Models of substitution:

nucleotide

amino acid

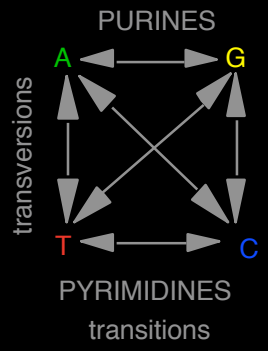
Why?

- with only 4 nt (A, C, G, T) the absolute difference or distance, between any 2 taxa is an underestimate, there may have been more changes that were over-written
- with only 4 nt (A, C, G, T) the observed branch length on a tree is an underestimate

What's wrong with absolute (p) distance?

If two taxa have random nt assigments
p(max) = 0.75 not 1.00
that is, random assignment of 4 nt = 25% identity
even when every site has had an infinitely high probability of substitution
p underestimates substitutions per site (e.g., multiple hits)

$$\text{Jukes Cantor: } a = -\,{}^{3}/_{4}\ln(1-(4p/3))$$



	A	C	G	T
A		a	a	a
C	a		a	a
G	a	a		a
T	a	a	a	

The JC model assumes all substitution types are the same

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

Jukes Cantor:

$$-\,{}^{3}/_{4}\,\ln(1-(4p/3))$$

But, if p > 0.75 JCdist is undefined.
This can happen if base compositions are unequal or
if transitions are more common than transversions.

Felsenstein81:

$$-\,(1-(\pi_a{}^2+\pi_c{}^2+\pi_g{}^2+\pi_t{}^2))\times\ln(1-(p/(\pi_a{}^2+\pi_c{}^2+\pi_g{}^2+\pi_t{}^2)))$$

π is base freq

Kimura 2 Parameter:

$$-\,{}^{1}/_{2}\,\ln({}^{1}/_{(1-2P-Q)})+{}^{1}/_{4}\,\ln\,({}^{1}/_{(1-2Q)})$$

P = proportion of changes that are transitions

Q = proportion of changes that are transversions

Mark Siddall
AMNH
2005

Species A	ATGGCTATTCTTATAGTACG
Species B	ATCGCTAGTCTTATATTACA
Species C	TTCACTAGACCTGTGGTCCA
Species D	TTGACCAGACCTGTGGTCCG
Species E	TTGACCAGTTCTCTAGTTCG

	A	B	C	D	E
Species A		0.20	0.50	0.45	0.40
Species B	0.23		0.40	0.55	0.50
Species C	0.87	0.59		0.15	0.40
Species D	0.73	1.12	0.17		0.25
Species E	0.59	0.89	0.61	0.31	

Uncorrected p

Kimura 2 paramater

Expressed as expected substitutions per site

Nucleotide Models - time reversible (i.e., symmetrical)

JUKES-CANTOR

	A	C	G	T
A	1-3k	k	k	k
C	k	1-3k	k	k
G	k	k	1-3k	k
T	k	k	k	1-3k

1 SUBSTITUTION TYPE

FELSENSTEIN 81

	A	C	G	T
A	1-3k	k	k	k
C	k	1-3k	k	k
G	k	k	1-3k	k
T	k	k	k	1-3k

$X \ \pi_{acgt}$

Because a site must either change or not change, the diagonal is the prob(no change) and is simply 1 less the sum of the prob of the 3 possible changes.

