THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 278, No. 48, Issue of November 28, pp. 47960–47970, 2003

Printed in U.S.A.

Evolution of Amino Acid Metabolism Inferred through Cladistic Analysis*

Received for publication, December 20, 2002, and in revised form, August 26, 2003
Published, JBC Papers in Press, August 29, 2003, DOI 10.1074/jbc.M213028200

Chomin Cunchillos‡ and Guillaume Lecointre§¶

From the ‡Institut Charles Darwin International, BP 70, 93230 Romainville, France and §Service de Systématique Moléculaire, Département de Systématique (UMR 7138 CNRS), Muséum National d'Histoire Naturelle, 43 rue Cuvier, 75231 Paris cedex 05 France

Because free amino acids were most probably available in primitive abiotic environments, their metabolism is likely to have provided some of the very first metabolic pathways of life. What were the first enzymatic reactions to emerge? A cladistic analysis of metabolic pathways of the 16 aliphatic amino acids and 2 portions of the Krebs cycle was performed using four criteria of homology. The analysis is not based on sequence comparisons but, rather, on coding similarities in enzyme properties. The properties used are shared specific enzymatic activity, shared enzymatic function without substrate specificity, shared coenzymes, and shared functional family. The tree shows that the earliest pathways to emerge are not portions of the Krebs cycle but metabolisms of aspartate, asparagine, glutamate, and glutamine. The views of Horowitz (Horowitz, N. H. (1945) *Proc. Natl. Acad. Sci. U. S. A.* 31, 153–157) and Cordón (Cordón, F. (1990) *Tratado Evolucionista de Biologia*, Aguilar, Madrid, Spain), according to which the upstream reactions in the catabolic pathways and the downstream reactions in the anabolic pathways are the earliest in evolution, are globally corroborated; however, with some exceptions. These are due to later opportunistic connections of pathways (actually already suggested by these authors). Earliest enzymatic functions are mostly catabolic; they were deaminations, transaminations, and decarboxylations. From the consensus tree we extracted four time spans for amino acid metabolism development. For some amino acids catabolism and biosynthesis occurred at the same time (Asp, Glu, Lys, Leu, Ala, Val, Ile, Pro, Arg). For others ultimate reactions that use amino acids as a substrate or as a product are distinct in time, with catabolism preceding anabolism for Asn, Gln, and Cys and anabolism preceding catabolism for Ser, Met, and Thr. Cladistic analysis of the structure of biochemical pathways makes hypotheses in biochemical evolution explicit and parsimonious.

Cellular metabolism is a complex process involving about a thousand chemical reactions catalyzed by globular proteins, enzymes. As any other biological phenomenon, metabolism has a structure that is the product of an evolutionary history. How can we infer that history? Because the history and interrelationships of living organisms is based on comparative anatomy, the history of metabolism must be reconstructed by the comparative analysis of the structure of its components (1–4). Biochemists recognized this necessity long ago but have never used comparative methods that formally controlled the consistency of biochemical evolutionary theories (3–8). In systematics, the science of classification of living things, there are now standardized comparative methods to infer events of the past with a measurement of the consistency of a theory.

Summarizing the Standard Comparative Method

The standard comparative method starts with a number of things to compare. These are known as "operational taxonomic units," "terminals," often "species," or "individuals." These words are not strictly equivalent. By convenience we often call them "taxons." Among those different things we detect that some structures are under different versions. To formalize that observation we create a column in a table called a "matrix" in which a given version is associated with 0, and the other version is associated with 1 in the list of things to compare (versions can be also "coded" 2, 3, etc. depending on the number found). The column is a character; its version as 0 or 1 is the character state. We bet that the versions are homologous; sameness must come from common ancestry. That bet is called "primary homology." Once the matrix is filled with all the characters we were able to detect for that collection of things, a phylogenetic tree is retained according to a criterion explained below. Onto that tree changes of character states become hypothetical events of transformation, and the tree yields a relative order of these events. That tree will show for each character whether that bet is won or lost. If a state appears only once in the tree, the bet is won, and it is called a secondary homology, or a synapomorphy; the state is present from common ancestry in those things that have it. If a state of a character is associated with more than one event, the bet is lost. It is homoplasy, *i.e.* sameness without common ancestry. But how do we choose a tree over others? For a given number of taxons there are a limited but high number of possible trees, which are different theories of interrelationships. In science theories are to be compared in terms of internal consistency. The theory that is the most consistent is the one that appeals the less ad hoc hypotheses. The most parsimonious tree is the best theory (against alternative trees) because it requires the smallest number of ad hoc hypotheses of transformation onto its branches. This is the reason why one always associates to the most parsimonious tree the number of transformations ("number of steps") and measurements of internal consistency, like the consistency index (C.I.)[1] and the retention index (R.I.). We propose the use of such a parsimony analysis (9–13) to infer the

* This work was supported by The Musée National d'Histoire Naturelle, Saint-Gobain, and La Maison des Sciences de l'Homme (to C. C.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

¶ To whom correspondence should be addressed. Tel.: 33-1-40-79-37-51; Fax: 33-1-40-79-37-71: E-mail: lecointr@mnhn.fr.

[1] The abbreviations used are: C.I., consistency index; R.I., retention index; KC, Krebs cycle; PLP, pyridoxal phosphate.

relative timing of emergence of a number of metabolic pathways, *i.e.* aliphatic amino acid metabolism and portions of the Krebs cycle, in order to shed light on the earliest pathways and enzymatic functions among them.

Metabolic Pathways and Evolution

In 1945 Horowitz (14) postulated that the earliest biosynthetic pathways evolved in a backward direction if life began in a rich soup of organic molecules. If primitive cells were using a particular external nutrient, soon this organic molecule would be depleted in the environment. A selective advantage could be obtained by organisms able to synthesize this nutrient from an available precursor. Each biosynthetic step was selected according to successive depletions of precursors in the environment. The first enzyme to appear in the biosynthetic pathway was, therefore, the most distal (*i.e.* downstream) in the pathway. Confluence of pathways was selected because it saved energy. This energy is used for other needs that will be more difficult to satisfy for competitor cells without confluence. This optimization of pathways is considered as a general basic rule of comparative biochemistry (1). For these early anabolisms, common enzymes or common reactions shared by two (or more) synthetic pathways are distal and are, therefore, evidence for common ancestry for these pathways. In terms of kinship pathways sharing these enzymes are judged to be closer to each other than to other pathways not using these enzymes.

In 1990, Cordón (2) proposed a symmetrical scenario of catabolic pathways. Early forms of life extracted energy from the degradation of substrates available in the environment into a product. Selective advantage was obtained for those able to produce a supplementary reaction of deeper degradation of this product, therefore obtaining more energy from the original substrate. Confluence is selected by obtaining the transformation of another substrate into an intermediate product already present in the protocell. The first reactions to appear in evolution of catabolism are proximal ones (*i.e.* upstream ones). The common distal elongation of two branched catabolic pathways is, therefore, a phenomenon whose final result will be evidence for common ancestry of these pathways. Two catabolic pathways sharing one or several distal portions of catabolism are supposed to be more closely related to each other than to other pathways. But there is a risk, which consists of the late connection of an "opportunistic" catabolic pathway once the early catabolism onto which it connects is already complete. In that case the common downstream portion is not evidence for common ancestry but just convergence obtained by recruitment, a phenomenon recognized for having played a role in biochemical evolution (5, 15, 16). It is sameness without common ancestry. When we reconstruct the past, homoplasy appears when there are character conflicts due to such similarities obtained by evolutionary convergence or reversion. The risk of homoplasy in our data just depends on the relative timing between events of distal elongation of a pathway and the connecting event of another pathway. An early connecting event followed by distal elongation will provide good phylogenetic indicators. A late connecting event (late in time and/or late in the pathway) will probably bring homoplasy. As in any other study of systematics where putative homologies are coded into a matrix, there are risks of homoplasy to carry on. We make the bet that this homoplasy will not swamp the phylogenetic signal.

According to Cordón (2) the symmetrical scenario is not rigid and can change from one type of metabolites to another; the order of development of a given pathway depends on the position and availability of the initial substrate and/or final product. The above timings in the genesis of reactions in anabolic pathways and catabolic pathways are valid because the final product and the initial substrate, respectively, are imposed from the outside to the cell, at least initially. Alternative scenarios can be obtained for transformations starting from products already integrated into the cellular metabolism. New biosynthetic pathways can develop in a forward direction by the addition of new enzymes and reactions to pre-existing pathways. For example, the urea cycle uses the biosynthesis of arginine (17). Thus, Cordón (2) proposed a forward development for amino acid catabolism, fatty acids anabolism, and glycogenesis and a backward development for amino acid anabolism, fatty acids catabolism, and glycolysis (confirmed in 1993 by Fothergill-Gilmore and Michels (18)). However, late opportunistic connections of pathways can complicate the general ordering of metabolism development as predicted by Cordón (2). For example, some catabolic pathways can develop backwards by opportunistic connecting to intermediate compounds from already developed pathways (*e.g.* Cordón's catabolism of serine). For the same reasons some anabolic pathways can develop forward (*e.g.* Cordón's anabolism of methionine and threonine). The complexity of the matter is increased by the fact that some biochemical pathways can be considered both catabolic and anabolic (pentose phosphate shunt, glycolysis, gluconeogenesis, the Krebs Cycle). Therefore, it is clear for all biochemists today that there is no direct link between being either anabolic or catabolic and having grown either forward or backward. This is one of the reasons why it is important to start the present methodological improvement with simple examples, like amino acids, for which in most cases the catabolic pathway is different from the anabolic pathway.

