J Mol Evol (1997) 44:552–572

JOURNAL OF MOLECULAR EVOLUTION

Molecular Evolutionary Analysis of the Thiamine-Diphosphate-Dependent Enzyme, Transketolase

Gerhard Schenk,[1] Roy Layfield,[2] Judith M. Candy,[3] Ronald G. Duggleby,[1] Peter F. Nixon[1]

[1] Department of Biochemistry, Centre for Protein Structure, Function and Engineering, The University of Queensland, St. Lucia, QLD 4072, Australia
[2] Centre for Molecular and Cellular Biology, The University of Queensland, St. Lucia, QLD 4072, Australia
[3] Centre for Molecular Biotechnology, School of Life Sciences, Queensland University of Technology, Gardens Point, Brisbane, QLD 4001, Australia

Received: 13 September 1995 / Accepted: 14 November 1996

Abstract. Members of the transketolase group of thiamine-diphosphate-dependent enzymes from 17 different organisms including mammals, yeast, bacteria, and plants have been used for phylogenetic reconstruction. Alignment of the amino acid and DNA sequences for 21 transketolase enzymes and one putative transketolase reveals a number of highly conserved regions and invariant residues that are of predicted importance for enzyme activity, based on the crystal structure of yeast transketolase. One particular sequence of 36 residues has some similarities to the nucleotide-binding motif and we designate it as the transketolase motif. We report further evidence that the recP protein from *Streptococcus pneumoniae* might be a transketolase and we list a number of invariant residues which might be involved in substrate binding. Phylogenies derived from the nucleotide and the amino acid sequences by various methods show a conventional clustering for mammalian, plant, and gram-negative bacterial transketolases. The branching order of the gram-positive bacteria could not be inferred reliably. The formaldehyde transketolase (sometimes known as dihydroxyacetone synthase) of the yeast *Hansenula polymorpha* appears to be orthologous to the mammalian enzymes but paralogous to the other yeast transketolases. The occurrence of more than one transketolase gene in some organisms is consistent with several gene duplications. The high degree of similarity in functionally important residues and the fact that the same kinetic mechanism is applicable to all characterized transketolase enzymes is consistent with the proposition that they are all derived from one common ancestral gene. Transketolase appears to be an ancient enzyme that has evolved slowly and might serve as a model for a molecular clock, at least within the mammalian clade.

Key words: Transketolase — Thiamine diphosphate — Transketolase motif — Evolution — Phylogenetic trees — Molecular clock

Introduction

Transketolase (EC 2.2.1.1) catalyzes two separate reactions of the nonoxidative branch of the pentose-phosphate pathway which, along with the enzyme transaldolase, provides the link between this pathway and glycolysis. This enables the recycling of pentose sugars under conditions where NADPH production is required for reductive biosynthesis. Transketolase (Datta and Racker 1961) is one of at least 14 enzymes requiring thiamine diphosphate (ThDP) and divalent cations for catalytic activity but is the only cytosolic ThDP-dependent enzyme in mammalian systems.

The enzyme has been purified from several organisms (Kochetov 1982), and the functional form isolated from baker's yeast (de la Haba 1955; Srere 1958) is a homodimer of 74-kDa subunits, each of which contains a

Correspondence to: P.F. Nixon

molecule of ThDP (Kochetov 1986). While the same appears to be true for the rat (Horecker et al. 1953), human (Heinrich and Wiss 1971), some plant (Bernacchia et al. 1995), and other yeast transketolases (Kiely et al. 1969; Waites and Quayle 1981), the enzyme apparnetly exists in monomeric form in spinach (Villafranca and Axelrod 1971) and as a tetrameric form in both *Candida boidinii* (Kato et al. 1982) and pig (Philippov et al. 1980). More recent reports have discounted the latter, demonstrating that the pig enzyme is also a homodimer (Voskoboev and Gritsenko 1981) and while it is possible that the *C. boidinii* enzyme is truly tetrameric, consideration must be given to the fact that transketolases from related *Candida* strains are confirmed dimers (Klein and Brand 1977). Yeast transketolase has been well characterized with respect to its chemical and catalytic properties (Kochetov 1986), and the three-dimensional structure of transketolase from *Saccharomyces cerevisiae* has recently been determined to 2.0-Å resolution (Nikkola et al. 1994).

Genes coding for dihydroxyacetone synthase (the specific transketolase of the methanol-utilizing yeast, *Hansenula polymorpha*) (Janowicz et al. 1985) and transketolase from mammals (Abedinia et al. 1992; McCool et al. 1993), yeasts (Sundström et al. 1993; Metzger and Hollenberg 1994), bacteria (e.g., Sprenger 1993; Schäferjohann et al. 1993; Chen et al. 1991), and plants (Bernacchia et al. 1995; Teige et al. 1995: GenBank) have been cloned and sequenced. A second transketolase, distinct from but closely related to the first, has been cloned in both *Escherichia coli* (Iida et al. 1993) and *S. cerevisiae* (Schaaff-Gerstenschläger and Zimmermann 1993), although any unique functions and metabolic implications of these enzymes remain unclear. In the dessication-tolerant plant *Craterostigma plantagineum*, three distinct forms of transketolase have been cloned (Bernacchia et al. 1995). One form (Cpl3) is expressed constitutively, while the remaining two forms (Cpl7 and Cpl10) appear to be involved specifically in the rehydration process.

