DISTINGUISHING HOMOLOGOUS FROM ANALOGOUS PROTEINS

WALTER M. FITCH

Abstract

Fitch, W. M. (Dept. Physiological Chem., U. Wisconsin, Madison 53706) 1970. Distinguishing homologous from analogous proteins. Syst. Zool., 19:99–113.—This work provides a means by which it is possible to determine whether two groups of related proteins have a common ancestor or are of independent origin. A set of 16 random amino acid sequences were shown to be unrelated by this method. A set of 16 real but presumably unrelated proteins gave a similar result. A set of 24 model proteins which was composed of two independently evolving groups, converging toward the same chemical goal, was correctly shown to be convergently related, with the probability that the result was due to chance being $<10^{-21}$. A set of 24 cytochromes composed of 5 fungi and 19 metazoans was shown to be divergently related, with the probability that the result was due to chance being $< 10^{-9}$. A process was described which leads to the absolute minimum of nucleotide replacements required to account for the divergent descent of a set of genes given a particular topology for the tree depicting their ancestral relations. It was also shown that the convergent processes could realistically lead to amino acid sequences which would produce positive tests for relatedness, not only by a chemical criterion, but by a genetic (nucleotide sequence) criterion as well. Finally, a realistic case is indicated where truly homologous traits, behaving in a perfectly expectable way, may nevertheless lead to a ludicrous phylogeny.

The demonstration that two proteins are related has been attempted using two different criteria. One criterion is to show that their chemical structures are very similar. An early example of this approach was the observation of the relatedness of the oxygen carrying proteins, myoglobin and hemoglobin (Watson and Kendrew, 1961). More recent is the relatedness of two enzymes in carbohydrate metabolism, lysozyme and alpha-lactalbumin (Brew, Vanaman and Hill, 1967). The other criterion is to show that underlying genetic structures of the proteins are more alike than one would expect by chance. This is now possible because our knowledge of the genetic code permits us to determine how many nucleotide positions, at the minimum, must differ in the genes encoding the two presumptively homologous proteins. One then compares the answer obtained to the number of differences one would expect for unrelated proteins. An example of this approach is the observation of the relatedness of plant and bacterial ferredoxins (Matsubara, Jukes and Cantor, 1969) for which added evidence has been produced (Fitch, 1970a). But regardless of the approach, the impulse, too powerful to resist, is to conclude that a particular pair of proteins had a common genic ancestor if they meet whichever criterion the observer uses.

Now two proteins may appear similar because they descend with *di*vergence from a common ancestral gene (i.e., are homologous in a time-honoured meaning dating back at the least to Darwin's *Origin of Species*) or because they descend with *con*vergence from separate ancestral genes (i.e., are analogous). And, if a common genic ancestor is to be the conclusion, a genetic criterion should be superior to a chemical criterion. This is because analogous gene products, although they have no common ancestor, do serve similar functions and may well be expected to have similar chemical structures and thereby be confused with homologous gene products. This danger can only be increased by using a chemical, as opposed to a genetic, criterion.

It is nevertheless possible that the restrictions imposed by a functional fitness may cause sufficient convergence to produce an apparent genetic relatedness. Therefore, the demonstration that two present-day sequences are significantly similar, by either chemical or genetic criteria, still must necessarily leave undecided the question whether their similarity is the result of a convergent process or all that remains from a divergent process. For example, it is at least philosophically possible to argue that fungal cytochromes *c* are not truly homologous to the metazoan cytochromes *c*, i.e., they just look homologous. Although I know of no one who believes they are only analogous, it is a view worth disproving, particularly in view of the recent statement that it can't be done (Winter, Walsh and Neurath, 1968). To show that the metazoan and fungal cytochromes are indeed homologous (i.e., are monophyletic in origin and have descended divergently) rather than analogous (i.e., are polyphyletic in origin and have descended convergently) one needs only to show that the ancestral cytochrome *c* sequences were more alike than are the present day representatives of these two groups.

The ancestral metazoan and fungal cytochrome *c* sequences have been reconstructed (i.e., intelligently guessed at) by a method previously described (Fitch and Margoliash, 1967). That method assumes divergent descent. In the paper which gives these ancestral sequences (Fitch and Margoliash, 1968), there is also provided an experimental demonstration of the validity and power of the method of reconstructing ancestral sequences assuming, again, a divergent descent. Present-day fungal cytochromes *c* are separated from their metazoan counterparts by an average of 63.2 nucleotide differences. Their reconstructed ancestral forms, however, were separated by only 19 differences. Each of these differences represents a mutation (nucleotide replacement) which has been fixed in the evolution of these genes since their common ancestor. Such differences may be termed "mutation distances." This 70% reduction in mutation distance as one goes back in time would prove that all eukaryotic cytochromes *c* are the result of a divergent process and therefore homologous were it not for the fact that behind this reasoning lies the initial assumption of a divergent relationship between these two groups. This paper demonstrates that the consequences of the assumption of divergence can be assessed. Thus, it is possible to decide if two groups of proteins are the result of convergent or divergent processes.

EFFECTS OF CONVERGENT VS. DIVERGENT EVOLUTION ON ANCESTRAL SEQUENCES

The rationale for deciding whether two groups of proteins are convergently or divergently related is shown in Figure 1. If one knows the nucleotide in a given position of the gene for each member of two groups and also knows the ancestral relationships of the gene within each group, one can then ask about the possible relation of their two ancestral genes. We impose the important restriction that we must account for the data in the fewest number of mutations. The result is that, in the upper trees of the figure, the descendants in both groups can be accounted for by postulating only two mutations each (shown by the arrows), but the necessary consequence is that this nucleotide position must be represented by a G[1] in the ancestral gene of both groups. The lower trees have the same ancestral relationships. They also have the same nucleotides present, but their location on the branch tips has been altered. Again, the descendants in both groups can be accounted by postulating only two mutations each but this

[1] Abbreviations used in this paper are A, adenine; C, cytosine; G, guanine; U, uracil, for the nucleotides of codons which are represented at the level of the messenger RNA rather than at the level of genic DNA.

