**February 15, 2005**

**Bio 107/207**
**Winter 2005**
**Lecture 13**
**Molecular population genetics. II. Natural selection**

- one approach for testing the neutral theory has been to study the adaptive significance of specific protein polymorphisms.
- it entails a multi-level research program with the following objectives:

1. document biochemical differences among allozyme genotypes.
2. show that the biochemical differences affect physiological performance.
3. demonstrate that physiological differences between genotypes affect fitness in nature.

- this has been successfully accomplished for a small number of polymorphisms.
- examples include the *Lap*-1 locus in the blue mussel *Mytilus edulis*, the *Pgi* polymorphism in *Colias* butterflies, and the *Ldh*-B locus in the killifish, *Fundulus heteroclitus*.
- in each case, strong evidence favoring selection has been obtained.

- it is unfortunate that examples of balancing selection acting at particular allozyme loci has not been able to definitively refute the neutral theory.
- most proponents of the neutral theory would simply say that these examples are **exceptions** to the general rule.
- what we require is not independent cases favoring selection in isolated species but an overall assessment of the role of selection **in one species**.
- so, for example, if 10 polymorphic enzyme loci were observed segregating in one species, how many of these are selectively maintained and what proportion are neutral?

**Testing the neutral theory at the DNA level**

- the neutralist-selectionist controversy has moved from the protein level to the DNA sequence level.
- are there any advantages provided by studying evolution at the DNA level?
- being one step removed from the protein level, the analysis of DNA sequence variation may not be capable of providing definitive answers about adaptive evolution.
- however, there are three distinct advantages of studying DNA sequence variation over protein variation.

**1. increased resolution**

- studying protein variation only provides information on (typically) a single amino acid substitution.
- usually, when these protein variants are sequenced, it is observed that the alleles being examined results from a single amino acid substitution caused by a single base pair change.
- it is obvious that moving to the nucleotide level allows many more sites to be examined for polymorphism within species or divergence among species.
- clearly, studying protein variation is a very crude estimate of the variation that is present at the level of DNA – in fact, allozyme alleles have been found to be heterogeneous.

**2. the ability to assess silent and replacement substitutions.**

- this is perhaps the most important advantage.
- the vital information is provided by examining "silent" changes with "replacement" changes.
- according to the neutral theory, most replacement changes in proteins are *de facto* neutral.
- because they are neutral they should evolve at similar rates, and show similar levels of polymorphism, to silent changes.
- the ability to compare the evolutionary dynamics of replacement and silent mutations allows for powerful tests of the neutral theory.

**3. the ability to infer the evolutionary histories of genes.**

- the evolutionary history of a gene is recorded its genealogy.
- certain types of selection will leave what is called a "footprint" on DNA sequences.
- for balanced polymorphisms, alleles persist for long periods of time in natural populations, much longer than expected for strictly neutral alleles.
- because of this long persistence, considerable silent changes will accumulate within alleles subject to long-term balancing selection.
- periods of rapid adaptive substitutions will also be apparent.
- here, the amount of silent variation will be less than expected under the strictly neutral model.

- coalescent theory provides a powerful foundation for inferring the effects of selection.
- this topic will be covered on Thursday.
- here, we will examine some of the approaches for testing the neutral theory.

**Ewens-Watterson test**

- this was one of the first tests of the neutral theory
- it is based on the infinite alleles model – that every new allele that enters a population is unique.
- it also assumes that the population is at mutation-drift equilibrium.
- according to the neutral theory the equilibrium heterozygosity at a locus is

$$H_e = 4Nu/(4Nu + 1)$$

- Ewens showed that at this equilibrium, the expected number of different alleles in a sample of size N is:

$$E(N) = \sum_{i=0}^{2N-1} 4Nu/(4Nu + i)$$

- the Ewens-Watterson test compares the observed Hardy-Weinberg homozygosity for a sample of size with n different alleles is significantly different from homozygosity expected under the neutral theory if the locus is at mutation-drift equilibrium.
- the allele frequency distribution can be too even (suggestive of balancing selection) or too uneven (suggestive of a recent selective sweep) compared with the neutral expectation.
- unfortunately, past demographic events like population bottlenecks can cause significant departures from neutrality.

- one way to deal with this problem is to examine multiple independent loci.
- if a bottleneck has occurred, then all loci should be affected to the same extent.
- if selection has occurred, then it should be idiosyncratic.

**Tajima's test**

- this test is similar to the Ewens-Watterson test but takes into account the extent of nucleotide differences between alleles in addition to their population frequencies.
- Tajima's test is based on comparing two estimates of nucleotide diversity – $\pi$ and $\theta$.
- the neutral theory predicts that these two estimates should be equal (at mutation-drift equilibrium).
- the Tajima D statistic is given by:

$$D = [k - S/a_1] / \sqrt{V(k - S/a_1)}$$

where the denominator is the difference in the standard deviations of the two estimates.