The evolutionary development of pathways has to do with the darwinian concept of descent with modification, justifying the use of cladistic analysis (19) for metabolic pathways. By using that concept we bet that present similarities in pathways detected through shared enzymes and enzymatic reactions can be interpreted as the result of pathway transformations through time. Comparison of pathways can, therefore, be followed by phylogenetic reconstruction of metabolic pathways.

The Interest of Amino Acid Metabolism

Why focus on amino acids and not on nucleic acids, fatty acids, or monosaccharides? The reasons are manyfold. Amino acids are among the earliest abiotic sources of energy and molecules for protocells without excluding other possible sources (for instance, sunlight for energy, pentoses). Consequently, their metabolism must have been involved among the earliest biochemical reactions. Simple amino acids have been obtained experimentally in an abiotic way for example by Miller in 1953 (20), who obtained glycine, alanine, glutamate, and aspartate from simple molecules like ammoniac, hydrogen, methane, and water. Serine can be obtained abiotically in the presence of formaldehyde (20). Even if the chemical environment of the primitive soup is today interpreted in a different manner than it was at the time of Miller, free amino acids are still found in abiotic and extraterrestrial environments, for example, in meteorites, like the famous carbonaceous chondrites, which are probably the source on primitive earth. Moreover, among classical candidates for early energy providers (fatty acids, monosaccharides, amino acids), amino acids are the only chemical precursors whose structure is complex enough to contain all the atoms and reactive groups necessary for most of the reactions necessary for central metabolism. Other compounds like monosaccharides, polysaccharides, and fatty acids have a poorer variety of atoms and groups and are rather monotonous. However, the strongest argument for early availability of amino acids to protocells might be the fact that any metabolism defined as a coordinated network of enzymatic

activities performed by proteins needs in its very early steps amino acid anabolism and catabolism for these proteins. Amino acid metabolism might, therefore, have preceded any other metabolism, even the most central metabolism like the Krebs cycle. However, this question is still debated.

The Krebs Cycle

Whatever the metabolic specialization found in diverse living organisms (heterotrophy, photosynthetic autotrophy, chemosynthetic autotrophy, various forms of respiration and fermentation), there is a universal core of about 50 metabolic pathways involving the anabolism and catabolism of amino acids, fatty acids, saccharides (the glycolysis and the glycogenesis, the pentose phosphate pathway), and the Krebs cycle. Because the Krebs cycle is the point of confluence of all other metabolic pathways, it is sometimes viewed as primitive. That view is, however, challenged by several lines of evidence. Molecules entering the Krebs cycle (oxo acids and acyl-CoAs) are intermediate metabolites that are diversely interpreted with regard to their availability in primitive abiotic environments. Some biochemists consider them as most likely the products of a peripheral cellular metabolism. For example, the origin of the Krebs cycle was thought to be secondary and composite by Schoffeniels (1, 21), Gest (22, 23), and Meléndez-Hevia *et al.* (7). Others consider them as most likely available right from the beginning; for instance, carbon dioxide, water, and sunlight provide oxalic acid through pyrite catalysis. Then, the primitive soup would have been rich in dicarboxylic acids. These debates led us to incorporate two portions of the Krebs cycle as taxons in the matrix along with the metabolic pathways of each of the 16 aliphatic amino acids in our data sets. By analyzing anabolism and catabolism all together following the protocol of Cunchillos and Lecointre (24, 25) for catabolism only, we compared portions of the Krebs cycle both to amino acid catabolism and anabolism. The ultimate question was to find the earliest pathways and enzymatic functions among peripheral amino acid metabolisms and portions of the Krebs cycle.

Taxons

This work is a cladistic analysis of the structure of a part of the cellular metabolism. This is allowed by taking each pathway as a taxon and shared enzymes, shared enzymatic functions, and shared cofactors as characters. Each taxon starts from the initial substrate and includes the pathway until its entry into the Krebs cycle or starts from the metabolites of the Krebs cycle and includes the pathway to the amino acid. For example, the taxon dASN1 is the catabolic pathway from asparagine to oxaloacetate, and the enzymes and functions along this pathway will be the characters of this taxon (Fig. 1, characters 2, 4, 11, 13, and 28). The taxon sASN1 is the anabolic pathway from oxaloacetate to asparagine (Fig. 1, characters 2, 4, 11, 28, 56, 106, and 107). In this example the pathways used for the degradation ("d") and the synthesis ("s") of the amino acid are very similar; however, this is not the case for all taxons. Some amino acids can be degraded or synthesized through several possible ways (cysteine, aspartate, asparagine, glutamate, glutamine, threonine). In these cases each way is taken separately as an additional taxon (for instance, dCYS1, dCYS2, dCYS3 for the degradation (d) of cysteine (Cys) after the pathways 1, 2, and 3, respectively). The Krebs cycle (KC) is considered in two portions, each beginning with an oxo acid, which is a point of entrance into the cycle and also a point of output. These two oxo acids are also, among the metabolites of the Krebs cycle, the closest to amino acids structurally speaking. The portion KC1 begins with oxaloacetate and ends with α-oxoglutarate; the portion KC2 begins with α-oxoglutarate and ends with oxaloacetate. Aromatic amino acids have not been considered at this stage of cellular evolution because their complex metabolism needs too much oxygen and is only possible once the metabolism of aliphatic amino acids is set.

Our aim is first to understand phylogenetic interrelationships of aliphatic amino acid catabolic and anabolic pathways, second to discover in the most parsimonious tree the earliest metabolic pathway, and third to discover the first enzymatic functions associated to them. This should answer the following connected questions; Is the Krebs cycle the first? Does amino acid catabolism appear before amino acid anabolism? Can we test the expected order of metabolism development as predicted by Horowitz (14) and Cordón (2)? In other words, does anabolism develop backwards and catabolism forwards?

Comparing differences in metabolisms of extant living organisms is of no help in reaching this aim because all the corresponding events are more differences in metabolism regulation than differences in structure of pathways (1) and, anyway, might have been posterior to the very early events we intend to infer. Comparing semantids of Zuckerkandl and Pauling (26), *i.e.* DNA sequences or protein sequences, is of no help because it leads to severe problems of linear sequence homology and would infer mutational changes that also are posterior to the occurrence of these early metabolisms. Problems with sequence data are even more complex. The present taxons are pathways. Each 0 and 1 coded into a cell of the matrix could have been replaced by the sequence of the corresponding enzyme in a given model organism. Then, comparing sequences would have mixed up different gene duplication histories, leading to no trees at all. Moreover, we have no theoretical model that links ordering events of gene duplication within a putative primitive genome with ordering the rise and organization of enzymatic activities in a protocellular metabolism. At last, the present work infers nothing about information storage, replication, and the RNA world.

Characters; Homology Criteria

If shared enzymes or similar enzymes are evidence for common ancestry of metabolic pathways, similarities in the structure of active sites would be sufficient to formulate putative homologies. However, only a few active sites (27) are known in detail compared with the large number of known enzyme types (4). We are, therefore, led to consider similarities in catalytic reactions and enzymatic mechanisms as reflecting similarities in active sites. The higher the specificity, the more accurate is this reflection. In the same way, considering the generally accepted idea that enzymes evolved from low specificities to high specificities (5, 15, 28–31), putative common ancestry of pathways can be postulated not only on the basis of shared enzymes with high specificities but also on the basis of very similar reactions. Similarity in function must correspond to an underlying similarity in structure of active sites, a structural similarity that comes from common ancestry. Recognizing that two metabolic pathways share the same reaction with high specificity for substrate uses a strict criterion of primary homology (32), whereas recognizing a common family of reaction relaxes this criterion with a risk of homoplasy obtained by convergence or recruitment. There is no reason for considering that this risk is higher in the present case than in aligned DNA sequences or in classical morpho-anatomical matrices, both usual in systematics (33). The criteria of primary homology is 4-fold: shared specific enzymatic activity (I), shared enzymatic function without shared specificity for substrate (IIa), shared coenzymes (IIb), and shared family of function (IIc).

