From previous lectures:

- binomial and multinomial probabilities
- Hardy-Weinberg equilibrium and testing HW proportions (statistical tests)
- estimation of genotype & allele frequencies within population
- maximum likelihood
- methods used to detect and observe genetic variation:
  - ✓ 1960s-1970s: genetic variation was first measured by protein electrophoresis (e.g., allozymes).
  - ✓ 1980s-2000s: genetic variation measured directly at the DNA level:
    - Restriction Fragment Length Polymorphisms (RFLPs)
    - Minisatellites (VNTRs)
    - DNA sequence
  - ✓ 1990s-2000s: PCR based methods and high-throughput genotyping:
    - Cleaved Amplified Polymorphism (CAP)
    - Single-stranded Conformation Polymorphism (SSCP)
    - Microsatellites (SSRs, STRs)
    - Random Amplified Polymorphic DNAs (RAPDs)
    - Amplified Fragment Length Polymorphisms (AFLPs)
    - Single Nucleotide Polymorphisms (SNPs)
  - ✓ **2007-now: Genotyping via high-throughput massively parallel sequencing**

Today (29 марта, среда):

- **how to measure and quantify genetic variation**

Levels of genetic variation for a single gene, multiple genes or an entire genome

- within individuals
- between individuals
- within populations
- between populations
- over the entire set of populations
- between different taxa (species, genera, families, etc.)

Descriptive measures of genetic variation

- **Polymorphism (*P*)**: proportion or % of loci or nucleotide positions showing more than one allele or base pair
- **Heterozygosity (*H*)**: proportion or % of heterozygous loci per individual, or proportion or % of individuals that are heterozygotes in a population
- **Allele/haplotype diversity (*h*)**: measure of number and diversity of different alleles/haplotypes within a population
- **Nucleotide diversity (π, Θ, etc.)**: measure of number and diversity of variable nucleotide positions within sequences of a population
- **Synonymous (K_S, d_S, etc.)** or **nonsynonymous substitutions (K_A, d_N, etc.)**: % of nucleotide substitutions that do not or do result in amino acid replacement
- **Genetic distance (*d*, *D*, etc.)**: measure of similarity or dissimilarity between two homologous sequences, individuals or populations
- **Genetic differentiation (G_{ST}, F_{ST}, R_{ST}, Φ_{ST}, etc.)**: measure of subdivision, differences among homologous sequences, individuals or populations

How to quantify genetic variation?

I. **Within individual variation:** Individual heterozygosity (observed proportion of heterozygose loci: $H_i = N_{\text{het loci}}/N_{\text{loci studied}}$

II. **Within population variation:**
- Proportion of polymorphic loci ($P = N_{\text{polymorphic loci}} / N_{\text{loci studied}}$; <99% or <95% of the most common allele criteria)
- Average number of alleles per locus (A)
- Effective number of alleles per locus ($A_e = 1/\Sigma p_i^2$)
- Heterozygosity - observed (H_O) and expected (H_E) , also referred to as "genetic diversity":
  - for 2 alleles: $H_E = 2p_1p_2$
  - for any number of alleles: $H_E = 1 - \Sigma p_i^2$
- Deviations from Hardy-Weinberg expectations (per locus and population)
- Inbreeding or fixation index $F = (H_e - H_o)/H_e = 1 - H_o/H_e$
- Nucleotide diversity (Θ and π)
- Assessment of non-random association of non-allelic genes or linkage disequilibrium (D, D', r^2, etc.)
- Estimates of N_e, effective population size (e.g., indirect from $\Theta = 4N_e\mu$)
- Pairwise individual genetic similarity or distance, allele-sharing indexes, relatedness

III. **Total variation over the entire set of populations:**
- P, A, A_E, H, and F are calculated with all the samples considered to constitute a single group.

