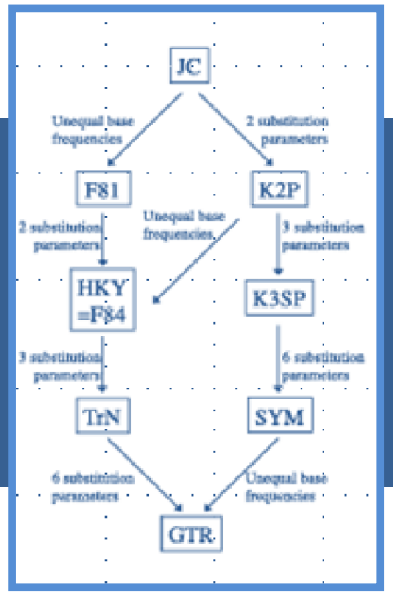
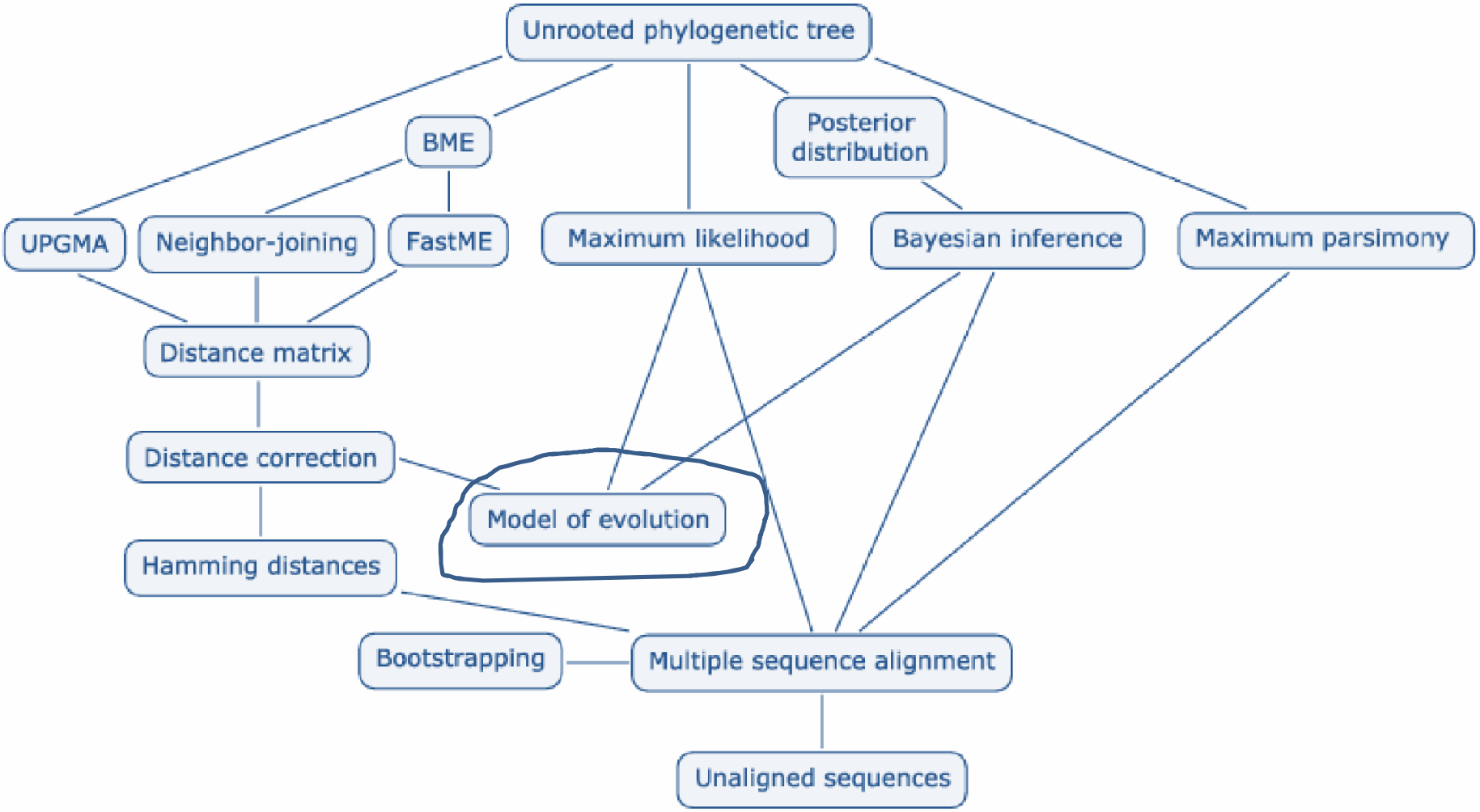
MODELLING SEQUENCE EVOLUTION



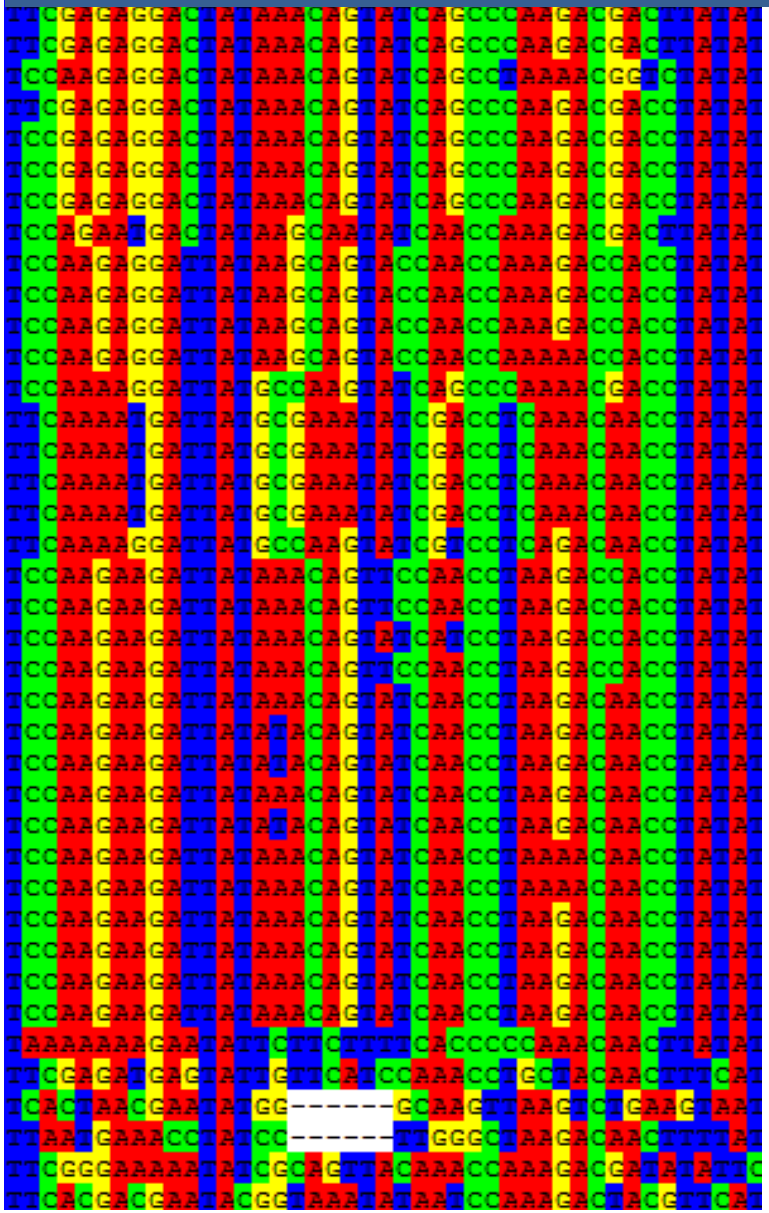
- Sequence differences that we observe now (among existing species or other items which are compared) are products of past mutation events, **substitutions**.
- Understanding the substitution process needs modelling.
- Historically, modelling started by the concept *molecular clock.*
- Parametrization: one parameter, *Jukes-Cantor model* → more realistic models
- Deleterious, non-deleterious, neutral mutations in genomes
    - Appendix: Markov models in general

