News: Allergy free cats go on sale in U.S.(October 2006)

- ALLERCA GD kitten at 12 weeks of age
- Complete and updated vaccinations
- Mandatory spaying or neutering
- Microchip Identifier implant
- Healthy and socialized

http://www.allerca.com

$5950 as of today

Booked for at least 12m months

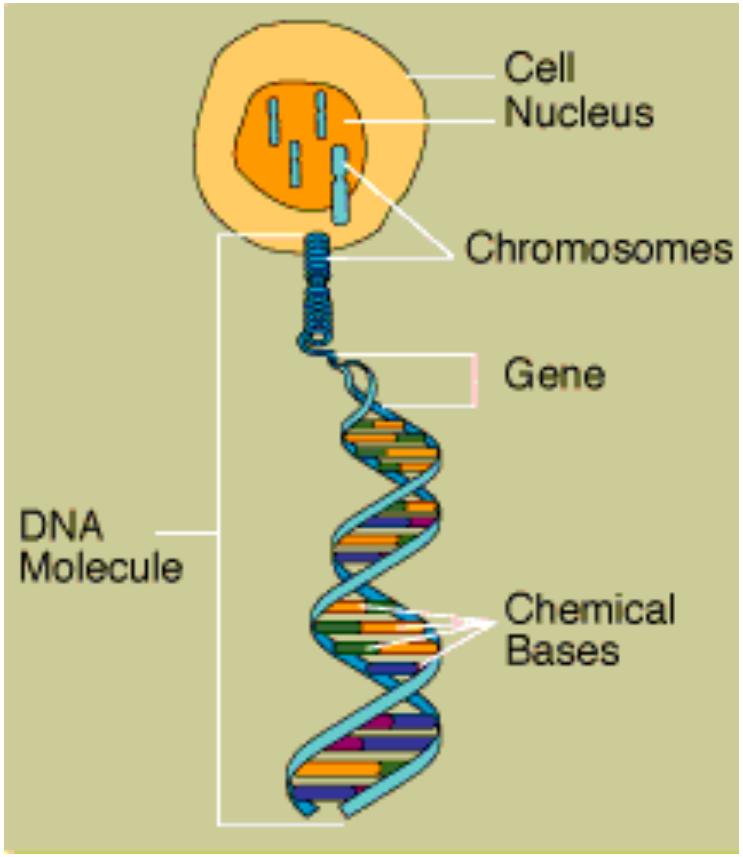
Expedited delivery available

Where is the computing?

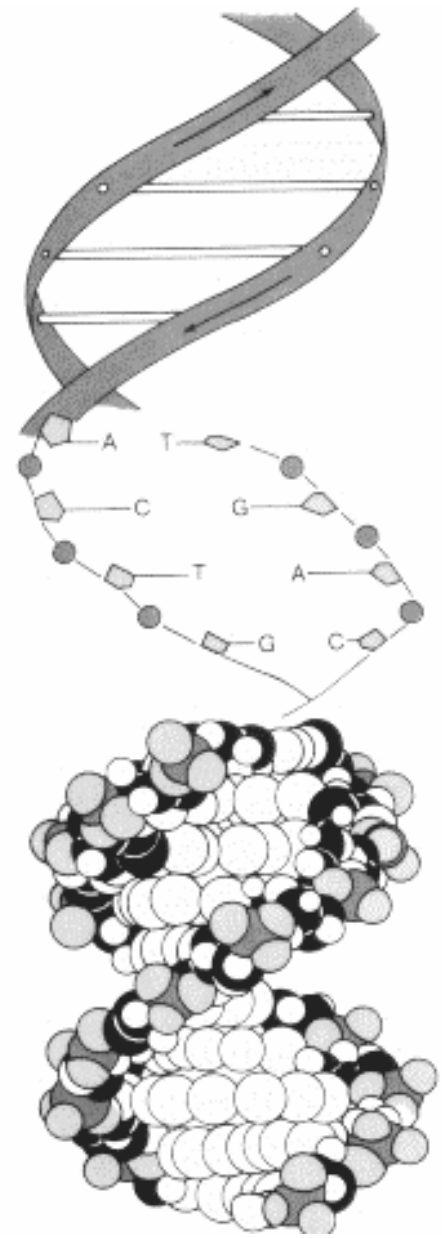
- "sophisticated bioinformatics" was used.
- A glycoprotein, Fel d 1, is the allergan.
- Sequence Fel d 1 gene from multiple cats and study naturally occurring divergences.
  - ➔ comparative genomics, multiple sequence alignment
- Target the divergences that could potentially alter structure of Fel d 1 protein.
  - ➔ sequence-structure relationship, protein threading.
- Carry out selective breeding to create the GD cat.

