quent pregnancies in families where there is already one offspring with XP. To minimize any familial idiosyncracies in nucleoside metabolism, amniocentetic cells should be compared to fibroblasts from parents and prior offspring.

JAMES D. REGAN, R. B. SETLOW
Biology Division,
Oak Ridge National Laboratory,
Oak Ridge, Tennessee 37830
MICHAEL M. KABACK
R. RODNEY HOWELL
Department of Pediatrics,
Johns Hopkins University School of
Medicine, Baltimore, Maryland 21205
EDMUND KLEIN
GORDON BURGESS
Department of Dermatology,
Roswell Park Memorial Institute,
Buffalo, New York 14240

References and Notes

[1] H. W. Siemens and E. Kohn, *Z. Indukt. Abstamm. Verebungsl.* **38**, 1 (1925).
[2] J. E. Cleaver, *Nature* **218**, 652 (1968); D. Bootsma, M. P. Mulner, F. Pot, J. A. Cohen, *Mutat. Res.* **9**, 507 (1970).
[3] R. B. Setlow, J. D. Regan, J. German, *Proc. Nat. Acad. Sci. U.S.* **64**, 1035 (1969).
[4] J. E. Cleaver, *ibid.* **63**, 428 (1969).
[5] M. M. Kaback, C. O. Leonard, T. H. Parmley, *Pediat. Res.*, in press.
[6] J. D. Regan, J. E. Trosko, W. L. Carrier, *Biophys. J.* **8**, 319 (1968).
[7] J. D. Regan, R. B. Setlow, R. D. Ley, *Proc. Nat. Acad. Sci. U.S.* **68**, 708 (1971).
[8] C. DeSanctis and A. Cacchione, *Riv. Sper. Freniat.* **56**, 269 (1932).
[9] M. M. Kaback and R. R. Howell, *N. Engl. J. Med.* **282**, 1336 (1970).
[10] S. E. Pfeiffer and L. J. Tolmach, *Cancer Res.* **27**, 124 (1967).
[11] Cells derived from amniocentesis may be in a much later stage of cellular aging than cells taken directly from biopsies of fetal or neonatal skin [see (*5*)]. The greater sensitivity of amniocentetic cells to 313-nm light after ultraviolet and incubation in thymidine compared to fetal or other skin culture cells may be a reflection of differential cellular age.
[12] We have tabulated weight-average rather than number-average molecular weights. For practical reasons all the samples were sedimented for the same length of time and the range of molecular weights in any particular gradient may be from 10^8 to 10^6. Since the molecular weights corresponding to fractions near the tcp of the gradient approach zero, the value of the number-average molecular weight is exquisitely dependent on the number of fractions at the top of the gradient which are excluded from the estimation of this average. (If all fractions were included, the number-average molecular weight would be zero.) The weight-average molecular weight is much less dependent upon this arbitrary decision. The use of the weight-average molecular weight obviously does not give the proper average number of breaks, but it does yield a reproducible set of numbers.
[13] W. D. Rupp and P. Howard-Flanders, *J. Mol. Biol.* **31**, 291 (1968); J. E. Cleaver and G. H. Thomas, *Biochem. Biophys. Res. Commun.* **36**, 203 (1969).
[14] The technical assistance of W. H. Lee and F. M. Faulcon is greatly appreciated. Research supported by the National Cancer Institute and by the AEC under contract with the Union Carbide Corporation, by NIH contract No. 69-79 and grants from the National Foundation–March of Dimes and the National Genetics Foundation to the Johns Hopkins University School of Medicine, and by grants from NIH (NIAID AI-09479-01) and the Office of Naval Research (ONR N00014-70-0105) to Roswell Park Memorial Institute.

18 April 1971

Amino Acid Composition of Proteins as a Product of Molecular Evolution

Abstract. *The average amino acid composition of proteins is determined by the genetic code and by random base changes in evolution. Small but significant deviations from expected composition can be explained by selective constraint on amino acid substitutions. In particular, the deficiency of arginine in proteins has been caused by constraint, during evolution, on fixation of mutations substituting arginine for other amino acids.*

The average amino acid composition of proteins can be predicted from the genetic code, if there is random arrangement of nucleotide bases within the genes (cistrons) (*1–3*). The predicted amino acid frequencies are accurate except for that of arginine, whose observed frequency is only half as large as that expected from random arrangement of bases (*2*, *3*). For other amino acids, the difference between observed and expected frequency is much smaller, and sometimes is negligible. According to King (*4*), the same amino acids occur more frequently than expected, and others less frequently than expected, whether proteins from mammals or from bacteria are sampled.