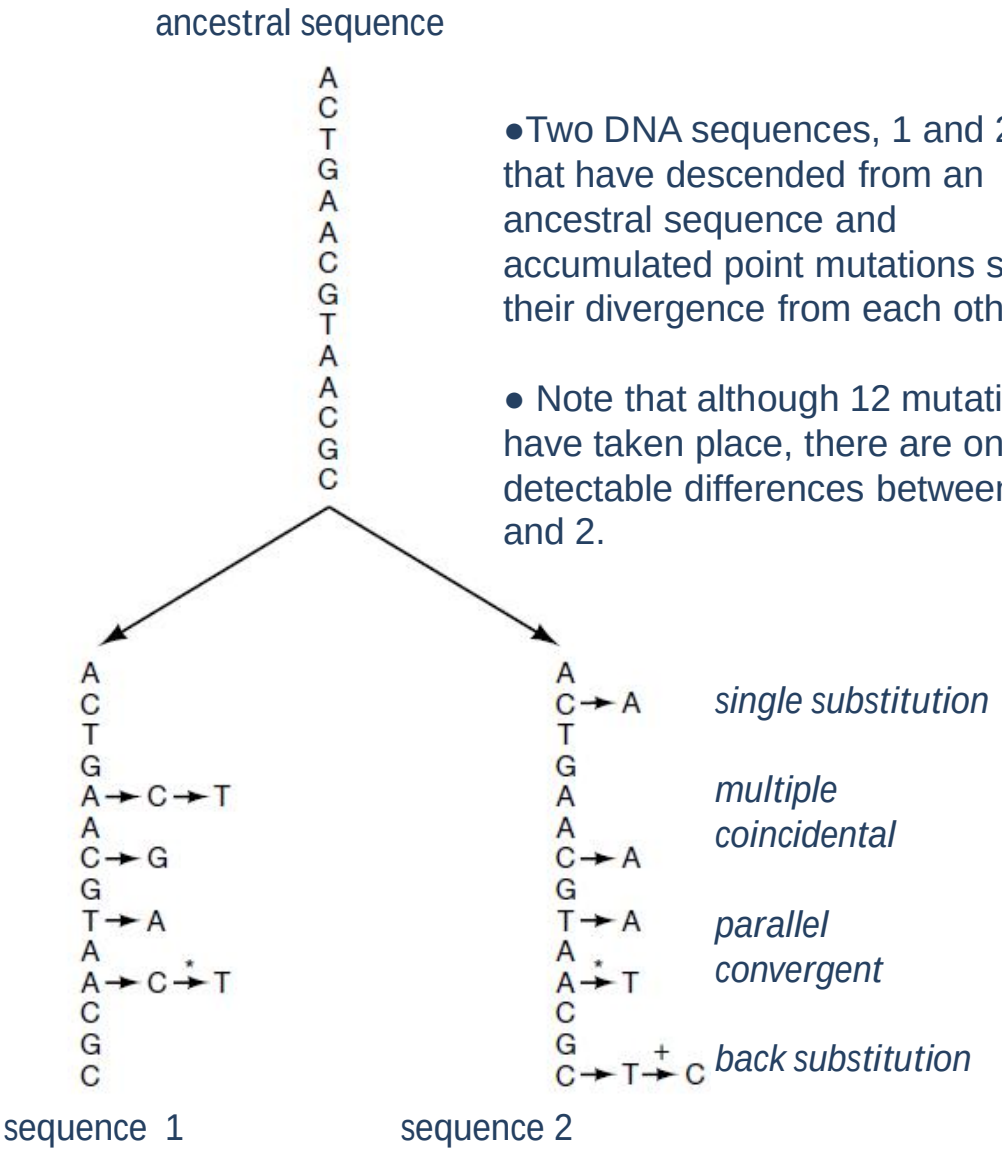
This is not a finalized version

ROLE OF MODELS IN PHYLOGENY RECONSTRUCTIONS



CURRENT DIFFERENCES ARE PRODUCTS OF PAST SUBSTITUTION EVENTS





- Two DNA sequences, 1 and 2, that have descended from an ancestral sequence and accumulated point mutations since their divergence from each other.
- Note that although 12 mutations have taken place, there are only 3 detectable differences between 1 and 2.

HISTORY: JUKES TELLS HOW THE IDEA AROSE

http://www.garfield.library.upenn.edu/classics1990/A1990CZ67100002.pdf

How Many Nucleotide Substitutions *Actually* Took Place?

Thomas H. Jukes
Department of Biophysics and Medical Physics
University of California
Berkeley, CA 94720

In 1965 I met Charles R. Cantor, who was 23 years old and was a graduate student in chemistry at the University of California. We started talking about molecular evolution and about comparing the polypeptide chains of homologous proteins. Charles said that a computer program should be written for searching for evidence of this, but that he was frightfully busy working on his PhD thesis, and he had no time for this. The next day he had written the program, and in 1966 we wrote two notes on its use. We resolved to write a textbook on molecular evolution together, and Charles left for Columbia University as an assistant professor in 1966. I received a request from Hamish N. Munro for a chapter in his forthcoming volume III of *Mammalian Protein Metabolism*. He asked us to write on "Evolution of protein molecules," and we sent him the manuscript that was to have been incorporated in our book. It was published in Munro's book in 1969, and the article has 110 printed pages. Citations to our long article relate only to the following short passage in it, written by Charles.

It can be shown that the mean number of base differences at a single position on the mRNA, μ, is related to the observed fraction of residues with single base differences, p, by the expression

$$\mu = \frac{3}{4} \ln \frac{3}{3-4p} \qquad (1)$$

The equation (1) assumes that all single base changes (nucleotide substitutions) are equally probable and that the frequencies of all four bases in DNA are the same. This gives me the chance to point out that (1) should be called the Cantor equation, not Jukes and Cantor. The formula came into wide use when rapid DNA and RNA sequencing became available. From then on molecular biologists became interested in comparing sequences of homologous genes to study evolution. For example, a portion of the two sequences of human α and β hemoglobin genes is

α gene A C C A A C G T C A A G G C C G
C C T G G G G T A A G G T T

β gene T C T G C C G T T A C T G C C C
T G T G G G G G A A G G T G

showing 12 nucleotide substitutions (40 percent). The mean number of substitutions that has actually occurred is greater than 12, because of revertants, such as A to C to A, and multiple changes, such as A to C to G. Equation (1) corrects for these, and the probable total number of substitutions is 17 (57 percent), *not* 40 percent.

The two genes diverged from a common ancestor at least 4×10^8 years ago. Sharks go back in the fossil record for 400 million years and sharks have α and β hemoglobins (but lampreys do not). The equation tells us that the average rate of substitution per year per nucleotide site is about $0.57 \div (4 \times 10^8) = 1.4 \times 10^{-9}$. We carry with us in every red blood cell the evidence that we are in a line of descent from an ancestor who lived 400 million years ago!

An example of the use of equation (1) is in the article by C.L. Manske and D.J. Chapman.[1] These authors used the equation to correct their comparisons of 5S ribosomal RNA sequences for rev[illegible] and convergent mutations. See also references 2 and 3 for similar usage.

Charles returned to Berkeley in 1989 to direct the human genome project at the Lawrence Berkeley Laboratory.

JUKES-CANTOR MODEL, ONE PARAMETER

● To study the dynamics of nucleotide substitution, assumptions on the probabilities of substitutions of one nucleotide by another are needed.

Assumption: all nucleotide substitutions occur with equal probabilites, α

● The rate of substitution for each nucleotide is 3α per unit time

	A	T	C	G
A		α	α	α
T	α		α	α
C	α	α		α
G	α	α	α	

● At time 0: Assumption that at a certain nucleotide site there is A, $P_{A(0)} = 1$
● Question: probability that this site is occupied by A at time t , $P_{A(t)}$?
● At time 1, probability of still having A at this site is

$$P_{A(1)} = 1 - 3\alpha \qquad (1)$$

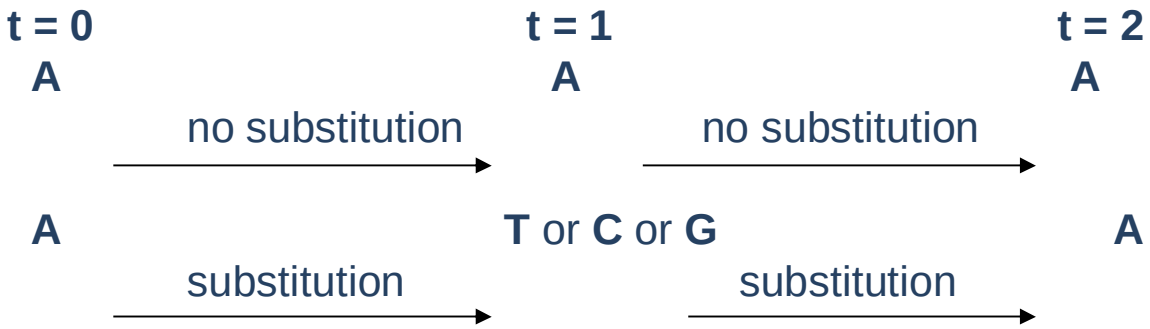
● 3α is the probability of A changing to T, C, or G

JUKES-CANTOR MODEL, ONE PARAMETER

- The probability of the site having A at time 2 is

$$P_{A(2)} = (1 - 3\alpha)P_{A(1)} + \alpha\,[1 - P_{A(1)}] \qquad (2)$$

- This includes two possible courses of events:



- The following recurrence equation holds for any t

$$P_{A(t+1)} = (1 - 3\alpha)P_{A(t)} + \alpha[1 - P_{A(t)}] \qquad (3)$$

Note that this holds also for $t = 0$, because $P_{A(0)} = 1$ and thus

$$P_{A(0+1)} = (1 - 3\alpha)\,P_{A(0)} + \alpha\,[1 - P_{A(0)}] = 1 - 3\alpha$$

which is identical with equation (1).

JUKES-CANTOR MODEL, ONE PARAMETER

- The amount of change in $P_{A(t)}$ per unit time, rewriting equation (3):

$$\Delta P_{A(t)} = P_{A(t+1)} - P_{A(t)} = -3\alpha P_{A(t)} + \alpha[1 - P_{A(t)}] = -4\alpha P_{A(t)} + \alpha \qquad (4)$$

- Approximating the previous discrete-time model by a continuous-time model, by regarding $\Delta P_{A(t)}$ as the rate of change at time *t.* With this approximation equation (4) is rewritten as

$$dP_{A(t)} / dt = -4\alpha P_{A(t)} + \alpha \qquad (5)$$

- The solution of this first-order linear differential equation is

$$P_{A(t)} = \tfrac{1}{4} + (P_{A(0)} - \tfrac{1}{4})e^{-4\alpha t} \qquad (6)$$

- The starting condition was A at the given site, $P_{A(0)} = 1$, consequently

$$P_{A(t)} = \tfrac{1}{4} + \tfrac{3}{4} e^{-4\alpha t} \qquad (7)$$

- Equation (6) holds regardless of the initial conditions, for example if the initial nucleotide is not A, then $P_{A(0)} = 0$, and the probability of having A at time t

$$P_{A(t)} = \tfrac{1}{4} + \tfrac{1}{4} e^{-4\alpha t} \qquad (8)$$

JUKES-CANTOR MODEL, ONE PARAMETER

- Equations (7) and (8) describe the substitution process. If the initial nucleotide is A, then $P_{A(t)}$ decreases exponentially from 1 to ¼ . If the initial nucleotide is not A, then $P_{A(t)}$ will increase monotonically from 0 to ¼ .

- Under this simple model, after reaching equilibrium, $P_{A(t)}=P_{T(t)}=P_{C(t)}=P_{G(t)}$ for all subsequent times.

