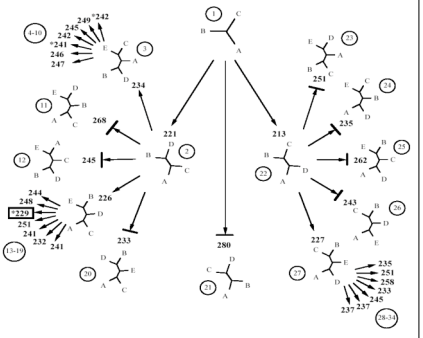
Searching for trees

Branch and bound algorithm:

The branch-and-bound algorithm for exact solution of the problem of finding an optimal parsimony tree. The search tree is the same as for exhaustive search, with tree lengths for a hypothetical data set shown in boldface type. If a tree lying at a node of this search tree has a length that exceeds the current lower bound on the optimal tree length, this path of the search tree is terminated (indicated by a cross-bar), and the algorithm backtracks and takes the next available path. When a tip of the search tree is reached (i.e., when we arrive at a tree containing the full set of taxa), the tree is either optimal (and hence retained) or suboptimal (and rejected). When all paths leading from the initial 3-taxon tree have been explored, the algorithm terminates, and all most-parsimonious trees will have been identified. Asterisks indicate points at which the current lower bound is reduced. See text for additional explanation, and circled numbers represent the order in which phylogenetic trees are visited in the search tree.



Adaptive Molecular Evolution

Nonsynonymous

vs

Synonymous

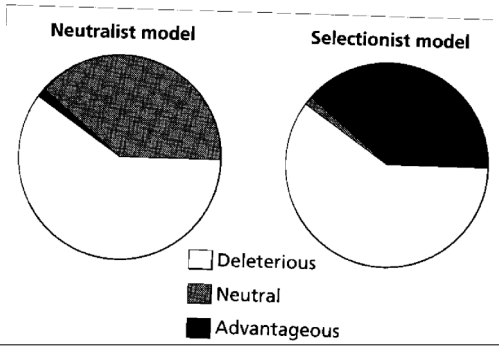
Reading for today

- Li and Graur chapter (PDF on website)
- Evolutionary EST paper (PDF on website)

Neutral theory

- The majority of substitutions are either deleterious or have no function.
  - Introns, pseudogenes, non-coding DNA.
- What about coding regions, promoters, centromeres.
  - How do we evolve differences between species

The neutral theory of molecular evolution

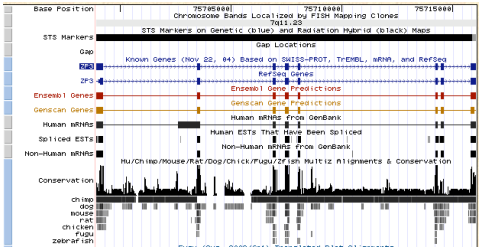


Predictions of neutral theory

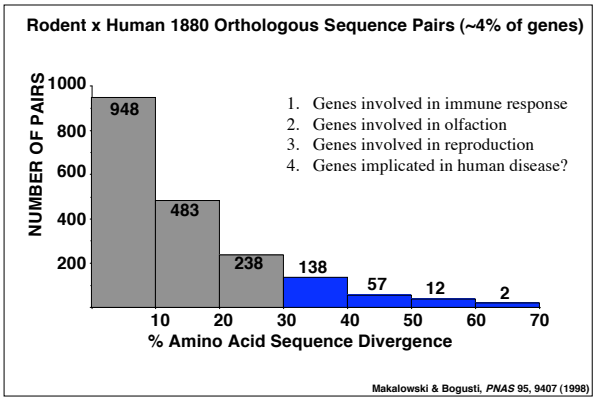
- There is an inverse correlation between rate of substitution and degree of functional constraint.
- Patterns of base composition and codon usage reflect mutational rather than selective pressures.
- There is a constant rate of molecular evolution.
- The level of within species variation is the product of population size and mutation rate and is correlated with levels between species.

Well accepted "rule":

Evolutionarily conservation of genes and regions implies functional importance



What molecular changes are different between species?



