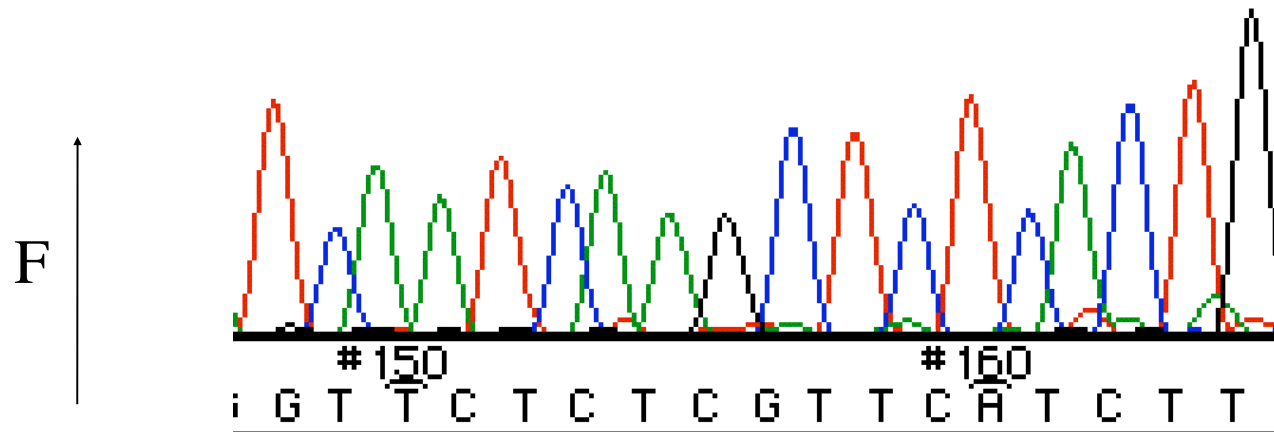


Bioinformatics 1-- Lecture 2

Experimental origins of sequence data

The Sanger dideoxynucleotide method



Each color is one lane of an electrophoresis gel.

base calling

- In Ugene: open data/samples/ABIF/A01.abi or download A01.abi from the link on the course web page (UGENE files)
- Look at the trace. Find beginning and ending of high-confidence sequence region.
- When should you tolerate uncertainty, and when does it matter?

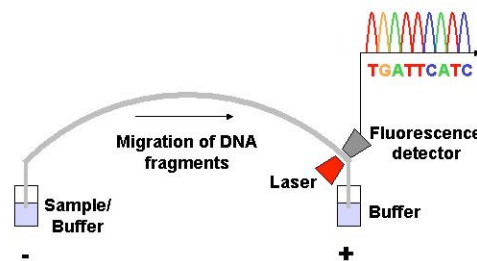
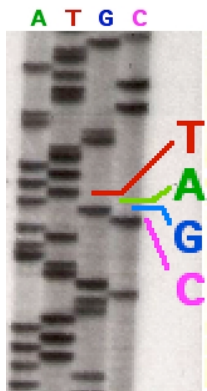
New technology: Pyrosequencing

- <http://www.youtube.com/watch?v=nFfgWGFe0aA&NR=1>
- ..or search youtube for “pyrosequencing”
- *Whole genome sequencing in < 1 day!!*