A broad range of substrates has been reported for the transketolases from yeasts, plants, and bacteria. Transketolase from *S. cerevisiae* can utilize sugars such as D-xylulose 5-phosphate, D-sedoheptulose 7-phosphate, D-fructose 6-phosphate, and D-erythrulose 4-phosphate, as well as dihydroxyacetone phosphate, dihydroxyacetone, and hydroxypyruvate as donors of glycoaldehyde. Acceptor substrates include D-ribose 5-phosphate, D-glyceraldehyde 3-phosphate, D-erythrose 4-phosphate, and glycoaldehyde (Kochetov 1986). The transketolase of *H. polymorpha* (dihydroxyacetone synthase) and also that of *C. boidinii* display an even larger range of substrate utilization, including all of the above-mentioned substrates as well as formaldehyde and acetaldehyde as acceptors (Kato et al. 1982; Janowicz et al. 1985). Transketolase from spinach leaves has a similar range of substrate specificity to the *S. cerevisiae* enzyme (Villafranca and Axelrod 1971) and can also catalyze the transfer of a two-carbon fragment from hydroxypyruvate to non-phosphorylated acceptor sugars. Recent findings indicate that transketolase purified from *E. coli* displays kinetic properties similar to those of the yeast and plant transketolases (Sprenger et al. 1995). In contrast, mammalian transketolases display a higher degree of substrate specificity, with only D-xylulose 5-phosphate, D-fructose 6-phosphate, and D-sedoheptulose 7-phosphate as donors, and D-ribose 5-phosphate, D-erythrose 4-phosphate, D-glyceraldehyde 3-phosphate, and glycoaldehyde as acceptor substrates (Waltham 1990; Schenk 1996).

Clearly there is evidence for functional diversity among members of the transketolase family, which could reflect different active-site fine structure. Given the range of sequences now available for study, it is possible to attempt the phylogenetic reconstruction of this family. Several reports have inferred the phylogeny of transketolase on the basis of dendograms calculated from relatively few sequences (Reizer et al. 1993; Sundström et al. 1993; Van Den Bergh et al. 1996), providing an interesting but limited insight into the evolution of this enzyme.

Here, we report the detailed analysis of two partial and 20 complete DNA sequences and translated amino acid sequences derived from mammals, yeast, bacteria, and plants, and thus provide a more detailed view of the divergent evolution of this family. Our investigation includes the recP gene from *S. pneumoniae* (Radnis et al. 1990). Previous studies have suggested that recP might code for a transketolase even though the product of this gene is thought to be involved in recombination (Sundström et al. 1993). We provide further evidence for this suggestion and identify a transketolase motif. Phylogenetic reconstructions are consistent with the proposition that transketolase represents an ancient ''housekeeping'' enzyme with a complex evolutionary history.

Methods

Sequence Alignments. The deduced amino acid sequences encoded by 21 different transketolase genes of *S. cerevisiae* (Schaaf-Gersternschläger and Zimmermann 1993; Sundström et al. 1993), *Pichia stipitis* (Metzger and Hollenberg 1994), *E. coli* (Sprenger 1993; Iida et al. 1993), *Alcaligenes eutrophus* (Schäferjohann et al. 1993), *Haemophilus influenzae* Rd (Fleischmann et al. 1995), *Rhodobacter sphaeroides* (Chen et al. 1991), *Rhodobacter capsulatus* (de Sury D'Aspremont et al. 1996), *Xanthobacter flavus* (Van Den Bergh et al. 1996), *Mycobacterium leprae* (Smith 1994: GenBank), *Mycoplasma genitalium* (Fraser et al. 1995), *Mus musculus* (mouse) (Schimmer et al. 1996), *Rattus norvegicus* (rat) (Kim et al. 1994: GenBank), adult *Homo sapiens* (this paper), fetal *Homo sapiens* (Jung et al. 1993: GenBank), *Solanum tuberosum* (potato) (Teige et al. 1995: GenBank), and *C. plantagineum* (Bernacchia et al. 1995), the formaldehyde transketolase of *H. polymorpha* (Janowicz et al. 1985), and the sequence of the putative transketolase encoded by the recP gene from *S. pneumoniae* (Radnis et al. 1990) were obtained from GenBank. These genes include representatives of mammals, yeast, bacteria, and plants. Additional

Table 1. Sequences used in this study[a]