KIMURA 2 PARAMETER

	A	C	G	T
A	1-2m-3k	m	k	m
C	m	1-2m-3k	m	k
G	k	m	1-2m-3k	m
T	m	k	m	1-2m-3k

2 SUBSTITUTION TYPES

HASEGAWA KISHINO YANO 85

	A	C	G	T
A	1-2m-3k	m	k	m
C	m	1-2m-3k	m	k
G	k	m	1-2m-3k	m
T	m	k	m	1-2m-3k

$X \ \pi_{acgt}$

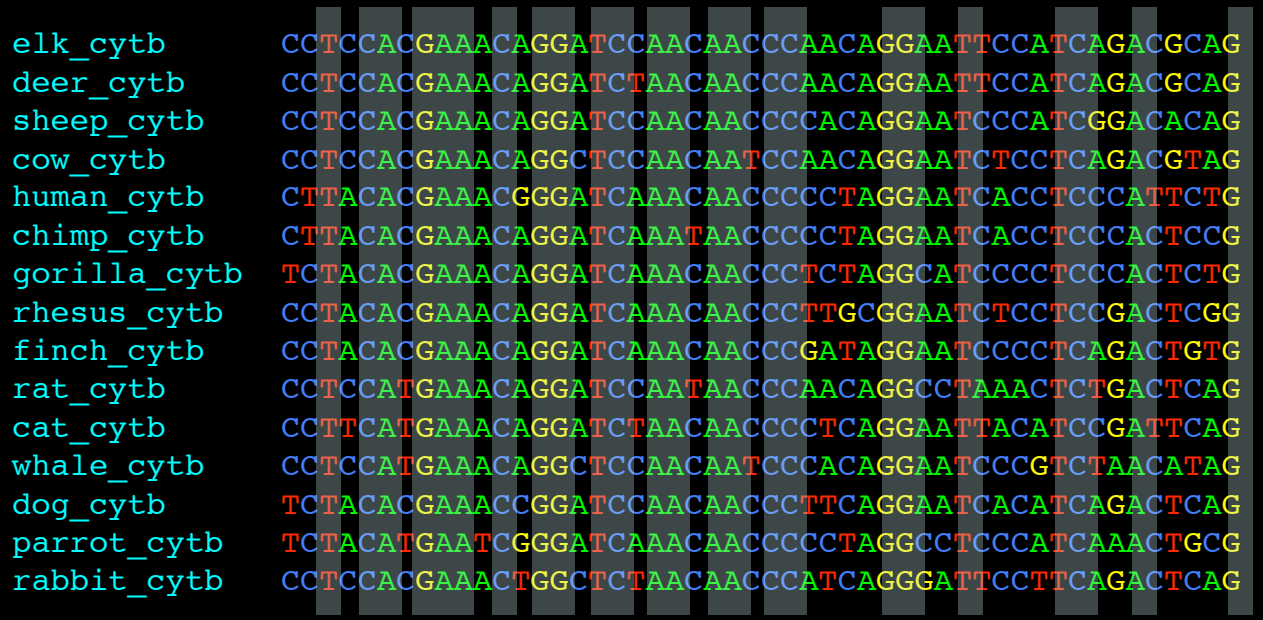
Nucleotide Models

GENERAL TIME REVERSIBLE

	A	C	G	T
A	1-k-m-x	k	m	x
C	k	1-k-d-q	d	q
G	m	d	1-m-d-h	h
T	x	q	h	1-x-q-h

6 SUBSTITUTION TYPES

Maybe there is evidence that some sites don't change