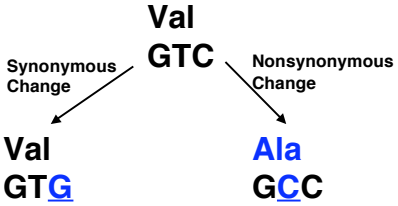
Potential causes of rapid evolution

- **Lack of constraint:**
  Selectively neutral evolution
- **Adaptive value for change:**
  Positive "Darwinian" selection

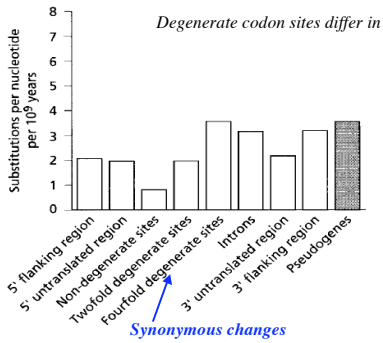
→ **Compare cDNA sequences.**

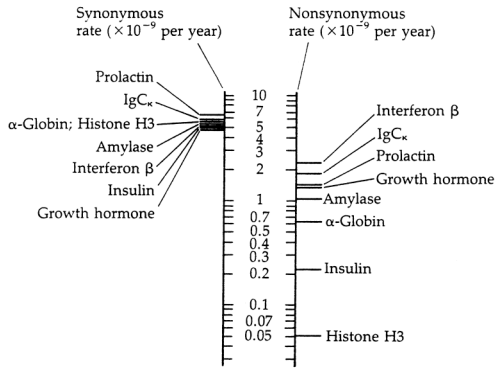
2 types of changes in codons

Synonymous = silent change (amino acid stays the same)
Nonsynonymous = replacement change (changes amino acid)



Rates of synonymous changes is similar to pseudogenes





Synonymous and nonsynonymous sites are both in coding regions. Synonymous sites are considered selectively neutral. Therefore, we can use synonymous sites as a "ruler" for nonsynonymous substitutions. When nonsynonymous changes exceeds synonymous changes, infer positive selection.

There are more nonsynonymous than synonymous sites in coding DNA

Prot. 1:	Ile	Cys	Ile	Lys	Ala	Leu	Val	Leu	Thr
DNA1:	ATA	TGT	ATA	AAG	CGA	GTC	CTG	TTA	ACA
DNA2:	ATA	TGT	ATA	AAG	CGA	GTC	CTG	TTA	ACA
Prot. 2:	Ile	Cys	Ile	Lys	Ala	Leu	Val	Leu	Thr

d_N = # nonsynonymous substitutions/# nonsynonymous sites
d_S = # synonymous substitutions/# synonymous sites

Test for selection by comparing d_N and d_S

$d_N/d_S = 1$: Neutral evolution

$d_N/d_S < 1$: Purifying selection

$d_N/d_S > 1$: Positive selection

The d_N/d_S ratio (ω) measures the selective pressure

Multiple methods for calculating d_N/d_S

- "Counting" methods
  - Nei and Gojobori
  - Li et al.
- Maximum likelihood methods (model of codon evolution)
  - Muse and Gaut
  - Neilsen and Yang

Codon degeneracy

- Non-degenerate
  - All mutations produce nonsynonymous change
- Two-fold degenerate
  - one of the three possible changes is synonymous
- Four-fold degenerate
  - all mutations produce synonymous change

When counting sites: GAG (Glu)

- Non-degenerate (1) [All mutations produce nonsynonymous change]
  - nonsynonymous
- Two fold degenerate (2) [one of the three possible changes is synonymous]
  - 1/3 synonymous and 2/3 nonsynonymous
- Four fold degenerate (4) [all mutations produce synonymous change]
  - synonymous

Note: Three fold degenerate treated as two-fold.

Example:

```
Degeneracy 1
                Asp  Thr  Ala  Val
Sequence 1      GAC  ACA  GCG  GTT
```

How many synonymous sites in sequence 1?
First, assign degeneracy to each codon position.

