Evolutionary Theory and Principles of Phylogenetics

Lucy Skrabanek
ICB, WMC
March 19, 2008
Theory of evolution

- Evolution: process of change over time
  - 2 competing models
    - Phyletic gradualism
    - Punctuated equilibrium

- Phyletic gradualism
  - Gradual process

- Punctuated equilibrium
  - Long periods of *stasis* (little change) interspersed with *revolutions* (short bursts of rapid change)
• Three major aspects to the theory of evolution
  – Common descent: All organisms descended from a single ancestor (or ancestral gene pool)
  – Mutation: Manifestation of novel traits in a lineage
  – Mechanisms that cause some traits to persist, while others perish
Ancestry of organisms

• Traits shared between organisms
  – Morphological
    • Fossils
  – Genetic
    • All organisms have DNA/RNA as their genetic material
    • All organisms use the same genetic code (with some small variations)
    • Sequence similarity between phylogenetically close organisms
Novel traits by mutations

• Evolution is the process by which novel traits arise in populations and are passed on from generation to generation.

• Mutations are permanent changes to the genetic material
  – Germline: can be passed on to the next generation
  – Somatic cells: affect a single cell only

• “Driving force” of evolution
  – Create new alleles

• Gene duplication
  – Introduce new genes
Types of mutation

Micromutation
  local base changes: substitutions, insertions, deletions

Macromutation

Heterozygosity

- Diploid nuclei have two alleles at every locus on homologous chromosomes (one on each chromosome)
- Homozygous: the two alleles are the same
- Heterozygous: the two alleles are different
- Heterozygosity: the state of being a heterozygote
- Heterozygosity (population genetics): the fraction of individuals in a population that are heterozygous at a particular locus
Linkage

• Alleles nearby on a chromosome tend to be propagated together
  – Measured by the co-occurrence of two alleles (linkage disequilibrium; LD)
  – A set of alleles commonly co-propagated are called a haplotype

• Recombination during meiosis may separate previously linked alleles
  – Allows mutations to be propagated independently
  – Allows unfavorable mutations to disappear from the population, and favorable mutations to be fixed more easily
  – Depends on how close the alleles are on the same chromosome
Survival of novel traits

• Natural selection
  – Differential survival and reproduction rates

• Genetic drift
  – Random fluctuations in genetic material
    (not selection-driven)

• Gene flow
  – Exchange of genetic material by interbreeding and migration
Natural selection

• Based on the fitness of an individual in a particular environment
• Organisms that survive and reproduce pass on their genetic material with greater frequency than those that do not survive
• Organisms that are more “attractive” reproduce more
• Consequences
  – Eliminates deleterious mutations
  – Increases frequency of beneficial mutations
  – Maintains variation in a population
    • Heterozygote advantage (e.g., sickle-cell anemia, cystic fibrosis)
    • Frequency-dependent selection (rare alleles have higher fitness; as they become more common, their fitness decreases)
Genetic drift

• Under the Neutral Theory of Evolution (Kimura)
  – Given a population of two equally fit states, over time one state will become extinct

• Genetic drift: describes the changes in allele frequency from one generation to the next due to sampling variance.
  – The frequency of an allele in a population will vary according to the probability distribution of the frequency of the allele in the parent generation
  – Over time, allele frequencies will tend to increase or decrease, eventually going to 0% or 100%
Genetic drift contd

• Fluctuations in allele frequency between generations may result in some alleles disappearing from the population
• At large population sizes, the number of generations needed for either allele to become ‘fixed’ is $2N$ ($N =$ population size)
• More important in small populations, since chance fluctuations between generations can make a bigger difference
Gene flow

• (Also called *admixture* or *migration*)
• Transfer of genetic material from one population to another
  – Migration of populations
• Effect:
  – Two populations become genetically more similar, while increasing their variation
• Speciation occurs when gene flow is impeded by non-geographic obstacles
• Interspecies transfer
  – Gene transfer: Moving genetic material across species boundaries (e.g., horizontal gene transfer)
Functional restraint