- Equation (7) can be rewritten in a more explicit form to take into account that the initial nucleotide is A and the nucleotide at time t is also A

$$P_{AA(t)} = ¼ + ¾ e^{-4\alpha t} \qquad (9)$$

- If the initial nucleotide is G instead of A, from equation (8)

$$P_{GA(t)} = ¼ + ¼ e^{-4\alpha t} \qquad (10)$$

Since all the nucleotides are equivalent under the Jukes-Cantor model, the general probability, $P_{ij(t)}$, that a nucleotide will become *j* at time *t*, given that it was *i* at time 0, equations (9) and (10) give the general probabilities $P_{ii(t)}$ and $P_{ij(t)}$, where $i \neq j$.

$$P_{ii(t)} = ¼ + ¾ e^{-4\alpha t} \quad \text{and} \quad P_{ij(t)} = ¼ + ¼ e^{-4\alpha t} \qquad (11)$$

JUKES-CANTOR MODEL -> NUCLEOTIDE DIVERGENCE BETWEEN TWO SEQUENCES

- We assume that all sites in sequence evolve at the same rate and follow the same substitution scheme. The number of sites compared between two sequences is denoted by L.

- Consider the probability that a nucleotide at a given site at time t is the same in both sequences. Suppose that the nucleotide at a given site was A at time point 0. At time t, the probability that a descendant sequence will have A at this site is $P_{AA(t)}$, and consequently the probability that two descendant sequences have A at this site is $P^2_{AA(t)}$. Similarly, the probabilities that both sequences have T, C or G at this site are $P^2_{AT(t)}$, $P^2_{AC(t)}$, and $P^2_{AG(t)}$

- The probability that the nucleotide at a given site at time t is the same in both sequences is

$$I_{(t)} = P^2_{AA(t)} + P^2_{AT(t)} + P^2_{AC(t)} + P^2_{AG(t)} \qquad (12)$$

- From equations (11) we obtain

$$I_{(t)} = \tfrac{1}{4} + \tfrac{3}{4} e^{-8\alpha t} \qquad (13)$$

JUKES-CANTOR MODEL -> NUCLEOTIDE DIVERGENCE BETWEEN TWO SEQUENCES

- Equation (13) also holds for T, C or G. Therefore, regardless of the initial nucleotide at a given site, $I_{(t)}$ represents the proportion of *identical* nucleotides between two sequences that diverged t time units ago. The probability that the two sequences are *different* at a site at time t is $p = 1 - I_{(t)}$. Thus

$$p = \tfrac{3}{4}(1 - e^{-8\alpha t}) \quad \text{or} \quad 8\alpha t = \ln(1 - (4/3)\,p) \qquad (14)$$

- The time of divergence between two sequences is usually not known, and thus estimation of α is not possible. Instead, it possible to calculate K, which is the number of substitutions per site since the time of divergence between the two sequences. In the case of the one-parameter model, $K = 2(3\,\alpha t)$, where $3\,\alpha t$ is the number of substitutions per site in a single lineage

$$K = 6\,\boldsymbol{\alpha} t = -\tfrac{3}{4}\ln(1 - (4/3)\,p) \qquad (15)$$

where p is the observed proportion of different nucleotides between the two sequences.

An example. Page 3 (book chapter page 143) in *Phylogeny methods based on distance matrices* (see course webpage, week 1) shows how Jukes-Cantor model serves like a *correction* to sequence diverge calculation.

TWO PARAMETERS, KIMURA´S MODEL

- The Jukes-Cantor –model was introduced in 1969 when virtually nothing was known about nucleotide substitution
- **In 1980 Motoo Kimura proposed different parameters for transitions and transversions**.
- Transition is a nucleotide change between purines, A and G, and pyrimidines, T and C. Transversion is a purine – pyrimidine change.
- The rate of transition change is α and transversion change is β per unit time

	A	T	C	G
A		**β**	**β**	**α**
T	**β**		**α**	**β**
C	**β**	**α**		**β**
G	**α**	**β**	**β**	

(In Assignment 4 one of the exercises is to derive this model in a similar way that was used for one-parameter model. You have to consider four courses of events: no substitution, transition, two different transversions. You can also derive the model in some other way, if you like.)

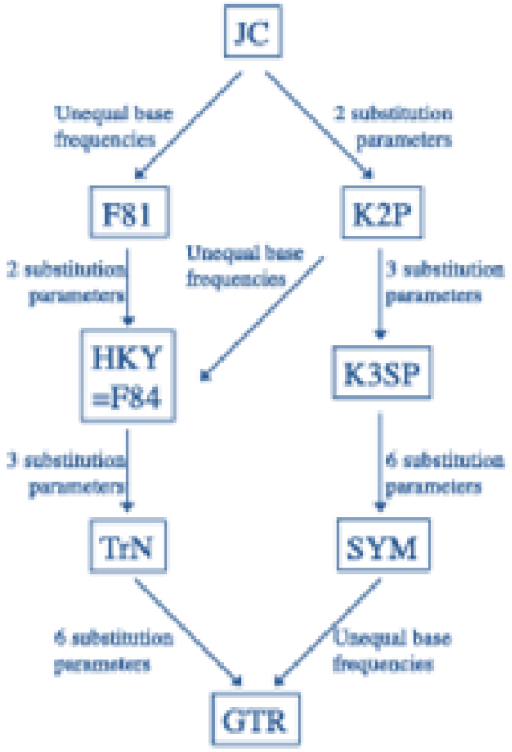
HISTORICAL LOOK AT NUCLEOTIDE CHANGE MODELLING

● Since 1980´s it has been known that misincorporation errors (mutations) during DNA replication or repair are facilitated if a base is replaced by similar one and thus **transitions** (purine replaced by a purine, or pyrimidine replaced by pyrimidine) occur more frequently than **transversions** (purine replaced by a pyrimidine or vv). Differences in mutation rate tend to decrease TA and CG dimers and to produce an excess of CT and TG dimers, and many other kinds of biased processes (cf. the constancy in the Jukes-Cantor model).

● The development of models of sequence evolution is an active field and there is a large number of models.

● Two main approaches to building models of sequence evolution: An *empirical* one, using properties calculated through comparisons of large numbers of observed sequences (for example, counting apparent replacements between many closely related sequences). Empirical models result in fixed parameter values which are estimated only once and then assumed to be applicable to other datasets (=> easy to use computationally). The alternative approach is to build models *parametrically* on the basis of chemical or biological properties of DNA and amino acids. For example, incorporating a parameter to describe the relative frequency of transition to and transversion substitutions in the sequences studied. Both methods result in **Markov process models (see the Appendix).**

FLOW-DIAGRAM OF THE MOST WIDELY USED SUBSTITUTION MODELS



JC	"Jukes-Cantor"
F81	"Felsenstein 81"
K2P	"Kimura 2-Parameter"
K3SP	"Kimura 3-Parameter"
HKY	"Hasegawa-Kishino-Yano"
F84	"Felsenstein 84"
TrN	"Tamura-Nei"
SYM	"Symmetric"
GTR	"General Time Reversible"

- Starting with the simple Jukes-Cantor model, more general models are obtained by allowing unequal nucleotide frequencies and/or more than one substitution parameter. The most general model of this type is the GTR model that allows unequal base frequencies and prescribes a different substitution parameter for each of the six pairs of different nucleotides.

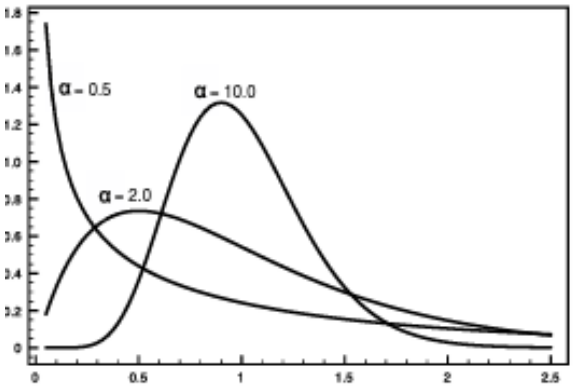
RATE HETEROGENEITY IN SUBSTITUTION PROCESS

● An important aspect in substitution process modelling is the consideration of heterogeneity of evolutionary rates among sites. The biological basis of heterogeneous mutation rate among sites may reflect the influence of the nearest neighbors on mutation rate. Stacking energies along the molecule, helix configuration (A, B, Z-DNA, triple helix), supercoiling, and DNA intrinsic curvature (that is sequence dependent) change the solvent accessibility and thus base reactivity. The fixation of any mutation depends on DNA and protein structure/function selection pressures (and on stochastic processes, of course). Protein coding and noncoding DNA regions show remarkably different mutation rates; moreover, each codon position is subject to different selection pressures.

● The incorporation of heterogeneity of evolutionary rates among sites has led to a new set of models that generally provides a better fit to observed data. Some authors have considered models in which a fraction of sites change at one rate, whereas the other sites are invariable. More popular and successful have been models based on a continuous distribution of rates. Modelling site rates using a Gamma distribution is a widely used approach. A continuous distribution in which every site may have a different rate seems to be the most biologically plausible model. It has been shown, however, that the **discrete Gamma model**, with as few as four categories of evolutionary rates chosen to approximate a Gamma distribution, performs very well. It is also very practical computationally.