Amino acid composition is a product of molecular evolution. The overall agreement between observed and expected amino acid compositions suggests that amino acid substitutions in evolution were produced by random fixation of selectively neutral or nearly neutral mutations (*2*, *5*, *6*). Deviations from the expected frequency, such as the deficiency of arginine, must have been caused either by nonrandom mutations or, more likely, by selective constraints. In fact, many mutations appear to be deleterious, and are eliminated from the population by selection, although neutral mutations predominate among those mutants that contribute to molecular evolution and enzyme polymorphism (*7*). Some evidence suggests that functionally similar amino acids are substituted more frequently than less similar ones (*8–10*). We now report our efforts at clarifying the nature of the nonrandomness in the evolution of the amino acid composition in proteins.

Probably the simplest way of treating the process of evolutionary change in amino acid composition is to use the method of Markov chains. We use a 20 by 20 matrix giving transition probabilities for any one of the 20 amino acids to any other during a unit of time. Starting from a given amino acid composition, we can then compute the expected composition after *n* units of time by taking the *n*th power of the matrix. From an estimation of the ancestral sequences and with the use of comparative data on amino acid sequences in homologous proteins, a transition probability matrix could be constructed. However, for such a matrix, a large body of data would have to be compiled. To avoid these difficulties, we have tentatively used the "mutation probability matrix" of Dayhoff *et al.* (*9*). These workers counted 814 "accepted point mutations" among closely related sequences from cytochrome c, globins, virus coat proteins, chymotrypsinogen, glyceraldehyde-3-phosphate dehydrogenase, clupeine, insulin, and ferredoxin. From these mutations, they constructed an "accumulated matrix of accepted point mutations" (*9*, figure 9-3).

They multiplied this matrix by the overall "mutability" of individual amino acids per unit time to obtain the mutation probability matrix. This matrix (M) is reproduced as Fig. 1.

In contrast, if we assume that mutant base substitutions are completely random, we can construct a corresponding transition probability matrix from the genetic code (assuming only one-step mutations). This matrix (R) is shown in Fig. 2.

There are several differences between M and R. Particularly noteworthy is the fact that the transition to Arg (*11*) from a number of amino acids, such as Cys, Gly, Ile, Leu, Pro, Ser, Thr, and Trp, is much lower in M than in R. This indicates that these single-step mutations were seldom accepted by natural selection. The reverse transitions are also somewhat restricted, but not so severely. These transition rates can account for the significantly lower frequency of Arg than is expected from random base arrangement. The biochemical explanation for such selective constraint is not known, but it is conceivable that, because Arg is unusually large and contains three amine groups, its insertion might disturb normal configuration of proteins.

Original amino acid (j)

Replacement amino acid (i)	Ala	Arg	Asn	Asp	Cys	Gln	Glu′	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	9730	0	31	24	5	34	37	42	5	3	5	18	19	5	54	99	45	0	0	32
Arg	0	9381	5	0	0	13	0	0	17	0	0	23	18	2	0	1	0	0	0	0
Asn	14	7	9701	36	0	20	7	10	24	4	2	19	1	0	10	51	17	0	0	4
Asp	13	0	45	9757	0	27	96	8	6	0	2	8	1	0	1	26	2	0	0	4
Cys	1	0	0	0	9928	0	0	1	0	2	0	0	11	0	0	12	3	0	0	6
Gln	12	14	15	16	0	9736	24	4	14	4	2	9	11	0	11	13	10	0	0	5
Glu	21	0	9	95	0	40	9726	13	4	4	4	13	1	0	17	15	12	0	0	7
Gly	40	0	22	13	3	11	22	9870	1	0	2	5	0	0	17	42	8	0	0	7
His	2	19	20	4	0	15	3	0	9865	4	3	6	0	3	0	10	5	11	4	1
Ile	1	0	3	0	3	4	3	0	4	9703	22	4	22	14	2	3	14	0	0	70
Leu	4	0	3	3	0	4	7	2	6	52	9899	6	99	19	0	5	7	0	0	24
Lys	17	65	37	13	0	23	21	5	14	9	6	9845	11	0	6	22	14	0	4	13
Met	2	7	0	0	5	4	0	0	0	7	14	2	9672	5	0	5	2	0	0	12
Phe	2	3	0	0	0	0	0	0	4	18	10	0	18	9879	0	5	2	30	74	2
Pro	23	0	9	1	0	13	13	7	0	3	0	3	0	0	9850	11	5	0	0	4
Ser	59	2	67	28	27	22	16	26	17	4	3	14	23	6	15	9598	69	0	0	7
Thr	30	0	25	3	8	20	14	6	8	24	5	10	11	3	8	76	9759	0	0	20
Trp	0	0	0	0	0	0	0	0	4	0	0	0	0	8	0	0	0	9941	7	0
Tyr	0	0	0	0	0	0	0	0	4	0	0	2	0	51	0	0	0	17	9909	0
Val	27	0	7	5	18	12	10	6	3	156	22	12	82	3	3	9	25	0	0	9733