Phyla	Code	Source	Accession no.
Mammalia	Hsa(ad)	*Homo sapiens* (adult)	U55017
	Hsa(ft)	*Homo sapiens* (fetal)	L12711
	Rno	*Rattus norvegicus* (rat)	U09256
	Mmu	*Mus musculus* (mouse)	U05809
Plant	Cpl7	*Craterostigma plantagineum*	Z46648
	Cpl10	*Craterostigma plantagineum*	Z46647
	Cpl3	*Craterostigma plantagineum*	Z46646
	Stu	*Solanum tuberosum* (potato)	Z50099
Yeast	Hpo	*Hansenula polymorpha*	X02424
	Sce1	*Saccharomyces cerevisiae*	X73224
	Sce2	*Saccharomyces cerevisiae*	X73532
	Pst	*Pichia stipitis*	Z26486
Bacteria	Eco1	*Escherichia coli* (γ-subdivision)	X68025
	Eco2	*Escherichia coli* (γ-subdivision)	D12473
	Rsp	*Rhodobacter sphaeroides* (α-subdivision)	M68914
	Rca	*Rhodobacter capsulatas* (α-subdivision)	L48803
	Xfl	*Xanthobacter flavus* (β-subdivision)	U29134
	Aeu	*Alcaligenes eutrophus* (β-subdivision)	M68905
	Hin	*Haemophilus influenzae* (γ-subdivision)	L45661
	Mle	*Mycobacterium leprae*	U00013
	Mge	*Mycoplasma genitalium*	U39686
	Spn	*Streptococcus pneumoniae* (recP)	M31296

[a] The codes used in figures and text, the source of each sequence and its GenBank accession number, plus phyla are listed. In the case of gram-negative bacteria, the subdivision is indicated in parenthesis. Sequences are listed in the same order as the alignment in Fig. 1

sequences encoding the human transketolase were also available from GenBank, including both a full-length sequence (McCool et al. 1993) and a partial sequence reported previously by us (Abedinia et al. 1992). Our laboratory has since cloned and sequenced a full-length cDNA for human transketolase which has some differences from that of McCool et al. (see Discussion). This sequence has now been deposited with the GenBank database and is the human sequence used for the purpose of alignment and phylogenetic reconstruction reported here.

Table 1 lists the sources of transketolase sequences used in our analyses. The entire translated peptide sequence was used for analysis in all cases except for the plant sequences Cpl3 and Stu for which only partial sequences (missing N-terminal residues) were available. Initial alignments were performed using Clustal W software (Thompson et al. 1994). The initial alignment was further refined by eye, bearing in mind the secondary structure of the yeast transketolase derived from crystallographic studies (Nikkola et al. 1994). Gaps were introduced into the sequences where necessary to improve the overall alignment, especially to allow for the larger size of dihydroxyacetone synthase from *H. polymorpha* (Janowicz et al. 1985) and blocks of divergence between the mammalian and other transketolases. Care was taken to ensure that no gaps disrupted the secondary structure of Sce1. Unless otherwise stated, the numbering of residues is derived from the *S. cerevisiae* transketolase Sce1 (Nikkola et al. 1994).

Phylogenetic Analysis. The nucleotide and the amino acid compositions of the 22 sequences in our comparison were assessed using the MEGA software package (version 1.0) (Kumar et al. 1993). Transition/transversion ratios were derived from pairwise comparisons between all sequences using the same program. The nucleotide and amino acid composition matrices were subject to contingency table tests in order to determine the heterogeneity/homogeneity of the data sets. Stationarity was checked according to the method of Saccone et al. (1990), assuming $\chi^2 \leqslant 1.5$ as the necessary stationarity criterion.

Using the MEGA software package (Kumar et al. 1993) and the Phylo Win program (Galtier and Gouy 1995) sequence distance matrices were established in pairwise comparisons for both character sets using a variety of algorithms: p-distance and Gamma distance ($a = 2$) (see Kumar et al. 1993) for protein sequences and distances derived by calculating the p-distance and applying the algorithms of Jukes and Cantor (1969), Kimura (1980), Galtier and Gouy (1995) and Lockhart et al. (1994) for DNA sequences. Euclidean distances were calculated in order to distinguish between the phylogenetic and the compositional signal (Lockhart et al. 1994).

All distance-matrix-based phylogenies were derived using the neighbor-joining (NJ) method (Saitou and Nei 1987) and the minimum evolution (ME) approach (Rzhetsky and Nei 1992). (The program METREE [Rzhetsky and Nei 1994] was kindly provided by Dr M. Nei, Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, USA.) Maximum likelihood (ML) (Felsenstein 1981) and maximum parsimony (MP) (Fitch 1971) trees were analyzed using the PHYLIP package (Felsenstein 1989) and the Phylo Win program (Galtier and Gouy 1995), respectively. NJ, ME, and MP trees were subjected to bootstrapping (1,000 replicates) (Felsenstein 1985). All trees were unrooted and no outgroup was nominated.

Variations in sequence lengths (Fig. 1) resulted in some gaps. Tree topologies were obtained from data sets after removal of gaps in pairwise comparisons and from data sets after global removal of gaps. Additionally, we analyzed an unambiguously aligned subset of regions. (The selected positions in the amino acid alignment corresponded to Sce1 residues 4–140, 149–268, 294–333, 339–349, 356–397, 409–519, 555–605, 629–639, and 640–664.)

The three dimensional structure of transketolase from the yeast *S. cerevisiae* has been reported previously (Nikkola et al. 1994). The secondary-structure elements are indicated in the alignment (Fig. 1). Secondary-structure predictions of the remaining sequences were derived by use of the programs PEPTIDESTRUCTURE (Jameson and Wolf 1988) and PREDICTPROTEIN (Rost and Sander 1993).

