT	F	F
167	V	AFILMVW	IMV	RTN	IMV	V
168	R	RNST	NST	AVY	NST	**T**
200	Y	YKHR	HKR	MRS	HKR	**H**
201	L	LFMW	FM	WTK	MILF	L
202	S	SKHR	HKR	MRS	HKR	**K**
204	Q	QNST	NST	AVY	NST	**T**
224	V	AFILMVW	IV	RTK	IV	V

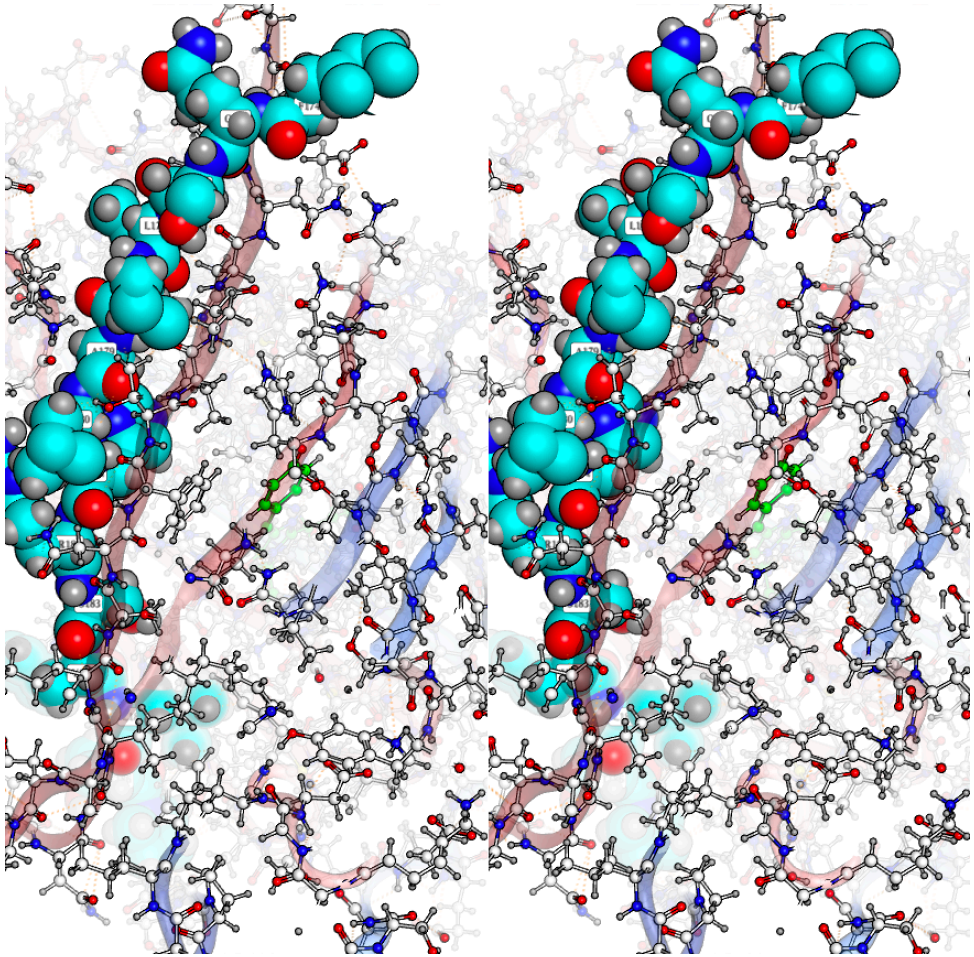
[a]Position numbers refer to the unpermuted superfolder GFP OPT sequence as shown in Figure S1. LOO7-HA4, a specific design with 7 mutations (bold), was one of approximately 75 glowing colonies. Degenerate codons are shown using IUPAC nucleotide notation.

Huang, Y. M., & Bystroff, C. (2009). Complementation and reconstitution of fluorescence from circularly permuted and truncated green fluorescent

Huang, Y. M., Banerjee, S., Crone, D. E., Schenkelberg, C. D., Pitman, D. J., Buck, P. M., & Bystroff, C. (2015). Toward computationally designed self-reporting biosensors using leave-one-out green fluorescent protein. *Biochemistry*, *54*(40), 6263-6273.

New, Manual design process.

1. Pick a target sequence.
2. Pick a strand to leave out (4,7,8,9, or 11)
3. Mutate the strand to the target sequence.
4. Mutate residues around the target until all criterea are met.



Protein Data Bank

- rcsb.org
- **4ms2** a voltage-gated calcium channel.

