Molecules provide an amazing view of evolution ...

Molecules usually confirm what has been found from careful analysis of morphology. Just as you can classify similar organisms based on similar morphologies, you can also classify them based on similar molecular details.

That is, we can determine the sequence of the molecules, examine how they have changed and determine a genetic distance between the organisms including potential time frames.

Some molecules evolve and change rapidly

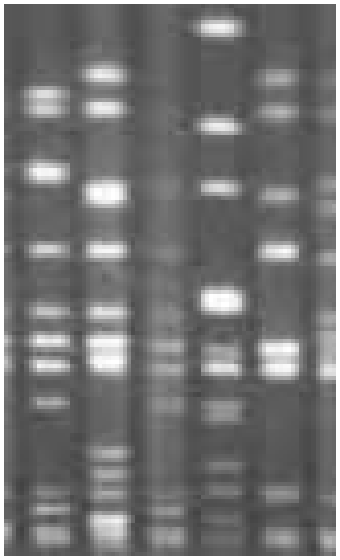
TCCTGGCATGC ACACACACACACACACACACACACAC TGCTAATA

Sequences such as microsatellites change very rapidly (usually by the deletion or insertion of another AC pair).

By examining several of these microsatellites it is possible to use the DNA sequence to distinguish between individuals, even between parent and offspring. This ability is of interest to ...

CODIS

Sir Alec Jeffreys



minisatellites

CODIS - use repetitive DNA

Microsatellite DNA

Unit - 2-4 bp (most 2).

Repeat - on the order of 10-100 times.

Location - Generally euchromatic.

Examples - Most useful marker for population level studies.
This example is from a water snake ...

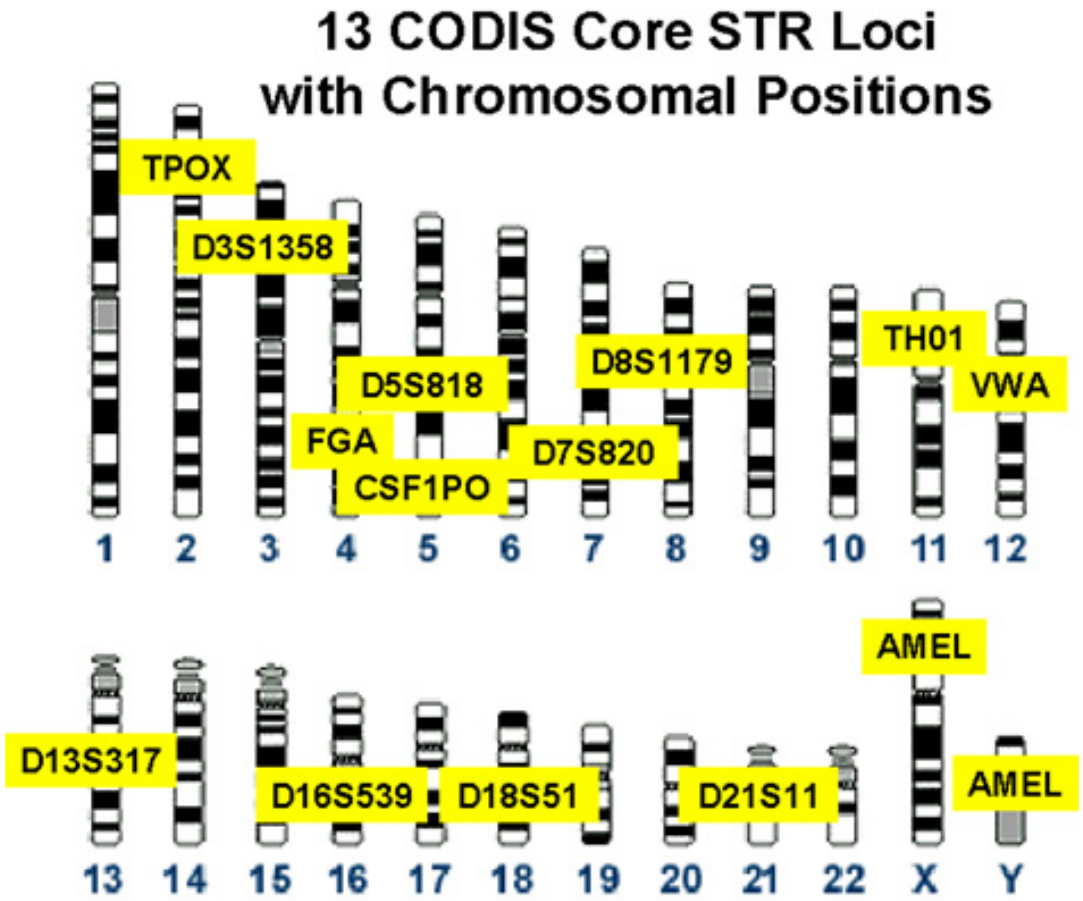
```
...TCCAGACAAGGTGGTGTGTGTGTGTGTGTG
   TGTGTGTGTGTGTTTCTCCAGTGAGATTTA...
```

CODIS - the database

- In 1990 the FBI established a pilot project.
- In 1994 CODIS (Combined DNA Index System) was established.
- In 1998 it was fully operational.
- By 2007 it contained more than 4.5×10^6 records.
- In May 2013 it had
  - 10,376,000 offender profiles,
  - 1,515,800 arrestee profiles and
  - 493,500 forensic profiles
- Records contain specimen numbers and DNA profiles but nothing else.