Results

Sequence Comparisons

Alignment of the amino acid sequences of 19 entire and two partial (Cpl3 and Stu) transketolase enzymes and an entire putative transketolase (corrected recP gene product from *S. pneumoniae*) is shown in Fig. 1. Previously, a high similarity between recP and bacterial transketolases was shown (Reizer et al. 1993). Correction of several assumed frame shifts in the N-terminal domain allowed Sundström to confirm that recP could be a transketolase (sites of assumed frame shifts are indicated by an X in Fig. 1) (Sunderström et al. 1993). We have discovered another potential frame shift that supports this suggestion (see below).

A high level of similarity was apparent between all studied sequences, with 50 totally invariant residues (Table 2A). Additionally, residues at many other positions are highly conserved throughout most of the pro-

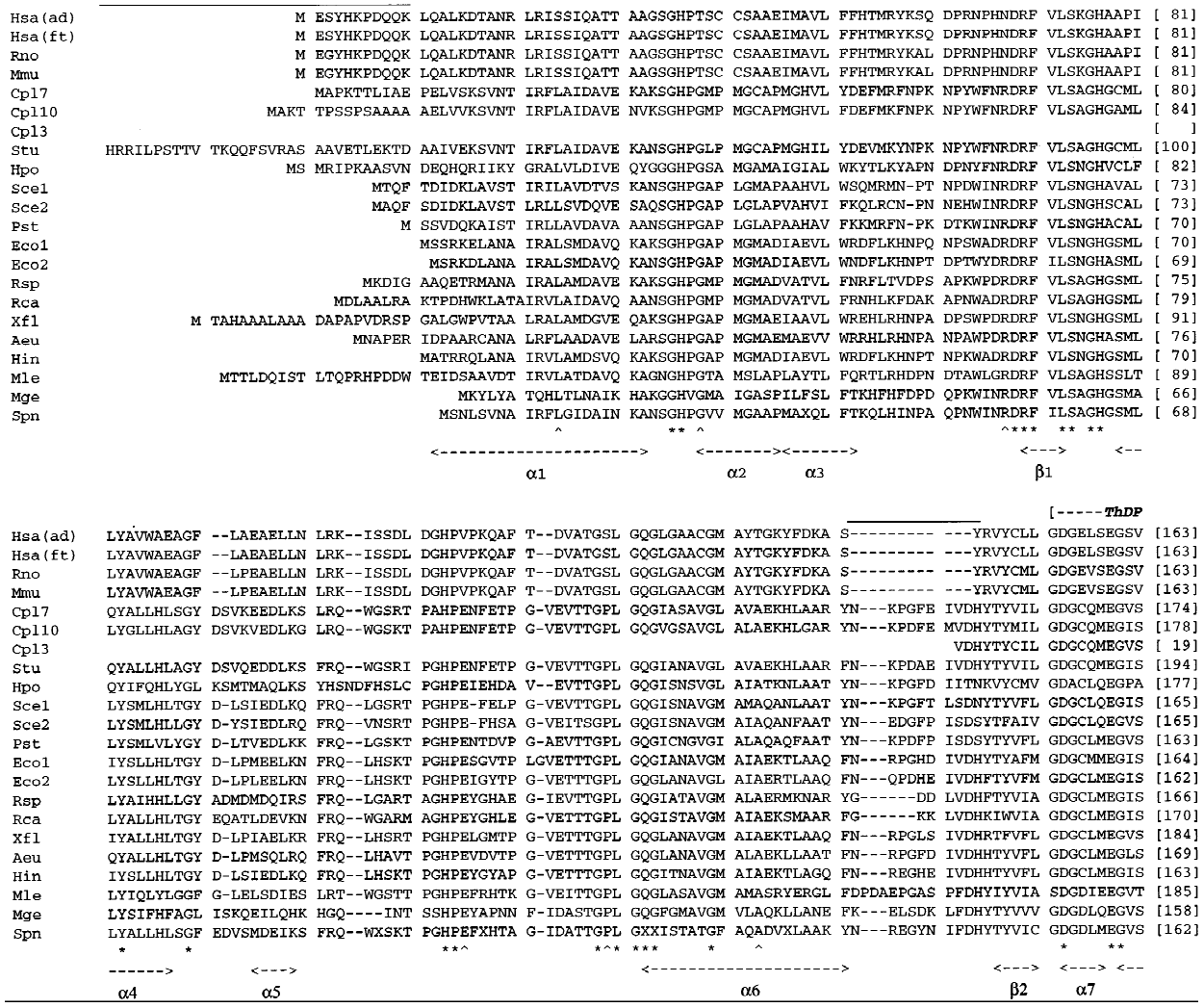

Fig. 1. Sequences of 21 transketolases and one putative transketolase (Spn) were aligned using ClustalW. The sources of the alignment and abbreviations are listed in Table 1. *Gaps* (−) were introduced to improve the overall alignment. A *colon* (:) denotes differences between the fetal (Hsa[ft]) and human (Hsa[ad]) sequences. Residues resulting from correction of an apparent sequencing error in the recP sequence of *S. pneumoniae* (see text) are *underlined and in italics.* Ten additional frame shifts in Spn (recP) suggested by Sundström are denoted by an *X* (Sundström et al.1993). Below the alignment, invariant residues are marked by an *asterisk* (*); those residues that are invariant in all but the mammalian sequences are marked by an *arrowhead* (^); and the secondary structure of Sce1 is indicated by *arrows. Numbers* of the residues in each sequence are shown in parenthesis. The regions not indicated by *bars* on top of the alignment were selected for separate phylogenetic analysis in addition to analyses of the complete sequence (see Methods).

tein, including 24 positions at which only the mammalian sequences differ from the others (Table 2B). Moreover, at these positions the amino acids are invariant within the mammalian sequences.

