Volume 10 Number 17 1982 Nucleic Acids Research

Hierarchical analysis of influenza A hemagglutinin gene sequences

David J.Lipman*, Temple F.Smith+, Richard J.Beckman† and Michael S.Waterman†

*Mathematical Research Branch, N.I.A.D.D.K.,N.I.H. Bethesda, MD 20205, +Northern Michigan University, Marquette, MI 49855, and †Los Alamos National Laboratory, Los Alamos, NM 87544, USA

Received 26 March 1982; Revised 22 June 1982; Accepted 26 July 1982

ABSTRACT

Five recently sequenced hemagglutinin genes from Influenza A virus strains are studied for similarities in a hierarchical fashion. The sequences are compared for similarity, first on the level of sequence homology, and then on several progressively more general levels. Though the HA1 subsequences contain regions where homology drops to that of a Monte Carlo generated reference value, subsequent tests reveal great similarity due to constraints on the level of amino acid sequence. Other tests detect statistically significant differences between subtypes due to constraints acting below the level of amino acid sequence, such as the 2° structure of the viral RNA, or involving translation of the mRNA. The general applicability of the hierarchical approach to sequence analysis is discussed.

INTRODUCTION

Influenza A virus, a negative strand RNA virus with a segmented genome, employs a rather unique strategy against its host. When the density of the susceptible host population drops to a critical level, a new strain appears. This strain can then infect previously resistant individuals. The major antigenic determinants of the virus (antibodies to which neutralize the virus) reside on the hemagglutinin (HA) protein, a glycoprotein of the viral envelope.

There are two mechanisms which can generate a new strain, antigenic shift and antigenic drift. Antigenic shift is responsible for new pandemic strains. The proposed mechanism is the acquisition of an HA gene segment from a strain previously infecting, for example, swine,by a strain capable of infecting humans. Drift, which is responsible for the epidemic strains appearing every 2-3 years, involves point mutations in the HA

0305-1048/82/1017-5375$2.00/0

gene segment causing sufficient alteration in the antigenic determinants of the protein to allow the virus escape from neutralization by host antibodies to the previous epidemic strain. Hemagglutinins are classified into subtypes; similar subtypes are thought to have evolved from the same pandemic strain (for review see 1).

The HA protein is cleaved into two segments, the HA1 and HA2. The HA1 has been found to vary more between strains and there is evidence that it contains the sites responsible for the antigenic differences between strains (2).

Recently the HA gene segments for five strains of influenza have been sequenced; A/Aichi/2/68(H3), (3), A/NT/60/68/29C(H3), (4), A/Victoria/3/75/(H3), (5), A/Japan/307/57(H2), (6), and FPV(Hav1), (7). In this paper we analyze these sequences with several methods of general use. Our approach is to progress from the most specific method of sequence analysis, homology analysis, to successively more global analyses. This hierarchical analysis allows us to dissect out various functional constraints acting on the sequences.

METHODS

Homology searches

Several algorithms are now available for finding an optimal alignment of two nucleic acid sequences (8,9). An alignment is determined to be optimal if it has the best score, under the scoring rules of the algorithm, among all possible alignments of the two sequences, allowing gaps in either sequence. These algorithms are useful in searching for the maximum global homology between two sequences. Smith & Waterman have recently developed an algorithm (10) which identifies maximally homologous subsequences among larger sequences. This algorithm is useful in searching for local homologies because the maximally homologous subsequences are not always aligned in a global homology search. The scoring rules for either global or local homology searches were as follows: Each nucleotide match increases the score by 1.0 and each mismatch decreases the score by 0.5 (or 0.333). Each gap of length k decreases the score by $1.0 + 0.5 * k$ (or $1.0 + 0.333 * k$).

A homology algorithm, by itself, can not assign a statistical significance to an alignment. The probability of an alignment score for two sequences occuring by chance alone can be determined as follows: Each of the sequences is randomly shuffled and the score of the optimal alignment is determined as before. This process is repeated a number of times in order to calculate a mean and standard deviation for the shuffled sequences. The mean ± the standard deviation is the Monte Carlo reference value. Statistical significance for the alignment can thus be estimated with respect to the Monte Carlo reference value.

We have examined the hemagglutinin sequences for both global and local maximum homologies. To reduce computer time, and to search for repeats, we have divided each HA gene into overlapping 240 base segments (1-240,121-360,241-480,...). When comparing genes, all pairwise comparisons of the 240 base segments were made. To calculate the Monte Carlo reference value, over 250 (240 by 240 length) shuffled sequence comparisons were made.