- in an equilibrium population D is expected to be zero.
- balancing selection is expected to give a positive D value.
- purifying selection is expected to produce a negative D value.
- however, once again demographic factors can be difficult to distinguish from selection.

**McDonald-Kreitman test**

- the neutral theory predicts that polymorphism is simply a transient phase of molecular evolution as neutral or nearly neutral alleles wander aimlessly through populations by random drift.
- the neutral theory thus predicts that the dynamics of silent and replacement polymorphism should be similar.
- after all, the theory was based on a class of variants (namely allozymes) that involved replacement changes in proteins that modified net charge properties.
- how can we test whether the dynamics of these two classes are similar?
- one option is by comparing patterns of polymorphism and divergence in closely related species.

- the rationale for this test of the neutral theory is that the same protein in closely related species should have a very similar, if not identical, function.
- therefore, the degree of constraint acting on that protein should be the same in the two species.
- the degree of constraint determines the level of polymorphism observed and also the rate of evolution.
- the approach is as follows.
- first, the homologous protein-coding locus is sequenced in two closely related species.
- a number of alleles are sequenced from both species so we can obtain estimates of the amount of polymorphism present.

```
                          N   A   E   R   T   R
Example: D. melanogaster  AATGCGGAACGGACTCGT
                          --C---------------
                          -----------T-----
         D. simulans      -----C-T-------A--
                          -------T-------A--
```

-------T-------A--

- mutations are then classified as either being **fixed** between species or **polymorphic** within species
- since we have sequence for a coding region, we can also can classify mutations as being **silent** or **replacement**.
- the numbers are put into a 2 x 2 matrix.

| | fixed | polymorphic |
|---|---|---|
| replacement | a | c |
| silent | b | d |

- the neutral theory predicts that polymorphism is simply a transient phase of molecular evolution that ultimately produces fixed differences between species.
- therefore, the proportion of replacement to silent changes that are polymorphic within species should be similar to the proportion of replacement to silent changes that are fixed between species.
- **in other words, the test predicts that the ratio of a:b will be identical to the ratio of c:d.**
- we can test this by means of a Chi-square or G-test of independence.

**Examples**:

**1. Alcohol dehydrogenase (Adh) locus in *Drosophila melanogaster* and *D. simulans*.**

| | fixed | polymorphic |
|---|---|---|
| replacement | 7 | 2 |
| silent | 17 | 42 |

- what does these numbers mean?
- we see 7 of 24, or 29%, of the fixed differences between species are replacement substitutions.
- only 2 of 44, or 5% of the polymorphisms occur at replacement sites.
- this is not what one would expect if the replacement substitutions are selectively neutral.
- the different proportions of fixed and polymorphic sites is significant ($G = 7.43$, $P = 0.006$).
- this suggests that natural selection has been an important mechanism directing evolution at the Adh locus among *Drosophila* species.
- in particular, it appears to have resulted in an excess number of replacement changes suggesting positive selection for advantageous alleles.

**2. Glucose-6-phosphate dehydrogenase (G6pdh) locus in *D. melanogaster* and *D. simulans*.**

| | fixed | polymorphic |
|---|---|---|
| replacement | 21 | 2 |
| silent | 26 | 36 |

- in this example, 21 of 47 fixed differences between species (45%) are replacement changes.
- only 2 of the 38 polymorphisms observed in either species (5%) involve replacement changes.
- the G-test here is highly significant ($G = 19.0$, $P < 0.0001$).

- this example is similar to that of *Adh* in suggesting that there has been an excess number of replacement changes occur among these species!
- again, the most likely explanation for this excess is not random drift but selection.

**Tests for positive selection**

- one of the most important principles of molecular evolution is that the majority of amino acid mutations are deleterious.
- the rate of amino acid evolution for a particular gene is thus almost always lower than the rate of silent evolution.
- we know that natural selection is usually vigilant in removing deleterious mutations from populations
- the question is whether the amino acid changes observed within and between natural populations are neutral, weakly selected, or adaptive.
- one indisputable sign of positive selection is an elevated rate of amino acid replacement changes relative to silent changes.
- in comparing individual positions within a gene, we can identify two classes of sites.

1. synonymous, or silent, sites.
2. nonsynonymous, or replacement, sites.