Structural studies have discerned three distinct domains in the transketolase monomer of *S. cerevisiae,* these being the N-terminal, middle, and C-terminal domains (Lindqvist et al. 1992; Nikkola et al. 1994). The N-terminal domain forms almost half of a subunit (residues 1–322) and contains the highly conserved ThDP-binding motif (Hawkins et al. 1989). Homology between the transketolase sequences shown here is greatest in this N-terminal domain, followed by the middle domain (residues 323–538), which contains a large and extremely well-conserved stretch of amino acids with the consensus sequence (S/T)H(D/C)(S/G)X_3**G**X_2**G**P(S/T)(Q/H)X_9RX_8(R/Y)PX**D** where X denotes any amino acid; the residues shown in bold are considered below. Previously, the C-terminal part of this sequence has been identified as being similar to a nucleotide-binding site (Abedinia et al. 1992) with the fingerprint sequence GXGXXG$X_{17–20}$D (Bernard et al. 1995; Kochhar et al. 1992), which has been reported for several enzymes such as malate dehydrogenase (Birktoft et al. 1989), glyceraldehyde 3-phosphate dehydrogenase (Skarzynski et al. 1987; Murthy et al. 1980), alcohol dehydrogenase (Eklund et al. 1976), D- and L-lactate dehydrogenases (Bernard et al. 1995; Abad-Zapatero et al. 1987), and formate dehydrogenase (Lamzin et al. 1992). It is thought that this motif is an absolute requirement in order for proteins

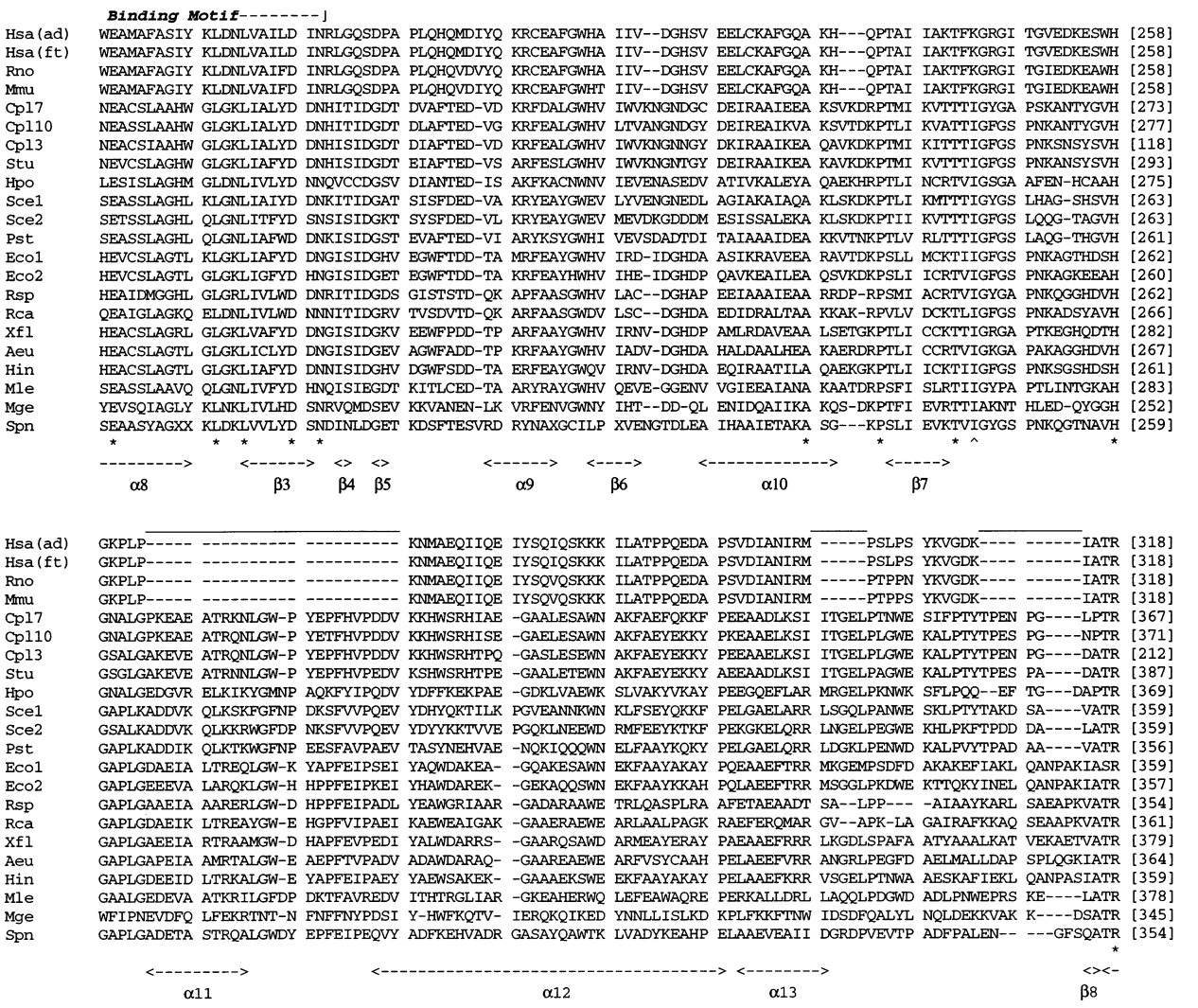

Fig. 1. Continued.