HYPANC	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sASPone	010?0?000?1000000000000000?10000000000000000000000000000000000000000000000000000000000000000000000000000000000
sASPtwo	101?0?000?0100000000000000?10000000000000000000000000000000000000000000000000000000000000000000000000000000000
sASNone	01010?000?1000000000000000?10000000000000000000000000000100000000000000000000000000000000000000000000000011000
sASNtwo	10110?000?0100000000000000?10000000000000000000000000000100000000000000000000000000000000000000000000000011000
sASNthree	01010?000?1000000000000000?10000000000000000000000000000010000000000000000000000000000000000000000000000011000
sASNfour	10110?000?0100000000000000?10000000000000000000000000000010000000000000000000000000000000000000000000000011000
sGLUone	010?0?000?0001000000000000?10000000000000000000000000000000000000000000000000000000000000000000000000000000000
sGLUtwo	101?0?000?0100000000000000?10000000000000000000000000000000000000000000000000000000000000000000000000000000000
sGLNone	01010?000?0001000000000000?10000000000000000000000000000001000000000000000000000000000000000000000000000011000
sGLNtwo	10110?000?0100000000000000?10000000000000000000000000000001000000000000000000000000000000000000000000000011000
sARGone	111?1?000?0001000000000000?10000000000000000000000000000000011111110000000000000000000000000000000000000001100
sARGtwo	101?1?000?0100000000000000?10000000000000000000000000000000011111110000000000000000000000000000000000000001100
sARGthree	111?1?000?0001000000000000?10000100000000000000000000000000000111100000000000000000000000000000000000000001000
sARGfour	101?1?000?0100000000000000?10000100000000000000000000000000000111100000000000000000000000000000000000000001000
sPROone	010?1?000?0001000000000000?10000010000000000000000000000000000001100000000000000000000000000000000000000001000
sPROtwo	101?1?000?0100000000000000?10000010000000000000000000000000000001100000000000000000000000000000000000000001000
sALAone	101?0?0011000000000000000111000000000000000000000000000000000000000010000000000000000000000000000000000000000000
sSERone	101?0?0011000000000000000011000000000000000000000000000000000000000001111100000000000000000000000000000000001000
sSERtwo	101?0?0011000000000000000011000000000000000000000000000000000000000001110001110000000000000000000000000000001000
sGLYone	101?0?0011000000100000000011000000000000000000000000000000000000000001111100000000000000000000000000000000001000
sGLYtwo	101?0?0011000000100000000011000000000000000000000000000000000000000001110001110000000000000000000000000000001000
sCYSone	101?0?0011000000000000000011000000000000000000000000000000000000000001111100000110000000000000000000000000001100
sCYStwo	101?0?0011000000000000000011000000000000000000000000000000000000000001110001110110000000000000000000000000001100
sMETone	110?1?000?1000000000000000?10000000001000000000000000000000000000000000000000000011111100000000000000000001100
sMETtwo	101?1?000?0100000000000000?10000000001000000000000000000000000000000000000000000011111100000000000000000001100
sTHRone	110?1?000?1000000000000000?10000000000000000000000000000000000000000000000000000011100010000000000000000001000
sTHRtwo	101?1?000?0100000000000000?10000000000000000000000000000000000000000000000000000011100010000000000000000001000
sILEone	111?1?100?1000010001000000?10000000000000000000000000000000000000000000000000000011100011110000000000000001000
sILEtwo	101?1?100?0100010001000000?10000000000000000000000000000000000000000000000000000011100011110000000000000001000
sLEU	101?0?0011000000000100000011000000000000000000000000000000000000000010000000000000000000001111110000000000000100
sVAL	101?0?0011000000000100000011000000000000000000000000000000000000000010000000000000000000001110000000000000000000
sLYSone	111?1?0011100000000000000011000000000000000000000000000000000000000000000000000000011000000000000111111100001100
sLYStwo	101?1?0011010000000000000011000000000000000000000000000000000000000000000000000000011000000000000111111100001100
sLYSthree	111?1?0011000000000000110011000000000000000000010000000000000000000000000000000000000000000000000000000011101101
dASPone	010??00?001000000000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000?00
dASPtwo	101??00?000100000000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000?00
dASNone	0101?00?001010000000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000?00
dASNtwo	1011?00?000110000000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000?00
dGLU	010??00?000001000000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000?00
dGLNone	0101?00?000001000000000000110000000000000000000000000000000000000000000000000000000000000000000000000000000?00
dGLNtwo	1011?00?000000000000000010110000000000000000000000000000000000000000000000000000000000000000000000000000000?00
dARG	111?100?000001000000000000010001100000000000000000000000000000000000000100000000000000000000000000000000000?00
dPRO	010?100?000001000000000000010000010000000000000000000000000000000000000100000000000000000000000000000000000?00
dALA	101??10?000000100000000000111000000000000000000000000000000000000000000000000000000000000000000000000000000?00
dSER	100??11?000000110000000000110000000000000000000000000000000000000000000000000000000000000000000000000000000?00
dGLY	100??11?000000111000000000110000000000000000000000000000000000000000000000000000000000000000000000000000000?00
dCYSone	100??11?000000100100000000110000000000000000000000000000000000000000000000000000000000000000000000000000000?00
dCYStwo	101??10?000000100000000000110000001100000000000000000000000000000000000000000000000000000000000000000000000?00
dCYSthree	101??10?000000100000000000111000000011000000000000000000000000000000000000000000000000000000000000000000000?00
dMET	100??1110000000001100000001100000000000110000000000000000000000000000000000000000000000001000000000000000001?10
dTHRone	100??1110000000100100000001100000000000000000000000000000000000000000000000000000000000000000000000000000001?10
dTHRtwo	100??11?000000111000000000110000000000000000000000000000000000000000000000000000000000000000000000000000000?00
dILE	101??1011000000000111000001100000000000001000000000000000000000000000000000000000000000000000000000000000001?00
dLEU	101??1011000000000011100001100000000000000111100000000000000000000000000000000000000000000000000000000000001100
dVAL	101??1011100000000111000001100000000000000000010000000000000000000000000000000000000000000000000000000000001?00
dLYS	111?110?110000000000011110110000000000000000000100000000000000000000000000000000000000000000000000000000000?00
KCone	?????0?0110000000000000000001?00000000000000000001110000000000000000000000000000000000000000000000000000?0100
KCtwo	?????1?00000000000000000001?000000000000000000000011111000000000000000000000000000000000000000000000000?0000

FIG. 1. **Matrix containing 59 taxons (*rows*) and 110 characters (*columns*).** Each taxon (*rows*) is a pathway of amino acid degradation (*d*) or a pathway of amino acid synthesis (*s*). *0* and *1* refer to the character state found in the corresponding taxon. Characters (*columns*) are numbered following their order from *left* to *right*, the character number *one* is the *first column*; the character number *26* is the 26th column. All these characters are named (Fig. 2) with the corresponding number of international nomenclature (*Enzyme Nomenclature*) for the sake of precision and homology types defined in the text.

Homology Type I

Several pathways share the same enzyme with high specificity for its substrate. The enzyme itself is the hypothesis of primary homology. For reversible enzymes this is valid either for a given reaction or the reverse reaction. The absence is coded 0, and presence is coded 1. For instance, the catabolism of aspartate and asparagine both use the aspartate aminotransferase, which transforms aspartate into oxaloacetate. The anabolism of these two amino acids use the same enzyme for the reverse reaction, which transforms oxaloacetate into aspartate. The character is called "aspartate aminotransferase" (character 12); it is coded 1 for taxons dASP2, dASN2, sASP2, and sASN2 and 0 for dGLU, sGLU1, dALA, and sALA, for example. This type of homology is used in characters 11–26, 29–105, 109, and 110.

Homology Type II

IIa; Shared Enzymatic Functions—Several pathways utilize the same enzymatic functions, *i.e.* exhibit the same kind of chemical transformation, without considering the specificity of

each enzyme for its substrate. The underlying hypothesis is that similarity in enzymatic function must correspond to similarity in the structure of active sites, with the hypothesis that enzymes must have evolved from generalist active sites to specialized ones (28–31). When the substrate is present, but the function not performed, the character state is coded 0. When the substrate is present and the function is performed, the character state is coded 1. When the required substrate is not available, the character state is coded with a question mark (?). For example, catabolism of alanine and aspartate perform transamination using exactly the same enzymatic mechanisms using pyridoxal phosphate by two similar enzymes that differ in their specificities for their respective substrates, alanine aminotransferase and aspartate aminotransferase. The reverse anabolic reactions use the same enzymes with their specificities for their respective oxo acids. The character state is coded 1 in dASP2, dALA, sASP2, sALA, and other pathways where aminotransferases occur, 0 in dSER, dGLY, sASP1, sASN1 for example (where transaminations occur with a different mechanism that uses NAD), and ? for portions of the Krebs cycle where transaminations are impossible. This type of homology is used in characters 2–10, 106, and 108.

