RESEARCH ARTICLES

Integrating the Universal Metabolism into a Phylogenetic Analysis

Chomin Cunchillos and Guillaume Lecointre†*

*Institut Charles Darwin International, Romainville, France; and †UMR7138 Systématique, Adaptation, Evolution, Département Systématique et Evolution, Muséum National d'Histoire Naturelle, Paris, France

The darwinian concept of "descent with modification" applies to metabolic pathways: pathways sharing similarities must have inherited them from an exclusive, hypothetical ancestral pathway. Comparative anatomy of biochemical pathways is performed using five criteria of homology. Primary homologies of "type I" were defined as several pathways sharing the same enzyme with high specificity for its substrate. Primary homologies of "type II" were defined as the sharing of similar enzymatic functions, cofactors, functional family, or recurrence of a set of reactions. Standard cladistic analysis is used to infer the evolutionary history of metabolic development and the relative ordering of biochemical reactions through time, from a single matrix integrating the whole basic universal metabolism. The cladogram shows that the earliest pathways to emerge are metabolism of amino acids of groups I and II (Asp, Asn, Glu, and Gln). The earliest enzymatic functions are mostly linked to amino acid catabolism: deamination, transamination, and decarboxylation. For some amino acids, catabolism and biosynthesis occur at the same time (Asp, Glu, Lys, and Met). Catabolism precedes anabolism for Asn, Gln, Arg, Trp, His, Tyr, and Phe, and anabolism precedes catabolism for Pro, Ala, Leu, Val, Ile, Cys, Gly, Ser, and Thr. The urea cycle evolves from arginine synthesis. Metabolism of fatty acids and sugars develops after the full development of metabolism of amino acids of groups I and II, and they are associated with the anabolism of amino acids of groups III and IV. Syntheses of aromatic amino acids are branched within sugar metabolism. The Krebs cycle occurs relatively late after the setting of metabolism of amino acids of groups I and II. One portion of the Krebs cycle has a catabolic origin, whereas the other portion has an anabolic origin in pathways of amino acids of groups III and IV. It is not possible to order glycolysis and gluconeogenesis with regard to the Krebs cycle, as they all belong to "period 6." Pentose-phosphate and Calvin cycles are later (periods 7 and 8, respectively). Cladistic analysis of the structure of biochemical pathways makes hypotheses in biochemical evolution explicit and parsimonious.

Introduction

Metabolic pathways are series of successive biochemical reactions catalyzed by enzymes. As the product of enzymatic activity, they can be seen as a part of the phenotype. Most of the phenotypic attributes are inherited and classically analyzed through standard phylogenetic methods to depict the distribution of character states among the living world. Classically, attributes of organisms are coded into a matrix to express primary homologies (de Pinna 1991), to infer synapomorphies (i.e., secondary homologies) through tree reconstruction, and to obtain a relative ordering through time of branches and character changes. Attributes of metabolic pathways can be coded the same way to order enzymatic innovations through time. From these premises, comparative anatomy of metabolic pathways has been developed by Cunchillos and Lecointre (2000, 2002). The darwinian concept of descent with modification was shown to apply to metabolic pathways, legitimizing the use of cladistic analysis to infer the history of metabolic development. Phylogenetic analysis of amino acid metabolism was performed by defining a new kind of taxon, new characters, and four criteria of primary homology (Cunchillos and Lecointre 2002, 2003). To achieve this goal, a taxon was defined as a "pathway"; that is, the series of reactions from the tip of the pathway to its point of connection into the Krebs cycle. This definition is valid either for catabolism, where the tip of the pathway is the nutrient, or for anabolism, where the tip is the final product. Characters were defined as presence of enzymes, type of reactions, or use of cofactors, according to four criteria of primary homology. Primary homologies of "type I" were defined as several pathways sharing the same enzyme with high specificity for its substrate. In such case, the name of the enzyme itself is the name of the character (a column in the matrix). Primary homologies of "type II" were defined as the sharing of similar enzymatic functions by different pathways (although the enzymes recognized can have different specificities), sharing the same cofactors or sharing the same functional family (see Cunchillos and Lecointre [2003] for details). Because the parts of metabolism analyzed are universal, the inferred history is located in time before the phylogenetic development of the species "tree of life" as we know it today.

Using these new concepts, the historical development of aliphatic amino acid catabolism was first inferred in relation to the development of the Krebs cycle, and then aliphatic amino acid anabolism was incorporated into the matrix. Obviously, such a methodological framework now has to incorporate the most widely shared pathways through life, called here "universal metabolism"; that is, anabolism and catabolism of the three fundamental kinds of biomolecules: amino acids, fatty acids, and saccharides. It is of interest because that framework offers new and transparent means to answer questions about metabolism evolution. One of them is the relative timing of the rise of glycolysis, the Krebs cycle, and amino acid biosynthesis. Meléndez-Hevia, Waddell, and Cascante (1996) considered glycolysis as earlier than both the amino acid biosynthesis and the Krebs cycle because glucose is implicitly considered as more profitable as the earliest nutrient than are amino acids. Amino acid catabolism was, therefore, thought by these

Key words: metabolism, metabolic pathways, biochemical evolution.

E-mail: lecointr@mnhn.fr.

Mol. Biol. Evol. 22(1):1–11. 2005
doi:10.1093/molbev/msh253
Advance Access publication September 8, 2004

authors to be secondary. Conversely, Cunchillos and Lecointre (2003) found the Krebs cycle as the product of metabolism of amino acids of groups I and II (in the sense of Cordón [1990], namely, Asp, Asn, Glu, Gln, Arg, and Pro). However, their matrix did not include glycolysis. For this reason, glycolysis and gluconeogenesis have been included here. Is the metabolism of monosaccharides (also including Calvin and pentose-phosphate cycles) the first? In the same framework, when did fatty acid metabolism rise, compared with other pathways? A number of other metabolic pathways have also been included in the present work to submit a greater diversity of widely shared reactions to a transparent procedure for inferring a more complete view of their temporal development. For instance, the development of the urea cycle is traditionally thought to be linked to the metabolism of arginine and might have been possible as early as arginine metabolism, and the metabolism of aromatic amino acids is classically thought as arising later than the appearance of aliphatic amino acid metabolism.

Earlier authors (e.g., Horowitz 1945; Cordón 1990) have speculated on the evolutionary timing of the development of pathways from a theoretical point of view. For example, because free aliphatic amino acids were considered one of the very first sources of molecules in abiotic environments, upstream reactions of amino acid catabolic pathways must have occurred before downstream reactions, whereas downstream reactions of amino acid anabolic pathways must have occurred before upstream reactions (see Cunchillos and Lecointre [2000]). The present methods have the power to test such hypotheses. By using those methods, those hypotheses were corroborated for some aliphatic amino acids only (Cunchillos and Lecointre 2003). The evolutionary timing of development of other pathways can now to be tested in the same way. In the present work, new data increasing complexity of characters required to face new methodological challenges.

Materials and Methods

Taxonomic Sampling

The present work focuses on the metabolism of the three main kinds of universal biochemical compounds, namely, amino acids, fatty acids, and monosaccharides. Analysis of the metabolism of these compounds requires the coding of the following assemblages of pathways: Krebs cycle, Calvin cycle, pentose-phosphate cycle, urea cycle, fatty acid anabolism and catabolism, amino acid anabolism and catabolism, glycolysis, and gluconeogenesis. These pathways are shared by all living things, at least primitively, with few exceptions (Calvin cycle), and discussion of secondary losses in some species or species groups is beyond the scope of the present work.