Background

- *DNA*
  - Computationally, a string over alphabet *{A,C,G,T}*
- *Genome*
  - Collection of all DNA in a cell
- *Gene*
  - Encodes the recipe for producing proteins
- *Protein*
  - A sequence of amino acids



Source: *http://rex.nci.nih.gov/behindthenews/ugt/05ugt/ugt05.htm*


A
T
C
G
T
A
G
C

Some challenges in Computational Biology

1. Compare DNA sequences and proteins sequences for similarity.
2. Study the evolution of sequences and species.
3. Obtain the genome of an organism.
4. Identify and annotate genes.
5. Find the sequences, three dimensional structures, and functions of proteins.
6. Find sequences of proteins that have desired three dimensional structures.

How to Compare Two Sequences?

- **Problem**:
  - Given two sequences s_1 and s_2 over a fixed alphabet Σ, what is the set of variations that best describes the genetic transformation from s_1 to s_2 (or equivalently, from s_2 to s_1)?

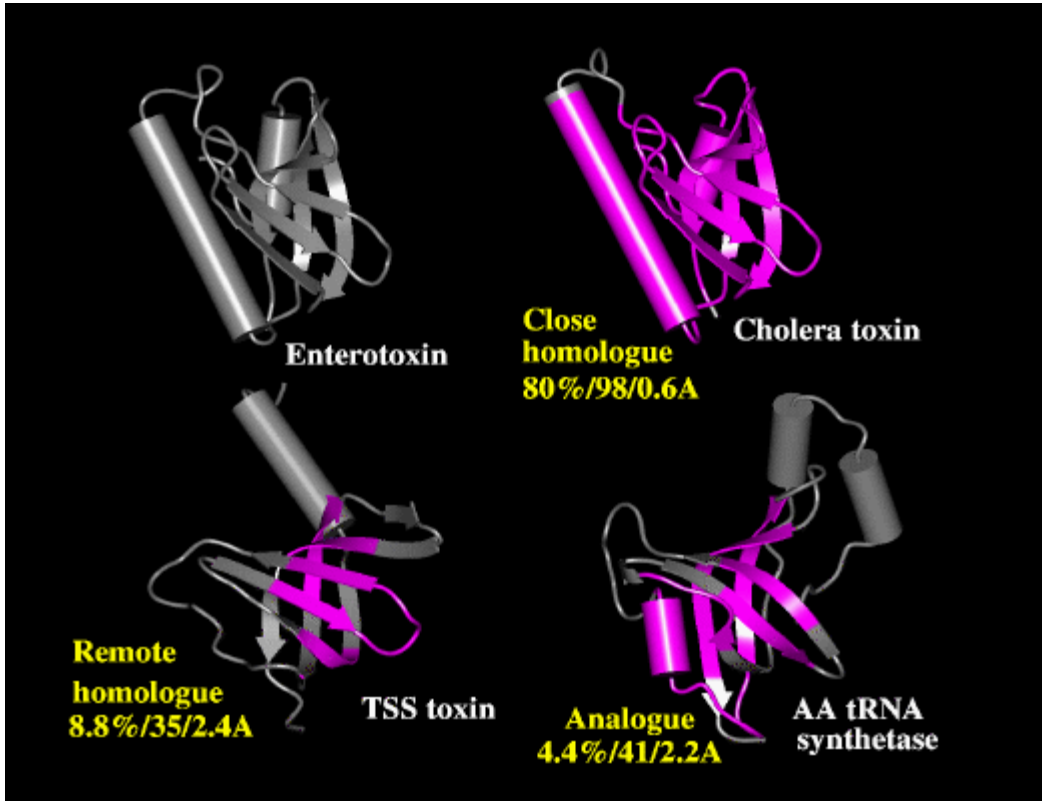
Combinatorial Optimality

- Based on either maximizing an *alignment score* or minimizing *edit distance*
- Standard dynamic programming techniques

Probabilistic Optimality

- Based on finding a most *probable* set of changes in aligning two sequences
- Hidden-Markov Model (HMM) techniques

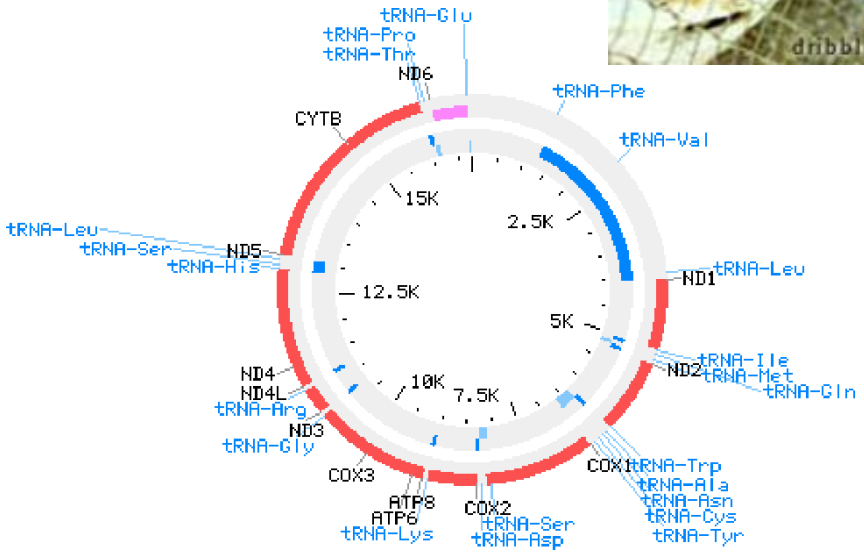
Sequence Comparison Caveats



Magenta regions are structurally equivalent with enterotoxin (top left).

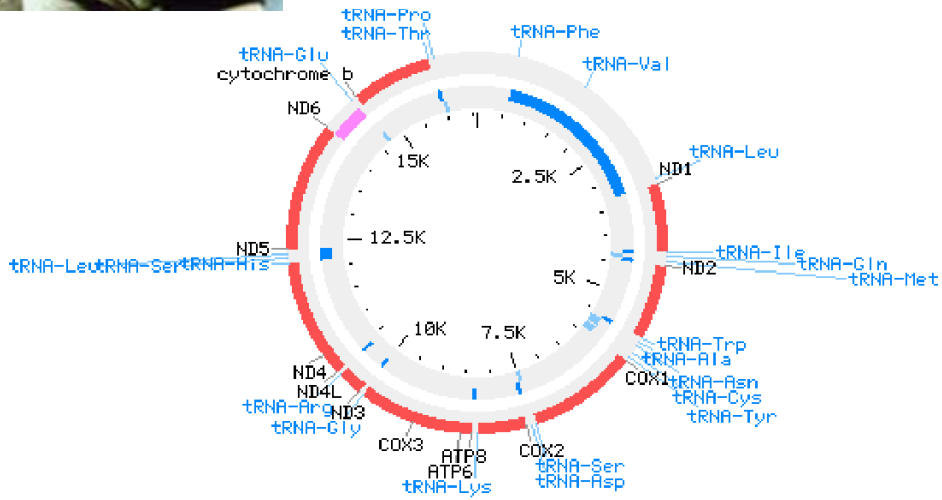
Comparative Genomics

Chicken



NCBI accession #NC_001323

Human



NCBI accession #NC_001807

Health or Disease?