IIb; Shared Cofactors—Shared cofactors reflect similarity in enzymatic mechanisms, which in this case does have the same functional meaning. If a common cofactor is used without similarity in enzymatic mechanisms, it is considered that the use of this cofactor has been gained independently, and each enzymatic mechanism is, thus, coded as homologies of type IIa. For example, this criterion does apply to the pyridoxal phosphate (PLP), in which the functional meaning is deamination/amination. The character state is coded 0 when the deamination or the amination is not performed although possible or performed using another cofactor, 1, when direct deamination or amination or a transamination uses PLP and ? for portions of the Krebs cycle where neither deamination/amination nor transamination is possible.

This criterion of homology is restricted in its application. First, when the use of a cofactor is too specific to an enzyme, coding such a cofactor into a character would lead to coding twice the character of homology type I associated to the enzyme. For instance, the use of biotin is restricted to the propionyl carboxylase. In such a case the character biotin will not be taken into account; otherwise, it would give a weight of 2 to the character propionyl carboxylase. Second, when a cofactor is specific to an enzymatic function, there is also a risk of overweighting a character of homology type IIa. For example, thiamin is specific to α-decarboxylations. Third, coding a ubiquitous cofactor brings risks of homoplasy. It is necessary in this case to consider the kind of reaction; that is, to come back to homology type IIa. For instance, NAD is used in a wide range of enzymatic functions; they are NAD-deamination, NAD-aldehyde acid dehydrogenation, NAD-β-oxidation, NAD-α-decarboxylation. Recoding each of these functions is useless; they are already coded as homologies IIa. Consequently, it appears that this criterion of homology is only useful for the PLP (character 1).

IIc; Shared Functional Family—In principle, this criterion of homology is the same as in IIa, just relaxed. It is an extension of the above idea that enzymes must have evolved from generalists to specialists. The character state is coded 0 when the reaction is not performed although it is possible, 1 when the reaction is performed, and ? when the reaction cannot be performed considering the chemical groups present in the metabolites of the pathway considered (? for deamination/amination in KC1 and KC2). This criterion actually concerns three main families of reactions, decarboxylation (character 27), deamination/amination (character 28), and phosphorylation (character 107).

How is the presence of reverse reactions in the same matrix to be managed? Because the biosynthesis of a given amino acid can follow a different path than its degradation, some enzymes are used for the catabolism; however, they have no meaning for the anabolism of the same amino acid. In that case the catabolic enzyme is coded with a question mark in the anabolic pathway. In the same way question marks are present for some anabolic enzymes in catabolic pathways where they have no meaning. The original matrix contains 59 taxons and 110 characters (Fig. 1).

EXPERIMENTAL PROCEDURES

Ordered Characters

In some cases several pathways share a couple of successive enzymes. Instead of coding these enzymes as separated type I homologies, their proximities have been incorporated into the matrix using ordered characters in Cunchillos and Lecointre (24, 25), where amino acid catabolism only was studied. The principle of coding successive enzymes into an ordered character was not followed here for two reasons. First, it leads to useless very complex coding situations when both degradation and biosynthetic pathways are present in the matrix and when the successive enzymes are reversible (*i.e.* as parts of the anabolism and catabolism as well). Second, a large number of successive enzymes with lateral branched pathways lead to some arbitrariness in defining the set of successive enzymes to code as ordered, *i.e.* unjustified delineation of such sets. These difficulties burden the matrix with a heavy set of doubtful assumptions. If coding successive enzymes into an ordered character was simple when only amino acid catabolism was studied, it must be abandoned in the present case to leave only shared individual enzymes as witnesses of the past. Consequently, the few characters involved in the catabolism treated as ordered in Cunchillos and Lecointre (24, 25) have been treated unordered here.

Tree Search

For the matrix of catabolic pathways (25 taxons and 60 characters), the low number of taxons allowed the exact search. The most parsimonious trees were obtained through the "branch and bound" search of PAUP4 (34). The number of trees is recorded after having condensed trees using the "amb-" command, *i.e.* collapsing branches of minimal length of zero. Strict consensus trees are shown. For larger matrices (biosynthesis, 37 taxons and 86 characters, and the whole data, 59 taxons and 110 characters) heuristic searches were conducted with NONA (35) as implemented into WINCLADA (36) using TBR branch swapping. For a better exploration of trees the Parsimony Ratchet (Hopper Islands, Nixon (37)) was used. The proportion of data to be re-weighed was set between 25 and 50%, and the number of iterations was progressively increased from 50,000 to 200,000 (option ambpoly =). This increase in iterations allowed a check that the number of supplementary MP trees gained each time decreased or was null. Each time, the number of trees was recorded after having collapsed all unsupported nodes in all trees ("hard collapse").

Rooting

The tree was rooted using an all-zero hypothetical ancestor (HYPANC), justified by the fact that, in the coding of character states, zero was given for the absence of enzymes, for the absence of performance of particular functions (even in presence of a putative suitable substrate), or for the absence of utilization of a cofactor. We have to keep in mind that such a rooting option will automatically put the simplest pathways closer to the root. However, this does not make any assumption on the nature of the corresponding enzymatic reactions. Production of explicit hypotheses remains the most important thing. If tomorrow another way of rooting is found, the character coding and trees produced by the present study will still be useful; they will have to be considered in the light of the new root. In other words, the topology (the way branches are connected to each other) could still be the same but just rooted differently.

Defining Time Spans

From the root to the tip of branches phylogenetic trees provide a relative order of transformations (here enzymatic innovations) through time. We define time spans in metabolism as the time along the tree

separating two character changes referring to homologies of type II, taking into account the following criteria:

The Order of Nodes—In absence of other criteria, two sister-nodes are of the same relative period if the period is limited downstream by the following period.

The Nature of Enzymatic Changes—If a branch is followed by downstream branches that do not bear changes in homologies of type II, the downstream branches are of the same period. For example (see Fig. 4) although posterior (topologically) to the red period of the branches I and J, the terminal branch leading to dGLN2 is in *red* because the pathway can be completely achieved without any new enzymatic functions, *i.e.* just with the already available general functions of transamination using PLP (node J) and amide deamination (node H). Indeed, characters occurring on the terminal branch of dGLN2 (characters 30 and 31) are only gains of specificity or particular substrates and, therefore, only refer to homologies of type I. It must be noted that the present criterion has the power to change the time span of the sister branch if supported by a new homology of type II. For example (Fig. 4), sister-nodes K and L are not of the same period because the node L is supported by the rise of a new enzymatic activity (homology of type II) completely absent from all upstream branches.

No Homoplasy—When a character exhibits homoplasy, classically one has the choice in localizing changes in the tree (convergences or reversions). Only transformations with unambiguous localization (*i.e.* in which the place on the tree does not depend on that choice) will be taken into account in the use of the second criterion.

RESULTS

The Succession of Enzymatic Innovations—Separate analyses of catabolism and anabolism were conducted to test the predictions of Horowitz (14) and Cordón (2), according to which the earliest catabolic enzymes in the course of evolution are the first in the catabolic pathway, and the earliest anabolic enzymes in evolution are the more downstream in the anabolic pathway. From the catabolic data matrix, the branch and bound search of PAUP4 (34) yielded 12 equi-parsimonious trees of 74 steps (C.I. = 0.81, R.I. = 0.84; strict consensus in Fig. 3). It is interesting to notice that character changes occurring on deep nodes A, B, D, are all uniquely derived characters (*i.e.* without homoplasy). The temporal succession of these character changes provides the succession of occurrence of enzymatic catabolic activities. First enzymatic catabolic activities must have been deamination, aldehyde dehydrogenation (branch A), and then decarboxylation (branch D). Within deamination, deamination of amides (A) must have preceded deamination using either pyridoxal phosphate or NAD (B). This has to be related to the fact that amide deamination does not use any cofactor. The earliest pathways to emerge are dASP1, dASN1, dGLU, dARG, dPRO, and dGLN1. This group of pathways corresponds to the first period of catabolic development as predicted by Cordón (2). Then decarboxylation appears, which is the second period according to Cordón (2). Although it is not shown by the tree, one can suppose that β-decarboxylation might have occurred early in this period because it does not use any cofactor (it occurs in the common branch of KC1 and dLYS, with a convergent occurrence in dVAL) (25). The two blocks of the Krebs cycle appear to be the product of early amino acid catabolism and not the reverse. Finally, it appears that deamination and then decarboxylation is the earliest enzyme to appear in the course of catabolism evolution and also the earliest in catabolic pathways, as predicted by Horowitz (14) and Cordón (2). From the anabolic data matrix, the progressive increase of iterations from 50,000 to 200,000 in the heuristic search using the Parsimony Ratchet of WINCLADA led to 10 equi-parsimonious trees (135 steps, C.I. = 0.63, R.I. = 0.75). The strict consensus of these trees is shown in Fig. 3. It is interesting to note that the tree fails to identify the earliest biosyntheses (there is a rake, also called polytomy). Under the assumption that catabolism preceded anabolism, basal polytomy can be interpreted as a rapid diversification of amination and decarboxylation once the corresponding enzymes were available through catabolism. Indeed, α-decarboxylation present in KC2 already occurs in one of the branches of the polytomy, and β-decarboxylation appears at the node G. The node E exhibits amination-NAD, amide amination, and homologies of type IIc like general decarboxylation and general amination. Amination is the last step of most amino acid biosyntheses; that these enzymes are first to emerge in the tree confirm Horowitz's hypothesis. The node F is supported by two homologies of type II, the use of PLP and the transamination using PLP, and one homology of type I, aspartate aminotransferase. The node G is supported by the same previous homologies of type II (redundant on that branch because of the polytomy) and by two other homologies of type II, β-degradation and β-decarboxylation.