IV. **Among population variation:**
- Differences among populations in P, A, A_E, and H. (Does one or more populations have unusually high or low values for any of the above?)
- F_{ST}, G_{ST}, R_{ST} – genetic variance measures. Hierarchical, if appropriate.
- Heterogeneity and differences in allele frequencies among populations
- Patterns of variation: clinal, ecotypic, and latitudinal correlations, etc.
- Assignment tests (how well do individuals match the population in which they were sampled?)
- Genetic distances (Cavalli-Sforza's, Nei's, etc.)
- Correlation between genetic distance and geographic distance (Mantel tests)
- Estimates of gene flow, effective population size ($\Theta = 4N_em$)
- Cluster analysis, phylogenetic tree-building
- Multivariate Statistics - Principal Components, Principal Coordinate and Factor Analysis, Multidimensional scaling
- Assessment of whether partitions (subpopulation structure) exist in the data (Bayesian approaches, tree-building analyses)

Main basic measures of genetic variation

- ***H***, heterozygosity
- ***F***, the inbreeding coefficient
- ***Θ*** and ***π***, nucleotide diversity
- ***I***, genetic identity
- ***D***, genetic distance
- F_{ST}, population subdivision

Heterozygosity

- H_o - observed heterozygosity (proportion heterozygotes) within a subpopulation.
- H_E - expected heterozygosity within a subpopulation. If there are ***i*** different alleles at a locus in a subpopulation, p_i is the frequency of the ***i***th allele:

$$H_E = 2\sum_{i,j} p_i p_j = 1 - \sum_i^n p_i^2$$

- Unbiased estimate: $H_E = \frac{2N}{2N-1}(1 - \sum_i^n p_i^2)$
- H_T - expected heterozygosity across the entire metapopulation assuming random mating: $H_T = 1 - \sum_i \overline{p_i}^2$ ($\overline{p_i}$ - average frequency of the ***i***th allele)
- Mean over ***m*** loci: $H = \frac{1}{m}\sum_{j=1}^{m} H_j$
- Sampling variance: $V_H = \frac{H(1-H)}{Nm}$

Inbreeding coefficient

- the inbreeding coefficient ***F*** (*aka* "fixation index") can measure the reduction in heterozygosity due to inbreeding
- in a simple two-allele system with inbreeding $H_o = 2pq - 2pqF = H_E - H_EF$ therefore $F = (H_E - H_o)/H_E = 1 - H_o/H_E$
- $\chi^2 = F^2N$ (df =1)

Nucleotide polymorphism or diversity

- **Expected nucleotide polymorphism or diversity under neutral theory of molecular evolution: Θ** (theta) = $4N_e u$
- **Θ can be estimated**

1) from observed number of segregating (polymorphic) sites divided by sequence length (***S***) in a sample of ***n*** sequences under neutral theory of molecular evolution:

$$\theta = \frac{S}{\alpha_1} \qquad a_1 = \sum_{i=1}^{n-1} \frac{1}{i},$$

where α_1 is an evolutionary coefficient

2) as an average observed proportion of nucleotide differences between all different pairs of sequences in the sample, π (pi):

$$\pi = \frac{1}{n(n-1)} \sum_{ij} p_{ij} \ ,$$

where p_{ij} is the proportion of different nucleotides between the ***i***th and ***j***th types of DNA sequences, and ***n*** is the number of sequences in the sample. The summation is taken over all ***ij*** pairs without repetition. Resampling (such as bootstrapping) is recommended for analysis of variance for Θ and π

Nucleotide sequence variation

Example: Assume that the following nucleotide sequences were found in four individuals:

```
ACTTGTCCTA
AGTTCTCCTA
TCTTGACGTA
AGTTGACCTA
```

1) from observed number of segregating (polymorphic) sites (***S***): $\theta = \frac{S}{\alpha_1}$

number of segregating substitution sites divided by sequence length $S = 5/10 = 0.5$

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i} = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} = 1.833 \qquad \theta = \frac{S}{\alpha_1} = \frac{0.5}{1.833} = 0.273$$

2) as an average observed proportion of nucleotide differences:

seq#	1	2	3	4
1		2/10	3/10	2/10
2	2/10		5/10	2/10
3	3/10	5/10		3/10
4	2/10	2/10	3/10	

$$\pi = \frac{\sum_{ij} p_{ij}}{n(n-1)}$$

$$\pi = \frac{0.2+0.3+0.2+0.2+0.5+0.2+0.3+0.5+0.3+0.2+0.2+0.3}{4\times(4-1)} = 0.283$$

Allele and SNP discovery: Douglas-fir case study

Isolation of mRNA

↓cDNA synthesis and either cloning for Sanger sequencing or direct multiplex sequencing↓

