87

2

Reading the Palimpsest:* Contemporary Biochemical Data and the RNA World

Steven A. Benner,[1] Mark A. Cohen,[1] Gaston H. Gonnet,[2]
David B. Berkowitz,[3] and Kai P. Johnsson[1]
[1] Laboratory for Organic Chemistry
[2] Department of Computer Science
E.T.H. Zürich, CH-8092 Switzerland

The chemical behavior of contemporary living systems contains vestiges of the history of living chemistry on planet Earth. To read this "palimpsest" is a challenge of the first order (Benner et al. 1989). Some of the historical record has been written over by the demands of natural selection that have forced the evolution of new chemical structures to meet new biological challenges. Some has been lost in the noise arising from random events, the "neutral drift" that characterizes the structural divergence of biological molecules following the divergence of their host organisms (King and Jukes 1969; Kimura 1982). Some has undoubtedly been confused by lateral transfer of genetic information between phylogenetically distant organisms (Doolittle et al. 1990) and by "sequence convergence," the independent emergence of polypeptide sequences that offer unique chemical solutions to particular biological problems.

Three advances of the past decade have greatly improved our ability to deconvolute the information about earlier life forms written in the biological chemistry of contemporary organisms. First, substantial progress has been made toward an integration of structural theory from chemistry and evolutionary theory from biology (Benner and Ellington 1990b). The integrated theory allows us to proceed past the classic ques-

*A palimpsest is a parchment that has been inscribed two or more times, with the previous texts imperfectly erased and therefore still partially legible.

[3] Present address: Department of Chemistry, University of Nebraska, Lincoln, Nebraska 68588.

The RNA World
© 1993 Cold Spring Harbor Laboratory Press 0-87969-380-0/93 $5 + .00

tions in biological chemistry (what happens at a chemical level in a living system) to ask *why* it happens. A number of questions in biological chemistry have been addressed within the context of the integrated theory, and models are now available to understand the evolutionary status of many molecular aspects of living systems. Most of these hypotheses are directed to specific behaviors in biological macromolecules: Catalytic activity, stereospecificity, and thermal stability are three examples.

Second, there has been an explosion of sequence data (Bairoch and Boeckmann 1992), both from individual investigators reporting sequences of individually chosen proteins and from genome projects now beginning to yield results (Oliver et al. 1992; Sulston et al. 1992). The recently completed organization and exhaustive matching of the protein sequence database (Gonnet et al. 1992) make systematic analysis of these sequence data possible in their entirety for the first time. This analysis has in turn provided better methods for aligning homologous protein sequences, one of the more important tools for understanding the evolution of living systems.

Finally, substantial progress has been made in developing a manipulative understanding of how biological macromolecules work. It is now possible to predict many details of the folded structure of proteins de novo from sequence data (Crawford et al. 1987; Benner 1989b; Bazan 1990; Niermann and Kirschner 1990; Benner and Gerloff 1991; Knighton et al. 1991; Thornton et al. 1991; de Vos et al. 1992). Furthermore, the design of polypeptides that fold in solution and catalyze reactions, at least at some rate, is now possible, and the first examples of designed peptides whose structure in solution has been rigorously proven have recently appeared (Johnsson et al. 1990; Osterhout et al. 1992).

With these tools, a detailed reconstructed history of the evolution of biological macromolecules back to the protogenome, the genome from the most recent common ancestor of today's archaebacteria, eubacteria, and eukaryotes (Benner and Ellington 1987, 1990a; Benner et al. 1987, 1989), becomes a tangible research goal. The organism most closely corresponding to this reconstructed ancestor lived over a billion years ago. Using chemical assumptions, a somewhat less detailed history back to the "breakthrough organism," the last organism in the RNA world to use RNA as the sole genetically encoded component of biological catalysis (Benner et al. 1989), is available. The breakthrough organism lived perhaps 2.5 billion years ago (Benner et al. 1989). In this chapter, we outline contemporary tools for reconstructing models for ancient forms of life and provide examples of the use of these tools to solve specific problems in the history of life.

THE TOOLS

Contemporary Methods for Analyzing Sequence Data

Central to historical reconstructions of earlier episodes in the history of life are alignments of the sequences of homologous proteins and nucleic acids. Although sequence alignments have been constructed for 30 years (Edwards and Cavalli-Sforza 1963; Zuckerkandl and Pauling 1965; Fitch and Margoliash 1967) and are routinely used in laboratories throughout the world, alignments available to most biochemists are in fact quite problematic (Thorne et al. 1992). In particular, the tools available for constructing alignments, both the matrices that allow the scoring of mutations and formulae for scoring gaps in alignments, remain primitive. As a result, much research is handicapped by suboptimal alignments, and the literature contains seemingly endless arguments over evolutionary issues that might be resolved by well-constructed alignments.

Recently, the protein sequence database was reconstructed using a "patricia tree" data structure (Gonnet et al. 1992). This made possible an exhaustive cross-matching of every subsequence in the database with every other subsequence. From this came 1.7 million aligned pairs of sequences of potentially homologous proteins. This large collection of aligned sequence pairs was the starting point for a broadly empirical study of macromolecular evolution, providing the most advanced tools to date for constructing alignments.

Mutation Matrices, Gap Penalties, and Statistically Rigorous Alignments

A "log-odds" (or Dayhoff) matrix is a 20 x 20 table that shows the logarithm of a probability (multiplied by 10) of each pair of the 20 proteinogenic amino acids being matched in an alignment (Table 1). The matrix is defined for a particular evolutionary distance between the two proteins. Thus, a 1% mutation matrix describes pairwise probabilities of amino acids in two protein sequences that have undergone one point mutation per 100 amino acid residues. In such a matrix, the sum of all the off-diagonal terms is equal to 1%.

Such matrices are named in honor of Margaret Dayhoff, who proposed their use some 20 years ago (Dayhoff et al. 1978). Versions of the 1978 matrix are used by most commercially available computer alignment packages. However, two methodological problems make the 1978 matrix suboptimal for the alignments that are most interesting in the context of this article. First, and unavoidably in 1978, the amount of data

Table 1 Dayhoff "log-odds" matrix

C	11.5																			
S	0.1	2.2																		
T	−0.5	1.5	2.5																	
P	−3.1	0.4	0.1	7.6																
A	0.5	1.1	0.6	0.3	2.4															
G	−2.0	0.4	−1.1	−1.6	0.5	6.6														
N	−1.8	0.9	0.5	−0.9	−0.3	0.4	3.8													
D	−3.2	0.5	−0.0	−0.7	−0.3	0.1	2.2	4.7												
E	−3.0	0.2	−0.1	−0.5	−0.0	−0.8	0.9	2.7	3.6											
Q	−2.4	0.2	0.0	−0.2	−0.2	−1.0	0.7	0.9	1.7	2.7										
H	−1.3	−0.2	−0.3	−1.1	−0.8	−1.4	1.2	0.4	0.4	1.2	6.0									
R	−2.2	−0.2	−0.2	−0.9	−0.6	−1.0	0.3	−0.3	0.4	1.5	0.6	4.7								
K	−2.8	0.1	0.1	−0.6	−0.4	−1.1	0.8	0.5	1.2	1.5	0.6	2.7	3.2							
M	−0.9	−1.4	−0.6	−2.4	−0.7	−3.5	−2.2	−3.0	−2.0	−1.0	−1.3	−1.7	−1.4	4.3						
I	−1.1	−1.8	−0.6	−2.6	−0.8	−4.5	−2.8	−3.8	−2.7	−1.9	−2.2	−2.4	−2.1	2.5	4.0					
L	−1.5	−2.1	−1.3	−2.3	−1.2	−4.4	−3.0	−4.0	−2.8	−1.6	−1.9	−2.2	−2.1	2.8	2.8	4.0				
V	−0.0	−1.0	0.0	−1.8	0.1	−3.3	−2.2	−2.9	−1.9	−1.5	−2.0	−2.0	−1.7	1.6	3.1	1.8	3.4			
F	−0.8	−2.8	−2.2	−3.8	−2.3	−5.2	−3.1	−4.5	−3.9	−2.6	−0.1	−3.2	−3.3	1.6	1.0	2.0	0.1	7.0		
Y	−0.5	−1.9	−1.9	−3.1	−2.2	−4.0	−1.4	−2.8	−2.7	−1.7	2.2	−1.8	−2.1	−0.2	−0.7	−0.0	−1.1	5.1	7.8	
W	−1.0	−3.3	−3.5	−5.0	−3.6	−4.0	−3.6	−5.2	−4.3	−2.7	−0.8	−1.6	−3.5	−1.0	−1.8	−0.7	−2.6	3.6	4.1	14.2

The terms represent ten times the logarithm of the probability of a pairwise matching of the indicated amino acids in two proteins separated by 250 PAM units. The terms were generated from data collected for all protein pairs separated by PAM distances between 6 and 100 PAM, extrapolated by exponential fitting to PAM 116.5, and then extrapolated directly to a PAM distance of 250 to conform to standard usage.

available to Dayhoff was small. Thus, the elements of the 1978 Dayhoff matrix have large variances.

Second, to ensure that the alignments yielding mutation data were of high quality, Dayhoff constructed matrices from pairs of proteins that were themselves quite similar (typically 90–95% identical) in sequence and then extrapolated the matrices to give a matrix suitable for aligning distant proteins. Data from the exhaustive matching of the contemporary sequence database (Gonnet et al. 1992) show that this extrapolation was problematic. The probability of certain substitution pairs turns out to be a strong function of evolutionary distance, apparently because constraints imposed by the genetic code influence the mutation matrix for closely related proteins but not distantly related proteins (Gonnet et al. 1992). A revised Dayhoff matrix suitable for aligning distant sequences is shown in Table 1.

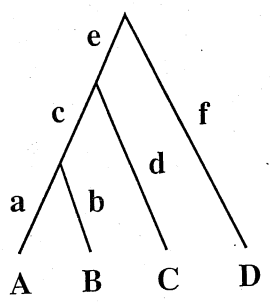
Insertions and deletions also create problems in constructing alignments. If gaps are introduced into an alignment at no cost, any random sequences can be aligned. Classically, therefore, gaps have been penalized. The penalty most commonly used involves a gap cost calculated from a formula of the form (*ak*+*b*), where *k* is the length of the gap and *a* and *b* are arbitrarily chosen parameters. At an intuitive level, such scoring is well known to be inadequate, because most biochemists "tweak" computer-produced alignments to achieve a more pleasing disposition of gaps. Data from the exhaustive matching show why such adjustment is necessary. The probability of a gap of length *k* declines as a function of $k^{-1.7}$ (Gonnet et al. 1992 and in prep.), not exponentially as implied by the gap cost calculated using the (*ak*+*b*) formula. Thus, the classic formula is an inadequate approximation for constructing alignments containing gaps.

PAM Distances and Evolutionary Trees

Reconstructions of ancient protein sequences generally require alignments of several sequence pairs and arrangement of the sequences on an evolutionary tree constructed from a measure of evolutionary distance between sequence pairs. Often, evolutionary distance is defined by the fraction (or percent) sequence identity. It is well known that this measure is imperfect. Inspection of the revised Dayhoff matrix (Table 1) shows, for example, that conservation of a cysteine is more significant than conservation of an alanine. Furthermore, due to reverse mutations and multiple mutations at individual residues, evolutionary distance is not a simple function of percent identity after modest evolutionary divergence.

Evolutionary distance between two homologous protein sequences is best measured in PAM units, indicating the number of accepted point mutations per 100 amino acid residues separating the two sequences. Two protein sequences 1 PAM unit distant differ by 1 accepted point mutation per 100 amino acid residues. For two sequences separated by a PAM distance of x, transformation of the first sequence via the 1% mutation matrix (see above) x times gives the second sequence with the highest probability. PAM distances and trees constructed from them can both be subjected to a rigorous statistical analysis.

Given a set of homologous proteins, a PAM distance table (Fig. 1) can be constructed tabulating the PAM distances between each pair of aligned sequences. A tree is constructed from these distances, where vertices in the tree represent ancestral sequences from ancient organisms. If three sequences are available, the length of individual segments of the tree (in PAM distance units) connecting these vertices can be obtained by solving the set of linear simultaneous equations derived from the PAM distance data. With more than three sequences, the solution is overdetermined, and the best tree is obtained from a least-squares fit of the data (Gonnet et al. 1992). An idealized example of the process for constructing these trees is shown in Figure 1. A representative tree for the superfamily of proteins that includes aspartate aminotransferases (AAT), tyrosine aminotransferases (YAT), aminocyclopropanecarboxylate synthase, and histidinol phosphate aminotransferases (HPAT) is shown in Figure 2. A segment of the corresponding alignment is shown in Figure 3.



A-B distance
$a+b = 12$

A-C distance
$a+c+d=47$

B-C distance
$b+c+d=49$

A-D distance
$a+c+(e+f)=76$

B-D distance
$b+c+(e+f)=74$

C-D distance
$d+(e+f)=78$

	B	C	D
A	12	47	76
B	49	74	
C	78		

$a \approx 5$
$b \approx 7$
$c \approx 18$
$d \approx 24$
$e+f \approx 53$

Figure 1 A hypothetical tree reconstructed from a hypothetical PAM distance table (below). Note that in the simultaneous equations, (*e*+*f*) always appear as a pair, and the data do not permit assignment of the relative magnitudes of *e* and *f* (that is, the data do not permit the rooting of the tree). However, the remaining distances are calculated with known statistical parameters.