The Quantico FBI Forensics Laboratory
From NY Times May 12 2009.

CODIS - the markers



From wikipedia (28/11/08).

DNA databases go too far

Under a 2005 US federal law, the FBI database will continue to include convicted felons, but it will also add genetic profiles of people who have been arrested but not convicted and of immigrant detainees for an estimated 1.3 million more profiles by 2012. The F.B.I. also accepts the DNA of missing persons.

As of 2007, the US Justice Department estimated the sample processing backlog at 600,000 to 700,000 samples.

In 2002, the F.B.I. was processing about 5,000 DNA samples each year. With the help of new robotic systems, analysts with the crime lab plan to process 90,000 samples each month by 2010.

From the NY Times, May 12 2009, page D3

Other molecules change at more moderate levels

Back to our question ...

For our question sequences are needed that change more slowly. In fact, too slowly to distinguish parent/offspring or even individuals of different races. But they are useful to distinguish different species.

In contrast, microsatellites between human and chimpanzee have changed so much in the intervening millions of years that all trace of similarity between the two species would have been lost.

We need to be able to measure genetic distance

Utility

- Estimation of the amount of variation between two sequences
- Used in alignments
- Used in phylogenetic reconstruction
- Used in population genetic history reconstruction
- Used in estimates of functionality

What is meant by a distance

A distance metric should satisfy ...

- Nonnegativity – $d_{AB} \geq 0$
- Uniqueness – $d_{AB} = 0 \quad \text{implies} \quad A = B$
- Symmetry – $d_{AB} = d_{BA}$
- Triangle inequality – $d_{AC} \leq d_{AB} + d_{AC}$

Types of distances

- **Hamming distance** – minimum number of substitutions
- **Levenshtein distance** – minimum number of insertions, deletions
- **Edit distance** – minimum number of substitutions, insertions, deletions
- **Weighted distance** – weight each event differently
- ?? others ?? – interchanges of characters ??

Types of distances - examples

- Hamming $d_{CGACG,\,GTCGA} = 5$

  Each position needs a substitution of one character for another character.

- Levenshtein $d_{CGACG,\,GTCGA} = 4$

	C	G	A	C	G	
1		G	A	C	G	
2		G	A	C	G	A
3		G		C	G	A
4		G	T	C	G	A

- Edit $d_{CGACG,\,GTCGA} = 3$

  The last two steps in the Levenshtein distance can be done by one substitution.

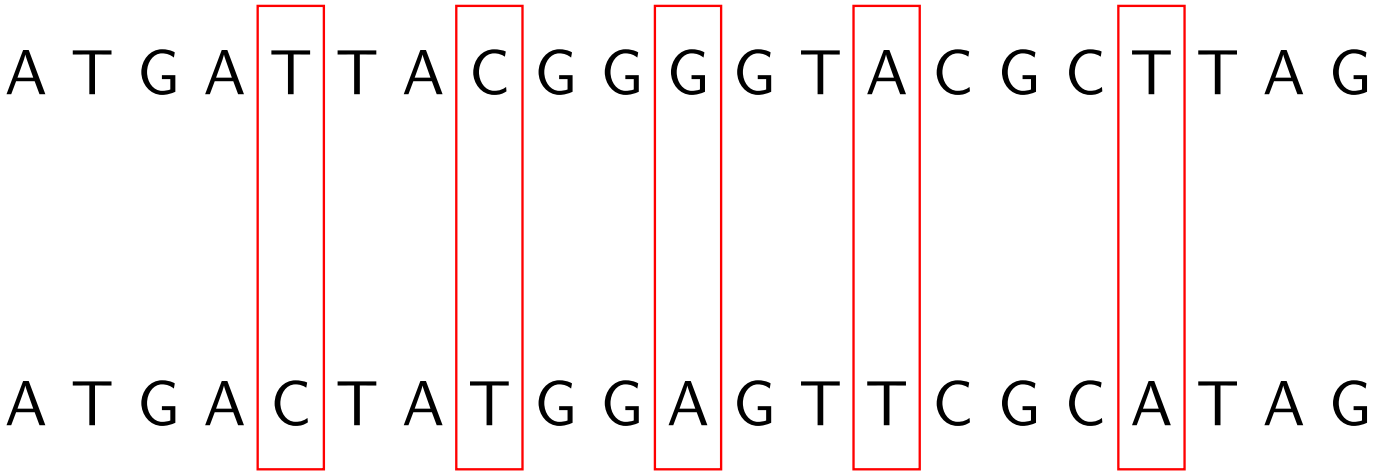
Hamming distance

Most straightforward distance:
number of differences between two sequences weighted by the length of the sequence

$$D = k/n$$

$$Var(D) = D(1 - D)/n$$

where,
k is the number nucleotide differences
n is the number of sites.



$$D = 5/21 = 0.238$$

$$Var(D) = 0.0086$$