1) visualize overall structure in NGL
2) view ligands, ions
3) view electron density
3) Look at ion gate.
4) Look at ion-selective channel.
5) Look at voltage sensitive domain

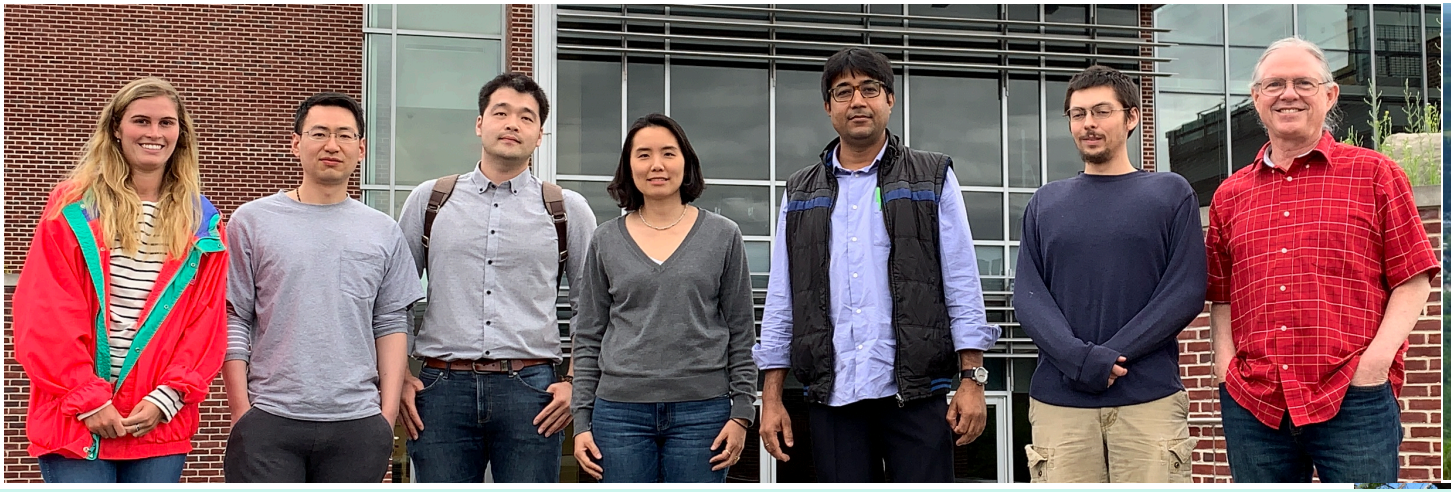
Next time:

- Read CatSper vaccine paper

KN, Jordan TB, Yuan X,Basore DA, Zagorevski D, Clarke C, Werner G, Hwang JY, Wang H, Chung J-J, McKenna A. Jarvis M, Singh G, Bystroff C. (2023) Bacterial production of recombinant contraceptive vaccine antigen from CatSper displayed on a human papilloma virus-like particle.. Vaccine, 41(46), 6791-6801

Read at least one part of the paper in detail. Bring a comment, suggestion, or question to class 2/19.