- the former are sites where mutations will not cause a change in amino acid structure.
- the latter are positions where mutations result in changes in amino acid structure.
- the number of synonymous or nonsynonymous sites in a gene depends on the specific amino acid composition of the gene.
- most positions are nonsynonymous.
- in comparing the divergence of a gene among species, we can then estimate the rate of substitutions for both classes of sites

**let $d_S$ = rate of synonymous substitutions per synonymous site**
**let $d_N$ = rate of nonsynonymous substitutions per nonsynonymous site**

- $d_S$ and $d_N$ will differ under different types of selection:

| | |
|---|---|
| Purifying selection | $d_S > d_N$ |
| No selection | $d_S = d_N$ |
| Positive selection | $d_N > d_S$ |

- how common are examples where $d_N > d_S$?
- not very - perhaps a handful of examples including lysin in abalones, the ABS in MHC class I and II loci, S alleles in plants, surface coat proteins in pathogens.

**Tests for selective sweeps**

- a recently proposed test by Fay and Wu (called the H test) is aimed at detecting recent selective sweeps.
- a selective sweep occurs when an advantageous allele occurs at a locus and directional selection causes the allele to rapidly go to fixation.

- because this process occurs quickly, the selected allele will exhibit a reduced amount of polymorphism.
- this because the selected allele because it has achieved its high frequency in the population far more rapidly than expected by a neutral allele experiencing drift.
- this test requires information from one or more outgroup species to be able to infer ancestral or derived alleles.
- if this information is available, the test appears to work well.

- a similar test for a selective sweep is to survey its predicted effects in a large chromosomal region.
- this has been called the long-range haplotype test.
- the logic behind this test is similar to the H test in that a recent sweep is expected to result in a loss of nucleotide polymorphism (mainly silent) in a chromosome.
- the window of decreased polymorphism is expected to be determined by the degree of recombination.
– if recombination is low, then the impact of selection will be far reaching and cause a large region of reduced polymorphism.
- however, if the degree of recombination is high, then the size of the region will be much smaller.
- with enough markers in a region (like SNPs), it is possible to detect windows of reduced polymorphism on large scales (i.e., tens of thousands of base pairs) that can only be produced by sweeps.

- it is important to note that balancing selection has exactly the opposite effect.
- the region surrounding a site experiencing balancing selection is expected to exhibit an increase in polymorphism (mainly silent) because the alleles persist for times exceeding that expected for a neutral allele.
- the region of elevated polymorphism is again determined by the extent of local recombination.
- if balancing selection occurs a region of low recombination, then it will elevated levels of silent polymorphism a fair distance away from the selected site.
- in regions of high recombination, the elevated window of polymorphism is expected to be very narrow.
- this appears to be the case for the *Adh* locus in *D. melanogaster*.
- there is an elevated level of polymorphism in the region around the mutation in the gene giving rise to the fast-slow allozyme polymorphism.
- however, this disappears about 200 bp away from this selected site.

**The molecular clock revisited**

- in Wednesday's class, I discussed the fact that the molecular clock was originally used by Kimura as evidence favoring the neutral theory.
- a more detailed study of DNA sequences has undermined the idea of a molecular clock.

- the substitution of amino acids in proteins over long evolutionary time periods is predicted to follow a Poisson distribution.
- a Poisson distribution is a discrete frequency distribution that describes the number of times a rare event occurs.
- for an event to be distributed in a Poisson fashion two features must be met.
- first, the Poisson variable must exhibit a small mean relative to its maximum possible rate in a given sampling period.
- second, the occurrence of the event must be independent of past events in the sampling period.

- in other words, the substitution of an amino acid at a site in a protein must have no effect on the probability or likelihood of further substitutions.
- these two conditions mean that “rare and random” events should be distributed in a Poisson fashion.

- let N(t) be our Poisson variable, the total number of amino acid substitutions at a locus over a period of t years.

$$N(t) = \text{total number of amino acid substitutions observed over a period of t years}$$

- the Poisson distribution is given by:

$$\text{Prob}\,(N(t) = i) = \frac{e^{-\lambda t}(\lambda t)^i}{i!}$$

- where $\lambda$ is the mean rate of substitution observed for the locus under study.
- this equation thus predicts the probability that 1, 2, or more amino acid substitutions will occur in a protein over a period of t years.

- one of the most important characteristics of the Poisson distribution is that its mean is equal to its variance.
- therefore, the ratio of the variance to the mean is expected to equal 1.
- this is sometimes called the "index of dispersion", estimated by R.

$$R = S^2/M,$$

where $S^2$ is the variance and M is the mean.

- the basis of the molecular clock is that molecular evolution is constant over time.
- if this rate is constant and follows a Poisson distribution, the expected value of R is 1.0.
- if, however, the rate of evolution fluctuates over time, the value of R should exceed 1 because the variance becomes greater than the mean.
- so here we have a powerful way to test the neutral theory by examining the index of dispersion for a number of proteins.