From the global data matrix the progressive increase of iterations from 50,000 to 200,000 in the heuristic search using the Parsimony Ratchet of WINCLADA led to 107 equi-parsimonious trees (182 steps, C.I. = 0.59, R.I. = 0.78). The strict consensus is shown Fig. 4. The reactions first to emerge (node H) are general deamination/amination (character 28) and amide deamination/amination (character 4). The group I is defined by amination/deamination using NAD (character 2). The group J shows the use of PLP (character 1) and transamination using PLP (character 3). The group K is defined by the use of aspartate aminotransferase (homology of type I, character 12). The group L is defined by general decarboxylation (homology of type IIc, character 27) and α-decarboxylation (character 6). The group M is supported by several type I homologies: phosphoenolpyruvate carboxykinase (character 70), enolase (71), phosphoglycerate phosphomutase (72), and homoplastic use of phosphorylation (character 107) and β-degradation (character 9). The group N is supported by two homologies of type I, aspartate kinase (82) and aspartate semialdehyde dehydrogenase (83), and the homoplastic use of homologies of type II, phosphorylation (character 107) and aldehyde dehydrogenase (character 5, in the latter case with a remarkable convergence in synthesis and degradation of arginine, proline, and lysine). Compared to the results of Cunchillos and Lecointre (25), it is remarkable that the introduction of biosynthetic pathways did not change the general order of the rise of reactions and enzymes. First, general deamination/amination and amide deamination/amination appear, then NAD-deamination/amination appears in the group I, whereas PLP transamination appears in its sister-group J. In a posterior period, decarboxylation becomes possible. Note that the group D of Cunchillos and Lecointre (25) is not recovered here, placing general decarboxylations and α-decarboxylations at the same node L. However, some more resolution is gained later on. Finally, the Krebs cycle (KC1 and KC2) is not basal in the tree but only a product of amino acid metabolism. It is not nested within a sub-group of biosyntheses or degradations; it is only branched into a polytomy. Compared to the development of catabolic enzymes and functions described in Cunchillos and Lecointre (25), two differences emerge; PLP deamination does not appear on the branch J but separately occurs on branches leading to the synthesis of isoleucine and degradation of serine, cysteine (dCYS1), methionine, threonine, and glycine. Aldehyde dehydrogenation does not occur on branch H but on two separate branches, leading to synthesis and degradation of arginine and proline on a branch leading to synthesis and degradation of lysine and on the branch N.

For some amino acids degradation is exactly the reverse synthetic pathway, so that the degradation pathway is the sister group of the anabolic pathway. This is the case of proline (sPRO1 is the sister group of dPRO) and lysine (sLYS3 and dLYS). For some others, the degradation and the synthesis are

[1] Pyridoxalphosphate enzymes; IIb.
[2] Deamination (NAD); IIa.
[3] Transamination (PLP); IIa.
[4] Amide deamination; IIa.
[5] Aldehyde dehydrogenation (NAD); IIa.
[6] α-Decarboxylation; IIa.
[7] Deamination (PLP); IIa.
[8] Carboxylation (biotin); IIa.
[9] β-Degradation; IIa.
[10] β-Decarboxylation; IIa.
[11] *Amino-acid dehydrogenase*: *1*.4.1.5; I.
[12] *Aspartate aminotransferase*: *2*.6.1.1; I.
[13] *Asparaginase*: *3*.5.1.1; I.
[14] *Glutamate dehydrogenase*: *1*.4.1.3; I.
[15] *Pyruvate dehydrogenase*: *1*.2.4.1; I.
[16] *Serine deaminase*: *4*.2.1.16; I.
[17] *Serine hydroxymethyltransferase*: *2*.1.2.1; I.
[18] *Cysteine desulfydrase*: *4*.4.1.1; I.
[19] *Propionyl-CoA carboxylase*: *4*.1.1.41; I.
[20] *Branched-chain-amino-acid aminotransferase*: 2.6.1.42; I.
[21] *Oxoisovalerate dehydrogenase*: *1*.2.1.25; I.
[22] *Acetyl-CoA acetyltransferase*: *2*.3.1.9; I.
[23] *Aminoadipate-semialdehyde dehydrogenase*: 1.2.1.31; I.
[24] *Aminoadipate aminotransferase*: *2*.6.1.39; I.
[25] *Ketoacid dehydrogenase*: *4*.1.1.1; I.
[26] *Alanine aminotransferase*: 2.6.1.2; I.
[27] Decarboxylation; IIc.
[28] Deamination; IIc.
[29] *Glutaminase*: *3*.5.1.2; I.
[30] *Glutamin-oxo-acid aminotransferase*: *2*.6.1.15; I.
[31] *ω-Amidase*: *3*.5.1.3; I.
[32] *Arginase*: *3*.5.3.1; I.
[33] *Ornithine-oxto-acid aminotransferase*: *2*.6.1.13; I.
[34] *Pyrroline-carboxylate reductase*: *1*.5.1.1; I.
[35] *Cysteine aminotransferase*: *2*.6.1.3; I.
[36] *Mercaptopyruvate sulphurtransferase*: *2*.8.1.2; I.
[37] *Cysteine dioxygenase*: *1*.13.11.20; I.
[38] *Aspartate decarboxylase*: *4*.1.1.12; I.
[39] *Methionine methyltransferase*; I.
[40] *Cystathionine synthase*: *4*.2.1.22; I.
[41] *α-Methyl-butyryl-CoA β-oxidase*; I.
[42] *Oxoisovaleryl-CoA dehydrogenase*; I.
[43] *Acyl-CoA carboxylase*; I.
[44] *Hydroxymethylglutaryl-CoA lyase*: *4*.1.3.4; I.
[45] *Ketoacid CoA-transferase*: *2*.8.3.5; I.
[46] *Isobutryl-CoA β-oxidase* ; I.
[47] *Lysine aminotransferase*: *2*.6.1.36; I.
[48] *Citrate synthase*: *4*.1.3.7; I.
[49] *Citrate dehydratase*; I.
[50] *Isocitrate dehydrogenase*: *1*.1.1.41; I.
[51] *α-Oxoglutarate dehydrogenase*: *1*.2.4.2; I.
[52] *Succinyl-CoA hydrolase*: *3*.1.2.3; I.
[53] *Succinate dehydrogenase*; I.
[54] *Fumarate hydratase*: *4*.2.1.2; I.
[55] *Malate dehydrogenase*; I.
[56] *Asparagine synthase (ADP-forming)*: *6*.3.1.4; I.
[57] *Asparagine synthase (glutamine-hydrolyzing)*: 6.3.5.4; I.
[58] *Glutamine synthetase*: *6*.3.1.2; I.
[59] *Amino-acid acetyltransferase*: *2*.3.1.36; I.
[60] *Acetylglutamate kinase*: *2*.7.2.8; I.
[61] *N-acetyl-glutamyl-P reductase*: *1*.2.1.38; I.
[62] *Acetylornithine aminotransferase*: *2*.6.1.11; I.
[63] *Acetylornithine deacetylase*: *3*.5.1.16; I.
[64] *Ornithine carbamoyltransferase*;2.1.3.3; I.
[65] *Argininosuccinate synthetase*: *6*.3.4.5; I.
[66] *Argininosuccinate lyase*: *4*.3.2.1; I.
[67] *Glutamate kinase*; I.
[68] *Glutamate semialdehyde dehydrogenase*; I.
[69] *Oxaloacetate decarboxylase*: *4*.1.1.3; I.
[70] *Phosphoenolpyruvate carboxykinase (GTP)*: *4*.1.1.32; I.
[71] *Enolase*: *4*.2.1.11; I.
[72] *Phosphoglycerate phosphomutase*; I.
[73] *Glycerate kinase*: *2*.7.1.31; I.
[74] *Hydroxypyruvate reductase*: *1*.1.1.81; I.
[75] *Serine-pyruvate aminotransferase*: *2*.6.1.51; I.
[76] *Phosphoglycerate dehydrogenase*: *1*.1.1.95; I.
[77] *Phosphoserine aminotransferase*: *2*.6.1.52; I.
[78] *Phosphoserine phosphatase*: *3*.1.3.3; I.
[79] *Glycine synthase*: *2*.1.2.10; I.
[80] *Serine acetyltransferase*: *2*.3.1.30; I.
[81] *Cysteine synthase*: *4*.2.99.8; I.
[82] *Aspartate kinase*: *2*.7.2.4; I.
[83] *Aspartate-semialdehyde dehydrogenase*: *1*.2.1.11; I.
[84] *Homoserine dehydrogenase*: *1*.1.1.3; I.
[85] *Homoserine succinyltransferase*; I.
[86] *Succinyl homoserine lyase*: *4*.2.99.9; I.
[87] *Cystathionine lyase*: *4*.4.1.8; I.
[88] *Homoserine kinase*: *2*.7.1.39; I.
[89] *Threonine synthase*: *4*.2.99.2; I.
[90] *Acetolactate synthase*: *4*.1.3.18; I.
[91] *Ketol-acid reductoisomerase*;1.1.1.86. I.
[92] *Dihydroxyacid dehydrase*: *4*.2.1.9. I.
[93] *Isopropylmalate synthase*: *4*.1.3.12; I.
[94] *Isopropylmalate isomerase*; I.
[95] *Isopropylmalate dehydrogenase*: *1*.1.1.85; I.
[96] *Dihydrodipicolinate synthase*: *4*.2.1.52 ; I.
[97] *Dihydrodipicolinate reductase*; I.
[98] *Piperidine-dicarboxilate succinyltransferase* ; I.
[99] *Succinyl-diaminopimelate aminotransferase*: *2*.6.1.17; I.
[100] *Succinyl-diaminopimelate desuccinylase*: *3*.5.1.18; I.
[101] *Diaminopimelate epimerase*: *5*.1.1.7; I.
[102] *Diaminopimelate decarboxylase*: *4*.1.1.20; I.
[103] *Homocitrate synthase*: *4*.1.3.21; I.
[104] *Homoaconitate hydratase*: *4*.2.1.36; I.
[105] *Homoisocitrate dehydrogenase* ; I.
[106] Acid-ammonia lyases; IIa.
[107] Phosphorylation; IIc.
[108] Acyl-CoA synthases; IIa.
[109] *α-Oxobutyrate dehydrogenase*; I.
[110] *Oxoglutarate decarboxylase*; I.