Example:

```
Degeneracy 1    1
                Asp  Thr  Ala  Val
Sequence 1      GAC  ACA  GCG  GTT
```

Example:

```
Degeneracy 1    11
                Asp  Thr  Ala  Val
Sequence 1      GAC  ACA  GCG  GTT
```

Example:

```
Degeneracy 1    112
                Asp  Thr  Ala  Val
Sequence 1      GAC  ACA  GCG  GTT
```

Example:

```
Degeneracy 1    112  1
                Asp  Thr  Ala  Val
Sequence 1      GAC  ACA  GCG  GTT
```

Example:

```
Degeneracy 1   112  11
               Asp  Thr  Ala  Val
Sequence 1     GAC  ACA  GCG  GTT
```

Example:

```
Degeneracy 1   112  114
               Asp  Thr  Ala  Val
Sequence 1     GAC  ACA  GCG  GTT
```

Example:

```
Degeneracy 1   112  114  1
               Asp  Thr  Ala  Val
Sequence 1     GAC  ACA  GCG  GTT
```

Example:

```
Degeneracy 1   112  114  11
               Asp  Thr  Ala  Val
Sequence 1     GAC  ACA  GCG  GTT
```

Example:

```
Degeneracy 1   112  114  114
               Asp  Thr  Ala  Val
Sequence 1     GAC  ACA  GCG  GTT
```

Example:

```
Degeneracy 1   112  114  114  1
               Asp  Thr  Ala  Val
Sequence 1     GAC  ACA  GCG  GTT
```

Example:

```
Degeneracy 1   112  114  114  11
               Asp  Thr  Ala  Val
Sequence 1     GAC  ACA  GCG  GTT
```

Example:

```
Degeneracy 1   112  114  114  114
               Asp  Thr  Ala  Val
Sequence 1     GAC  ACA  GCG  GTT
```

Example:

```
Degeneracy 1   112  114  114  114
               Asp  Thr  Ala  Val
Sequence 1     GAC  ACA  GCG  GTT
```

How many nonsynonymous sites in sequence 1?

Example:

```
Degeneracy 1   112  114  114  114
               Asp  Thr  Ala  Val
Sequence 1     GAC  ACA  GCG  GTT
```

How many nonsynonymous sites in sequence 1?
8 nondegenerate sites 1 two fold degenerate site = 8.66 nonsynonymous sites

Example:

```
Degeneracy 1   112  114  114  114
               Asp  Thr  Ala  Val
Sequence 1     GAC  ACA  GCG  GTT
```

How many synonymous sites in sequence 1?

Example:

```
Degeneracy 1   112  114  114  114
               Asp  Thr  Ala  Val
Sequence 1     GAC  ACA  GCG  GTT
```

How many synonymous sites in sequence 1?
3 four fould degenerate sites, 1 two fold = 3.33 synonymous sites.

Example:

Degeneracy 1	112	114	114	114
	Asp	Thr	Ala	Val
Sequence 1	GAC	ACA	GCG	GTT
Sequence 2	GCC	ACT	TCG	GTT
	Ala	Thr	Ser	Val
Degeneracy 2	114	114	114	114

Sequence 2 has 8 nonsynonymous sites and 4 synonymous sites.

For this comparison, we average number from both sequences.
Nonsynonymous sites = (8.66 + 8)/2 = 8.33
Synonymous sites = (3.33 + 4) = 3.67

Example:

Degeneracy 1	112	114	114	114
	Asp	Thr	Ala	Val
Sequence 1	GAC	ACA	GCG	GTT
Sequence 2	GCC	ACT	TCG	GTT
	Ala	Thr	Ser	Val
Degeneracy 2	114	114	114	114

There are 2 nonsynonymous changes,
So dn = 2/8.33 = 0.24

There is 1 silent change,
So ds = 1/3.67 = 0.27

dn/ds = 0.24/0.27 = 0.88
< 1 despite having more nonsynonymous changes.