Enzymes acting on complex molecules have been excluded from that universal core of enzymatic activities. Complex molecules are assemblages of those elementary molecules whose metabolic evolution is being studied. Complex molecules include, for instance, coenzymes, triglycerids, phospholipids, nucleosids, and polymers (e.g., glycogen, starch, proteins, DNA, and RNA). They all are secondary products of the core metabolism studied here. Purines and pyrimidines are considered as complex molecules themselves because they are never synthesized as such in vivo. Indeed, their precursors are already attached to other compounds (riboses or ribose-phosphates), so their recognition as isolated compounds have no biological basis. This delineation leads us to sample the following pathways for phylogenetic analysis.

As in Cunchillos and Lecointre (2000), taxa are defined from the tip of the pathway to its point of contact into the Krebs cycle. To name pathways, prefixes "d" and "s" are used to refer to degradation and synthesis, respectively. For example, dGLN is the set of enzymatic activities involved in converting glutamine to oxoglutarate, whereas sGLN is the synthetic pathway from oxoglutarate to glutamine. When degradation or synthesis of a compound can occur in different ways, they are numbered in the name of the pathway. For instance, cysteine can be degraded via mercaptopyruvate (dCYS2) or directly through pyruvate and acetyl-CoA (dCYS1). Taxons are listed in tables 1 and 2. Fatty acid catabolic and anabolic pathways stop at acetyl-CoA. Monosaccharide anabolic pathways (gluconeogenesis and pentose-phosphate cycle) stop at oxaloacetate, and monosaccharide catabolic pathways (glycolysis and Calvin cycle) stop at acetyl-CoA. Amino acid anabolism and catabolism stop at different points of the Krebs cycle or at acetyl-CoA, depending on the amino acid (oxoglutarate, succinyl-CoA, oxaloacetate, or acetyl-CoA). The Krebs cycle is divided into two parts, designated using the two main points of entrance: KC1 from oxaloacetate to oxoglutarate and KC2 from oxoglutarate to oxaloacetate. The urea cycle is not delineated in reference to the Krebs cycle, but independently, including the reactions of its own cycle.

Characters and Homologies

Characters are defined according to four criteria of homology (see Cunchillos and Lecointre [2003] for details). Primary homologies of "type I" were defined as several pathways sharing the same enzyme with high specificity for its substrate. The name of the enzyme itself is then given to the character. Primary homologies of "type II" were defined as the sharing of similar enzymatic functions by different pathways (although the enzymes recognized can have different specificities, "IIa"), sharing the same cofactors ("IIb"), or being of the same functional family ("IIc"). For the coding of homologies of type II, in a number of taxa, the coding involves reactions embedded within the Krebs cycle. The anabolism of alanine (sALA), valine (sVAL), serine (sSER1 and sSER2), glycine (sGLY1 and sGLY2), and cysteine (sCYS1 and sCYS2), gluconeogenesis, and the Calvin cycle all start with β-decarboxylation of oxaloacetate. However, there cannot be any β-decarboxylation without a preceding β-dehydrogenation of an alcohol to obtain the carbonyl precursor. For the pathways considered here, this is done within the Krebs cycle, from malate to oxaloacetate. So these pathways (taxons) have to incorporate β-dehydrogenation as a type II homology (coding "1" at character 27: alcohol NAD-dehydrogenation), as if they were starting from malate. For symmetrical reasons, the second portion of the Krebs cycle (from oxoglutarate to oxaloacetate, "KC2") does include the presence of a β-decarboxylation from

Table 1
Names of Characters, with the Corresponding Number of International Nomenclature

[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 (biotine); IIa
[9] No Alcohol, no aldehyde NAD dehydrogenases; IIa
[10] β-Decarboxylation; IIa
[11] Decarboxylation; IIc
[12] Deamination; IIc
[13] Acid-ammonia ligases; IIa
[14] Phosphorylation; IIc
[15] Acyl-CoA synthases; IIa
[16] Thiamine pyrophosphate enzymes; IIb
[17] Transketolases; IIa
[18] β-Oxidation/reduction sequence repeat; IId
[19] β-Oxidation sequence; IIa
[20] No niacin, no flavin Oxidases; IIa
[21] Tetrahydrofolate enzymes; IIb
[22] Transaldolases; IIa
[23] Acetyltransferases; IIa
[24] Succinyltransferases; IIa
[25] Malonyltransferases; IIa
[26] Ammonia lyases; IIa
[27] Alcohol dehydrogenation (NAD); IIa
[28] FAD dehydrogenation; IIa
[29] Hydratases; IIa
[30] Isomerases; IIa
[31] Hydrolases; IIa
[32] Phosphoribosyltransferase; IIa

[33] *Amino-acid dehydrogenase*: 1.4.1.5; I
[34] *Aspartate aminotransferase*: 2.6.1.1; I
[35] *Asparaginase*: 3.5.1.1; I
[36] *Glutamate dehydrogenase*: 1.4.1.3; I
[37] *Pyruvate dehydrogenase*: 1.2.4.1; I
[38] *Serine deaminase*: 4.3.1.19; I
[39] *Serine hydroxymethyltransferase*: 2.1.2.1; I
[40] *Cysteine desulphhydrase*: 4.4.1.1; I
[41] *Propionyl-CoA carboxylase*: 6.4.1.3; I
[42] *Branched-chain-amino-acid aminotransferase*: 2.6.1.42; I
[43] *Branched-chain ketoacid dehydrogenase:* 1.2.4.4; I
[44] *Acetyl-CoA acetyltransferase*: 2.3.1.9; I
[45] *Aminoadipate-semialdehyde dehydrogenase*: 1.2.1.31; I
[46] *Aminoadipate aminotransferase*: 2.6.1.39; I
[47] *Saccharopine dehydrogenase 1*: 1.5.1.9; I
[48] *Alanine aminotransferase*: 2.6.1.2; I
[49] *Glutaminase*: 3.5.1.2; I
[50] *Glutamine-oxo-acid aminotransferase*: 2.6.1.15; I
[51] ω-Amidase: 3.5.1.3; I
[52] *Arginase*: 3.5.3.1; I
[53] *Ornithine-oxto-acid aminotransferase*: 2.6.1.13; I
[54] *Pyrroline-carboxylate reductase*: 1.5.1.2; I
[55] *Cysteine aminotransferase*: 2.6.1.3; I
[56] *Mercaptopyruvate sulphurtransferase*: 2.8.1.2; I
[57] *Cysteine dioxygenase*: 1.13.11.20; I
[58] *Aspartate decarboxylase*: 4.1.1.12; I
[59] *Homocysteine methyltransferase*; 2.1.1.10; I
[60] *Cystathionine synthase*: 4.2.1.22; I
[61] *Hydroxymethylbutyryl-CoA dehydrogenase*: 1.1.1.178; I
[62] *Methylcrotonoyl-CoA carboxylase*: 6.4.1.4; I
[63] *Methylglutaconyl-CoA hydratase*: 4.2.1.18; I
[64] *Hydroxymethylglutaryl-CoA lyase*: 4.1.3.4; I
[65] *Ketoacid CoA-transferase*: 2.8.3.5; I
[66] *Methylmalonate-semialdehyde dehydrogenase*: 1.2.1.27; I
[67] *Saccharopine dehydrogenase 2*: 1.5.1.7; I
[68] *Citrate synthase*: 2.3.3.1; I
[69] *Citrate dehydratase*; 4.2.1.3; I