FIG. 2. **List of characters of the matrix (Fig. 1).** Each number and name associated corresponds to a column in the matrix, with the corresponding number of international nomenclature (Enzyme Nomenclature) and homology types defined in the text. For example, the 26th character is the alanine aminotransferase; it corresponds to the 26th column in Fig. 1, the international nomenclature is 2.6.1.2, and in the text it is a primary homology of type I.

nested into a group of limited extent, showing closely related catabolism and anabolism as very similar reverse pathways (for example, all pathways within the group I). Other amino acids show a catabolism radically different from the anabolism, and in that case we must consider epochs as defined under "Experimental Procedures."

Epochs of Metabolic Development—The general cladogram (Fig. 4) provides a temporal succession of enzymatic innovations. In such a rooted tree, nodes correspond to time spans in which enzymatic innovations occurred. These time spans or "epochs" have been defined according to three criteria (see "Experimental Procedures") and are in color on the corresponding branches (Fig. 4) as well as on the corresponding reactions or sets of reactions in Fig. 5. The cladogram does not order events within an epoch but orders events between epochs. The first one is the red epoch of the rise of general deamination/amination and transamination (nodes H, I, and J). In the red epoch, cysteine transforms into mercaptopyruvate, alanine into pyruvate, asparagine into aspartate (dASN1, dASN2), aspartate into oxaloacetate (dASP1), glutamine into glutamate (dGLN1, dGLN2), and glutamate into oxoglutarate (dGLU). Valine, leucine, and isoleucine can be transformed into their respective oxo acids, and lysine can be transformed into oxoadipate. Syntheses being the reverse reactions of previously described degradations can be achieved, like sASP1 and sGLU1. These syntheses do not use phosphorylation. They are simpler syntheses, with phosphorylation part of a later step. Innovations occurring on red nodes generally preceded innovations occurring during the next epoch, on pink nodes, but not always. Indeed, if we follow the principle by which general functions occurred before specificities for a substrate, transamination using PLP was possible (character 3 on node J) before the rise of aspartate aminotransferase (character 12 on node K). Therefore, the complete syntheses and degradations of aspartate (second pathway, sASP2 and dASP2) and glutamate (second pathway, sGLU2) could be achieved before the rise of the specificity of the aspartate aminotransferase for its substrate (the same for dASN2). Therefore, the corresponding pathways are members of the red period, although the corresponding branches are distal to the branch pink branch K. The branch O really corresponds to a new period, because a com-

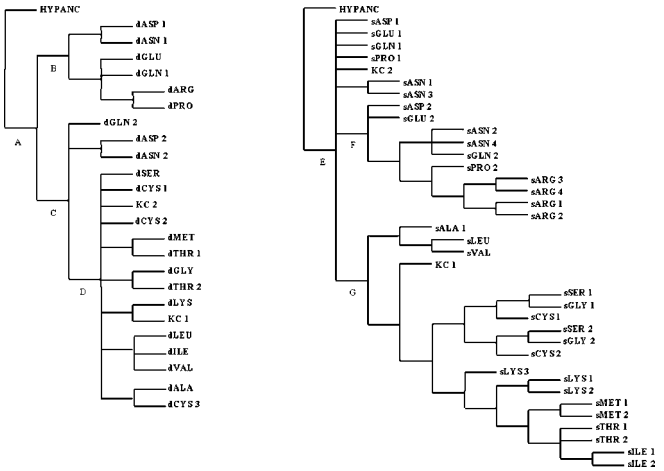

FIG. 3. *Left*, strict consensus of 12 equi-parsimonious trees obtained through the branch and bound search of PAUP4 calculated on a matrix made of catabolic pathways only (25 taxons, 60 characters). Each tree is of 74 steps (C.I. = 0.81, R.I. = 0.84). *Right*, strict consensus of 10 equi-parsimonious trees obtained through the Parsimony Ratchet of WINCLADA calculated on a matrix made of anabolic pathways only (37 taxons, 86 characters). Each tree is of 135 steps (C.I. = 0.63, R.I. = 0.75).

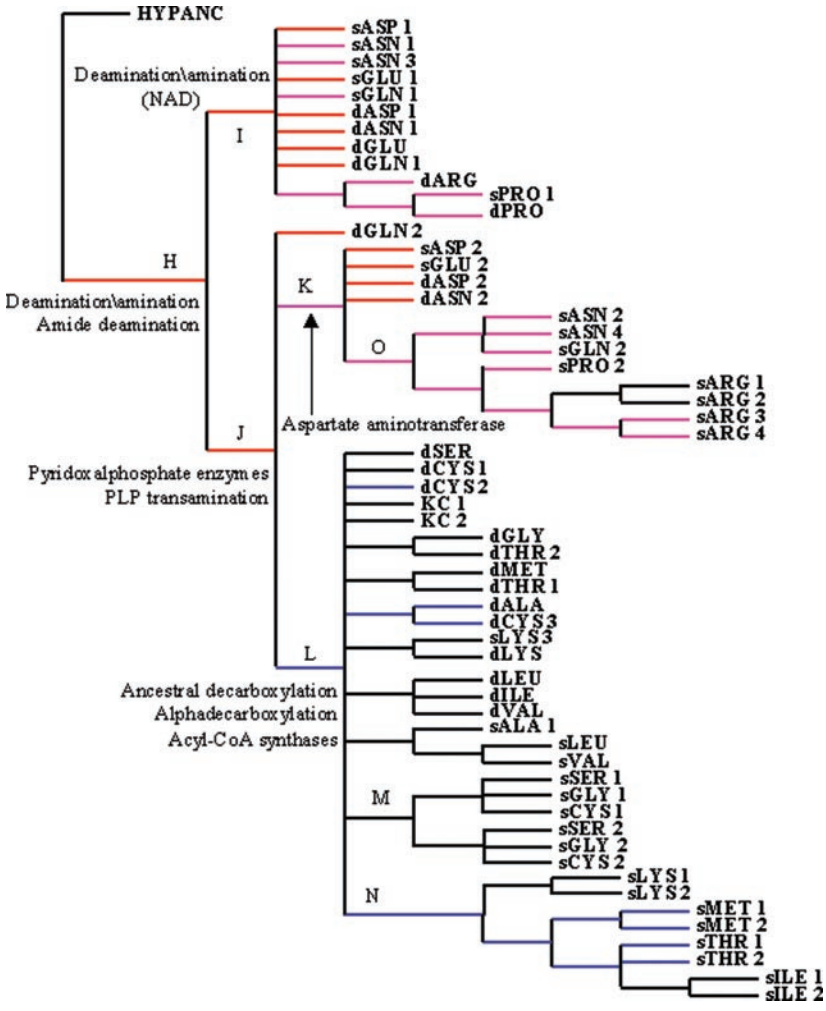

FIG. 4. **Strict consensus of 107 equi-parsimonious trees obtained through the Parsimony Ratchet of WINCLADA calculated on a matrix made of anabolic and catabolic pathways (59 taxons, 110 characters).** Each tree is of 182 steps (C.I. = 0.59, R.I. = 0.78). Some groups are named with *letters* to allow comments in the text; the name is put on the corresponding internal branch. For clarity only a few character changes (*i.e.* enzymatic innovations) have been indicated onto the tree at the node where they were located by the parsimony criterion. Characters that change only once in the tree are indicated.