Other factors can effect calculation of d_N/d_S

- Transition/transversion ratio
  - Transitions typically more frequent
- Pathway of substitution
- Codon bias

Counting differences

Two pathways between CCT and CAG:

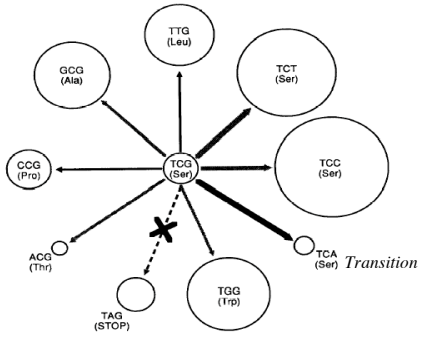
Pathways	Syn	Nonsyn
CCT (Pro) ↔ CAT (His) ↔ CAG (Gln)	0	2
CCT (Pro) ↔ CCG (Pro) ↔ CAG (Gln)	1	1
Average	0.5	1.5

Nearly all counting methods assume all pathways are equally likely.

Codon Bias

- Unequal codon usage results in reduced number of effective codon sites.
- Ignoring codon bias leads to underestimate of ds.

Maximum likelihood methods incorporate models of codon evolution / bias.



Problems with dn/ds for detecting selection

- Positive selection acting only on a few sites (binding cleft).
- Burst of positive selection followed by purifying selection (lineage specific events).
- Positive selection in promoter and non-coding regions.
- Positive selection for post-translational modification (glycosylation).

Reproductive proteins and ESTs coupled with d_N/d_S analysis

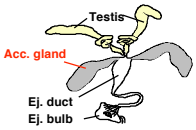
Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*

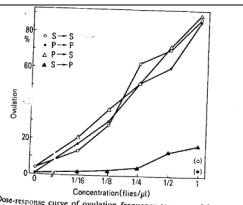
Willie J. Swanson*†, Andrew G. Clark‡, Heidi M. Waldrip-Dail‡, Mariana F. Wolfner*, and Charles F. Aquadro*

*Department of Molecular Biology and Genetics, Biotechnology Building, Cornell University, Ithaca, NY 14853-2703; and ‡Institute of Molecular Evolutionary Genetics, Department of Biology, Pennsylvania State University, 208 Mueller Lab, University Park, PA 16802-5301

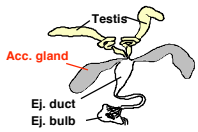
Elevation of egg laying rate is regulated by seminal fluid protein in flies

- Seminal fluid by by accessory glands
- Seminal fluid can be artificially injected into female following mating to test for between species differences.



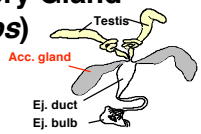


Dose-response curve of ovulation frequency to paragonial extract injection. Methanolic extract of the paragonial substances were prepared as follows: 1000 virgin male flies (5–10-day-old) that had been stored at −20°C were homogenized in 10 ml 80% methanol, and homogenates were centrifuged at 3500 rpm for 10 min. The supernatants were evaporated, the residues dissolved in 1 ml saline, and centrifuged again at 15,000×g for 10 min. The supernatants, containing approximately 1 fly equivalent per μl, were diluted to various concentrations, and 0.25 μl each were injected into the abdominal cavity of 6-day-old virgin females. Two days after injection, the females were dissected and examined for the presence of ova in the uterus. Ovulation frequency is the proportion of females retaining ova to total females dissected. ○, Extract from *D. suzukii* male injected into *D. suzukii* female; ●, from *D. pulchrella* to *D. suzukii*; △, from *D. pulchrella* to *D. suzukii*; ▲, from *D. suzukii* to *D. pulchrella*. Points shown in parentheses represent the corresponding figures obtained by the injections of virgin female extracts. Each point is based upon 98–103 females.



Injection of accessory gland extracts from species P into species S induces ovulation

***Drosophila* Accessory Gland Proteins (*Acps*)**

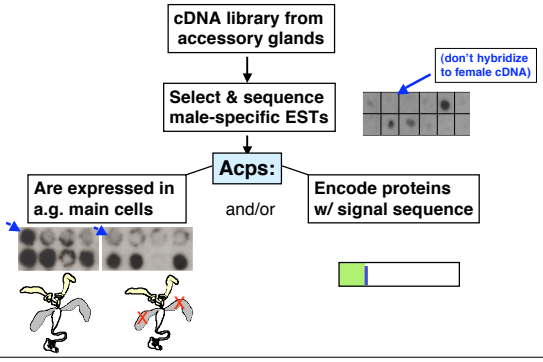


- **Induce egg laying/ovulation.**
- **Sperm storage.**
- **Toxic to females (byproduct of other function?).**
- **Reduce remating rate.**
- **18 known genes from estimated 100 *Acps*.**

How to find Acps?

