Gene Mapping, D. Bergsma, Ed. (vol. 10, No. 3 of Birth Defects: Original Article Series) (Symposia Specialists, Miami, 1974), p. 132.
18. T. R. Chen, F. A. McMorris, R. Creagan, F. Rucciuti, J. Tischfield, F. H. Ruddle, *Am. J. Hum. Genet.* **25**, 200 (1973).
19. J. A. Tischfield, R. P. Creagan, F. Ricciuti, F. H. Ruddle, in *Human Gene Mapping*, D. Bergsma, Ed. (vol. 10, No. 3 of Birth Defects: Original Article Series) (Symposia Specialists, Miami, 1974), p. 164.
20. V. G. Dev, P. A. Miller, P. W. Allderdice, O. J. Miller, *Exp. Cell Res.* **73**, 259 (1972).
21. C. B. Laurell, *Anal. Biochem.* **15**, 45 (1966).
22. L. R. Weitkamp, D. L. Rucknagel, H. Gershowitz, *Am. J. Hum. Genet.* **18**, 559 (1966).
23. M. L. Petras, *Biochem. Genet.* **7**, 237 (1972).
24. F. Ricciuti and F. H. Ruddle, *Nature (Lond.)* **241**, 186 (1973).
25. C. Boone, T. R. Chen, F. H. Ruddle, *Proc. Natl. Acad. Sci. U.S.A.* **69**, 510 (1972).
26. F. H. Ruddle and E. A. Nichols, *In Vitro* **7**, 120 (1971).
27. F. A. McMorris, T. R. Chen, F. Ricciuti, J. Tischfield, R. Creagan, F. H. Ruddle, *Science* **179**, 1129 (1973).
28. Y. H. Tan, J. Tischfield, F. H. Ruddle, *J. Exp. Med.* **137**, 317 (1973).
29. A. Korner and J. R. Debro, *Nature (Lond.)* **173**, 1067 (1956).
30. This work was supported by grants from PHS (GM 09966) and NSF (GB 34303). H.P.B. is the recipient of awards from the Swiss National Foundation and the PHS (TW-01896). We thank Elizabeth Nichols, Annelis Bernhard, and Patricia Eager for their excellent technical assistance.

11 March 1974; revised 14 May 1974

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.*

I present here an improved formula for difference between amino acids which identifies the chemical factors that individually correlate best with evolutionary exchangeability of protein residues. I also estimate the extent to which observed exchanges can be explained by conservation of these factors.

Disagreement exists over what mainly directs gene evolution at the molecular level. Essentially the controversy is between randomness and physicochemical determinism. Randomness proponents (*1*) believe that proteins have evolved by chance fixation of "neutral mutations" (substitutions of one amino acid for another such that the original and substituted genes have equal adaptive values). The other viewpoint is that physicochemical forces are the principal determinants of molecular evolution. There has not been enough effort, though, to specify these forces quantitatively. I now relate overall chemical difference, approximated from side chain properties, to evolutionary difference between amino acids, implied by their "mutation rates." Although precise measures are yet to come, Dayhoff (*2*) and McLachlan (*3*) have estimated frequencies of exchange between protein residues. For the correlations in this report we use relative substitution frequency (*RSF*) of McLachlan, which is the largest sampling so far (*3*).

Methods for assessing total difference between amino acids are few. Sneath's index (*4*) contains too many characters for satisfying correlations [although Clarke (*5, 6*) improved this among amino acids whose codons have two common bases by judicious weighting of the characters], while Epstein's formulas (*7*) considering only size and polarity class yield identical differences for many pairs of amino acids. The number of base changes (in the messenger RNA's, derived by decoding the proteins) needed to give the same amino acid sequence is used to construct phylogenetic trees. But amino acid substitution frequencies cannot be rationalized this way (*2, 3*). Further, the minimum base change method poorly reflects homology and consistently underestimates the total number of fixations inferred from phyletic data (*8*). The Sneath or Epstein difference correlates better against *RSF* than does the minimum number of base changes, but their correlation coefficients remain weaker than −.5 (see below). This leaves considerable room for chance determination of residue exchanges.