to have a βαβ-fold (Fig. 2A) with ADP-binding properties (Wierenga et al. 1986). However, slight deviations from this motif can still lead to the formation of a βαβ-fold (Lamzin et al. 1992). Only the second glycine and the aspartate of the NADH-binding motif seem to be totally invariant in NADH-dependent enzymes (Fig. 2A). Site-directed mutagenesis studies on this particular aspartate residue in D-lactate dehydrogenase revealed that this residue discriminates between NADH and NADPH binding (Bernard et al. 1995), consistent with its binding to the 2′-hydroxyl group of the adenosine ribose of NADH. In some enzymes containing the nucleotide-binding motif—for example, formate dehydrogenase—structural studies have shown that it is the ribose phosphate moieties of the pyridine nucleotide, alone, which bind across the βαβ-fold identified by the motif (Lamzin et al. 1992) (Fig. 2A). The characters in bold in the consensus sequence indicate residues found in the fingerprint sequence mentioned above. Most transketolases contain the first glycine of the fingerprint sequence although in mammals it is replaced by a serine, in *H. polymorpha* by an asparagine, in *M. genitalium* by a glutamine, and in *S. pneumoniae* by an alanine (Figs. 1 and 4). The number of variable residues between the last glycine and the aspartate is higher in transketolases (24) (Fig. 4) than in the fingerprint sequence (17–20). The aspartate residue (Asp503), which was shown to be required for the formation of a βαβ-fold in D-lactate dehydrogenase (Bernard et al. 1995), is invariant in all but the recP sequence from *S. pneumoniae* (Spn). However, correction of an assumed frameshift sequencing error leads to the invariant aspartate and to a threonine that is present in 17 out of the 22 sequences. (The six amino acids changed by this frame shift are underlined and italicized in Fig. 1.)

The crystal structure of Sce1 (Nikkola et al. 1994) indicates that transketolase contains a structure that is similar to the βαβ-fold reported for NADH-binding enzymes (Wierenga et al. 1986); however, the loop between the first β-strand and the α-helix is extended (Fig. 2B). Although transketolase is not an enzyme that requires the binding of nucleotides for its function, one