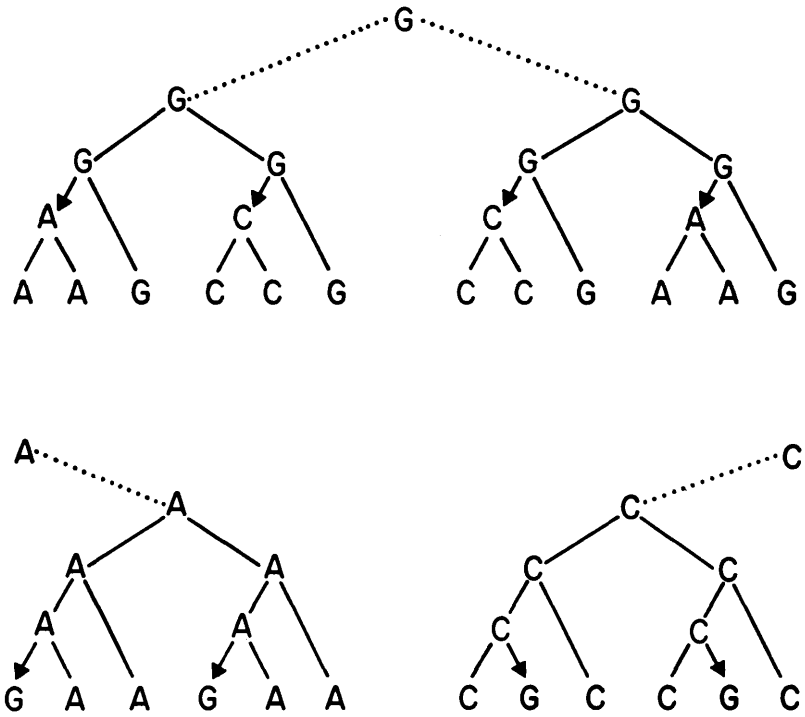

FIG. 1.—Distinguishing convergent from divergent types of nucleotide replacement patterns. Given are two groups of species (related within each group as shown by the solid lines) together with the nucleotide present at a specific position of the gene for each member species as shown at the branch tips. Given also the requirement that the ancestral nucleotide must permit the descendant nucleotides to be obtained in the minimum number of replacements, the ancestral nucleotide of the upper two groups must be set as G, with the required replacements indicated by the arrows. Were one to postulate a common ancestor for the two groups, no new mutations would need to be assumed; hence, this kind of pattern is called the divergent types. The lower two groups are identical except for rearranging the nucleotides at the branch tips, but now, in order to account for descendants in only four nucleotide replacements, the ancestral nucleotide of the lower two groups must be A and C. To postulate a common ancestor for these two groups would require, unlike the upper pair, an additional mutation. This situation shows different ancestral characters apparently converging toward the same descendant character, and hence is called the convergent type. One can calculate the frequency with which one might expect each type to be found in examining a large number of such nucleotide positions and compare that value to what is in fact found for a particular set of proteins. An abnormally large number of either type is evidence favoring that type of relation between the two groups examined.

time the necessary consequence is that the ancestral nucleotide of one tree must be an A and of the other tree a C. The result shown for the upper tree is of the type one would expect to be more common where the two gene groups had diverged from a common ancestral gene (in this nucleotide position, four mutations diverging from G). The result shown for the lower tree is of the type one would expect to be more common where the two gene groups had converged from separate ancestral genes (in this nucleotide position, four mutations converging to G).

Now, if the process of reconstructing the ancestral sequences is properly described, one can calculate the probability that the result in a randomly selected nucleotide position would be of the convergent type if the starting sequences are all unrelated to

each other. This probability, multiplied by the total number of nucleotide positions to be examined, is the number of positions which one would expect to be of the convergent type. Each such position implies that a mutation must have been fixed to account for the difference between the ancestral nucleotide sequences if they, in turn, had had a common ancestor. Should there prove to be a statistically significant excess of the convergent type over that predicted, we may conclude that the two gene groups are probably of independent origin. If, on the other hand, there proves to be a statistically significant excess of the divergent type over that predicted, we may conclude that the two genes probably had a common ancestor. This amounts to saying that we can estimate the number of mutations which will separate the reconstructed "ancestors" of two unrelated groups of unrelated proteins. If, however, the proteins within the two groups are related, we may hope to show that the ancestral genes of these two groups are significantly more mutations apart than expected where convergent evolution has occurred, or significantly less mutations apart than expected where divergent evolution has occurred.

In the material to follow, I shall show in order: *i*, the logical structure used to reconstruct the nucleotide sequence of an ancestral gene; *ii*, the computation of the expected mutation distance between any two genes; *iii*, that when the procedures are employed using unrelated proteins, their "ancestral" genes are separated by a number of mutations not significantly different from expectation; *iv*, how a model set of convergent proteins was constructed; *v*, that when the procedures are employed using the model convergent proteins, their ancestral genes are separated by a number of mutations significantly greater than expectation; and *vi*, that when the procedures are employed using the cytochromes *c*, their ancestral genes are separated by a number of mutations significantly less than expectation. This will be followed by a discussion of the limitations and implications of this work.

SIMPLIFIED ANCESTRAL GENE RECONSTRUCTION

The reconstruction rules given previously (Fitch and Margoliash, 1967) are designed to create ancestral sequences which reflect, as much as possible, biological reality. A greatly simplified version, which is still biologically reasonable, yet lends itself more easily to statistical treatment, follows. A given nucleotide position is defined by the set of nucleotides which may occur there, given the amino acid(s) that may be encoded in the particular sequence under consideration. For example, the first coding position for aspartate is G and the third is C/U, i.e., either pyrimidine. Thus, G and C/U are the nucleotide sets for positions 1 and 3 of the aspartate codon. Given the two descendant nucleotide sets for a given position, their ancestral nucleotide set is defined as the collection of those nucleotides common to both descendants. If they have none in common, then their ancestral nucleotide set is defined as the entire collection of nucleotides found in either of them. The process is illustrated in Figure 2. Mathematically, this amounts to defining the ancestral nucleotide set as the intersection of the descendant sets if that intersection does not give an empty set (which is to say, if the two descendant sets are not disjoint); otherwise it is the union of the descendant (or "combining") sets. A mutation (nucleotide replacement) will be required to account for the data every time the intersection gives an empty set. When this process is carried out for all positions and all ancestors, the total number of mutations found (i.e., disjoint sets combined) is the true minimum for the particular starting sequences used and the particular phylogeny assumed when the ancestral forms were constructed. Other arrangements of nucleotides can normally be found which will account for

