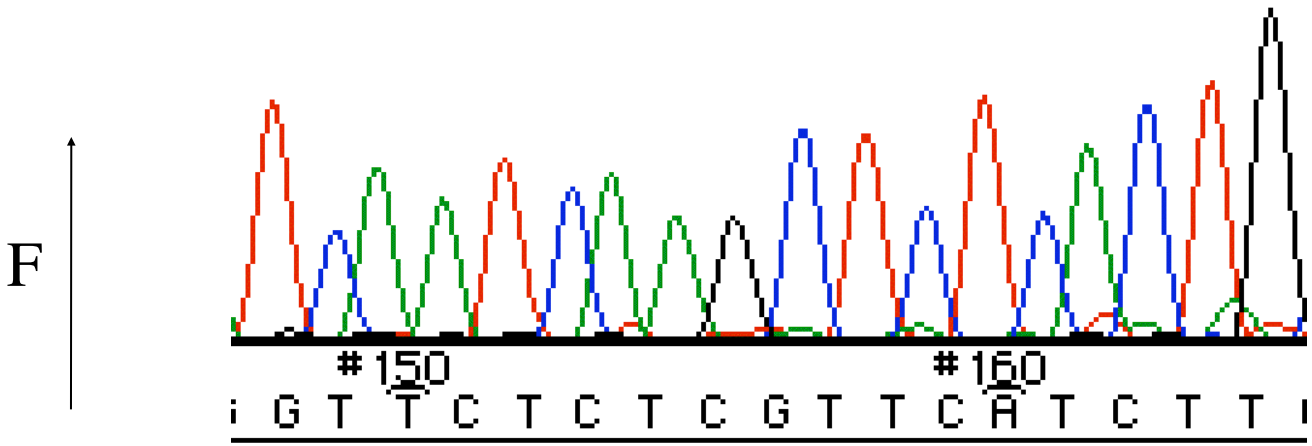
# Bioinformatics 1-- Lecture 2

- Follow-up of lecture 1

# Experimental origins of sequence data

The Sanger dideoxynucleotide method




Each color is one lane of an electrophoresis gel.

# base calling

- In Geneious: select Sample Documents --> Contig Assembly
- Follow along.

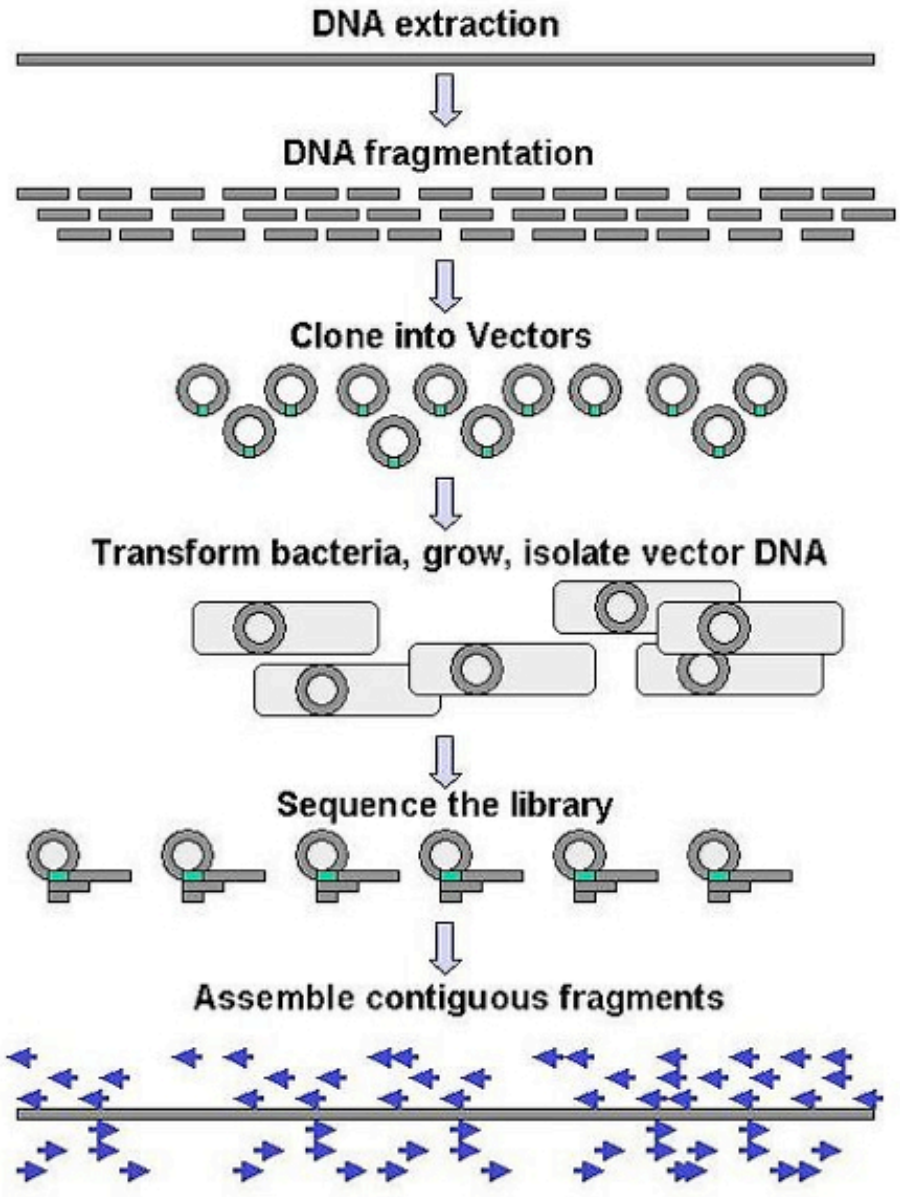
  *Is Geneious intuitive enough*?