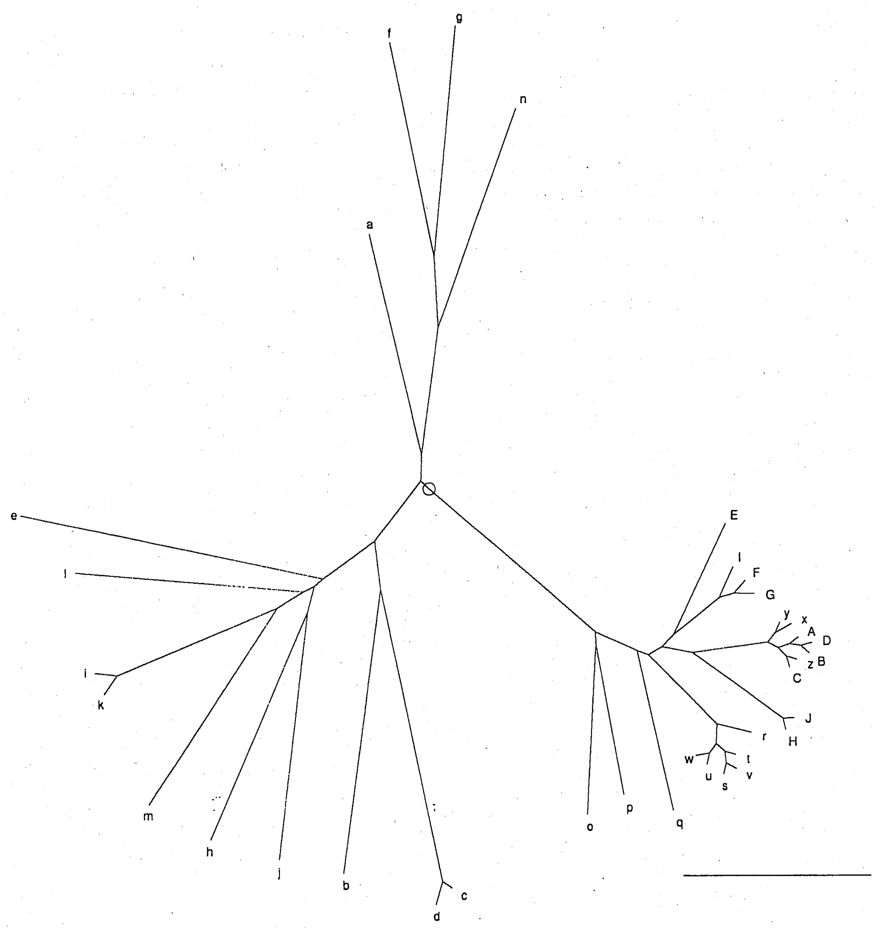


Figure 2 The most probable evolutionary tree relating HPAT, AAT, YAT, and a number of other enzymes. The bar represents an evolutionary distance of 50 PAM units. Key: (a) 1-Aminocyclopropanecarboxylate-1C synthetase, tomato; (b) AAT, *Sulfolobus solfataricus*; (c) YAT, human; (d) YAT, rat; (e) CobC protein, *Pseudomonas*; (f) lysine decarboxylase, *Hafnia alvei*; (g) ornithine decarboxylase, *E. coli*; (h) HPAT, *Bacillus subtilis*; (i) HPAT, *E. coli*; (j) HPAT, *Haloferax volcanii*; (k) HPAT, *Salmonella typhimurium*; (l) HPAT, *Streptomycete*; (m) HPAT, *Saccharomyces cerevisiae*; (n) hypothetical protein, *B. subtilis*; (o) aromatic amino acid aminotransferase, *E. coli*; (p) AAT, *E. coli*; (q) AAT, *S. cerevisiae*, cytoplasmic; (r) AAT, chicken, cytoplasmic; (s) AAT, horse, cytoplasmic; (t) AAT, human, cytoplasmic; (u) AAT, mouse, cytoplasmic; (v) AAT, pig, cytoplasmic; (w) AAT, rat, cytoplasmic; (x): AAT, chicken, mitochondrial; (y) AAT, turkey, mitochondrial; (z) AAT, horse, mitochondrial; (A) AAT, human, mitochondrial; (B) AAT, mouse, mitochondrial; (C) AAT, pig, mitochondrial; (D) AAT, rat, mitochondrial; (E) AAT, lupine, P2-root; (F) AAT, lupine, P1-root; (G) AAT, alfalfa, leaf; (H) AAT, *Panicum miliaceum*, isoenzyme 1; (I) AAT, *P. miliaceum*, isoenzyme 2; (J) AAT, *P. miliaceum*, isoenzyme 3.

```
     223 ..296
b - HNPTGTLFSPNDVKKIVDISR-DNKIILLSDEIYDNFVYEGKMRSTLE--DSDWRDF-LIYVNGFSKTFSMTGWRLG
c - SNPCGSVFSKRHLQKILAVAA-RQCVPILADEIYGDMVFSDCKYEPLA--TLSTDVP-ILSCGGLAKRWLVPGWRLG
d - SNPCGSVFSKRHLQKILAVAE-RQCVPILADEIYGDMVFSDCKYEPLA--NLSTNVP-ILSCGGLAKRWLVPGWRLG
e - NNPTGRALAPAELLAI-AARQKASGGLLLVDEAFGDLE-P---QLSVA--GHASGQGNLIVFRSFGKFFGLAGLRLG
h - NNPTGTYTSEGELLAFLE--RVPSRVLVVLDEAYYEYV-TAEDYPETV--PLLSKYSNLMILRTFSKAYGLAALRVG
j - HNPTGSVLPREELVELAE--SVEEHTLLVVDEAYGEFA----EEPSAI--DLLSEYDNVAALRTFSKAYGLAGLRIG
i - NNPTGQLINPQDFRTLLELTRGK--AIVVADEAYIEFC-PQA---SLA--GWLAEYPHLAILRTLSKAFALAGLRCG
k - NNPTGQLINPQDLRTLLELTRGK--AIVVADEAYIEFC-PQA---TLT--GWLVEYPHLVILRTLSKAFALAGLRRG
m - GNPTGAKIKTSLIEKVLQ-NWDN--GLVVVDEAYVDFC-GGS---T-A--PLVTKYPNLVTLQTLSKSFGLAGIRLG
l - NNPTGTAVPAETVLALYEAAQAAKPSMVVVDEAYIEFS-HGA---SLL--PLLDGRPNLVVSRTMSKAFGAAGLRLG
G - HNPTGVDPTLEQWEQIRQLIRSKSL-LPFFDSAYQGFASGSLDADAQPVRLFVADGGELLVAQSYAKNMGLYGERVG
F - HNPTGVDPTTEQWEQIRKLLRSKAL-LPFFDSAYQGFASGSLDIDAQAVRLFVADGGELLLAQSYAKNMGLYGERVG
I - HNPTGVDPTIDQWEQIRQLMRSKSL-LPFFDSAYQGFASGSLDKDAQPVRMFIADGGELLMAQSYAKNMGMYGERVG
E - HNPTGIDPTPEQWEKIADVIQEKNH-IPFFDVAYQGFASGSLDEDAASVRLFVARGLEVLVAQSYSKNLGLYAERIG
z - HNPTGVDPRPEQWKEIATLVKKNNL-FAFFDMAYQGFASGDGNKDAWAVRYFIEQGINVCLCQSYAKNMGLYGERVG
C - HNPTGVDPRPEQWKEMATLVKKNNL-FAFFDMAYQGFASGDGNKDAWAVRHFIEQGINVCLCQSYAKNMGLYGERVG
A - HNPTGVDPRPEQWKEIATVVKKRNL-FAFFDMAYQGFASGDGDKDAWAVRHFIEQGINVCLCQSYAKNMGLYGERVG
B - HNPTGVDPRPEQWKEIASVVKKKNL-FAFFDMAYQGFASGDGDKDAWAVRHFIEQGINVCLCQSYAKNMGLYGERVG
D - HNPTGVDPRPEQWKEMAAVVKKKNL-FAFFDMAYQGFASGDGDKDAWAVRHFIEQGINVCLCQSYAKNMGLYGERVG
x - HNPTGVDPRQEQWKELASVVKKRNL-LAYFDMAYQGFASGDINRDAWALRHFIEQGIDVVLSQSYAKNMGLYGERAG
y - HNPTGVDPRPEQWKEMATLVKKNNL-FAFFDMAYQGFASGDINRDAWAVRHFIEQGINVVLSQSYAKNMGLYGERAG
J - HNPTGVDPTEEQWREISHQFKVKKH-FPFFDMAYQGFASGDPERDAKAIRIFLEDGHQIGCAQSYAKNMGLYGQRVG
H - HNPTGVDPTEEQWREISHQFKVKKH-FPFFDMAYQGFASGDPERDAKAIRIFLEDGHQIGCAQSYAKNMGLYGQRVG
r - HNPTGTDPTPDEWKQIAAVMKRRCL-FPFFDSAYQGFASGSLDKDAWAVRYFVSEGFELFCAQSFSKNFGLYNERVG
s - HNPTGTDPTPEQWKQIASVMKRRFL-FPFFDSAYQGFASGNLDRDAWAVRYFVSEGFELFCAQSFSKNFGLYNERVG
v - HNPTGTDPTPEQWKQIASVMKRRFL-FPFFDSAYQGFASGNLEKDAWAIRYFVSEGFELFCAQSFSKNFGLYNERVG
t - HNPTGIDPTPEQWKQIASVMKRRFL-FPFFDSAYQGFASGNLERDAWAIRYFVSEGFEFFCAQSFSKNFGLYNERVG
u - HNPTGTDPTPEQWKQIAAVMQRRFL-FPFFDSAYQGFASGDLEKDAWAIRYFVSEGFELFCAQSFSKNFGLYNERVG
w - HNPTGTDPTEEEWKQIAAVMKRRFL-FPFFDSAYQGFASGDLEKDAWAIRYFVSEGFELFCPQSFSKNFGLYNERVG
q - HNPTGLDPTSEQWVQIVDAIASKNH-IALFDTAYQGFATGDLDKDAYAVRXXLSTVSPVFVCQSFAKNAGMYGERVG
p - HNPTGIDPTLEQWQTLAQLSVEKGW-LPLFDFAYQGFARG-LEEDAEGLRAFAAMHKELIVASSYSKNFGLYNERVG
o - HNPTGADLTNDQWDAVIEILKAREL-IPFLDIAYQGFGAG-MEEDAYAIRAIASAGLPALVSNSFSKIFSLYGERVG
    .**.*.  ..... ..... .... ... .*.... ..  ...  ... ... ..  ... ......*.......*.*

αβ            aaaaaaaa        bbbbbbbbbbbb            aaaaaaaabbbbbb    bbbbbbbbb  α2 1
αβ           aaaaaaaaaaaaaa       bbb          aaaaaa        bbbbbb            bb  α2 2
αβ          aaaaaaaaaaaaaaaabbbbbbbbb       aaaaaaaaaaaaaa   bbbbbbbb         bbb  α2 3
```

Figure 3 Segment of a multiple alignment of the proteins in the tree in Fig. 2, excluding four (proteins a, f, g, and n) that do not give significant alignments in this region. The key is as in Fig. 2. A deletion is represented by a dash. Secondary structures predicted for proteins b, c, d, e, h, j, k, l, and m (the left branch of the tree in Fig. 2) are shown below the alignment (**1**), above secondary structures determined by crystallographic analysis for proteins p (**2**) and r (**3**). Especially noteworthy is the prediction of an αβ unit preceding, and an α_2 unit following, the segment where the alignment is statistically significant. The alignment in these regions is not shown.

Reconstructed Sequences of Ancient Proteins

Points in an evolutionary tree correspond to protein intermediates in the evolution of the protein family. Using the appropriate Dayhoff matrix, sequences for these ancient proteins can be reconstructed in a probabilistic form. In a reconstructed sequence, each position is represented by a vector of unit length in 20-dimension space, where the component of the vector in each of the 20 dimensions is the probability that each of the 20 amino acids was present in this protein at this position. Part of the probabilistic sequence for the reconstructed histidinol phosphate aminotransferase from the most recent common ancestor of *Bacillus* (a eubacterium), *Haloferax* (an archaebacterium), and *Saccharomyces* (a eukaryote) (the left branch of the tree in Fig. 2) is given in Table 2.[1]

With many contemporary sequences, a high branching order of the tree, and a slow rate of divergence, ancestral sequences can often be reconstructed with remarkable precision. For example, for the protein family that includes elongation factor 1α, more than 71% of the 478 positions in the reconstructed proto-eukaryotic sequence are reconstructed with more than 90% probability (Table 3). An ancient protein with a well-defined probabilistic sequence can be made and studied in the laboratory. The first example of an ancestral protein from an extinct organism to be made and studied was a ribonuclease from the most recent common ancestor of swamp buffalo, river buffalo, and ox, corresponding in the fossil record approximately to the fossil organism *Pachyportax latidens* (Stackhouse et al. 1990).

Predicting and Designing Protein Folds

Converting sequence data into conformational data has been among the most difficult challenges in structural biology. In fact, there are two challenges. The first is to design de novo proteins that fold in solution in a productive way (i.e., to form stable folded structures and catalyze reactions). The second is to predict de novo (i.e., without input of crystallographic data) the folded structure of natural proteins. Both problems have now been addressed in specific cases, and there is reason to believe that the approach that yielded the best solutions will apply, if not to all protein structures, at least to a large subset of them. For the purposes of this chapter, it is important only to know that solutions to these problems have improved substantially and to understand how they might be used in reconstructing ancient forms of life.

In protein design, the first polypeptides designed to fold in solution and catalyze reactions have been prepared, and their structure in solution has been proven (Johnsson et al. 1990). In one case, the catalytic mechanism of the designed enzyme has been explored in detail (Johnsson et al. 1990; K. Johnsson, unpubl.). These experiments provide a chemical basis for assumptions regarding how catalytically active proteins originated (see below). Likewise, the design of catalytic proteins yields

[1] Sequence data allow the calculation only of the most probable sequences for specified points in an evolutionary tree. These trees are, however, "unrooted"; the sequence data alone do not define a point on the tree that is geologically the most ancient protein sequence. Such points can only be specified given additional biological information and are less precisely defined for more ancient branch points than for more recent branch points. As a convention, we reconstruct here ancient sequences indicated by the circle on the tree, representing the center of gravity of the tree. This is a formalism; probabilistic sequences corresponding to all other points are also available. However, this point generally falls within the region of the tree where the root is likely to lie.