RATE HETEROGENEITY IN THE SUBSTITUTION PROCESS

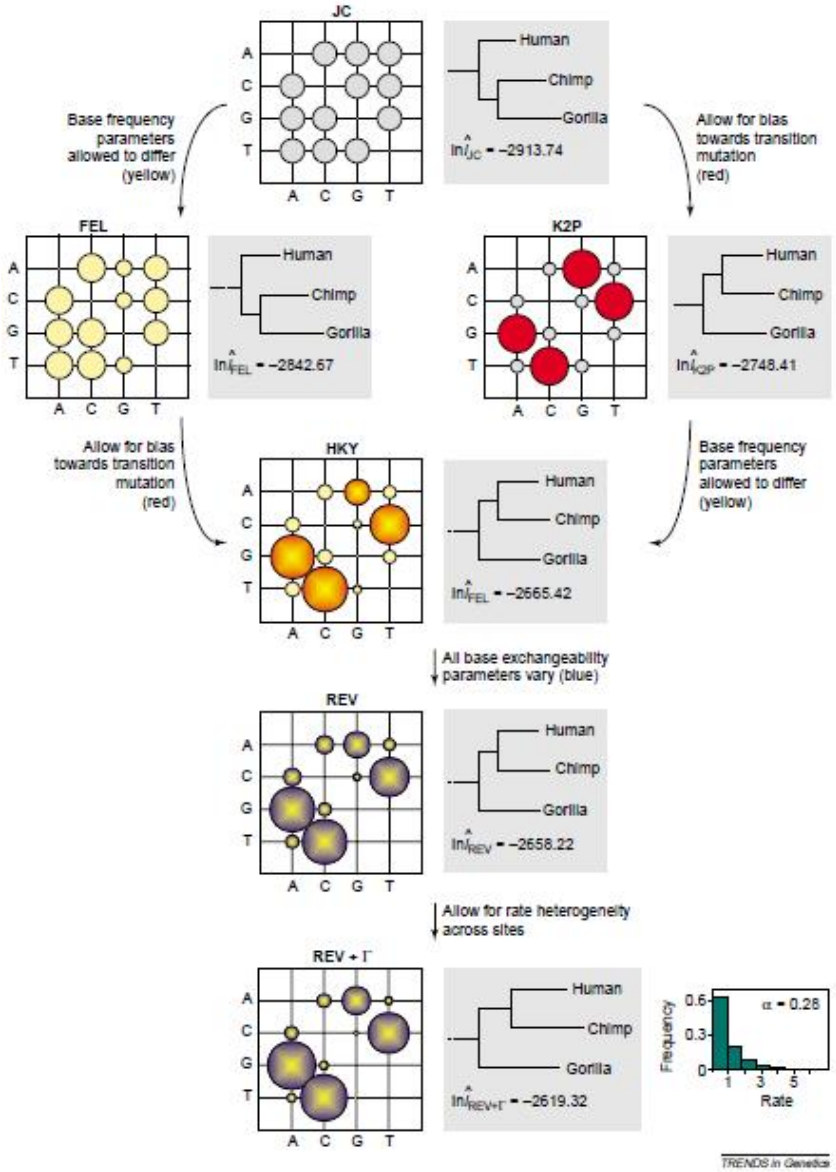
- When using Gamma distribution, it is assumed that the rate of substitution for each site is drawn from this distribution with shape parameter α.



- If α is <1, the distribution implies that there is a relatively large amount of rate variation, with many sites evolving very slowly but some sites evolving at a high rate.
- For values of α>1, the shape of the distribution changes qualitatively, with less variation and most sites having roughly similar rates.
- It is known that the range of distributional shapes available under the permitted values of $0 < \alpha < \infty$ is well able to describe the variation found in DNA sequences.

- Next page gives an example of comparisons of substitution models, including the gamma assumption. Maximum likelihood phylogeny inference (see below) is inferred, using different model assumptions for a given sequence dataset.

.RELATIONSHIPS AMONG SUBSTITUTION MODELS IN A PRACTICAL EXAMPLE



- The sequence studied is a part of mitochondrial genome. Mitochondrial sequences are known to have highly biased transitions vs. transversions.

- The models JC , FEL, K2P, REV, REV +Γ (the inferred shape parameter value is α=0.28) are presented in a flowchart showing relationships between them. For each model, the matrix of rates of substitutions between nucleotides is represented by a bubble plot where the area of each bubble indicates the corresponding rate. The models become more advanced moving down the figure, as illustrated in the bubble plots by their increasing flexibility in estimating relative replacement rates and as reflected by increasing log-likelihoods.

- For the REV+Γ model the reverse-J shape of the graph indicates that the majority of sites have low rates of evolution, with some sites having high rates of evolution.

- Note how the inferred maximum likelihood phylogeny changes significantly as the models become more advanced. (compare JC with K2P); inferred branch lengths also tend to increase (compare REV to REV+Γ). Arrows show where models are nested within each other; that is, where the first model is a simpler form of the next. For example, the JC model is nested within the K2P model (it is a special case arising when κ is fixed equal to 1), but the K2P model is not nested with the FEL model.

MUTATIONS IN GENOMES

● Various kind of mutations, the basic **evolutionary factors**, produce evolutionary raw material. Other evolutionary factors - recombination, natural selection and random drift - dictate the fates of mutations.

● Comparative approaches, involving data from multiple different species, are suitable for detecting past selection. One important tool used to detect selection from genome data is to compare the ratio of nonsynonymous mutations per nonsynonymous site to the number of synonymous mutations per nonsynonymous site (d_N/d_S) (see the genetic code).

● This set of slides is about current results on profiling, through an evolutionary ”telescope” , amino acid mutations in human genome data.

● Which amino acids can be replaced by which - through mutations? An old problem and extremely relavant in many kind of biological and medical questions! Next page shows the historical first step for quantifying the problem.

AMINO ACID DIFFERENCES - GRANTHAM DISTANCE

A classical paper, *Science 185:862-864, 1974*: Grantham distance.

Amino Acid Difference Formula to Help Explain Protein Evolution

Abstract. *A formula for difference between amino acids combines properties that correlate best with protein residue substitution frequencies: composition, polarity, and molecular volume. Substitution frequencies agree much better with overall chemical difference between exchanging residues than with minimum base changes between their codons. Correlation coefficients show that fixation of mutations between dissimilar amino acids is generally rare.*

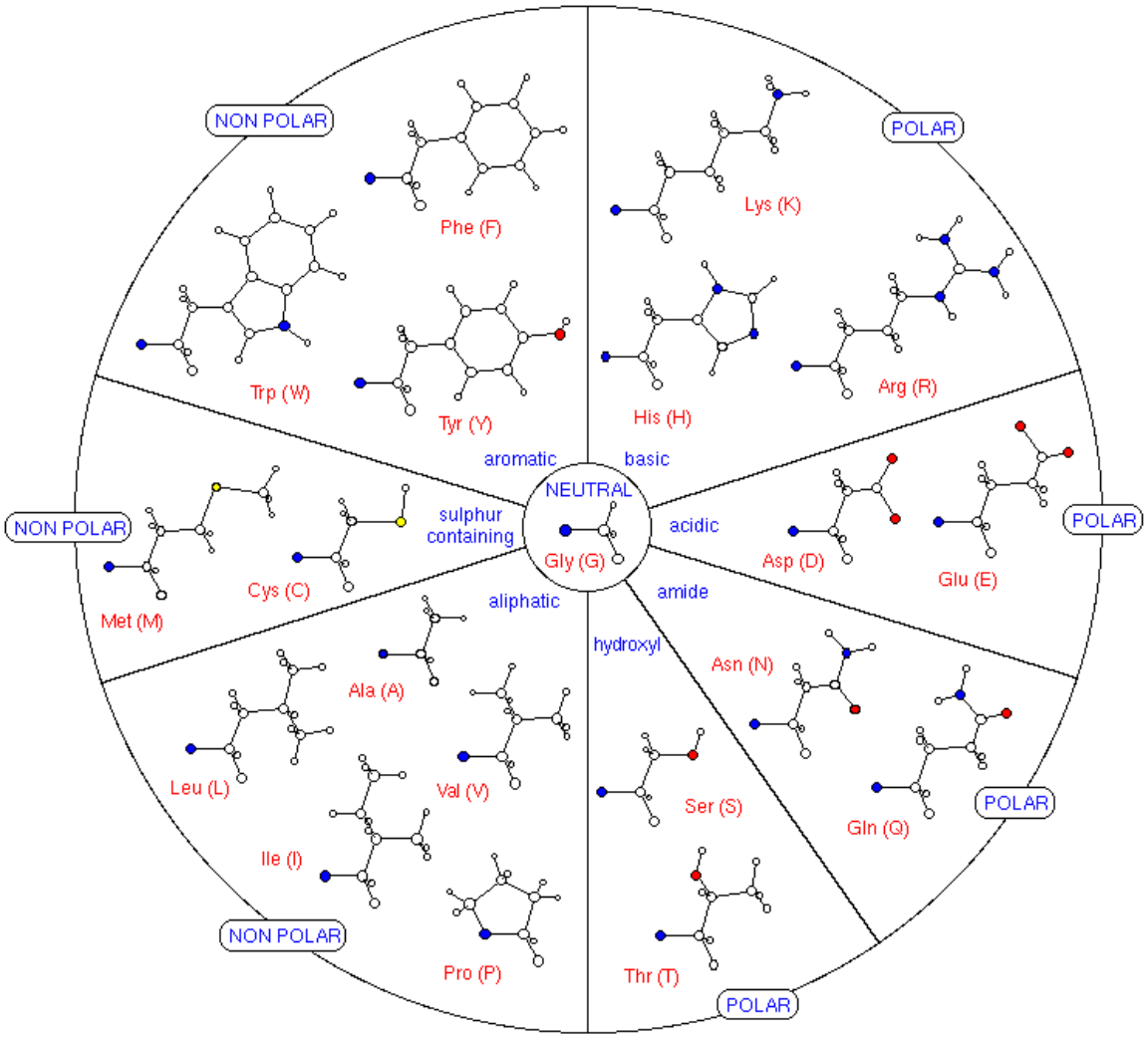
R. GRANTHAM

Laboratoire de Biométrie,
Université Lyon I,
69 Villeurbanne, France

Arg	Leu	Pro	Thr	Ala	Val	Gly	Ile	Phe	Tyr	Cys	His	Gln	Asn	Lys	Asp	Glu	Met	Trp	
110	145	74	58	99	124	56	142	155	144	112	89	68	46	121	65	80	135	177	Ser
	102	103	71	112	96	125	97	97	77	180	29	43	86	26	96	54	91	101	Arg
		98	92	96	32	138	5	22	36	198	99	113	153	107	172	138	15	61	Leu
			38	27	68	42	95	114	110	169	77	76	91	103	108	93	87	147	Pro
				58	69	59	89	103	92	149	47	42	65	78	85	65	81	128	Thr
					64	60	94	113	112	195	86	91	111	106	126	107	84	148	Ala
						109	29	50	55	192	84	96	133	97	152	121	21	88	Val
							135	153	147	159	98	87	80	127	94	98	127	184	Gly
								21	33	198	94	109	149	102	168	134	10	61	Ile
									22	205	100	116	158	102	177	140	28	40	Phe
										194	83	99	143	85	160	122	36	37	Tyr
											174	154	139	202	154	170	196	215	Cys
												24	68	32	81	40	87	115	His
													46	53	61	29	101	130	Gln
														94	23	42	142	174	Asn
															101	56	95	110	Lys
																45	160	181	Asp
																	126	152	Glu
																		67	Met

Table 2. Difference D for each amino acid pair (*10*). The mean chemical distance from the three-property formula (see text) $\bar{D}_{cpv} = 100$ (D_{ij} values have been multiplied by 50.723 to make this mean possible). Linear regression of *RSF* and log *RSF* on these D values gives correlation coefficients of −.66 and −.72, respectively. Previous difference indexes give correlation coefficients against *RSF* of −.34 (minimum base changes), −.42 (Sneath difference), and −.49 (Epstein formula). In each case, correlation is between the two sets (difference and *RSF*) of 190 values (*3*, *4*, *7*).