• Genes whose products are essential to survival pick up mutations very slowly
• Different regions of the genome accumulate changes at very different rates
  – Reflect the extent to which they are functionally constrained
Synonymous vs. nonsynonymous

• Synonymous:
  – $K_s$ : rate of synonymous substitution
  – Mutations which do not change the coding sequence
  – Usually at the third site in the codon
  – $K_s$ tends to reflect the actual substitution rate

• Nonsynonymous:
  – $K_a$ : rate of nonsynonymous substitution
  – Mutations which change the coding sequence
  – Under natural selection
  – $K_a$ is much more variable than $K_s$
  – $K_a$ usually much lower than $K_s$
What is phylogenetics?

Visualize relationships among objects

Li and Graur, 1997

Species tree

Gene tree
Phylogenetics - beyond the obvious

Maximum parsimony tree of 12S and 16S rRNA genes

Stanhope et al, MPE, 1998
Uses

- Gene family identification
- Inferring gene function
- Identifying origins of genetic disease
- Characterization of polymorphisms
Simple example

- Seq1: AGA
- Seq2: AGG
- Seq3: TGA
- Seq4: TGG

Possible number of unrooted trees: \((2n - 5)!\) == 3x5x…x(2n-5)
Possible number of rooted trees: \((2n - 3)!\)

OR

AND 13 OTHER TREES
Some terminology

- Tree, topology
  - Root
  - Branches
  - Nodes
- Rooted versus unrooted
  - outgroup
- Homologs
  - Orthologs
  - Paralogs
  - Xenologs
- Homoplasy
  - Convergence
  - Parallelism
  - Reversal
Orthologs vs. paralogs
But… bacterial evolution

- Lateral or horizontal gene transfer

  ![E. coli and S. typhimurium Transfer](image)

- How does the mechanism of lateral gene transfer affect the concept of a tree of life, and the concept of prokaryote species?
Buzzword central:
(how to sound as if you know what you’re talking about)

• OTU = operational taxonomic unit
• UPGMA = unweighted pair group method using arithmetic mean
• FM = Fitch Margoliash
• NJ = neighbor joining
• ME = minimum evolution
• MP = maximum parsimony
• ML = maximum likelihood
• JC = Jukes Cantor
• K2P = Kimura 2-parameter
Approaches

• Distance-based methods
  – UPGMA
  – Fitch-Margoliash
  – Neighbor-joining

• Sequence-based methods
  – Maximum parsimony
  – Maximum likelihood

• Probabilistic methods
UPGMA
(Unweighted Pair Group Method using Arithmetic mean)

- Construct a distance matrix
  - Genetic
  - Biochemical
  - Morphological
- Correct for multiple substitutions
  - Jukes-Cantor
  - Kimura
- Assumes a molecular clock

<table>
<thead>
<tr>
<th>OTU</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>$d_{AB}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>$d_{AC}$</td>
<td>$d_{BC}$</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>$d_{AD}$</td>
<td>$d_{BD}$</td>
<td>$d_{CD}$</td>
</tr>
</tbody>
</table>
Distance calculations

- Hamming distance: count number of mismatches between pairs of sequences

<table>
<thead>
<tr>
<th>Species</th>
<th>Human/cow</th>
<th>Human/rat</th>
<th>Cow/rat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of mismatches</td>
<td>6</td>
<td>9</td>
<td>11</td>
</tr>
</tbody>
</table>

Human PPY: atggctgccgcacgtctctgccctctccccctgtgctgtcctgtgctctctgt
Cow PPY: atggctgccgctgcaccgctgcctcttctgctccttttctctgt
cow PPY: atggcctcgcatactactgcctctccccctgttttctctcat
Percent difference