DNA Sequence

Person 1 **A A A T T T**

Normal protein

Some DNA variations have no negative effects

Person 2 **A A T T T T**

Low or nonfunctioning protein

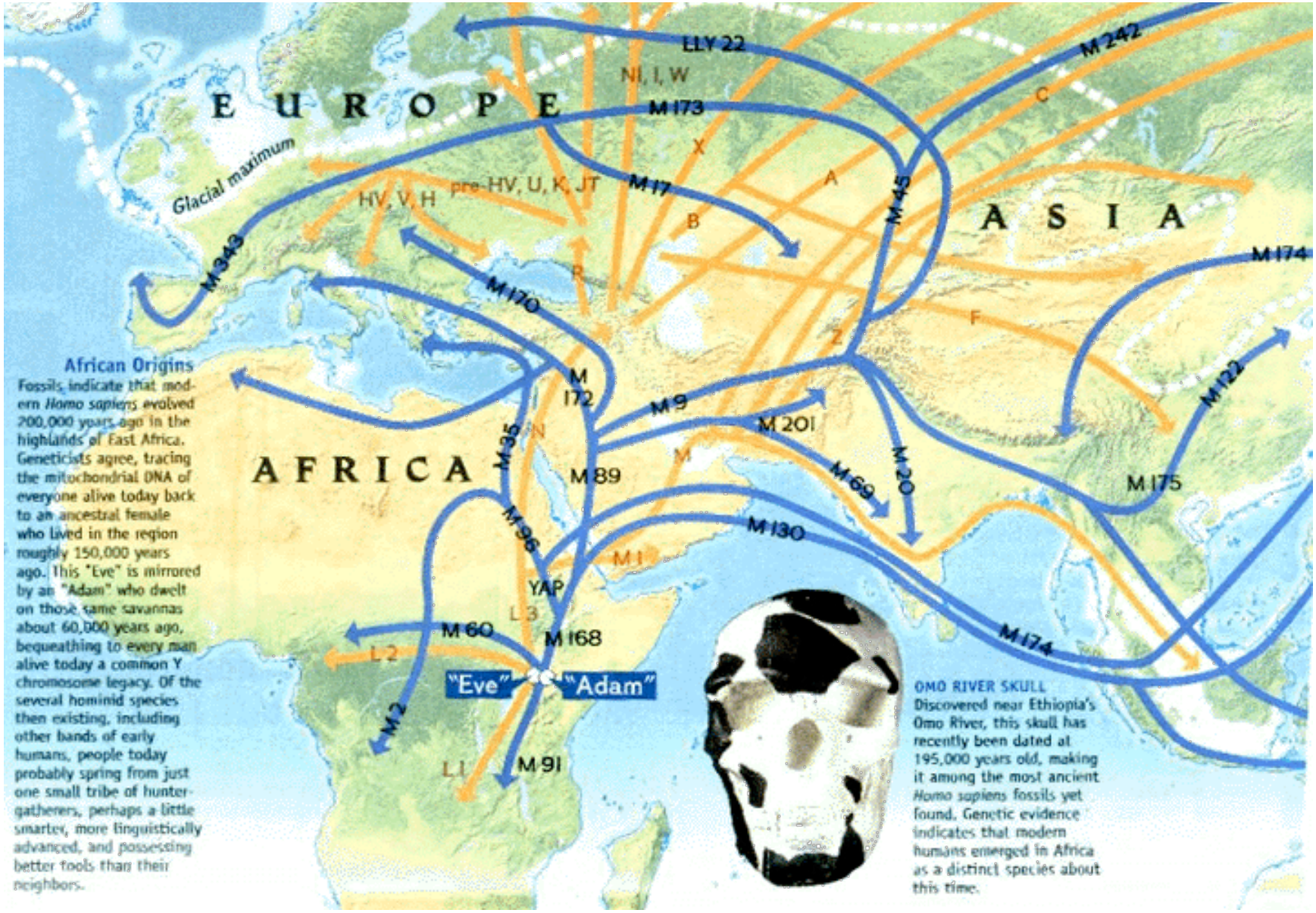
Person 3 **A A C T T T**

Other variations lead to disease (e.g., sickle cell) or increased susceptibility to disease (e.g., lung cancer)

