Comparing Mouse vs Human Genomes

Comparisons at the genome level are a much harder computational and theoretical problem.
At the finer scale, we can start to see patterns.

From Gregory et al. (2002), Nature.
Within the genome of a single species, there are many duplications, translocations, and inversions.

How genomes involve through duplication.

From Deonier, Tavaré and Waterman, 2005.
How much of the genome is conserved?

- Yeast genome contains 70% coding sequences.
- Human genome contains 1.2% protein coding sequence.

Figure 10 | Relative proportion of different annotations among constrained sequences. The 4.9% of bases in the ENCODE regions identified as constrained is subdivided into the portions that reflect known coding regions, UTRs, other experimentally annotated regions, and unannotated sequence.
Does the stationarity assumption work?

Definition of Terms

- **Homology (of genes)** = similarity due to common ancestry. There are two types of homology, the distinction depends on ordering of speciation and gene duplication dates.
- **Orthologues** = the “same” gene in different organisms, that is, common ancestry goes back to a speciation event.
- **Paralogues** = different genes in the same organism, that is, common ancestry goes back to a gene duplication.
- There are other forms of homology, such as lateral gene transfer.
Synteny

- **linked genes** = genes that reside on the same chromosome.
Synteny

- **linked genes** = genes that reside on the same chromosome.
- **conserved synteny** = a group of linked genes that are highly conserved and hypothesized to be homologous.
Synteny

- **linked genes** = genes that reside on the same chromosome.
- **conserved synteny** = a group of linked genes that are highly conserved and hypothesized to be homologous.
- **syntenic segment** = A group of landmarks that appear in the same order on a single chromosome in each of the two species.
Synteny

- **linked genes** = genes that reside on the same chromosome.
- **conserved synteny** = a group of linked genes that are highly conserved and hypothesized to be homologous.
- **syntenic segment** = A group of landmarks that appear in the same order on a single chromosome in each of the two species.
- **syntenic block** = A set of adjacent syntenic segments.
Synteny

A. Conserved synteny

Chromosome i, species B

\[ g_{2B} \quad g_{1B} \quad g_{3B} \]

Chromosome j, species C

\[ g_{1C} \quad g_{2C} \quad g_{3C} \]

B. Syntenic blocks and segments

Chromosome i, species B

\[ g_{5B} \quad g_{4B} \quad g_{1B} \quad g_{2B} \quad g_{3B} \]

Chromosome j, species C

\[ g_{4C} \quad g_{5C} \quad g_{1C} \quad g_{2C} \quad g_{3C} \]

Syntenic segment

Syntenic block
Synteny
Genome Alignment

1. To align a whole genome we assume that the syntenic regions have already been found through homologous genes. Next, the vast non-coding regions need to be aligned.
Genome Alignment

1. To align a whole genome we assume that the syntenic regions have already been found through homologous genes. Next, the vast non-coding regions need to be aligned.

2. Alignment of non-coding regions is much harder, due to the low conservation.
Genome Alignment

1. To align a whole genome we assume that the syntenic regions have already been found through homologous genes. Next, the vast non-coding regions need to be aligned.

2. Alignment of non-coding regions is much harder, due to the low conservation.

3. To combine speed and sensitivity, most programs use use an anchored-alignment approach: In a first step, a fast search tool is used to identify a chain of high-scoring sequence similarities. These similarities are then used as anchor points for the final alignment, where a more sensitive method aligns those regions that are left over between the identified anchor points.
Genome Alignment

1. To align a whole genome we assume that the syntenic regions have already been found through homologous genes. Next, the vast non-coding regions need to be aligned.

2. Alignment of non-coding regions is much harder, due to the low conservation.

3. To combine speed and sensitivity, most programs use an anchored-alignment approach: In a first step, a fast search tool is used to identify a chain of high-scoring sequence similarities. These similarities are then used as anchor points for the final alignment, where a more sensitive method aligns those regions that are left over between the identified anchor points.