Before 2007: Generation of cDNA clonal library and Sanger sequencing

After 2007: High-throughput massively parallel multiplex sequencing

↓ sequencing

Generation of EST sequence library, contigs and unigenes

↓ deposition ↕ annotation via BLAST

Genbank

↓

Selection of gene of interest

↓

Design of PCR primers & amplification of a subsample

↓

Direct sequencing of amplicon or sequencing after cloning PCR product

(direct sequence of PCR amplified DNA from 24 or 32 individual megagametophytes representing unrelated Douglas-fir trees from 6 regions in Washington and Oregon)

↓

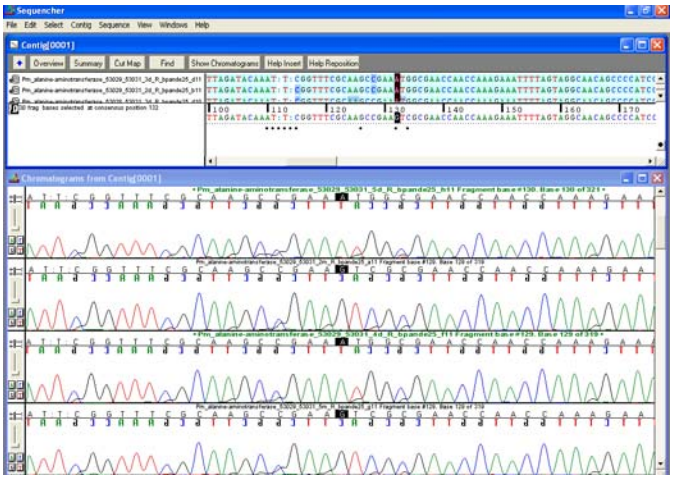
Sequence processing (quality control, editing, alignment)

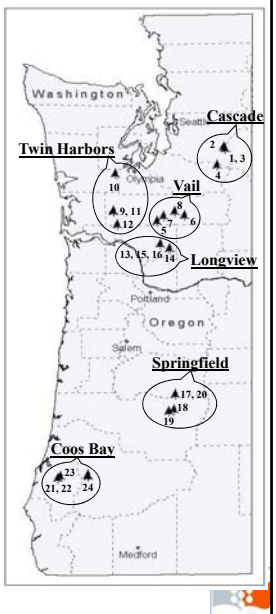
↓

Nucleotide polymorphism analysis and genotyping

Allele and SNP discovery

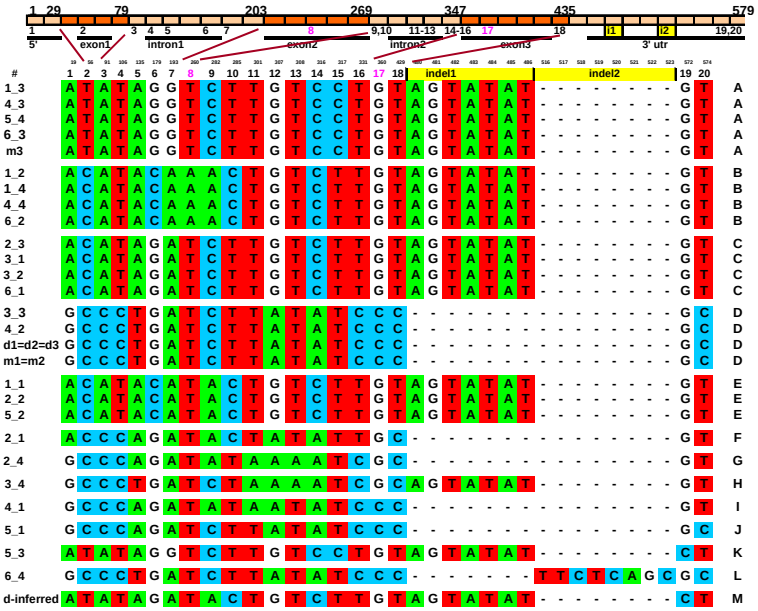
Direct sequence of PCR amplified DNA from 24 individual megagametophytes representing unrelated Douglas-fir trees from 6 regions in Washington and Oregon





Metallothionein-like (MT) protein (complete)

20 single nucleotide polymorphic (SNP) sites and 2 indels in 13 haplotypes representing 24 individual *Pseudotsuga menziesii* trees from 6 regions and two parents