YGG-00-0480

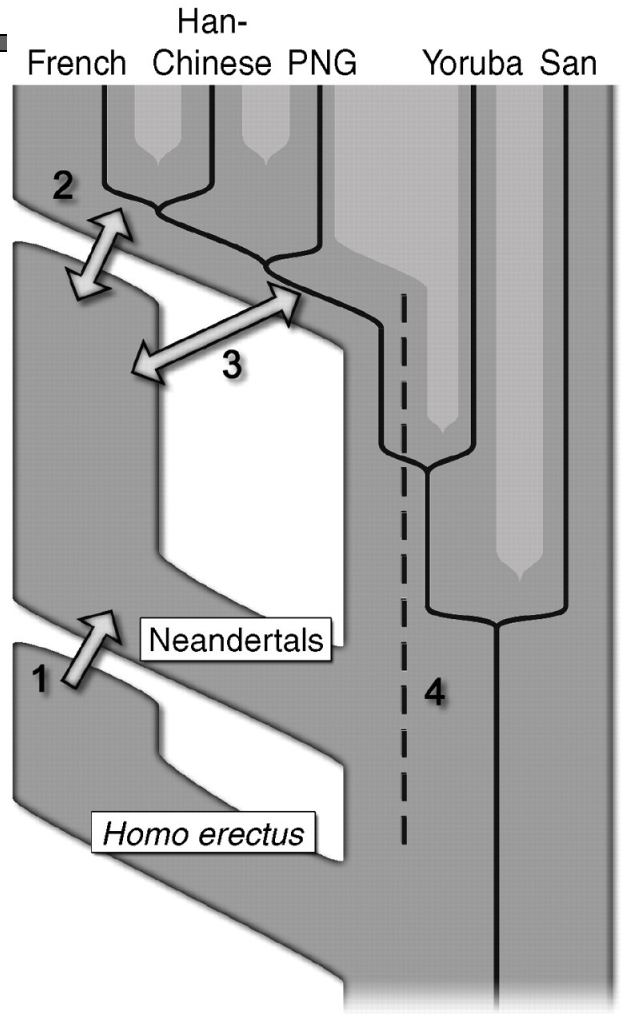
B73 F1 Mo17 B73 F1 Mo17

B73/Mo17-derived hybrids are widely grown; B73 genome has been sequenced.



Map from Genographic project

Fig. 6 Four possible scenarios of genetic mixture involving Neandertals



R. E. Green et al., Science 328, 710-722 (2010)

Published by AAAS


PHASE TWO : INTERPRETATION
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME
SHERMAN Philadelphia Inquirer

Sequence Alignment

Today

- We'll discuss a simple but highly used Dynamic Programming solution to a biological problem
- Arguably one of the most important algorithms in bioinformatics; over 40 years old.
- The ultimate goal of alignment is to describe sequence similarity, or how closely two sequences match each other.
  - Can be a score (number)
  - Can also be an "alignment" (visual)

Applications

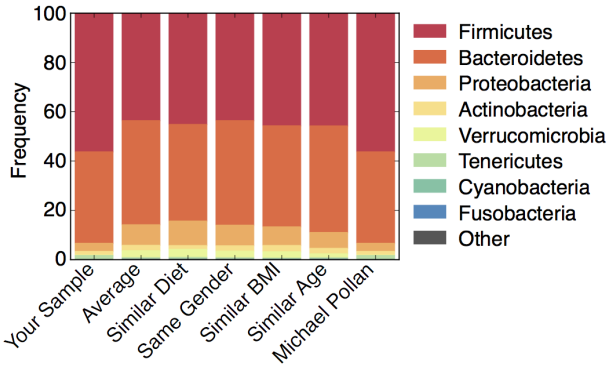
- Prediction on function
  - Commonalities among sequences can imply similar functions
- Database searching (BLAST)
  - Find interesting genes in a new genome
- Sequence divergence
  - Look at evolutionary relationships
- Sequence assembly
  - Making a big sequence from a bunch of small ones

american gut

YOUR AMERICAN GUT SAMPLE

MICHAEL POLLAN

What's in your American Gut sample?