- we can go further, however, and compare values of R for **silent** and **replacement** mutations.
- not only does the neutral theory predict that R should be 1, it also predicts that R estimated for silent and replacement positions should be similar.
- if values of R are found to exceed 1, then molecular evolution proceeds in a non-clocklike fashion.
- if the values of R estimated for silent and replacement sites differs, it may suggest that the evolutionary forces acting on these sites differs.
- this would also contradict the neutral theory.

- these are the predictions to test, now just what group of taxa to examine?
- the ideal group would represent what is called a “star phylogeny”.
- a star phylogeny is composed of a number of taxa that have radiated from a similar ancestor at similar times.

- an excellent star phylogeny to use are various mammal orders - artiodactyls, rodents, and primates which have diverged from a common ancestor in the late Mesozoic about 60-65 mya.
- comparisons can be made among homologous proteins in various species belonging to the star phylogeny.
- the expectation is that the rate of evolution should be similar among the different branches.
- the analysis proceeds as follows.
- construct a matrix of the mean number of amino acid differences (M) observed among pairs of species in the star phylogeny.
- then we must correct for multiple substitutions (for the mammal groups being compared, this has a minor effect on the estimates).
- we then compute variance for the number of substitutions ($S^2$).
- we then calculate $R = S^2/M$.

- detailed study of many proteins following the seminal work of Zuckerkandl and Pauling proposed the molecular clock have shown that rates of evolution vary considerably.
- for example, the rate of substitution of the insulin gene in rodents varies by a factor of 30, primate hemoglobins by a factor of 10, and primate cytochrome c by a factor of 25.
- what are values of R?
- the most extensive data has been obtained for primates, rodents and artiodactyls (pigs, sheep, etc.) and has been analyzed by Gillespie (1991).

- the table below presents estimates of $R = S^2/M$ for 20 loci sequenced in primates, rodents and artiodactyls (from Gillespie 1991)

| Locus | Replacement | Silent |
|---|---|---|
| Prolactin | 4.32 | 1.02 |
| Parathyroid | 1.04 | 4.58 |
| Proenkaphalin | 2.18 | 9.05 |
| Proglucagon | 2.82 | 9.39 |
| alpha-globin | 1.62 | 4.68 |
| beta-globin | 1.78 | 0.73 |
| Thyrotropin B | 2.31 | 4.89 |
| POMC | 2.14 | 1.46 |
| Growth hormone | 60.25 | 17.10 |
| GPHA | 34.67 | 2.52 |
| Luteinizing B | 12.85 | 2.16 |
| Relaxin | 3.29 | 0.28 |
| Interleukin-2 | 8.84 | 17.19 |
| Signal peptides | 0.82 | 7.52 |
| CCK | 0.21 | 1.20 |
| ACHRG | 2.20 | 3.71 |
| UPA | 19.64 | 0.25 |
| ANF | 2.53 | 2.17 |
| beta-crystallin | 4.39 | 0.36 |
| Na, K-ATPase | 5.50 | 2.48 |

| | | |
|---|---|---|
| MEAN | 8.67 | 4.64 |

- for replacement sites, the mean R for 20 genes is 8.67 (range from 0.21 to 60.25 for growth hormone).
- for silent sites, the mean R over the same loci is 4.64 (range 0.25 to 17.2 for interleukin-2).

**Conclusions**:

**1. the molecular clock is not very clocklike at all!**
- the index of dispersion - a measure of the variation in rate of molecular evolution - is too great to be accounted for by a strictly neutral model.
- proteins diverge over time, but the rate at which they diverge is not constant.

**2. the dynamics of silent and replacement sites are different.**
- this contradicts a fundamental prediction of the neutral theory.
- replacement substitutions are presumed to largely immune to selection, but the only conclusion we can draw from this result is that selection must be involved in affecting the rate of replacement changes.

**3. the clock, if it exists, is "episodic".**
- Gillespie's interpretation of the data is bursts of amino acid substitutions occur.
- these are interspersed with periods in which little evolution happens.
- is this pattern consistent with neutral theory?

**Current status of the neutral theory**

- the neutral theory enjoyed widespread support throughout the 1970's and 1980's.
- it is currently being strongly challenged by DNA sequence data from a wide diversity of organisms.
- this is leading to a re-appraisal of the role that natural selection may play in molecular evolution.
- back in the 1960's when Kimura proposed the neutral theory, he thought the number of substitutions being recorded among species were too great to be caused by natural selection.
- in fact, this is just what these new tests are discovering – a greater than expected number of replacement changes!
- is the rate of adaptive evolution at the molecular level that high?
- do advantageous mutations arise much more frequently than usually assumed?
- these appear to be the emerging questions in the field.