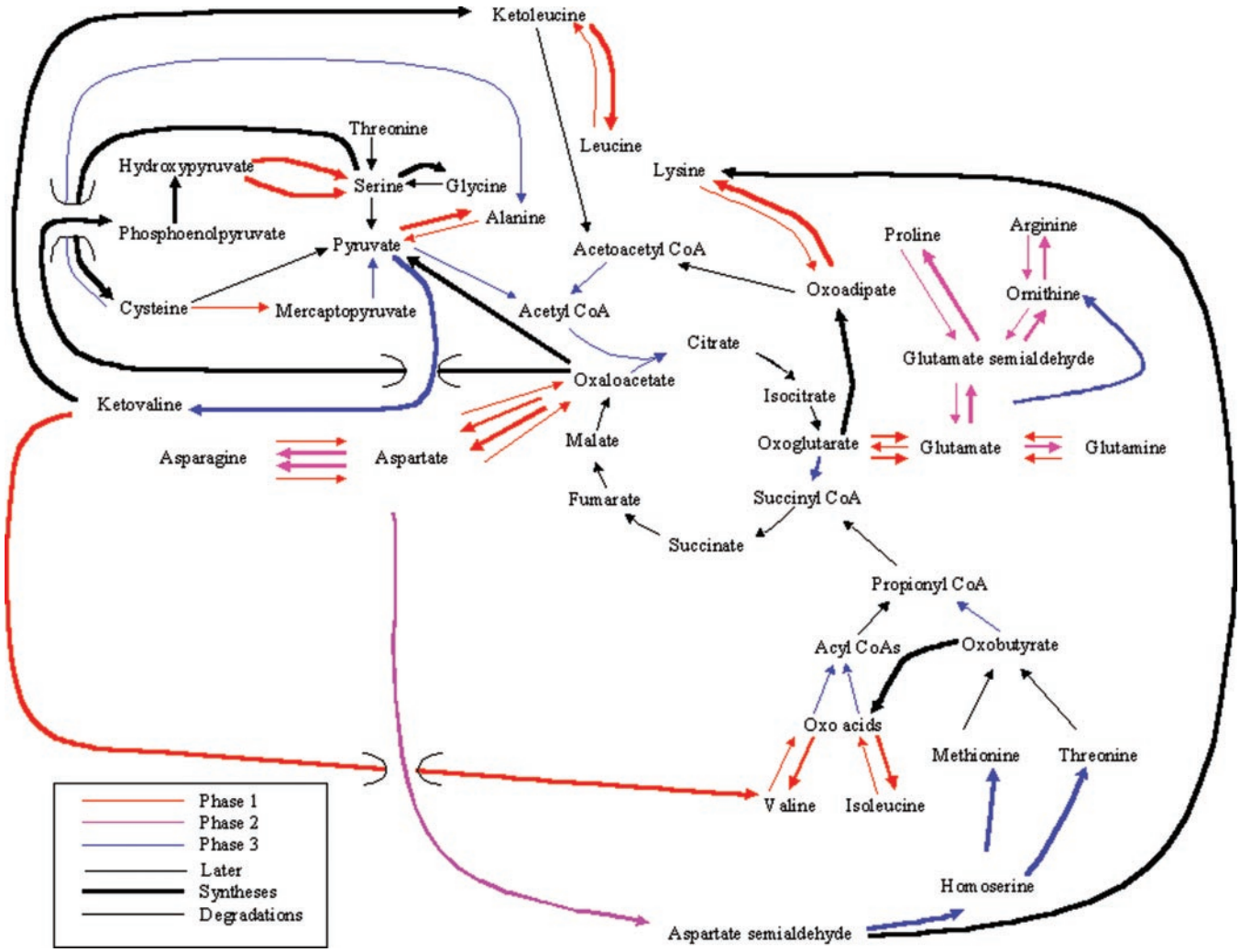

FIG. 5. **General view of amino acid metabolic pathways and their connecting points to the Krebs cycle.** Colors are successive time spans (or epochs) as inferred from the tree: *red* first, then *pink*, *deep blue*, *black*. Reactions in *black* do not mean that all these reactions appeared at the same time. These are just reactions posterior to the blue epoch.

pletely different kind of reaction becomes available. In this red epoch, some other amino acids can be implicated in reversible reactions; their very early anabolism is possible at the same time as their very early catabolism without being connected to the Krebs cycle right from the beginning. These amino acids are lysine, leucine, alanine, valine, and isoleucine.

During the next epoch (in pink), new enzymatic activities occur (character 5, aldehyde dehydrogenases using NAD; character 107, phosphorylation), making possible the complete achievement of metabolism of proline and arginine (except sARG1 and sARG2), asparagine, and glutamine, given some supplementary specificities for particular substrates, however, of the same functional family. This epoch closes the complete metabolisms of amino acids belonging to the groups I and II identified by Cordón (2) (aspartate and asparagine for group I; glutamate, glutamine, arginine, and proline for group II). At this step all amino acids can be deaminated, leaving available the corresponding carbonic skeletons. The meaning for catabolism is that the upstream reactions are the earliest in the tree, whereas for anabolism the downstream reactions of amination are also the earliest.

Then arrives the next epoch, in blue, when decarboxylation became possible, and acyl-CoA synthases became available (Fig. 4). Some pathways can be achieved without new families of enzymatic activities (*i.e.* without new type II homologies) just by gaining specificities for particular substrates within families of reactions already available. These pathways are catabolism of alanine (dALA), cysteine (two pathways, dCYS2 and dCYS3), and anabolism of arginine (sARG1 and sARG2) and methionine (sMET1 and sMET2). The Krebs cycle develops from oxo acids available once the metabolism of amino acids of groups I and II is set. The development of the Krebs cycle begins during the blue epoch.

The next phase (in black) are later steps involving new families of enzymatic activities (new homologies of type II), among which, however, the strict consensus tree cannot further distinguish time spans because of the polytomy. These further activities are deamination using PLP, β-degradation, β-decarboxylation, and carboxylation using biotin.

DISCUSSION

Events Difficult to Order—β-Decarboxylation is simpler than α-decarboxylation because it does not require any cofactor and can easily be obtained spontaneously in acid environments. β-Decarboxylation might have appeared first, *i.e.* before α-decarboxylation, for instance in the decarboxylation of oxaloacetate into pyruvate. Indeed, only one substrate suitable for a β-decarboxylation is present at the end of the red epoch, oxaloacetate. Oxaloacetate then provides either pyruvate or phosphoenolpyruvate, the first biosynthetic step of the Cordón group III of amino acids (Ala, Ser, Cys, Gly) that could arise before α-decarboxylation. Later on, among the pathways con-

sidered here, all the metabolites produced at the end of the pink epoch of deamination are α-oxo acids. Among new decarboxylation, only α-decarboxylation was, thus, possible in the blue step, and other β-decarboxylation was possible just after, as soon as possible.

Ordering Catabolism and Anabolism—Does amino acid catabolism appear before amino acid anabolism? It depends on the amino acid (Fig. 5). For some amino acids, it seems that the earliest reactions of anabolism occurred at the same time of the earliest catabolic reactions. In the red epoch, this is true for asparagine and glutamate (both directly connected to the Krebs Cycle) and lysine, leucine, alanine, valine, and isoleucine (completely connected to the Krebs cycle later on). In the pink epoch, these amino acids are proline and arginine. For some others (asparagine, glutamine, and cysteine) the first upstream reaction of catabolism occur in the red epoch before the first downstream reaction of anabolism becomes possible. Some amino acids have their first downstream anabolic reaction before the first upstream catabolic one; serine is synthesized in the red epoch before its degradation becomes possible, and methionine and threonine are synthesized in the blue epoch before their degradation become possible.