Your most abundant microbes:

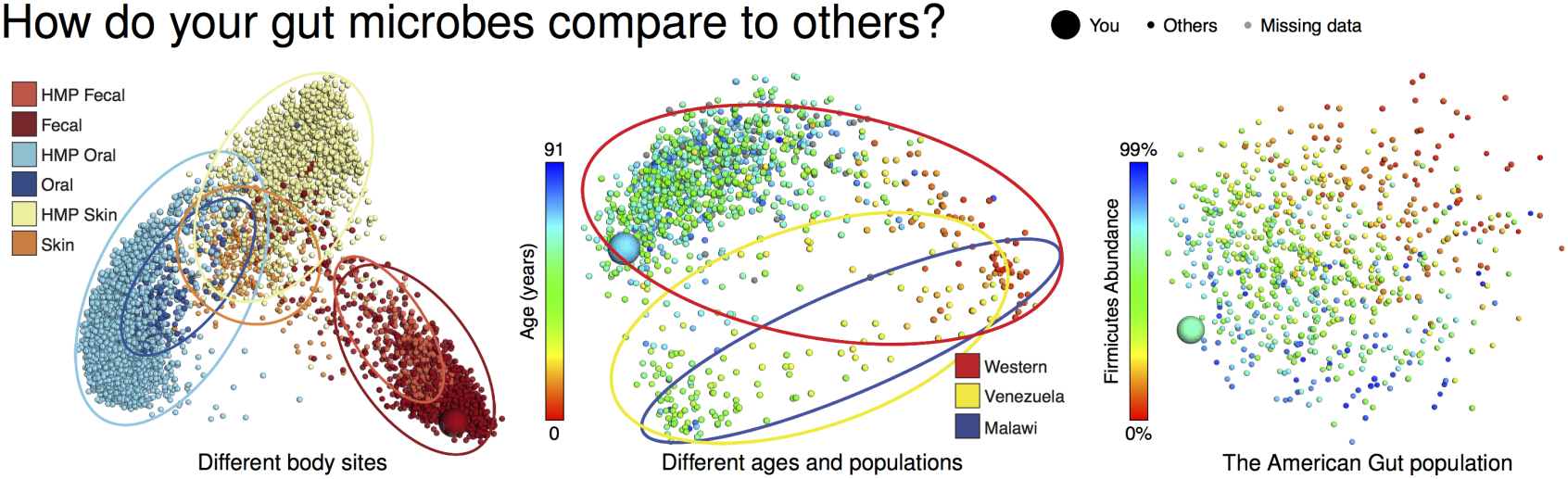
Taxonomy	Sample
Genus *Prevotella*	24.9%
Family Ruminococcaceae	13.4%
Family Lachnospiraceae	10.1%
Genus *Bacteroides*	10.0%

Your most enriched microbes:

Taxonomy	Sample	Population	Fold
Genus *Clostridium*	2.5%	0.3%	7x
Genus *Finegoldia*	0.7%	0.0%	17x
Genus *Prevotella*	24.9%	2.6%	9x
Genus *Collinsella*	0.9%	0.1%	8x

This sample included the follow rare taxa: Genus *Varibaculum*, Genus *Neisseria*, Genus *Campylobacter*, Order ML615J-28

How do your gut microbes compare to others?



Alignment overview

- Computationally, naïve alignments grow exponentially with n : not good
  - There are 10^{17} alignments for two length 30 sequences.

- Luckily, a tried and true method for solving similar problems (we'll provide an overview today) comes to the rescue.