4. This is what the fast pair-wise alignment algorithms BLAST and FASTA. For genome alignment, the programs differ by how the details of how the anchors are strung up, how many anchors to use, etc.
For example, CHAOS, which was developed here by Batzouglou’s group, uses the following seed-and-extension scheme.
Questions to think about

1. How should one frame the null hypothesis in genome alignment? Is it relevant?
Questions to think about

1. How should one frame the null hypothesis in genome alignment? Is it relevant?
2. How should one choose the parameters for the alignment?
Questions to think about

1. How should one frame the null hypothesis in genome alignment? Is it relevant?
2. How should one choose the parameters for the alignment?
3. How sensitive is the “optimal” alignment to the alignment parameters?
Questions to think about

1. How should one frame the null hypothesis in genome alignment? Is it relevant?
2. How should one choose the parameters for the alignment?
3. How sensitive is the “optimal” alignment to the alignment parameters?
4. What does “homology” mean when it applies to non-coding regions? What is the unit of measurement? Can it possibly be inferred at the nucleotide level?
Siepel et al. aligned 5 vertebrate species (human, mouse, rat, chicken, and Fugu), 4 insect species, 2 worm species, and 7 yeast species. They used a program called PhastCons to search for highly conserved elements (HCE). The HCE cover 3% - 8% of the genome, many lying outside of known exons (1.5%). Why are these regions conserved? What do they do?
Genome Conservation
Finding Highly Conserved Regions

1. Start with a multiple alignment of the genome. Siepel et al. used MULTIZ program, which builds a multiple alignment through progressive local pairwise alignments of a designated reference genome. The multiple alignments consisted of blocks covering:
   - 40% of human genome.
   - 86.9% of fly genome.
   - 43.8% of worm genome.
   - 96.6% of yeast genome.

2. Fit the PhastCons HMM.
Kimura’s 2 parameter model for nucleotide substitution

Transition probabilities:
- Horizontal: $a$
- Diagonal & vertical: $b$
- Self: $c = 1 - a - 2b$
Multiplication rule for Markov Chains

$S_t$ is a Markov chain on $\{S_1, \ldots, S_N\}$ with transition probability matrix $P$. Then,

$$\begin{align*}
pr(\text{state after next is } S_k \mid \text{current state is } S_i) \\
= \sum_j pr(\text{state after next is } S_k, \text{next state is } S_j \mid \text{current state is } S_i) \quad [\text{addition rule}] \\
= \sum_j pr(\text{next state is } S_j \mid \text{current state is } S_i) \times pr(\text{state after next is } S_k \mid \text{current state is } S_j) \quad [\text{multiplication rule}] \\
= \sum_j p(i,j) \times p(j,k) \quad [\text{Markov assumption}] \\
= (i,k)-\text{element of } P^2, \text{ where } P=(p(i,j)).
\end{align*}$$

More generally,

$$pr(\text{state } t \text{ steps from now is } S_k \mid \text{current state is } S_i) = i,k \text{ element of } P^t$$
The Continuous Time Version

For any \( s, t \) write \( p_{ij}(t) = pr(S_j \text{ at time } t+s \mid S_j \text{ at time } s) \) for the stationary (time-homogeneous) transition probabilities.

Write \( P(t) = (p_{ij}(t)) \) for the matrix of \( p_{ij}(t) \)'s.

Then for any \( t, u \): \( P(t+u) = P(t) P(u) \).

It follows that \( P(t) = \exp(Qt) \), where \( Q = P'(0) \) is the derivative of \( P(t) \) at \( t = 0 \). 