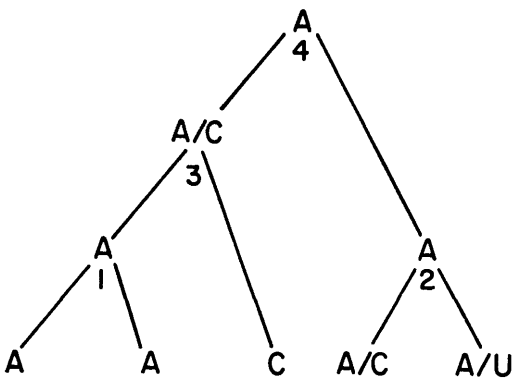

FIG. 2.—Reconstruction of ancestral gene coding. This tree illustrates for a single nucleotide position the simplified method used for reconstructing ancestral nucleotide sequences and is designed to facilitate a statistical analysis of the results. At apex 1, nucleotide A is selected because both descendants are A and no mutation would be required. At apex 2, both descendants are ambiguous but both contain an A and so A is chosen at this apex. The descendants of apex 3 have no nucleotide(s) in common and so the ambiguous A/C is recreated here, indicating we cannot judge which is the better choice at this point. In such cases, where no common nucleotide(s) exists, a mutation will be required. At apex 4, A is again the only nucleotide common to both immediate descendants at apices 2 and 3. Thus, a minimum of one nucleotide replacement is required to account for the 5 descendants shown and this mutation was discovered in formulating the ancestor at apex 3. This is as far as the analysis needs to proceed for statistical purposes but it is clear, once the full tree is at hand, that ambiguous apex 3 must in fact be A in this case and the mutation assigned to the right-hand line of descent. Were it not so, more than the one minimum replacement would be required.

the present day sequences in the same total number of nucleotide replacements, but none will reduce that total. The ancestral sequences so derived provide several important results: *i*) the sequences at an ancestral node are easily derived and completely independent of the investigator's bias; *ii*) as a consequence, the location and number of mutations revealed in joining two sequences to form an ancestral sequence is similarly free of the investigator's bias; and *iii*) the probability that a given number of mutations would be required if the two nucleotide sequences were unrelated can be calculated.

CALCULATION OF EXPECTED MUTATION DISTANCES

The calculation of the probability that two randomly chosen nucleotide sets will be disjoint and therefore necessitate the assumption of a mutation requires the prior enumeration of the frequency of each kind of nucleotide set. There are 15 of them, four of which (A, C, G and U) are discrete, the remaining eleven of which (A/C, A/C/G, etc.) possess varying degrees of ambiguity. If the protein is 100 amino acids long, there would be 300 nucleotides in the messenger RNA of each of the two species which are to be considered as the immediate descendants of the ancestor to be reconstructed. The total number of ways of selecting a pair of nucleotide sets, one from each descendant sequence, is 300^2. If f_i and f_j' are the number of occurrences of the i^{th} nucleotide set in one of the sequences and of the j^{th} nucleotide set in the other, then $f_i \times f_j'$ is the number of ways that particular combination can be selected. If $i \neq j$, there are also $f_j \times f_i'$ ways of getting that particular pair of nucleotide sets by reversing the respective sequences from which the two sets are obtained. Therefore $\sum_{i,j=1}^{15} f_i f_j$ is all possible ways of selecting two nucleotide sets, and this sum must equal L^2, where L is the length of the messenger RNA. If that summation process is repeated with the proviso that this time only those cases are to be included for which i and j are disjoint sets, the sum will represent the number of ways that nucleotide sets may be selected which require the assumption of a mutation. That sum, divided by L^2, is the probability, p_k, that any randomly selected pair of nucleotide sets will require the assumption of a mutation. The subscript k denotes that this probability applies only to the formation of the k^{th} ancestral se-

quence from its immediate descendants and that p must be recalculated for each ancestral sequence to be reconstructed on the basis of the frequencies in its immediate descendants. If the sequence is L long, the total expected number of mutations in the formation of the k^{th} ancestor will be Lp_k and the variance, σ^2, will be $Lp_k(1-p_k)$.

We let the distance (nucleotide differences) found between two sequences be d and the mean expectation of that distance be $\mu = Lp_k$. The standard deviation of the distances is $\sigma = \sqrt{Lp_k(1-p_k)}$. Therefore, the number of standard deviations d is from expectation is $s = (\mu - d)/\sigma$. The probability of a value as large as s can be found in any table of normal probability.

The expected composition of nucleotide sets in the reconstructed ancestor comes from the same calculation, since i and j will lead, either by intersection or union, to a defined nucleotide set with a frequency $f_i f_j'$. If all the $f_i f_j'$ are summed according to the nucleotide set formed from i and j, one gets the expected distribution of nucleotide sets in the ancestor. It is important to determine this expected distribution since it, rather than the exact reconstructed composition, must be used to calculate further mutation probabilities when these reconstructed ancestors are considered as descendants from which still more remote ancestors are reconstructed. Because the sequences of the more remote ancestors are dependent upon the reconstruction of the sequences of less remote ancestors, expectations regarding the mutational differences of two remote ancestors must be similarly dependent.

It should be noted that a change in the third coding position can never lead to more than one other amino acid being encoded. As might be suspected from Figure 1, this procedure appears to be less sensitive in distinguishing convergence from divergence where there are only two (and sometimes but one) character states. Therefore, the following analyses were performed using only the first two of the three nucleotide positions in each codon.

Also, the computer program at its present stage of development has serine represented as A or U in the first coding position and G or C in the second, with the consequence that cysteine, which is represented as U in the first and G in the second position, is not recognized as being one nucleotide replacement away. The statistical computations make proper allowance for this fact.

VALIDITY OF COMPUTATIONS: UNRELATED PROTEINS

To demonstrate that the statistical methods correctly predict the expected result when the sequences are truly unrelated, 16 independent sequences of 100 amino acids were assigned randomly to the tips of a tree whose structure (topology) was as if each line of descent from the ultimate ancestor successively divided exactly and symmetrically three more times. The statistical computation indicated that the two reconstructed penultimate ancestors would be expected to average 70.2 ± 6.8 mutations distant in the first and second coding nucleotides. Upon their reconstruction, they were found to be 77 mutations distant or $(70.2 - 77)/6.8 = -1.0$ standard deviation from the mean expectation. Since approximately ⅔ of all random events fall within 1 σ of the mean expectation, the difference is not statistically significant and there is thus no evidence that the two groups of eight have any relationship—as indeed should be the case. Varying the number of the random sequences analyzed did not significantly affect the variation of the found distance about the mean expectation as shown by the circles in Figure 3.

Since it could be argued that the random assignment of these sequences to an *a priori* topology is unfair because in the normal case one first seeks out the best fitting tree for the sequence data, the 16 random sequences were also tested this way. The best tree is shown in Figure 4.