https://www2.bioinfo.rpi.edu/bystrc/pub/Nand2023.pdf



The Contraceptive Vaccine Project

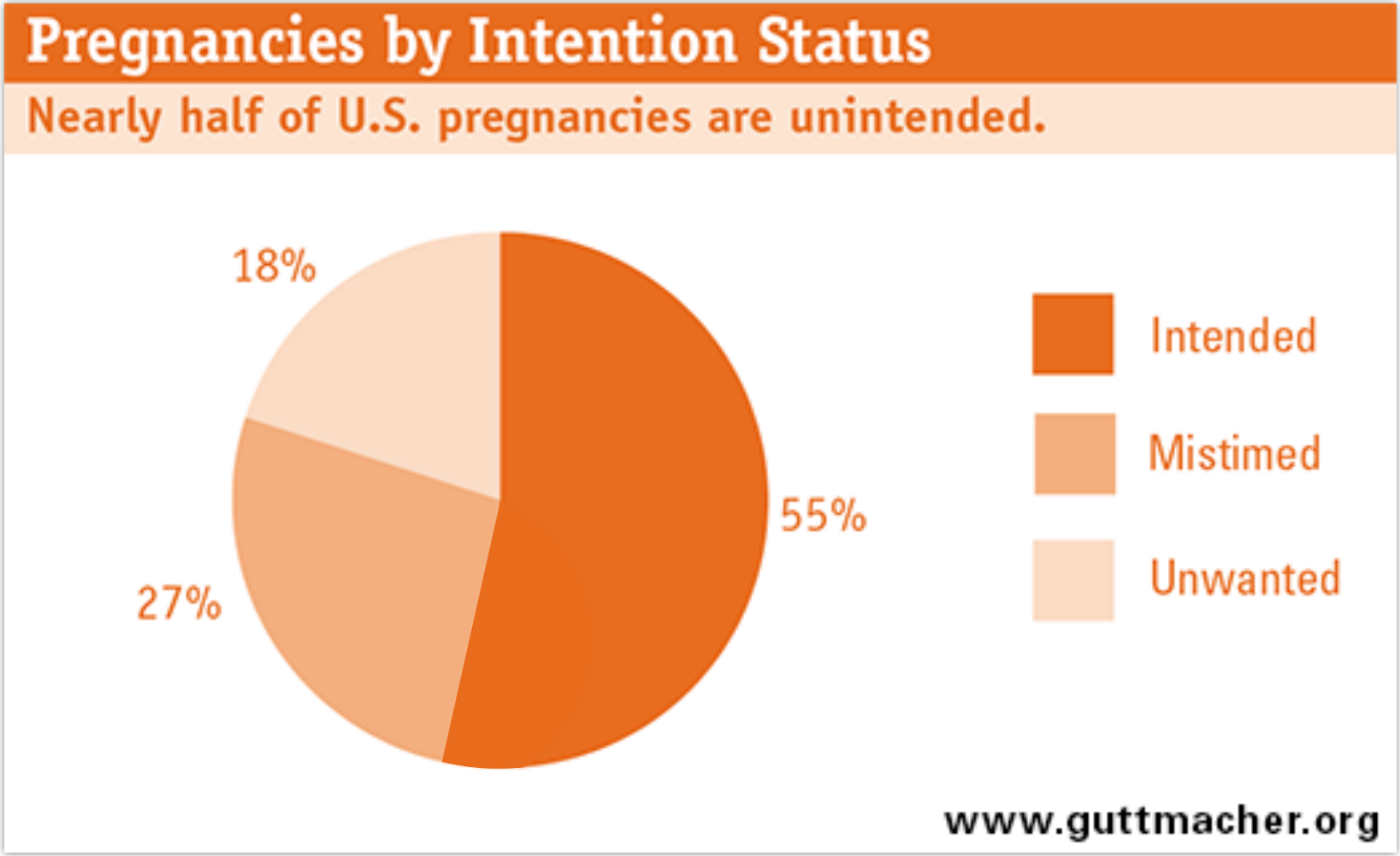