Fig. 1. Transition probability matrix M with each element multiplied by 10,000. Quoted from figure 9-7 of Dayhoff *et al.* (*9*).

Original amino acid (j)

Replacement amino acid (i)	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	9825	0	0	29	0	0	29	29	0	0	0	0	0	0	29	19	29	0	0	29
Arg	0	9835	0	0	29	29	0	44	29	10	19	29	29	0	29	29	15	58	0	0
Asn	0	0	9767	29	0	0	0	0	29	19	0	58	0	0	0	10	15	0	29	0
Asp	15	0	29	9767	0	0	58	15	29	0	0	0	0	0	0	0	0	0	29	15
Cys	0	10	0	0	9796	0	0	15	0	0	0	0	0	29	0	19	0	58	29	0
Gln	0	10	0	0	0	9796	29	0	58	0	10	29	0	0	15	0	0	0	0	0
Glu	15	0	0	58	0	29	9796	15	0	0	0	29	0	0	0	0	0	0	0	15
Gly	29	29	0	29	29	0	29	9832	0	0	0	0	0	0	0	10	0	29	0	29
His	0	10	29	29	0	58	0	0	9767	0	10	0	0	0	15	0	0	0	29	0
Ile	0	5	29	0	0	0	0	0	0	9796	19	15	87	29	0	10	22	0	0	22
Leu	0	19	0	0	0	29	0	0	29	39	9840	0	58	87	29	10	0	29	0	44
Lys	0	10	58	0	0	29	29	0	0	10	0	9825	29	0	0	0	15	0	0	0
Met	0	5	0	0	0	0	0	0	0	29	10	15	9738	0	0	0	7	0	0	7
Phe	0	0	0	0	29	0	0	0	0	19	29	0	0	9767	0	10	0	0	29	15
Pro	29	19	0	0	0	29	0	0	29	0	19	0	0	0	9825	19	29	0	0	0
Ser	29	29	29	0	58	0	0	15	0	19	10	0	0	29	29	9820	44	29	29	0
Thr	29	10	29	0	0	0	0	0	0	29	0	29	29	0	29	29	9825	0	0	0
Trp	0	10	0	0	29	0	0	7	0	0	5	0	0	0	0	5	0	9796	0	0
Tyr	0	0	29	29	29	0	0	0	29	0	0	0	0	29	0	10	0	0	9825	0
Val	29	0	0	29	0	0	29	29	0	29	29	0	29	29	0	0	0	0	0	9825

Fig. 2. Transition probability matrix R constructed using the genetic code table by assuming completely random single-step mutations.

Other notable differences between M and R are the transitions from Ile to Val and from Ser to Ala. This may be due to the inclusion of two-step amino acid substitutions in M. Smaller differences between M and R may not be significant. Although M is based on 814 observations, the data are still insufficient to obtain reliable values for all elements in such a large transition matrix.

Several investigators have pointed out that substitutions of similar amino acids have a higher chance of being accepted by natural selection (*8–10*). In contrast, substitutions of dissimilar amino acids are likely to be rejected. Clarke, using the data compiled by Dayhoff *et al.* (*9*), obtained significant correlation between amino acid similarity and relative occurrence of substitutions (*10*). The differences in the two matrices reflect this fact.