Several amino acid side chain properties correlate appreciably with *RSF*. The three strongest correlators are composition, polarity, and molecular volume. These last two properties are from published data [see (*9*)]. Composition, *c*, is defined as the atomic weight ratio of hetero (noncarbon) elements in end groups or rings to carbons in the side chain. Such ratios are a simple and sensitive way of reflecting composition differences between amino acids. As an example, for the

Table 1. Values for properties in amino acid difference formula and correlation results (*10*). The correlation coefficient R_{19} was obtained from linear regression of $\log RSF$ on D within each group of 19 amino acid pairs; R_{190} was obtained with all 190 pairs by regression of $\log RSF$ on D given by indicated single property or combination of properties; R_S is the Spearman rank correlation coefficient (*12*). For interproperty correlations, the 190 differences for one property were ranked against those for the other property; the interproperty R_S values are *cp*, .435; *cv*, .092; and *pv*, .008. The $\overline{D}$ below each property column is the average chemical distance given by that property alone; thus $\overline{D}_c = \Sigma[(c_i - c_j)^2]^{1/2}/190$. The inverse mean weighting factors are $\alpha = (1/\overline{D}_c)^2 = 1.833$; $\beta = (1/\overline{D}_p)^2 = 0.1018$; $\gamma = (1/\overline{D}_v)^2 = 0.000399$.

Amino acid	Property			$-R_{19}$	Formula	$-R_{190}$	$-R_S$
	c	*p*	*v*				
Ser	1.42	9.2	32	.76	*c*	.49	.49
Arg	0.65	10.5	124	.68	*p*	.47	.55
Leu	0	4.9	111	.92	*v*	.37	.33
Pro	0.39	8.0	32.5	.67			
Thr	0.71	8.6	61	.63	*cp*	.61	
Ala	0	8.1	31	.75	*cv*	.61	
Val	0	5.9	84	.86	*pv*	.63	
Gly	0.74	9.0	3	.66			
Ile	0	5.2	111	.89	*cpv*	.72	.765
Phe	0	5.2	132	.83			
Tyr	0.20	6.2	136	.64			
Cys	2.75	5.5	55	.31			
His	0.58	10.4	96	.53			
Gln	0.89	10.5	85	.79			
Asn	1.33	11.6	56	.87			
Lys	0.33	11.3	119	.76			
Asp	1.38	13.0	54	.93			
Glu	0.92	12.3	83	.82			
Met	0	5.7	105	.58			
Trp	0.13	5.4	170	.58			
$\overline{D}$	0.739	3.134	50.06				

Ser (*10*) side chain –COH, $c = 17/12$ (atomic weight of hydroxyl over that of carbon). For Lys –CCCCNH$_2$, $c = 16/48$ (amino over four carbons). The composition difference ($c_{Ser} - c_{Lys}$) is, therefore, $1.42 - 0.33 = 1.09$. The 190 ($c_i - c_j$) values (or those for any other property) are calculated and then correlated against the corresponding 190 *RSF* values. The following formula, by combining the three properties, estimates overall difference, *D*, between any two protein residues *i* and *j*

$$D_{ij} = [\alpha(c_i - c_j)^2 + \beta(p_i - p_j)^2 + \gamma(v_i - v_j)^2]^{1/2}$$

where c = composition, p = polarity, and v = molecular volume. The dependence of one property on another is shown in Table 1 as interproperty correlations. In a Euclidean space having these properties as axes, D_{ij} would be the distance between the amino acids. The properties are not assumed to be mutually independent; the axes are made orthogonal to facilitate distance calculations. Each property is weighted by dividing by the mean distance found with it alone in the formula. Thus, the constants α, β, and γ are squares of the inverses of the *D*'s below columns *c*, *p*, and *v* of Table 1 (*9*).