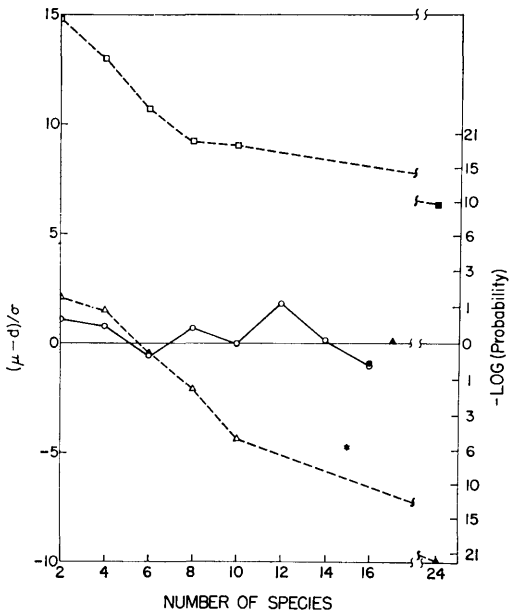

FIG. 3.—Effect of number of species on ascertaining convergent and divergent evolution. The number of species in the two groups being compared is shown on the x-axis. The y-axis shows the number of standard deviations, s, that a given result is from expectation. Wherever open symbols are used the species are divided equally between the two groups. The circles represent random sequences of amino acids (two unrelated groups of unrelated sequences), the triangles represent sequences from the convergent model (two unrelated groups of related sequences), and the squares represent known cytochromes *c* (two related groups of related sequences). The closed circle represents the random sequences calculated accordnig to the tree in Figure 4. For the cytochromes, the set of two species is composed of man and saccharomyces iso-1. The set of four adds duck and saccharomyces iso-2 to the preceding group. The set of six adds tuna and candida to the preceding group. The set of eight adds drosophila and debaromyces to the preceding group. The set of ten adds wheat and Neurospora to the preceding group. The set of 24 cytochromes *c* (closed square, species divided unequally) is composed of two groups of 5 fungal and 19 metazoan species as shown in Figure 5. The closed triangle at 24 is for the set of similarly divided, convergent proteins as shown in Figure 6. The closed triangle at 17 is a set of convergent proteins containing only FUB from the FUN group and 16 randomly selected sequences from the MET group. The axis on the right side of the figure gives the negative exponent of 10 (k) for the probability that by chance a result may be as

It is clearly unlike the trees one usually obtains using real homologous sequences or even the artificial sequences that resulted from the model convergent process. What this tree essentially depicts is that every sequence is approximately equidistant from every other sequence. For this random model, it would be expected that the penultimate ancestors would average 71.0 ± 6.8 mutations distant in the first and second coding nucleotides. Upon their reconstruction by the above procedure, they were found to be 77 mutations distant or (71–77)/6.8 = 0.88 σ. Since the initial random sequences are ~150 mutations distant while the reconstructed ancestors are only 77 mutations apart, there is the surface appearance of a divergent process. But since the calculation tells us that the reconstruction procedure would create this much apparent divergence (indeed a little more) we are in no danger of concluding that these random sequences are the result of a divergent evolution. This verifies once again that unrelated sequences lead to ancestral sequences whose mutation distance upon reconstruction by a carefully prescribed process can be reasonably estimated.

Finally, to be certain that this result is due to the unrelatedness of the sequences rather than a randomness of the amino acids which might not be present in real protein sequences, the entire procedure

←

many standard deviation units away from the mean as the corresponding values of s shown on the left side. The k are very nearly equidistant and are all the sums of consecutive numbers starting at 1 with the result that the n^{th} number up is $k = n(n + 1)/2$. If n is set equal to 0.609 s and k calculated, then the probability of a value s σ from the mean is 10^{-k}. The probability value is precise for s = 1.644 ($p = 10^{-1}$). At s = 7, p is calculated to be 6.62×10^{-12} when it is really 2.55×10^{-12} so the result is conservative. The relative error continues to increase as the probability decreases but the result is always conservative. Thus, we have a simple way of estimating p for large values of s in the absence of extensive tables for the normal distribution.

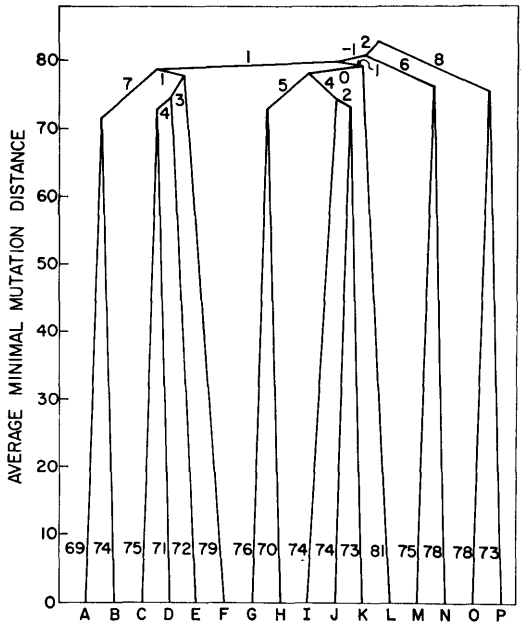

FIG. 4.—"Phylogenetic Tree" for Random Sequences of Amino Acids. Tree depicts the result when the computer was forced to "discover" the evolutionary relationships among 16 random sequences of 100 amino acids each. The "% SD" is only 3.4. All three codon nucleotide positions were used in formulating this tree.

was repeated using the first 100 amino acids of 16 presumably unrelated proteins whose sequences have been determined.[2] The sequences were randomly assigned to the branch tips of a tree with the same topology as the first random case above. Calculation showed that the ancestors of the two groups of eight would be expected to be 71 mutations apart and were in fact 64 mutations distant or 1.07 standard deviation units from the mean, i.e., not significantly different from expectation.

The statistical procedure is such that the nucleotide most frequent in two descendants is expected to be even more frequent in their immediate ancestor. This process, if continued through a sufficient number of ancestors, could conceivably cause some difficulty in estimating the mutation distance expected between the reconstructed "ancestors" of two very large unrelated groups of unrelated nucleotide sequences. The results with the random sequences, as shown in Figure 3, fail to give any evidence that the present numbers of sequences cause any real difficulty.