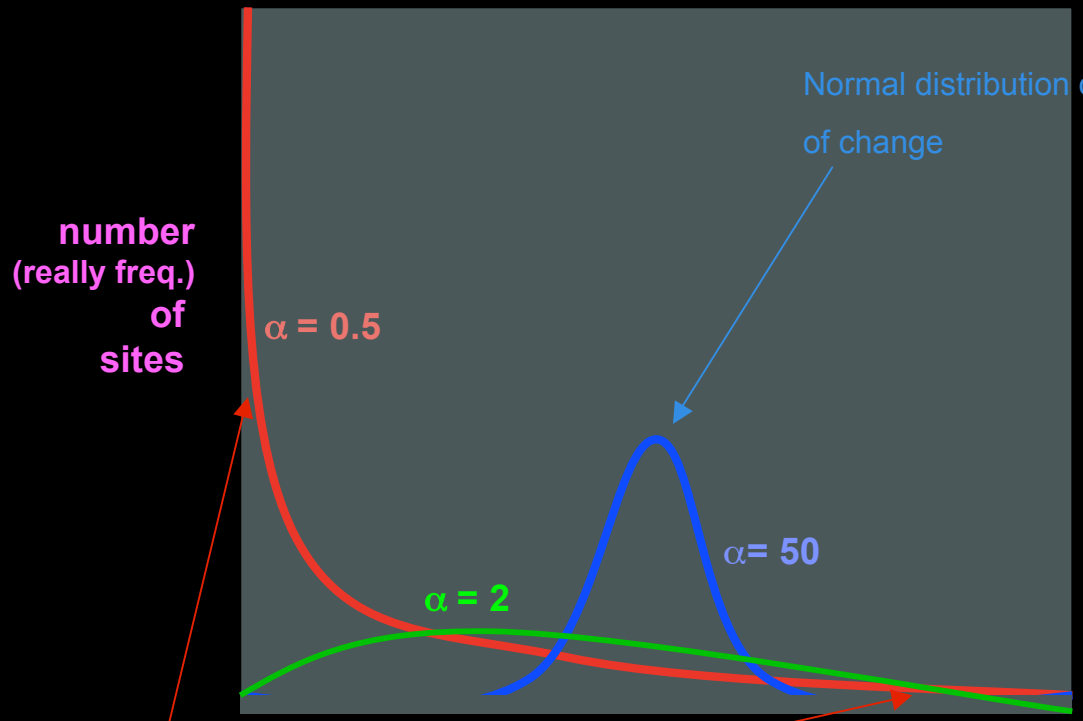


some proportion invariant

You may assign a proportion invariant (PINVAR or I)

Nucleotide Models

Maybe not all sites have same rate
the gamma distribution

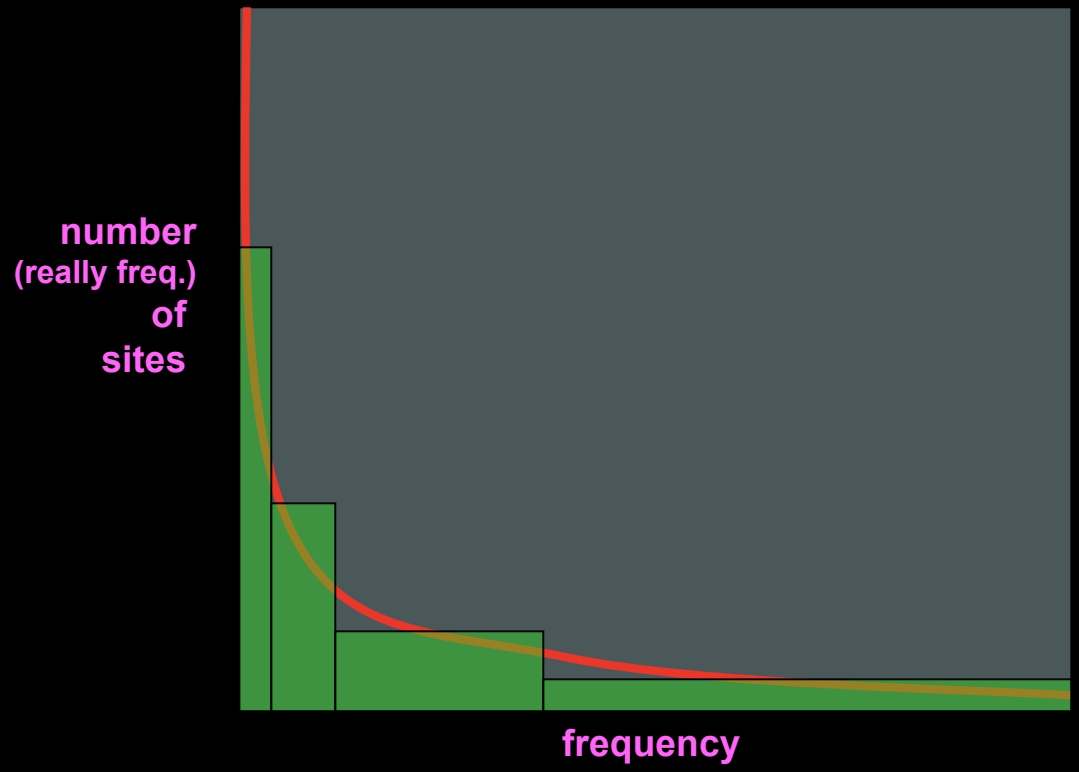


Usually:

$0.4 < \alpha < 2$

Nucleotide Models

rate heterogeneity
the gamma distribution



In practise, it's much too difficult to use a continuous distribution so it is approximated with rate bins of equal # of characters (above) or of equal width (not shown).