AMINO ACID DIFFERENCES, CHEMICALLY



nSNV´s IN HUMAN GENOMES – GENETIC PROFILING

This chapter is based on Kumar et al. 2011, Trends in Genetics 27: 277-386

● Thousands of individuals in the general public have begun to gain access to their genetic variation profiles by using direct-to-consumer DNA tests available from commercial vendors, which profile hundreds of thousands of genomic markers for low costs.

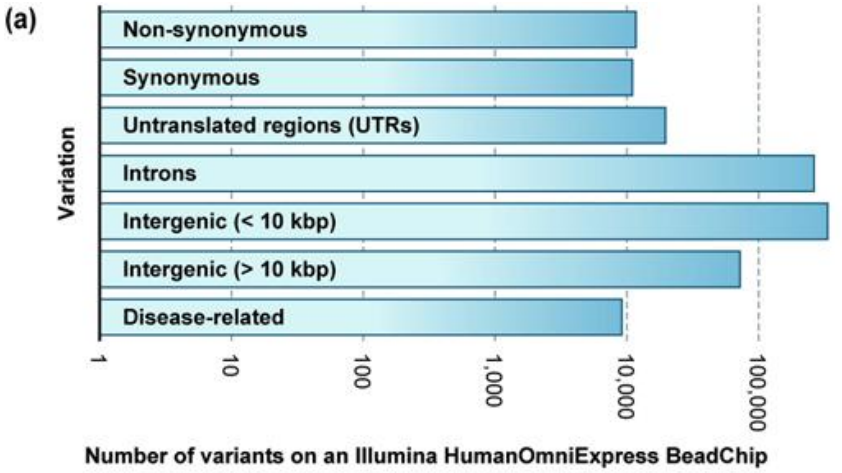
● Through this genetic profiling, individuals hope to learn about not only their ancestry, but also genetic variations underlying their physical characteristics and predispositions to diseases.

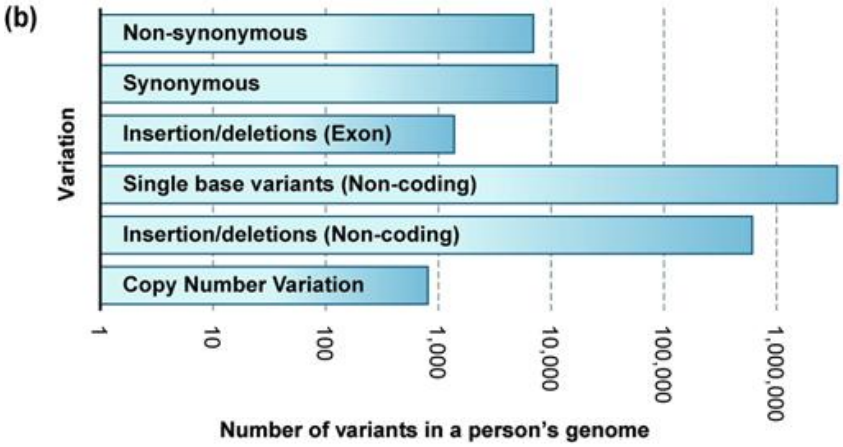
● Biomedicine scientists have been profiling variations at genomic markers in healthy and diseased individuals at genome scale in a variety of disease contexts and populations:
Discovery of thousands of disease associated genes and DNA variants.

● Any one personal genome contains more than a million variants, the majority of which are **single nucleotide variants, SNVs.**

● Majority of the known disease-associated variants are found within protein-coding genes with genome-wide association studies beginning to reveal also thousands of non-coding variants. Proteins are encoded in genomic DNA by exon regions, which comprise just ~1% of the genomic sequence, **Exome**. This is best understood part: how DNA blueprint sequence relates to function, and is arguably the best chance to connect genetic variations with disease pathophysiology. **A person's exome carries about 6,000 – 10,000 amino-acid-altering nonsynonymous SNVs, nSNVs,** known to be associated with more than a thousand major diseases **.**

nSNV´s IN HUMAN GENOMES – INDIVIDUAL GENETIC PROFILING



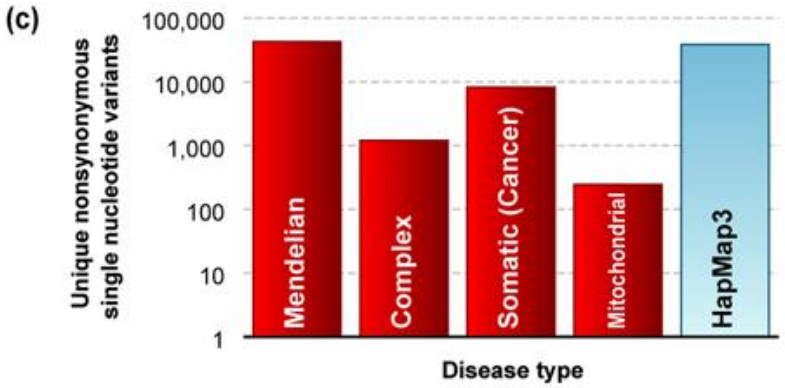


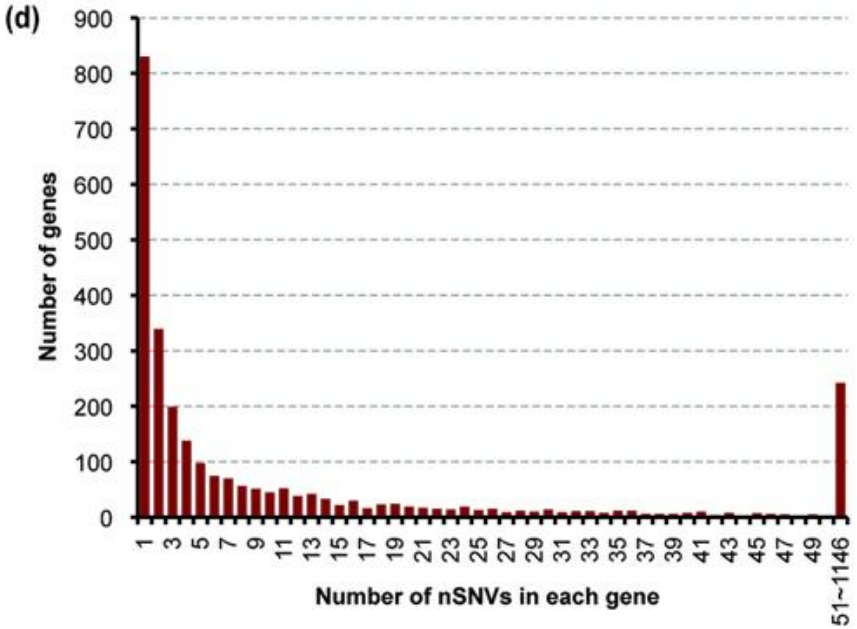
Profiles of personal and population variations.

(a) Counts of various types of genetic variants profiled by 23andMe using the Illumina HumanOmniExpress BeadChip. 733,202 SNP identifiers (rsIDs), retrieved from the Illumina website and mapped to the dbSNP database.

(b) The numbers of different types of variants found per human genome.

nSNV´s IN HUMAN GENOMES – INDIVIDUAL GENETIC PROFILING





(c) The numbers of known non-synonymous single nucleotide variants (nSNVs) in the human nuclear and mitochondrial genomes that are associated with Mendelian diseases, complex diseases, and somatic cancers. Compared to complex diseases and somatic cancers, nSNVs related to Mendelian diseases account for the most variants discovered to date.

(d) The number of nSNVs in each gene related to Mendelian diseases. The majority of genes have only one or a few mutations, while there are some genes hosting hundreds or even more than 1000 mutations.

The numbers of variants in panels {a–c} (a,b in previous page) are in $\log_{10}$ scale. Information for disease associated variants is shown in red and the personal and population variations are shown in blue.

nSNV´s IN HUMAN GENOMES – INDIVIDUAL GENETIC PROFILING

● Translating a personal variation profile into useful phenotypic information (e.g., relating to predisposition to disease, differential drug response, and other health concerns) is a grand challenge in the field of genomic medicine. Genomic medicine is concerned with enabling healthcare that is tailored to the individual based on genomic information.

● **Phylomedicine:** Through multispecies comparisons of data from various animals in “the tree of life”, it is possible to mine this information and evaluate the severity of each variant computationally (*in silico*).

● With the availability of large number genomes from the tree of life, it is becoming clear that evolution can serve as a kind of telescope for exploring the universe of genetic variation. In this evolutionary telescope, the degree of historical conservation of individual position (and regions) and the sets of substitutions permitted among species at individual positions serve as two lenses. This tool has the ability to provide first glimpses into the functional and health consequences of variations that are being discovered by high-throughput sequencing efforts.

● Phylomedicine is an important discipline at the intersection of molecular evolution and genomic medicine with a focus on understanding of human disease and health through the application of long-term molecular evolutionary history. Phylomedicine expands the purview of contemporary evolutionary medicine to use evolutionary patterns beyond the short-term history (e.g., populations) by means of multispecies genomics.