Table 1
continued

[70] *Isocitrate dehydrogenase*: 1.1.1.41; I
[71] α-*Oxoglutarate dehydrogenase*: 1.2.4.2; I
[72] *Succinyl-CoA hydrolase*: 3.1.2.3; I
[73] *Succinate dehydrogenase*: 1.3.99.1 I
[74] *Fumarate hydratase*: 4.2.1.2; I
[75] *Malate dehydrogenase*: 1.1.1.37; I
[76] *Asparagine sinthase* (ADP-forming): 6.3.1.4; I
[77] *Asparagine synthase* (glutamine-hydrolysing): 6.3.5.4; I
[78] *Glutamine synthetase*: 6.3.1.2; I
[79] *Amino-acid acetyltransferase*: 2.3.1.1; I
[80] *Acetylglutamate kinase*: 2.7.2.8; I
[81] *N-acetyl-glutamyl-P reductase*: 1.2.1.38; I
[82] *Acetylornithine aminotransferase*: 2.6.1.11; I
[83] *Acetylornithine deacetylase*: 3.5.1.16; I
[84] *Ornithine carbamoyltransferase*; 2.1.3.3; I
[85] *Argininosuccinate synthetase*: 6.3.4.5; I
[86] *Argininosuccinate lyase*: 4.3.2.1; I
[87] *Glutamate kinase*: 2.7.2.11; I
[88] *Glutamate semialdehyde dehydrogenase*: 1.5.1.12; I
[89] *Oxaloacetate decarboxylase*: 4.1.1.3; I
[90] *Phosphoenolpyruvate carboxykinase* (GTP): 4.1.1.32; I
[91] *Enolase*: 4.2.1.11; I.
[92] *Phosphoglycerate phosphomutase*:5.4.2.1; I
[93] *Glycerate kinase*: 2.7.1.31; I
[94] *Hydroxypyruvate reductase*: 1.1.1.81; I
[95] *Serine-pyruvate aminotransferase*: 2.6.1.51; I
[96] *Phosphoglycerate dehydrogenase*: 1.1.1.95; I
[97] *Phosphoserine aminotransferase*: 2.6.1.52; I
[98] *Phosphoserine phosphatase*: 3.1.3.3; I
[99] *Acyl-CoA dehydrogenase*: 1.3.99.3; I
[100] *Serine acetyltransferase*: 2.3.1.30; I
[101] *Cysteine synthase*: 2.5.1.47; I
[102] *Aspartate kinase*: 2.7.2.4; I
[103] *Aspartate-semialdehyde dehydrogenase*: 1.2.1.11; I
[104] *Homoserine dehydrogenase*: 1.1.1.3; I
[105] *Homoserine succinyltransferase*:.3.1.46; I
[106] *Cystathionine synthase*: 2.5.1.48; I
[107] *Cystathionine lyase*: 4.4.1.8; I
[108] *Homoserine kinase*: 2.7.1.39; I
[109] *Threonine synthase*: 4.2.3.1; I
[110] *Acetolactate synthase*: 2.2.1.6; I
[111] *Keto-acid reductoisomerase*:1.1.1.86. I
[112] *Dihydroxyacid dehydratase*: 4.2.1.9. I
[113] *Isopropylmalate synthase*: 2.3.3.13; I.
[114] *Isopropylmalate isomerase*: 4.2.1.33 ; I
[115] *Isopropylmalate dehydrogenase*: 1.1.1.85; I
[116] *Dihydrodipicolinate synthase*: 4.2.1.52 ; I
[117] *Dihydrodipicolinate reductase*: 1.3.1.26; I
[118] *Piperideindicarboxylate succinyltransferase*: 2.3.1.117; I
[119] *Succinyl-diaminopimelate aminotransferase*: 2.6.1.17; I
[120] *Succinyl-diaminopimelate desuccinylase*: 3.5.1.18; I
[121] *Diaminopimelate epimerase*: 5.1.1.7; I
[122] *Diaminopimelate decarboxylase*: 4.1.1.20; I
[123] *Homocitrate synthase*: 2.3.3.14; I
[124] *Homoaconitate hydratase*: 4.2.1.36; I
[125] *Homoisocitrate dehydrogenase*: 1.1.1.155; I
[126] *Phosphoglycerate kinase*: 2.7.2.3; I
[127] *Triosephosphate dehydrogenase*: 1.2.1.12; I
[128] *Triose-phosphate isomerase*: 5.3.1.1; I
[129] *Fructose-bisphosphate aldolase*: 4.1.2.13; I
[130] *Fructose-bisphosphatase*: 3.1.3.11; I
[131] *Glucose-6-phosphate isomerase*: 5.3.1.9; I
[132] *6-Phosphofructokinase*: 2.7.1.11; I
[133] *Pyruvate kinase*: 2.7.1.40; I
[134] *Glycolaldehydetransferase*: 2.2.1.1; I
[135] *Sedoheptulose diphosphatase*: 3.1.3.37; I
[136] *Ribulose-phosphate 3-epimerase*: 5.1.3.1; I
[137] *Ribulose-5-phosphate isomerase*: 5.3.1.6; I
[138] *Phosphoribulokinase*: 2.7.1.19; I
[139] *Ribulose diphosphate carboxylase*: 4.1.1.39; I
[140] *Glucose-6-phosphate 1-dehydrogenase*: 1.1.1.49; I

Table 1
continued

[141] *Phosphogluconolactonase*: 3.1.1.31; I
[142] *Phosphogluconic acid dehydrogenase*: 1.1.1.44; I
[143] *Dihydroxyacetonetransferase*: 2.2.1.2; I
[144] *Dihydrolipoamide dehydrogenase*: 1.8.1.4; I
[145] *Acetate thiokinase*: 6.2.1.13; I
[146] *Acyl-CoA oxidase*: 1.3.3.6; I
[147] *Enoyl-CoA hydratase*: 4.2.1.17; I
[148] β-*Hydroxyacyl dehydrogenase*: 1.1.1.35; I
[149] *Acetyl-CoA C-acyltransferase*: 2.3.1.16; I
[150] *Trans-2-enoyl-CoA reductase (NAD)*: 1.3.1.44; I
[151] *Acyl-ACP-hydrolase*: 3.1.2.14; I
[152] *Acetyl-CoA carboxylase*: 6.4.1.2; I
[153] *Acyl-carrier-protein] S-malonyltransferase*: 2.3.1.39; I
[154] *Acyl-carrier-protein] S-acetyltransferase*: 2.3.1.38; I
[155] β-*Ketoacyl synthetase*: 2.3.1.41; I
[156] β-*Ketoacyl reductase*: 1.1.1.100; I
[157] *Enoyl acyl carrier protein hydratase*: 4.2.1.58; I
[158] *Acyl-ACP dehydrogenase*: 1.3.1.10; I
[159] β-*Hydroxydecanoate dehydrase*: 4.2.1.60; I
[160] *Glutaryl-CoA dehydrogenase*:1.3.99.7; I
[161] *Phenylalanine 4-monooxygenase*:1.14.16.1; I
[162] *Hydroxyphenylpyruvate dioxygenase*:1.13.11.27; I
[163] *Dihydroxyphenylacetate 2,3-dioxygenase*:1.13.11.15; I
[164] *Maleylacetoacetate isomerase*: 5.2.1.2; I
[165] *Fumarylacetoacetase*: 3.7.1.2; I
[166] *Tryptophan 2,3-dioxygenase*: 1.13.11.11; I
[167] *Kynurenine formamidase*: 3.5.1.9; I
[168] *Kynurenine hydroxylase*: 1.14.13.9; I
[169] *Kynureninase*: 3.7.1.3; I
[170] *Hydroxyanthranilate oxygenase*: 1.13.11.6; I
[171] *Picolinic acid carboxylase*: 4.1.1.45; I
[172] *Aminomuconate-semialdehyde dehydrogenase*: 1.2.1.32; I
[173] *Histidine ammonia-lyase*: 4.3.1.3; I
[174] *Urocanate hydratase*: 4.2.1.49; I
[175] *Imidazolone propionic acid hydrolase*: 3.5.2.7; I
[176] *Glutamate formyltransferase*: 2.1.2.5; I
[177] *Aromatic amino acid aminotransferase*: 2.6.1.57; I
[178] *Deoxy-7-phosphoheptulonate synthase*: 2.5.1.54; I
[179] *Dehydroquinate synthase*: 4.2.3.4; I
[180] *Dehydroquinate dehydratase*: 4.2.1.10; I
[181] *Shikimate 5-dehydrogenase*: 1.1.1.25; I.
[182] *Shikimate kinase*: 2.7.1.71; I
[183] *Riboflavin synthase*: 2.5.1.9; I
[184] *Anthranilate synthase*: 4.1.3.27; I
[185] *Anthranilate phosphoribosyltransferase*: 2.4.2.18; I
[186] *Phosphoribosylanthranilate isomerase*: 5.3.1.24; I
[187] *Indoleglycerol phosphate synthetase*: 4.1.1.48; I
[188] *Tryptophan synthase*: 4.2.1.20; I
[189] *Prephenate dehydrogenase*: 1.3.1.12; I
[190] *Chorismate mutase*: 5.4.99.5; I
[191] *Prephenate dehydratase*: 4.2.1.51; I
[192] *Methylmalonyl-CoA epimerase*: 5.1.99.1; I.
[193] *Methylmalonyl-CoA mutase*: 5.4.99.2; I
[194] *Carbamate kinase*: 2.7.2.2; I
[195] *ATP phosphoribosyltransferase*: 2.4.2.17; I
[196] *Phosphoribosyl-ATP diphosphatase*: 3.6.1.31; I
[197] *Phosphoribosyl-AMP cyclohydrolase*: 3.5.4.19; I
[198] *Phosphoribosylformiminoaminophosphoribosylimidazole-carboxyamide isomerase*: 5.3.1.16; I
[199] *Imidazoleglycerol-phosphate dehydratase*: 4.2.1.19;I
[200] *Histidinol-phosphate transaminase*: 2.6.1.9; I
[201] *Histidinol-phosphatase*: 3.1.3.15; I
[202] *Histidinol dehydrogenase*: 1.1.1.23; I