Let us denote the relative frequencies of 20 amino acids by a column vector $\mathbf{v}$. If we multiply the initial vector $\mathbf{v}_0$ by transition matrix $\mathbf{M}_N$, the resulting vector $\mathbf{v}_N$ represents the average amino acid composition after N units of time.

$$\mathbf{v}_N = \mathbf{M}_N\mathbf{v}$$

The eigen vector $\mathbf{v}$ which satisfies the equation, $\mathbf{v} = \mathbf{M}\mathbf{v}$, gives the equilibrium amino acid composition.

According to Dayhoff *et al.* (*9*), $\mathbf{v}$ reflects the average amino acid composition of proteins used for the construction of the matrix M (*9*, figure 9-6). We now examine the hypothesis that this model of protein evolution can be generalized. Let us compare $\mathbf{v}$ with the average amino acid composition of many protein families. We have used the average values obtained from Smith (*12*), who compiled amino acid compositions of 80 proteins from vertebrates, bacteria, and viruses. Figure 3 illustrates the relation between observed and equilibrium ($\mathbf{v}$) compositions. As may be seen from Fig. 3, the agreement between the two is satisfactory, indicating the generality of the model. In discussing the evolutionary implications of the present result, we must first emphasize that the observed nonrandomness in the amino acid composition is not incompatible with the neutral mutation–random drift theory of molecular evolution first put forward by Kimura (*5*). As shown above, the observed amino acid composition represents a state of quasi-equilibrium of "accepted point mutations." From the standpoint of the neutral mutation–random drift theory, most of these accepted point mutations are mutant substitutions due to random frequency drift of selectively neutral or nearly neutral mutations (*2*, *6*). The strongest evidence for the theory is the remarkable constancy, among diverse species, of the substitution rate in evolution of each cistron. The neutral mutation–random drift theory is so far the most plausible hypothesis to explain this fact (*13*). In terms of this theory, amino acid compositions of contemporary organisms represent quasi-equilibrium states of substitutions among selectively neutral or nearly neutral mutations.

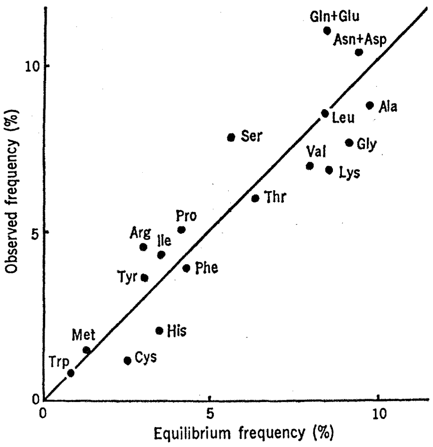

Fig. 3. Correlation between the observed and expected equilibrium amino acid frequencies. The straight line represents equality between the two.

That the asymmetry of substitution frequency, particularly between Arg and some other amino acids, is due to selection rather than mutation becomes obvious when we note that, in 147 human hemoglobin variants, there are four cases where Gly is replaced by Arg (*14*, table 8-1). In contrast, there is no transition (that is, evolutionary substitution) from Gly to Arg in the matrix M (Fig. 1). Hemoglobin variants containing the changes Leu → Arg and Pro → Arg are similar examples. All these variants must eventually be eliminated from the population in the course of evolution. The nonrandomness in the DNA base arrangement (*15*), in particular the relative rarity of the doublet C-G (*11*), might also be explained by the same mechanism. King and Jukes (*2*) suggested that, since C and G are the first and the second letters of the Arg codon, the lower frequency of C-G may be related to the lower Arg composition. A similar discrepancy is observed for T-A, whose observed and expected frequencies are 5.3 and 8.1 percent, respectively. Since T and A are the first and the second DNA letters of the two terminating codons (UAA and UAG in the RNA code), evolutionary pressure selects against T-A within a cistron.

The higher frequency of T compared to that of A, reported by Ohta and Kimura (*3*) for the informational DNA strand of cistrons, may also be due to selective constraint. The nonrandomness of the base arrangement merely reflects the nonrandomness of the amino acid composition.

Fitch (*16*) reported an exceptionally high mutation rate for C → T in his analyses of human hemoglobin variants and amino acid substitutions in cytochrome c. The same tendency was reported by Vogel (*17*) for human hemoglobin variants. They attributed this phenomenon to different mutation rates for the four bases. We believe that a more likely cause is the differential survival of randomly occurring mutations. The high C → T mutation rate is mainly due to three types of substitutions: Asp → Asn, Glu → Lys, and Gly → Asp (*17*, tables 2 to 5). The variants produced by changes in the reverse direction are very few indeed. From the random matrix R and from the actual amino acid composition of hemoglobins, we should expect variants with the following changes to occur with high frequency.