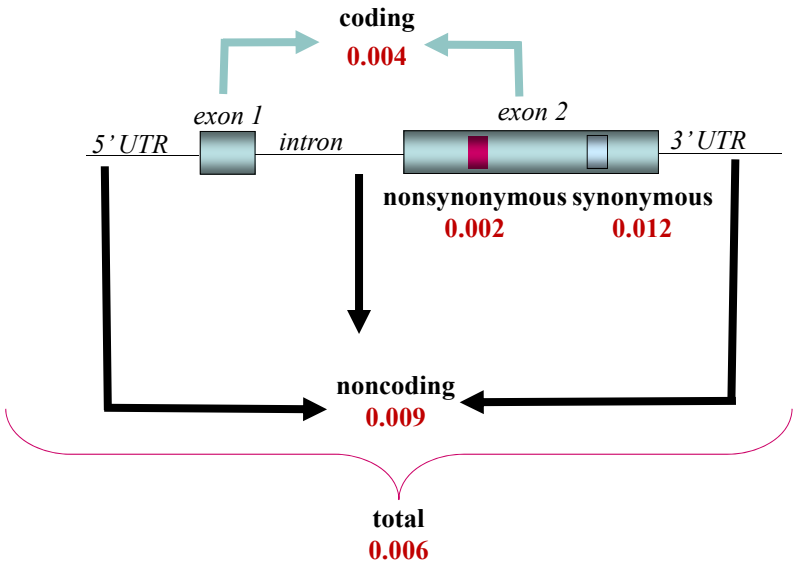


Protein (68 a/a):
MSSDGKDCGCADPTOCDKKGNSLGVEMVETSYDYNKNMSFGFEYEMETVAAENGCKSGASSKYSNRCN

Nucleotide diversity in 20 Douglas-fir candidate genes

Gene	Total sites, bp	SNPs	bp per SNP	Pars. SNPs	*h*	π	Θ	Tajima's *D*
EF1A	1072	14	77	9	0.940	0.00274	0.00339	-0.656
TBE	2954	58	51	36	0.963	0.00516	0.00626	-0.723
F3H1	365	14	26	4	0.690	0.00528	0.00988	-1.576
F3H2	647	14	46	12	0.828	0.00629	0.00562	0.150
Formin-like	337	3	112	3	0.585	0.00480	0.00229	1.498
AT	2578	93	28	66	0.966	0.00936	0.00935	-0.037
LEA-II	504	18	28	13	0.884	0.00647	0.00878	-0.862
MT-like	579	20	29	20	0.907	0.01334	0.00911	1.639
60S-RPL31a	609	21	29	18	0.701	0.01011	0.00891	0.479
LEA-EMB11	545	33	17	26	0.950	0.01378	0.01594	-0.593
40S-RPS3a	500	12	42	10	0.810	0.00601	0.00617	-0.336
PolyUBQ	898	17	53	15	0.840	0.00544	0.00494	0.357
ERD15-like	646	14	46	12	0.598	0.00438	0.00563	-0.757
ABA-WDS	344	9	38	5	0.825	0.00662	0.00672	-0.048
LP3-like	481	16	30	13	0.866	0.00662	0.00848	-0.713
CHS	762	11	69	5	0.569	0.00281	0.00371	-1.011
4CL-1	628	8	79	3	0.841	0.00268	0.00316	-0.460
4CL-2	629	10	63	7	0.814	0.00237	0.00378	-1.128
ADF	634	2	317	0	0.140	0.00023	0.00081	-1.511
APX	867	26	33	17	0.884	0.00636	0.00789	-0.700
Mean	**829.0**	**20.7**	**40**	**14.7**	**0.780**	**0.00604**	**0.00654**	**-0.349**
Total	**16579**	**413**		**294**				

Mean nucleotide diversity π in different gene regions for 20 Douglas-fir genes



Mean nucleotide diversity (π and Θ per site) in different nucleotide sites or gene regions for 20 Douglas-fir genes

Sites	π	Θ
all	0.00604	0.00654
coding	0.00424	0.00460
noncoding	0.00925	0.01044
nonsynonymous	0.00194	0.00240
synonymous	0.01187	0.01238
silent	0.00979	0.01068