NOTE.—Homology types defined in the text. For characters 1 to 32, homologies are of type II. For characters 33 to 204, homologies are of type I (Enzyme Nomenclature 1973).

oxaloacetate to pyruvate or phosphoenolpyruvate in the coding of character 10.

A new kind of homology of type II was used in the present work. Pathways can be similar in the recurrence of a set of reactions made by the same enzymes from different substrates. This homology is used in the metabolism of fatty acids (character 18) and called "IId." All characters are named in table 1, with the number from enzyme nomenclature and the type of homology involved. In figures 1 and 2, characters involving homologies of type II are numbered from 1 to 32. It must be stressed that, if homologies of type I are named strictly following the international enzymatic nomenclature, homologies of type II are not. For example, hydratase (character 29) is used here in a wider meaning than in the international nomenclature. A reaction involving the pyridoxal-phosphate is coded as involving a hydratase because a molecule of water is implied, whereas it is not the case for international nomenclature. Conversely, our delineation of homologies of type II is more precise for dehydrogenases (or reductases) involving NAD. Alcohol-NAD-dehydrogenases, aldehyde-NAD-dehydrogenases, deaminase-NAD-dehydrogenases, and FAD-dehydrogenases have been separated into different characters.

A number of characters referring to homologies of type II contain question marks. These are included when the taxon-pathway does not exhibit the appropriate type of substrate for the enzymatic function, the coenzyme, or the functional family at hand. A question mark is also used when the enzymatic function to be coded is meaningless for the taxon. For instance, the coding of functions of degradation into anabolic pathways is meaningless: there is no way for an α-decarboxylation into a biosynthetic pathway.

Phylogenetic Reconstruction

Tree Search

The matrix contains 75 taxons and 202 characters (table 2). Characters are treated as unordered and unweighted. Heuristic searches were conducted with NONA (Goloboff 1998) as implemented into WINCLADA (Nixon 1999*a*), using TBR branch swapping. For a better exploration of trees, the Parsimony Ratchet (Hopper Islands [Nixon 1999*b*]) was used. The proportion of data to be reweighted was set between 25% and 50% and the number of iterations progressively increased from 25,000 to 150,000 (option amb- poly=). This increase in iterations was used to check that the number of supplementary MP trees gained each time was decreasing or 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). This is justified by the fact that, in the coding of character states, zero was given to the absence of enzymes or to the absence of performance of particular functions (even in presence of a putative suitable substrate) or to absence of use of a cofactor. Such a rooting option will automatically put the simplest pathways closer to the root. However, this does not make any assumption of the nature of the corresponding enzymatic reactions.

Table 2
Matrix Containing 75 Taxons (Rows) and 202 Characters (Columns)

HYPANC	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sASP1	010?0?000??100000??000000000101010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sASP2	101?0?100??100000??000000000101001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sASN1	01010?000??111000??000000000101010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sASN2	10110?100??111000??000000000101001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sASN3	01010?000??111000??000000000101010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sASN4	10110?100??111000??000000000101001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sGLU1	010?0?000??100000??000000000101000010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sGLU2	101?0?100??100000??000000000101001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sGLN1	01010?000??111000??000000000101000010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sGLN2	10110?100??111000??000000000101001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sARG1	111?1?100??101100??000100000101000010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sARG2	101?1?100??101100??000100000101001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sARG3	111?1?100??101000??000000000101000010000000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000000
sARG4	101?1?100??101000??000000000101001000000000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000000
sPRO1	010?1?001??101000??000000000101000010000000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000000
sPRO2	101?1?101??101000??000000000101001000000000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000000
sALA1	101?0?10011100000??000000010101000000000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sSER1	101?0?10011101000??000000010111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sSER2	101?0?10011101000??000000010111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sGLY1	101?0?10011101000??010000010111000000010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sGLY2	101?0?10011101000??010000010111000000010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sCYS1	101?0?10011101100??000100010111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sCYS2	101?0?10011101100??000100010111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sMET1	110?1?000??101100??010010000101010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sMET2	101?1?100??101100??010010000101010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sTHR1	110?1?000??101000??000000000101010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sTHR2	101?1?100??101000??000000000101001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sILE1	111?1?100??101010??000000000101010000100010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sILE2	101?1?100??101010??000000000101001000100010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sLEU	101?0?10011100110??000100010111000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sVAL	101?0?10011100010??000000010101000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sLYS1	111?1?10011101100??000010010111010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sLYS2	101?1?10011101100??000010010111001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sLYS3	111?1?10111101100??000100010101000000000000011100000000000000000001000000000000000000000000000000000000000000000000000000000000000000
dASP1	010??00?000100?00??000000000101010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dASP2	101??01?000100?00??000000000101001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dASN1	0101?00?000100?00??000000000101010100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dASN2	1011?01?000100?00??000000000101001100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dGLU	010??00?000100?00??000000000101000010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dGLN1	0101?00?000100?00??000000000101000010000000000001000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dGLN2	1011?01?000100?00??000000000101000000000000000011000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dARG	010?100?000100?00??000000000101000010000000000000011000000000000000000000000000000000000000000000000000000000000000000000000000000000
dPRO	010?100?100100?00??000000001101000010000000000000001000000000000000000000000000000000000000000000000000000000000000000000000000000000
dALA	101??11?001100?10??000000001101000001000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dSER	100??11?001100?10??000000001000000001100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dGLY	100??11?001100?10??010000001100000001110000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dCYS1	100??11?001100?10??000000001100000001001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dCYS2	101??11?001100?10??000000001101000001000000000000000011000000000000000000000000000000000000000000000000000000000000000000000000000000
dCYS3	101??11?001100?10??100000001101000001000000000010000000110000000000000000000000000000000000000000000000000000000000000000000000000000
dMET	100??111001101?10??010000001110000001001100000000000000001100000000000000000000000000000000000000000000000000000000000000000000000000
dTHR1	100??111001101?10??000000001010000001100100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dTHR2	100??11?001100?10??010000001100000001110000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dILE	101??111001101?10?100000001111100000000011100000000000000000100001000000000000000000000000000000000000000000000000000000000000000010
dLEU	101??111001101110??000110011101000000000011100000000000000000111110000000000000000000000000000000000000000000000000000000000000000010
dVAL	101??111011101?10?100000001111100000000011100000000000000000100000000000000000000000000000000000000000000000000000000000000000000010
dLYS	111?111?111100010?100010001110100000000000011110000000000000000000010001000000000000000000000000000000000000000000000000000000000000
1Krebs Cycle	?????0?0011???100??000100010100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
2Krebs Cycle	?????1?0011??0010?100000001110100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Glycolysis	????11??001??1010??001000001110000001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Gluconeogenesis	????10?0011??1000??001000010111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Calvin cycle	????11??011??1011??001000011111000001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Pentose ph cycle	????10?0011??1011??001000010111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dFatty acids	?????????000??1100110001000111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
s2Fatty acids	????????00????1100110001000101010000000000001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
s1Fatty acids	????????11????1100110001010101010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dTYR	101??11?001100000??100000000111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dPHE	101??11?001100000??100000000111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dTRP	111?111?101100010??100100011101000001000001000100000000000000000000000010000000000000000000000000000000000000000000000000000000000000
dHIS	101??01?000100?00??010000100101001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sTYR	1?1?1?10111??1011??001000010111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sPHE	1?1?1?10111??1011??001000010111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sTRP	1?1?1?10111??1011??001000010111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sHIS	1?1?1?10111??1011??001000010111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Urea Cycle	111?1?100??101000??000000000101000010000000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000000