Table 2 Reconstructed probabilistic ancestral sequences

Position		Position	
223	N=59.4 H=35.0	179	N=98.3
224	N=98.9	180	N=99.9
225	P=99.8	181	P=100.0
226	T=98.3	182	T=99.8
227	G=99.7	183	G=100.0
228	T=64.2 V=17.6 S=7.4	184	T=68.1 S=23.8 A=5.2
229	A=39.5 V=17.0 D=14.1 E=5.9	185	V=35.8 L=24.1 A=14.8 I=6.3
230	I=32.2 L=27.6 P=17.8 V=11.2	186	I=64.1 V=25.8 L=7.4
231	S=57.9 T=31.2	187	P=80.0 S=9.6
232	P=99.3	188	A=24.2 T=15.4 E=14.1 P=13.1 S=9.2 R=8.7 K=6.5 Q=5.6
233	E=90.8 Q=5.1	189	E=74.2 Q=14.7
234	E=64.8 Q=24.2 D=6.4	190	E=78.3 D=12.7
235	L=65.7 W=21.5	191	L=82.1 I=12.0
236	Q=31.8 L=25.9 K=22.5 E=5.8 R=5.3	192	L=73.1 V=13.4 I=5.5
237	A=65.2 K=10.4 Q=8.2 T=6.6	193	A=66.8 T=9.9 E=8.8 S=5.8
238	I=61.6 L=33.0	194	L=95.3
239	L=66.5 A=14.5 V=8.1	195	L=93.7
240	E=86.2 Q=8.1	196	E=99.2
241	V=53.2 L=21.9 I=14.1 A=8.3	197	L=37.6 A=28.0 V=9.7 I=5.2
242	T=27.0 A=23.6 S=21.3 M=6.7 V=5.4 I=5.1	198	T=34.9 A=21.9 S=15.8 N=13.6
243	R=67.3 K=24.2 Q=6.5	199	R=88.7 Q=6.5
244	A=68.6 S=12.2 E=6.9	200	A=48.8 V=40.5
245	K=78.8 E=6.9	201	E=26.0 K=23.5 A=13.2 P=8.5 N=8.2 Q=6.1 S=5.7
246	N=57.1 S=16.9 E=7.8 K=6.4	202	E=30.7 K=30.2 S=13.9 Q=6.0
247	H=40.6 R=18.3 L=11.3 K=7.2 Q=5.5	203	R=28.1 H=25.8 P=18.0 K=7.7 Q=7.4
248	A=55.1 S=13.6 T=9.1 V=8.7	204	A=42.5 S=26.8 T=23.2
249	L=84.9 I=10.7	205	L=86.9 M=7.5 I=5.2

250	V=54.6 P=19.3 L=8.4 I=5.9
251	V=42.1 L=26.8 F=19.4 I=7.3
252	V=62.1 F=18.0 I=7.2 L=6.4
253	D=98.7
254	E=66.2
255	A=98.1
256	Y=99.3
257	G=53.1 Q=18.9 A=5.3
258	E=40.7 D=27.2 G=21.3
259	F=99.1
260	A=78.6 V=11.0 C=6.3
261	Y=26.3 S=24.2 F=11.9 A=8.3 T=5.6
262	P=35.1 G=30.0 S=16.4 A=6.6
263	G=72.4 D=16.2
264	A=58.4 L=18.9 V=5.2
265	E=67.6 D=27.1
266	E=46.2 Q=21.5 K=15.1 R=5.6
267	P=46.9 D=25.4 E=9.0 S=5.7
268	S=63.7 A=24.7 T=7.8
269	L=64.2 Y=12.2 F=7.8 W=5.7
270	A=97.6
271	V=39.5 I=26.3 L=14.5
272	R=64.7 K=10.2
273	P=60.0 A=24.0
274	L=70.4 F=24.4
275	A=44.3 L=33.7 V=11.1
276	S=92.2
277	E=68.6 K=13.2 D=6.2 Q=5.0
278	Y=47.2 G=18.8 H=6.2
279	P=70.2 L=15.0

206	V=95.6
207	V=99.2
208	V=93.9
209	D=99.8
210	E=99.7
211	A=99.8
212	Y=99.9
213	I=55.0 V=27.4
214	E=98.6
215	F=97.9
216	C=61.6 A=23.4 S=6.8 V=6.1
217	T=23.9 P=20.1 H=15.2 S=13.0 A=6.3 G=6.1
218	G=84.4 A=10.3
219	A=77.2 S=10.3 E=7.0
220	D=46.1 E=42.1
221	Y=28.6 E=23.4 H=6.9 D=6.0 Q=5.0
222	P=97.2
223	S=83.0 T=14.3
224	L=62.1 A=11.1 V=7.8 T=7.5
225	V=55.5 I=19.3 L=11.3 A=10.7
226	P=99.4
227	L=99.2
228	L=98.9
229	S=75.6 T=14.4 A=6.7
230	E=56.3 K=37.0
231	Y=99.9
232	P=86.9 S=6.0

Table 2 Reconstructed probabilistic ancestral sequences (*continued*)

Position		Position	
280	N=63.8 E=16.9 D=8.2	233	N=99.7
281	L=93.6	234	L=97.3
282	V=46.5 I=29.7 L=19.0	235	V=82.8 A=9.1
283	V=90.6 I=8.1	236	V=51.8 I=43.7
284	L=61.8 A=18.8 V=7.2	237	L=98.8
285	R=69.7 Q=14.2 K=5.6	238	R=99.6
286	T=55.6 S=39.3	239	T=99.8
287	F=94.7	240	F=58.8 L=32.7 M=7.3
288	S=96.2	241	S=99.6
289	K=98.2	242	K=99.7
290	A=56.5 N=17.0 S=9.7	243	A=99.1
291	F=98.8	244	F=64.2 Y=35.7
292	G=99.6	245	G=99.9
293	L=98.0	246	L=98.8
294	A=61.2 Y=17.0	247	A=99.8
295	G=99.7	248	G=99.6
296	L=61.7 E=13.4	249	L=98.5
297	R=98.9	250	R=99.9
298	L=65.4 V=21.8 I=9.8	251	L=51.7 I=26.3 V=19.6
299	G=99.7	252	G=100.0

The sequences presented correspond to the segment of the multiple alignment presented in Fig. 3. The column on the left is the probabilistic ancestral sequence for all the proteins in the multiple alignment (see footnote 1). The column on the right is the probabilistic ancestral sequence for only the histidinol phosphate aminotransferases (sequences h–m in Fig. 3). The numbering on the left corresponds to the position in the alignment in Fig. 3. The positions on the right indicate the equivalent positions in the multiple alignment of the histidinol phosphate aminotransferases alone. Only probabilities >5% are shown.

Table 3 Ancestral sequences of elongation factors 1α and Tu

			Probability of most probable amino acid[a]		
Protein	No. of sequences in family	Length of ancestral sequence	90–99%	99%	100%
Proto-eukaryotic EF-1α	26	478	36	243	65
Proto-eubacterial ET-Tu	19	475	85	115	0
Proto-archaeal EF-1α/Tu	6	442	35	226	2
Protogenomic EF-1α/Tu[b]	51	530	115	41	0

[a]Value indicates the percent probability of the most probable amino acid at this position in family ancestral sequence.
[b]See footnote 1.

an understanding of which reactions are easy to catalyze, which are difficult to catalyze, and what chemical structures are needed for catalysis. This brings at least some much-needed chemical rigor to speculations concerning primitive catalysts (see below).

Likewise, methods for predicting de novo the folded structure of proteins have advanced substantially over the past 5 years. To date, structures of at least three protein families have been predicted *before* crystallographic data were available where the predictions were shown to be remarkably accurate by subsequently determined crystal structures: the αβ-barrel domain of tryptophan synthase (Crawford et al. 1987; Hyde et al. 1988), the catalytic domain of protein kinase (Benner and Gerloff 1991; Knighton et al. 1991), and the extracellular domain of the human growth hormone receptor (Bazan 1990; de Vos et al. 1992). The only successful methods to date for predicting de novo the folded structure of proteins start with an alignment of homologous sequences. The most reliable of these methods extracts information from patterns of variation and conservation within a set of aligned homologous sequences (Benner 1989b; Benner and Gerloff 1991). Secondary structure can now be predicted with reasonable certainty for proteins for which a number of homologous sequences are known. Assignments of active-site residues, covariation analysis, and biochemical information (positions of S–S bonds and cross-linking studies) allow the predicted secondary structural units to be assembled to yield a model of the folded protein.

It is well known that secondary and tertiary structure in proteins diverges less rapidly than primary structure (Chothia and Lesk 1986). Motifs and consensus sequence elements also can indicate distant relationships that are sometimes not identified by sequence comparisons (Dever et al. 1987; Bairoch 1992), although these methods have proven on occasion to be quite unreliable (Bork 1992). If crystal structures are available for two families of proteins, alignments emphasizing structural similarities can establish homology in cases where sequence alignments alone might be unconvincing (Taylor and Orengo 1989; Valencia et al. 1991). Reliable structure predictions can replace crystal structures in this process.

For example, the evolutionary tree in Figure 2 joins proteins in two connected components[2] of the protein sequence database. Homology between these two sets of protein sequences is not strongly supported by sequence similarity, as evident from inspection of the segment of the multiple alignment (Table 2) produced by computational analysis. No crystallographic data are available for any protein in the left branch of the tree. It is possible, however, to predict de novo a secondary structure for proteins in this branch using methods developed in Zürich (Benner 1989b; Benner and Gerloff 1991). The predicted secondary structure for the aligned and flanking regions is shown beneath the alignment in Figure 3. The secondary structure predicted for the left branch of the tree (Fig. 2) corresponds well to the secondary structure obtained crystallographically for aspartate aminotransferases from eubacteria and eukaryotes from the right branch. The correspondence between a reliable predicted secondary structure in one branch of the tree with crystallographically determined secondary structures of representative proteins in a second branch secures what would otherwise be a marginal hypothesis regarding homology between the two branches.

Models for Understanding the Evolution of Protein Structure and Behavior

Natural selection targets the properties of the folded sequence that contribute to survival and reproduction in the host organism. Issues related to function and survival are formulated in three ways (Benner and Ellington 1988). First and most classically, the sequence of a macromolecule determines its behavior, and much work in contemporary biochemistry focuses on the relationship between sequence and behavior. Second, behavior determines survival value. Finally, the survival value is itself determined by protein sequence. This final relationship has been the most difficult to probe, although much of the discussion of macromolecular evolution (e.g., molecular clocks) has been based on assumptions regarding this relationship (Lewin 1988).

It is now possible to distinguish (at least at the level of hypothesis) adaptive, neutral, and historical behaviors in biological macromolecules (Benner and Ellington 1990b). The stereospecificity of metabolic trans-

[2] Connected components are families of proteins within the database related by specific criteria, in this case a PAM distance of less than 200, a similarity score of >125, and a match length of >80 positions.

formations (Benner et al. 1990), the kinetic details of enzymatic reactions (Albery and Knowles 1976; Benner 1989a), and the form and nature of metabolic pathways (Benner et al. 1989) are three that have been examined in special detail. Adaptive differences influence survival. Neutral differences do not. Historical behaviors are those that, although not themselves adaptive in the contemporary world, are vestiges of ancient structures that were adaptive or that arose as a result of particular constraints in the evolution of earlier forms of life. The last are the most relevant here, as constraints imposed in a world where RNA was the sole genetically encoded macromolecule, or those imposed by nonbiological chemical reactions, can greatly influence models of the organisms that arose in these environments.

The Reactivity of Organic Molecules

Structure-reactivity theory from organic chemistry constrains speculations concerning what types of reactions are plausible in primitive catalytic systems. Such constraints are often lacking in the discussion of catalysis, especially by RNA molecules. A few comments are relevant in the present context.

First, most functionalized macromolecules catalyze reactions. The catalytic power depends on what kind of functional groups the macromolecule bears and what the reaction is. Some reactions are considerably more difficult to catalyze than others, and an understanding of structure-reactivity theory, in particular stereoelectronic theory, allows one to evaluate which reactions belong to which classes (Benner 1988). Proteins presumably came to be the predominant catalysts in the modern world because they carry more encoded functional groups than RNA molecules (Benner et al. 1987). Even with RNA, however, structure theory suggests that the problem is to control catalytic power in RNA, not to create it.

An understanding of chemical reactivity is also crucial for reconstructing ancient forms of life. For example, the most convincing case for an RNA world comes from the fact that many biological cofactors contained RNA fragments *that play no role* in the chemical reactivity of the cofactor in the contemporary world (White 1976; Visser and Kellogg 1978; Benner et al. 1989). Such behavior is presumably a vestige of a world where RNA was the primary biological macromolecule, because the absence of a function in the contemporary world (and, indeed, abundant examples of alternative structures in contemporary organisms that performed the same chemistry without the RNA fragment) precludes any functional reasons to create such structures in any other environment.

Finally, chemistry becomes a still more powerful tool by allowing scientists to ask "What if?" and "Why not?" Many of these questions are directed toward nucleic acid structure. For example, Figure 4 shows six isomorphic Watson-Crick base pairs constructed from 12 different purine and pyrimidine base analogs. In principle, oligonucleotides could incorporate all 12 bases, potentially providing an RNA molecule with much of the structural versatility of proteins (Switzer et al. 1989; Piccirilli et al. 1990). The only way to learn why Nature did not avail itself of this structural banquet is to make, by chemical synthesis, the bases themselves and to study their chemistry. Hexose DNA (Eschenmoser and Loewenthal 1992), floppy DNA (Schneider and Benner 1990), DNA with unnatural bases, DNA with altered linking groups (Huang et al. 1991), and a variety of other unusual structures (Van der Woerd et al. 1987) now grace the chemical literature, all synthesized to answer the question "Why not?"

GENERAL VIEWS CONCERNING DIVERGENT EVOLUTION IN PROTEINS

A systematic analysis of divergent evolution of protein macromolecules casts new light on a wide range of widely held beliefs concerning macromolecular evolution. We cannot review more than a few of these, but the discussion below should encourage caution throughout.

Molecular Clocks

If the rate of accumulation of point mutations is a constant function of time, it should be possible to assign a chronological date to the divergence of two lineages simply by a comparison of protein sequences. In the absence of dated fossils, this is one of the few approaches available for obtaining such chronology. Not surprisingly, many have hoped that such a "molecular clock" exists (Lewin 1988).

Although the divergence of nonfunctional (neutral) aspects of macromolecular sequences (e.g., codon use) may display clock-like behavior, it is clear from rigorously constructed evolutionary trees (e.g., Fig. 2) that this is not generally the case for functional sequences (see also Jukes and Holmquist 1972). In such trees, the lengths of the lines (representing the number of accepted point mutations) between the ends of the branches (representing contemporary sequences) and the common branch points are variable, even though the time interval in each case is identical. Thus, assigning chronological dates for branch points in an evolutionary tree remains problematic (see below).