Nucleotide Models

NOTE: A more complex model (one with more parameters, will always fit the data better. However, using more parameters will take more time and may create instances of inconsistency (if the model is wrong).

In practise one should use the model with no more than the number of necessary parameters such that adding another paramter does not significantly improve the performance.

Likelihood ratio tests (LRT) work for nested models (where the simpler model is just a sepcial case of the more compex model) but may not work in cases where the models are not nested.

JC is a special case of K2P

JC is a special case of F81

F81 and K2P are not cases of each other - are not nested

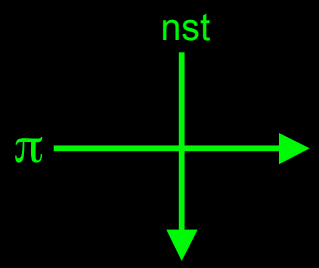
JC, F81 and K2P are special cases of HKY85

Nucleotide Models

$\frac{p(e|h_{simple})}{p(e|h_{complex})}$ ... likelihood ratio test

approximately χ^2 distributed if models are nested

JC F81



K2P HKY85

Akaike Information Criterion

$$AIC = -2 \ln L + 2K$$

$$AIC_c = -2 \ln L + \frac{2K(K+1)}{n-K-1}$$

K = number of sites

Mark Siddall
AMNH
2005

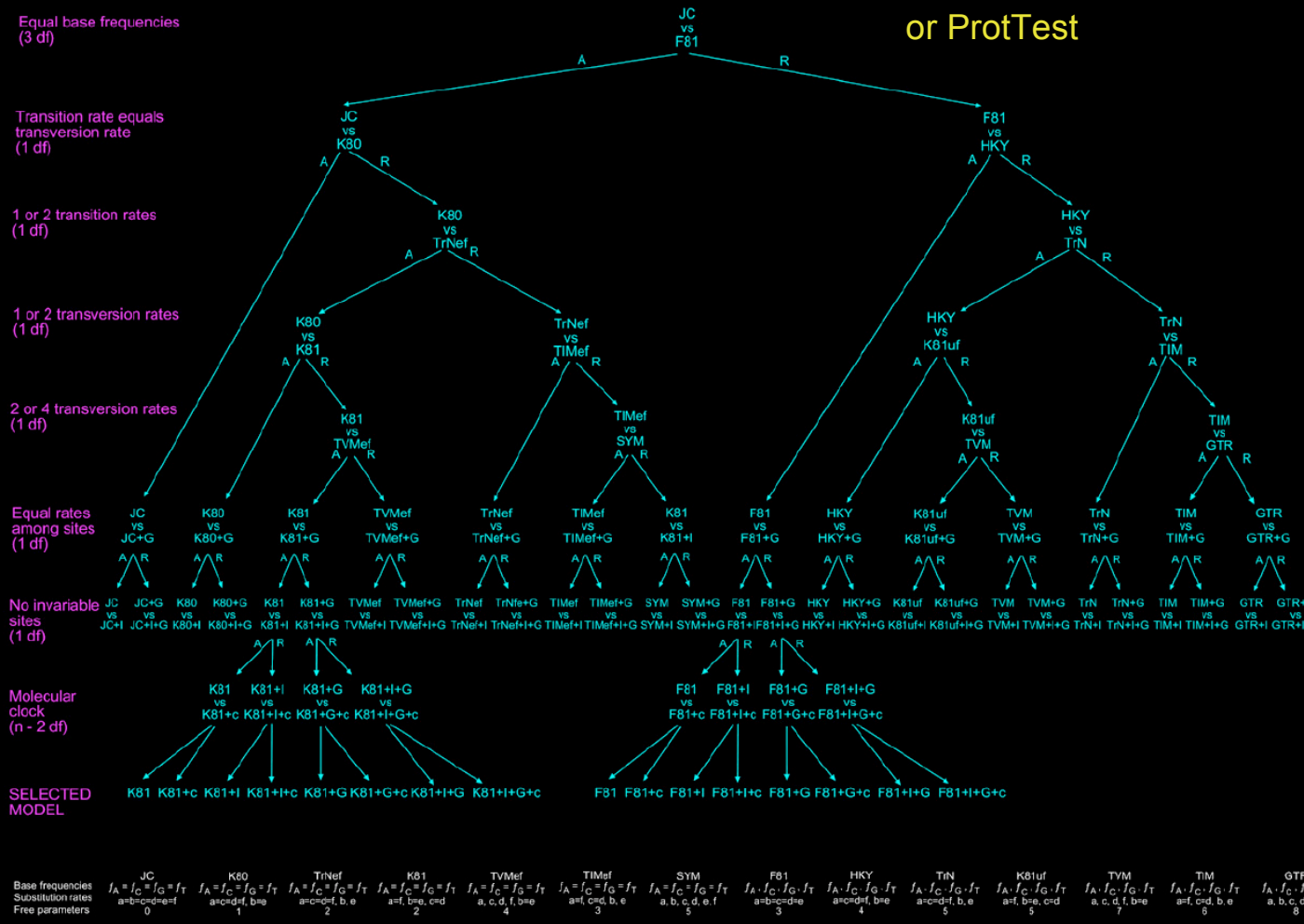
Nucleotide Models

Another pet peev:

Authors specifying a model without a justification

Use ModelTest! (Posada and Crandall)

or ProtTest



	JC	K80	TrNef	K81	TVMef	TIMef	SYM	F81	HKY	TrN	K81uf	TVM	TIM	GTR