The aim was to produce explicit hypotheses so that the character coding and trees produced by the present study will still be useful even if another way of rooting is used. They will just have to be considered again 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. Because a different rooting yields different phylogenetic interpretations, the present work would still be useful for such a new interpretation.

Defining Time Spans

Criteria

From the root to the tip of branches, phylogenetic trees provide a relative order of transformations (here enzymatic innovations) through time. We call an "upstream node" a deep, more inclusive internal branch or clade and a "downstream node" a more terminal, less inclusive internal branch or clade. Time spans ("periods" or "phases") in metabolism are defined as the time along

Table 2
continued

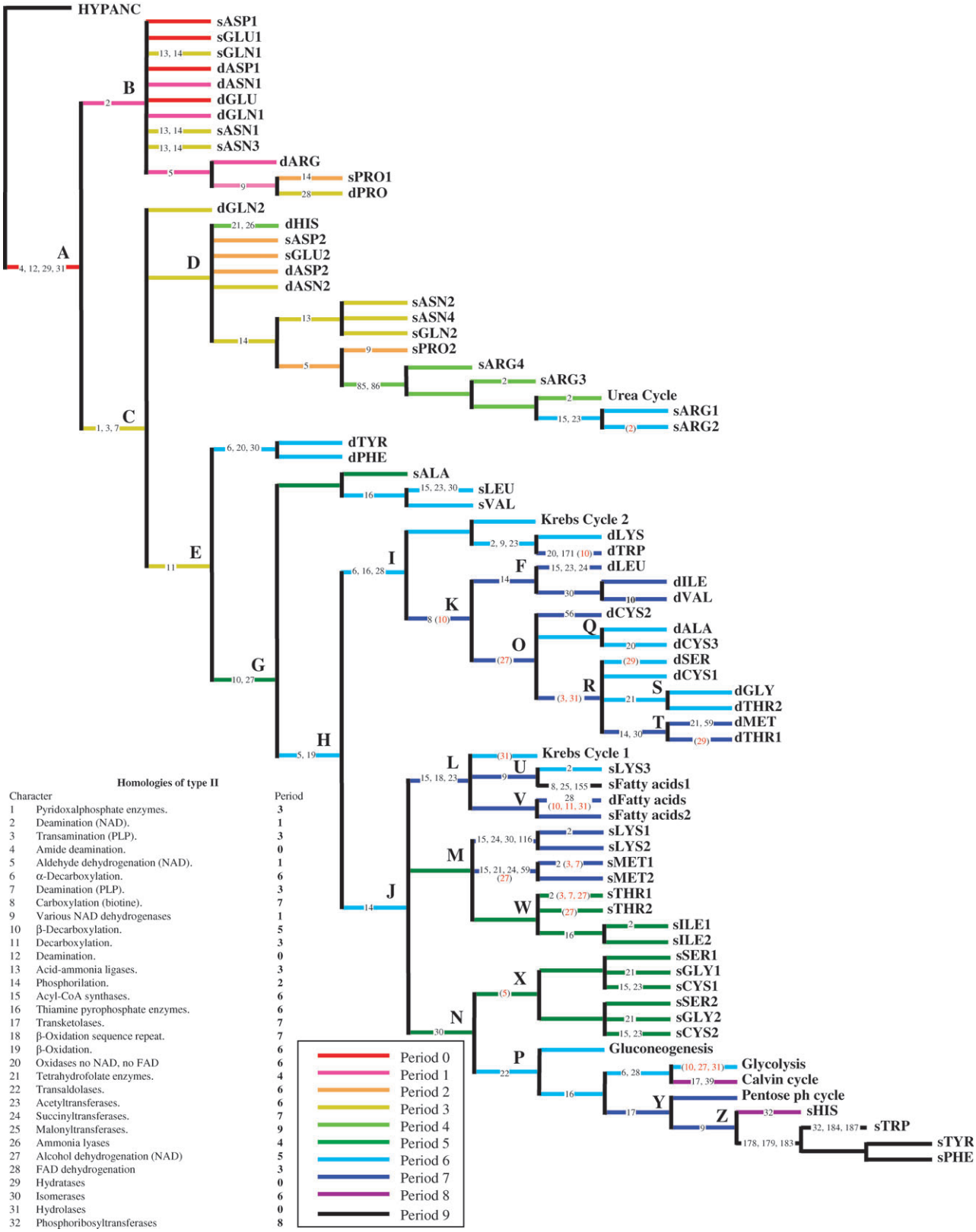
HYPANC	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sASP1	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sASP2	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sASN1	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sASN2	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sASN3	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sASN4	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sGLU1	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sGLU2	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sGLN1	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sGLN2	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sARG1	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sARG2	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sARG3	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sARG4	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sPRO1	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sPRO2	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sALA1	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sSER1	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sSER2	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sGLY1	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sGLY2	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sCYS1	10000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sCYS2	10000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sMET1	01111110000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sMET2	01111110000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sTHR1	01110001100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sTHR2	01110001100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sILE1	01110001111100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sILE2	01110001111100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sLEU	00000000011111100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sVAL	00000000011100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sLYS1	01100000000000111111100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sLYS2	01100000000000111111100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sLYS3	00000000000000000000011100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dASP1	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dASP2	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dASN1	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dASN2	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dGLU	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dGLN1	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dGLN2	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dARG	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dPRO	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
dALA	00000000000000000000000000000000000000000000000010000000000000000000000000000000000000000000000000000000000000000000
dSER	00000000000000000000000000000000000000000000000010000000000000000000000000000000000000000000000000000000000000000000
dGLY	00000000000000000000000000000000000000000000000010000000000000000000000000000000000000000000000000000000000000000000
dCYS1	00000000000000000000000000000000000000000000000010000000000000000000000000000000000000000000000000000000000000000000
dCYS2	00000000000000000000000000000000000000000000000010000000000000000000000000000000000000000000000000000000000000000000
dCYS3	00000000000000000000000000000000000000000000000010000000000000000000000000000000000000000000000000000000000000000000
dMET	00000010000000000000000000000000000000000000000010000000000000000000000000000000000000000000000000000011000000000
dTHR1	00000000000000000000000000000000000000000000000010000000000000000000000000000000000000000000000000000011000000000
dTHR2	00000000000000000000000000000000000000000000000010000000000000000000000000000000000000000000000000000000000000000
dILE	00000000000000000000000000000000000000000000000010010000000000000000000000000000000000000000000000000011000000000
dLEU	00000000000000000000000000000000000000000000000010000000000000000000000000000000000000000000000000000000000000000
dVAL	00000000000000000000000000000000000000000000000010010000000000000000000000000000000000000000000000000011000000000
dLYS	00000000000000000000000000000000000000000000000010011000000000010000000000000000000000000000000000000000000000000
1Krebs Cycle	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
2Krebs Cycle	00000000000000000000000000000000000000000000000010000000000000000000000000000000000000000000000000000000000000000000
Glycolysis	00000000000000000000000000011101110000000001000000000000000000000000000000000000000000000000000000000000000000000000
Gluconeogenesis	00000000000000000000000000111111000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Calvin cycle	00000000000000000000000001111111111111100010000000000000000000000000000000000000000000000000000000000000000000000000
Pentose ph cycle	00000000000000000000000001111110010110011110000000000000000000000000000000000000000000000000000000000000000000000000
dFatty acids	00000000000000000000000000000000000000000000001111100000000000000000000000000000000000000000000000000000000000000000
s2Fatty acids	00000000000000000000000000000000000000000000000011111000000000000000000000000000000000000000000000000000000000000000
s1Fatty acids	00000000000000000000000000000000000000000000000011111111100000000000000000000000000000000000000000000000000000000000
dTYR	00000000000000000000000000000000000000000000000000000001111100000000001000000000000000000000000000000000000000000000
dPHE	00000000000000000000000000000000000000000000000000000000111100000000001000000000000000000000000000000000000000000000
dTRP	00000000000000000000000000000000000000000000001001100000000001000001111111000000000000000000000000000000000000000000
dHIS	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
sTYR	00000000000000000000000001111110010110011110000000000000000000000000000000000000000000000000000000000000000000000000
sPHE	00000000000000000000000001111110010110011110000000000000000000000000000000000000000000000000000000000000000000000000
sTRP	00000000000000000000000001111110010110011110000000000000000000000000000000000000000000000000000000000000000000000000
sHIS	00000000000000000000000001111110010110011110000000000000000000000000000000000000000000000000000000000000000000000000
Urea Cycle	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000