CREATION OF CONVERGENTLY EVOLVING PROTEINS

To demonstrate that proteins which evolved convergently are not given the appearance of having evolved divergently when the previously described method of

[2] The following 16 amino acid sequences were used to construct a tree of unrelated proteins: **Human β hemoglobin**—Braunitzer, G., Gehring-Muller, R., Hilschmann, N., Hilse, K., Hobom, G., Rudloff, V. and Wittmann-Liebold, B., (1961) Z. Physiol. Chem. *325*, 283; **Pig cytochrome *c***—Stewart, J. W. and Margoliash, E., (1965) Can. J. Biochem. *43*, 1187; ***Leucaena glauca* ferredoxin**—Benson, A. M. and Yansunobu, K. T., (1969) J. Biol. Chem. *244*, 955; **Lobster glyceraldehyde-3-phosphate dehydrogenase**—Davidson, B. E., Sajgo, M., Noller, H. F. and Harris, J. I., (1967) Nature *216*, 1181; **Bovine trypsinogen**—Mikes, O., Holeysovsky, V., Tomasek, V. and Sorm, F., (1966) Biochem. Biophys. Res. Commun. *24*, 346; **Bovine ribonuclease**—Smyth, D. G., Stein, W. H. and Moore, S., (1963) J. Biol. Chem. *238*, 227; **Chicken lysozyme**—Canfield, R., (1963) J. Biol. Chem. *238*, 2698; **Tobacco mosaic virus *Vulgare* coat protein**—Anderer, F. A., Wittmann-Liebold, B. and Wittmann, H. G., (1965) Z. Naturforschg. *20B*, 1203; **Papaya papain** (fragment)—Light, A., Frater, R., Kimmel, J. R. and Smith, E. C., (1964) Proc. Natl. Acad. Sci. U. S., *52*, 1276; **Human Bence Jones kappa CUM**—Hilschmann, N., (1967) Z. Physiol. Chemie, *348*, 1718; ***Escherichia coli* tryptophan synthetase alpha**—Yanofsky, C., Drapeau, G. R., Guest, J. R. and Carlton, B. C., (1967) Proc. Natl. Acad. Sci. U. S. *57*, 296; ***Pseudomonas fluorescens* azurin**—Ambler, R. P. and Brown, L. H., (1967) Biochem. J. *104*, 784; ***Escherichia coli* aspartate transcarbamylase regulatory polypeptide chain**—Weber, K., (1968) Nature *218*, 1116; ***Golfingia gouldii* hemerythrin**—Klippenstein, G. L., Holleman, J. W. and Klotz, I. M., (1968) Biochemistry *7*, 3868; **Human growth hormone**—Li, C. H., Liu, W-K. and Dixon, J. S., (1966) J. Amer. Chem. Soc. *88*, 2050; **Human haptoglobin 2 alpha 1,S**—Black, J. A. and Dixon, G. H., (1968) Nature *218*, 736. The ferredoxin sequence is only 96 amino acids long and so the C-terminal end was treated as if it had had a deletion of 4 amino acids.

ancestral sequence reconstruction is used, one could test the method on convergently evolved proteins. Since none of these are available, a model set of proteins was created, and the method tested on these. For this purpose it was decided that the general structure of their evolution should be as much as possible like the evolution of cytochrome *c* shown in Figure 5. To this end, the fungal and metazoan portions of this tree were exactly reproduced in the convergent model but only in so far as the topology and the number of mutations on each descending leg is concerned.

First of all, two ancestral genes called FUN and MET, 104 codons long, were obtained by random selection where each codon's probability of being selected was equal to its presumed frequency in the primordial cytochrome *c* gene.

Using the sequence of human cytochrome *c* as the standard of fitness, potential mutations were randomly generated for various positions in the two sequences. Any mutation which would not increase the fitness of a sequence was not accepted. The general requirements of acceptability for a mutation were as follows. The new amino acid encoded as a result of a mutation and the previously encoded (old) amino acid were compared to the desired amino acid in that position of human cytochrome *c*. The following convergence rules for phenotypic convergence were used and the mutation, and therefore the new amino acid, was not accepted whenever:

1. the old amino acid was already identical to the desired amino acid;

2. the codon for the old amino acid was as few nucleotide replacements from the codon of the desired amino acid as was the new codon[3];

3. the old amino acid was of the same polarity as the desired amino acid but the new amino acid was not (for this purpose, the non-polar amino acids were ala,cys,phe, ileu,leu,met,pro,trp,tyr, and val; the other ten were regarded as polar);

4. the old amino acid had the same charge as the desired amino acid but the new one did not[3];

5. the old amino acid was uncharged and polar, the desired amino acid was non-polar but the new amino acid was charged;

6. the old amino acid was a member of the same subgroup as the desired amino acid but the new amino acid was not (the subgroups are A, tyr and phe; B, val, met, leu, and ileu; C, gly,ala and val; and D, ser and thr)[3];

7. the desired amino acid was polar, and the new amino acid was a larger non-polar amino acid than the old one (for this purpose, size was set as trp>tyr>phe>met> leu = ileu>val = pro>cys>ala);

8. the total charge on the molecule moved out of or away from the range of 10.5 ± 1 excess of positive over negative charges (his was given half a charge).

Finally, for simplicity, it was decided not to permit a given nucleotide position to mutate more than once in the descent from an ancestor to its next immediate descendant and that while silent mutations (which don't change the amino acids encoded) were allowed, they were not counted among those mutations necessary to reach a descendant.

COMPUTATIONS ON CONVERGENTLY EVOLVED PROTEINS

Table I presents a general summary of the minimum mutation distances between the various proteins being considered in the convergent model. The original FUN and MET ancestral sequences were 156.5

[3] This rule number 2, unlike the others, selects at the level of the genotype rather than at the level of the phenotype. It was adopted solely to assure that the amount of convergence obtained per mutation accepted was maximized. When used, rules number 4 and 6 become redundant. The convergent model which produced the sequences compared in Figure 7 did not employ rule 2 and hence convergence was more gradual. It was necessary to omit rule 2 in this case because the intent was to show that selection at the level of the phenotype leads to convergence at the level of the genotype.