Bystroff Lab@ RPI

Funding:

NIH

GM-099827

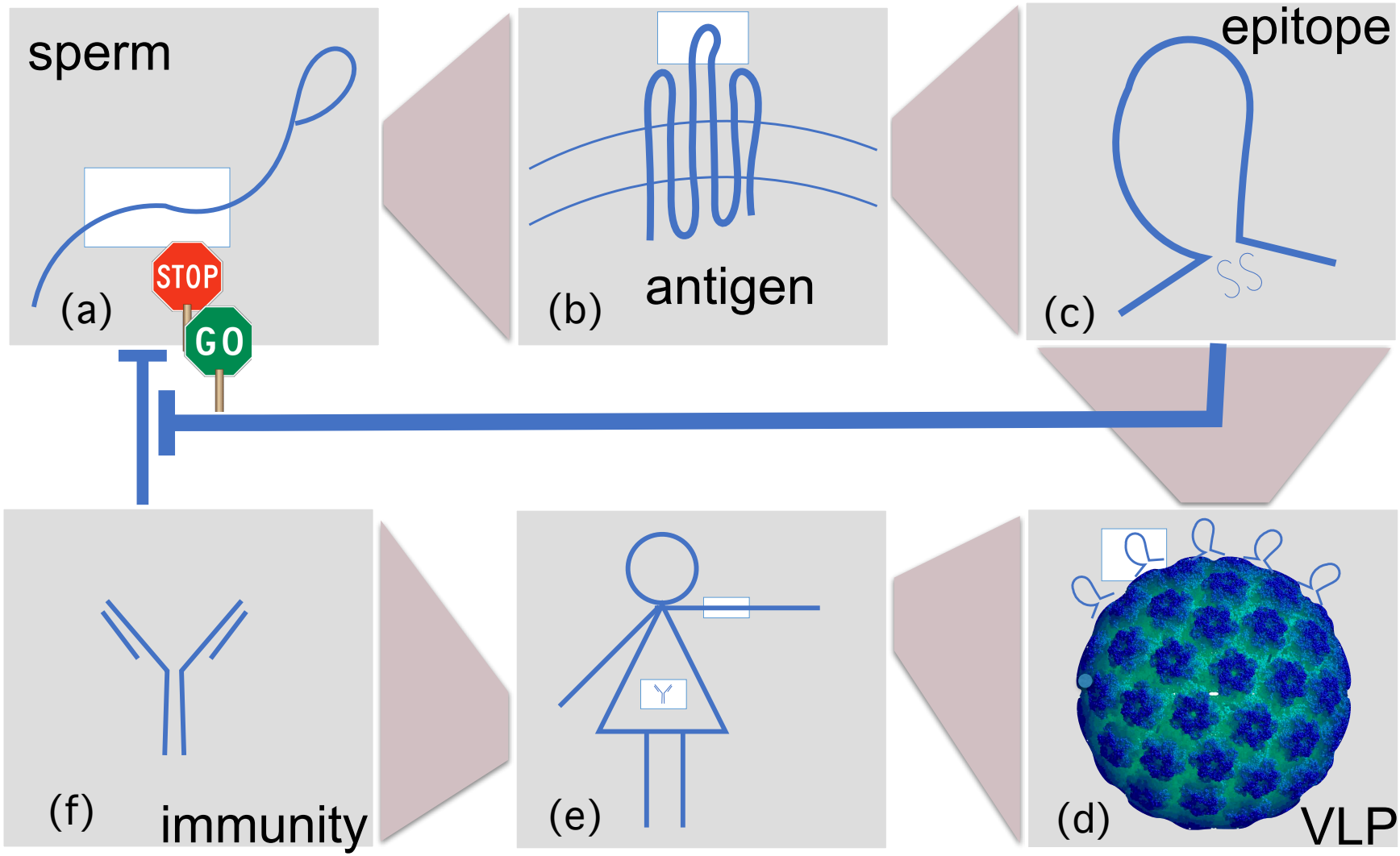
Almost half of all pregnancies are unintended.