This equation has been solved for each possible pair of amino acids by substituting in the appropriate property values from Table 1. The resulting $\bar{D}$'s (Table 2) were then correlated with *RSF* (*3*). Correlation coefficients (R_{19}) for each set of 19 amino acid pairs appear in Table 1, as do those (R_{190}) for all 190 pairs obtained with single parameters or combinations of parameters. Coefficients for Spearman rank correlation, which does not depend on linearity between *D* and *RSF*, are also shown. Correlations given by this formula are contrasted (see Table 2) to the weak agreement other difference indexes show with *RSF*.

A circularity may seem to exist in the manner of building the formula and interpreting the correlation results, but at the outset one has no guarantee of high correlation between *RSF* and *D*. That is, in the absence of an appropriate evolutionary theory there is no reason why any amino acid property should correlate with *RSF* or why certain properties could be combined to give higher correlations. Thus, the correlations obtained do show that evolutionary amino acid replacements depend highly on chemical factors.

The variation in correlation coefficients (R_{19} in Table 1) of the 20 groups suggests that the formula is not yet perfect, nine groups having *R* weaker than −.75. The extent to which substitutions are determined by chance could, of course, vary between groups. Thus Cys, with the weakest *R*, would "just happen" in protein sites more often, relatively, than other residues because it forms more neutral pairs. More likely, though, Cys has properties important to its function (ability to form disulfide bridges, for instance) that are not well reflected in the formula. The weak correlations are, therefore, not necessarily proof that Cys has often been fixed or replaced in protein by chance. Conversely, *R* of −.75 or better for the other 11 groups must mean that randomness has had small part in their evolutionary exchanges.

As seen in the legend to Table 2, there is some agreement between codon relatedness and amino acid substitution rate. However, difference based on *c*, *p*, or *v* alone correlates better with *RSF* than does minimum number of base changes. Protein evolution is more related to property differences between substituted and substituting amino acids than to a priori probabilities for minimizing codon changes (*11*). Many base changes due to chromosomal and replication errors evidently occur and reoccur before any one is fixed in the population, so that mutations are not strongly selected by the codon transformations they require. Note that two (Tyr-Trp and Phe-Trp) of the three highest *RSF* values found by McLachlan (*3*) are for amino acid pairs without two common bases in their codons.

The conclusions stated above are subject to the limitations of statistical inference, but correlations of this magnitude with all 190 amino acid pairs cannot be due to chance. The formula correlates better against log *RSF* ($R = -.72$) than *RSF* ($R = -.66$) since log *RSF* has a less skewed distribution than raw *RSF* (*D* values approximate a normal distribution). The distribution-independent Spearman rank coefficient is −.765 for *RSF* and the *D* values of Table 2. With 190 degrees of freedom this leaves no doubt of a strong constraint on protein residue exchanges by chemical factors. In fact, rank correlation between *c*, *p*, or *v* alone and *RSF* is such (see Table 1) that this doubt is already extremely small (*12*).

Can this correlation be improved? Sample size, homology uncertainties, and methods of calculating substitution frequencies (*2*, *3*) weaken the precision of *RSF* values, which are averages over several subclasses of protein structure from any organisms. Surface conserva-

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*).

tion is less restrictive than that in the interior, and structural regularities (helixes, sheets, and bends) each prefer certain residues. This introduces indeterminacy since residue replacement rules are probably not identical in all subclasses. It is also conceivable that the link between amino acid properties and exchange rate varies somewhat with time and type of organism. In any case, the formula is an improvement, not a final solution. Adding other parameters helps correlations somewhat. The present parameters are the best set of three, but are not only ones that could appear in the formula.