\( Q \) is called the infinitesimal matrix of \( P(t) \), and satisfies

\[ P'(t) = QP(t) = P(t)Q. \]
Interpretation of $Q$

Roughly, $q(i,j)$ is the rate of transitions of $i$ to $j \neq i$, while $q(i,i) = -\sum_j q(i,j)$, so each row sum is 0. If under some initial conditions, we have a Markov chain evolving in continuous time with infinitesimal matrix $Q$, and $p_j(t) = \text{pr}(S_j \text{ at time } t)$, then

$$p_j(t+h) = \sum_i \text{pr}(S_i \text{ at } t, S_j \text{ at } t+h)$$

$$= \sum_i \text{pr}(S_i \text{ at } t)\text{pr}(S_j \text{ at } t+h \mid S_i \text{ at } t)$$

$$= p_j(t)x(1+hq_{jj}) + \sum_{i \neq j} p_i(t)x \cdot hq_{ij}$$

i.e., $h^{-1}[p_j(t+h) - p_j(t)] = p_j(t)q(j,j) + \sum_{i \neq j} p_i(t)q(i,j)$

which becomes $P' = QP$ as $h \downarrow 0$.

**Important approximation:** when $t$ is small, $P(t) \approx I + Qt$. 
Jukes Cantor Model

\[ Q = \begin{bmatrix}
-3\alpha & \alpha & \alpha & \alpha \\
\alpha & -3\alpha & \alpha & \alpha \\
\alpha & \alpha & -3\alpha & \alpha \\
\alpha & \alpha & \alpha & -3\alpha
\end{bmatrix} \]

\[ P(t) = \begin{bmatrix}
r & s & s & s & s \\
s & r & s & s \\
s & s & r & s \\
s & s & s & r
\end{bmatrix} \]

**Exercise:** \( r = \frac{(1+3e^{-4\alpha t})}{4}, \quad s = \frac{(1-e^{-4\alpha t})}{4}. \)
PhastCons: a two-state phylogenetic HMM

The phylogenetic model associated with the two states are identical except for a scaling parameter $\rho$. 

$\begin{align*}
  x &= \text{TGGGCGATATACGA} \ldots \\
  &= \text{TTGGGGCATGTGGGT} \ldots \\
  &= \text{AGCAGACGTCCGCAA} \ldots
\end{align*}$
PhastCons Model

The nonconserved phylogenetic model is defined as

\[ \psi_n = (Q, \pi, \tau, \beta) : \]

- \( Q \), a 4 \( \times \) 4 substitution rate matrix.
The nonconserved phylogenetic model is defined as
\[ \psi_n = (Q, \pi, \tau, \beta) : \]

- \( Q \), a 4 \( \times \) 4 substitution rate matrix.
- \( \tau \), a tree topology (given).
PhastCons Model

The nonconserved phylogenetic model is defined as
\( \psi_n = (Q, \pi, \tau, \beta) \):

- \( Q \), a \( 4 \times 4 \) substitution rate matrix.
- \( \tau \), a tree topology (given).
- \( \pi \) Background (equilibrium probabilities for \{A, C, G, T\} (estimated by all bases in data set).

In addition there are \( \mu \) and \( \nu \) that control transitions between the conserved and non-conserved states.
PhastCons Model

The nonconserved phylogenetic model is defined as

\[ \psi_n = (Q, \pi, \tau, \beta) : \]

- \( Q \), a 4 \( \times \) 4 substitution rate matrix.
- \( \tau \), a tree topology (given).
- \( \pi \) Background (equilibrium probabilities for \{A, C, G, T\} (estimated by all bases in data set).
- \( \beta \) branch lengths of the phylogeny in expected substitutions per site.
The nonconserved phylogenetic model is defined as $\psi_n = (Q, \pi, \tau, \beta)$:

- $Q$, a $4 \times 4$ substitution rate matrix.
- $\tau$, a tree topology (given).
- $\pi$ Background (equilibrium probabilities for $\{A, C, G, T\}$ (estimated by all bases in data set).
- $\beta$ branch lengths of the phylogeny in expected substitutions per site.