Fewer births if every child is a wanted child = 41.3 million/year (2016)

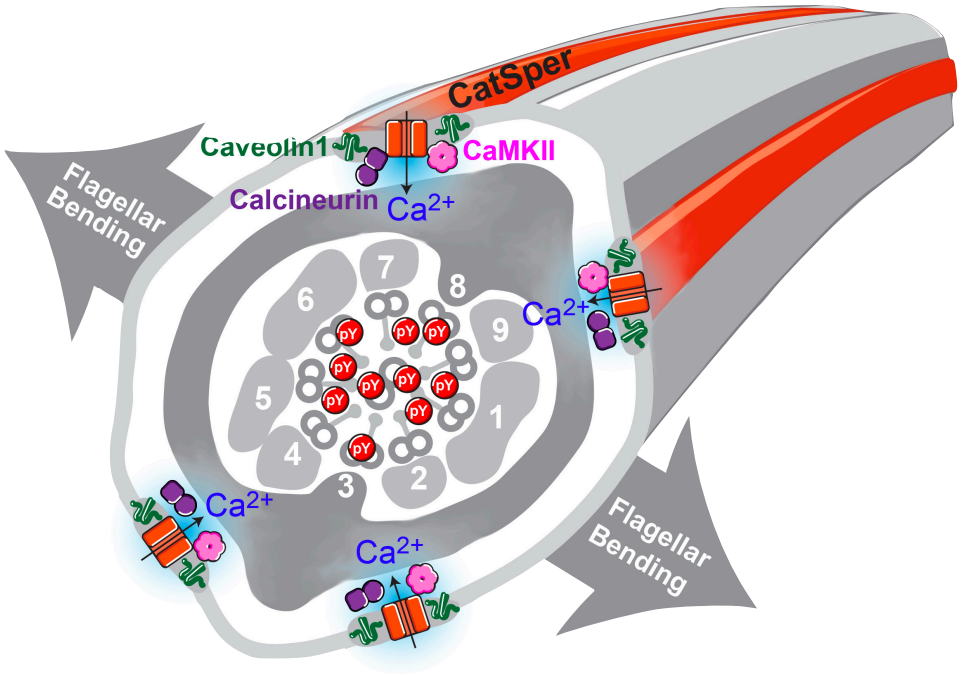
- On by default. No action needed.
- Reversible on demand
- Safe
- Cheap to produce
- 100% effective
- Non-hormonal
- No surgury needed
- One shot

A contraceptive vaccine

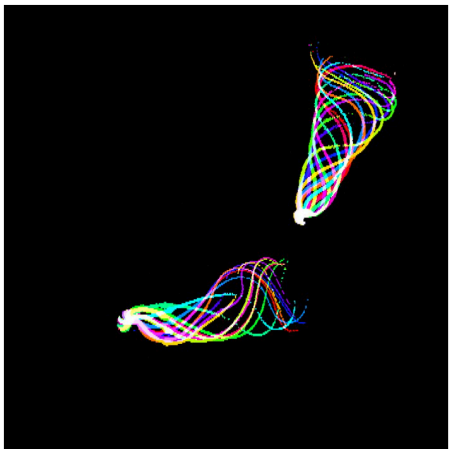


long-term. cheap. hands-off. no side effects. reversible.