Base frequencies	$f_A = f_C = f_G = f_T$	$f_A = f_C = f_G = f_T$	$f_A = f_C \neq f_G = f_T$	$f_A = f_C = f_G = f_T$	$f_A = f_C = f_G = f_T$	$f_A = f_C = f_G = f_T$	$f_A = f_C = f_G = f_T$	f_A, f_C, f_G, f_T	f_A, f_C, f_G, f_T	f_A, f_C, f_G, f_T	f_A, f_C, f_G, f_T	f_A, f_C, f_G, f_T	f_A, f_C, f_G, f_T	f_A, f_C, f_G, f_T
Substitution rates	a=b=c=d=e=f	a=c=d=f, b=e	a=c=d=f, b, e	a=f, b=e, c=d	a, c, d, f, b=e	a=f, c=d, b, e	a, b, c, d, e, f	a=b=c=d=e	a=c=d=f, b=e	a=c=d=f, b, e	a=f, b=e, c=d	a, c, d, f, b=e	a=f, c=d, b, e	a, b, c, d, e, f
Free parameters	0	1	2	2	4	3	5	3	4	5	5	7	6	8

FIG. 2.—Hierarchical likelihood ratio tests (ηLTRs). Hypotheses tested are indicated on the left. The acceptance or rejection of each LRT (by default when $P < 0.01$) determines the path. F clarity purposes, only the full paths corresponding to the K81 and F81 model families are indicated. Mean features of the models of evolution compared are summarized at the bottom. JC (Jukes a Cantor 1969); K80 (Kimura 1980); K81 (Kimura 1981); TrNef (TrN model with equal base frequencies); TIMef (TIM model with equal base frequencies); TVMef (TVM model with equal ba frequencies); SYM (Zharkikh 1994); F81 (Felsenstein 1981); HKY (Hasegawa, Kishino, and Yano 1985); K81uf (K81 model with unequal base frequencies); TrN (Tamura and Nei 1993); TI (transitional model: $rAC = rGT \neq rAT = rCG \neq rAG \neq rGT$); TVM (transversional model: $rAC = rCT \neq rAG \neq rAT \neq rCG \neq rGT$); GTR (Rodríguez et al. 1990). I = invariable sites; G gamma distribution; c = molecular clock enforced.