# New technology: Pyrosequencing

- http://www.youtube.com/watch?v=nFfgWGFe0aA&NR=1
- ..or search youtube for “pyrosequencing”

- *Whole genome sequencing in < 1 day!!*

# 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

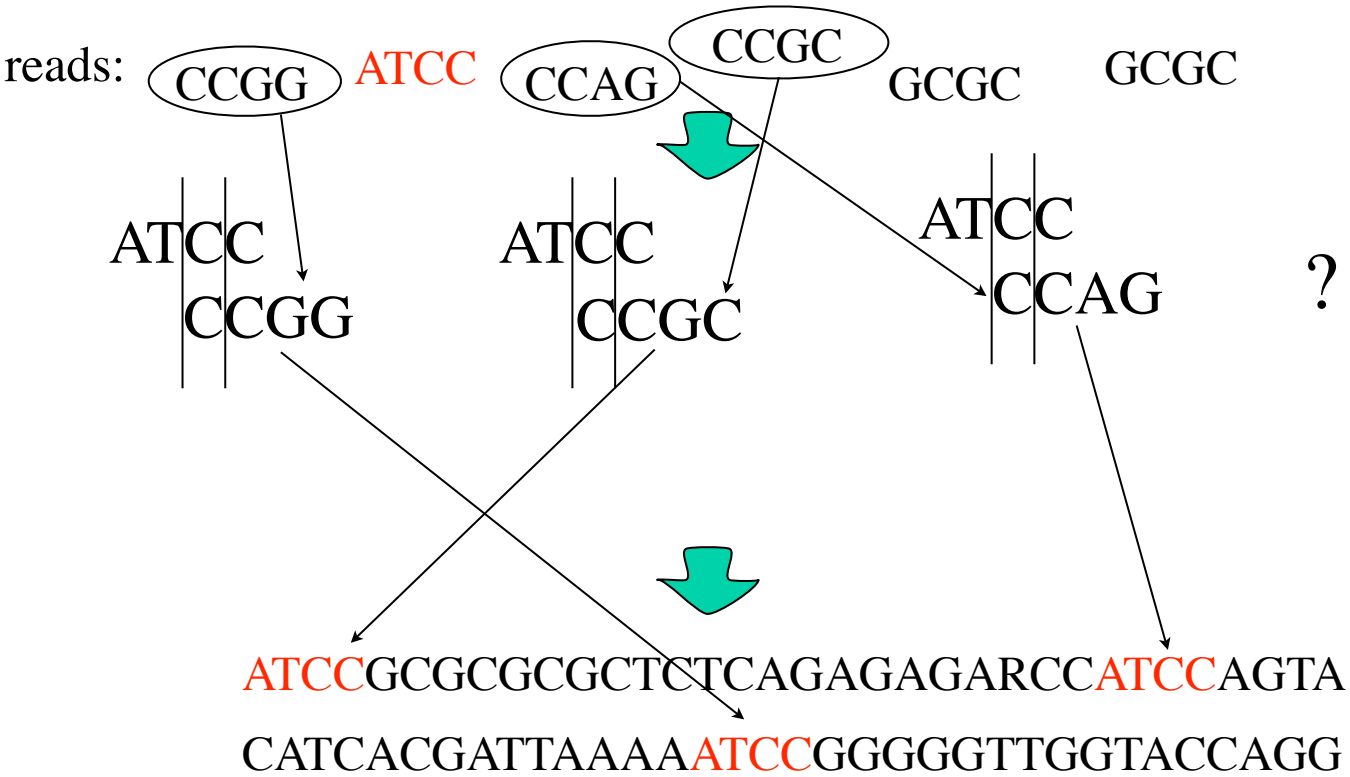
aligned reads:

ATCC
CCGC
GCGC
GCGC...

assembled sequence: ATCCGCGCGCGCTCTCAGAGAGARCCATCCAGTA
CATCACGATTAAAAATCCGGGGGTTGGTACCAGG

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

# Assembly

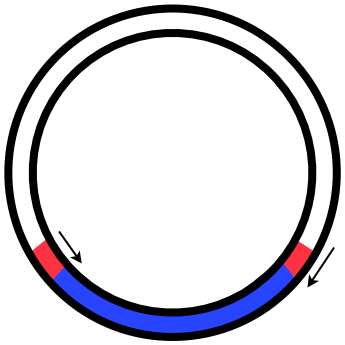


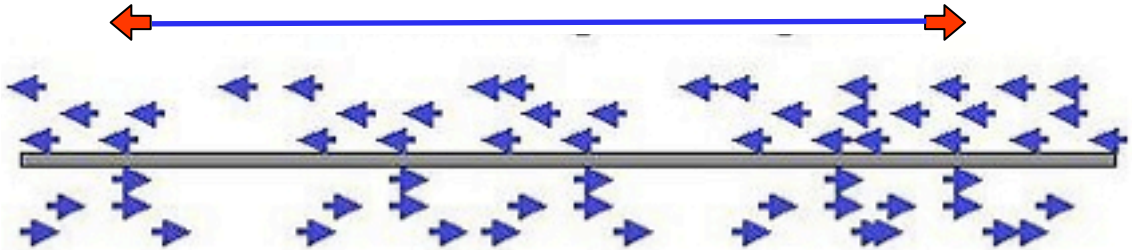

Genomes contain repeats and duplications, making the assembly ambiguous.