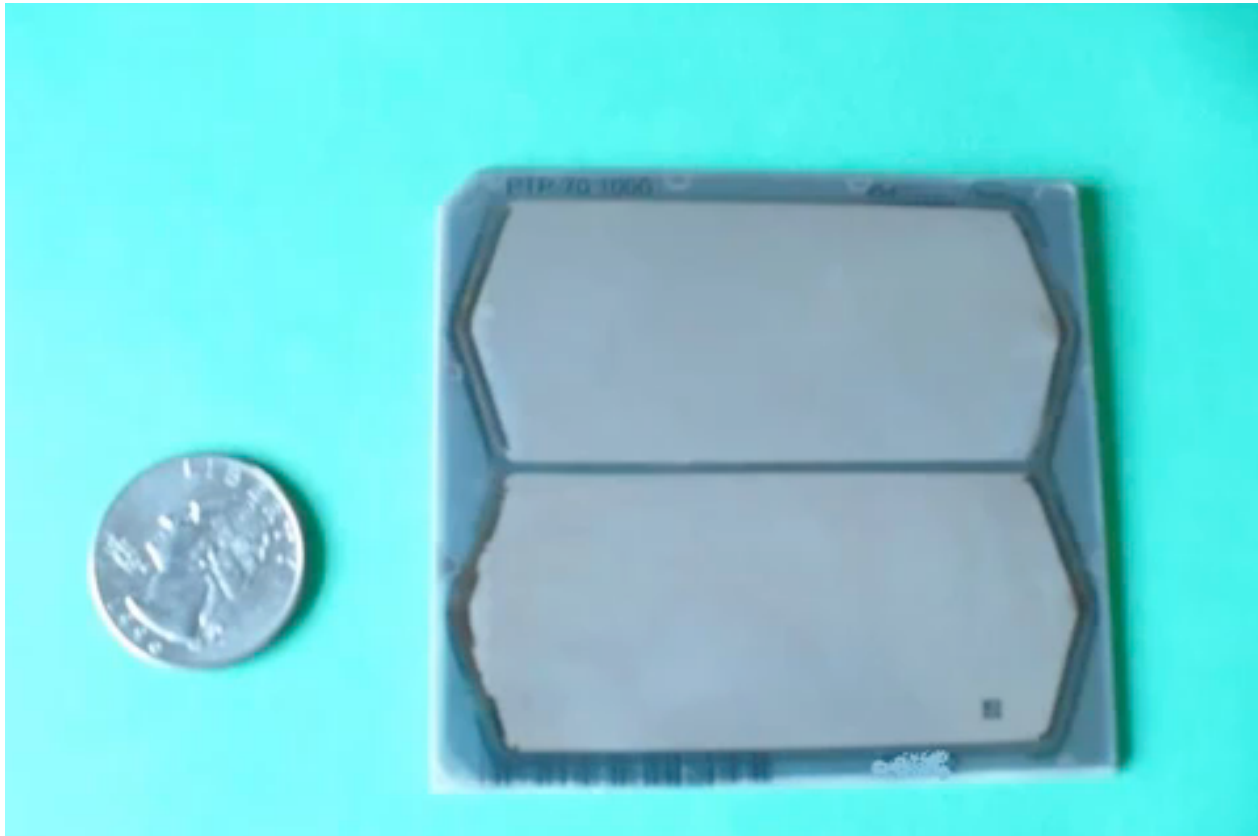
glass plate
200 bp in 3 days

capillary
500 bp

pyrosequencing
 10^9 bp

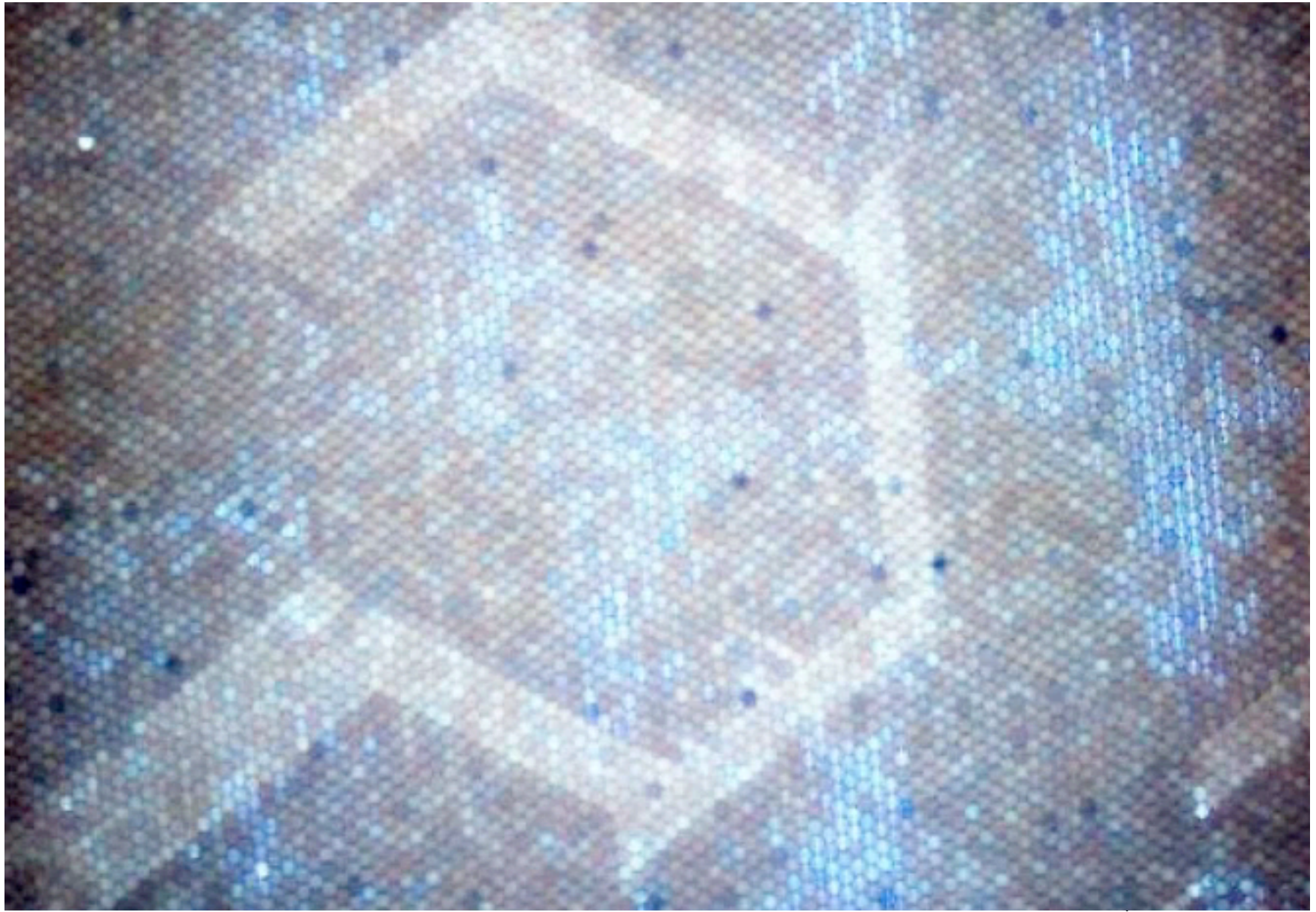


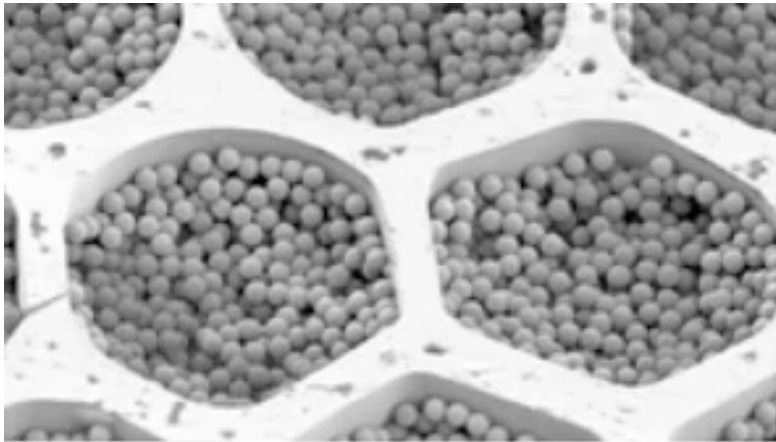
454 sequencing (Hoffman La Roche)



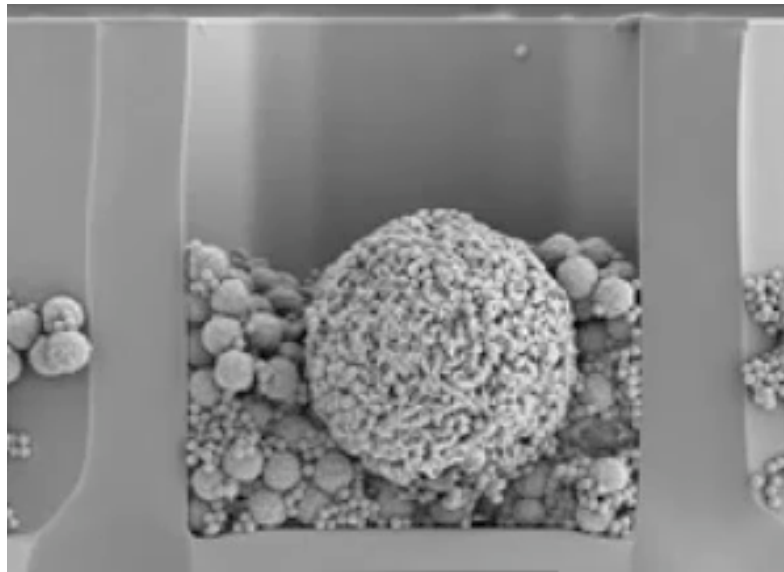
zooming in...