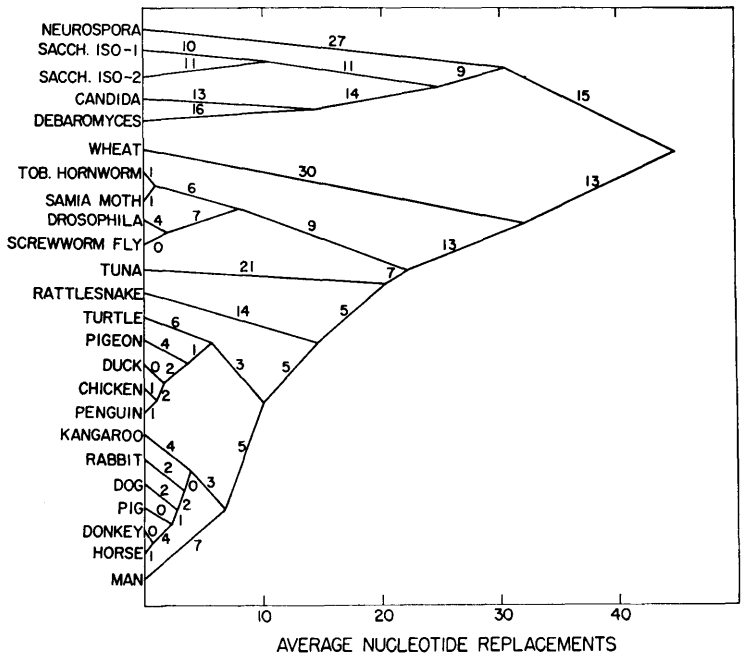

FIG. 5.—Phylogeny of 24 species of Cytochrome *c*. The tree is the best-fitting one found for the sequences[4] employed. All three codon nucleotide positions were used in formulating this tree. The number on the various legs indicate the number of nucleotide replacements assigned to that particular descent upon reconstruction of the ancestral sequences according to previously described rules (Fitch and Margoliash, 1967).

[4] The following cytochromes *c*, in the order they appear in Figure 5, were used in constructing the tree: *Neurospora crassa*—Heller, J. and Smith, E. L., (1966) J. Biol. Chem. *241*, 3165; *Saccharomyces cerevisiae* iso-1—Yaoi, Y., Titani, K. and Narita, K., (1966) J. Biochem. (Tokyo) *59*, 247; *Saccharomyces* iso-2—Stewart, J. W., Putterman, G. J. and Margoliash, E., unpublished; *Candida krusei*—Narita, K. and Titani, K., (1965) Proc. Japan Acad. *41*, 831; *Debaryomyces kloeckeri*—Titani, K., unpublished; wheat—Stevens, F., Glazer, A. N. and Smith, E. L., (1967) J. Biol. Chem. *242*, 2764; *Protogarce sexta*—Chan, S. K., unpublished; *Samia cynthia*—Chan, S. K. and Margoliash, E., (1966) J. Biol. Chem. *241*, 335; *Drosophila melanogaster*—Nolan, C., Weiss, L. J., Adams, J. J. and Margoliash, E., unpubl; *Haematobia irritans*—Chan, Tulloss, Margoliash, unpubl; Tuna—Kreil, G., (1963) Z. Physiol. Chem. *334*, 154; *Crotalus adamanteus*—Bahl, O. P. and Smith, E. L., (1965) J. Biol. Chem. *240*, 3585; *Chelydra serpentina*—Chan, S. K., Tulloss, I. and Margoliash, E., (1966) Biochem. *5*, 2586; Pigeon and *Anas platyrhynchos*—Chan, S. K., Tulloss, I. and Margoliash, E., unpublished; Chicken—Chan, S. K. and Margoliash, E., (1966) J. Bio. Chem. *241*, 507; *Aptenodytes patagonica*—Chan, S. K., Tulloss, I. and Margoliash, E., unpublished; *Macropus canguru*—Nolan, C. and Margoliash, E., (1966) J. Biol. Chem. *241*, 1049; Rabbit—Needleman, Saul B. and Margoliash, E., (1966) J. Biol. Chem. *241*, 853; Dog—McDowell, M. A. and Smith, E. L., (1965) J. Biol. Chem. *240*, 4635; Pig—Stewart, J. W. and Margoliash, E., (1965) Can. J. Biochem. *43*, 1187; Donkey—Walasek, O. F. and Margoliash, E., unpublished; Horse—Margoliash, E., Smith, E. L., Kreil, G. and Tuppy, H., (1961) Nature *192*, 1125; Man—Matsubara, H. and Smith, E. L., (1963) J. Biol. Chem. *238*, 2732. Most of the sequences which are listed as unpublished have nevertheless been provided by the authors to the *Atlas of Protein Sequence and Structure*, 1967–8, Dayhoff, M. O. and Eck, R. V., Eds. (Nat. Biomed. Res. Fdn., Silver Spring, Md.).

TABLE I. MINIMAL MUTATION DISTANCES BETWEEN VARIOUS PROTEIN SEQUENCES OF THE CONVERGENT MODEL.[1]

		A	B	C	D	E	F	G
FUA → E	A	—						
MEF → Y	B	131.6	—					
FUN	C	32.2		—				
MET	D		37.3	149	—			
"FU"	E			42		—		
"ME"	F				36	105	—	
Human Cyt *c*	G	118.6	123.0	152	161	112	125	—

[1] The minimal mutation distance between the genes encoding two amino acid sequences is the minimum number of nucleotide differences that must be postulated to account for the observed differences in their sequence. The implication is that, since their common ancestor if they had one, this number of mutations or nucleotide replacements would have to have been fixed to explain their divergence. FUA → E and MEF → Y are the two major groups formed by the descent from the FUN and MET ancestral genes respectively. FUN and MET are names derived from a stretched analogy to fungi and metazoan ancestral cytochromes *c* which led to this model case. The ancestors, "FU" and "ME," were reconstructed from the descendant species using the "found" phylogeny in Figure 6 and the reconstruction rules given in the text. Human cytochrome *c* is the "standard" of fitness toward which the FUN and MET ancestral sequences were made to evolve. The numbers in columns A and B are the average of 5 and 19 values respectively. Distances were calculated on the basis of all three coding positions.

± 4.5 nucleotide replacements distant from the human cytochrome *c* toward which they were to converge. This is almost exactly 1.5 mutations/coding position, which is about what one would expect from a pair of random proteins. The descendant sequences FUA through MEY average 121 mutations away from human cytochrome *c*. Thus, the true net convergence has amounted to about 35 nucleotide replacements, demonstrating that the model does produce convergence. However, the ancestral forms, FUN and MET, were initially only 149 mutations apart from each other (as opposed to their 156 replacements away from cytochrome *c*) and their descendants were still, on the average, 131.6 mutations apart from each other. This gives a relative convergence between the two groups of only 17 mutations. With real proteins it is unlikely that anything other than relative convergence would be observable.

The convergent model proved interesting in part because the computer program, in attempting to reconstruct the phylogeny of the FUN and MET descendants, found a "better" tree than the actual tree. The "% SD" (standard deviation), which is an estimate of the error between the measured mutation distances between species and those distances computed for the tree under consideration, was 5.6% for the actual tree (see Fitch and Margoliash, 1967 for details). The error was reduced to 5.1% for the best fitting (or found) tree. Figure 6 shows a comparison between these two trees. The solid lines are identical for both trees. Where the trees differ, the structure of the actual tree is shown by dashed lines while the structure of the best-fitting found tree is shown by dotted lines. The found, rather than the actual, tree was used in reconstructing ancestral sequences because similar discrepancies must occur when using real protein sequences.