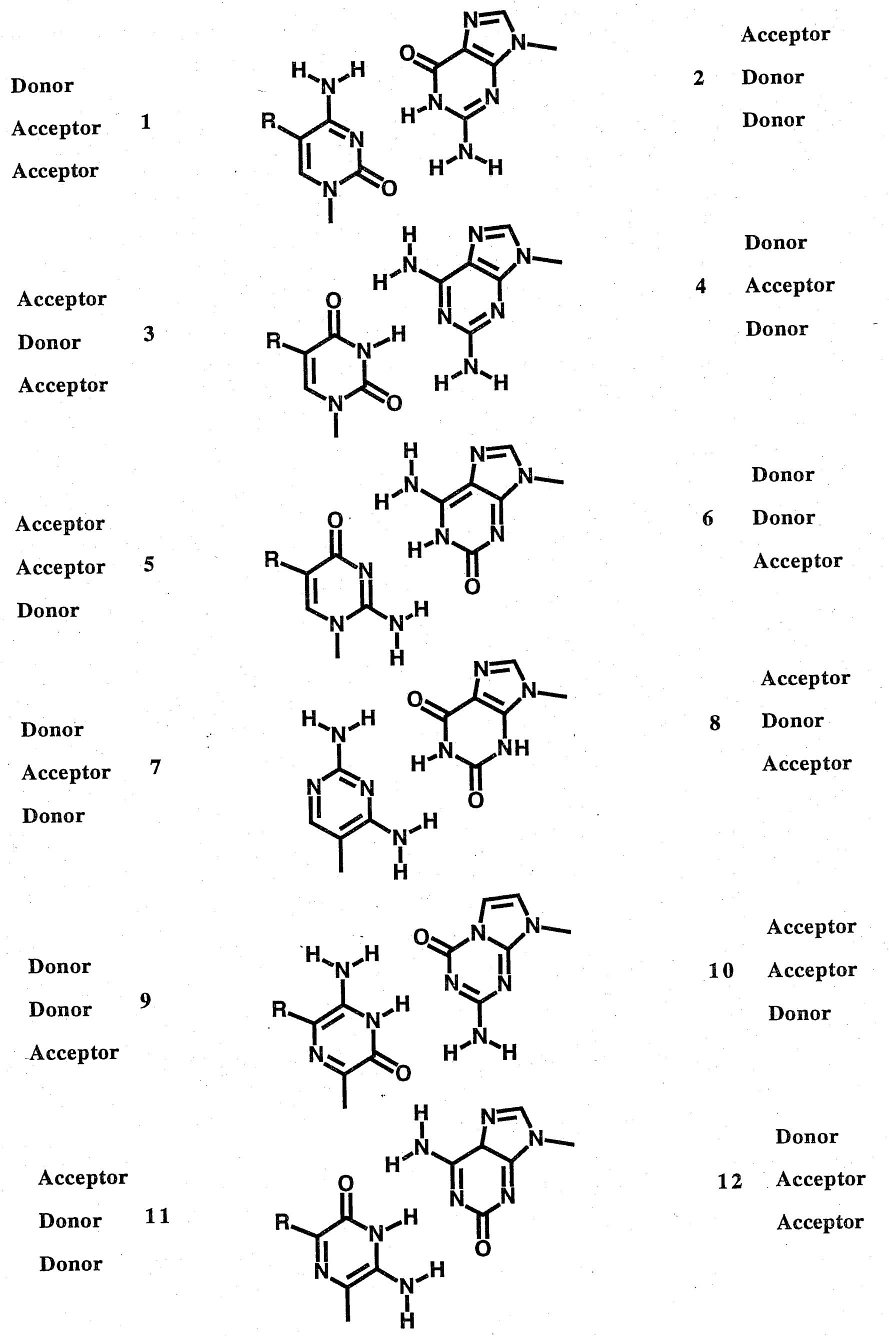


Figure 4 Six isomorphic Watson-Crick base pairs can be constructed from 12 purine and pyrimidine base analogs.

Exon Shuffling

In essentially all protein families that participate in central metabolism, divergence of function has been accomplished through the accumulation of point mutations, insertions, and deletions. Overall, these processes behave approximately statistically. The family of proteins represented by the tree in Figure 2 is an excellent example of this.

There is essentially no evidence in metabolic enzymes for modular behavior (e.g., domain shuffling), either for developing primitive catalysts or for altering the function of advanced enzymes. Rather, modular behavior is observed primarily in proteins involved in advanced regulatory systems in advanced organisms; in particular, proteins involved in the immune system, blood clotting, and other regulatory pathways that emerged only within the last 400 million years. The particular attention paid to these types of proteins by contemporary molecular biologists has, we believe, created the illusion that domain shuffling is more widespread (and more ancient) than it actually is.

Recruitment of Proteins and Deletion-replacement Events

Protein families containing two (or more) types of catalysts suggest a process of "recruitment," where an enzyme performing one function is recruited (often following gene duplication) to perform a second function. Recruitment presents special challenges in reconstructive efforts, because it is difficult to decide which of the two functions was performed by the common ancestral protein (i.e., which function is "primitive" and which is "derived"). Preparation of the ancestral proteins in the laboratory (see above) is possible only in special cases. Furthermore, whereas the derived function might have been created for the first time by recruitment, it is also possible that the ancestral organism had catalysts for both reaction types. If so, the recruitment was the second step of a "deletion-replacement" event, where the gene encoding the enzyme for one reaction was lost and a replacement was obtained by recruitment. Deletion-replacement events occurring on the laboratory time scale have been well studied (Li et al. 1983).

The frequency of deletion-replacement events is unknown, although inspection of the connected components obtained from the exhaustive matching (Gonnet et al. 1992) can provide a guess. Of the approximately 70 connected components that contain an archaebacterial sequence and a sequence from at least one other kingdom (Table 4, A–E), 16 (22%) contain enzymes catalyzing more than one type of reaction. It is difficult to know whether this overestimates or underestimates actual recruitment events. The criteria used to construct the connected components are quite liberal (footnote 2), and several of these trees almost certainly include proteins that are not homologs as a result (see notes to Table 4). This suggests that we will overestimate the frequency of deletion-replacement events by using this sample. However, the connected components may not include some homologous proteins that perform different functions where significant sequence similarity no longer remains. Furthermore, the sequence database is far from complete. These factors suggest that we have underestimated the frequency of deletion-replacement events.

Nevertheless, deletion-replacement events follow rules (Benner et al. 1989). For an existing enzyme performing a specific biological function to be replaced, it first must be deleted to yield an organism that lacks, at least for a time, this function. The more "lethal" the deletion is, the more difficult a deletion-replacement event. Furthermore, deletion-replacement events are more likely in an organism that already contains an enzyme that catalyzes a reaction chemically similar to that catalyzed by the deleted enzyme; this enzyme is ready to be recruited following a relatively small number of structural alterations. This implies that deletion-replacement events occur most rapidly in an organism that has a wide variety of enzyme types.

Although the literature occasionally focuses on conservation of binding sites (Yeh and Ornston 1980), conservation of catalytic mechanism is far more important during divergent evolution. For example, in the evolutionary tree indicated in Figure 2, at least three reaction types are represented. All require pyridoxal cofactors, however. The tree also suggests a solution to a particular biochemical puzzle in the literature: What is the enzymatic function of the cobC (entry e in Fig. 2), a genetic entity involved in the biosynthesis of vitamin B_{12} in *Pseudomonas*? (Crouzet et al. 1990). The literature suggests, on the basis of complementation and other biological data, that cobC encodes an amidase. However, there is clear homology between the cobC protein and several aminotransferases and decarboxylases dependent on pyridoxal. This evolutionary information coupled with chemical information concerning the biosynthesis of B_{12} makes it more likely that cobC protein is a threonine decarboxylase.

RECONSTRUCTING EARLY FORMS OF LIFE

With these tools in hand, we can begin to reconstruct detailed models for the biochemistry of ancient forms of life. This reconstruction proceeds stepwise; the metabolism of the "proto-mammal," the "proto-animal," the "proto-plant," the "proto-fungi," and the "proto-eukaryote" are reconstructed in this order. The prefix "proto-" is used exactly as in the field of

Table 4 Connected components containing archaebacterial sequences

Component	Size	Comments
		(A) Enzymes with at Least One Representative from Each Primary Kingdom
33[a]	16	pyridoxal enzymes (aminocyclopropanecarboxylate synthase, tyrosine aminotransferase, aspartate aminotransferase, histidinol phosphate aminotransferase, lysine decarboxylase, ornithine decarboxylase, CobC, and a hypothetical protein
332	17	ATP-citrate lyase, citrate synthase, succinyl CoA synthase
349[b]	126	oxidoreductases (acyl CoA desaturase, cytochrome b5, ferredoxin, flavodoxin, nitrate reductase, and others)
1183	8	argininosuccinate synthase
1277[c]	74	ATP synthase α,β and vacuolar subunits
1685	12	β-galactosidase, β-glucosidase, lactase-phlorizin hydrolase precursor
1735[a]	8	biotin sulfoxide reductase, dimethyl sulfoxide reductase, formate dehydrogenase, NAD-reducing hydrogenase, respiratory nitrate reductase, NAD-ubiquinone reductase
1859[a]	29	various oxidoreductases
2181[a]	51	carbamoyl-phosphate synthase, anthranilate synthase, D-ala-D-ala ligase, PAB synthase, isochorismate synthase, propionyl CoA carboxylase
4018	46	*O*-acetylserine sulfhydrylase, serine dehydratase, threonine dehydratase, threonine synthase, tryptophan synthase α and β subunits
4637	21	dihydrofolate reductase
5847	21	various hydrogenases
5942	51	glyceraldehyde 3-phosphate dehydrogenase
6465	43	glutamine synthase
13113	26	phosphoglycerate kinase
13288	5	DNA photolyase
13903	5	pyrroline-5-carboxylate reductase
14385	7	phosphoribosylaminoimidazole carboxylase
16521	23	superoxide dismutase Fe/Mn/Fe-Mn

(B) Ribosomal Proteins with at Least One Representative from Each Primary Kingdom

Component	Size	Comments
10487	5	S5 and putative LLREP3
12533	6	L6, L9, and outer membrane binding protein
14487	18	S14, S11, CRP-2, RP59
15018	6	L11, YL15
15028	16	L14, L17A
15089	19	L2, KD4, K37
15133	15	L23, L25
15180	10	L3
15250	10	39A, L5
15384	24	S12
15834	7	S10, S20
15904	15	S19, S15
15914	6	S16, S9
15976	12	S22, S8
16039	20	S7

(C) Ribosomal Proteins with at Least One Archaebacterial and One Eukaryotic Sequence and No Eubacterial Representative

Component	Size	Comments
26	4	S13, S15, and an unknown 17.4-kD protein
8352	4	HS3, S4E, S4
15005	6	L1, L1A, L1B, L2, L4
15047	6	L15, L29, L27A
15074	5	L18, L5, L5A, L5B
15078	5	L19, L19E, and probable ribosomal protein (OrfE)
15125	3	L22, L23

Table 4 Connected components containing archaebacterial sequences (*continued*)

(C) Ribosomal Proteins with at Least One Archaebacterial and One Eukaryotic Sequence and No Eubacterial Representative (continued)

Component	Size	Comments
15151	3	L24, L26
15192	7	L30, L7
15194	4	L30, L32, probable ribosomal protein (Orf1), hypothetical 11.5-kD protein
15204	5	L32, RP49, and probable ribosomal protein (OrfD)
15211	4	L32, L35A, and a hypothetical 9.7-kD protein
15258	4	YL4, L7A, HS6
15842	5	S11, S17
15865	4	HS12, S15A, S19
15958	4	S19, S24

(D) Enzymes with at Least One Archaebacterial and One Eubacterial Sequence and No Eukaryotic Representative

Component	Size	Comments
855	34	nitrogenase Fe-(Fe/Mo/V) protein; nitrogenase Fe-Mo cofactor biosynthesis protein NIFE
1616	2	bacterio-opsin activator/nitrogen fixation regulatory protein
4095	18	DNA-binding protein and integration host factor
5846	2	8-hydroxy-deazaflavin-reducing hydrogenase/hypothetical protein
8003	7	phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase; HISF protein (cyclase)
17808	6	indole-3-glycerol phosphate

(E) Archaebacterial Sequences from Connected Component 1[d]

Archaebacterial enzyme	Similarity	Related enzymes[e]
DNA gyrase subunit B	120	DNA topoisomerase II (E), DNA topoisomerase large subunit (V), DNA gyrase subunit B (B)
Acidic ribosomal protein	135	ribosomal protein L20/L12 (A), acidic ribosomal protein P0 (E), protein homolog P0
Ribosomal protein L20	120	ribosomal protein A/L12 (A), acidic ribosomal protein P1/P2 (E)
Ribosomal protein L12	120	ribosomal protein A/L20 (A), acidic ribosomal protein P1/P2 (E)

DNA-directed RNA polymerase A	120	DNA-directed RNA, polymerases (A,E,B and V)
DNA-directed RNA polymerase B	120	DNA-directed RNA, polymerases (A,E,B and V)
DNA-directed RNA polymerase C	130	DNA-directed RNA, polymerases (A,E,B and V)
Elongation factors[f]		elongation factors (A,E, and B)
	(F) Archaebacterial Singletons (Unconnected Entries)[g]	
22.6-kD Protein		*Desulfurococcus mobilis*
24-kD Flagellin (frag)		*Methanospirillum hungatei*
30S Ribosomal protein hs13		*Halobacterium marismortui*
30S Ribosomal protein s14		*Methanococcus vannielii*
31-kD Flagellin (frag)		*Methanococcus voltae*
50S Ribosomal protein hl5 (frag)		*Halobacterium cutirubrum*
50S Ribosomal protein hl5 (frags)		*H. marismortui*
50S Ribosomal protein hl9 (frag)		*H. cutirubrum*
50S Ribosomal protein hl10 (frag)		*H. cutirubrum*
50S Ribosomal protein hl16 (frag)		*H. cutirubrum*
50S Ribosomal protein hl21/hl22		*H. marismortui*
50S Ribosomal protein hl29 (l19)		*H. marismortui*
50S Ribosomal protein hl30 (frag)		*H. cutirubrum*
50S Ribosomal protein hl31 (frag)		*H. cutirubrum*
50S Ribosomal protein hl32		*H. marismortui*
50S Ribosomal protein l5 (hl19) (frag)		*H. cutirubrum*
50S Ribosomal protein l6 (hmal6) (hl10) (frag)		*H. marismortui*
50S Ribosomal protein l9 (frag)		*H. marismortui*
50S Ribosomal protein l12 (frag)		*H. marismortui*
50S Ribosomal protein l13 (frag)		*H. cutirubrum*
50S Ribosomal protein l22 (hl23) (frag)		*H. cutirubrum*
50S Ribosomal protein l29 (hmal29) (hl33)		*H. marismortui*
50S Ribosomal protein l29		*Halobacterium halobium*
50S Ribosomal protein l31		*H. marismortui*