Comparative nucleotide diversity

Species	No. loci	θ_T (total per nucleotide site)	θ_C (per site in coding regions)	θ_{NC} (per noncoding site including introns and untranscribed regions)	θ_S (per synonymous site in coding regions)	θ_{NS} (per nonsynonymous site in coding regions)	Reference
Human [a]	75	8 ± 2	8 ± 2	9 ± 2	15 ± 4	6 ± 1	Halushka *et al.* 1999
	106	5 ± 1	5 ± 1	5 ± 1	12 ± 3	3 ± 1	Cargill *et al.* 1999
Soybean	143	5 ± 2			10 ± 4	4 ± 2	Zhu *et al.* 2003
Douglas-fir	20	65 ± 27	46 ± 22	107 ± 46	124 ± 60	24 ± 17	Krutovsky & Neale 2005
Drosophila [a]	24	70 ± 58	40 ± 31	105 ± 80	130 ± 92	15 ± 14	Moriyama & Powell 1996
Maize	21	96 ± 32	72 ± 25	111 ± 37	173 ± 61	39 ± 14	Tenaillon *et al.* 2001

θ values are multiplied by 10^4

[a] as compiled in Zwick *et al.* (2000)

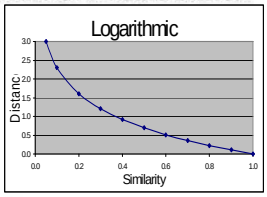
Genetic identity and distance

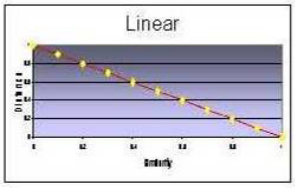
- **Genetic identity (*I*) or similarity (*S*) and distance (*D*)** are measures of the similarity and dissimilarity, respectively, of genetic material between different individuals, populations or species.
- The genetic identity and distance between two individuals, populations, species or any samples are described as the proportion of genetic elements (alleles, genes, gametes, genotypes) that they share or do not share, respectively.
- ***D*** = 0 (***I*** or ***S*** = 1) when two samples are absolutely identical; ***I*** or ***S*** = 0 when they have no genetic elements in common (***D*** = 1 or $\rightarrow\infty$).

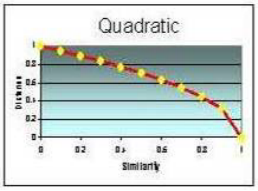
Genetic similarity and distance

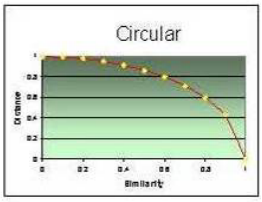
Depending on the similarities of individuals, **four** representation types of distance (D) are possible:

- D = 1 – S, known as linear distance, because it assumes that the relationship with similarity is linear.
- D = √(1 – S), known as quadratic distance because it assumes that the relationship with similarity follows a quadratic function, so that, to make it linear, the square root must be calculated.
- D = √(1 – S²), known as circular distance.
- **D = –ln(S), known as logarithmic distance**





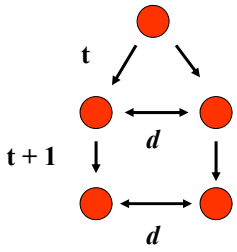




Genetic identity and distance

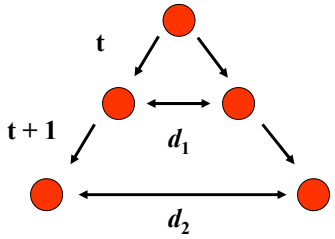
Calculation of distance, or dissimilarity, follows one of two possible models:

Equilibrium model



Distance remains constant over time (equilibrium exists between migration and genetic drift)

Disequilibrium model



Distance changes with time through migration and genetic drift

Two alternative distances exist for the disequilibrium model

- **Geometric distance**
  - does not take into account evolutionary processes
  - based only on allele frequencies
  - divergence time cannot be directly inferred from distance
- **Genetic distance**
  - takes into account evolutionary processes
  - distance increases from the time of separation from an ancestral population
  - a genetic model of evolution is needed

When should we use **geometric** or **genetic distance**?

- **Geometric distance** is used in studies of closely related individuals, accessions or populations. It can be used with any markers, but often is used with dominant markers (RAPDs, AFLPs) whose molecular evolution is unknown. Because evolutionary aspects are not considered, the dendrograms obtained cannot be interpreted as phylogenetic trees giving information about evolution or divergence among groups.
- **Genetic distance**, in contrast, considers evolutionary models and can be incorporated into phylogeny studies. It can be used with both codominant and dominant markers, although, with the latter, information is incomplete.