His → Asn, Leu, and Pro
Lys → Asn, Gln, and Thr

Yet only a few of these changes are actually found. The amino acids Lys and His are functionally important in the hemoglobin molecule; they occupy, respectively, 16 and 8 invariant sites within the α and β chains (*18*). Therefore, mutations at those sites might immediately be eliminated. Although human hemoglobin variants have not been "accepted" by natural selection, they have not been eliminated, and may be products of differential survival among mutants.

Another interesting fact is that the variance of G + C content from different cistrons is larger than expected from random arrangement of bases (*3*, *19*). The larger variance may be caused, in part, by natural selection acting through functional requirements of individual cistrons. Although the average amino acid or base composition represents a quasi-equilibrium state of neutral or nearly neutral mutations, small DNA segments must retain their characteristic compositions.

We have endeavored to establish the

view that the amino acid composition is determined largely by the existing genetic code and the random nature of base changes in evolution. Small but significant deviations from such expectation might be accounted for satisfactorily by assuming selective constraint of amino acid substitutions.

TOMOKO OHTA
MOTOO KIMURA

*Department of Population Genetics,
National Institute of Genetics,
Mishima, Shizuoka-ken 411, Japan*

References and Notes

1. M. Kimura, *Genet. Res.* **11**, 247 (1968).
2. J. L. King and T. H. Jukes, *Science* **164**, 788 (1969).
3. T. Ohta and M. Kimura, *Genetics* **64**, 387 (1970).
4. J. L. King, personal communication.
5. M. Kimura, *Nature* **217**, 624 (1968).
6. ———, *Proc. Nat. Acad. Sci. U.S.* **63**, 1181 (1969).
7. ——— and T. Ohta, *Nature* **229**, 467 (1971).
8. C. J. Epstein, *ibid.* **215**, 355 (1967).
9. M. O. Dayhoff, R. V. Eck, C. M. Park, in *Atlas of Protein Sequence and Structure*, M. O. Dayhoff, Ed. (National Biomedical Research Foundation, Silver Spring, Md., 1969), p. 75.
10. B. Clarke, *Nature* **228**, 159 (1970).
11. Abbreviations used are as follows: Ala, alanine; Arg, arginine; Cys, cysteine; Gly, glycine; Ile, isoleucine; Leu, leucine; Pro, proline; Ser, serine; Thr, threonine; Trp, tryptophan; Val, valine; A, adenosine; C, cytidine; G, guanosine; T, thymidine; U, uridine.
12. M. H. Smith, *J. Theor. Biol.* **13**, 261 (1966).
13. M. Kimura and T. Ohta, *J. Mol. Evol.*, in press; T. Ohta and M. Kimura, *ibid.*, in press.
14. M. R. Sochard and M. O. Dayhoff, in *Atlas of Protein Sequence and Structure*, M. O. Dayhoff, Ed. (National Biomedical Research Foundation, Silver Spring, Md., 1969), p. 61.
15. J. Josse, A. D. Kaiser, A. Kornberg, *J. Biol. Chem.* **236**, 864 (1961).
16. W. M. Fitch, *J. Mol. Biol.* **26**, 499 (1967).
17. F. Vogel, *Humangenetik* **8**, 1 (1969).
18. M. O. Dayhoff, *Atlas of Protein Sequence and Structure 1969* (National Biomedical Research Foundation, Silver Spring, Md., 1969), D-218 and D-219.
19. N. Sueoka, *J. Mol. Biol.* **3**, 31 (1961); M. Kimura, *Genet. Res.* **2**, 127 (1961); H. Yamagishi, *J. Mol. Biol.* **49**, 603 (1970).
23. We thank Dr. Kazutoshi Mayeda for stimulating discussions and for correcting the English.