Contingency table analysis

The doublet frequencies of a sequence are the basis of a more global method of sequence analysis. Doublet frequencies were often used to compare nucleic acids before rapid sequence analysis methods were available (see 11). More recently, characteristic patterns in the doublet frequencies of sequenced nucleic acids have been noted by Nussinov (12). Smith,et al. (13) have used doublet frequencies in characterizing functional domains of nucleic acid sequences.

A 4 X 4 contingency table for a sequence can be constructed by tabulating the overlapping doublets in a sequence. Sequences are then compared by their respective contingency tables. The simplest and most straightforward approach for comparing sequences by their contingency tables is the chi square method described by Elton (14). This method takes into account the fact that the doublet counts are not independent (since they overlap), but are really transition counts, i.e. the contingency tables tell how often a T, for example, is followed by each of the 4 bases. Thus, several sequences are compared by their contingency tables using the conventional chi square formula:

$$\text{chi square} = (O - E)^2 / E$$

where O = the observed doublet counts in the contingency table of a sequence;

E = the estimated doublet counts utilizing the pooled contingency tables of all of the sequences in the comparison.

A more sophisticated approach for comparing sequences by their contingency tables employs the loglinear model (15). This approach also calculates a chi square value, but allows more flexibility in comparing sequences. For example, one could simultaneously compare the doublet and triplet count contingency tables of several sequences. Both of these methods have an advantage over homology analysis in that one may simultaneously compare any number of sequences. For the comparisons presented in this paper, either method would be satisfactory. The results we report are due to the loglinear model, but we have also used the simpler chi square test and found nearly identical chi square values in all cases. We have used the loglinear model because of the added flexibility of this method, and because it is available in the B.M.D., and S.P.S.S. statistical packages.

A variation of the analysis of sequences by their contingency tables allows one to dissect out the constraints on a sequence due to the amino acid sequence. For a sequence which codes for protein, one may construct three new sequences. One constructs a sequence comprised of only the first (or second or third) base of each consecutive codon. Thus constructed, these sequences will be referred to as Position 1, Position 2, or Position 3 sequences. In turn, one can then construct doublet frequency contingency tables from these position specific sequences. These new tables reflect constraints on the protein coded by the gene to varying degrees. For example, changes in the third position of the codon affect amino acid sequence very little. Thus, the table from the third position sequence reflects constraints on the gene independent of protein function, such as the secondary structure requirements of the viral RNA. By dissecting out the effect of the third codon position, the tables for first or second position sequences are strongly

influenced by amino acid sequence and thus more specifically reflect constraints on gene sequence due to protein function.

Analysis by informational constraints

One may compare sequences by measures which characterize perhaps the most general constraints acting on a sequence. These are the constraints toward nonuniform base composition and on the ordering of the bases. These measures are based on the statistical properties of the singlet and doublet counts of a sequence. Lipman & Maizel (16) used these measures to characterize a broad range of coding sequences and found that the general constraints of mitochondrial sequences were different from all other sequences examined. The eukaryotic coding sequences could be divided into two groups by their general constraints; those with introns, and those without introns.

Using the singlet and doublet counts of a sequence one can calculate two measures, D1, the divergence from 25% each base, and D2, the divergence from statistical independence of the bases. Independence of the bases is defined as:

$$p_{ij} = p_i * p_j$$

where p_{ij} = the probability of doublet ij, p_i = the probability of singlet i, and p_j = the probability of singlet j. The formulae for D1 and D2 are derived from Information Theory (17), and are as follows:

$$D1 = 2 - \sum_{i=1}^{4} (n_i/N)\log(n_i/N)$$

$$D2 = \sum_{i,j=1}^{4} (n_{ij}/N)\log[(n_{ij}*N)/(n_i*n_j)]$$

where n_i = number of singlet i;
n_j = number of singlet j;
n_{ij} = number of doublet ij;
N = number of bases in the sequence.

If the logarithms are to base 2 then D1 and D2 are measured in bits of information. D1 equals 0 if the base composition is 25% each base. If a sequence consists of all A's then D1 is equal to 2 bits. The upper limit for D2 is also 2 bits. If the bases in a sequence are independent, then D2 = 0. As D2 increases, the

probability of, for example, a T at a particular sequence position increasingly depends on at least its nearest neighbors. Thus D1 characterizes the constraints on a sequence toward nonuniform base composition, and D2 characterizes the constraints on the ordering of the bases.