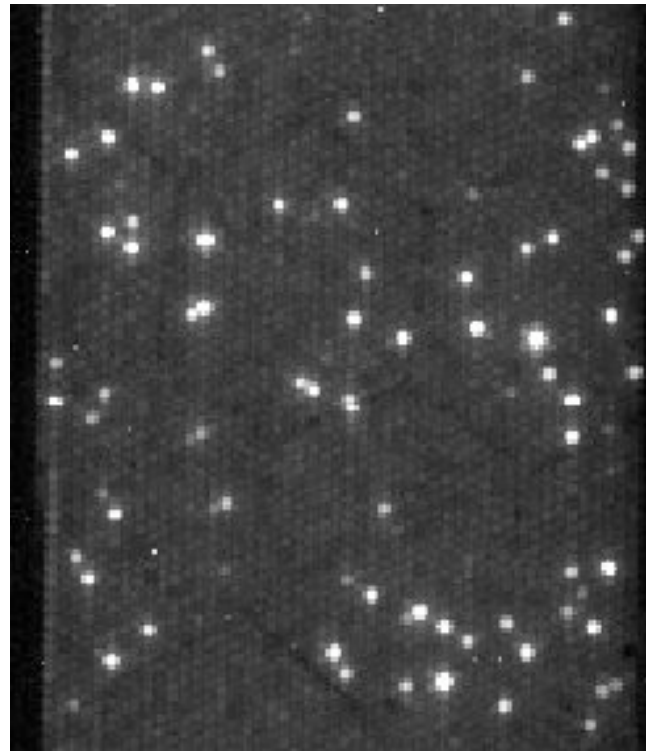




picoliter chambers



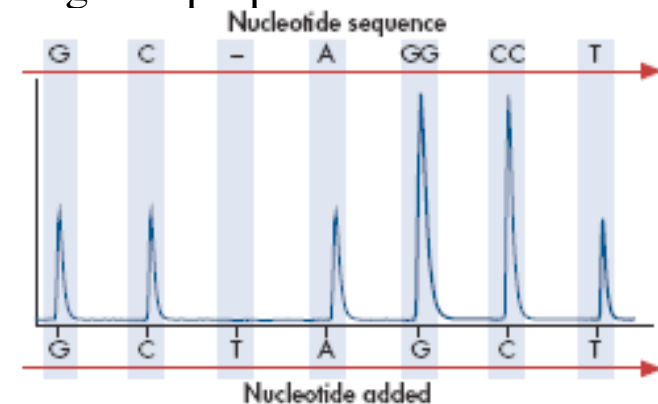
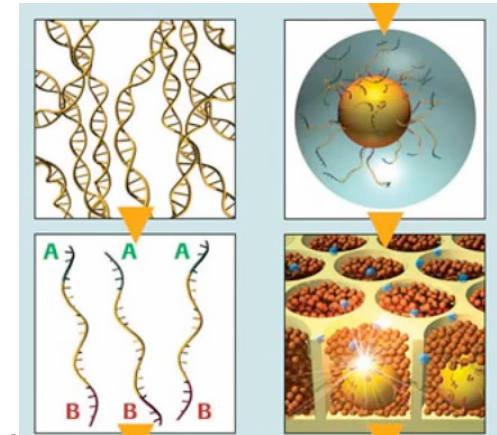
1 sequencing bead per chamber



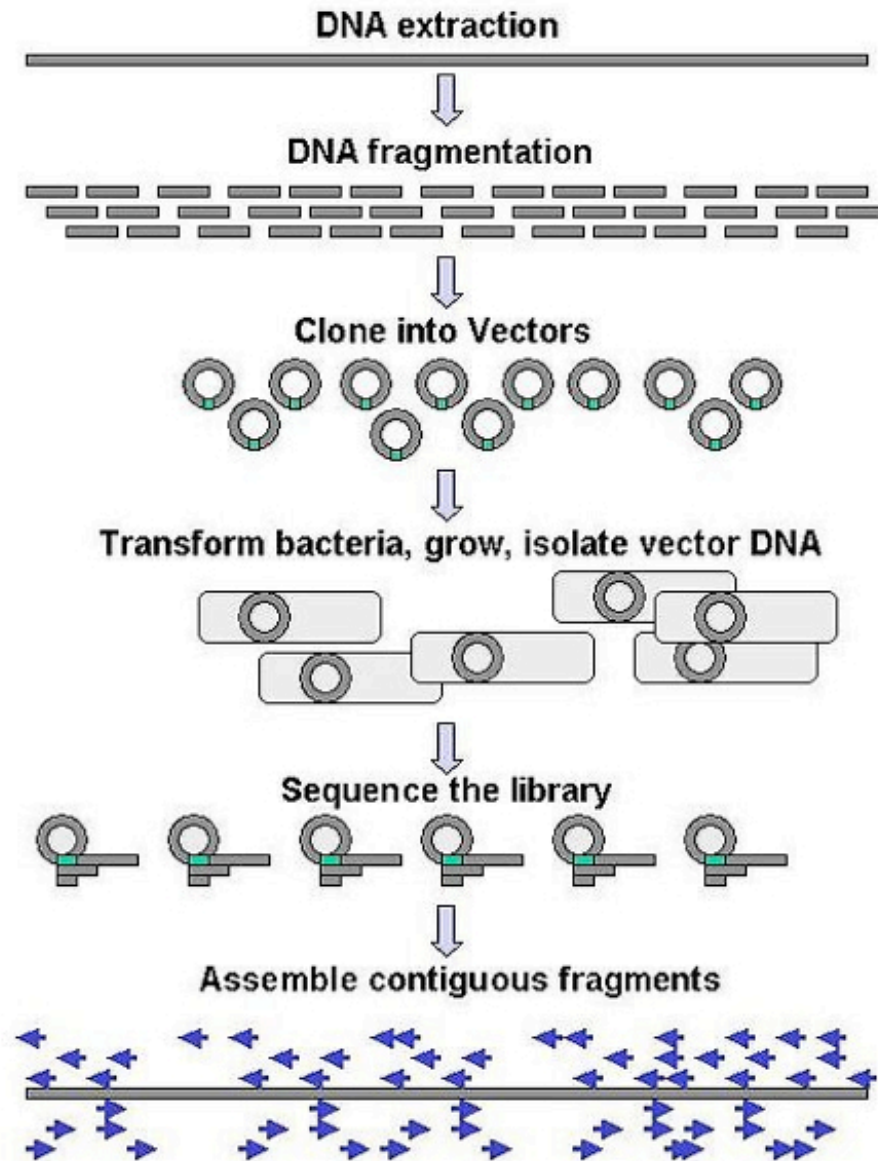
CCD image showing chemoluminescence from chambers after adding dCTP. Brighter dots added more C.

Pyrosequencing

- DNA is sheared or cut. Poly-A tag added.
- Individual DNA strands are attached to poly-T linked beads. 1 strand/bead.
- Beads are added to a microarray matrix. 1 bead/chamber. (picometer size).
- DNA is amplified in place. New copies bind to the bead. +strand only. -strand washed away.
- Add enzymes: DNA polymerase, ATP sulfurylase, apyrase, luciferase. Adding dNTP releases PPi. Apyrase chews up left-over dNTP. ATP sulfurylase catalyzes $PPi + APS \rightarrow ATP$. Luciferase emits light in proportion to ATP.
- Light emission detected by CCD. Each pixel produces a “pyrogram” (see fig).