# “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.

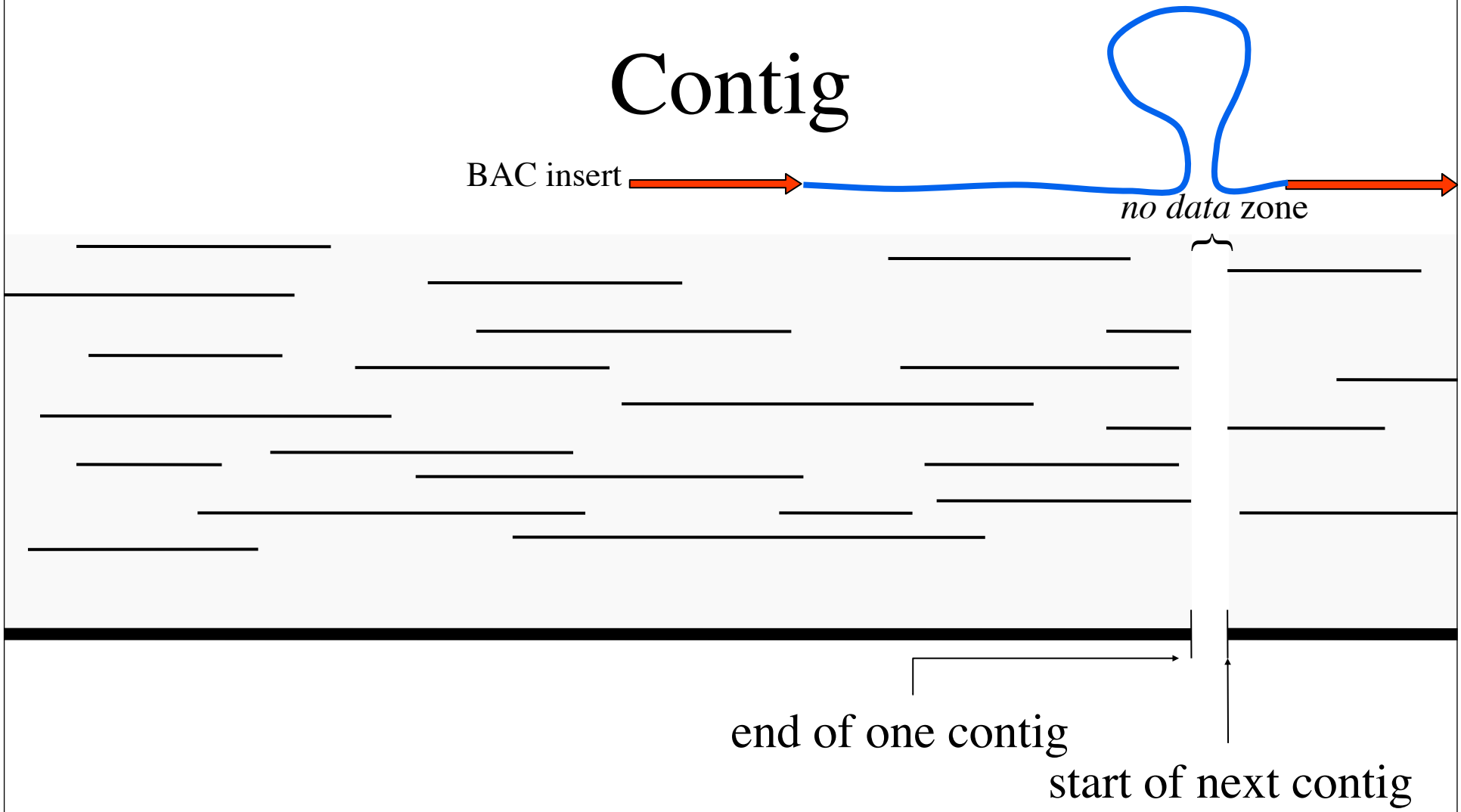




•The size of the insert is known. So the sequence separation of the two reads is known.

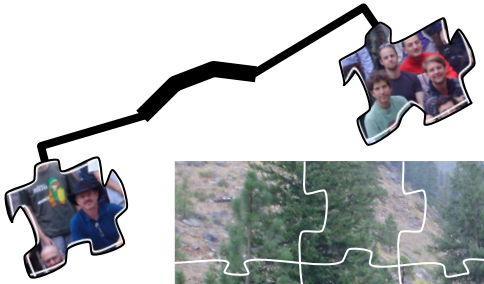
•Largest fragment insertable into BAC = 700 kbp, YAC = 3000 kbp.

# Contig




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.