RESULTS

Homology analysis

As described in Methods, for each pair of hemagglutinin genes, we have analyzed all pairwise comparisons of overlapping 240 base blocks for local homology. In all cases, for each block of one gene, the maximally homologous subsequence was found in the corresponding 240 base block of any other gene. Thus block 241-480 of FPV aligned best with block 241-480 of all other sequences. In Table I we report the local homology scores of the corresponding 240 base subsequences in the indicated pairwise comparisons. For example, a local homology score of 43.1 was obtained from the comparison of subsequence 121-360 (centered at position 240) of A/Japan versus subsequence 121-360 (centered at position 240) of FPV. The Monte Carlo reference value was 30 ± 4.5. Thus a homology score of 75, for example, is 10 standard deviations greater than the Monte Carlo reference. The search for global homology gave virtually identical results, and no repeated sequences of significance were noted. The local homology values for the three pairwise comparisons of A/Victoria(H3), A/Japan(H2), and FPV(Hav1) are plotted versus subsequence midpoint on figure 1. From Table I, and figure 1, it can be seen that regions of gene coding for the HA2 are very similar in all strains, while comparison between subtypes of the HA1 subsequences reveals regions only slightly more similar than the reference value.

On the average the most poorly conserved subsequences had midpoint positions at base #360 and base #480. This corresponds to a region spanning bases #240 to #600 and includes the probable antigenic sites of the HA protein (18). There is an abrupt rise in conservation moving from the HA1 to the HA2 subsequence. There is also a drop in conservation at the end corresponding to the region containing the hydrophobic tail of the HA2.

TABLE 1 Homology Scores of Subsequences *

Position of Subsequence Midpoint	Japan(H2) vs. FPV(Havl)	Victoria(H3) vs. FPV(Havl)	Victoria(H3) vs. Japan(H2)	Japan(H2) vs. NT(H3)	FPV(Havl) vs. NT(H3)	FPV(Havl) vs. Aichi(H3)	Aichi(H3) vs. Japan(H2)	Victoria(H3) vs. Aichi(H3)	Aichi(H3) vs. NT(H3)	Victoria(H3) vs. NT(H3)
120	50	59.3	44.1	44.6	59.6	57.7	43.6	222.4	233.7	216.1
240	43.1	43.8	48.3	49.5	45.4	45.4	49.5	224	237	221.0
360	50.7	24.9	31.3	30.2	25.7	25.3	30.2	225.6	237	222.6
480	43.6	32.6	33.5	31.8	30.8	29.3	33	219.2	233.8	217.8
600	49	38.7	58.2	51.7	35.6	37.0	51.3	212.8	232.2	213
720	43.2	44.8	59.7	55.9	45.2	44.2	50.7	218.2	232.2	214.6
840	41.5	34.8	54.2	59.3	34.6	32.8	55.2	222.4	233.8	216.2
960	60.6	59.5	69.3	75.5	61.6	60.2	72.5	224	237.0	221
1080	87.8	108	82.8	82.2	109.7	108.3	81.2	232	237.0	229
1200	79.2	103.8	71.4	69.5	105.8	105.8	68.2	235.2	237	232.2
1320	83.2	100.8	76.2	80.4	100.4	98.6	79.0	232.0	237	229
1440	101	122.4	94.3	97.5	124.4	124.6	95.9	227.2	233.8	227.4
1560	75.1	89.0	68.0	69.3	87.8	90.6	66.8	230.4	232.2	229
1680	36.6	75.6	38.0	35.7	57.1	71.3	39.2	198.1	168.4	163.6

* Monte Carlo reference value = 30 ± 4.5

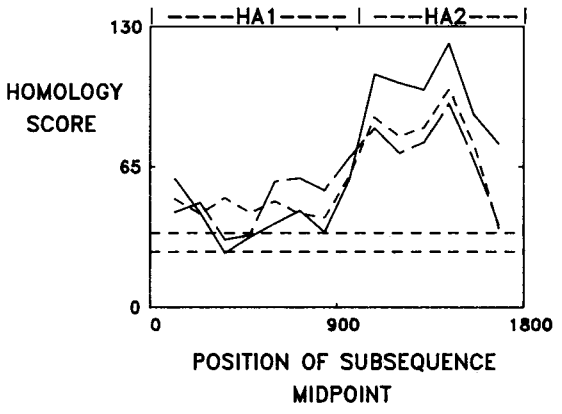

Figure 1
Homology score vs. position of subsequence midpoint. A/Victoria(H3) vs. FPV(Hav1), solid line; A/Japan(H2) vs. FPV(Hav1), dashed line; A/Victoria(H3) vs. A/Japan(H2), broken line. Horizontal dashed lines indicate Monte Carlo reference value of 30±4.5. Subsequences are 240 bases long and overlap by 120 bases.