Whole genome shotgun sequencing protocol



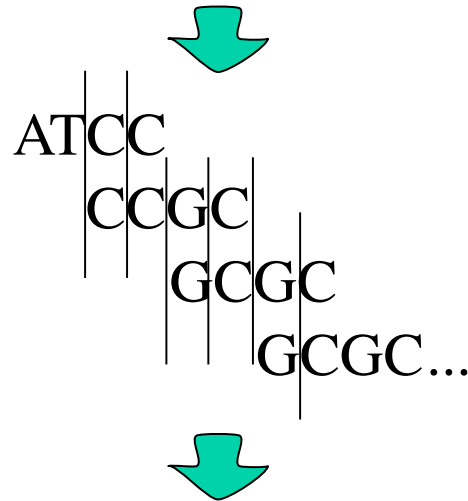
Whole genome shotgun strategy

- Sequence at least 10 times as much DNA as contained in the genome. i.e. If the genome has 4.6 Mb (mega-bases) then sequence 46 Mb. This is called "10-fold redundancy".
- Find all overlapping sequences. (sometimes the overlap is ambiguous)
- If the overlap is ambiguous on one end of the BAC or YAC, the ambiguity can be resolved using the other end.
- Errors in assembly can still occur in **highly repetitive regions** of the genome (such as near the centromeres).

Assembly

reads: CCGG ATCC CCAG CCGC GCGC GCGC

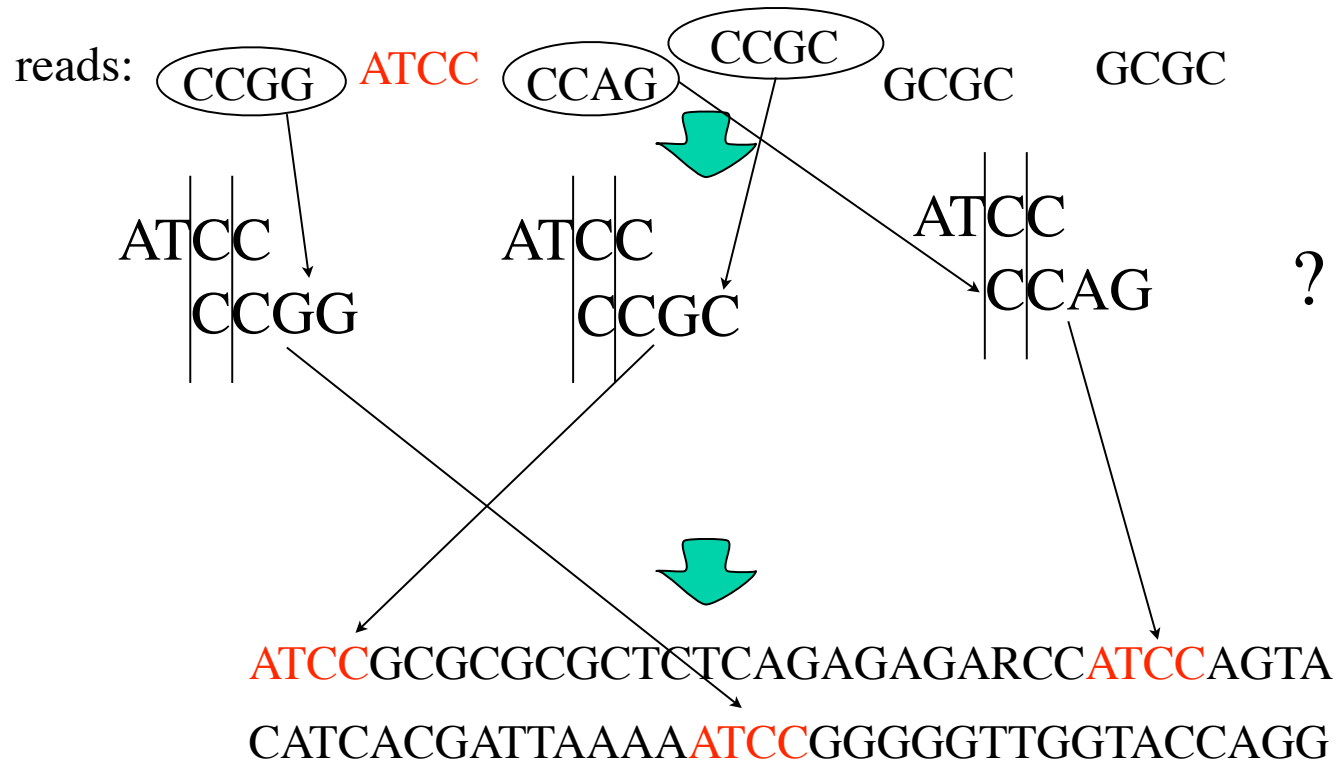
aligned reads:



assembled
sequence: **ATCCGCGCGC**GCTCTCAGAGAGARCCATCCAGTA
 CATCACGATTA AAAATCCGGGGGTTGGTACCAGG

Sequence reads are assembled by aligning the overlap regions. This is easy if all reads are *unique*. But they are not.

Assembly

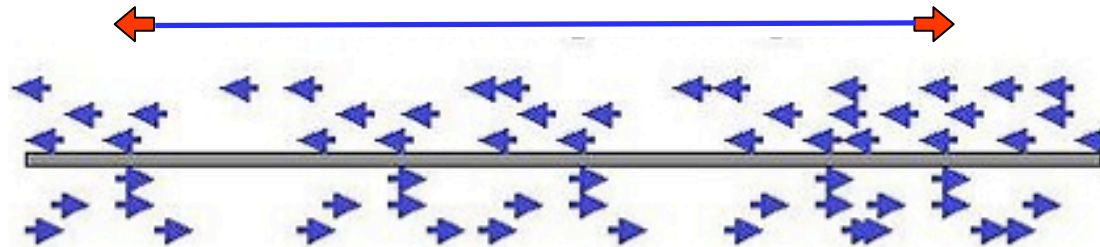
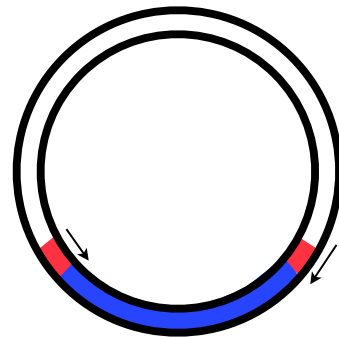


Genomes contain repeats and duplications, making the assembly ambiguous.

“Scaffolding” for disambiguity

- **Large** fragments are cloned into [yeast artificial chromosomes \(YAC\)](#) or [bacterial artificial chromosomes \(BAC\)](#).

- These are grown up, and just the **ends** are sequenced.