R. GRANTHAM

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

References and Notes

1. M. Kimura and T. Ohta, *Nature (Lond.)* **229**, 467 (1971); T. Ohta and M. Kimura, *ibid.* **233**, 118 (1971); T. Yamazaki and T. Maruyama, *Science* **178**, 56 (1972). See also B. Clarke (book review), *ibid.* **180**, 600 (1973).
2. M. O. Dayhoff, *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Silver Spring, Md., 1972), p. 89.
3. A. D. McLachlan, *J. Mol. Biol.* **61**, 409 (1971); *ibid.* **64**, 417 (1972), where *RSF* for each amino acid pair is given on p. 435.
4. P. H. A. Sneath, *J. Theor. Biol.* **12**, 157 (1966).
5. B. Clarke, *Science* **168**, 1009 (1970); *Nature (Lond.)* **228**, 159 (1970).
6. T. H. Jukes and J. L. King, *ibid.* **231**, 114 (1971).
7. C. J Epstein, *Nature (Lond.)* **215**, 355 (1967).
8. R. E. Dickerson, *J. Mol. Biol.* **57**, 1 (1971); T. Uzzell and K. W. Corbin, *Science* **172**, 1089 (1971).
9. Properties must be weighted for combining. If the values in Table 1 for the three properties are used in the formula without weighting ($\alpha = \beta = \gamma = 1$), correlation against *RSF* will give nearly the same result as when difference is determined by v alone. This is because expressing v in Å^3 inflates its importance in the correlation relative to c and p. Inverse mean weighting is a scale change in the distance space to accommodate to the arbitrary units of each property. Inverse mean weighting (legend to Table 1) gives $\alpha^{\frac{1}{2}} : \beta^{\frac{1}{2}} : \gamma = 68.7 : 16 : 1$ and $R = -.72$ for the D : log *RSF* regression. When the formula is optimized by computer to maximize the correlation between D and log *RSF*, the square roots of the weighting factors = 54 : 17.6 : 1 and $R = -.74$. Thus, the inverse mean scale change gives nearly optimum weights. Incidentally, no linear combination of c, p, and v correlates better with *RSF* than this distance formula. Glycine, which is undefined by the hetero element/carbon ratio (see text), has an arbitrarily assigned c (the mean of the 14 nonring side chains). Proline, whose delta C joins to alpha N, is counted as having an atom of N in its side chain. Polarity is averaged: $p = (PR + PA)/2$, where PR is the "polar requirement" of C. R. Woese [*Naturwissenschaften* **60**, 447 (1973)] and PA (adjusted to same scale as PR) = 13.66 — 14.85 R_F, where R_F is the amino acid mobility of A. A. Aboderin [*Int. J. Biochem.* **2**, 537 (1971)]. Side chain molecular volume is the residue volume (Å^3) of D. E. Goldsack and R. C. Chalifoux [*J. Theor. Biol.* **39**, 645 (1973)] minus the constant peptide volume (Gly residue — Gly side chain volume, the latter estimated as 3 Å^3). Proline's side chain sweeps out relatively less volume because it is pinned to alpha N; we therefore estimate the effective v_{Pro} as one-half its formal value. Ignoring the three refinements (taking into account the side chain volume of Gly in v determinations and the influence of Pro's side chain particularly on c and v) weakens the correlations slightly.
10. Amino acid residue abbreviations used in this report are: Arg, arginine; Leu, leucine; Pro, proline; Thr, threonine; Ala, alanine; Val, valine; Gly, glycine; Ile, isoleucine; Phe, phenylalanine; Tyr, tyrosine; Cys, cysteine; His histidine; Gln, glutamine; Asn, asparagine; Lys, lysine; Asp, aspartic acid; Glu, glutamic acid; Met, methionine; and Trp, tryptophan.
11. Further evidence is found by separating the 190 amino acid pairs into three classes according to codon relatedness and correlating D with *RSF* in each class. The Spearman rank coefficient (—.70) for correlation of D with *RSF* among the 101 pairs in the class with one common base is not greatly lower than that (—.82) among the 75 pairs having two common bases; however, that (—.43) among the 14 pairs with no common bases does help confirm a tendency for the code's structure to facilitate mutations to similar amino acids.
12. For rank correlation see: M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics* (Griffin, London, 1961), vol. 2, p. 476; S. Siegel, *Nonparametric Statistics for the Behavioral Sciences* (McGraw-Hill, New York, 1956), p. 202.
13. D. Chessel, J. Estève, C Gautier, J. M. Legay, and A. Pavé collaborated in this work. I also thank R. Jeanneau, E. Kahane, G. Labourdette, J. Parello, and H. Vogel (Montpellier); B. C. Clarke (Nottingham); A. D. McLachlan (Cambridge); B. Thomas (Lund); and C. Biémont, J. David, M. Jarry, J. Pontier, and C. Souchier (Lyon) for help.