Contingency table analysis

The homology search in the previous section detected statistically significant homology between subtypes in almost all segments of the HA2. Far less homology was detected between subtypes in virtually all segments of the HA1. In Table II are the results of a contingency table analysis of the HA1 and HA2 subsequences from A/Japan(H2),FPV(Hav1),and A/Victoria (H3). The null hypothesis is that the contingency tables of the indicated subsequences represent random samples from the same distribution. In the first case, we compared contingency tables from all three HA2 subsequences and found them statistically indistinguishable (p=.45). This was to be expected from the results of the homology analysis. In the second case, we compared the contingency tables of all three HA1 subsequences and found them statistically indistinguishable as well (p=.79). This is an unexpected result considering the findings of the homology analysis. In the third case, we compared the contingency table constructed from all three HA1 subsequences to the contingency table constructed from all three HA2 subsequences. The null hypothesis can be rejected (p=.001), and thus the HA1 and HA2

Table II

Comparison of contingency tables for intact subsequences: Sequences are A/Victoria(H3),FPV(Hav1),A/Japan(H2).

Null Hypothesis: The contingency tables constructed from the indicated subsequences are drawn from the same distributions.

CASE	Chi Sq.	df[1]	p[2]
All three HA2 subsequences	24.11	24	.45
All three HA1 subsequences	18.22	24	.79
All three HA1 subsequences versus all three HA2 subsequences	32.6	12	.001

1. df = degrees of freedom
2. P = The probability of observing, if the null hypothesis is true, a chi sq. value at least this large

subsequences are statistically distinguishable by their contingency tables. The doublets which contributed most to the differences between HA1 and HA2 contingency tables are shown in Table III. "CC" was involved in every case.

Position 1, Position 2, and Position 3 sequences were constructed from each of the above HA subsequences. The contingency tables for these position specific sequences were analyzed and are shown in Table IV. For the HA2 subsequences, the contingency tables from the Position 2 sequences are most similar and the tables from the Position 3 sequences are least similar, however, in no case are there statistically significant differences between the contingency tables of the HA2 position specific sequences. For the HA1 subsequences, this pattern also holds: Position 2 tables were most similar and Position 3 tables were least similar, however, in this case, the Position 3 contingency tables are statistically distinguishable (p=.06).

The opposite of this pattern is seen when the position specific contingency tables for the HA1 subsequences are compared with the tables for the HA2 subsequences. The Position 2 contingency tables are most dissimilar while the Position 3 tables are most similar. There are statistically significant

Table III

Doublets making the largest contribution to the differences in the HA1 and HA2 contingency tables

VIRUS	Doublets			
FPV(Hav1)	CC	GC	UC	CU
A/NT(H3)	CC	AC	UA	CG
A/Victoria(H3)	CC	UU	UA	CG
A/Aichi(H3)	CC	AC	UA	CG
A/Japan(H2)	CC	AC		CU

Table IV

Comparison of contingency tables for position specific sequences: Sequences are A/Victoria(H3),FPV(Hav1),and A/Japan(H2)

CASE	Position	Chi Sq.	df	P
All 3 HA2 subsequences	3	18.17	24	.79
	2	8.24	24	.99
	1	15.02	24	.94
All 3 HA1 subsequences	3	35.67	24	.06
	2	5.66	24	1.0
	1	19.61	24	.72
HA1 vs. HA2 for all 3 subsequences	3	6.06	12	.91
	2	36.84	12	.00
	1	20.23	12	.06

differences between the Position 1 and Position 2 tables ($p=.06, p<.001$ respectively) while the Position 3 contingency tables are statistically indistinguishable ($p=.91$).