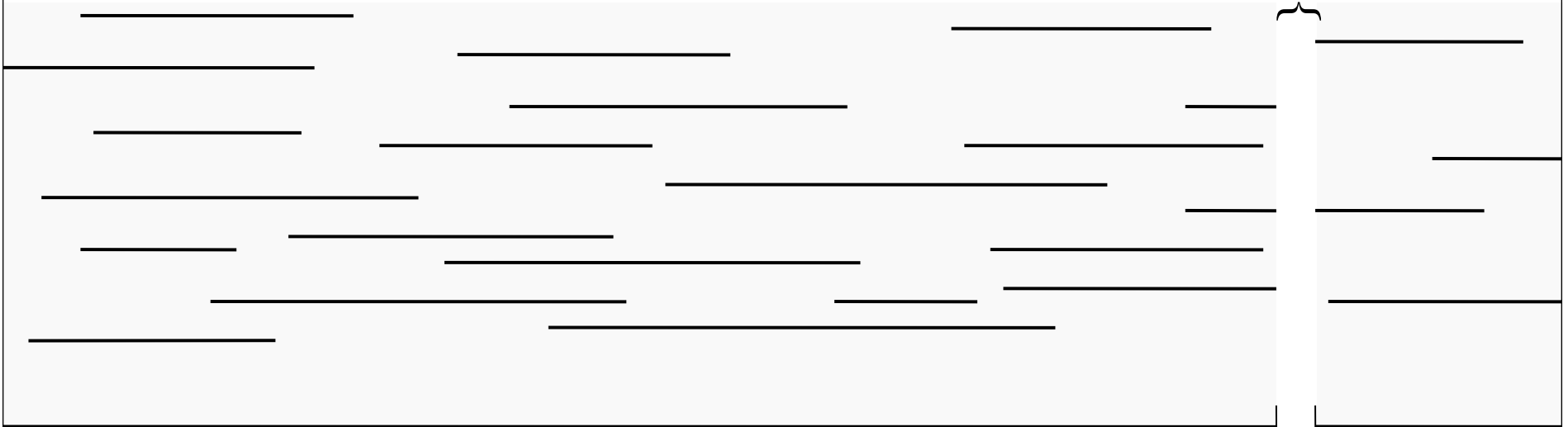
- The size of the insert is known. So the sequence separation of the two reads is known.
- Largest fragment insertable into BAC = 700 kbp, YAC = 3000 kbp.

Contig

BAC insert



no data zone

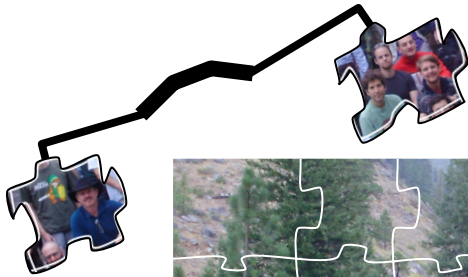


end of one contig

start of next contig

Throughout a "contig" there is a continuous tiling of overlapping fragments with no gaps in the data. Size of "no data" zone is determined using YAC-ends or BAC-ends.

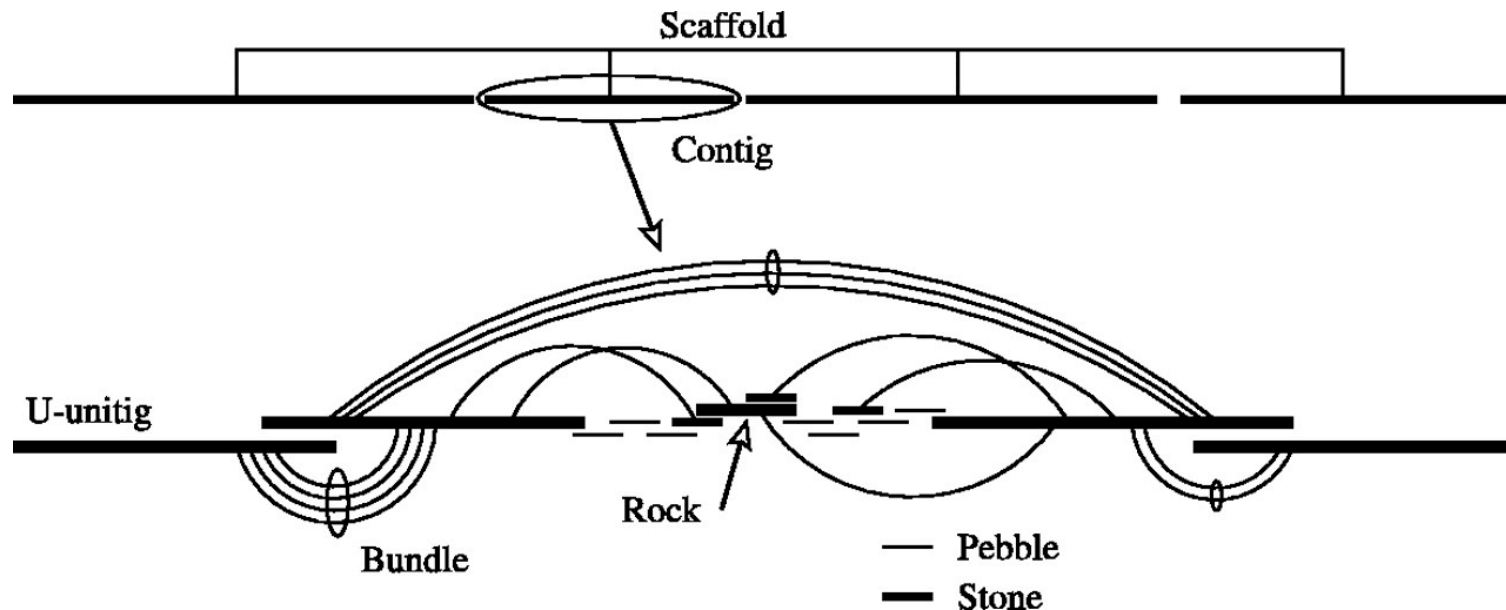
assembly with scaffolds



...is like solving a puzzle with linked pieces.

Assembly algorithm w/scaffolds

First used for the drosophila genome, 2000



Sequence placement order:

1. “Unitigs” = contiguous confidently assembled reads
2. “Scaffold” = 2 or more Unitigs connected by bundles of re-enforcing BAC-ends
3. “Rocks” = unitigs connected by 2 or more BAC-ends
4. “Stones” = unitigs linked by one BAC-end to a Scaffold.
5. “Pebbles” = un-linked Unitigs.

Myers *et al.* **Science** 24 March 2000:
Vol. 287. no. 5461, pp. 2196 - 2204

Warehouses of sequence data

NCBI Washington,DC

www.ncbi.nlm.nih.gov

EMBL Heidelberg, Germany

www.embl-heidelberg.de

DDBJ Shizuoka-ken, Japan

www.ddbj.nig.ac.jp

Members of **International Nucleotide Sequence Database Collaboration**

Assembly viewer in UGENE

Download <http://www.bioinfo.rpi.edu/bystrc/courses/biol4540/ugene/chrM.sorted.bam.ugenedb>

On course web page, click “UGENE files”, Click `chrM.sorted.bam.ugenedb`

Open in UGENE.

Follow along.

- Locate high confidence, low-confidence regions.
- Identify possible polymorphisms.
- Find beginning and end of a *contig*.

Flat files are machine readable

Properties that aid parsing of "machine readable" files...

- generally keyworded
- space delimited fields
- contain special characters like /, :,=,{}, etc (/product)
- contain database identifiers, accession number (gi:123456789)
- sometimes have a checksum, to guard against corruption.
- Not easily human readable...

Exact pattern matching

- DNA
 - Identity matching, uses only A, T, C, G
 - Degenerate base matching, uses IUPAC codes
- Protein
 - Identity matching, 20 aa's
 - Prosite pattern matching. May be variable in length.