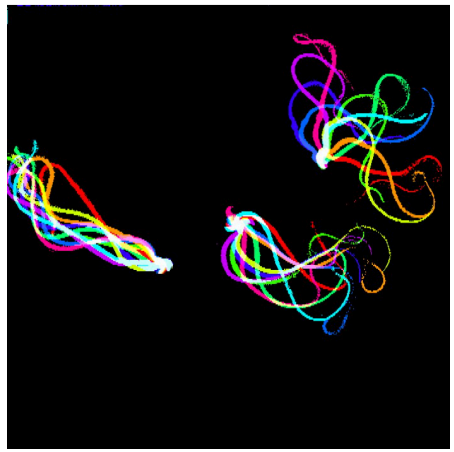
CatSper is sperm-specific, extracellular and required for fertility

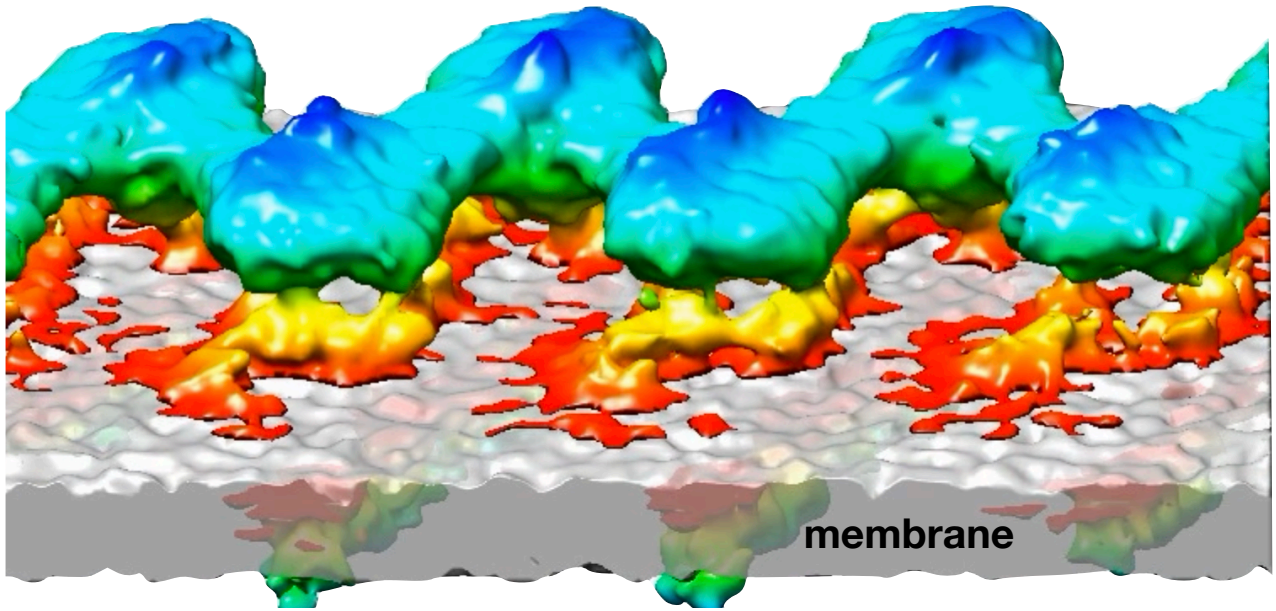


Activated



Hyperactivated