Mendelian (monogenic) diseases

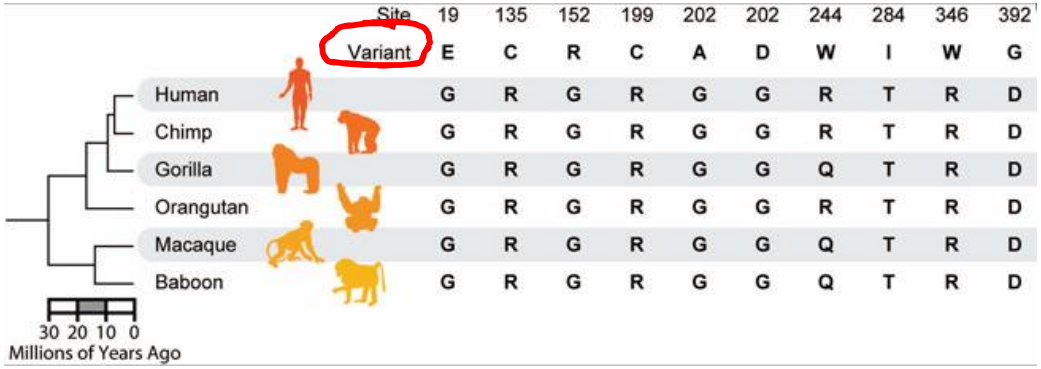
● For centuries it has been known that particular diseases run in families, notably in some royal families where there was a degree of inbreeding. Once Mendel's principles of inheritance became widely known in the early 1900s it became evident from family genealogies that specific heritable diseases fit Mendelian predictions.

● Over the last three decades, mutations in single (candidate) genes in many families have been linked to individual Mendelian diseases. Sometimes more than a hundred SNVs in the same gene have been implicated in a particular disease. For example, by the turn of this century, individual patient and family studies revealed over 500 nSNVs in the Cystic fibrosis transmembrane conductance regulator (*CFTR*) gene for cystic fibrosis (CF). *This enabled first efforts to examine evolutionary properties of the positions harboring CFTR nSNVs.*

- *The disease-associated nSNVs were found to be overabundant at positions that had permitted only a very small amount of change over evolutionary time.*
- This trend was confirmed at the proteome scale in analyses of thousands of nSNVs from hundreds of genes.
- These patterns were in sharp contrast to the variations seen in non-patients, which are enriched in the fast evolving positions. In population polymorphism data, faster evolving positions also show higher minor allele frequencies than those at slow evolving positions, which translates into *an enrichment of rare alleles in slow-evolving and functionally important genomic positions.*

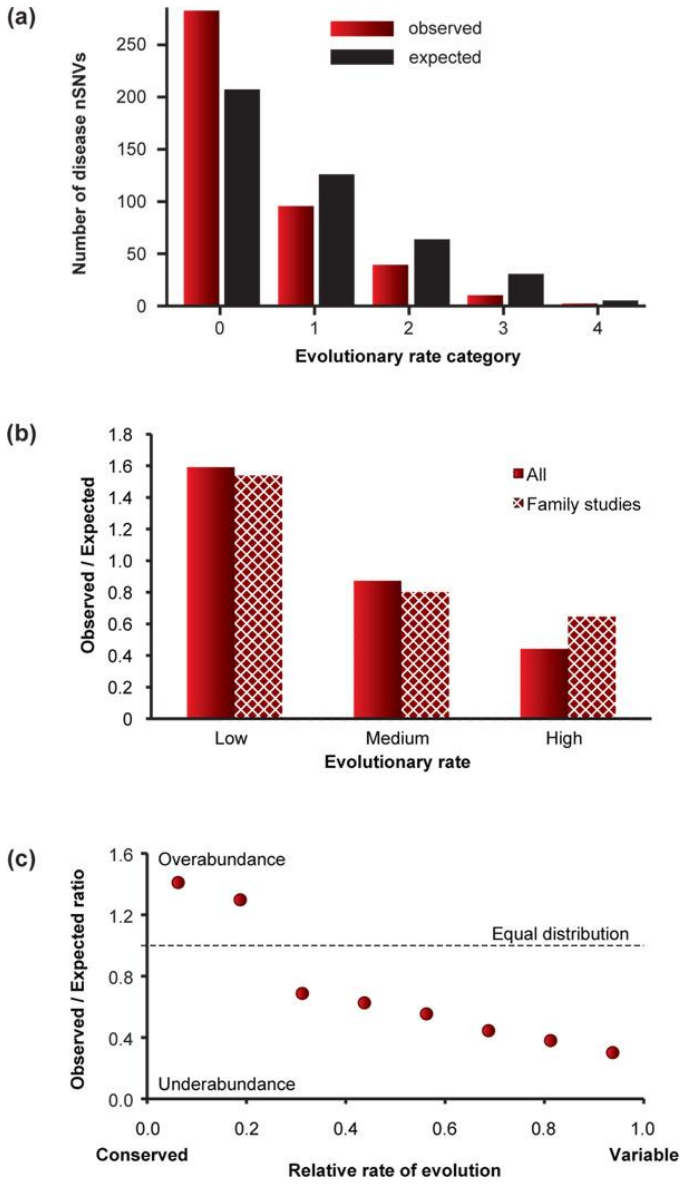
An example

Miller syndrome is a rare genetic disorder characterized by distinctive craniofacial malformations that occur in association with limb abnormalities. It is a typical Mendelian disease that is inherited as an autosomal recessive genetic trait. By sequencing the exomes of four affected individuals in three independent kindreds, ten mutations in a single candidate gene, *DHODH*, were found to be associated with this disease. They are in slow-evolving sites that are highly conserved not only in primates, but also among distantly related vertebrates. Specifically, 50% of these mutations are found at completely conserved positions among 46 vertebrates, including human. The average evolutionary rate for sites containing these disease-related mutations is 0.50 substitutions per billion year, which is ~40% slower than those sites hosting four non-disease-related population polymorphisms of DHODH available in the public databases.



Ten amino acid altering mutations at sites 19, 135, etc. referring to the protein sequence positions

nSNV´s IN HUMAN GENOMES – MENDELIAN DISEASES



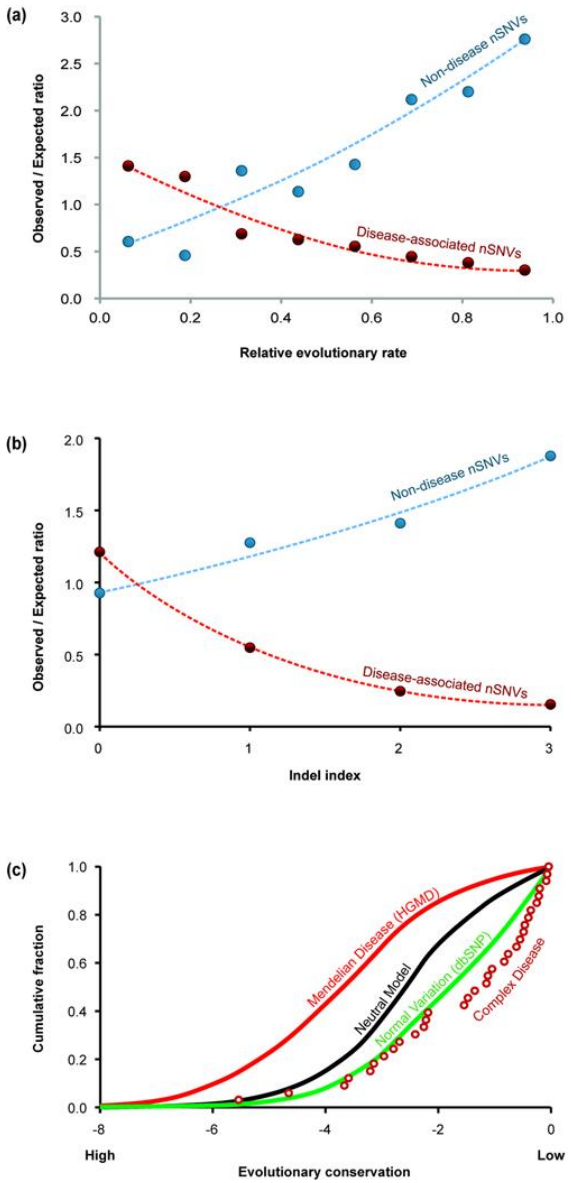
Evolutionary properties of positions afflicted with disease-associated nonsynonymous single nucleotide variants (nSNVs).

(a) The observed and expected numbers of disease associated nSNVs in positions that have evolved with different evolutionary rates in the *CFTR* protein (cystic fibrosis). The disease associated nSNVs are enriched in positions evolving with the lowest rates, which belong to the rate category 0.

(b) The ratio of observed to expected numbers of nSNVs in different rate categories for all *CFTR* variants (solid pattern; 431 variants) and those reported in publications profiling one or more families (hatched pattern; 59 variants).

(c) The proteome-scale relationship of the observed/expected ratios of Mendelian disease-associated nSNVs in positions that have evolved with different evolutionary rates. The results are from an analysis of disease associated nSNVs from 2,717 genes (public release of HGMD). Just as for individual diseases, nSNVs are enriched in positions evolving with the lowest rates.

nSNV´s IN HUMAN GENOMES – MENDELIAN DISEASES



The enrichment of disease-associated nSNVs (red) and the deficit of population polymorphisms (blue) in human amino acid positions

(a) evolving with different rates and

(b) with differ degrees of insertion-deletions. In both cases, smaller numbers on the x axis correspond to more conserved positions. There is an enrichment of disease associated nSNVs and a deficit of population nSNPs in conserved positions. This trend is reversed for the fastest evolving positions.

(c) The cumulative distributions of the evolutionary conservation scores for nSNVs associated with Mendelian diseases (solid red line), complex diseases (open red circles), and population polymorphisms (green line). The shift towards the left in Mendelian nSNVs indicates higher position specific evolutionary conservation. Conversely, a shift towards the right in complex disease nSNVs indicates lower evolutionary conservation, which overlaps with normal variations observed in the population. Data for the neutral model (black line) is from a simulation.

nSNV´s IN HUMAN GENOMES – MENDELIAN DISEASES

- Patterns of evolutionary retention at positions, another type of evolutionary conservation, a similar pattern is noticed: positions preferentially retained over the history of vertebrates were more likely to be involved in Mendelian diseases as compared to the patterns of natural variation. Somatic mutations in a variety of cancers have also been found to occur disproportionately at conserved positions. A similar pattern has emerged for mitochondrial disease-associated nSNVs.