Exact matching algorithms provide a yes/no answer, no scoring.

Useful reference tables

The Genetic Code

	U	C	A	G	
U	UUU Phenylalanine UUC Phenylalanine UUG Leucine UUA Leucine	UCU Serine UCC Serine UCA Serine UCG Serine	UAU Tyrosine UAC Tyrosine UAA Stop UAG Stop	UGU Cysteine UGC Cysteine UGA Stop UGG Tryptophan	U C A G
C	CUU Leucine CUC Leucine CUA Leucine CUG Leucine	CCU Proline CCC Proline CCA Proline CCG Proline	CAU Histidine CAC Histidine CAA Glutamine CAG Glutamine	CGU Arginine CGC Arginine CGA Arginine CGG Arginine	U C A G
A	AUU Isoleucine AUC Isoleucine AUA Isoleucine AUG Methionine	ACU Threonine ACC Threonine ACA Threonine ACG Threonine	AAU Asparagine AAC Asparagine AAA Lysine AAG Lysine	AGU Serine AGC Serine AGA Arginine AGG Arginine	U C A G
G	GUU Valine GUC Valine GUA Valine GUG Valine	GCU Alanine GCC Alanine GCA Alanine GCG Alanine	GAU Aspartic acid GAC Aspartic acid GAA Glutamic acid GAG Glutamic acid	GGU Glycine GGC Glycine GGA Glycine GGG Glycine	U C A G

IUPAC nucleotide codes

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

Exercise: Can you write the IUPAC expression for the set of all STOP codons

Pattern matching in DNA: Write a single IUPAC expression for

...the set of all STOP codons



...the set of all ARG codons



IUPAC nucleotide codes

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	puRine A or G
Y	pYrimidine C or T or U
S	strong G or C
W	weak A or T or U
K	Keto G or T or U
M	aMino A or C
B	not A C or G or T or U
D	not C A or G or T or U
H	not G A or C or T or U
V	not T,U A or C or G
N	aNything any base
. or -	gap

Pattern matching in proteins:

Functional motifs -- ProSite

ProSite motifs are created by using experimental data, then extending it using sequence data.

Example: A conserved histidine is required for function.

*

ALRDFATHDDF

SMTAEATHDSI

ECDQAATHEAS

Based on the homolog sequences, starting with the His, a pattern of conservation is found.

If it is too specific, the pattern is selective but not sensitive.

If it is too vague, the pattern is not selective.

Motifs exist due to selective pressure

Selective pressure on proteins for:

folding -- some proteins must be stable

others are turned over

function --

active site residues

binding to other proteins

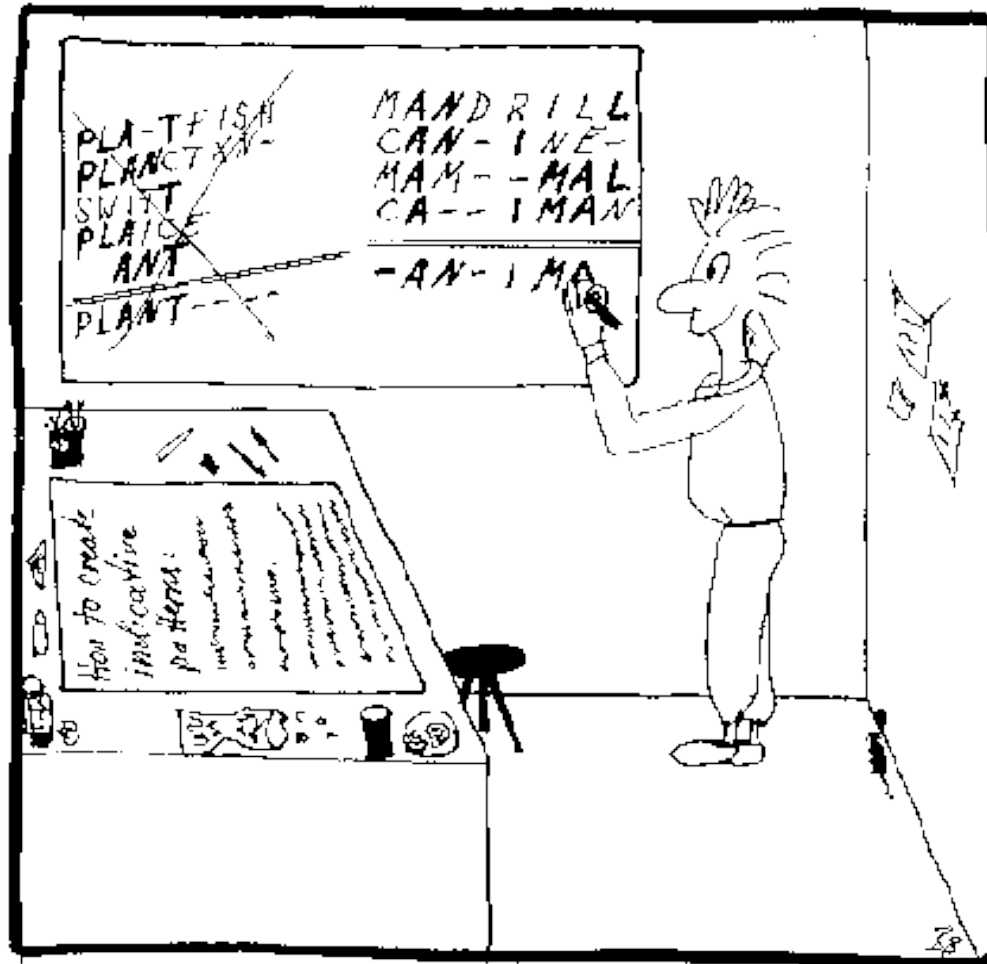
as a substrate for --

signal sequences, intra-cellular transport, export

post-translational modification,...

a joke

How we develop Prosite patterns!



Brigitte Boeckmann / 1995

Syntax for motif patterns

- $x(n)$ Any amino acid. If n is specified, then n amino acids.
 n may be a range or a list.
- X Amino acid X , only.
- $[XY]$ **Either** X or Y .
- $\{XY\}$ **NOT** X, Y . Anything but X or Y .