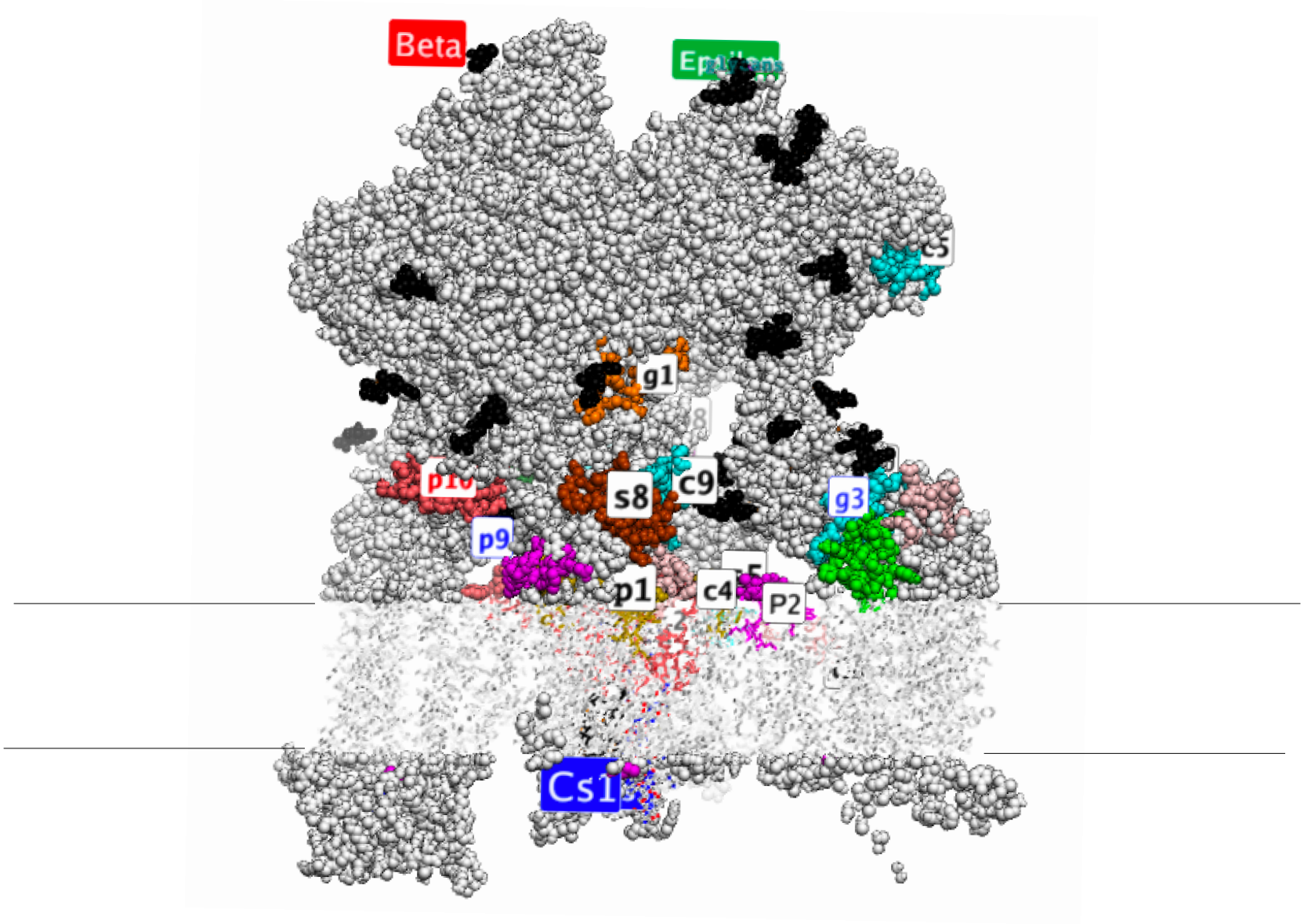




CatSper

Jean-Ju Chung, Yale

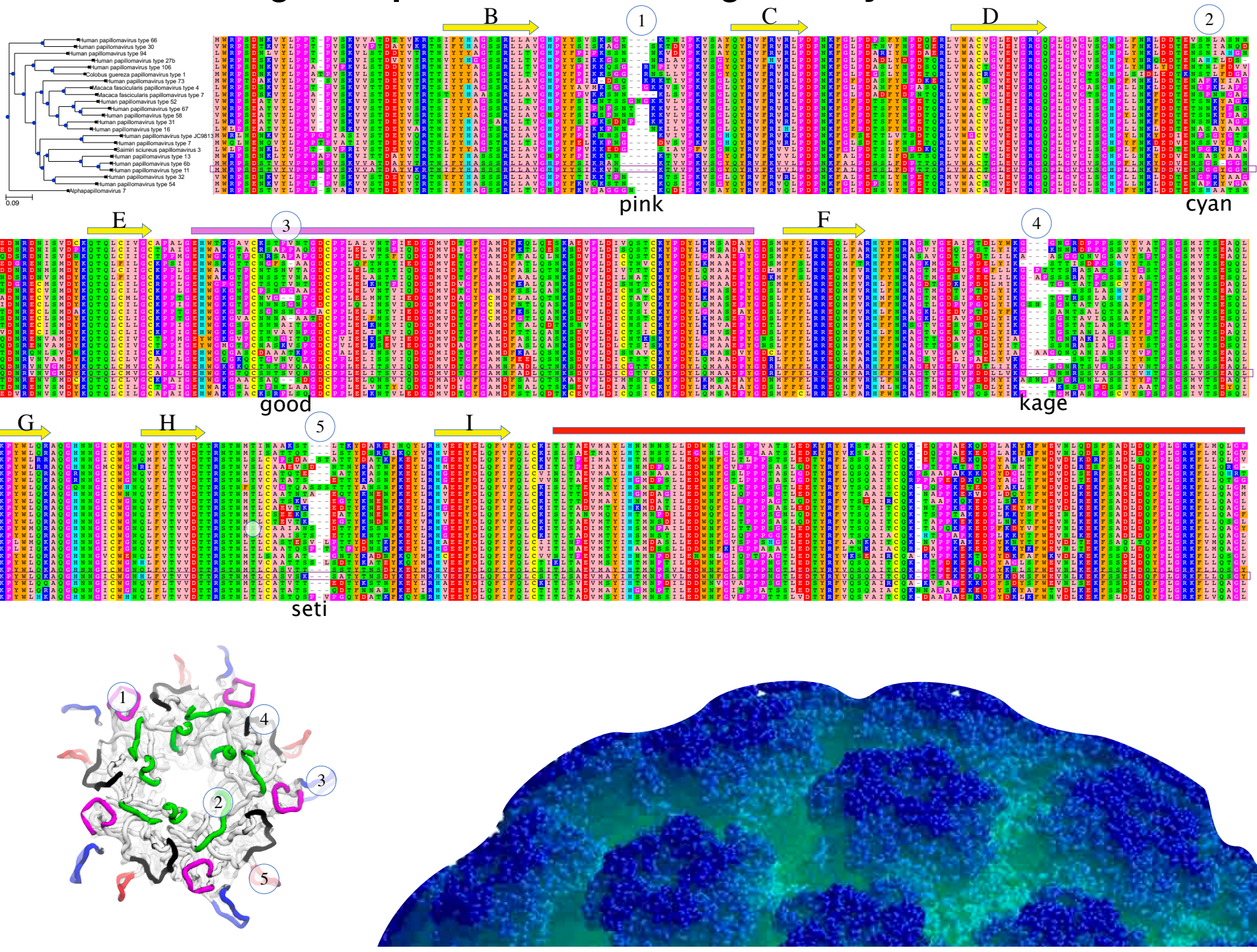
Identification of calcium-blocking epitopes on the CatSpersome