- The relationship between evolutionary conservation and disease association has been explained by the effect of natural selection:
    - There is a high degree of purifying selection on variation at highly conserved positions because of their potential effect on inclusive fitness (fecundity, reproductive success) due to the functional importance of the position.
    - At the faster-evolving positions, many substitutions have been tolerated over evolutionary time in different species.
    - This points to the “neutrality” of some mutations that spread through the population primarily by the process of random genetic drift and appear as fixed differences between species.
    - Therefore, fewer mutations are culled at fast-evolving positions, producing a relative under-abundance of disease mutations at such positions. Of course, the above arguments hold true only when the functional importance of a position has remained unchanged over evolutionary time, an assumption that is expected to be fulfilled for a large fraction of positions in orthologous proteins.

Identification of deleterious mutations within three human genomes

Sung Chun[1] and Justin C. Fay[1,2,3]

[1]*Computational Biology Program, Washington University, St. Louis, Missouri 63108, USA;* [2]*Department of Genetics, Washington University, St. Louis, Missouri 63108, USA*

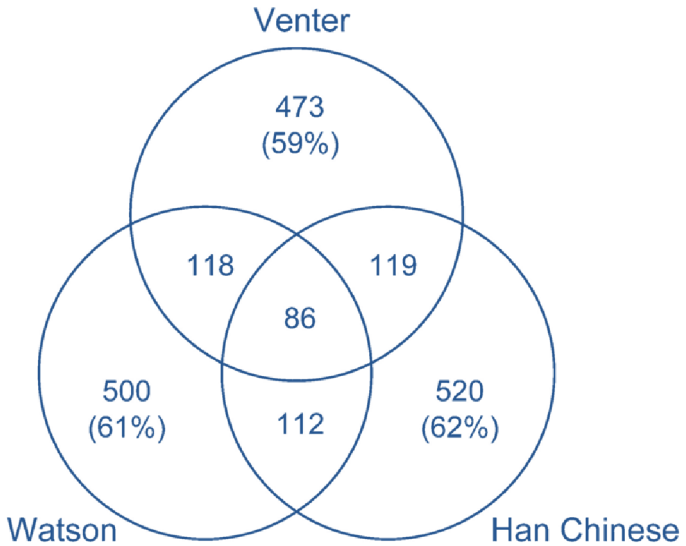
Each human carries a large number of deleterious mutations. Together, these mutations make a significant contribution to human disease. Identification of deleterious mutations within individual genome sequences could substantially impact an individual's health through personalized prevention and treatment of disease. Yet, distinguishing deleterious mutations from the massive number of nonfunctional variants that occur within a single genome is a considerable challenge. Using a comparative genomics data set of 32 vertebrate species we show that a likelihood ratio test (LRT) can accurately identify a subset of deleterious mutations that disrupt highly conserved amino acids within protein-coding sequences, which are likely to be unconditionally deleterious. The LRT is also able to identify known human disease alleles and performs as well as two commonly used heuristic methods, SIFT and PolyPhen. Application of the LRT to three human genomes reveals 796–837 deleterious mutations per individual, ~40% of which are estimated to be at <5% allele frequency. However, the overlap between predictions made by the LRT, SIFT, and PolyPhen, is low; 76% of predictions are unique to one of the three methods, and only 5% of predictions are shared across all three methods. Our results indicate that only a small subset of deleterious mutations can be reliably identified, but that this subset provides the raw material for personalized medicine.

19:1553–1561 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

DELETERIOUS MUTATIONS IN THREE HUMAN GENOMES

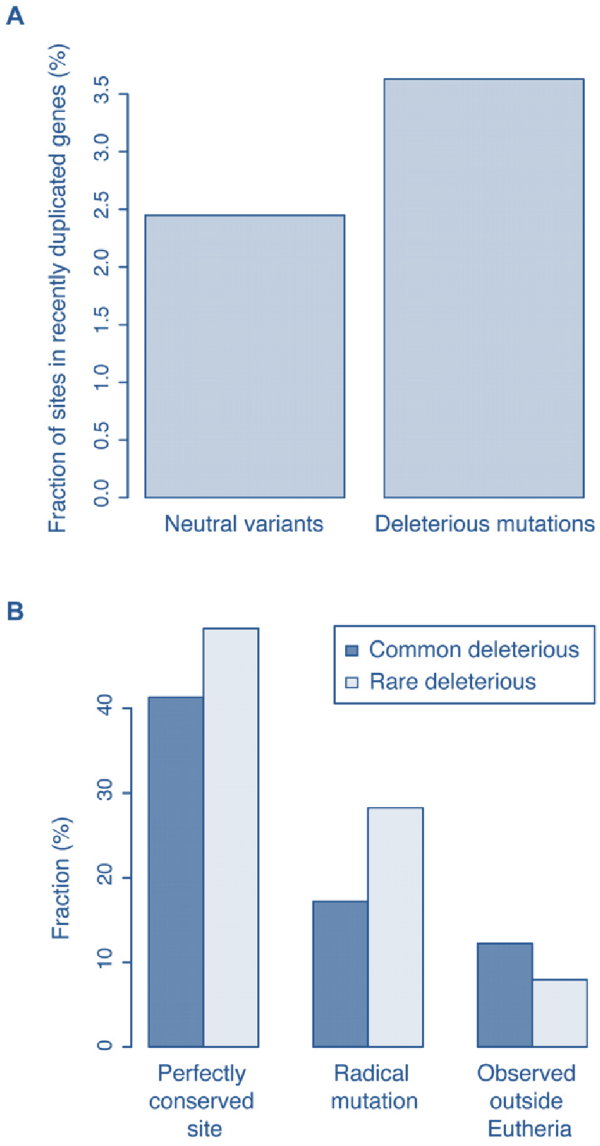
- The paper has data from a complete catalog of SNPs from J. Craig Venter (Google who he is if you don´t know), from a Han Chinese male from their respective websites (http://www.jcvi.org/cms/research/projects/huref/ and http://yh.genomics.org.cn), and for James D. Watson (from "Watson – Crick")

- Nonsynonymous and synonymous SNPs were identified using known genes in Ensembl release 49. Coding SNPs in ambiguous reading frames, due to overlap of adjacent genes or frame shifts between known splice variants, or in known pseudogenes, were excluded.



- The percentage of individual-specific deleterious mutations found in each genome is shown in parentheses.

DELETERIOUS MUTATIONS IN THREE HUMAN GENOMES



Characteristics of deleterious mutations.

(*A*) Deleterious mutations (*n* = 1928) are more likely to occur in recently duplicated genes relative to neutral variants (*n* = 8287).

(*B*) Mutations at perfectly conserved sites, mutations that cause radical amino acid changes, defined by BLOSUM62 ≤ −2, and mutations to amino acids that are not observed outside of eutherian mammals are more frequent among rare (*n* = 807) compared with common deleterious mutations (*n* = 1121).

EXAMPLE 1

OPEN ACCESS Freely available online

PLoS GENETICS

Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome

Adam R. Boyko[1,2], Scott H. Williamson[1], Amit R. Indap[1], Jeremiah D. Degenhardt[1], Ryan D. Hernandez[1], Kirk E. Lohmueller[1,2], Mark D. Adams[3], Steffen Schmidt[4], John J. Sninsky[5], Shamil R. Sunyaev[4], Thomas J. White[5], Rasmus Nielsen[6], Andrew G. Clark[2], Carlos D. Bustamante[1]*

Abstract

Quantifying the distribution of fitness effects among newly arising mutations in the human genome is key to resolving important debates in medical and evolutionary genetics. Here, we present a method for inferring this distribution using Single Nucleotide Polymorphism (SNP) data from a population with non-stationary demographic history (such as that of modern humans). Application of our method to 47,576 coding SNPs found by direct resequencing of 11,404 protein coding-genes in 35 individuals (20 European Americans and 15 African Americans) allows us to assess the relative contribution of demographic and selective effects to patterning amino acid variation in the human genome. We find evidence of an ancient population expansion in the sample with African ancestry and a relatively recent bottleneck in the sample with European ancestry. After accounting for these demographic effects, we find strong evidence for great variability in the selective effects of new amino acid replacing mutations. In both populations, the patterns of variation are consistent with a leptokurtic distribution of selection coefficients (*e.g.*, gamma or log-normal) peaked near neutrality. Specifically, we predict 27–29% of amino acid changing (nonsynonymous) mutations are neutral or nearly neutral ($|s|<0.01\%$), 30–42% are moderately deleterious ($0.01\%<|s|<1\%$), and nearly all the remainder are highly deleterious or lethal ($|s|>1\%$). Our results are consistent with 10–20% of amino acid differences between humans and chimpanzees having been fixed by positive selection with the remainder of differences being neutral or nearly neutral. Our analysis also predicts that many of the alleles identified via whole-genome association mapping may be selectively neutral or (formerly) positively selected, implying that deleterious genetic variation affecting disease phenotype may be missed by this widely used approach for mapping genes underlying complex traits.

EXAMPLE 2

Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations

Sudhir Kumar,[1,2,3] Michael P. Suleski,[1] Glenn J. Markov,[1] Simon Lawrence,[1] Antonio Marco,[1] and Alan J. Filipski[1]