Corroborating Cordón's Scenario—Cordón (2) distinguished four groups of amino acids based on the structure of catabolic pathways (3). The first phase of evolution according to Cordón is the period of deamination and transamination (*red* and *pink*, Fig. 5), leading to the development and the complete achievement of the pathways of amino acids of groups I (dASP1, dASN1, dASP2, dASN2) and II (dGLU, dGLN1, dARG, dPRO, dGLN2). These two groups appear as the most basal; however, their components are not exclusively related to each other (*i.e.* I and II are paraphyletic). It is also interesting to confirm Cordón's views on the direction of catabolic pathways evolution. The first reactions of amino acid degradation in the pathway (deamination, transamination) are the first in the course of evolution; deamination and transamination occur at the deepest nodes of our tree. We corroborate the views of Horowitz (14) and Cordón (2), *i.e.* biosynthetic pathways develop backwards (for example, sALA, sSER, sVAL) and catabolic pathways develop forwards (dALA, dCYS2, dVAL, dILE). However, as confluence is selected, opportunistic connecting events allow some pathways the benefit from available substrates, leading to exceptions in the temporal order of the rise of reactions already mentioned by Cordón (2). Indeed, once Cordón's groups I and II are developed, a set of reactions of a pathway develops in the previous expected directions within each other group (III: cysteine, alanine, serine, glycine; IV: threonine, methionine, isoleucine, valine). Then this set becomes the core on which other pathways will develop opportunistic connections, leading to pathways with more complicated temporal developments. For example, within Cordón's group II, the catabolism of proline and arginine do not follow a forward development; the upstream reactions are in pink, and the most downstream reaction is in red. In the same way syntheses of arginine and proline do not follow a backward development. The development of pathways of proline and arginine are opportunistic, connecting on the already available glutamate pathways. In the same way sGLN does not follow a backward development by connecting the glutamate already available. Within the group I, it is the same for sASN connecting the available aspartate. Within Cordón's group III such situations are also found in dCYS1, dCYS3, dGLY, and dSER (which connect to the already available dALA) and sCYS1, sCYS2, sCYS3, sGLY1, and sGLY2 (which connect to the already available sSER). Within Cordón group IV, such situations are found in dMET, dTHR1 (connected to dVAL and dILE), dTHR2 (connected to dALA), sMET, sTHR, sILE, sLEU, dLEU, dLYS1, and dLYS2. A number of these pathways have complex temporal developments, as they do not exhibit any correlation between the position of the reactions in the pathway and the temporal order. Nevertheless, it is striking that such late opportunistic connections can lead to a perfectly opposite temporal development, *i.e.* a biosynthetic pathway with a perfect forward development as in sTHR and sMET. For the sake of simplicity, the complications in temporal development of sLYS are not shown in Fig. 5.

Obviously, opportunistic connections of pathways lead to complicate Horowitz and Cordón's directions of metabolic development; some amino acid catabolic pathways do not develop in a forward direction, and some anabolic pathways do not develop in a backward direction. Going into details into Cordón's scenario, it is striking that this author did include these complex events. His views are compatible with our findings. One must keep in mind that none of Cordón's ideas have been taken into account in the coding of character states. One could object that the obtained result depends on our rooting option. Indeed, the simplest pathways are basal because of the all-zero hypothetical ancestor. But this makes no assumption on the nature of the simplest pathways. The fact that the simplest pathways contain only deamination (and not, for instance, decarboxylation) is not a matter of arbitrary rooting but just a fact written in the structure of these catabolic pathways. Anyway, objectivity is not a matter of arbitrariness but a matter of justification of choices and transparency of the procedures.

Methodological Interest—Pathways can easily evolve by recruitment of enzymes or portions of pathways so that one could think that depicting interrelationships of pathways under the form of a tree might be unsuitable. Such processes should typically provoke homoplasy in trees by opportunist late connections of pathways, at least detectable by a strong decrease of the C.I. and the R.I. However, values obtained for C.I. and R.I. are rather high, not compatible with the idea of complete reticulate mode of evolution. In the same way one could have been discouraged in using trees because of the possibility of gaining the same enzymatic functions from different structures or because similar enzymatic mechanisms could have arisen a number of times. Nevertheless, this risk is the burden of any comparative approach, and our homoplasy measurements indicate that these homoplastic events have not been dominant. By proposing a new kind of taxon, the anabolic pathway, and using the enzymes, enzymatic functions, cofactors, and families of enzymatic functions as characters, it is now possible to propose a phylogeny of metabolic pathways that has the virtues of being explicit (through cladistic coding and standard procedures of systematics), parsimonious, and based neither on sequence data (26, 38) nor distance approaches (38). This phylogeny independently confirms the views of Cordón (2) who did not use a cladistic analysis. This work illustrates the new methodological framework within which other metabolic pathways now can be analyzed; hypotheses in biochemical evolution become parsimonious, explicit, testable, and based on patterns only.

Acknowledgments—We thank Marc Silberstein and Patrick Tort for help.

REFERENCES

1. Schoffeniels, E. (1984) *Biochimie Comparée*, Masson, Paris
2. Cordón, F. (1990) *Tratado Evolucionista de Biologia*, Aguilar, Madrid, Spain
3. Cunchillos, C. (1997) in *Pour Darwin* (Tort, P., ed) pp. 425–447, Presses Universitaires de France, Paris
4. Michal, G. (1999) *Biochemical Pathways*, John Wiley & Sons, Inc., New York
5. Jensen, R. A. (1976) *Annu. Rev. Microbiol.* **30,** 409–425
6. Cavalier-Smith, T. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52,** 805–824
7. Meléndez-Hevia, E., Waddell, T. G., and Cascante, M. (1996) *J. Mol. Evol.* **43,** 293–303

8. Martin, W., and Müller, M. (1998) *Nature* **392,** 37–41
9. Hennig, W. (1950) *Grundzüge einer Theorie der Phylogenetischen Systematik*, Deutscher Zentralverlag, Berlin
10. Hennig, W. (1966) *Phylogenetic Systematics*, University of Illinois Press, Urbana and Chicago, IL
11. Farris, J. S. (1983) in *Advances in Cladistics* (Platnick, N., and Funk V. A., eds) Vol. 2, pp. 1–36, Columbia University Press, New York
12. Forey, P. L., Humphries, C. J., Kitching, I. L., Scotland, R. W., Siebert, D. J., and Williams, D. M. (1992) *Cladistics: A Practical Course in Systematics*, Clarendon Press, Oxford
13. Darlu, P., and Tassy, P. (1993). *Reconstruction Phylogénétique: Concepts et Méthodes*, Masson, Paris
14. Horowitz, N. H. (1945) *Proc. Natl. Acad. Sci. U. S. A.* **31,** 153–157
15. Petsko, P., Kenyon, G. L., Gerlt, J. A., Ringe, D., and Kozarich, J. W. (1993) *Trends Biochem. Sci.* **18,** 372–376
16. Copley, S. D. (2000) *Trends Biochem. Sci.* **25,** 261–265
17. Takiguchi, M., Matsubasa, T., Amaya, Y., and Mori, M. (1989) *BioEssays* **10,** 163–166
18. Fothergill-Gilmore, L. A., and Michels, P. A. M. (1993) *Prog. Biophys. Mol. Biol.* **59,** 105–235
19. Kluge, A. G. (2001) *Cladistics* **17,** 199–210
20. Miller, S. A. (1953) *Science* **117,** 528
21. Schoffeniels, E. (1981) *Les Cahiers de Biochimie* (*Paris*), Vaillant-Carmanne Maloine, Paris
22. Gest, H. (1981) *FEMS Microbiol. Lett.* **12,** 209–215
23. Gest, H. (1987) *Biochem. Soc. Symp.* **54,** 3–16
24. Cunchillos, C., and Lecointre, G. (2000) in *Biosystema 18* (Barriel, V., and Bourgoin, T. eds.) pp. 87–106, Société Française de Systématique, Paris
25. Cunchillos, C., and Lecointre, G. (2002) *C. R. Biologies* **325,** 119–129
26. Zuckerkandl, E., and Pauling, L. (1965) *J. Theor. Biol.* **8,** 357–366
27. Knowles, J. R. (1991) *Nature* **350,** 121–124
28. Holden, J. T. (1968) *J. Theor. Biol.* **21,** 97–102
29. Jallon, J. M. (1983) in *L'Evolution des Protéines* (Hervé G., ed) pp. 92–103, Masson, Paris
30. Pierard, A. (1983) in *L'Evolution des Protéines* (Hervé G., ed) pp. 53–66, Masson, Paris
31. O'Brien, P. J., and Herschlag, D. (1999) *Chem. Biol.* (*Lond.*) **6,** 91–105
32. de Pinna, M. C. C. (1991) *Cladistics* **7,** 367–394
33. Sanderson, M. J., and Hufford, H. (1996) *Homoplasy*, Academic Press, Inc., New York
34. Swofford, D. L. (1999) *PAUP*. Phylogenetic Analysis Using Parsimony* (**and Other Methods*), Version 4, Sinauer Associates, Sunderland
35. Goloboff, P. (1998) *Nona Computer Program and Software*, published by the author, Tücuman, Argentina
36. Nixon, K. C. (1999) *Winclada* (*BETA*), Version 0.9.9, published by the author, Ithaca, New York
37. Nixon, K. C. (1999b) *Cladistics* **15,** 407–414
38. Forst, C. V., and Schulten, K. (2001) *J. Mol. Evol.* **52,** 471–489