- Count fraction of sites where aligned sequences differ

<table>
<thead>
<tr>
<th>Species</th>
<th>Human/cow</th>
<th>Human/rat</th>
<th>Cow/rat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6/40</td>
<td>9/40</td>
<td>11/40</td>
</tr>
<tr>
<td>Fraction of</td>
<td>= 0.15</td>
<td>= 0.225</td>
<td>= 0.275</td>
</tr>
<tr>
<td>different</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sites</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Human PPY: atggctgccgcacgcctctgctccctcttccttgctgctcctgt
Cow PPY: atggctgccgcgcaccgctgcctctttcctgtctccctgt
Rat PPY: atggccgtcgcatatactactgcctctccctgttttctcctat
Another method: convert a similarity score (e.g., from Smith-Waterman) into a distance score
- Shuffle sequences, align and score
  - Do this many times and get an average score for unrelated sequences
- Align sequences to themselves
  - Average the two self-assigned scores (for each pairwise comparison), and get a maximum score
- Normalized similarity score:
  - Score = \((S_{\text{actual}} - S_{\text{random}}) / (S_{\text{max}} - S_{\text{random}})\)

Note! Dissimilarity \(\neq 1 - \text{Score}\)
Substitutions

- Expected number of substitutions is expected to follow the Poisson distribution
- The Poisson distribution is a discrete distribution
- Commonly used to model the number of random occurrences of some rare phenomenon in a specified unit of space or time

\[ p(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad \text{for } \lambda > 0 \text{ and } x = 0, 1, 2... \]

\( \lambda \) = mean number of events per unit time
\( x \) = number of events

http://astronomy.swin.edu.au/~pbourke/analysis/distributions/
Similarity $\rightarrow$ Dissimilarity

- Some substitutions in an alignment represent one change, some may be sequential changes.
- To account for possible multiple substitutions, use the Poisson distribution:
  - $\lambda = \text{dissimilarity} = \text{mean number of substitutions}$
  - From $p(x) = \frac{e^{-\lambda} \lambda^x}{x!}$ for $\lambda > 0$ and $x = 0, 1, 2...$
  - The probability of zero changes $P(0) = e^{-\lambda}$
  - Score (proportional to the probability of one or more changes) $= 1 - e^{-\lambda}$
  - And taking the logs of both sides and rearranging $\lambda = -\ln(1 - \text{Score})$
UPGMA

\[
d_{(AB)C} = \frac{d_{AC} + d_{BC}}{2} \\
d_{(AB)D} = \frac{d_{AD} + d_{BD}}{2}
\]
UPGMA contd

**Table: OTU Distances**

<table>
<thead>
<tr>
<th>OTU</th>
<th>AB</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0.6</td>
<td>0.66</td>
</tr>
</tbody>
</table>

**Formula:**

\[d_{(ABC)D} = \frac{d_{AD} + d_{BD} + d_{CD}}{3}\]

**Diagram:**

- **Node A:** 0.05
- **Node B:** 0.05
- **Node C:** 0.2
- **Node D:** 0.31
- **Leaf AB:** 0.15
- **Leaf ABC:** 0.15
- **Leaf BD:** 0.11
- **Leaf CD:** 0.2
- **Leaf D:** 0.62
Distance calculation corrections
(compensate for multiple substitutions and reversions)

- **Jukes-Cantor model**
  - There is the same probability of change at each sequence position, and once a position has been mutated, it is just as likely to change again

- **Kimura two-parameter model**
  - Takes into account that transitions should occur more often than transversions

- **Kimura three-parameter model**
  - Extends the two-parameter model to also distinguish between $A \leftrightarrow T / G \leftrightarrow C$ and $A \leftrightarrow C / G \leftrightarrow T$ transversions
Jukes-Cantor

• Simplest model of correction
  – There is the same probability of mutation at every position
  – Once a site has mutated, it still has the same probability that it did before of changing again
  – For equal base frequencies
    \[ f_A = f_G = f_C = f_T \]
    \[ K_{AB} = -\frac{3}{4} \ln[1 - \frac{4}{3} d_{AB}] \]
  – Unequal base frequencies
    \[ K_{AB} = -B \ln[1 - \frac{d_{AB}}{B}] \]

where \( B = 1 - (f_A^2 + f_G^2 + f_C^2 + f_T^2) \)

e.g., 2 sequences of length 30, which differ by 3 sites, assuming equal base frequencies:
\[ d_{AB} = \frac{3}{30} = 0.1 \]
\[ K_{AB} = -\frac{3}{4} \ln \left[ 1 - \frac{4}{3} (0.1) \right] = 0.107 \]
### Jukes Cantor derivation