Example:

$C-[AHY]-x(2,4)-G-\{DERKH\}-[GN]$

matches the sequences:

CAFIN TGIN

CHQ--SGFN

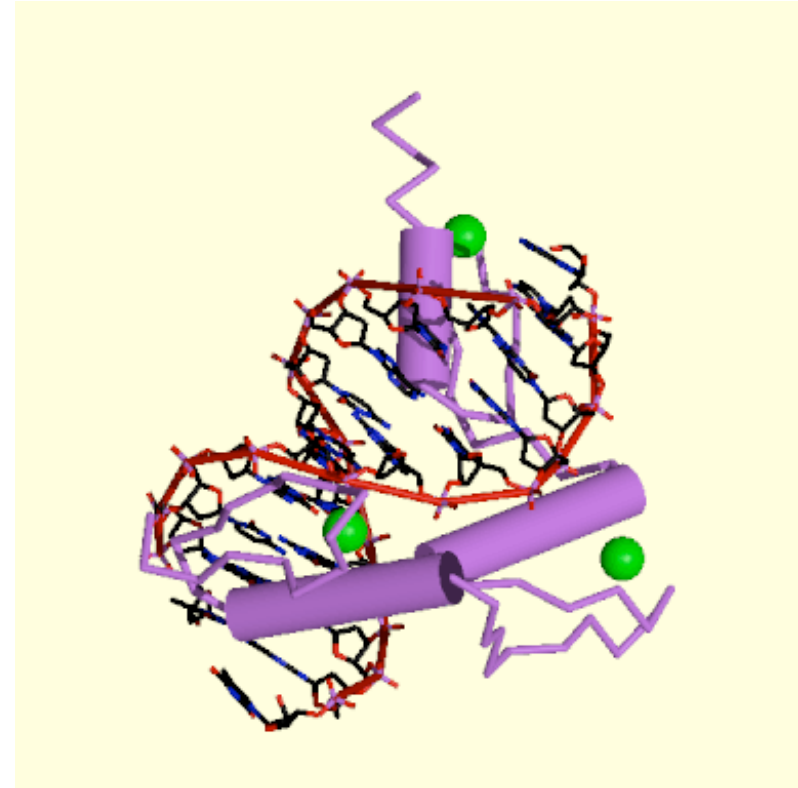
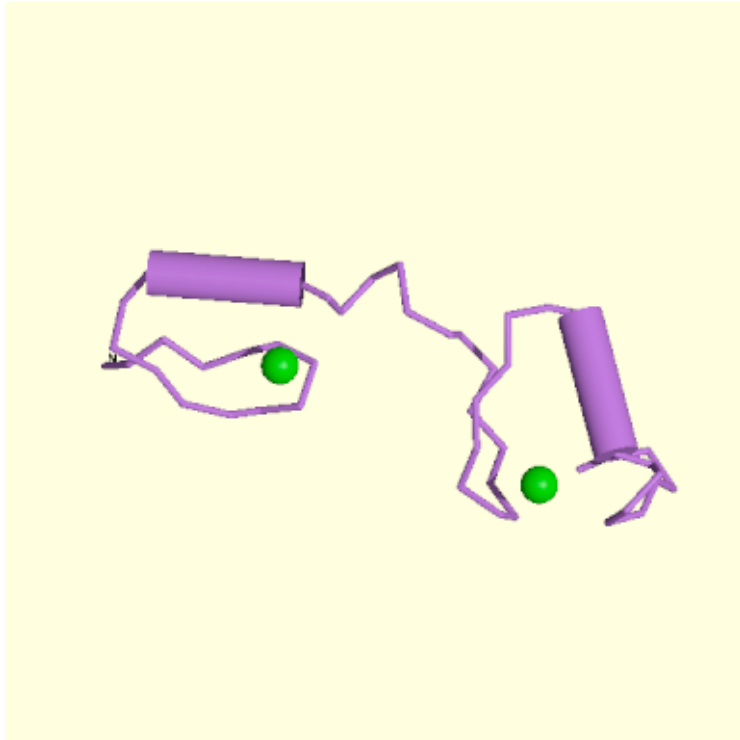
CY--MLGMG

CAHDNAGTN

Can you find it?

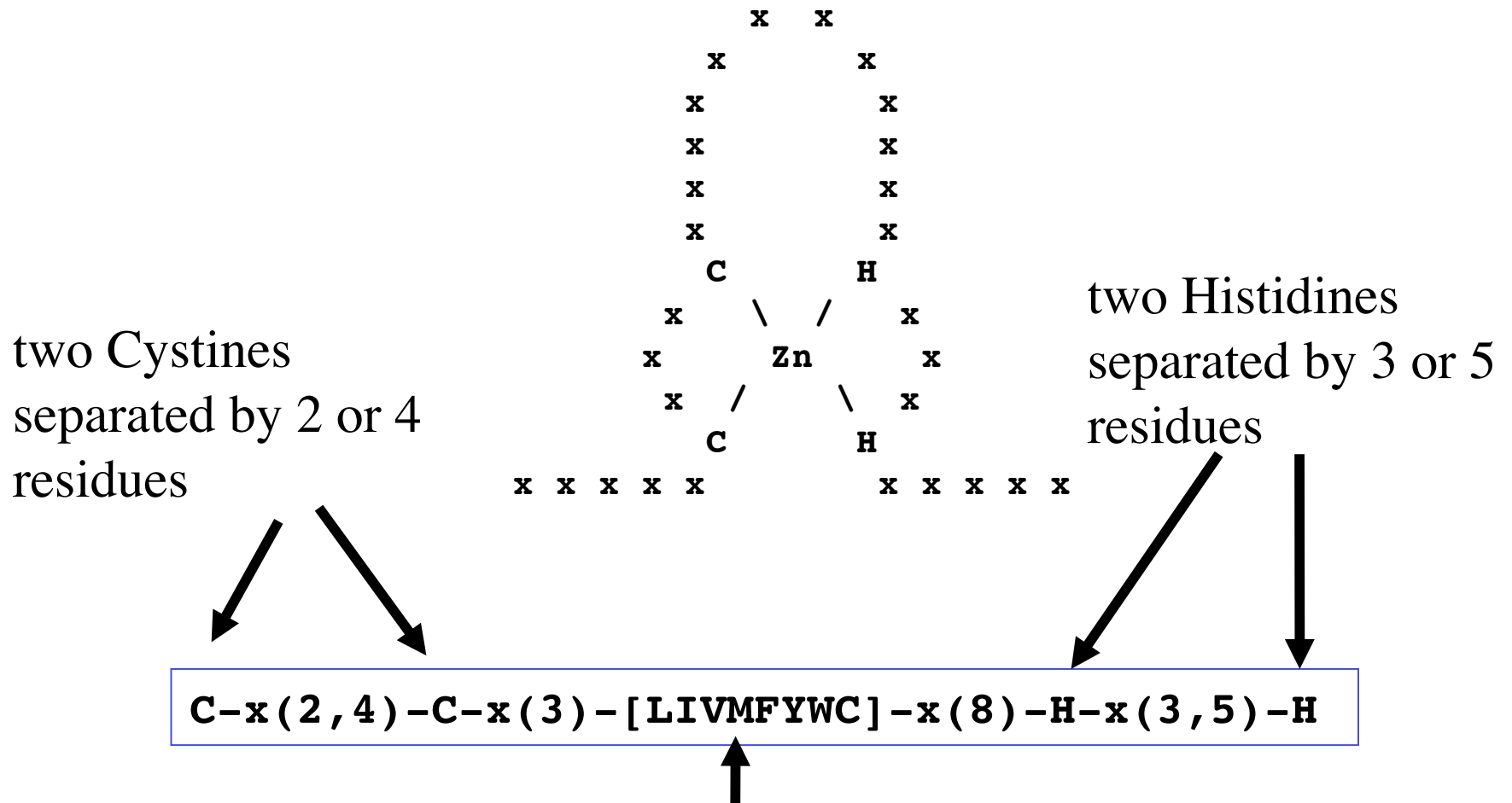
CAAAAWGYGAHCGQTKGENCYHAGDGCYCYGLNPKGL

Zn finger structure



The helix side of the finger makes H-bonds to the nucleotides. So that side is highly variable.