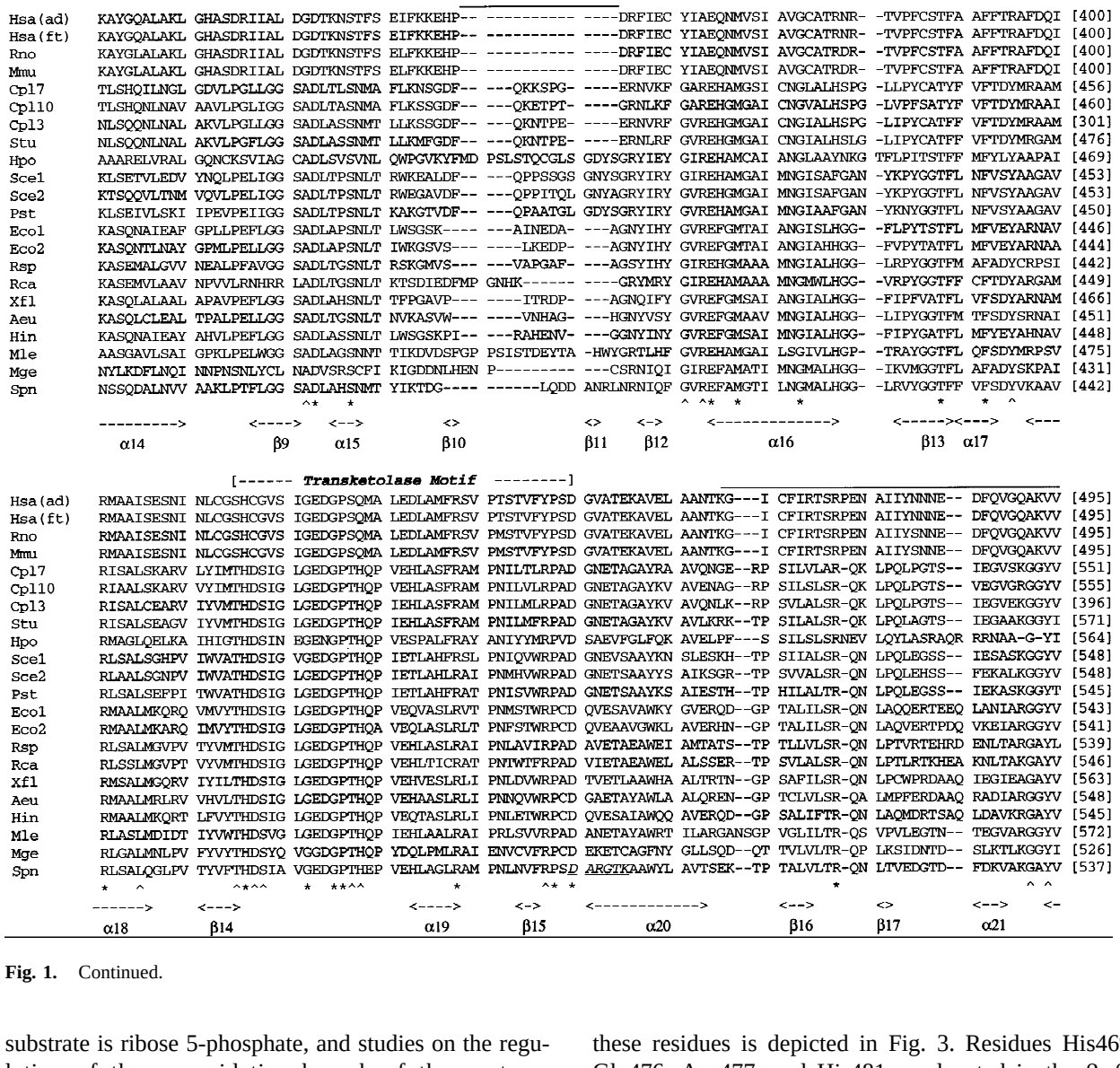

Fig. 1. Continued.

substrate is ribose 5-phosphate, and studies on the regulation of the nonoxidative branch of the pentose-phosphate pathway indicate that ADP and phosphoribosyl pyrophosphate (PrPP) competitively inhibit the activity of transketolase (Hosomi et al. 1989). Additionally, transketolase can be obtained in pure form by affinity chromatography using the dye Cibacron Blue F3G-A (Waltham 1990; Booth 1991). Although the interactions between dye and protein are not fully understood it has been shown that liver alcohol dehydrogenase binds the dye in the βαβ-fold in a manner similar to that of nucleotides (Biellmann et al. 1979). These results suggest that transketolase can bind the ribose phosphate moiety of ADP and PrPP, and it contains a βαβ-fold very similar to the nucleotide-binding fold of NADH-dependent enzymes.

A potential substrate channel has been identified in the structure of yeast transketolase (Nikkola et al. 1994), crystallized in the absence of substrate. Table 3 lists all residues identified from the crystal structure as contributing to potential substrate binding. The arrangement of these residues is depicted in Fig. 3. Residues His469, Glu476, Asp477, and His481 are located in the βαβ-fold. Analysis of the nucleotide binding in formate dehydrogenase shows that the invariant aspartate in the fingerprint sequence could bind to the 2′-hydroxyl group of the adenosine ribose (Lamzin et al. 1992) (Fig. 2A). By analogy Asp503 may be involved in binding the acceptor substrate in transketolase (Fig. 2B). However, the crystal structure of Sce1 shows that this could be the case only if significant conformational changes occur upon substrate binding (Fig. 3).

Searching the databases with the consensus sequence mentioned above and allowing for up to three mismatches retrieved all the transketolases and only transketolases. We suggest that the term ‘‘transketolase motif’’ appropriately describes this consensus sequence (Fig. 4). While the ThDP-binding motif is common to all thiamine-dependent enzymes, the transketolase motif unambiguously identifies all transketolases within this group. This definition would classify recP from *S. pneumoniae* as a transketolase. Functional constraints on the