Informational constraints

In Figure 2 we have plotted D1, the constraints toward

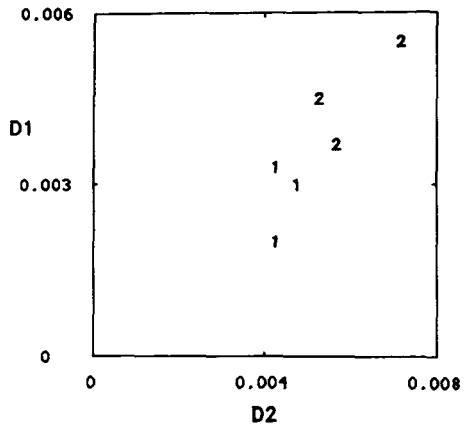

Figure 2 D1 versus D2 (in bits) for the HA1 and HA2 subsequences of A/Victoria(H3), FPV(Hav1), and A/Japan(H2). HA1,"1"; HA2,"2".

nonuniform base composition versus D2, the constraints on the ordering of the bases, for the HA1 and HA2 subsequences of A/Victoria (H3), A/Japan (H2), and FPV (Hav1). The HA1's are labelled with a "1", and the HA2's are labelled with a "2". There appears to be a consistent relationship of the D1 to the D2 for all of the subsequences, however, D1 and D2 are greater for the HA2 subsequences.

Base composition

Before discussing the above results, an interesting observation on the base composition of the entire influenza genome should be noted. J.L.King has calculated an estimate of the base composition which has the potential to code for each amino acid with equal probability (19). In Table V , the resulting base composition is seen opposite the base composition of A/PR8/34 (20). The close agreement in the base compositions is striking and without ready explanation. It is likely that the answer may be found in the strategy of influenza virus, that of periodic change. One has the most potential messages (proteins) when the components of the messages (amino acids) are equiprobable. Thus, this base composition could be explained teleologically. How this base composition is actually selected for and maintained is unknown.

DISCUSSION

There is a useful relationship between the homology,

Table V

Base composition of the entire genome of A/PR8 vs. that of a gene having the potential to code for each amino acid with equal probability (King base composition) (see text)

BASE	A/PR8	King Base Composition
C	23.6±.2	25.0
A	23.3±.8	23.3
G	21.2±.6	19.4
U	31.8±1.6	32.3

contingency table and informational constraints analyses. If two sequences have high homology values, their contingency tables will be statistically indistinguishable. Two sequences may not have high homology values, the HA1 subsequences for example, but their contingency tables may be statistically indistinguishable. The contingency table analysis focuses on information about the sequences such as particular amino acids that are strongly selected for or against, or certain constraints on the evolution of the nucleic acid sequence not dependent on the amino acid sequence.

If two contingency tables are statistically indistinguishable, differences in the D1 and D2 values calculated from them are not likely to be statistically significant. However, two sequences whose contingency tables have statistically significant differences may have very similar D1 and D2 measures. For example, a particular secondary structure may be satisfied by a wide variety of sequences as long as proper areas are complementary and the total structure has the same free energy. Though the homology and contingency table analyses may not find any similarity between the sequences, the D1 and D2 measures may detect this similarity in underlying constraints.

The HA1 subsequences account for most of the differences in homology between the strains studied. Indeed, the regions coding for the HA1's are, in many sections, no more similar than the

Monte Carlo reference. Despite this, the chi square value for the HA1 contingency tables is smaller than that for the HA2 contingency tables. Thus, the HA1 subsequences show similar constraints not apparent in the homology analysis. The results of the position specific analysis suggest these similar constraints are acting on the level of the amino acid sequence because the Position 2 tables were most similar, while the Position 3 tables were most dissimilar. Min Jou, et.al. (5), noted similarities in the HA1 amino acid sequences: conservation of cysteine positions and conservation of position and length of hydrophobic stretches. They concluded that the overall architecture of the hemagglutinin was not flexible. Therefore, despite great differences in amino acid sequence, the functional and structural similarities in the HA1 proteins are the basis for similarities detected by the contingency table analysis.

Statistically significant differences were found between the HA1 and HA2 contingency tables. Statistically significant differences were also found between the HA1 and HA2 Position 2 (and Position 1) contingency tables. However, the Position 3 contingency tables were not statistically distinguishable. This is because the HA1 and HA2 Position 3 contingency tables for <u>each</u> HA gene show,on the average, the least differences (data not shown). The above results suggest that the HA1 and HA2 subsequences show differences due to constraints acting at the protein level, which is expected since the HA1 and HA2 proteins certainly evolve under different functional constraints. However, the HA1 and HA2 subsequences show similarities due to constraints acting below the level of amino acid sequence. Such constraints may involve the 2° structure of the vRNA or the rate and/or accuracy of translation. Grantham, in his analysis of codon frequencies (21), has found the third position most powerful in distinguishing mRNAs. He has also hypothesized that the third position was involved in the control of the transcription rate (22). If this is indeed true, then the three subtypes may differ in the rate of transcription of the hemagglutinin gene - this may be related to differences in virulence between subtypes.