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1-3α</td>
<td>α</td>
<td>α</td>
<td>α</td>
</tr>
<tr>
<td>C</td>
<td>α</td>
<td>1-3α</td>
<td>α</td>
<td>α</td>
</tr>
<tr>
<td>G</td>
<td>α</td>
<td>α</td>
<td>1-3α</td>
<td>α</td>
</tr>
<tr>
<td>T</td>
<td>α</td>
<td>α</td>
<td>α</td>
<td>1-3α</td>
</tr>
</tbody>
</table>

Assume rates of substitution between bases are equal, denoted by $\alpha$.
- Substitution rate = $3\alpha$
- Expected number of substitutions per site = $6\alpha t$

Probability that two seqs differ at a site:
$P_{\text{diff}} = \frac{3}{4} \left( 1 - e^{-8\alpha t} \right)$

Rearranging:
$\alpha t = -\frac{1}{8} \ln \left( 1 - \frac{4}{3} P_{\text{diff}} \right)$

Multiply both sides by 6:
$P_{\text{sub}} = -\frac{3}{4} \ln \left( 1 - \frac{4}{3} P_{\text{diff}} \right)$

Estimate expected number of observed substitutions by actual number ($d_{AB}$):

$P_{\text{sub}} = -\frac{3}{4} \ln \left( 1 - \frac{4}{3} d_{AB} \right)$

$P_{\text{diff}} = \text{expected number of observable substitutions}$

$P_{\text{sub}} = \text{expected number of actual substitutions per site}$

Probabilities of having C at time t given C as starting base, and not C as starting base, respectively:

$P_c(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$

$P_c(t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}$

Approximate as a continuous time model and solve the first order linear differential equation.
Kimura two-parameter model

- Assumes that transitions have a different frequency from transversions
- Assumes equal base frequencies
- Calculate proportion of transitions and transversions between the two sequences

\[
a = \frac{1}{1 - 2d_{ABts} - d_{ABtv}} \\
b = \frac{1}{1 - 2d_{ABtv}}
\]

\[
K_{AB} = \frac{1}{2} \ln(a) + \frac{1}{4} \ln(b)
\]

e.g., 2 sequences of length 30, which differ by 3 sites (2 transitions, 1 transversion), assuming equal base frequencies:

Non-corrected distance: \(d_{AB} = 3/30 = 0.1\)
Jukes-Cantor correction: \(K_{AB} = -3/4 \ln [1 - 4/3 (0.1)] = 0.107\)
Kimura 2-parameter: \(K_{AB} = 1/2 \ln (1.2) + 1/4 \ln (1.07) = 0.107\)
Distance correction

- The JC, K2P and K3P methods can also be used to correct protein distances.

- Jukes-Cantor:

  \[ B = 1 - \sum (f_{aai})^2 \]

- K2P

  \[ K = -\ln (1 - D - 0.2 D^2) \]

  Where:
  
  - \( D = \frac{\text{number of exact matches between the seqs}}{\text{total number matched residue in the alignment}} \)
A molecular clock?

• For a given protein, the rate of sequence evolution is approximately constant across lineages (Zuckerkandl and Pauling, 1965)

• Often used to estimate speciation events

• Relative rate test
  \[ K_{AC} - K_{BC} \approx 0 \]
Fitch-Margoliash

• Find the most similar pair, group all other sequences and recalculate the distance from each sequence in the pair to the average of all the other sequences

• Solve algebraically:

(Eq1) $a + b = \text{distance between } a \text{ and } b$
(Eq2) $a + c = \text{distance between } a \text{ and } c$
(Eq3) $b + c = \text{distance between } b \text{ and } c$
(Eq4) $Eq3 - Eq2 = b + c - a - c = b - a$
$Eq4 + Eq1 = b - a + a - b = 2b$
Divide result by 2 = $b$
And substituting into 1, get $a$
And then get $c$ from either 2 or 3