- First efficient algorithm published in 1970 by Needleman and Wunch, improved by Smith and Waterman in 1981.

Global alignment

- Also called a *pairwise alignment*.

- Intuitive goal: related sequences will share many (most?) characters. To maximize this we introduce gaps represented by "-"

Two simple rules

- Rule #1:
  - A gap must be aligned to a nongap, i.e., "-" can not align to "-"
- Rule #2:
  - To distinguish good alignment from not so good ones, we introduce a scoring function *E*. Some functions have biological meaning, some are arbitrary.
- Consequence #1:
  - Alignment length can be no longer than sum of two sequences!

Scoring functions

- Here is a basic scoring function that rewards 1 for a match and -1 for a mismatch gap

$$E(-,a) = E(a,-) = E(a,b) = -1 \quad \forall a \neq b$$
$$E(a,b) = 1 \quad \forall a = b$$

- Can also be represented as a substitution matrix.

Measuring similarity

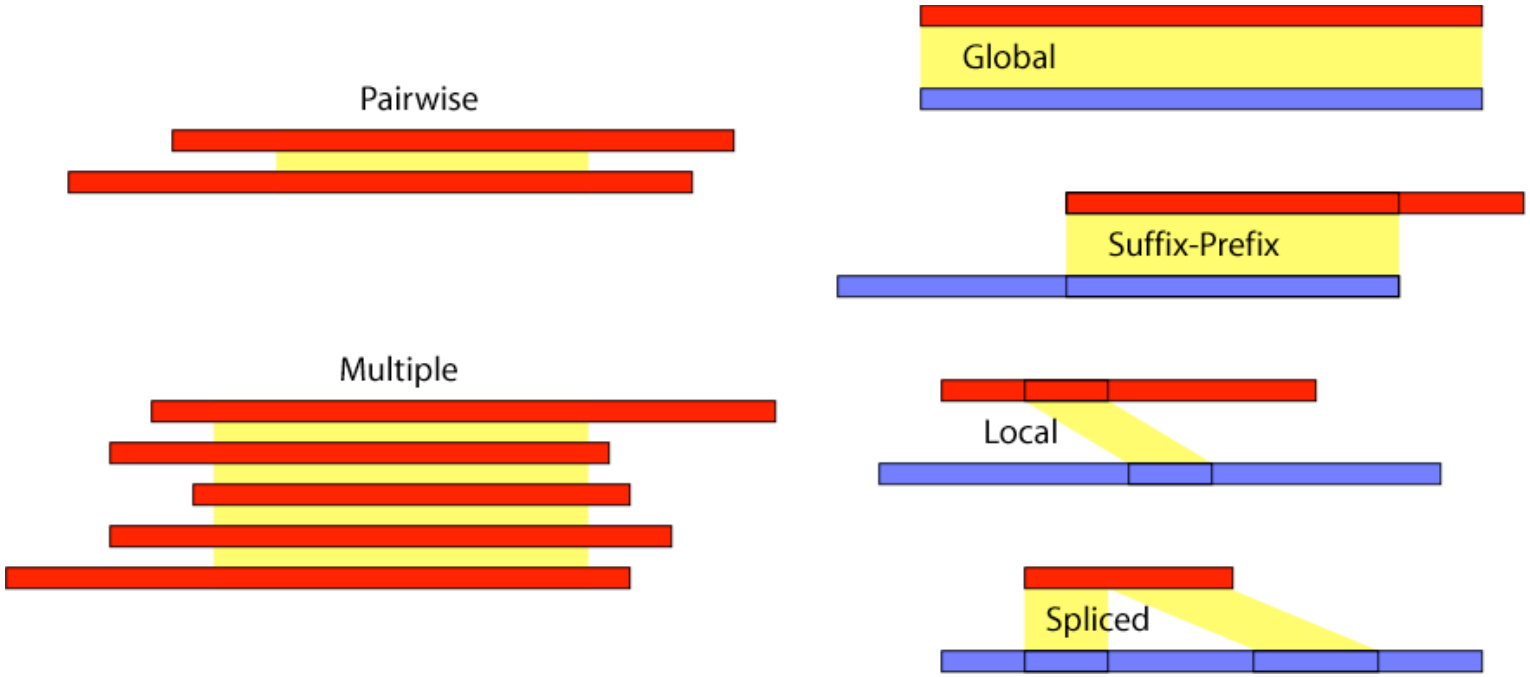
Score: A measure of alignment quality

```
  C  A  T  -  T  C  A  -  C
  C  -  T  C  G  C  A  G  C
-------------------------------
 10 -5 10 -5 -2 10 10 -5 10
```