Table 4 Connected components containing archaebacterial sequences (*continued*)

(F) Archaebacterial Singletons (Unconnected Entries)[g] (continued)	
50S Ribosomal protein l34 (hl30)	*H. marismortui*
50S Ribosomal protein l46e	*Sulfolobus solfataricus*
50S Ribosomal protein lc12 (frag)	*H. marismortui*
50S Ribosomal proteins hl46e	*H. marismortui*
Cell surface glycoprotein precursor (csg)	*H. halobium*
Deoxyribodipyrimidine photolyase (E.C. 4.1.99.3) (frag)	*Methanobacterium thermoautotrophicum*
DNA binding protein hmf-2 (hmfb)	*Methanothermus fervidus*
DNA gyrase subunit a (E.C. 5.99.1.3) (gyra) (frag)	*Haloferax sp. (strain aa 2.2)*
DNA-binding protein 7e	*Sulfolobus acidocaldarius*
DNA-binding proteins 7a, 7b, and 7d	*S. acidocaldarius*
Ferredoxin	*Methanococcus thermolithotrophicus*
Ferredoxin	*Methanosarcina barkeri*
Ferredoxin	*S. acidocaldarius*
Formylmethanofuran--THMP formyltransferase (E.C. 2.3.1.101)	*M. thermoautotrophicum*
Gas vesicle protein, chromosomal (c-vac) (gvp) (gvpa)	*H. halobium*
Gas vesicle protein, plasmid (p-vac) (gvp) (gvpa)	*H. halobium*
Glucose dehydrogenase (E.C. 1.1.1.47) (frag)	*Thermoplasma acidophilum*
Gvpd protein (gvpd)	*H. halobium*
Gvpe protein (gvpe)	*H. halobium*
Hypothetical 7.1-kD protein in hmal3 5′ region (Orf2)	*H. marismortui*
Hypothetical 8.5-kD protein (Orf75)	*H. halobium*
Hypothetical 9.9-kD protein in rpob 5′ region (Orf88)	*S. acidocaldarius*
Hypothetical 10.3-kD protein in ribosomal proteins operon (Orf3)	*H. marismortui*
Hypothetical 10.9-kD protein in rs10 3′ region (Orf4)	*M. vannielii*
Hypothetical 11.3-kD protein in gapdh 3′ region (Orfb)	*Pyrococcus woesei*
Hypothetical 11.5-kD protein in rs17 5′ region (Orfa)	*M. vannielii*
Hypothetical 11.6-kD protein in rs17 5′ region (Orfb)	*M. vannielii*
Hypothetical 13.7-kD protein in transposable element ish50	*H. halobium*

Hypothetical 15.2-kD protein	*M. thermoautotrophicum*
Hypothetical 15.3-kD protein in rs10 3′region (Orf3)	*M. vannielii*
Hypothetical 15.6-kD protein in phr 5′region	*H. halobium*
Hypothetical 16.7-kD protein	*M. thermoautotrophicum*
Hypothetical 18.7-kD protein in ribosomal RNA operon	*Thermofilum pendens*
Hypothetical 23-kD protein in ribosomal protein gene cluster (nab)	*H. cutirubrum*
Hypothetical 23.1-kD protein in hmal3 5′region (Orf1)	*H. marismortui*
Hypothetical 24.4-kD protein in lacs 3′region	*S. solfataricus*
Hypothetical 24.7-kD protein in gapdh 5′region (Orfa)	*P. woesei*
Hypothetical 28.5-kD protein in 7s RNA 5′region (Orf260)	*M. fervidus*
Hypothetical 30.7-kD protein	*M. thermoautotrophicum*
Hypothetical 31-kD protein in transposable element ish50	*H. halobium*
Hypothetical 38-kD protein in 23s RNA operon	*Thermoproteus tenax*
Hypothetical 46.2-kD protein (Orfb)	*Methanobrevibacter smithii*
Hypothetical 60.5-kD protein	*M. thermoautotrophicum*
Hypothetical 80.2 kd protein (Orf4)	*Haloferax sp. (strain aa 2.2)*
Hypothetical protein in ftr 5′region (frag)	*M. thermoautotrophicum*
Hypothetical protein in gapdh 3′region (Orfx)	*P. woesei*
Hypothetical protein in gapdh 5′region (Orfz) (frag)	*P. woesei*
Hypothetical protein in glna 3′region (frag)	*M. voltae*
Hypothetical protein in hmal29 3′region (frag)	*H. marismortui*
Hypothetical protein in nifh 3′region (frag)	*M. thermolithotrophicus*
Hypothetical protein in nifh 5′region (frag)	*M. thermolithotrophicus*
Hypothetical protein in ribosomal l1 protein 5′region (frag)	*M. vannielii*
Hypothetical protein in ribosomal protein s11 3′region (frag)	*H. marismortui*
Hypothetical protein in ribosomal proteins operon (Orf1) (frag)	*H. marismortui*
Insertion element ism1 Hypothetical 48.3-kD protein (Orfis)	*M. smithii*
Malate dehydrogenase (E.C. 1.1.1.37) (E.C. 1.1.1.82) (mdh)	*M. fervidus*
Malate dehydrogenase (E.C. 1.1.1.37) (frag)	*S. acidocaldarius*
Membrane-associated ATPase γ chain (E.C. 3.6.1.34) (atpd)	*S. acidocaldarius*

Table 4 Connected components containing archaebacterial sequences (*continued*)

(F) Archaebacterial Singletons (Unconnected Entries)[g] (continued)	
Membrane-associated ATPase γ chain (E.C. 3.6.1.34)	*S. acidocaldarius*
Methyl CoM methylreductase α subunit (frag)	*Methanosarcina thermophila*
Methyl CoM methylreductase β subunit (frag)	*M. thermophila*
Methyl CoM methylreductase γ subunit (frag)	*M. thermophila*
Methyl CoM reductase II α subunit (frag)	*M. thermoautotrophicum*
Methyl CoM reductase II β subunit (frag)	*M. thermoautotrophicum*
Methyl CoM reductase II γ subunit (frag)	*M. thermoautotrophicum*
N-(5′-phosphoribosyl)anthranilate isomerase (trpf) (frag)	*M. voltae*
Nitrogenase Mo-Fe protein β chain (E.C. 1.18.6.1) (frag)	*M. thermolithotrophicus*
Ribosomal protein "a" (frag)	*M. thermoautotrophicum*
Thermopsin precursor (E.C. 3.4.23.-)	*S. acidocaldarius*

From SwissProt Release 19 (November 1991). Connected components are defined as sets of sequences where every sequence in the set is connected to at least one other sequence in the set by a PAM distance less than 200 provided that the similarity of the match is greater than 125 and extends over at least 80 positions. These were generated using the DARWIN system (Gonnet and Benner 1991). The connected components are unedited and may include at this PAM distance grouping of nonhomologous families. In connected components that contain more than one functional class of enzyme, the name of the class that includes an archaebacterial sequence is underlined.

[a] Only the archaebacterial aminotransferase is in this connected component. Eubacterial and eukaryotic AATs are in a separate connected component.

[b] These connected components arise from matches that may be insignificant statistically. They are properly resolved into several probably nonhomologous protein families.

[c] The α and β subunits form a single connected component at this level. However, they are more properly treated as separate components because the duplication resulting in the two chains is believed to have occurred before the divergence of the 3 primary kingdoms (Gogarten et al. 1989).

[d] Connected Component 1 (the "black hole") contains a large number (>5527) of protein entries joined by repetitive sequences, possibly arising from convergent sequence evolution. In these repetitive sequences, a small number of amino acid types are overrepresented compared to the database as a whole, and independent variation at each position according to the standard Dayhoff matrix can no longer be plausibly assumed and statistical analysis is problematic. In the protein pairs reported here, the similarity criteria noted above were achieved outside the region of the repetitive sequence.

[e] Related enzymes are archaebacterial (A), eukaryotic (E), eubacterial (B), or viral (V) enzymes as indicated.

[f] A matching of archaebacterial elongation factor sequences against the database as a whole shows matches likely to be nonsignificant at a level that excludes some nonarchaebacterial elongation factors.

[g] Protein sequences unconnected with other proteins at PAM 200 units. Many of these sequences are fragments and therefore may be excluded from a relevant connected component because of the size constraint. Of these 93 sequences, 35 are labeled hypothetical.

historical linguistics (Lehmann 1973; Benner and Ellington 1990a), where it designates a language reconstructed from descendant languages using a principle of parsimony and rules for transformation (e.g., "proto-Indoeuropean").

Most interesting in the context of this volume is the organism containing the "protogenome," the genome in the most recent common ancestor of archaebacteria, eubacteria, and eukaryotes (Benner and Ellington 1990a). We therefore jump directly back to this organism, a leap that takes us back over 1 billion years and omits, unfortunately, much important discussion of evolution in the intervening period.

The Protogenome: The Most Recent Common Ancestor of Archaebacteria, Eubacteria, and Eukaryotes

Reconstruction of the protogenome (Benner and Ellington 1990a) begins with sequence data for proteins performing analogous functions in the three kingdoms of life (Balch et al. 1977; Woese and Fox 1977). Two situations can exist. First, proteins having analogous biological functions in archaebacteria, eubacteria, and eukaryotes might be found in homologous forms in all three kingdoms. In this case, the protogenomic sequence can be reconstructed (see above), and the function can be placed in the protogenome. The possibility that the protein appears in homologous form in all three kingdoms via lateral transfer of genetic information is generally discounted, unless the gene is plasmid-encoded or if lateral transfer is indicated on other grounds. For example, the gas vacuolar protein is homologous in the archaebacterium *Halobacterium halobium* and the eubacterium *Pseudoanabaena* (Horne and Pfeifer 1989). However, because a version of the gene is plasmid-borne (Dassarma et al. 1987), the contact between these organisms in brines is intimate, and the gene is unconnected with a metabolic pathway, the possibility of lateral transfer cannot be ignored.

Second, proteins performing analogous biological roles in archaebacteria, eubacteria, and eukaryotes might *not* be homologous in all three kingdoms. In this case, either the protogenome did not encode the function at all (implying that the function arose by independent convergent evolution in all three kingdoms or by lateral transfer between kingdoms), or it did encode the function, but the molecule encoded in the protogenome was replaced in one or more of the derived kingdoms.

Multiple deletion-replacement events in the three lineages derived from the protogenome are normally considered improbable, except in one particular case. If a reaction was catalyzed in the protogenome by a riboenzyme, the possibility of multiple deletion-replacement events is presumably larger. Protein enzymes are (again presumably) more efficient than riboenzymes because they have a larger repertoire of encoded functional groups (see above). Thus, if a metabolic pathway is clearly assigned to the protogenome on independent grounds, but is catalyzed in the descendant kingdoms by nonhomologous proteins, and if the function is critical and the reaction unusual (implying that deletion-replacement events are slow), the hypothesis that a riboenzyme encoded by the protogenome was replaced independently in the three kingdoms is considered plausible. This is the case, for example, with ribonucleotide reductases (Benner et al. 1989).

Table 4 lists connected components in the current database containing sequences from both archaebacteria and at least one other kingdom. These are therefore families of proteins that might have common ancestors encoded by the protogenome. The list suggests two conclusions of immediate significance. First, an analysis of the multiple alignments for these connected components permits the reconstruction of more than 6,000 amino acids derived from at least 18,000 base pairs in the protogenome.[3] The encoded proteins include at least 16 ribosomal proteins and 20 enzymes, including proteins involved in replication (DNA-gyrase), protein synthesis (DNA-directed RNA polymerases, elongation factors), amino acid biosynthesis (tryptophan synthase, histidinol phosphate aminotransferase), glycolysis (phosphoglycerate kinase), the urea cycle (argininosuccinate lyase, carbamoyl phosphate synthase), redox reactions (ferredoxins, NADH-ubiquinone oxidoreductase), and ATP synthesis (ATP synthase).

Table 4 also establishes that the metabolism encoded by the protogenome was rather complex (Fig. 5). Two enzymes from the glycolytic pathway leading to the citric acid cycle (phosphoglycerate kinase and citrate synthase) are rigorously reconstructed in the protogenome; a third (glyceraldehyde-3-phosphate dehydrogenase) is tentatively assigned. These reconstructions suggest that the protogenome coded for the full glycolytic pathway as phosphoglycerate (the product of the reconstructed phosphoglycerate kinase) is not likely to be an end product of metabolism, nor is acetyl CoA (one substrate for the reconstructed citrate synthase) likely to have been obtained from the diet. Such a hypothesis, of course, is experimentally testable. For example, archaebacterial lactate dehydrogenases and enolases should exist and should be homologous to their eubacterial and eukaryotic counterparts.

[3] To obtain this estimate, positions in the multiple alignments deleted in any of the proteins were ignored. It must be noted that the 6000 amino acids representing 36 proteins do not indicate an average protein size of 166 amino acids for these proteins in the last common ancestor of the three kingdoms, but rather that the amino acids can be reliably assigned in an average of 166 positions per protein.

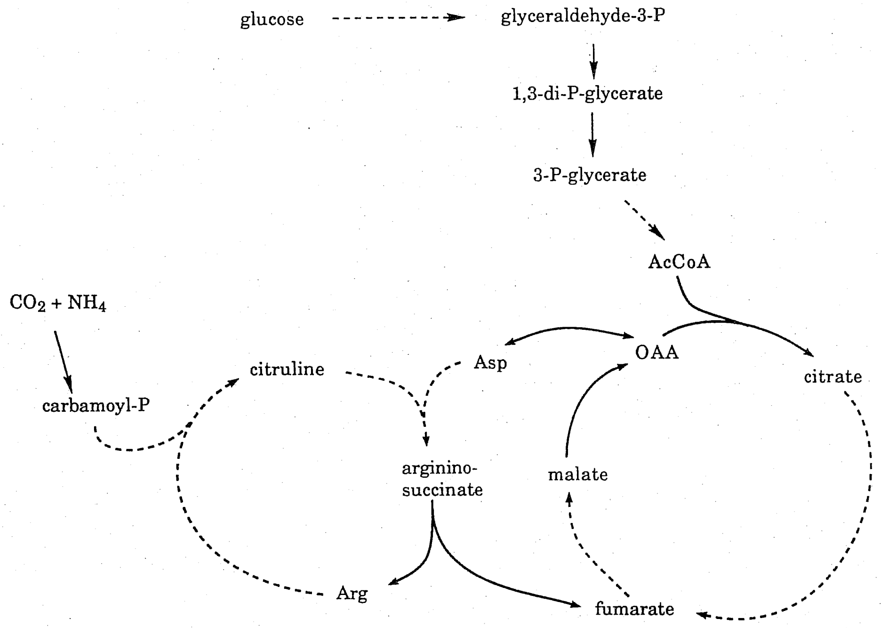


Figure 5 Diagram of the central metabolism encoded by the protogenome, the most recent common ancestor of the genome found in archaebacteria, eubacteria, and eukaryotes. Solid lines indicate metabolic reaction steps catalyzed by enzymes that can be reconstructed in the most recent common ancestor of archaebacteria, eubacteria, and eukaryotes. Broken lines indicate metabolic steps implied on chemical grounds from the reconstructed enzymes but catalyzed by enzymes yet to be reconstructed by sequence comparisons of homologous enzymes in all three kingdoms. The reconstruction of an ancestral malate dehydrogenase enzyme catalyzing the oxidation of malate to oxaloacetate is only poorly reconstructed.