Disequilibrium models: geometric distance

- This measures the direct relationship between the similarity index (***S***) and distance (***D*** = 1 — ***S***)

- Different variables are possible, for example:
  - ✓ binary variables (e.g., RAPD, AFLP, SNPs)
  - ✓ quantitative variables
  - ✓ mixed types of variables

Similarity coefficients for binary variables (e.g., RAPD, AFLP, SNPs)

	Author	*Expression (S =)*	*Example of the coefficient value if a = 3, b = 1, c = 3, d = 2*
S1	Russel and Rao (1940)	a/n	0.333
S2	Simpson	a/min[(a + b),(a + c)]	0.750
S3	Braun-Blanquet	a/max[(a + b),(a + c)]	0.500
S4	***Dice (1945); Nei and Li (1979)***	***a/[a + (b + c)/2]***	***0.600***
S5	Ochiai (1957)	a/[(a + b)(a + c)]1/2	0.612
S6	Kulczynski 2	(a/2)([1/(a+b)] + [1/(a+c)])	0.625
S7	***Jaccard (1900, 1901, 1908)***	***a/(a + b + c)***	***0.429***
S8	Sokal and Sneath 5 (1963)	a/[a +2(b + c)]	0.273
S9	Kulczynski 1 (1928)	a/(b + c)	0.750
S10	***Sokal and Michener (1958)***	***(a + d)/n***	***0.556***
S11	Rogers and Tanimoto (1960)	(a + d)/[a + d + 2(b + c)]	0.385
S12	Sokal and Sneath 1 (1963)	(a + d)/[a + d + (b + c)/2]	0.714
S13	Sokal and Sneath 3 (1963)	(a + d)/(b + c)	1.250

Simple Matching (S10), Jaccard (S7) and Nei-Li (S4) are the most common indices

		Indiv.j		
		1	0	
Indiv.i	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	n

Disequilibrium models: genetic distance

- Measures the difference between two genes, proportional to the time of separation from a common ancestor
- Several models are possible:

Mutation of infinite alleles (e.g., Nei's genetic distance, allozymes, SNPs).

Each mutation event gives rise to a new allele.

- If 2 genes are the same, no mutation has occurred. If 2 genes are different, the **mean number of mutations, μ** since time t when they diverged from an ancestor is $\mu = 2ut$, where u is the rate of mutation and is multiplied by **2** because we are dealing with **2** independent genes.
- The Poisson probability that no mutations (i=**0**) has occurred in 2 genes (or they are identical) after time t is $P_i = (e^{-\mu}\mu^i)/i!$ and assuming that mean number of mutations (μ) in any particular time (t) interval $\mu = 2ut$, then $P_0 = [e^{-2ut}(2ut)^0]/0! = e^{-2ut}$). Then, probability of mutation $P_m = 1-P_0 = 1-e^{-2ut}$, $2ut$ is divergence or distance between two sequences, let call it d, and observed proportion of nucleotide difference, let call it p, is a proxi of P_m, then $p = 1-e^{-d}$, and $d = -\ln(1-p)$

Stepwise mutation model (e.g. distance using microsatellites)

- **Alleles mutate in a step-wise fashion** (in multiples of the repeat unit). Alleles that are closer in size are more related than alleles that show larger size differences.
- In the case of microsatellites (*aka* SSRs), mutation is assumed to change the number of repeats, increasing or decreasing step by step. It can be shown that the square of the difference in the number of repeats between 2 microsatellites is proportional to the time of divergence from a common ancestor.

Mutation in the nucleotide sequence

- Some methods assume the probability of transition (purine →purine or pyrimidine → pyrimidine) and transversion (purine → pyrimidine or pyrimidine → purine)

23

p-distance for nucleotide and amino acid sequence data

- If nucleotide or amino acid sequences are available, then the proportion (p) of different amino acids or nucleotides between sequences can be used for comparing of sequence divergence $p = S/N$, where S is the number of different (segregating) sites, and N is the total number of sites
- This proportion is called the p-distance
- If sites are subject to substitution with equal probability, then S follows the binomial distribution, and, therefore, the variance of p is given by $V_p = p(1-p)/N$

Nei M. & Kumar S. 2000 Molecular Evolution and Phylogenetics. Oxford University Press, New York.

p-distance for nucleotide and amino acid sequence data

- However, the relationship between divergence time ***t*** and ***p***-distance is nonlinear due to multiple repetitive amino acid or nucleotide substitutions occuring at the same site, so the discrepancy between observed ***S*** and the actual number of substitutions gradually increases.
- Poisson correction (**PC**) helps to address this problem.