NOTE.—Each taxon (row) 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 in table 1.

the tree separating two character changes of type II, taking into account the following criteria as in Cunchillos and Lecointre (2003):

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. This restriction is simply because of the last period (9), for which it cannot be said that reactions are of the same period: they are just reactions later than period 8.

The nature of enzymatic changes: If a branch is followed by downstream branches that do not bear changes



Homologies of type II

Character		Period
1	Pyridoxalphosphate enzymes.	3
2	Deamination (NAD).	1
3	Transamination (PLP).	3
4	Amide deamination.	0
5	Aldehyde dehydrogenation (NAD).	1
6	α-Decarboxylation.	6
7	Deamination (PLP).	3
8	Carboxylation (biotine).	7
9	Various NAD dehydrogenases	1
10	β-Decarboxylation.	5
11	Decarboxylation.	3
12	Deamination.	0
13	Acid-ammonia ligases.	3
14	Phosphorilation.	2
15	Acyl-CoA synthases.	6
16	Thiamine pyrophosphate enzymes.	6
17	Transketolases.	7
18	β-Oxidation sequence repeat.	7
19	β-Oxidation.	6
20	Oxidases no NAD, no FAD	6
21	Tetrahydrofolate enzymes.	4
22	Transaldolases.	6
23	Acetyltransferases.	6
24	Succinyltransferases.	7
25	Malonyltransferases.	9
26	Ammonia lyases	4
27	Alcohol dehydrogenation (NAD)	5
28	FAD dehydrogenation	3
29	Hydratases	0
30	Isomerases	6
31	Hydrolases	0
32	Phosphoribosyltransferases	8

FIG. 1.—Strict consensus of 20 equiparsimonious trees obtained through the Parsimony Ratchet of WINCLADA calculated on the matrix. Each tree is of 347 steps (CI = 0.58, RI = 0.79). Character reversions in red.

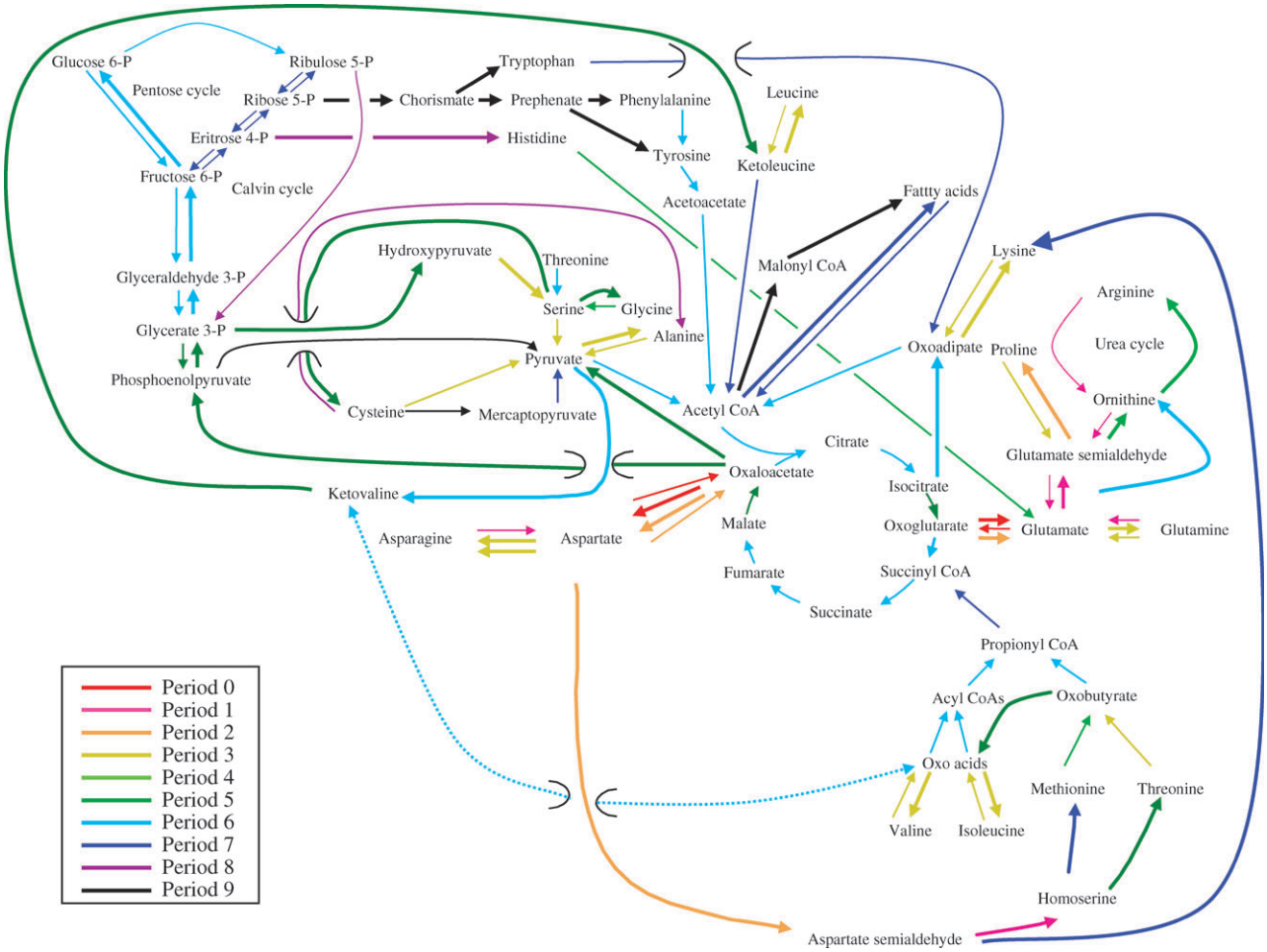


FIG. 2.—General view of metabolic pathways and their connecting points to the Krebs cycle. Colors are successive time spans (or "periods") as inferred from the tree. Reactions in black (period 9) do not mean that all these reactions appeared at the same time. These are just reactions later than the purple period (period 8).

in homologies of type II, the downstream branches are of the same period. When a new homology of type II occurs on a branch, it defines the next period. When the homology of type II occurs in the next branch but is already present in earlier periods (homoplasy), it does not define a new period. For example, the node H is light blue (fig. 1, period 6), and the downstream node J exhibits a change in homology of type II (homology II n°14) while it remains in light blue. There is no new period here because the phosphorylation is already available in a previous period (period 2, within the clade B: sPRO1).