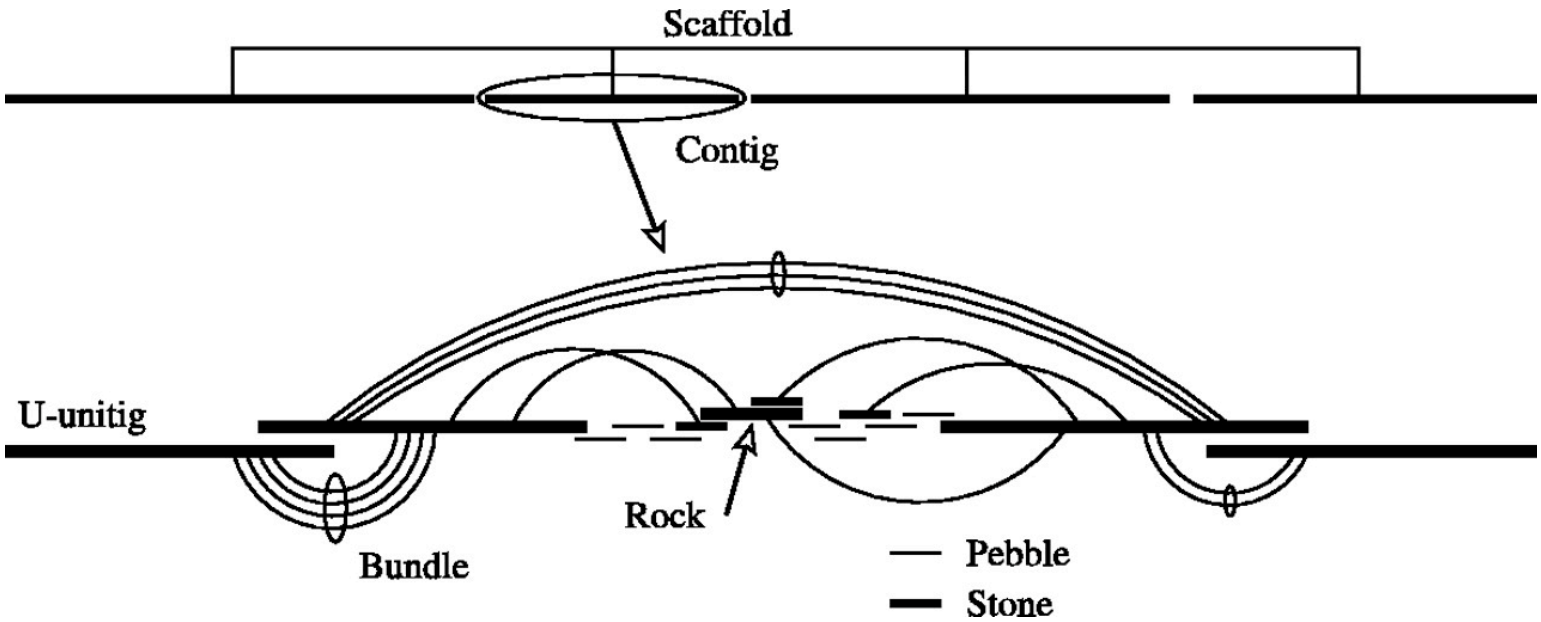
# assembly with scaffolds




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

# Assembly algorithm w/scaffolds

First used for the drosophila genome, 2000




<u>Sequence placement order:</u>

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**

# Follow along: short NCBI tour

Set the browser to

**www.ncbi.nlm.nih.gov**

And follow along….

# Flat files

Z&B p. 57

# Properties 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 very human readable...

Generally it is better to let the machine do the reading!

# In-class exercise: Finding patterns in DNA

- In Geneious: make a new folder called “inclass” under “Local”
- Select NCBI-->Nucleotide
  - Search for “influenza A virus H1N1 Puerto Rico mRNA”
  - Select the first one and drag it to “inclass”
  - Select “inclass”. Select the seq. Download if necessary.
  - Sequence-->Find motif...
    - enter pattern: WWWWW. Name it “AT-rich”
  - Sequence-->Find motif...
    - enter pattern: SSSSS. Name it “GC-rich”
    - Delete all motif annotations by double-clicking on the legend.
- Translate the sequence in all 6 frames.
  - Which is the correct frame based on the absence of STOP codons?