Furthermore, the reconstructed tryptophan synthase, aspartate aminotransferase, and histidinol phosphate transaminase support the assumption that the protogenome encoded the biosynthesis of the full repertoire of amino acids. This assumption is, of course, also suggested by the fact that the reconstructed proteins encoded by the protogenome include all 20 proteinogenic amino acids found in contemporary organisms.

The reconstructed protogenome all but disproves the prevailing view that the most recent common ancestor of archaebacteria, eubacteria, and eukaryotes was a "progenote," an organism with highly imprecise mechanisms for replicating and translating genetic information (Woese and Fox 1977). Genome size and fidelity of handling of genetic information are directly correlated (Eigen and Schuster 1977). A genome of a size needed to encode such a metabolism is sustainable only by reasonably precise mechanisms for copying and translating genetic information.

Rather interesting evidence that the protogenome did not live in a progenote comes from the field of peptide design (see above). It has proven relatively easy to prepare relatively small polypeptides that catalyze the decarboxylation of oxaloacetate (Johnsson et al. 1990). It therefore appears that a substantial fraction of random functionalized polypeptides will catalyze destruction of oxaloacetate. An organism without reasonable control over what peptides it makes would therefore find oxaloacetate a problematic metabolic intermediate, because oxaloacetate would not survive in a cell where proteins with imprecise structures were being prepared. From the reconstruction, however, it is clear that the protogenome lived in a cell with oxaloacetate as a metabolic intermediate; at least two enzymes that use oxaloacetate as a substrate (citrate synthase and aspartate aminotransferase) can be reconstructed and placed in the protogenome. This implies that the protogenome encoded proteins that effected a metabolism involving oxaloacetate as a metabolic intermediate, suggesting in turn that the protogenome was not in a progenote.

Many pathways in contemporary organisms are missing from the reconstructed protogenome, however. Enzymes requiring biotin are absent in the reconstruction at this point. Pathways for the biosynthesis of straight-chain fatty acids are also absent; presumably, the protogenome resided in cells with terpenoid membranes, as do contemporary archaebacteria. In contrast, there is no evidence to reconstruct enzymes involved in methanogenesis, the classic metabolic pathway found in a major branch of archaebacteria. Thus, archaebacteria appear to display only some primitive traits, a statement that applies to each of the main kingdoms, suggesting a certain inappropriateness of the name "archaebacteria" (and, unfortunately, its more recently suggested alternative, "Archaea") (Woese et al. 1990).

Assigning a chronological date to the protogenome is, of course, extremely difficult in view of the absence of a reliable molecular clock (see above). It is possible, however, to reconstruct certain enzymes in the protogenome that suggest that it lived in an environment where molecular oxygen was available. Superoxide dismutase, gas vacuolar proteins, and the C5 pathway involved in the synthesis of chlorophyll (Friedmann et al. 1987; Kannangara et al. 1988) can all be assigned to the protogenome with varying degrees of reliability. Each of these is related in some way to oxygen in the atmosphere. Superoxide dismutase removes toxic products of oxygen metabolism. The gas vacuolar protein allows an organism to float at a position in water relative to an oxygen-containing atmosphere. Chlorophyll is essential (at least in the contemporary world) in the photosynthetic generation of molecular oxygen. Regardless of whether these arguments are definitive, making a connection between the protogenome and molecular oxygen is important in efforts to determine when the protogenome lived. Geological records suggest that molecular oxygen appeared on earth approximately 2.5 billion years ago. Fossil records make almost certain that the three kingdoms diverged before 1 billion years ago. Thus, if the protogenome lived in an oxygen atmosphere, it must have lived within these two limiting dates.

The Last Ribo-organism

In the metabolism encoded by the reconstructed protogenome, RNA is a more important component of catalysis than in contemporary organisms. RNA cofactors of all sorts are reliably placed in the protogenome, as is ribosomal RNA. Additional catalytic activities (e.g., ribonucleotide reductase) are also assigned to the protogenome in a riboenzymatic form (Benner et al. 1989). It is now well recognized that these facts are consistent with the presumption that the protogenome is itself derived from a more ancient genome that encoded RNA catalysts exclusively (Rich 1962; White 1976; Visser and Kellogg 1978; Gilbert 1986; Orgel 1986). It is useful to punctuate the episode in the history predating the protogenome by a breakthrough: the invention of a translation apparatus to translate an encoded mRNA (Benner et al. 1989). The breakthrough organism separates life that possessed an mRNA from life that had RNA as the sole genetically encoded component of biological catalysts.

Because information concerning only a single descendant lineage of the breakthrough organism (that leading to the protogenome) is available, rules of parsimony cannot assist reconstructions of the breakthrough organism. Rather, reconstructions must rely on rules of transformation alone. The simplest of these identifies components of the protogenomic metabolism that involve RNA performing roles where RNA is not an optimal chemical solution to the particular biochemical problem (Benner et al. 1989). RNA used in this capacity is presumed to be a vestige of an earlier time where RNA was the only available encoded molecule: the RNA world. Thus, these roles are assigned to a reconstructed metabolism for the breakthrough organism.

By applying this simple rule, one is forced to conclude that the last ribo-organism had a relatively complex metabolism that included oxidation and reduction reactions, aldol and Claisen condensations, transmethylations, porphyrin biosynthesis, and an energy metabolism based on nucleoside phosphates, all catalyzed by riboenzymes (Benner et al.

1989). It should be noted that this reconstruction cannot be weakened without losing much of the logical and explanatory force of the RNA world model. Nevertheless, the reconstructed metabolism of the breakthrough organism has proven to be the most controversial aspect of the "palimpsest" model (Maizels and Weiner 1987).

Many of the ribozymes in the breakthrough organism evidently were replaced by protein-based enzymes during the time separating the breakthrough from the protogenome. These are deletion-replacement events, governed by the rules noted above. The rate of deletion-replacement processes was determined by two factors: the degree to which a deletion of the riboenzyme would be lethal, and the possibility of finding a protein already participating in the metabolism that catalyzed a reaction chemically similar to the reaction of the deleted riboenzyme. In all cases, the rate of deletion replacement undoubtedly accelerated as the protein-based metabolism became more versatile. Some riboenzymes proved to be extremely difficult to replace; those involved in the ribosome-based synthesis of proteins apparently have survived until the present day (Noller et al. 1992). Others, such as the riboenzymes that encoded the conversion of ribonucleotides to 2′-deoxyribonucleotides (the ribonucleotide reductases) apparently resisted replacement until well after the divergence of the descendants of the protogenome (Benner et al. 1989).

It is impossible to estimate the time interval separating the protogenome from the breakthrough organism, other than to say that it must have been substantial. Each of the 36 or more proteins (with a total of 6000 encoded amino acids) apparently developed de novo in the interval separating the protogenome from the breakthrough organism. Thus, there is no reason to view the protogenome and the breakthrough genome as being close, either in time or metabolically.

Earlier in the RNA World

The reconstructions presented here bring us only to the end of the RNA world, the focus of this volume. Extrapolation further back in time is extremely difficult. As with the breakthrough organism, parsimony is unavailable to help us reconstruct events in the RNA world. Furthermore, transformation rules (e.g., mutation matrices) for reconstructing the evolution of catalytically functional nucleic acids are simply unavailable at present. Transformation rules might at best come from a better understanding of ribosomal RNA, RNAs from snRNPs, and other catalytic RNAs that are presumably vestiges of the RNA world. Even here, the process is risky, because many popular proposals that contemporary RNA molecules are vestiges of the RNA world, including the selfsplicing introns from *Tetrahymena* (as a primitive vestige of the first self-replicating RNA molecule) and genomic tags (Weiner and Maizels 1987), are at best only weakly supported at this time by rigorous evolutionary analysis.

Nevertheless, the reconstructed metabolism of the breakthrough organism can influence models of earlier events and structures in the RNA world itself. For example, in most models, translation was assembled in an environment that is either prebiotic or primitive metabolically. The reconstructed metabolism of the breakthrough organism suggests that this view is incorrect. Rather, translation apparently arose in a relatively complex metabolic background provided by the breakthrough organism. It is, of course, much easier to design a new catalyst (including a catalyst for translating an mRNA) in a complex metabolic environment than in a simple one. Because none of the reactions involved in ribosome-based translation are exceptional (phosphate anhydride exchange, carboxylate ester formation, and aminolysis of a carboxylate ester being the only types), and because the reconstructed metabolism of the breakthrough organism contains several pathways that must have involved riboenzymes catalyzing similar reactions, there is no need to struggle to build models describing the origin of translation in the primitive world.

Indeed, specific RNA-catalyzed metabolic processes in the reconstructed breakthrough organism are plausible precursors for components in the first ribozyme. For example, the charging of amino acids is chemically similar to the first step in the pathway for preparing porphyrins reconstructed first in the protogenome on biochemical evidence (Friedmann et al. 1987), and from there to the breakthrough organism (Benner et al. 1989).[4]

THE PYRIDOXAL PARADOX

Reconstructions take on additional significance when directed toward specific problems in biological chemistry. Pyridoxal is perplexing in the context of the palimpsest model. It is present in all three kingdoms (Noll and Barber 1988) and therefore might be assigned to the protogenome according to a rule of parsimony. However, unlike nicotinamide cofactors, flavin cofactors, CoA, and ATP, for example, pyridoxal does not

[4] Sequence data suggest that the pyridoxal-dependent enzymes catalyzing the condensation of glycine with succinyl CoA in eubacteria and eukaryotes are not homologous, suggesting that this competing pathway for preparing aminolevulinate (without the need for an RNA cofactor) was not encoded by the protogenome (Li et al. 1989).

have an RNA substituent that identifies it as a cofactor originating in the RNA world (Orgel 1968; White 1976; Visser and Kellogg 1978). In this respect, it is similar to biotin. However, pyridoxal is unlike biotin in that its chemistry is easily modeled, a characteristic suggested by Visser and Kellogg (1978) for cofactors that originated in the RNA world. Although pyridoxal is formally accessible by combination of ribose and glyceraldehyde phosphate, its biosynthesis in contemporary organisms remains obscure (Hill et al. 1987), as do possible routes for nonbiological synthesis.

The enzymology of pyridoxal presents similar paradoxes. With ATPases, dehydrogenases (e.g., dihydrofolate reductase), enzymes that catalyze Claisen condensations (e.g., citrate synthase), and many other classes of enzymes (e.g., polymerases), analogous enzymes in the three kingdoms are generally homologous.[5] Building both trees and the reconstructed protogenomic sequences is direct. When enzymes catalyzing different reactions have been recruited from enzymes within these families, the derived function is generally apparent and not encoded by the protogenome (e.g., the sulfoxide reductases of connected component 1735 and the lactase-phlorizin hydrolase of connected component 1685; Table 5). There is little evidence for deletion-replacement events in central metabolic enzymes, and there is no evidence for large amounts of lateral transfer of genetic information between kingdoms.

This is not the case with pyridoxal-dependent enzymes. Consider just one family of pyridoxal-dependent enzymes, those in Figure 2. Aspartate aminotransferases (AAT) in this tree are found in all three kingdoms. However, the tyrosine aminotransferase (YAT) and AAT from two different kingdoms (archaebacteria and eukaryotes) are more closely related than YAT and AAT from the same kingdom (eukaryotes) (Cubellis et al. 1989). Furthermore, the AAT of *Escherichia coli* shares a common ancestor with the aromatic amino acid aminotransferase of *E. coli* (Fig. 2) that acts on tyrosine. Parsimony suggests that the protoeubacterial protein transaminated aspartate (Fig. 6). Therefore, the common ancestor of the entire family was, according to parsimony, also an AAT, and YATs

[5] Enzymes such as malate dehydrogenase and glyceraldehyde-3-phosphate dehydrogenase are found in all three kingdoms; sequence analysis indicates in one case that there may be no significant homology and in the other that the region of significant homology between all three kingdoms is very small. Although these activities might be tentatively assigned to the last common ancestor, complete reconstruction of ancestral malate and glyceraldehyde-3-phosphate dehydrogenases is not yet possible. Recent results from the *Caenorhabditis elegans* genome project indicate that an expressed protein in *C. elegans* is homologous to archaebacterial malate dehydrogenase (Waterston et al. 1992). This suggests that a new class of malate dehydrogenase is present in at least two of the kingdoms. Elucidation of the structure of this second class of malate dehydrogenases (either by crystallography or prediction after more homologous sequences have been obtained) should resolve the question of the relationship between the two classes of malate dehydrogenase.

were derived more than once from the ancestral AAT. Similarly, histidinol phosphate aminotransferases were also derived, in the most parsimonious model, from the ancestral AAT.

This implies either that there was no enzyme specifically catalyzing transamination reactions involving tyrosine or histidinol phosphate in the protogenome or that the enzymes catalyzing these transformations were deleted and replaced by an enzyme derived from the ancestral aspartate aminotransferase. Because the biosynthesis of tyrosine and histidine was almost certainly encoded by the protogenome (see above), the second case is more probable. This conclusion could, of course, be profoundly altered by additional sequence data; in particular, those for an archaebacterial YAT. However, we might predict that archaebacterial YATs were derived from archaebacterial AAT, if they are members of this family of proteins at all.