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

Classically, homologies of type II involve several enzymes with different specificities. However, for 14 cases, single enzymes with the same high specificity innovate a new enzymatic mechanism. They are recorded as homologies of type I, but because they also correspond to enzymatic changes used to define homologies of type II, they are taken into account for defining periods. In figure 1, they are shown with asterisks.

Polytomies

It must be stressed that both the tree topology and the new nonhomoplastic changes in homologies of type II are involved in the definition of periods. How this can be used when parts of the tree are unresolved? A polytomy is just an absence of resolution. A polytomy normally contains branches of different periods, simply because there are different numbers of new homologies of type II occurring on them. For example, within the clade B (fig. 1), a polytomy involves branches of the same period as the origin of the clade B (period 1, pink) and branches of later periods 2 (orange) and 3 (yellow). Some of these branches are of period 3 because they exhibit two additional homologies of type II (13 and 14). Homologies of type II are, therefore, cumulative in defining periods.

The Problem of Question Marks

Question marks bring complications in the definition of time spans. As a consequence of using the parsimony criterion, some homologies of type II are optimized onto a node deeper than the ones where the derived state is really observed, because of question marks in the rest of terminals of the clade. For example, character 8 (carboxylations using biotin) is optimized to occur on the node K. Among the 12 pathways included within K, only five (dLEU, dILE, dVAL, dMET, and dTHR1) really exhibit this function; all others are coded "?" for that function. Thus, the early position of the gain of character 8 is because of the parsimony criterion used to manage question marks in character optimization, whereas real observation of this function would optimize two gains of character 8 onto nodes T and F only. This causes no problem, except for the ordering the time spans. K is of period 7 because of the gain of this homology of type II. However, seven downstream pathways (dALA, dCYS2, dCYS3, dSER, dCYS1, dGLY, and dTHR2) do not really have to wait for that period to be achieved, because they do not need to use this function, for which they are coded "?". Therefore, assigning their complete development to period 7 is an artifact of optimization affecting time-span assignment. Because these pathways are actually possible as early as the previous period (period 6), they have been assigned to that period. This is the reason why there can be an apparent contradiction in observing a node or a branch to which is assigned a period earlier than an upstream (more inclusive) branch (dCYS1 is possible in period 6 while node K is in period 7). Such situations are met four times in the analysis and are always caused by question marks.

Results

From our data matrix, the heuristic search using the Parsimony Ratchet stabilized the number of equiparsimonious trees to 20 (347 steps, CI = 0.58, RI = 0.79). The strict consensus is shown figure 1. The first enzymes to occur (node A) are hydrolases, hydratases, and those involving deaminations of amino acids. Therefore, degradation of amino acids must have preceded all other kind of reactions.

Cordón (1990) defined four groups of amino acids, according to common reactions in their metabolism and the point of entry into the Krebs cycle: group I enters by oxaloacetate, group II by ketoglutarate, group III by pyruvate and acetyl-CoA, and group IV by succinyl-CoA. In our tree (fig.1), amino acids of Cordón's groups I and II develop first. At period 6 (light blue), it is possible to synthesize and degrade all these amino acids (Asp, Asn, Arg, Pro, Gln, and Glu). Based on the full development of the synthesis of arginine (sARG3 and sARG4), the urea cycle is fully developed by the period 4 (light green). The complete metabolism of Cordón's groups III (Ser, Gly, Cys, and Ala) and IV (Thr, Val, Ile, and Met) starts at period 5 (deep green) and mostly develops during periods 5 (for syntheses) and 6 to 7 (for degradations). Aromatic amino acids are degraded as early as period 4 for histidine and period 6 or later for others, whereas their synthesis is so sophisticated that it is only possible very late in periods 8 and 9.

The two parts of the Krebs cycle develop within period 6. Period 5 (deep green) is the period of termination of almost all aliphatic amino acid syntheses (except for lysine and methionine, whose syntheses are achieved in period 7). Period 6 (light blue) is very rich; that is, it contains many events that cannot be ordered: the closing of the Krebs cycle, the termination of the degradation of all aliphatic amino acids of groups III (except dCYS2), and the full development of glycolysis and gluconeogenesis. Later in period 7 (deep blue) almost all aliphatic amino acid metabolism (except dCYS2), degradation of fatty acids, and one of the two syntheses of fatty acids (the "intramitochondrial pathway," here numbered "2") are all possible. It is noticeable that the relative order of glycolysis and the closure of the Krebs cycle cannot be clarified; both of them appear to be the consequence of amino acid metabolic pathways. Period 8 (purple) closes the complete development of all aliphatic amino acid metabolisms and develops the Calvin cycle. Period 10 is arbitrarily defined as containing all events later than period 9. This includes syntheses of aromatic amino acids and fatty acids along the "extramitochondrial pathway," here numbered "1".

Some branches are in a period earlier than one of their upstream branches because of question mark optimization (see above). This is the case for optimizations of characters 4, 8, 18, and 19. Questions marks in character 8 (carboxylations using biotin) imply that terminal branches dALA, dCYS3, dSER, dCYS1, dGLY, dTHR2, and branches Q and S are in period 6 while their upstream branches K, O, and R are in period 7. Question marks of character 18 (β-oxidation sequence repeat) in KC1 and sLYS3 place terminals KC1 and sLYS3 in period 6 while their upstream nodes L and U are in period 7. Question marks in character 19 (β-oxidations) for pathways of clade H (except KC2, dLYS, dILE, dVAL, and synthesis and degradation of fatty acids) place the gain of β-oxidation onto the branch H. That branch is assigned to period 6 because of the new gain of β-oxidation (character 5, already available from period 2, is not new). All pathways that exhibit a question mark for character 19 are assigned one period earlier. Branch M, all members of clade W, branch N, and all members of clade X are, therefore, assigned to period 5, which is earlier than period 6 of H. The same reasoning is followed for character 4, which is coded "1" in all pathways of asparagine and glutamine and coded "?" in all others. All others are assigned one period earlier. Question marks explain why one can find new homology of type II in a branch without changing the period. The assignment of the period involves one period forward because of the new type II homology and one period earlier because of a question mark for another character. For example, within the clade B, the node on which character "5" occurs remains in period 1 (pink) because one step forward for the new homology "5" and one step earlier because of question marks on character "4" for dARG, sPRO1, dPRO. Another example is the node E, which remains in period 3 despite the new homology "11".

Discussion

The rise of different enzymes and enzymatic functions are plotted onto the tree (fig. 1), and therefore each event is associated to a given period. Then, the different parts of a given pathway can belong to different periods according to the enzymes involved. This allows a first view on the temporal development of each pathway, as shown in figure 2. For instance, degradation of valine (dVAL) starts in period 2 and ends in period 7.