The reconstructed ancestors of the convergent model prove to have only 117 mutations separating them in the first and second nucleotide positions, compared to 122 for the descendants, so that we have, seemingly, a small divergence where we know convergence has occurred. Thus, finding that the mutation distance between ancestral sequences reconstructed by the methods presented is less than the distance between the present day, descendant sequences is not sufficient to demonstrate the presence of an homologous relationship.

The statistical calculation on the "found" tree of the convergent model indicates, assuming the sequences are all unrelated,

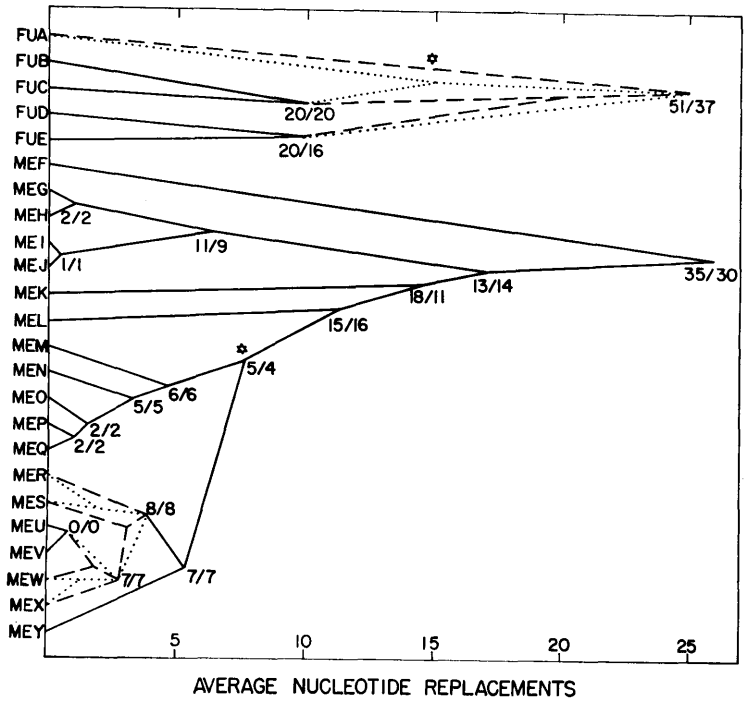

FIG. 6.—Model Convergent Evolution. The two random sequences, FUN and MET, were allowed to evolve according to the topology shown by the solid plus dashed lines (the actual tree) and under the rules given in the text for selection for convergence. Using all three coding nucleotides, the best-fitting tree was reconstructed according to previously described procedures (Fitch and Margoliash, 1967). This found tree, where topologically identical to the actual tree, is shown by solid lines; where different, by dotted lines. The topology and the number of mutations accepted into each segment of the actual tree conforms to that for a set of 24 cytochromes *c* shown in Figure 5. The mutation numbers at the nodal points are only those mutations in the first and second coding positions. The numbers represent those mutations occurring between the node at which the numbers appear and the two immediate descendants of that node. The number to the left of the slash is the number of mutations incorporated into the first two coding positions in the actual tree, the number to the right of the slash is the number of mutations discovered upon reconstruction by the present method.

that the reconstructed FUN and MET ancestral sequences would be expected to average 54.1 ± 6.3 mutations distant in the first and second nucleotides. Reconstruction according to the procedure discussed in connection with Figure 2 finds them 117 mutations distant. This is (54–117)/6.3 = –9.95 standard deviation units from the mean expection. The probability of this result occurring by chance alone is less than 10^{-21}. The result is almost identical if the actual rather than the found tree is used.

The triangles in Figure 3 show how the number of convergent proteins involved in the computation affects the calculation. No significant result was obtained until 10 species were included. However, as shown by the closed triangle, even 17 species is insufficient if one of the two groups contains only one species. There must be multiple representation in both groups to detect a significant relationship.

It might be objected that our ability to detect convergence was dependent upon testing against each other two reconstructed ancestral sequences which were originally as remote as two random sequences. Dr.

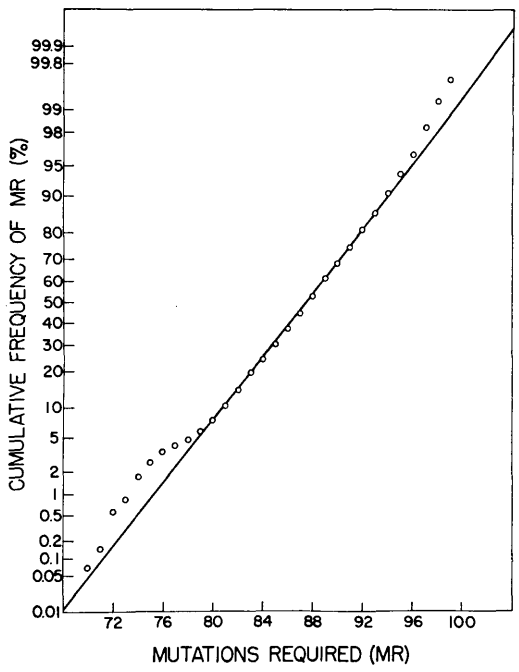

FIG. 7.—Comparison of Two Convergent Proteins for Coding Relatedness. Length of sequence examined is 60, number of sequences compared is 2916, number of independent observations is 107. The value of χ^2_{app} is 3.94 for which the probability that this result could be a chance event is 0.048. The method for determining χ^2_{app} has been submitted for publication. The two sequences are from a model case not otherwise discussed but which started with the same ancestral pair of sequences. More mutations were accepted, but since rule 2 regarding acceptability of mutations was omitted, the total amount of convergence was less than that of the convergent model in Figure 6.

M. Susman pointed out that if much of the convergence had already occurred prior to the point in time represented by the two ancestral forms being compared, the task would not be so easy. This is true, but the power of the method is indicated by eliminating species FUD through MEL from the data so that one is now reconstructing the ancestral forms only as far back as shown by the stars in Figure 6. Thus over 50% of the net convergence in the two groups had occurred prior to the appearance of the two ancestral forms to be compared. Nevertheless, the convergence was detected with a probability that the result could de due to chance equal to 2×10^{-6} as shown by the star on Figure 3.