The HA1 polypeptide contains a large region constituting the

immunogenic sites of the virus (2). In this region, favorable mutations are those which cause antigenic difference. This is a constantly varying constraint which should be contrasted with those on a region responsible for catalyzing a reaction or maintaining a specific structure. This "constraint" has a randomizing effect on the HA1 nucleic acid subsequences. Thus, though there is a similar relationship of D1 (constraints toward nonuniform base composition) to D2 (constraints on the ordering of the bases) for the HA1 and HA2 subsequences, the magnitude of oth constraints is less for the HA1 subsequences.

CONCLUSION

Conventional comparative sequence analysis, involving a search for homologies, would reveal the high degree of sequence conservation in the HA2 subsequences, and the relatively low degree of conservation in the HA1 subsequences. Conventional analyses, however, would not reveal, in such a direct manner, the strong similarities in constraints on the HA1 subsequences, nor could they assign these constraints to the level of protein function. Conventional analyses would certainly reveal the difference between the HA1 and HA2 subsequences, but could not reveal the significant similarities in the constraints operating below the level of protein function.

Hierarchical analysis provides a simple, integrated approach for comparative sequence analysis, and may prove even more useful in systems where the biological properties are not so well understood.

ACKNOWLEDGEMENTS

D.J.L. would like to thank Dr.P.Palese for the opportunity to develop some of the ideas presented in this paper.

REFERENCES

1. Webster,R.,Laver,W.,(1975),The Influenza Viruses and Influenza,Academic Press,pp.269-314.
2. Laver,W.,Air,G.,Webster,R.,Gerhard,W.,Ward,C.,Dopheide,T., (1980),in Structure and Variation in Influenza Virus,Elsevier North Holland,Inc.,pp.295-306.
3. Verhoeyen,M.,Fong,R.,Min Jou,W.,Devos,R.,Huylebroeck,D., Saman,E.,Fiers,W.,(1980),Nature,286,pp.771-776.
4. Both,G.,Sleigh,M.,(1980),Nuc.Ac.Res.,8,pp.2561-2575.

5. Min Jou,W.,Verhoiyen,M.,Devos,R.,Saman,E.,Fong,R., Huylbroeck,D.,Fiers,W.,(1980),Cell,19,pp.683-696.
6. Gething.M.,Bye,J.,Skehel,J.,Waterfield,M.,(1980),in Structure and Variation in Influenza Virus,Elsevier North Holland,Inc., pp.1-10.
7. Porter,A.,Barber,C.,Carey,N.,Hallewell,R.,Threfall,G.,Emtage, J.,(1979),Nature,282,pp.471-477.
8. Needleman,S.,Wunsch,C.,(1970),J.Mol.Bio.,48,pp.443-453.
9. Sellers,P.H.,(1979),Proc.Nat.Acad.Sci.,USA 76,p.3041.
10. Smith,T.F.,Waterman,M.S.,(1981),J.Mol.Bio.,147,pp.195-197.
11. Russell,G.,McGeoch,D.,Elton,R.,Subak-Sharpe,J. (1973),J.Mol.Evol.,2,pp.277-292.
12. Nussinov,R.,(1980),Nuc.Acids Res.,8,pp.4545-4562.
13. Smith,T.,Waterman,M.,Sadler,J.,(1982), submitted to Cell.
14. Elton,R.,(1975), J.Mol.Evol.,4,pp.323-346.
15. Bishop,Y.,Feinberg,S.,Holland,P.,(1978),Discrete Multivariate Analysis,MIT Press.
16. Lipman,D.,Maizel,J. (1982),Nuc.Acids Res.,10,pp.2723-39.
17. Gatlin,L.,(1972),Information Theory and the Living System, Columbia University Press.
18. Wilson,I.A.,Skehel,J.J.,Wiley,D.C.(1981),Nature 289,pp.373-378.
19. King,J.,(1972),Proc. of the Berkeley Symp. on Math.Stat. and Prob.,V,U. of California Press.
20. Ritchey,M.,Palese,P.,Kilbourne,E.,(1976),J.of Virology,18 pp.738-744
21. Grantham,R.,Gautier,C.,Gouy,M.,Mercier,R.,Pave,A.,(1980) Nucleic Acids Research,8,pp. r49-62.
22. Grantham,R.,Gautier,C.,Gouy,M.,Jacobzone,M.,Mercier,R., (1981),Nucleic Acids Research,9,pp. r43-74.