A small number of deletion-replacement events in a lineage are not, of course, excluded in the palimpsest model (Benner et al. 1989). However, we have invoked no fewer than three deletion-replacement events to account for the sequences in this tree. Still more must be invoked in the most parsimonious model describing other families of pyridoxal-dependent enzymes. For example, connected component 4018 (Table 5) contains enzymes essential for the biosynthesis of tryptophan, threonine, cysteine, and branched-chain amino acids. Again, multiple deletion-replacement events are required to account for these data, because enzymes involved in the biosynthesis of all these amino acids were presumably encoded by the protogenome.

The situation becomes still more complicated when other pyridoxal reactions are included. For example, the tree in Figure 2 includes some amino acid decarboxylases dependent on pyridoxal cofactors; other amino acid decarboxylases are found in other connected components (Table 5). The pattern has perplexed many authors (Martin et al. 1988; Mehta et al. 1989). Although decarboxylation of amino acids cannot as a pathway be reconstructed in the protogenome, recruitments were apparently still more prevalent within pyridoxal enzymes. To make matters worse, many decarboxylases do not use pyridoxal cofactors, even though the intrinsic chemistry involved would be well suited for pyridoxal chemistry. Instead, they use an enzyme-bound pyruvoyl residue (Recsei and Snell 1984). These include enzymes that decarboxylate aspartate, histidine, and *S*-adenosyl methionine and that are found in both eubacteria and eukaryotes. Furthermore, at least one enzyme, glycogen phosphorylase (from eubacteria and eukaryotes), uses pyridoxal phosphate as a Bronsted acid (Madsen and Withers 1986). In chemical terms, the function of this catalytically essential pyridoxal phosphate could be fulfilled

Table 5 Connected components containing pyridoxal-dependent enzymes

Connected component	Size	Representative entries and comments
1	5972	The "black hole";[a] glutamine decarboxylase
33	14	histidinol phosphate aminotransferase; lysine decarboxylase; ornithine decarboxylase, prokaryotic; 1-aminocyclopropane-1-carboxylate synthase; tyrosine aminotransferase; aspartate aminotransferase, archaebacterial; CobC protein; hypothetical protein, *B. subtilis*
216	14	aspartate aminotransferase, cytoplasmic and mitochondrial
223	1	aspartate aminotransferase, rabbit cytoplasmic (fragment)
225	1	aspartate aminotransferase, bovine mitochondrial (fragment)
231	1	aspartate aminotransferase, rabbit mitochondrial (fragment)
629	1	alanine aminotransferase, pig (fragment)
709	4	alanine racemase
713	1	amino acid racemase, *Pseudomonas* (fragment)
1076	13	acetylornithine aminotransferase, *E. coli*; 2,2-dialkylglycine decarboxylase, *Pseudomonas*; adenosylmethionine-8-amino-7-oxononanate aminotransferase; 4-aminobutyrate amino transferase; glutamate-1-semialdehyde (2,1)-aminomutase; ornithine aminotransferase, eukaryotic
1533	1	valine-pyruvate aminotransferase, *E. coli*
1717	1	adenosylmethionine-8-amino-7-oxononanate aminotransferase, *Salmonella*
1719	1	succinyldiaminopimelate aminotransferase, *Salmonella* (fragment)
1728	12	7-8-amino-7-oxononanate synthase; 5-aminolevulinic acid synthase; 2-amino-3-ketobutyrate CoA ligase
2515	6	cystathionine γ-lyase; methionine γ-lyase; cystathionine β-lyase, cystathionine γ-synthase
4018[b]	46	tryptophan synthase α and β; threonine synthase; L-serine dehydratase; threonine dehydratase; *O*-acetylserine sulfhydrylase

4074	3	D-alanine aminotransferase; branched-chain amino acid aminotransferase
4088	1	succinyldiaminopimelate aminotransferase, *E. coli* (fragment)
4131	13	aromatic amino acid decarboxylase; glutamate decarboxylase; histidine decarboxylase; α-methyl-DOPA hypersensitive protein; *Drosophila*
4137	11	diaminopimelate decarboxylase; arginine decarboxylase; ornithine decarboxylase, eukaryotic
4369	7	soluble hydrogenase, *Anabaena* and *Synechococcus*; NIFS protein *Anabaena*, *Synechococcus*, and *Azobacter*; serine-pyruvate aminotransferase, rat and human
5087	3	phosphoserine aminotransferase
6280	1	glycine dehydrogenase (decarboxylating), chicken (fragment)
6608	4	serine hydroxymethyltransferase
7741	3	5-aminolevulinic acid synthase
7747	1	5-aminolevulinic acid synthase, *Rhizobium* (fragment)
13296	11	glycogen phosphorylase, maltodextrin phosphorylase
16252	1	*S*-carboxymethylcysteine synthase, *E. coli* (fragment)
16330	1	D-serine dehydratase, *E. coli* (fragment)
16331	1	L-serine dehydratase, *E. coli* (fragment)
17353	1	threonine dehydratase, *Salmonella* (fragment)
17504	1	tryptophan synthase, *E. coli* (fragment)
17800	1	tryptophan synthase, *Klebsiella* (fragment)

SwissProt Version 19 identifies 153 enzymes identified as containing pyridoxal phosphate; SwissProt 22 identifies 179. The connected components generated at PAM <200, cost >125, and match length >80 place the 153 entries in SwissProt Version 19 in 33 connected components, as indicated below. Some sequences are singletons because they are only recorded as fragments and fail to fulfill the length requirement.

[a] See note d in Table 4.

[b] The two chains of tryptophan synthase fall in separate connected components. They are grouped together here by concatenation.

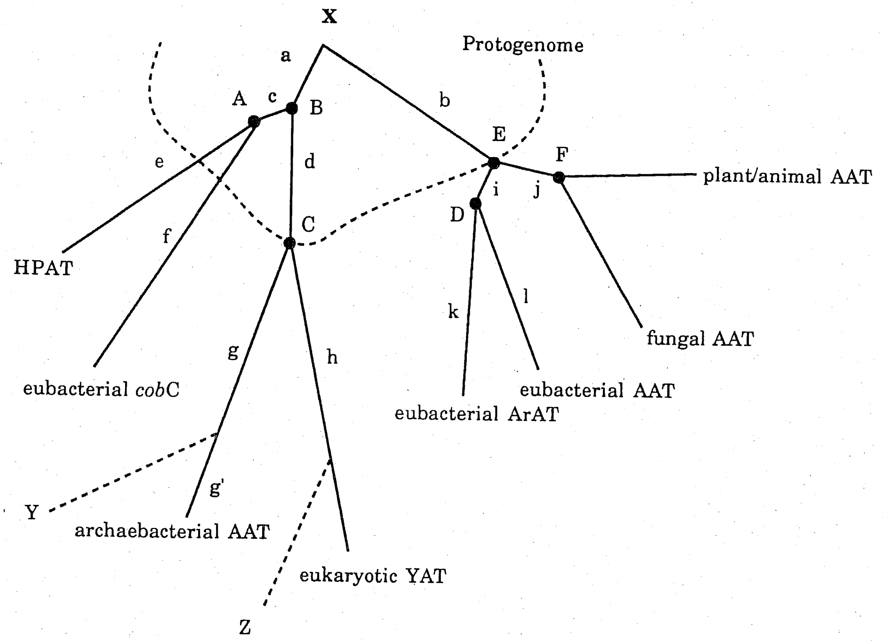


Figure 6 Schematic representation of the evolutionary tree in Fig. 2. Biological information suggests that the most recent common ancestor X lies above the curved dotted line. Functional parsimony indicates that X transaminated aspartate. Y and Z connected to the tree by broken lines indicate possible positions where an as yet not determined archaebacterial tyrosine aminotransferase might be treed. Introduction of an archaebacterial sequence at position Z does not change the most parsimonious interpretation of function on the tree, but it does indicate that the protogenome encoded a YAT. If an archaebacterial YAT sequence is treed at position Y, then the most functionally parsimonious reconstruction of X is as a tyrosine aminotransferase. However, no YAT can be assigned to the protogenome.

just as well by a simple phosphate ester such as a nucleoside phosphate or a phosphoserine residue.

More sequences, especially from archaebacteria, could of course alter the most parsimonious representation of the data presently available. Should this representation hold, however, there are three ways to account for the paradoxical behavior of pyridoxal enzymes. First, it may be intrinsically easy to delete and replace protein enzymes that use pyridoxal as cofactor. In this view, the protogenome encoded a full set of protein-based pyridoxal-dependent enzymes that were involved in multiple deletion-replacement events in the derived kingdoms. Supporting this view are several facts. Mutagenesis studies suggest that altering substrate specificity in pyridoxal enzymes is relatively easy (Cronin and Kirsch

1988). Pyridoxal-dependent decarboxylases could arise readily from a pyridoxal-dependent transaminase (and vice versa) (Smith et al. 1991). In fact, some enzymes (e.g., aspartate-β-decarboxylase; Novogrodsky and Meister 1964) have been shown to decarboxylate some amino acids while transaminating others. Amino acids might readily be obtained in the diet while recruitment is under way.

A second possibility derives from the hypothesis, noted above, that deletion-replacement events are likely to be especially facile if the enzyme being deleted and replaced is a riboenzyme. In this view, placing pyridoxal in the protogenome by a rule of parsimony is correct, but the protogenome-encoded aminotransferases dependent on pyridoxal cofactors were riboenzymes. The large number of deletion-replacement events in the derived lineages thus reflect this fact. In this view, RNA-based transaminases encoded by the protogenome were vestiges of the RNA world and were found in the breakthrough organism.

However, a third possibility should be considered. It is possible that pyridoxal was not in fact present in the most recent common ancestor, but rather was invented later and appeared in the three kingdoms followed by extensive lateral transfer once it was available and its unique chemical properties were appreciated. Just as with random peptides that destroy oxaloacetate (see above), considerations of the chemical reactivity of pyridoxal phosphate support this view. In the absence of any protein catalyst, pyridoxal catalyzes many irreversible reactions (Longenecker et al. 1957; Senda et al. 1977); this implies that sophisticated protein-binding sites (Visser and Kellogg 1978) are needed to control the reactivity of the cofactor. Furthermore, if unconstrained by a binding site, pyridoxal phosphate should lose the 5-phosphate group during its catalytic cycle (Benner 1988). Finally, the existence and evolutionary positions of pyruvoyl-dependent decarboxylases fit this view well, as do available data concerning the biosynthesis of pyridoxal.

We are unable to say at this time which model is correct, although the last seems to us to be the most plausible. If it is true, the suggestion of massive lateral transfer is most interesting, both for future efforts to use simple parsimony analysis for the reconstruction of models for ancient organisms and for evolutionary biology in general.

CONCLUSIONS AND PERSPECTIVES

A rigorous analysis of the divergent evolution of protein sequences and an integrated theory combining structural theory from chemistry and evolutionary theory from biology have provided the tools needed to address several important problems in structural biology. We can now construct high-quality alignments of protein sequences, predict reasonably well the conformation of proteins, design catalytic peptides de novo, and manipulate macromolecular structure in both proteins and nucleic acids. These and other tools sustain construction of preliminary models of the history of life on earth following the breakthrough that produced the first translation systems. The results of these reconstructions alter, often profoundly, our view of how life has evolved. If sequencing, enzymological, and design efforts continue at their present pace, and especially if they include an increased focus on archaebacterial systems, it should be possible at the end of the next decade to obtain definitive models for the protogenome and, perhaps, the breakthrough organism, tracing the molecular history of life back approximately 2.5 billion years before present.

ACKNOWLEDGMENTS

We are indebted to the Swiss National Science Foundation and Sandoz AG for support of the work described here. M.A.C. was supported by a fellowship from the Wellcome Trust. A portion of this work was presented at the meeting on Molecular Biology and the Origin of Life, organized by the Institute for Advanced Study in Biology, Berkeley, California, July 1990.

REFERENCES

Albery, W.J. and J.R. Knowles. 1976. Evolution of enzyme function and the development of catalytic efficiency. *Biochemistry* **15:** 5631–5640.

Bairoch, A. 1992. Prosite: A dictionary of sites and patterns in proteins. *Nucleic Acids Res.* **20:** 2013–2018.

Bairoch, A. and B. Boeckmann. 1992. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* **20:** 2019–2022.

Balch, W.E., L.J. Magrum, G.E. Fox, R.S. Wolfe, and C.R. Woese. 1977. An ancient divergence among the bacteria. *J. Mol. Evol.* **9:** 305–311.

Bazan, J.F. 1990. Structural design and molecular evolution of a cytokine receptor superfamily. *Proc. Natl. Acad. Sci.* **87:** 6934–6938.

Benner, S.A. 1988. Stereoelectronic analysis of enzymatic reactions. *Mol. Struct. Energ.* **9:** 27–74.

———. 1989a. Enzyme kinetics and molecular evolution. *Chem. Rev.* **89:** 789–806.

———. 1989b. Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. *Adv. Enzyme Regul.* **28:** 219–236.

Benner, S.A. and A.D. Ellington. 1987. Return of the "last ribo-organism." *Nature* **332:** 688–689.

———. 1988. Interpreting the behavior of enzymes: Purpose or pedigree? *CRC Crit. Rev. Biochem.* **23:** 369–426.

———. 1990a. "Progenote" or "Protogenote"? *Science* **248:** 943.

———. 1990b. Evolution and structural theory: The frontier between chemistry and biochemistry. *Bioorg. Chem. Front.* **1:** 1–70.

Benner, S.A. and D. Gerloff. 1991. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: The catalytic domain of protein kinases. *Adv. Enzyme Regul.* **31:** 121–181.

Benner, S.A., A.D. Ellington, and A. Tauer. 1989. Modern metabolism as a palimpsest of the RNA world. *Proc. Natl. Acad. Sci.* **86:** 7054–7058.

Benner, S.A., A. Glasfeld, and J.A. Piccirilli. 1990. Stereospecificity in enzymology: Its place in evolution. *Top. Stereochem.* **19:** 127–207.

Benner, S.A., R.K. Allemann, A.D. Ellington, L. Ge, A. Glasfeld, G.F. Leanz, T. Krauch, L.J. MacPherson, S. Moroney, J.A. Piccirilli, and E. Weinhold. 1987. Natural selection, protein engineering, and the last riboorganism: Rational model building in biochemistry. *Cold Spring Harbor Symp. Quant. Biol.* **52:** 53–63.

Bork, P. 1992. Mobile modules and motifs. *Curr. Opin. Struct. Biol.* **2:** 413–421.