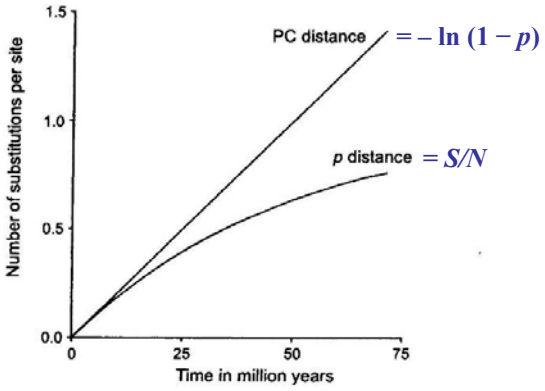


FIGURE 2.2. Relationships of the p distance and the Poisson correction (PC) distance with time. The rate of amino acid substitution (r) is assumed to be 10^{-8} per site per year. (Nei M, & Kumar S. 2000 Molecular Evolution and Phylogenetics. Oxford Univ. Press)

p-distance for nucleotide and amino acid sequence data

- The concept of the Poisson distribution helps to estimate the number of substitutions more accurately: $P = e^{-\mu}\mu^i/i!$, where μ – mean, i – number of occurrences
- If u is the rate of amino acid or nucleotide substitution (mutations) per year or generation, then the mean number of amino acid or nucleotide substitutions μ after a period of t years or generations is ut ($\mu = ut$)
- Then, the probability of occurrence of i amino acid or nucleotide substitutions ($i = 0, 1, 2, 3, ...$) is given by $P = e^{-ut}(ut)^i/i!$, where ut – mean number of substitutions
- If no substitutions have occurred, then $i = 0$ and $P(0;t) = e^{-ut}(ut)^0/0! = e^{-ut}$
- No substitutions for two sequences $(e^{-ut})(e^{-ut}) = (e^{-ut})^2 = e^{-2ut}$
- Respectively, probability of any substitutions $p = 1 - e^{-2ut}$

p-distance for nucleotide and amino acid sequence data

- Respectively, probability of any substitutions

$$p = 1 - e^{-2ut}$$

- ***2ut*** is divergence or distance between two sequences, let call it ***d***
- Then, $p = 1 - e^{-2ut} = 1 - e^{-d}$
- Let's solve it for ***d:***

$$1 - p = e^{-d}$$

$$\ln(1 - p) = \ln e^{-d}$$

$$\ln(1 - p) = -d \ln e$$

$$\ln(1 - p) = -d$$

$$d = -\ln(1 - p)$$

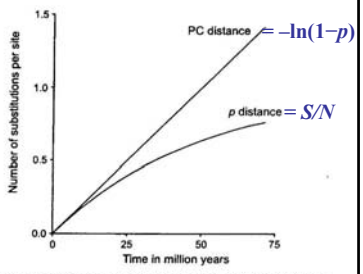


FIGURE 2.2. Relationships of the p distance and the Poisson correction (PC) distance with time. The rate of amino acid substitution (r) is assumed to be 10^{-9} per site per year.

p-distance for nucleotide and amino acid sequence data

- ***d*** in $d = 2ut$ (divergence or distance between two sequences) can be inferred from $d = -\ln(1 - p)$
- If we know ***d*** and time of divergence (***t***) between two sequences from other information (for instance, paleontological), then we can infer mutation or amino acid substitution rate $u = d/(2t)$
- Likewise, if we know mutation or substitution rate (***u***) from other information, then we can infer time of divergence $t = d/(2u)$

(Nei & Kumar 2000)

p-distance for nucleotide and amino acid sequence data

Table 3.1 Sixteen different types of nucleotide pairs between sequences X and Y.