• Group the pair and recalculate distance matrix as for UPGMA

• Repeat until all sequences have been put on the tree
Neighbor-joining

• Start off with a star-shaped topology

• Calculate the sum of all branch lengths for this tree
  – Add all distances, divide by (N - 1) since each length is used N - 1 times
NJ contd

- Iteratively combine all pairs of sequences to find the pair that minimizes the (new) sum of the branch lengths
- Use the FM method to calculate the branch lengths of the chosen pair
- Group the pair and calculate a new distance matrix
- Repeat until all sequences are in the tree
Maximum parsimony

• Occam’s Razor
  – The best tree is the one which requires the least number of substitutions

• Have to check every possible topology!

Entia non sunt multiplicanda praeter necessitatem
Pluralitas non est ponenda sine necessitate
- William of Occam (c. 1285-1349)
Maximum Parsimony method

1. Construct multiple alignment
2. Identify informative sites
3. Identify best topology for each informative site
4. Identify best topology overall

- Sequences should be few and not too variable
- Informative sites are not conserved and should have the same character in at least two sequences
- Select the topology or topologies with the least number of changes (weighted versus traditional)
- Select the topology or topologies with the least number of changes overall
Maximum Likelihood

- Similar to maximum parsimony
- Allows for introduction of distance correction methods such as JC or K2P
  - Usable for more diverse sequences than MP
- Assumes a model of evolution

<table>
<thead>
<tr>
<th>Base</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$-\mu(a\pi_C + b\pi_G + c\pi_T)$</td>
<td>$\mu a\pi_C$</td>
<td>$\mu b\pi_G$</td>
<td>$\mu c\pi_T$</td>
</tr>
<tr>
<td>C</td>
<td>$\mu g\pi_A$</td>
<td>$-\mu(g\pi_A + d\pi_G + e\pi_T)$</td>
<td>$\mu d\pi_G$</td>
<td>$\mu e\pi_T$</td>
</tr>
<tr>
<td>G</td>
<td>$\mu h\pi_A$</td>
<td>$\mu j\pi_C$</td>
<td>$-\mu(h\pi_A + j\pi_C + f\pi_T)$</td>
<td>$\mu f\pi_T$</td>
</tr>
<tr>
<td>T</td>
<td>$\mu i\pi_A$</td>
<td>$\mu k\pi_C$</td>
<td>$\mu l\pi_G$</td>
<td>$-\mu(i\pi_A + k\pi_C + l\pi_G)$</td>
</tr>
</tbody>
</table>

$\mu$ = mutation rate (number of substitutions per unit of time)
$a$-$l$ = frequency of change of one base to any other
$\pi$ = base frequency
ML contd

• Start off with an alignment
• Substitutions in each column made into a tree
  – Probability of each tree is the product of the mutation rates in each branch ($= \text{product of rate of substitution in each branch by the branch length}$)
  – $L = p (\text{data} \mid \text{tree, branch lengths, model})$
    • All possible ways of getting the data are taken into account
• Logs of the likelihoods of each column in the alignment for each tree are added to give the likelihood for that tree for the whole alignment
Long branch attraction

• Happens for all methods except ML
• Occurs when sequences are very divergent
• Longer branches become artificially grouped because the number of non-homologous similarities between the sequences exceeds the number of homologous similarities that the sequences have retained with their true closest homologues
Bootstrapping

- For more divergent sequences, where an optimal alignment is not always obvious, where multiple changes may have taken place in any given column
- Randomly choose vertical columns from the aligned sequences to produce an alignment of the same length (some columns may be used more than once, some none at all)
- Redo analysis
- Repeat above two steps, usually 100 or 1000 times
- Branching topology should occur in at least >70% of the resampled trees to be considered significant (95% confidence)
Jack-knifing

- Similar to bootstrapping, except that some fraction of the aligned columns is taken, and no column can be sampled more than once in any given resampling
- The jack-knife value is the percentage of resampled alignments that give you back that internal branch
Bremer support index