The conserved phylogenetic model is simply $\psi_c = (Q, \pi, \tau, \rho \beta)$. In addition there are $\mu$ and $\nu$ that control transitions between the conserved and non-conserved states.
Fitting the HMM

- Given $\psi_n$, $\psi_c$, and a column $X$ of the multiple alignment, $P(X|\psi_n)$ and $P(X|\psi_c)$ can be computed. These are the emission probabilities.
Fitting the HMM

- Given $\psi_n, \psi_c$, and a column $X$ of the multiple alignment, $P(X|\psi_n)$ and $P(X|\psi_c)$ can be computed. These are the emission probabilities.
- The optimal path (sequence of conserved/nonconserved) can be computed using the Viterbi algorithm.

Many details swept under the rug.
Fitting the HMM

- Given $\psi_n$, $\psi_c$, and a column $X$ of the multiple alignment, $P(X|\psi_n)$ and $P(X|\psi_c)$ can be computed. These are the emission probabilities.
- The optimal path (sequence of conserved/nonconserved) can be computed using the Viterbi algorithm.
- For each position $t$, $P(\text{Conserved at } t|\text{alignment})$ can be computed by forward backward algorithm. This is referred to as the conservation score.
- The parameters $\beta$, $\rho$, $Q$ can be estimated using Baum Welch Algorithm.
Fitting the HMM

- Given $\psi_n$, $\psi_c$, and a column $X$ of the multiple alignment, $P(X|\psi_n)$ and $P(X|\psi_c)$ can be computed. These are the emission probabilities.
- The optimal path (sequence of conserved/nonconserved) can be computed using the Viterbi algorithm.
- For each position $t$, $P(\text{Conserved at } t|\text{alignment})$ can be computed by forward backward algorithm. This is referred to as the conservation score.
- The parameters $\beta$, $\rho$, $Q$ can be estimated using Baum Welch Algorithm.
- Missing bases (e.g. assembly gaps, deletions) were handled by marginalizing over missing bases.
Fitting the HMM

- Given $\psi_n$, $\psi_c$, and a column $X$ of the multiple alignment, $P(X|\psi_n)$ and $P(X|\psi_c)$ can be computed. These are the emission probabilities.

- The optimal path (sequence of conserved/nonconserved) can be computed using the Viterbi algorithm.

- For each position $t$, $P($Conserved at $t|alignment)$ can be computed by forward backward algorithm. This is referred to as the conservation score.

- The parameters $\beta$, $\rho$, $Q$ can be estimated using Baum Welch Algorithm.

- Missing bases (e.g. assembly gaps, deletions) were handled by marginalizing over missing bases.

- Lod scores were defined as $\log \frac{P(X|\psi_c)}{P(X|\psi_n)}$. Many details swept under the rug.
Fitting the HMM

- Given $\psi_n$, $\psi_c$, and a column $X$ of the multiple alignment, $P(X|\psi_n)$ and $P(X|\psi_c)$ can be computed. These are the emission probabilities.
- The optimal path (sequence of conserved/nonconserved) can be computed using the Viterbi algorithm.
- For each position $t$, $P(\text{Conserved at } t|\text{alignment})$ can be computed by forward backward algorithm. This is referred to as the conservation score.
- The parameters $\beta$, $\rho$, $Q$ can be estimated using Baum Welch Algorithm.
- Missing bases (e.g. assembly gaps, deletions) were handled by marginalizing over missing bases.
- Lod scores were defined as $\log \frac{P(X|\psi_c)}{P(X|\psi_n)}$.

Many details swept under the rug.
Branch lengths to chicken and fish significantly under estimated. Why?

Siepel et al. claims that inaccuracies in branch length estimation does not significantly effect results.
Genome Conservation
A surprisingly large fraction of highly conserved elements are 3’ UTRs (9.6% of top 5000), compared to 5’ UTRs (1.1%).
Location and Composition of Conserved Elements
Figure S3: Coverage of vertebrate conserved elements by feature type for various score classes. The numbers 1–10 indicate disjoint classes of equal numbers of conserved elements, ranging from the highest scoring 10% of elements (1) to the lowest scoring 10% of elements (10). Subsets of class 1 consisting of the top-scoring 5000, 1000, and 100 elements are also shown.