Class	Nucleotide Pair				
Identical nucleotides	AA	TT	CC	GG	Total
Frequency	O_1	O_2	O_3	O_4	O
Transition-type pair	AG	GA	TC	CT	Total
Frequency	P_{11}	P_{12}	P_{21}	P_{22}	P
Transversion-type pair	AT	TA	AC	CA	
Frequency	Q_{11}	Q_{12}	Q_{21}	Q_{22}	
	TG	GT	CG	GC	Total
Frequency	Q_{31}	Q_{32}	Q_{41}	Q_{42}	Q

bols given in Table 3.1. The total frequencies of identical pairs, transition-type pairs, and transversion-type pairs are denoted by O, P, and Q, respectively. Obviously, we have the relationship $p = P + Q$.

(Nei & Kumar 2000)

p-distance for nucleotide and amino acid sequence data

Table 3.2 Models of nucleotide substitution.

	A	T	C	G	A	T	C	G
	(A) Jukes-Cantor model				(E) HKY model			
A	-	α	α	α	-	βg_T	βg_C	αg_G
T	α	-	α	α	βg_A	-	αg_C	βg_G
C	α	α	-	α	βg_A	αg_T	-	βg_G
G	α	α	α	-	αg_A	βg_T	βg_C	-
	(B) Kimura model				(F) Tamura-Nei model			
A	-	β	β	α	-	βg_T	βg_C	$\alpha_1 g_G$
T	β	-	α	β	βg_A	-	$\alpha_2 g_C$	βg_C
C	β	α	-	β	βg_A	$\alpha_2 g_T$	-	βg_G
G	α	β	β	-	$\alpha_1 g_A$	βg_T	βg_C	-
	(C) Equal-input model				(G) General reversible model			
A	-	αg_T	αg_C	αg_G	-	$a g_T$	$b g_C$	$c g_G$
T	αg_A	-	αg_C	αg_G	$a g_A$	-	$d g_C$	$e g_G$
C	αg_A	αg_T	-	αg_G	$b g_A$	$d g_T$	-	$f g_G$
G	αg_A	αg_T	αg_C	-	$c g_A$	$e g_T$	$f g_C$	-
	(D) Tamura model				(H) Unrestricted model			
A	-	$\beta\theta_2$	$\beta\theta_1$	$\alpha\theta_1$	-	a_{12}	a_{13}	a_{14}
T	$\beta\theta_2$	-	$\alpha\theta_1$	$\beta\theta_1$	a_{21}	-	a_{23}	a_{24}
C	$\beta\theta_2$	$\alpha\theta_2$	-	$\beta\theta_1$	a_{31}	a_{32}	-	a_{34}
G	$\alpha\theta_2$	$\beta\theta_2$	$\beta\theta_1$	-	a_{41}	a_{42}	a_{43}	-

Note: An element (e_{ij}) of the above substitution matrices stands for the substitution rate from the nucleotide in the i-th row to the nucleotide in the j-th column. g_A, g_T, g_C, and g_G are the nucleotide frequencies. $\theta_1 = g_G + g_C$, $\theta_2 = g_A + g_T$.

(Nei & Kumar 2000)

Nei's Genetic Distance (1972) based on allele frequencies:

Infinite Alleles Model: each mutation event gives rise to a new allele, and number of alleles is infinite

The standard Nei's genetic distance is based on the concept of genetic identity:

$$I_{XY} = \frac{J_{XY}}{(J_X J_Y)^{1/2}} \qquad D_{XY} = -\ln(I_{XY})$$ where,

J_X = average homozygosity in population **X** = $^1/_m \sum_{m=1}^{m} \sum_{i=1}^{n} p_{ix}^2$

J_Y = average homozygosity in population **Y** = $^1/_m \sum_{m=1}^{m} \sum_{i=1}^{n} p_{iy}^2$

J_{XY} = average interpopulation homozygosity = $^1/_m \sum_{m=1}^{m} \sum_{i=1}^{n} p_{ix} p_{iy}$ (n - # alleles, m - # loci)

- Such that, $I_{XY} = 1$, if two populations have the same allele frequencies in all sampled loci ($J_X = J_Y = J_{XY}$), $I_{XY} = 0$, if two populations do not share the same allele frequencies in all sampled loci ($J_{XY} = 0$).
- The value of D_{XY} varies from **0** (where populations have identical allele frequencies) to **infinity** (∞, where populations do not share any alleles).
- It assumes that the rate of substitution per locus is equal among all loci and populations.
- This distance estimates the genetic differences per locus between two populations.