Such convergent sequences also provide an answer to the question asked earlier, i.e., might not the constraints on an optimal functional fitness be sufficiently severe that a convergent evolution will produce two analogous proteins which will appear related, not only chemically but by the criterion of genetic relatedness previously set forth. The answer is yes. Figure 7 shows a comparison between two sequences obtained in a convergent model which give the appearance of being homologous. The probability that this much similarity would occur by chance is less than .05 (Fitch, 1970b). It required examination of an extraordinarily large length of sequence (60 amino acids) to detect this, but then greater convergence would have reduced the length required.

COMPUTATIONS ON REAL CYTOCHROMES *C*

The preceding computations show that when two sets of proteins have been caused to converge to somewhat similar chemical structures, the reconstructed ancestral sequences are less alike than would have been the case had all the presumed "descendant" sequences been in fact totally unrelated. What is the result when the same computations are performed for cytochrome *c* which has, presumably, the same number of mutations placed on the corresponding segments of a topology identical to the one used in the convergent model? The calculations were performed using 24 species and the tree shown in Figure 5. The statistical calculation indicates, assuming the present day fungal and metazoan cytochromes are all unrelated, that the reconstructed ancestral sequences for the fungal and metazoan groups would be expected to average 63.4 ± 6.8 mutations distant in the first and second coding nucleotides. Reconstruction finds them only 21 mutations distant. This

is $(63.4 - 21)/6.8 = 6.3$ standard deviations from the mean expectation. The probability of this result occurring by chance alone is less than 6×10^{-10}. Thus both convergent and divergent processes are shown to give statistically significant departures from the expectation and furthermore these departures are at opposite extremes of the distribution.

DISCUSSION

It may appear to some that this procedure contradicts a basic philosophical principle that, given information on a system at only one point in time, one can not tell in which direction the system is moving in the time coordinate. The principle is sound, and it should be clearly recognized that my procedure can not distinguished between 24 cytochromes *c* which have diverged from a single ancestral gene and 24 independently arising cytochromes *c* which are converging on a single future descendant form. These are the pure forms of convergence and divergence and are identical except for a reversal of the time axis. As Dr. R. L. Metzenberg pointed out to me, what we have ruled out in the case of cytochrome *c* is a specific mixture of the two. And this can be done because the assumption of divergence permits (within limits) one to describe past states of the system knowing the present state (i.e., nucleotide sequences). Thus, if one accepts divergence for the genes within two select groups, say within the artiodactyl and carnivore hemoglobins, one can ask if their past states, which are describable under the *a priori* assumption of divergence, plus their present states are together consistent with the groups being convergent or divergent. This we can do because we now have character states at two points in time. The result is that we may show that the present-day sequences are consistent with pure divergence but not with mixtures of convergence and divergence, with a monophyletic origin but not with a biphyletic origin. One can maintain convergence in the cytochrome *c* gene as a logical possibility only by going all the way and assuming that there must have been a very large number of origins (perhaps as many as 24) to the 24 cytochromes *c* that were analyzed in this study. But if such a position is to be advocated, one must also explain how so many independently arising genes should have, by themselves, led to a phylogeny of these species (Figure 5) which is so similar to the phylogeny biologists have produced using other characters. The explanation will become more tedious as other genes produce similar results until, like the geocentric view of the solar system, it collapses under the burden of epicycles of epicycles.

It has been stated by Winter, Walsh and Neurath (1968) that "The evolutionary biochemist . . . can show the similarity of two or more protein structures but he has not and cannot have any independent [internal] experimental evidence relating to the question of ancestral genes." Their point was emphasized by including a quotation from Alice in Wonderland, which, in the context of their statement, can only be interpreted to mean:

> "Homology? Homology?" and sometimes "Analogy?" for you see, as she couldn't know the answer, it didn't much matter which word she used for it.

I believe that Alice, Winter, Walsh and Neurath are correct in what we can know, only in the most extreme representations of these choices. I suggest this work shows that within the limits of reasonable alternatives, we certainly can determine from present-day amino acid sequences whether two groups of peptidases, for example, have a common or an independent origin, that is, whether they are homologous or analogous in the biological and genetic meaning of these words and that molecular biologists have not rendered irrelevant the distinction between them.

But the problems involved in the usage

of homology are not restricted to its correct usage. It is not sufficient, for example, when reconstructing a phylogeny from amino acid sequences that the proteins be homologous. It has been pointed out before that a phylogeny of birds and mammals based upon a haphazard mixture of α and β hemoglobins would be biological nonsense since the initial dichotomy would be on the distinction between the α and β genes rather than between the birds and the mammals (Fitch and Margoliash, 1967). Therefore, there should be two subclasses of homology. Where the homology is the result of gene duplication so that both copies have descended side by side during the history of an organism, (for example, α and β hemoglobin) the genes should be called *paralogous* (para = in parallel). Where the homology is the result of speciation so that the history of the gene reflects the history of the species (for example α hemoglobin in man and mouse) the genes should be called *orthologous* (ortho = exact). Phylogenies require orthologous, not paralogous, genes. Note that the present method does not permit us to conclude that the fungal and metazoan cytochromes are orthologous. We would have attained the same result had the two gene groups been paralogous. But where paralogy is the desired conclusion, one is on even stronger grounds. If one is willing to accept the trypsins as orthologous and to accept the chymotrypsins as orthologous, one can indeed decide, given say 5 sequences of each, whether these 2 gene groups had a common ancestor and are therefore paralogous or of independent origin and only analogous.

ACKNOWLEDGMENTS

This project received support from grants from the National Science Foundation GB-7486 and the National Institutes of Health NB-04565. The University of Wisconsin Computer Center, whose facilities were used, also receives support from NSF and other U. S. government agencies. The very valuable assistance of Mrs. K. Gould is recognized as is the most helpful discussions by Drs. R. L. Metzenberg, M. Susman, B. Harris and O. Smithies.

REFERENCES

BREW, K., T. C. VANAMAN, AND R. L. HILL. 1967. J. Biol. Chem., 242:3747.
FITCH, W. M. 1970a. J. Mol. Biol., 49 in press.
FITCH, W. M. 1970b. J. Mol. Biol., 49 in press.
FITCH, W. M., AND E. MARGOLIASH. 1967. Science, 155:279.
FITCH, W. M., AND E. MARGOLIASH. 1968. Brookhaven Symp. in Biol., 21:217.
MATSUBARA, H., T. H. JUKES, AND C. R. CANTOR. 1969. Brookhaven Symp. in Biol., 21:201.
WATSON, H. C., AND J. C. KENDREW. 1961. Nature, 190:670.
WINTER, W. P., K. A. WALSH, AND H. NEURATH. 1968. Science, 162:1433.

Department of Physiological Chemistry, University of Wisconsin, Madison, Wisconsin, 53706.