# Useful reference tables

## The Genetic Code

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

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

# Motifs exist due to selective pressure

<u>Selective pressure on proteins for:</u>

**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,...

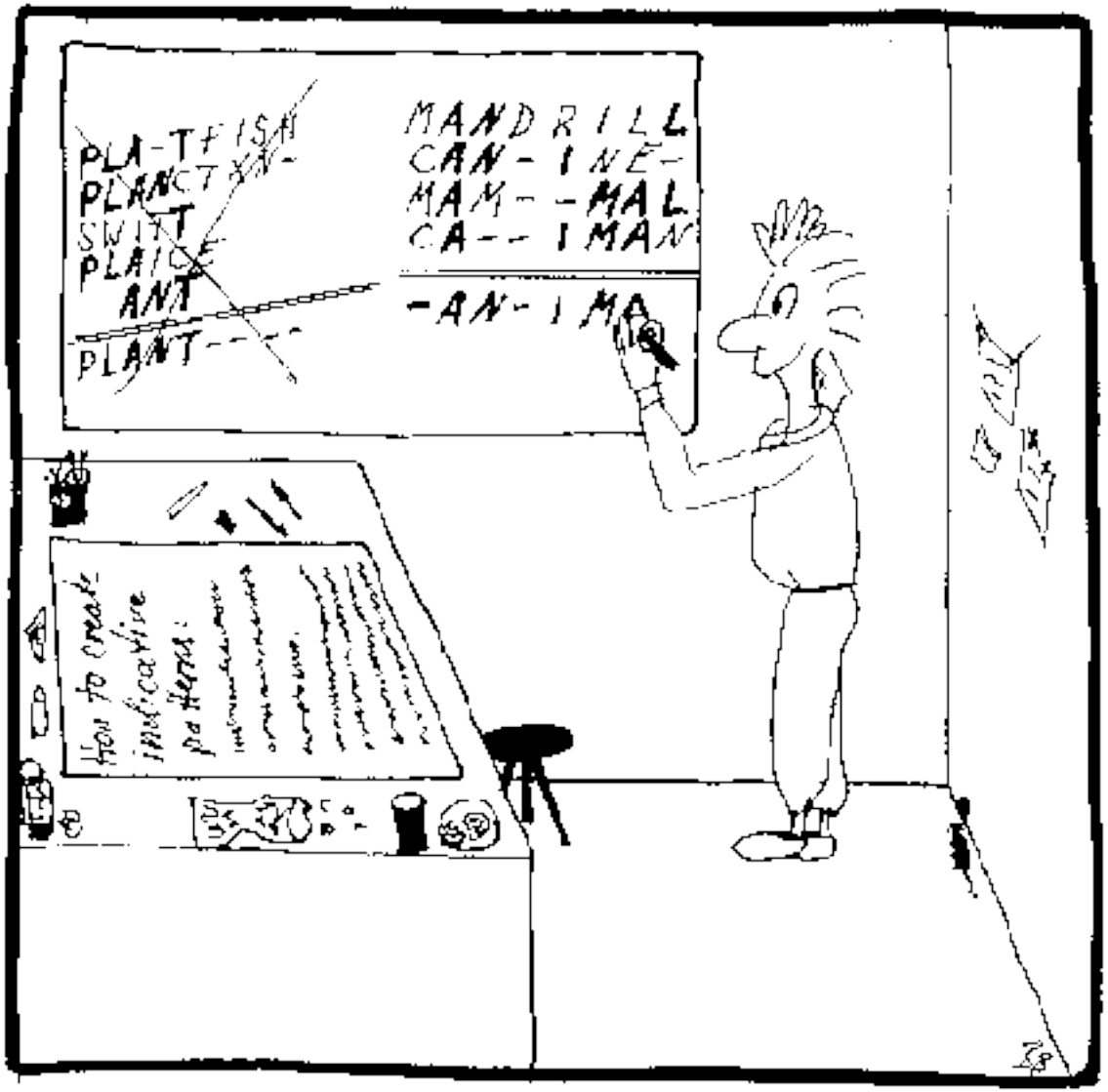
# Functional motifs -- ProSite

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.

# 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]`

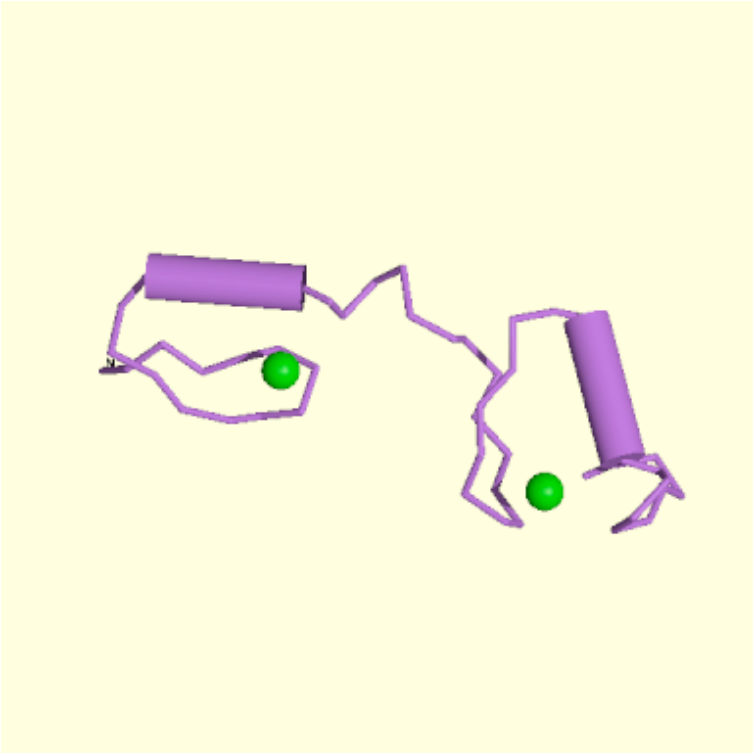
matches the sequences:

```
CAFINTGIN
CHQ--SGFN
CY--MLGMG
CAHDNAGTN
```