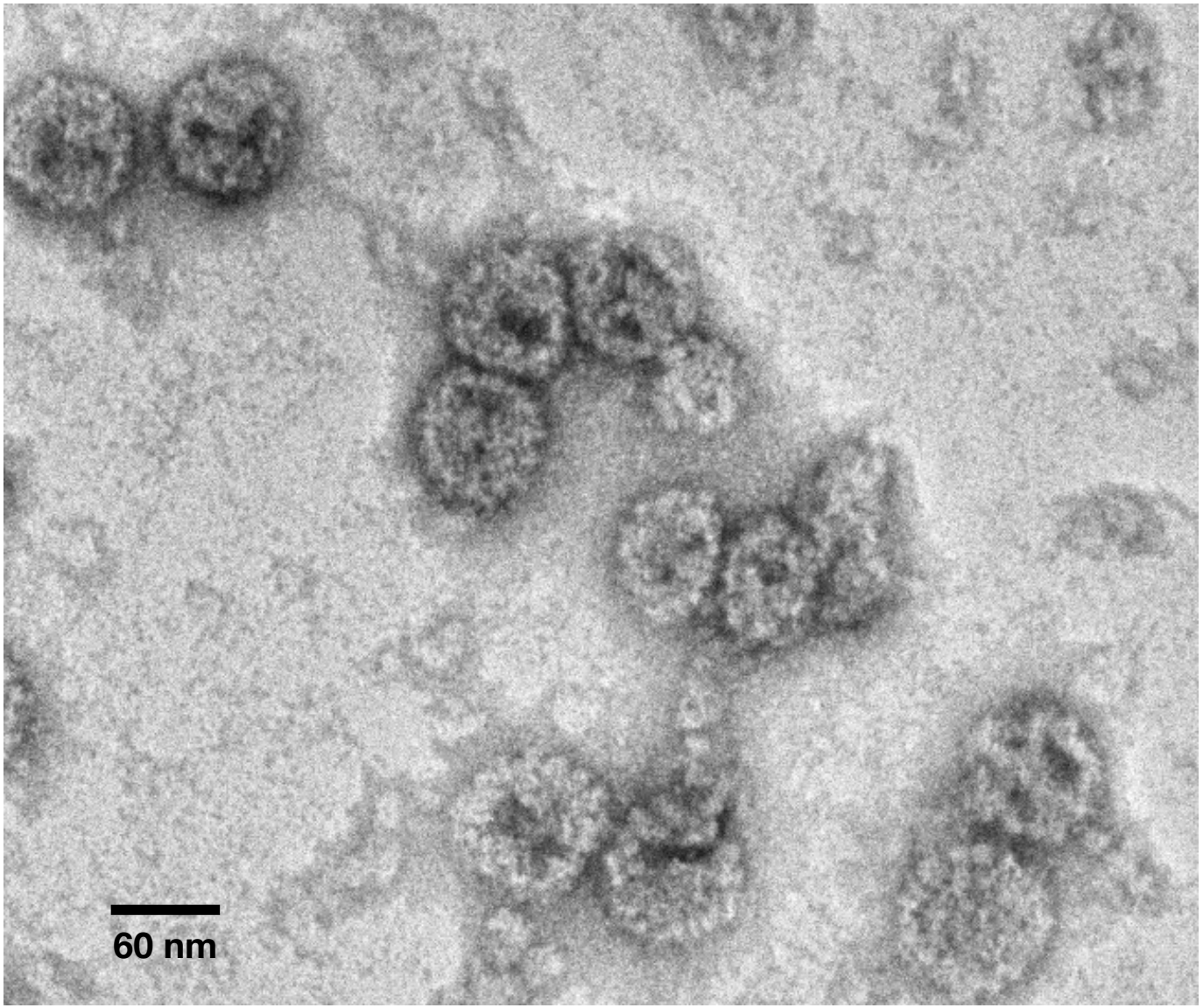
Rensselaer

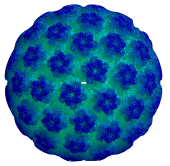
The Bystroff Lab

Placement of antigen loops within HPV L1 is guided by structure and evolution

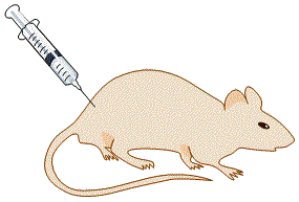


Formation of virus-like particles (VLP) from refolded protein.





Round 5 vaccine testing in mice.



			Mix 1	Mix 2	Mix 3	Mix 4	Mix 5	Arch
p1_klh	mCs1_s3s4	NSLSYSFYNHSLFR	p1_klh			p1_klh		GE
c4_klh	mCs4_Ps6_3	DFQMDEREYAME	c4_klh		c4_klh			GE
p2_klh	mCs2_s1s2	EIELMESTNTALWP	p2_klh			p2_klh		GE
s5_klh	mCs1_ps6	YIDNRAQGAWYII	s5_klh			s5_klh		ED
c6_klh	mCs2_s3s4	LVVL**LG**V**W**LQ	c6_klh	c6_klh				ED
s8_klh	mCsG_YTE	RDY**TEE**EIF**R**YNSPLD**TT**NSLI	s8_klh	s8_klh				GE
p8_klh	mCsD_STF	WSTFES**DIENE**EPFLW	p8_klh	p8_klh				DB
c10_kln	mCsE_QTW	QTWDSMIEENPDIPLDDVWG	c10_kln		c10_kln		c10_kln	ED
c5_cklh	mCSE_331-348	CDGTVYLRTEDEFT**K**LDES	c5_cklh	c5_cklh				ED
c7_cklh	mCsB_HIP	CHIPP**ED**WISG**V**H**K**D**S**QGFNM	c7_cklh		c7_cklh	c7_cklh	c7_cklh	DB
g2_cklh	mCsE_VID	CVID**K**ELLRESLS**DN**LK	g2_cklh			g2_cklh		ED
c8_cklh	mCsD_HED	CHEDNKDNPLL	c8_cklh	c8_cklh		c8_cklh		DB
p10_cklh	mCsB_NEP	CNEPENVKKMKH**Y**LE**PLL**K**T**PVYNPLG	p10_cklh	p10_cklh	p10_cklh		p10_cklh	BG
p11_cklh	mCsD_KGY	CKGYSYDYY**E**NTW**R**KL**E**GISEP	p11_cklh		p11_cklh	p11_cklh		ED
c9_cklh	mCsG_TKV	CTKV**ER**TT**E**D**K**KFY**I**MS**HE**SPG	c9_cklh	c9_cklh			c9_cklh	GE
p9css_cklh	mCsG_FDC	CFDSV**IKD**AEMPSF	p9css_cklh	p9css_cklh		p9css_cklh		BG
g1a_cklh	mCsG_IDK	CID**KKRASEQ**GMIG**R**NIKKT	g1a_cklh		g1a_cklh			GE
s9a_cklh	mSLC_GRQ	GRQSIELAMEENEKRNIIC	s9a_cklh		s9a_cklh			ED
g3	mSLC_s5s6	FQPPPKEKTVEIEPAKVYQLL						
s4	mCs4_s1s2	ALRTNSYLGQKHY						
		ROUND 5	Mix 1	Mix 2	Mix 3	Mix 4	Mix 5	PBS
	sea level	zero-pup animals	0	1	0	4	1	0
		percent efficacy	0	12%	0	50%	12%	n/a