- Find genes expressed in accessory gland
- Look for male specific genes
- Analyze for evidence of secreted proteins

EST screening for Acp genes



EST Project Results

- Following differential hybridization;
  285 ESTs sequenced.
- 212 independent genes identified.
  35% (75) not in other *Drosophila* EST datasets
  47%(100) do not match non-*Drosophila* GenBank
  8% (18) not identified by Celera annotation
  24% (51) have predicted signal sequences

With 200+ genes, how decide which might be important for species differences?

It was an "evolutionary EST" screen

cDNA library made from *D. simulans* accessory glands

It was an "evolutionary EST" screen

cDNA library made from *D. simulans* accessory glands

↓ *Blast to complete D. mel genome*

Compare to *D. melanogaster* genomic sequence
(*simulans* and *melanogaster* are ~10% divergent at silent sites)

It was an "evolutionary EST" screen

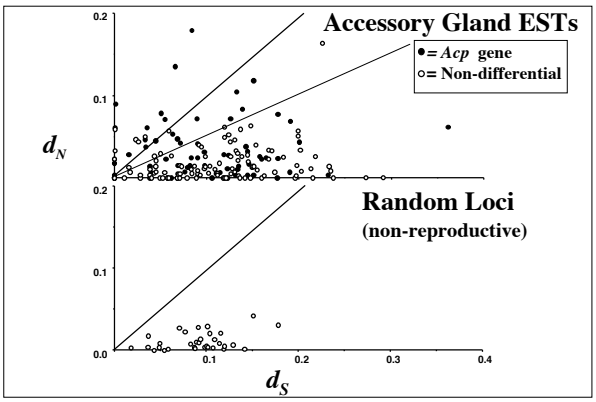
cDNA library made from *D. simulans* accessory glands

↓

Compare to *D. melanogaster* genomic sequence
(*simulans* and *melanogaster* are ~10% divergent at silent sites)

↓

Estimate d_N/d_S ratios to identify candidate rapidly-evolving genes
(d_N/d_S >1 OR 0.5)



2X increase in d_N for *Acp genes compared to non-Acps*, identical d_S

Compare d_N:

Accessory gland specific d_N= 0.052
Non-differential d_N= 0.024
$F_{1,180}$ = 62.22, p < 0.0001

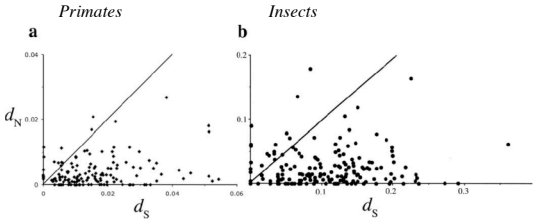
Compare d_S:

Accessory gland specific d_S= 0.11
Non-differential d_S= 0.10
$F_{1,180}$ = 0.53, p = 0.47

Is this limited to invertebrates

- Human versus Chimpanzee
- Data available from prostate ESTs
- Blast human versus Chimpanzee
- Calculate d_N/d_S

Similar pattern observed for human/Chimpanzee



Clark and Swanson, PLoS Genetics 1:e35

Other genome wide scans

- What adaptive changes occurred along human lineage
- What genes have been subject to recent selective sweeps, and in what populations.

1. My gene is the fastest evolving gene between two organisms I am studying, it is therefore been subjected to positive Darwinian selection. Describe why this statement is true or false.
2. Does the neutral model of molecular evolution allow for deleterious mutations?
3. Define two fold and four fold degenerate sites.
4. Define a synonymous substitution.
5. Define a nonsynonymous substitution
6. What is codon bias?
7. Do two fold degenerate sites evolve at the same rate as pseudogenes?
8. Describe one way to test for positive Darwinian selection.
9. Describe the difference between the neutralist model of molecular evolution and the selectionist model of molecular evovlution.
10. Between two closely related species, my gene has 2 nonsynonymous and 1 synonymous difference, it is therefore been subjected to adaptive evolution. Is this statement correct?