*Can you find it?*

`CAAAAAWGYGAHCGQTKGENCYHAGDGCYCYGLNPKGL`

# Zn finger structure



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

# Zinc finger motif

```
        x  x
       x      x
      x        x
      x        x
      x        x
      x        x
       C      H
     x   \  /   x
    x     Zn     x
     x  /    \  x
       C      H
x x x x x        x x x x x
```

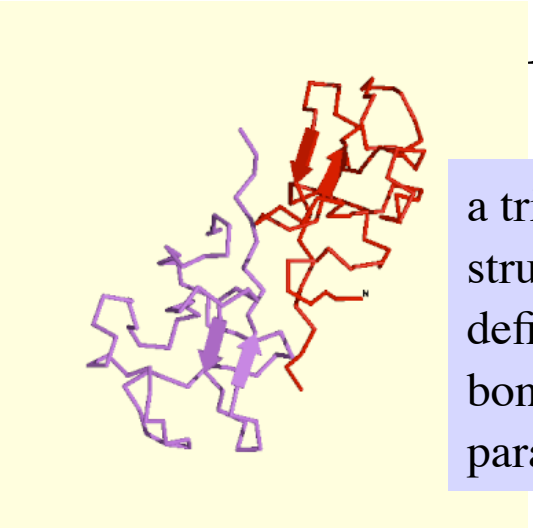
two Cystines separated by 2 or 4 residues

two Histidines separated by 3 or 5 residues

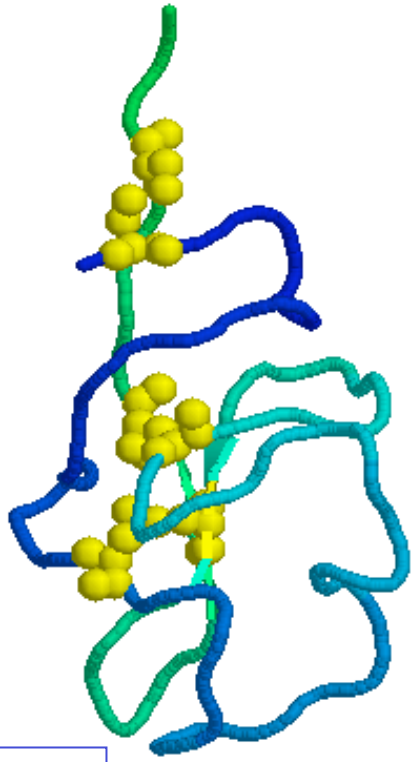
**C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H**

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`**

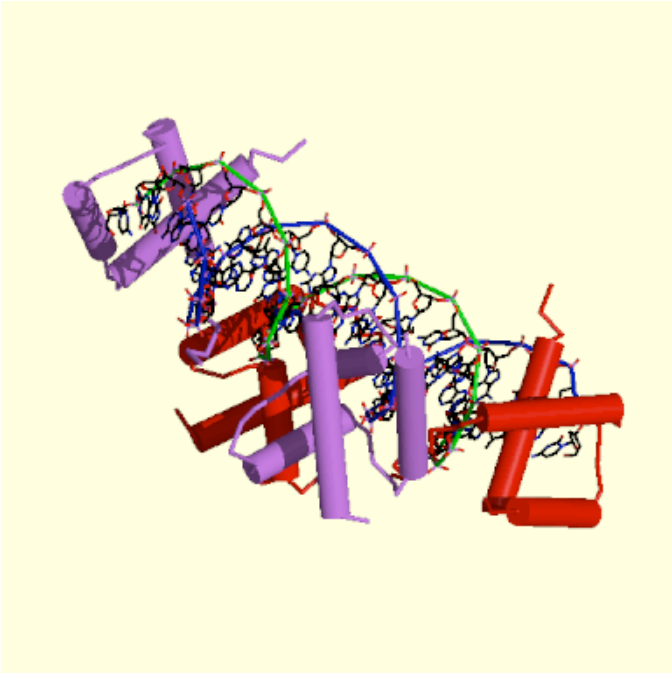
**The two C's are involved in a disulfide bonds.**

# 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 |

## 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.

```
                          ← ITVSY →
MSEHILYQGKPSICKKLQEAPNVIGIVSLTFNWPYAKAVAINLEE
```

# Amino acid substitution matrices

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

# Review, 2 Sep 2009

- Sequencing
  - Sanger sequencing, base calling
  - pyrosequencing
  - whole genome shotgun assembly
- Sequence warehouses
- Expressing DNA/protein patterns
- Inexact matching

# HW1, due Sep 9

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