19 October 1973; revised 12 June 1974 ■

Reverse Transcriptase in Normal Rhesus Monkey Placenta

Abstract. *Particles with the morphology of type C virus have been identified from primate placentas by electron microscopy. A reverse transcriptase (RNA-dependent DNA polymerase) was isolated and purified from microsomal pellets of two fresh placentas of rhesus monkeys in the early stages of gestation. This enzyme was biochemically similar yet immunologically distinct from the reverse transcriptases of known tumorigenic type C RNA viruses isolated from primates, but was immunologically related to a reverse transcriptase isolated from a type C virus obtained from normal baboon placenta. These particles may represent endogenous viruses and may function in the transfer of genetic information during embryogenesis.*

Several investigators, using electron microscopy, have reported that particles with the appearance of type C viruses bud from placental syncytial trophoblasts of the rhesus monkey (*1*), and of other primates (*2, 3*). Despite the presence of structures morphologically similar to oncogenic viruses, no mention has been made of any disease state in any of the maternal sources of the placentas (*1–3*).

The function of these type C particles is still unknown. Nor is it known whether they are biochemically similar to oncogenic type C viruses, whether they represent congenital transplacental infection by a tumor virus, or whether they are endogenous viruses that may be nononcogenic to the host.

We have confirmed the presence of type C viral particles in rhesus monkey placenta by electron microscopy; we now report the presence of a viral reverse transcriptase (RNA-dependent DNA polymerase) in two fresh placentas obtained from rhesus monkeys in early gestation. This enzyme is biochemically similar to but immunologically distinct from reverse transcriptases of known primate type C RNA tumorigenic viruses (woolly monkey sarcoma and gibbon ape lymphosarcoma viruses); nevertheless, it is immunologically related to a reverse transcriptase isolated from a virus propagated by cocultivation of normal baboon placental extract with heterologous cells.

Two rhesus monkey placentas (3.0 and 4.2 g) were obtained from normal mothers by cesarian section at days 33 and 36 of gestation, respectively (normal gestation is 165 days). In separate experiments, these tissues were homogenized in isotonic buffer, the nuclei and mitochondria were removed by differential centrifugation, and the resulting supernatants were centrifuged through a 25 percent sucrose "cushion" to obtain a microsomal pellet. The pellets were solubilized in high salt and detergent, and the nucleic acids were removed by passage of the extract over a fibrous diethylaminoethyl (DEAE)-cellulose column (*4*).

The placental extracts were then passed over a microgranular DEAE-cellulose column. Under certain conditions (*5*), viral reverse transcriptase and DNA polymerase II ($s_{20,w} = 3.4S$) elute from microgranular DEAE-cellulose columns at low salt concentrations (below 0.075*M* KCl) and may thus be separated from DNA polymerase I ($s_{20,w} = 6S$ to $8S$) and DNA-polymerase III (R-DNA polymerase) which elute at a higher ionic strength (between 0.10*M* and 0.35*M* KCl) (*6*). By virtue of their removal at this step, cellular DNA polymerases I and III are not sources of confusion in the sub-