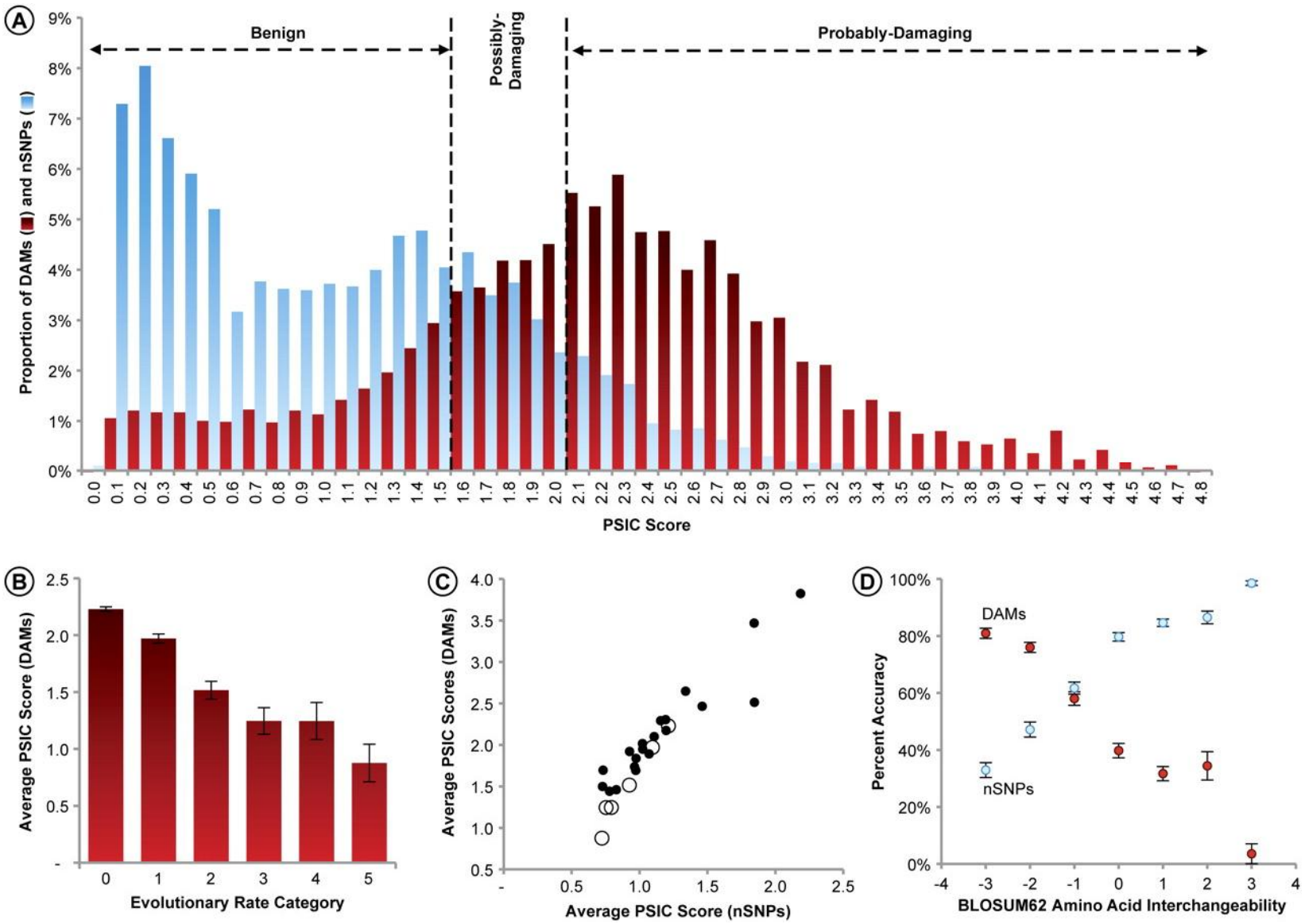
[1]*Center for Evolutionary Functional Genomics, Biodesign Institute, Arizona State University, Tempe, Arizona 85287-5301, USA;*
[2]*School of Life Sciences, Arizona State University, Tempe, Arizona 85287-4501, USA*

As the cost of DNA sequencing drops, we are moving beyond one genome per species to one genome per individual to improve prevention, diagnosis, and treatment of disease by using personal genotypes. Computational methods are frequently applied to predict impairment of gene function by nonsynonymous mutations in individual genomes and single nucleotide polymorphisms (nSNPs) in populations. These computational tools are, however, known to fail 15%–40% of the time. We find that accurate discrimination between benign and deleterious mutations is strongly influenced by the long-term (among species) history of positions that harbor those mutations. Successful prediction of known disease-associated mutations (DAMs) is much higher for evolutionarily conserved positions and for original–mutant amino acid pairs that are rarely seen among species. Prediction accuracies for nSNPs show opposite patterns, forecasting impediments to building diagnostic tools aiming to simultaneously reduce both false-positive and false-negative errors. The relative allele frequencies of mutations diagnosed as benign and damaging are predicted by positional evolutionary rates. These allele frequencies are modulated by the relative preponderance of the mutant allele in the set of amino acids found at homologous sites in other species (evolutionarily permissible alleles [EPAs]). The nSNPs found in EPAs are biochemically less severe than those missing from EPAs across all allele frequency categories. Therefore, it is important to consider position evolutionary rates and EPAs when interpreting the consequences and population frequencies of human mutations. The impending sequencing of thousands of human and many more vertebrate genomes will lead to more accurate classifiers needed in real-world applications.

1562 Genome Research
www.genome.org

19:1562–1569 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

UNDERSTANDING EVOLUTIONARY PATTERNS OF MUTATIONS



■ Consider a stochastic model for DNA or amino acid sequence evolution.

■ Assume independence of evolution at different sequence sites => sites can be considered one by one.

■ At any single site, the model works with probabilities $P_{ij}(T)$ that base i will have changed to base j after a time T.

■ The subscripts i and j take the values 1,...,4 to represent the nucleotides A, T, C, G for DNA sequences and 1,...,20 for amino acid sequences.

■ Given a stochastic variable $X(t)$ describing the evolution through time t of a site in one sequence, **the Markov assumption** asserts that

$$P_{ij}(T) = \Pr[X(s+T) = j | X(s) = i] \text{ is independent of } s \geqslant 0$$

■ This means that subsequent to any time s it does not matter how the process reached state i by time s and the future course of evolution depends only on i. The process is *memoryless*.

■ The probabilities of transition from one base to another, $P_{ij}(T)$, can be written as a matrix $\mathbf{P}(T)$,

$$\mathbf{P}(T + dT) = \mathbf{P}(T)(\mathbf{I} + \mathbf{Q}dT)$$

NOTE: transition (probability) here has nothing to do with the terminology
transitions vs. transversions which are for purine-purine and pyrimidine-pyrimidine vs, purine-pyrimidine substitutions

■ dT represents a small time
■ **I** is the identity matrix.
■ The matrix **Q** is the *instantaneous rate matrix* and has off-diagonal entries Q_{ij} equal to the rates of replacement of i by j. The diagonal entries, Q_{ii}, are defined by a requirement that the row sums are all zero.

■ Solving the equation gives

$$\mathbf{P}(T) = e^{T\mathbf{Q}} = \mathbf{I} + T\mathbf{Q} + (T\mathbf{Q})^2/2! + (T\mathbf{Q})^3/3! + \ldots$$

■ Diagonalization (spectral decomposition) of **Q:** calculating the matrix $\mathbf{P}(T)$

$$\mathbf{P}(T) = \mathbf{U} \cdot \text{diag}\{e^{\lambda 1 T}, \ldots, e^{\lambda n T}\} \cdot \mathbf{U}^{-1}$$

■ **U** contains the eigenvectors of **Q,** the λ_i are the eigenvalues of **Q** and diag{ } denotes the diagonal matrix of the elements.

■ The components $P_{ij}(T)$ can be written as $P_{ij}(T) = \Sigma_k c_{ijk} e^{\lambda k T}$ where the sum is over k = 1, .., 4 for DNA sequences, c_{ijk} is a function of **U** and $\mathbf{U}^{-1}$
■ T and **Q** are confounded $T\mathbf{Q} = (T/\gamma)(\gamma\mathbf{Q})$ for any $\gamma \neq 0$ (half the time at twice the rate has the same result).
■ *Time is not absolute, but scaled to units of expected substitutions per site.*

- A Markov process can have three important properties:

  - **Homogeneity.** The rate matrix is independent of time which means that patterns of nucleotide substitution (or amino acid replacement) remain the same in different parts of the phylogenetic tree. A homogeneous process has an equilibrium distribution that is also the limiting distribution when time approaches infinity.

  - **Stationarity** means that the process is at that equilibrium, that is, nucleotide frequencies have remained more or less the same during the course of evolution.

  - **Reversibility** means that $\pi_i P_{ij}(T) = \pi_j P_{ji}(T)$ for all i, j, and T where π_i are the frequencies of occurrence for each base. A consequence of reversibility is that the process of sequence evolution is theoretically indistinguishable from the same process watched in reverse.

PROPERTIES OF MARKOV MODELS IN NUCLEOTIDE SUBSTITUTION MODELLING

■ Models in widespread use typically assume homogeneity, yet this is rarely likely to be fully appropriate, for example, because of the dependence of mutation on local sequence context.

■ Stationarity is not a consequence of a Markov model but of its application; this, too, is generally assumed in phylogenetics, although when base frequencies are quite different in different species this assumption is clearly violated. Genomes show large differences in base compositions. For example, the genome of one bacterium is 74% G+C content, whereas the genome of another is only 25% G + C content.

‘■ Reversibility, too, is generally assumed, with little justification other than that numerical calculations are simplified considerably. Assumptions such as those of homogeneity, stationarity, and reversibility are typical of the approximations that have to be made to render the wide knowledge of molecular biology into a mathematically tractable form.

■ The Jukes-Cantor model (see above) is defined by $Q_{ij} = \alpha$ for all $i,j = 1,...,4$; $i \neq j$, meaning that each base is substituted by any other at equal rate α. A consequence of this model is that the base frequencies (π_i) are all assumed equal to 0.25.

■ Kimura´s two-parameter model considers the difference in transition and transversion rates. The instantaneous rate matrix is given in page 9. In this the order of the bases for columns and rows are A, T, C, G, and the (i,j) entry represents Q_{ij}, the rate ($i \rightarrow j$) at which a base i is replaced by a base j.

■ After Kimura, several authors have proposed models with increasing numbers of parameters. For example,

- asymmetry for some reciprocal changes: $i \rightarrow j$ has a different substitution rate from $j \rightarrow I$,
- with or without reversibility assumption, different kind of biases allowed, etc.
- The most general model has 12 independent parameters.

■ The Jukes-Cantor model (see above) is defined by $Q_{ij} = \alpha$ for all $i,j = 1,...,4$; $i \neq j$, meaning that each base is substituted by any other at equal rate α. A consequence of this model is that the base frequencies (π_i) are all assumed equal to 0.25.

■ Kimura´s two-parameter model considers the difference in transition and transversion rates. The instantaneous rate matrix is given in page 9. In this the order of the bases for columns and rows are A, T, C, G, and the (i,j) entry represents Q_{ij}, the rate ($i \rightarrow j$) at which a base i is replaced by a base j.

■ After Kimura, several authors have proposed models with increasing numbers of parameters. For example,

- asymmetry for some reciprocal changes: $i \rightarrow j$ has a different substitution rate from $j \rightarrow I$,
- with or without reversibility assumption, different kind of biases allowed, etc.
- The most general model has 12 independent parameters.