Chothia, C. and A. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5:** 823–826.

Crawford, I.P., T. Niermann, and K. Kirschner. 1987. Prediction of secondary structure by evolutionary comparison: Application to the α subunit of tryptophan synthase. *Proteins* **2:** 118–129.

Cronin, C.N. and J.F. Kirsch. 1988. Role of arginine 292 in the substrate specificity of aspartate aminotransferase as examined by site-directed mutagenesis. *Biochemistry* **27:** 4572–4579.

Crouzet, J., L. Cauchois, F. Blanche, L. Debussche, D. Thabaut, M.-C. Rouyez, S. Rigault, J.-F. Mayaux, and B. Cameron. 1990. Nucleotide sequence of *Pseudomonas denitrificans* 5.5 kbase DNA fragment containing five cob genes and identification of structural genes encoding S-adenosyl-L-methionine: Uroprophyinogen III methyltransferase and cobrynic acid a,c-diamide synthase. *J. Bacteriol.* **172:** 5968–5979.

Cubellis, M.V., C. Rozzo, G. Nitti, M.I. Arnone, G. Marino, and G. Sannia. 1989. Cloning and sequencing of the gene coding for aspartate aminotransferase from the thermoacidophilic archaebacterium *Sulfolobus solfataricus*. *Eur. J. Biochem.* **186:** 375–381.

Dassarma S., T. Damerval, J.G. Jones, and N. Tandeau de Marsac. 1987. A plasmid-encoded gas vesicle protein gene in a halophilic archaebacterium. *Mol. Microbiol.* **1:** 365–370.

Dayhoff, M.O., R.M. Schwartz, and B.C. Orcutt. 1978. A model of evolutionary change in proteins. In *Atlas of protein sequence and structure* (ed. M.O. Dayhoff), vol. 5, suppl. 3, p. 345–352. National Biomedical Research Foundation, Washington, D.C.

Dever, T.E., M.J. Glynias, and W.C. Merrick. 1987. GTP-binding domain: Three consensus elements with distinct spacing. *Proc. Natl. Acad. Sci.* **84:** 1814–1818.

de Vos, A.M., M. Ultsch, and K.K. Kossiakoff. 1992. Human growth hormone and extracellular domain of its receptor: Crystal structure of the complex. *Science* **255:** 306–312.

Doolittle, R.F., D.F. Feng, K.L. Anderson, and M.R. Alberro. 1990. A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. *J. Mol. Evol.* **31:** 383–388.

Edwards, F.W. and L.L. Cavalli-Sforza. 1963. The reconstruction of evolution. *Ann. Hum. Genet.* **27:** 104.

Eigen, M. and P. Schuster. 1977. The hypercycle, a principle of natural self-organization. *Naturwissenschaften* **64:** 341–369.

Eschenmoser, A. and E. Loewenthal. 1992. Chemistry of potentially prebiological natural products. *Chem. Soc. Rev.* **21:** 1–16.

Fitch, W. and E. Margoliash. 1967. Construction of phylogenetic trees. *Science* **155:** 279–284.

Friedmann, H.C., R.K. Thauer, S.P. Gough, and C.G. Kannangara. 1987. Δ-Aminolevulinic acid formation in the archae bacterium *Methanobacterium thermoautotrophicum* requires tRNA^{Glu}. *Carlsberg Res. Commun.* **52:** 363–371.

Gilbert, W. 1986. The RNA world. *Nature* **319:** 618.

Gogarten, J.P., H. Kibak, P. Diettrich, L. Taiz, E.J. Bowman, B.J. Bowman, M.F. Manolson, R.J. Poole, T. Date, T. Oshima, J. Konishi, K. Denda, and M. Yoshida. 1989. Evolution of the vacuolar H$^+$-ATPases: Implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci.* **86:** 6661–6665.

Gonnet, G.H., and S.A. Benner. 1991. Computational biochemistry at ETH. Technical report 151. ETH, Zürich.

Gonnet, G.H., M.A. Cohen, and S.A. Benner. 1992. Exhaustive matching of the entire protein sequence database. *Science* **256:** 1443–1445.

Hill, R.E., A. Iwanow, B.G. Sayer, W. Wysocka, and I.D. Spenser. 1987. The regiochemistry and stereochemistry of the biosynthesis of vitamin B6 from triose units. *J. Biol. Chem.* **262:** 7463–7471.

Horne, M. and F. Pfeifer. 1989. Expression of two gas vacuole proteins in *Halobacterium halobium* and other related species. *Mol. Gen. Genet.* **218:** 437–444.

Huang, Z., K.C. Schneider, and S.A. Benner. 1991. Building bocks for analogs of ribo- and deoxyribonucleotides with dimethylene-sulfide, -sulfoxide, and -sulfone groups replacing phosphodiester linkages. *J. Org. Chem.* **56:** 3869–3882.

Hyde, C.C., S.A. Ahmed, E.A. Padlan, E.W. Miles, and D.R. Davies. 1988. Three-dimensional structure of the tryptophan synthase $\alpha_2\beta_2$ multienzyme complex from *Salmonella typhimurium*. *J. Biol. Chem.* **263:** 17857–17871.

Johnsson, K., R.K. Allemann, and S.A. Benner. 1990. Designed enzymes: New peptides that fold in aqueous solution and catalyze reactions. In *Molecular mechanisms in bioorganic processes* (ed. C. Bleasdale and B.T. Golding), pp. 166–187. Royal Society of Chemistry, Cambridge, England.

Jukes, T.H. and R. Holmquist. 1972. Evolutionary clock: Nonconstancy of rate in different species. *Science* **177:** 530–532.

Kannangara, C.G., S.P. Gough, P. Bruyant, J.K. Hoober, A. Kahn, and D. von Wettstein. 1988. tRNA^{Glu} as a cofactor in δ-aminolevulinate biosynthesis. *Trends Biochem. Sci.* **13:** 139–143.

Kimura, M. 1982. Molecular evolution, protein polymorphism, and the neutral theory. Springer-Verlag, Berlin.

King, J.L. and T.H. Jukes. 1969. Non-Darwinian evolution. *Science* **164:** 788–798.

Knighton, D.R., J. Zheng, L.F. Ten Eyck, V.A. Ashford, N.H. Xuong, S.S. Taylor, and J.M. Sowadski. 1991. Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* **253:** 407–414.

Lehmann, W.P. 1973. *Historical linguistics.* Holt, Rinehard and Winston, New York.

Lewin, R. 1988. Molecular clocks turn a quarter century. *Science* **239:** 561–563.

Li, B.F.L., S. Osborne, and M.L. Sinnott. 1983. Catalytic consequences of experimental evolution. Part 2. Rate-limiting degalactosylation in the hydrolysis of aryl beta-D-galactopyranosides by the experimental evolvants ebg(a) and egb(b). *J. Chem. Soc. Perkin Trans.* **II:** 365–369.

Li, J.-M., C.S. Russel, and S.D. Cosloy. 1989. Cloning and structure of HemA gene of *Escherichia coli* K-12. *Gene* **82:** 209–217.

Longenecker, J.B., M. Ikawa, and E.E. Snell. 1957. The cleavage of α-methylserine and α-methylolserine by pyridoxal and metal ions. *J. Biol. Chem.* **226:** 663–666.

Madsen, N.B. and S.G. Withers. 1986. Glycogen phosphorylase. In *Vitamin B6, pyridoxal phosphate*, Part B (ed. D. Dolphin et al.), pp. 355–389. Wiley, New York.

Maizels, N. and A.M. Weiner. 1987. The last riboorganism was no breakthrough. *Nature* **330:** 616.

Martin, C., B. Cami, P. Yeh, P. Stragier, C. Parsot, and J.-C. Patte. 1988. *Pseudomonas aeruginosa* diaminopimelate decarboxylase: Evolutionary relationship with other amino acid decarboxylases. *Mol. Biol. Evol.* **5:** 549–559.

Mehta, P.K., T.I. Hale, and P. Christen. 1989. Evolutionary relationships among aminotransferases. *Eur. J. Biochem.* **186:** 249–253.

Niermann, T. and K. Kirschner. 1990. Improving the prediction of secondary structure of "TIM-barrel" enzymes. *Protein Eng.* **4:** 137–147.

Noll, K.M. and T.S. Barber. 1988. Vitamin contents of archaebacteria. *J. Bacteriol.* **170:** 4315–4321.

Noller, H.F., V. Hottarth, and L. Zimniak. 1992. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science* **256:** 1416–1419.

Novogrodsky, A. and A. Meister. 1964. Control of aspartate-b-decarboxylase activity by transamination. *J. Biol. Chem.* **239:** 879–888.

Oliver, S.G., Q.J.M. van der Aart, M.L. Agostoni-Carbone, M. Aigle, L. Alberghina, D. Alexandraki, G. Antoine, R. Anwar, J.P.G. Ballesta, P. Benit, G. Berben, E. Bergatino, N. Biteau, P.A. Bolle, M. Bolotin-Fukahara, A. Brown, A.J.P. Brown, J.M. Buhler, C. Carcano, G. Carignani, H. Cederberg, R. Chanet, R. Contreras, M. Crouzet, and B. Daignan-Fornier. 1992. The complete DNA sequence of yeast chromosome III. *Nature* **357:** 38–46.

Orgel, L.E. 1968. Evolution of the genetic apparatus. *J. Mol. Biol.* **38:** 381.

———. 1986. RNA catalysis and the origins of life. *J. Theor. Biol.* **123:** 127–149.

Osterhout, J.J., Jr., T. Handel, G. Na, A. Toumadje, R.C. Long, P.J. Connolly, J.C. Hoch, W.C. Johnson, Jr., D. Live, and W.F. DeGrado. 1992. Characterization of the structural properties of $\alpha_1\beta$, a peptide designed to form a four-helix bundle. *J. Am. Chem. Soc.* **114:** 331–337.

Piccirilli, J.A., T. Krauch, S.E. Moroney, and S.A. Benner. 1990. Extending the genetic alphabet: Enzymatic incorporation of a new base pair into DNA and RNA. *Nature* **343:** 33–37.

Recsei, P.A. and E.E. Snell. 1984. Pyruvoyl enzymes. *Annu. Rev. Biochem.* **53:** 357–387.

Rich, A. 1962. On the problems of evolution and biochemical information transfer. In *Horizons in biochemistry* (ed. M. Kasha and B. Pullman), pp. 103–126. Academic Press, New York.

Schneider, K.C. and S.A. Benner. 1990. Oligonucleotides containing flexible nucleoside analogs. *J. Am. Chem. Soc.* **112:** 453–455.

Senda, S., K. Hirota, T. Asao, and K. Maruhashi. 1977. A model for an intermediate in pyridoxal-catalyzed γ-elimination and γ-replacement reactions of amino acids. *J. Am. Chem. Soc.* **99:** 7356–7359.

Smith, D.M., N.R. Thomas, and D. Gani. 1991. A comparison of pyridoxal 5′-phosphate-dependent decarboxylase and transaminase enzymes at a molecular level. *Experientia* **47:** 1104–1118.

Stackhouse, J., S.R. Presnell, G. McGeehan, K.P. Nambiar, and S.A. Benner. 1990. The ribonuclease from an extinct bovid. *FEBS Lett.* **262:** 104–106.

Sulston, J., Z. Du, K. Thomas, R. Wilson, L. Hillier, R. Staden, N. Halloran, P. Green, J. Thierry-Mieg, L. Qiu, S. Dear, A. Coulson, M. Craxton, R. Durbin, M. Berks, M. Metzstein, T. Hawkins, R. Ainscough, and R. Waterston. 1992. The *C. elegans* genome sequencing project: A beginning. *Nature* **356:** 37–41.

Switzer, C.Y., S.E. Moroney, and S.A. Benner. 1989. Enzymatic incorporation of a new base pair into DNA and RNA. *J. Am. Chem. Soc.* **111:** 8322–8323.

Taylor, W.R. and C.A. Orengo. 1989. Protein structure alignment. *J. Mol. Biol.* **208:** 1–22.

Thorne, J.L., H. Kishino, and J. Felsenstein. 1992. Inching toward reality: An improved likelihood model of sequence evolution. *Mol. Biol. Evol.* **34:** 3–16.

Thornton, J.M., T.P. Flores, D.T. Jones, and M.B. Swindells. 1991. Prediction of progress at last. *Nature* **354:** l05–106.

Valencia, A., M. Kjeldgaard, E.F. Pai, and C. Sander. 1991. GTPase domains of ras p21 oncogene protein and elongation factor Tu: Analysis of three-dimensional structures, sequence families, and functional sites. *Proc. Natl. Acad. Sci.* **88:** 5443–5447.

Van der Woerd, R., C.G. Bakker, and A.W. Schwartz. 1987. Synthesis of P1,P2 dinucleotide pyrophosphates. *Tetrahedron Lett.* **28:** 2763–2766.

Visser, C.M. and R.M. Kellogg. 1978. Biotin. Its place in evolution. *J. Mol. Evol.* **11:** 171–178.

Waterston, R., C. Martin, M. Craxton, C. Huynh, A. Coulson, L. Hillier, R.K. Durbin, P. Green, R. Shownkeen, N. Halloran, T. Hawkins, R. Wilson, M. Berks, Z. Du, K. Thomas, J. Thierry-Mieg, and J. Sulston. 1992. A survey of expressed genes in *Caenorhabditis elegans*. *Nature Genet.* **1:** 114–123.

Weiner, A.M. and N. Maizels. 1987. tRNA-like structures tag the 3′ends of genomic RNA molecules for replication: Implications for the origin of protein synthesis. *Proc. Natl. Acad. Sci.* **84:** 7383–7387.

White III, H.B. 1976. Coenzymes as fossils of an earlier metabolic state. *J. Mol. Evol.* **7:** 101–104.

Woese, C.R. and G.E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci.* **74:** 5088–5090.

Woese, C.R., O. Kandler, and M.L. Wheelis. 1990. Towards a natural system of organisms: Proposal for the domains *Achaea*, *Bacteria*, and *Eucarya*. *Proc. Natl. Acad. Sci.* **87:** 4576–4579.

Yeh, W.K. and L.N. Ornston. 1980. Origins of metabolic diversity: Substitution of homologous sequences into genes for enzymes with different catalytic activities. *Proc. Natl. Acad. Sci.* **77:** 5365–5369.

Zuckerkandl, E. and L. Pauling. 1965. Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins* (ed. V. Bryson and H. J. Vogel), pp. 97–166. Academic Press, New York.