Total = 33

Scored as E(C,C) E(A,-), E(T,T), E(-,C), etc.

Various types



Example from text

- How do we align these proteins:
  - VIVALASVEGAS
  - VIVADAVIS

In class example

```
S:  CATCAC
T:  CTCCAGC
```

$$E(-,a) = E(a,-) = E(a,b) = -1 \quad \forall a \neq b$$

$$E(a,b) = 1 \quad \forall a = b$$

Global alignment

- Dynamic programming (DP) will save the day!
- DP is a general technique used when a large problem can be broken into smaller, easier problems like this.
- To solve sequence alignment, we will fix two substrings and find the best way to add the next character from at least one string.

Requirements

- We will need four things to compute a global alignment:
  1. Substitution matrix (parameters)
  2. Recurrence relation
  3. Filling up a table
  4. Traceback

Basic intuition

- Suppose we have an optimum alignment of size L. Is the following true?

- $A^* = A^*(s_1 \ldots s_i, t_1 \ldots t_j) + A^*(s_{i+1} \ldots s_n, t_{j+1} \ldots t_m)$
  - Where $|s| = n$ and $|t| = m$

- If so, what would happen if $i = n - 1$ and $j = m - 1$?

Visualization

Case 1: Match *s*[n] w/ *t*[m]

```
                    n - 1  n
s: C  A  T  T  C  A | C
t: C  -  T  T  C  A | G
                    m -1 m
```

Case 2: Match *t*[m] w/ gap

```
                    n - 1
s: C  A  T  T  C  A | -
t: C  -  T  T  C  A | G
                    m -1 m
```

Case 3: Match *s*[n] w/ gap

```
                    n - 1  n
s: C  A  T  T  C  A | C
t: C  -  T  T  C  A | -
                    m-1
```

Pairwise Global Alignment

$T[i,j]$ = Score of optimally aligning first i bases of s with first j bases of t.

$$T[i,j] = \max \begin{cases} T[i-1,j-1] + score(s[i], t[j]) \\ T[i-1,j] + g \\ T[i,j-1] + g \end{cases}$$

	λ	C	T	C	G	C	A	G	C
λ	0	-5	-10	-15	-20	-25	-30	-35	-40
C	-5	10	5						
A	-10								
T	-15								
T	-20								
C	-25								
A	-30								
C	-35								

+10 for match, -2 for mismatch, -5 for space (rowwise)

	λ	C	T	C	G	C	A	G	C
λ	0	-5	-10	-15	-20	-25	-30	-35	-40
C	-5	10	5	0	-5	-10	-15	-20	-25
A	-10	5	8	3	-2	-7	0	-5	-10
T	-15	0	15	10	5 *	0	-5	-2	-7
T	-20	-5	10 *	13	8	3	-2	-7	-4
C	-25	-10	5	20	15	18	13	8	3
A	-30	-15	0	15	18	13	28	23	18
C	-35	-20	-5	10	13	28	23	26	33

Traceback yields both optimal alignments in this example

Some Results

- Most pairwise sequence alignment problems can be solved in $O(mn)$ time. Some speedups exist, most notably the Four Russians technique.

- Space requirement can be reduced to $O(m+n)$, while keeping run-time fixed [Myers88].

- Two highly similar sequences can be aligned in $O(dn)$ time, where d is a measure of the distance between the sequences [Landau86].