24 February 1970; revised 4 May 1971 ■

Iron- and Riboflavin-Dependent Metabolism of a Monoamine in the Rat in vivo

Abstract. *n-Pentylamine enters into intermediary metabolism by the action of monoamine oxidase. $[1\text{-}^{14}C]$Pentylamine injected into rats is rapidly converted to $^{14}CO_2$. The rate of catabolism decreases progressively in the course of nutritional iron deficiency, reaching about 60 percent of control values in 3 weeks. Feeding with iron yields control levels within 6 days. The catabolism of amyl alcohol, which shares a common pathway with n-pentylamine by way of valeric aldehyde, is not significantly affected by the deficiency. The results demonstrate that the maintenance of normal monoamine oxidase activity in vivo depends upon an adequate supply of dietary iron.*

Information about possible cofactors of mitochondrial monoamine oxidase (MAO) is necessary for an understanding of the mechanism of action of this widely distributed enzyme. Thus far, no functional metal component has been detected in it, although it is known that MAO activity of rat liver, as measured in vitro, declines significantly during long-term nutritional deficiency of iron, as monitored by body weight changes, hemoglobin levels, and hepatic concentrations of iron (*1*). However, the decreases observed (*1*) with three different substrates were only moderately large and perhaps were insufficient to affect the disposition of monoamines in the intact animal. Hence, it was considered important to assess also the function of MAO in vivo. It has now been observed that the rate of oxidation of a standard substrate for this enzyme is subnormal in the iron-deficient rat.

We used the following compounds labeled in the carbon-1 (Mallinckrodt) position: pentylamine (*n*-amylamine) hydrochloride (specific activity, 1.0 mc/mmole) and *n*-amyl alcohol (specific activity, 1.76 mc/mmole); and unlabeled amylamine (Eastman) which we distilled before use, unlabeled *n*-amyl alcohol (Analyzed Reagent, Baker).

Each compound was injected intraperitoneally into male Sprague-Dawley rats (*2*) in a dose of 100 mg per kilogram of body weight; the injected material consisted of 5 μc of the respective labeled substances per kilogram, suitably diluted with the corresponding unlabeled compound. The animals were immediately placed in individual glass metabolism cages, and the $^{14}CO_2$ in the expired gases was collected in a 1 : 2 mixture of ethanolamine and ethylene glycol monomethyl ether, as described (*3*). About 30 to 40 percent of the administered radioactivity was recovered in the first hour after the labeled amine was injected into normal adult rats (170 g) fed on Purina Chow. A further 30 to 35 percent of the injected ^{14}C was collected over the next 2 hours.

The oxidation of pentylamine is initiated by MAO (*4*), as was demonstrated by our method with rats given the amine 16 hours after an intraperitoneal injection of a MAO inhibitor. With tranylcypromine (5 mg/kg, Smith Kline & French); the oxidation of pentylamine was inhibited by 55 percent during the first hour. With iproniazid phosphate (100 mg/kg; Hoffmann-LaRoche), the amount of injected radioactivity recovered as $^{14}CO_2$ during the first 3 hours was only 3 to 7 percent of that administered.

Control groups of rats, that were fed a semisynthetic diet prepared in this laboratory and containing 312 mg of added ferric citrate hexahydrate per kilogram of feed (Fe-supplemented diet) (*1*), metabolized the injected radioactive pentylamine as readily as those consuming the mixed natural diet (Purina). Other rats that were fed the semisynthetic diet for varying periods of time (Table 1), except for omission of the iron salt (Fe-deficient diet), oxidized the amine at a very much lower rate. Iron deficiency thus caused a reduced rate of metabolism of pentylamine, as judged from the rate of recovery of administered radioactivity, during the first hour. This reduction is apparent as early as 9 days (Table 1) and seems to be fully developed by 3 weeks, at approximately three-fifths of the control rate.

The iron dependency of this phenomenon was demonstrated by the following parallel experiments. Groups of rats were allowed to consume the iron-deficient diet for 28 days; then they were changed over to the control diet —one supplemented with ferric citrate. In tests carried out during the refeeding phase, these animals showed an increased rate of oxidation of pentylamine within 3 days and attained the same rate of metabolism as the control animals by day 6. There appeared to be a parallel, but slower restoration of the hemoglobin.

The immediate oxidation product of pentylamine is valeric aldehyde (*6*). *n*-Amyl alcohol, which is readily oxidized by the rat in vivo (*7*), also enters metabolism after enzymic conversion to valeric aldehyde. When the alcohol was administered to iron-deficient and iron-supplemented rats there was no significant difference in the respective rates of oxidation, even at 10 weeks (Table 1).