• Also called:
  – Decay index
  – Branch support
  – Length difference
  – Decay analysis
• Estimates the reliability of phylogenetic groupings
• Usable with maximum parsimony only
• Example: The most parsimonious tree groups A, B and C together, and consists of 189 changes. The next most parsimonious tree that does not group A, B and C together has 214 changes. The Bremer support index for the grouping of A, B and C is 214-189=25
# Method comparison

<table>
<thead>
<tr>
<th></th>
<th>UPGMA</th>
<th>NJ</th>
<th>MP</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Branch lengths</td>
<td>Branch lengths (may be negative)</td>
<td>No branch lengths</td>
<td>Branch lengths</td>
<td></td>
</tr>
<tr>
<td>Rooted</td>
<td>Unrooted</td>
<td>Unrooted</td>
<td>Unrooted</td>
<td></td>
</tr>
<tr>
<td>Molecular clock</td>
<td>n/a</td>
<td>n/a</td>
<td>Explicit model of evolution</td>
<td></td>
</tr>
<tr>
<td>fast</td>
<td>fast</td>
<td>slow</td>
<td>slow</td>
<td></td>
</tr>
<tr>
<td>Bad with very divergent sequences</td>
<td>Bad with very divergent sequences</td>
<td>Good for closely related sequences</td>
<td>Can be used with any set of related sequences</td>
<td></td>
</tr>
<tr>
<td>Long branch attraction</td>
<td>Long branch attraction</td>
<td>Long branch attraction</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>Not really biologically meaningful</td>
<td>Very fast, pretty accurate</td>
<td>Quite accurate</td>
<td>Most accurate, very slow</td>
<td></td>
</tr>
</tbody>
</table>
When do I use what?

Multiple sequence alignment (CRITICAL STEP)

- Strong sequence similarity?
  - Yes: Maximum parsimony
  - No: Somewhat similar sequences?

- Somewhat similar sequences?
  - Yes: Distance methods
  - No: Maximum likelihood

- Maximum likelihood

Bootstrap

From Bioinformatics, Mount, p247
Outgroups

• Used to root trees
  – Root is where the outgroup connects to the rest of tree

• Choice of outgroup
  – Should be any OTU not descended from the same common ancestor as the rest of the taxa in the tree, but should be closely related
  – Outgroups too closely related could be part of the tree
  – Outgroups too distantly related may not align well, or may be subject to long branch attraction
Warning!

• The trees that we construct are unlikely to exactly reflect reality
  – Unknown model of evolution
  – Variability in mutation rates
    • Between sites in the sequence
    • Between species used in the analysis
    • At different time points
  – Recombination or horizontal transfer may have occurred
Conservation genetics

Hybrid between grey wolf and coyote

Dusky seaside Sparrow
*(Ammodramus maritimus nigrescens)*

Red wolf
HIV strains

• HIV-1
  – Chimpanzee

• HIV-2
  – sooty mangabey
    (eaten and/or kept as pets)
HIV-2

- Not as prevalent as HIV-1
- Seven subtypes (A - G)
  - A: Senegal and Guinea-Bissau
  - B, G: Ivory Coast
  - C, D, E, F: Sierra Leone and Liberia
- Subtypes more related to SIVsm strains found in each area than to other SIVsm strains
  - Different subtypes are the result of multiple independent cross-species transmissions
HIV-1

- HIV-1 is more diverse than HIV-2
  - 3 main groups - M, N, O
    - M (main/major)
      - 8 subtypes in group M
    - N (non-M/non-O)
    - O (outlier)
  - Chimpanzee subspecies: *Pan troglodytes troglodytes*
  - Estimate of origin of HIV-1 group M radiation ~ 1930
  - Subtypes of HIV similar to subtypes of SIVcpz
  - Origins of subtypes can be localized to specific parts of Africa
HIV-1 recombination

– Many recombinant forms
  • Leads to the possibility of already infected humans being infected with more distant subtypes of SIV, recombination occurring, and those ‘new’ strains of HIV spreading (N?)
  • Causes potential problems with live-attenuated virus vaccines
Peeters et al, AIDS, 2003

(also see Sharp et al, Phil Trans R Soc Lond, 2001)