Zinc finger motif



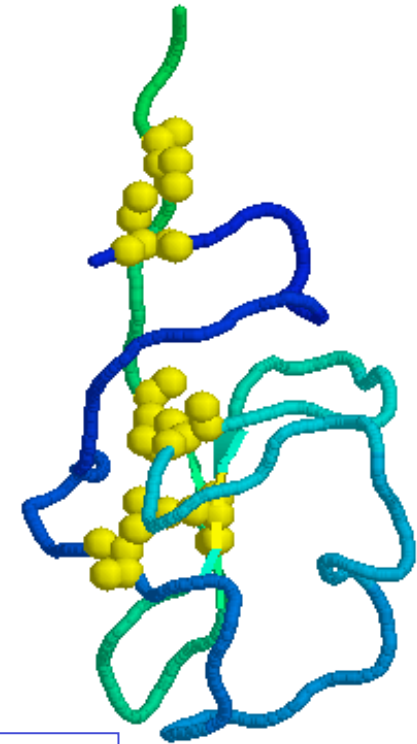
Loop must be length 12.

4th position in loop must be hydrophobic

Kringle domain



a triple loop, 3-disulphide bridge structure, whose conformation is defined by a number of hydrogen bonds and small pieces of anti-parallel -sheet.



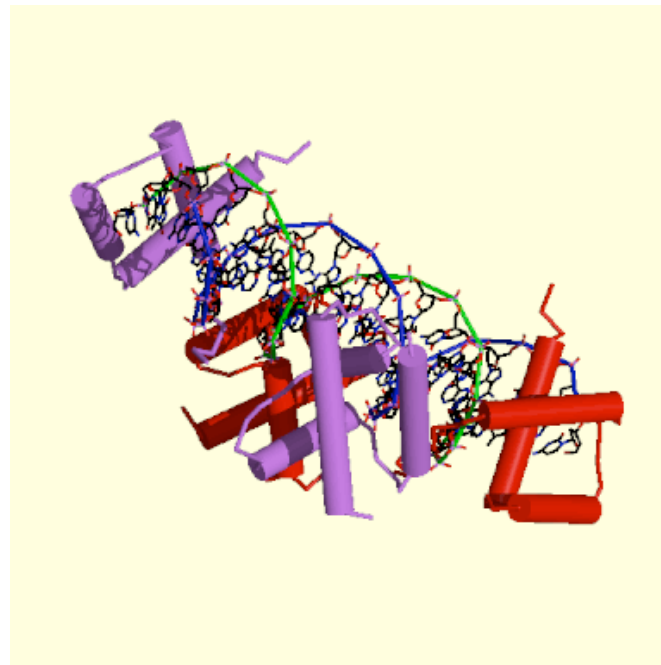
[FY] - C - [RH] - [NS] - x(7, 8) - [WY] - C
The two C's are involved in a disulfide bonds.

Homeobox

Found in transcription factors.

L-M-A-[EQ]-G-L-Y-N

Helix-turn-helix protein.
C-terminal helix
interacts with DNA, and
contains the signature.



ER targeting sequence

[KRHQSA] – [DENQ] – E – L

Proteins that permanently reside in the lumen of the endoplasmic reticulum (ER) have the C-terminal sequence Lys-Asp-Glu-Leu (KDEL). While KDEL is the preferred signal in many species, variants of that signal are used by different species.

Signal	Species
--------	---------

KDEL	Vertebrates, Drosophila, Caenorhabditis elegans, plants
HDEL	Saccharomyces cerevisiae, Kluyveromyces lactis, plants
DDEL	Kluyveromyces lactis
ADEL	Schizosaccharomyces pombe (fission yeast)
SDEL	Plasmodium falciparum

PTMs

N-glycosylation

$N-\{P\}-[ST]-\{P\}$

Tyrosine phosphorylation

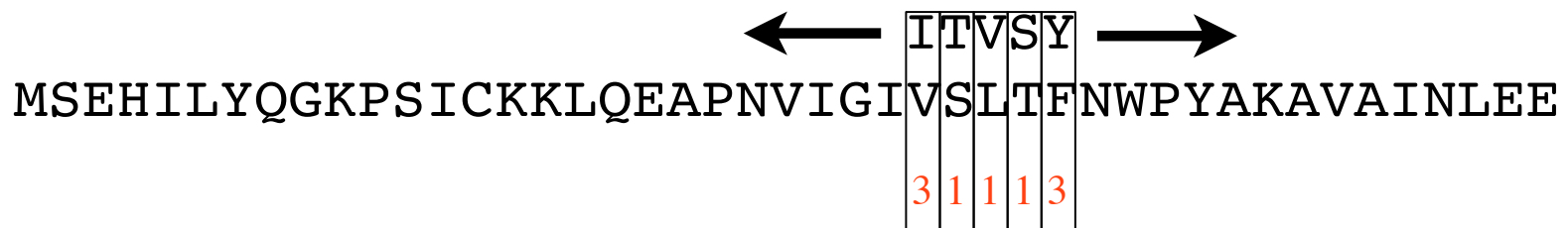
$[RK]-x(2)-[DE]-x(3)-Y$ or $[RK]-x(3)-[DE]-x(2)-Y$

C-terminal prenylation

$C-\{DENQ\}-[LIVM]-x$

Inexact pattern matching

- Exact matching is black/white.
- Most applications use inexact matching.
 - Requires a *mismatch score*.



Amino acid substitution matrices for inexact pattern matching

Two 20x20 substitution matrices are used: BLOSUM & PAM.

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

Each number is the score for aligning a single pair of amino acids.

What is the score for this alignment?:

ITVSY
VSLTF

BLOSUM62 substitution matrix

HW1, due Sep 10

- Find motifs in DNA and Protein, using IUPAC and Prosite notation, respectively.
- *Write a program to search for motifs.*
- **details in HW1 pdf file.**

In-class UGENE exercise: DNA dotplot

- Select NCBI-->Nucleotide
 - Search “Nucleotide” for “influenza A virus H1N1 Puerto Rico mRNA”
 - Select the first one.
 - Get the accession number.
- In UGENE, use File/Open remote database.
 - Paste the accession number.
 - Right-click the sequence window, a menu opens. Select analyze, select dotplot. Do a “self” dotplot.
 - Right-click on dotplot image to change %identical and length.
 - Find the locations of the longest repeat sequence

Review

- How does Sanger sequencing work?
- How does pyrosequencing work?
- What kind of sequence would cause errors in the pyrosequencing method?
- What is shotgun sequencing?
- What is sequence assembly?
- What kind of sequences cause errors in assembly?
- Do you know the 1-letter codes of the amino acids?
- Do you know the IUPAC nucleotide codes?
-