Amino acid metabolism is basal in the phylogenetic tree, and all other metabolisms develop from that background. The priority of amino acid degradation over syntheses depends on the amino acid. Full degradations are earlier than full syntheses for amino acids of Cordón's groups I (Asp and Asn) and II (Glu, Gln, Arg, and Pro, excepting proline, for which full synthesis is possible in period 2 while degradation appears in period 3) and for aromatic amino acids (His, Tyr, and Phe), contradicting Cordón's views on priority of aromatic amino acid syntheses over their degradations. For amino acids of Cordón's groups III (Ala, Cys, Gly, and Ser) and IV (Leu, Val, Ile, Lys, Thr, and Met) full syntheses are possible before full degradations (except for Trp), contradicting Cordón's views. In groups I and II, degradation and synthesis tend to be reverse processes, whereas in groups III and IV, they are different pathways. This observation is also true for aromatic amino acids, as the syntheses are different and very late in comparison to degradations. Our results do not fully corroborate predictions of Horowitz (1945) and Cordón (1990) according to which amino acid catabolic pathways develop forwards and which anabolic pathways develop backwards. Indeed, dPRO, dHIS, dALA, dTRP, and dMET develop backwards, whereas none of the amino acid anabolisms are found to develop backwards, except sLYS. Furthermore, in some cases, opportunistic late connections of pathways lead to complex developments neither forwards nor backwards as in sLEU, sVAL, dILE, sGLY, sTHR, sMET, dLEU, and dGLY. Horowitz (1945) and Cordón's (1990) predictions were not based on a data matrix, but (1) on the hypothesis that amino acids of abiotic origin might have been one of the very first sources of energy and structural components available to primitive forms of life (which has been highly corroborated since), (2) on a theoretical model of selective pressure acting on protocells during the course of biochemical evolution, and (3) on consistent comparative reasoning from structures of metabolic pathways and their compounds (see Cunchillos and Lecointre [2000]). No computerized parsimony was used, and it is not surprising that a formalized reconstruction algorithm using parsimony can deal with complex issues of homoplastic late opportunistic pathway connections better than a human mind can.

Metabolism of fatty acids and saccharides develop after the full development of metabolism of amino acids of groups I and II, and they are associated with the anabolism of amino acids of groups III and IV. Sugar metabolisms are within the clade N, with anabolic pathways of the amino acids of group III, as predicted by Cordón (1990). Syntheses of aromatic amino acids are branched within sugar metabolism. From the present data, it is neither possible to draw conclusions about forward or backwards development of fatty acid and sugar metabolisms nor possible to draw conclusions about priority of either their anabolism or their catabolism. Note that the extramitochondrial synthesis of fatty acids precedes (period 7) the intramitochondrial synthesis (period 9).

Period 6 is very rich, a period during which events are difficult to order and the two portions of the Krebs cycle take place after the setting of metabolisms of amino acids of groups I and II. Interestingly, one portion of the Krebs cycle has a catabolic origin and the other an anabolic origin: KC2 is associated with catabolism of amino acids of groups III and IV, and KC1 is associated with anabolism of the same groups. Despite the late branching of glycolysis in the tree, it is difficult to order the rise of full glycolysis and the rise of full Krebs cycle: they are all embedded within period 6. So the views of Meléndez-Hevia et al. (1996), that glycolysis must have preceded the Krebs cycle because it provided energy, are neither confirmed nor contradicted. However, relative position of a catabolic pathway must not only be assessed in terms of energy but also in the light of the compounds provided. Availability of glucose in early abiotic environments is much more speculative than availability of amino acids (Cunchillos and Lecointre 2000, 2002). For that reason, because the Krebs cycle is derived from amino acid metabolisms of groups III and IV, the hypothesis that the Krebs cycle arose earlier than glycolysis is much more likely.

From the present data, the ordering of gluconeogenesis compared with glycolysis is not possible: both are in period 6. Pentose-phosphate and Calvin cycles arose later (periods 7 and 8, respectively) than glycolysis and gluconeogenesis, as predicted by Cordón (1990). The urea cycle arises as early as the synthesis of arginine in period 4.

The present inferences partially corroborate Cordón's scenario, which, after all, is not surprising, because they are only based on the comparative anatomy of metabolic pathways, whereas Cordón incorporated much biological input. It is very interesting that his scenario was embedded into a consistent and well-documented DNA-free conception of natural selection of enzymatic specificity, leading to the evolution of metabolism through natural selection of proteins without any DNA information storage. This is one of the reasons why the present work has nothing to do with theories of early information storage: to test Cordón's scenario, such storage was not necessary. Concerning information storage in nucleic acids, it should be stressed that the compounds involved in the pathways studied here are basic elements for construction of nucleic acids of metabolic origin. Therefore, those nucleic acids arose later than the pathways. Biological purines and pyrimidines are the products of several already complex molecules; they are never synthesized as such in cells, because their parts are already linked to chemically different compounds such as riboses. From the evolutionary point of view, they are later chemical compounds than all the metabolic pathways compared here. The question of knowing whether there might have been DNAs or RNAs of nonmetabolic origin encoding "information" for enzymatic activities studied cannot be answered here. Cordón's scenario for the rise of protein complexity and enzymatic specificity by natural

selection without invoking DNA/RNA "information" storage is more parsimonious than explanations that require such a storage and is well worth further consideration for positive input in the increasing criticisms of the current uses of notions of genetic control and information (Kupiec and Sonigo 2000; Segal 2003).

Acknowledgments

We thank Patrick Tort for help to C.C., and John Macdonald and Craig Marshall for help in improving the manuscript. The Muséum National d'Histoire Naturelle and the company Saint-Gobain are acknowledged for support.

Literature Cited

Cordón, F. 1990. Tratado evolucionista de biologia. Aguilar, Madrid.

Cunchillos, C. and G. Lecointre. 2000. L'histoire du catabolisme des acides aminés aliphatiques inférée par l'analyse cladistique de deux nouveaux types de caractères : l'enzyme et la réaction enzymatique. Pp. 87–106 *in* V. Barriel and T. Bourgoin, eds. Caractères. Biosystema 18, Société Française de Systématique, Paris.

———. 2002. Early steps of metabolism evolution inferred by cladistic analysis of the structure of amino acid catabolic pathways. C. R. Biologies **325**:119–129.

———. 2003. Evolution of amino acid metabolism inferred through cladistic analysis. J. Biol. Chem. **278**:47960–47970.

de Pinna, M. C. C. 1991. Concepts and tests of homology in the cladistic paradigm. Cladistics **7**:367–394.

Enzyme Nomenclature. 1973. Recommendations (1972) of the International Union of pure and applied chemistry and the International Union of biochemistry. Elsevier Scientific Publishing, Amsterdam.

Goloboff, P. 1998. Nona: computer program and software. Published by the author, Tücuman, Argentina.

Horowitz, N. H. 1945. On the evolution of biochemical syntheses. Proc. Natl. Acad. Sci. USA **31**:153–157.

Kupiec, J. J. and P. Sonigo. 2000. Ni Dieu ni gène. Seuil, Paris.

Meléndez-Hevia, E., T. G. Waddell, and M. Cascante. 1996. The puzzle of the Krebs citric acid cycle: assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. J. Mol. Evol. **43**:293–303.

Nixon, K. C. 1999*a*. Winclada (BETA). Version 0.9.9. Published by the author, Ithaca, New York.

———. 1999*b*. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics **15**:407–414.

Segal, J. 2003. Le zéro et le un, histoire de la notion scientifique d'information au 20ème siècle. Syllepse, Paris.

Geoffrey McFadden, Associate Editor

Accepted August 12, 2004