Amino Acid Models

- 20 x 20 rate matrix

- rates, based on genetic code (min or ave substitutions between each)
  (e.g., Protpars in phylip is min number of nt changes between each amino acid)

- mutational data matrix (MDM)
  empirical "models" based on large numbers of datasets to determine "population" freq of change

  some are based on sequences (PAM, BLOSUM, JTT, WAG)
  some are based on structures (STR)

  some are based pairwise (BLOSUM)
  some are tree based
  of which some are parsimony based (PAM, JTT)
  and some are likelihood based (metREV, WAG)

Amino Acid Models

```
begin assumptions;
            usertype protpars = 23
            [This matrix gives the minimum number of amino acid
            replacement substitutions needed to convert one amino
            acid to another, based on the genetic code used in
            nuclear genes of most organisms and chloroplast genes
            in plants).  It was computed using a program written
            by David Swofford.
            ]
                 A C D E F G H I K L M N P Q R 1 2 T V W Y * -
            [A]  0 2 1 1 2 1 2 2 2 2 2 2 1 2 2 1 2 1 1 2 2 2 3
            [C]  2 0 2 2 1 1 2 2 2 2 2 2 2 2 1 1 1 2 2 1 1 1 3
            [D]  1 2 0 1 2 1 1 2 2 2 2 1 2 2 2 2 2 2 1 2 1 2 3
            [E]  1 2 1 0 2 1 2 2 1 2 2 2 2 1 2 2 2 2 1 2 2 1 3
            [F]  2 1 2 2 0 2 2 1 2 1 2 2 2 2 2 1 2 2 1 2 1 2 3
            [G]  1 1 1 1 2 0 2 2 2 2 2 2 2 2 1 2 1 2 1 1 2 1 3
            [H]  2 2 1 2 2 2 0 2 2 1 2 1 1 1 1 2 2 2 2 2 1 2 3
            [I]  2 2 2 2 1 2 2 0 1 1 1 1 2 2 1 2 1 1 1 2 2 2 3
            [K]  2 2 2 1 2 2 2 1 0 2 1 1 2 1 1 2 2 1 2 2 2 1 3
            [L]  2 2 2 2 1 2 1 1 2 0 1 2 1 1 1 1 2 2 1 1 2 1 3
            [M]  2 2 2 2 2 2 2 1 1 1 0 2 2 2 1 2 2 1 1 2 3 2 3
            [N]  2 2 1 2 2 2 1 1 1 2 2 0 2 2 2 2 1 1 2 3 1 2 3
            [P]  1 2 2 2 2 2 1 2 2 1 2 2 0 1 1 1 2 1 2 2 2 2 3
            [Q]  2 2 2 1 2 2 1 2 1 1 2 2 1 0 1 2 2 2 2 2 2 1 3
            [R]  2 1 2 2 2 1 1 1 1 1 1 2 1 1 0 2 1 1 2 1 2 1 3
            [1]  1 1 2 2 1 2 2 2 2 1 2 2 1 2 2 0 2 1 2 1 1 1 3
            [2]  2 1 2 2 2 1 2 1 2 2 2 1 2 2 1 2 0 1 2 2 2 2 3
            [T]  1 2 2 2 2 2 2 1 1 2 1 1 1 2 1 1 1 0 2 2 2 2 3
            [V]  1 2 1 1 1 1 2 1 2 1 1 2 2 2 2 2 2 2 0 2 2 2 3
            [W]  2 1 2 2 2 1 2 2 2 1 2 3 2 2 1 1 2 2 2 0 2 1 3
            [Y]  2 1 1 2 1 2 1 2 2 2 3 1 2 2 2 1 2 2 2 2 0 1 3
            [*]  2 1 2 1 2 1 2 2 1 1 2 2 2 1 1 1 2 2 2 1 1 0 3
            [-]  3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 0
            ;
            typeset *a = protpars:all;
endblock;
```

PROTPARS

(different matrices for different genetic codes)

BLOSUM62 Amino Acid Log-odd Substitution Matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C sulfhydryl
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P small hydrophilic
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D acid, acid-amide
E	-4	0	-1	-1	-1	-2	0	2	5												E and hydrophilic
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R basic
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I small hydrophobic
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y aromatic
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

xij < 0 freq. less than chance
xij = 0 freq. expected by chance
xij > 0 freq. great than chance

Note:
prob(no change) is high (>0)
prob(change to similar aa) is neutral (~0)
prob(change to different size aa) is low (<0)
prob(change to different charge aa) is low (<0)

ProtTest