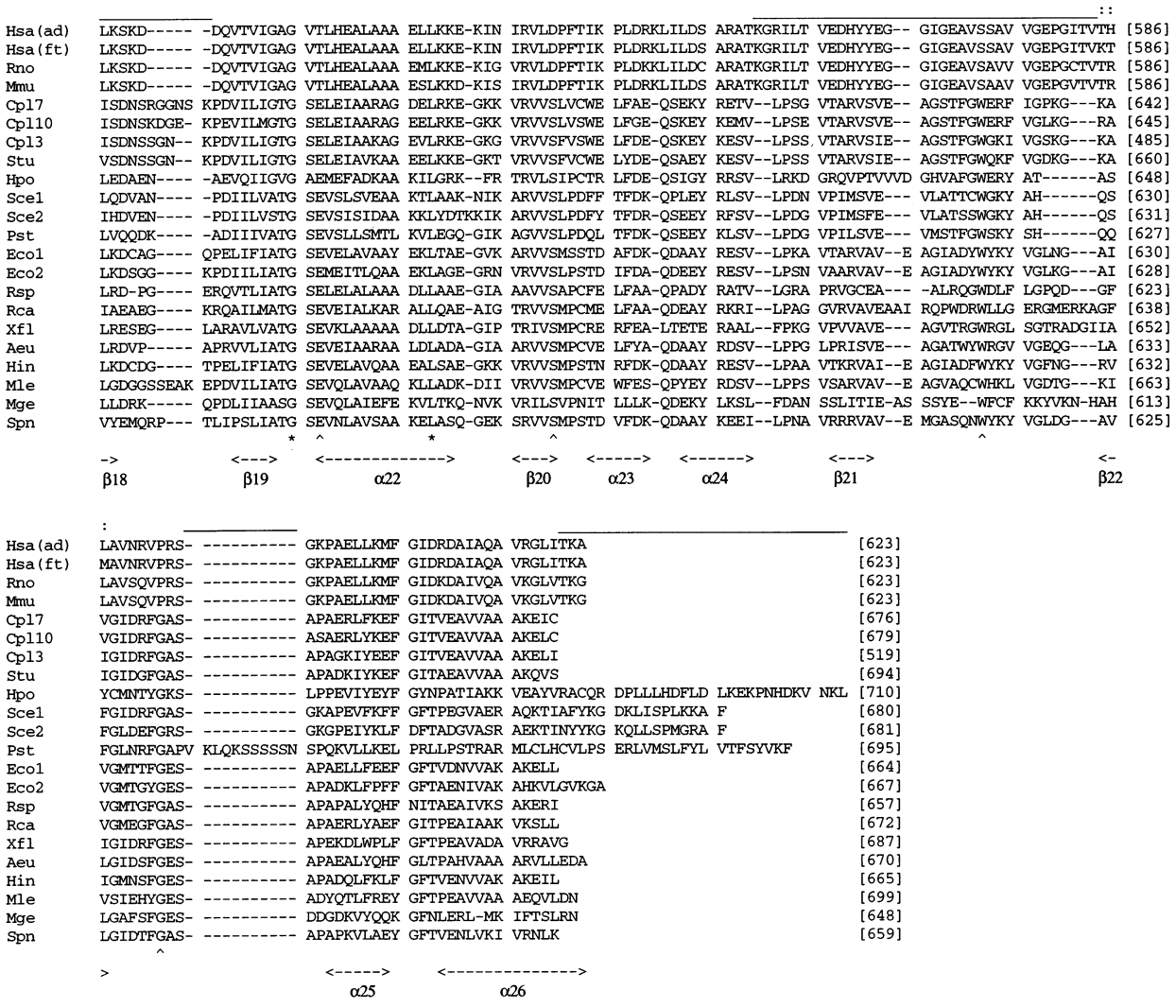

Fig. 1. Continued.

Table 2A. Residues that are totally invariant among 21 transketolase sequences and one putative transketolase[a]

Gly29	Gly82	Glu162	His263	His469
His30	His103	Gly163	Arg359	Gly475
Asp61	Pro104	Glu167	Asp382	Gly478
Arg62	Gly116	Leu177	Ser386	Pro479
Phe63	Leu118	Leu180	Glu418	Arg491
Leu65	Gly119	Asp185	Met421	Pro501
Ser66	Gln120	Asn187	Gly428	Asp503
Gly68	Gly121	Ala234	Thr441	Arg528
His69	Gly127	Pro241	Phe445	Gly563
Tyr75	Asp157	Thr248	Arg454	Leu573

diversification of this motif through evolution are reflected by comparing crystallographic data (Lindqvist et al. 1992; Nikkola et al. 1994), biochemical studies on the regulation of the nonoxidative branch of the pentose-phosphate pathway (Hosomi et al. 1989), and the very high level of sequence conservation through all four compared phyla.

There is a distinct lack of sequence homology in the C-terminal domain (residues 539–680), although this domain is well conserved within kingdoms. This might reflect some functional differences that evolved in a kingdom-dependent manner (e.g., substrate specificity, subunit interactions).

At only 623 residues in length, the mammalian sequences are the smallest of the observed transketolase enzymes, being 87 residues shorter than the largest sequence from *H. polymorpha.* This is a considerable size difference and might be related to the more specific substrate utilization found in mammalian transketolases (Waltham 1990). It is interesting to note that a comparison of the fetal human (Jung et al. 1993; GenBank) and our own human transketolase sequences reveals several base differences, resulting in three putative amino acid substitutions (Lys585 to Thr, Thr586 to His, and Met587 to Leu, with residues numbered according to the human sequences) (Fig. 1). It is not clear why there should be any difference between the two sequences, and all three

Table 2B. Residues that differ only between the mammalian sequences and all others (note that the listed amino acids are invariant within each group of sequences—nonmammalian and mammalian)

Nonmammalian	Mammalian	Nonmammalian	Mammalian	Nonmammalian	Mammalian
Leu18	Ser	Gly415	Tyr	His481	Gln
Gly32	Thr	Arg417	Ala	Arg500	Tyr
Arg60	Asn	Tyr448	Arg	Gly545	Ala
Glu105	Val	Leu458	Ile	Tyr547	Ala
Pro117	Ser	Thr468	Ser	Glu565	Thr
Ala131	Thr	Asp470	Cys	Ser587	Asp
Ile260	Lys	Ser471	Gly	Trp623	Ser
Ala381	Gly	Thr480	Ser	Gly637	Pro

substitutions occur in noncritical residues, suggesting that there would be little if any effect on enzyme activity. Since all of these substitutions occur in a single block, we wonder whether compounded sequencing errors may provide a simple explanation for the discrepancy. However, consideration must also be given to the fact that several organisms have more than one transketolase gene (e.g., *E. coli* (Sprenger 1993; Iida et al. 1993), *S. cerevisiae* (Schaaf-Gerstenschläger and Zimmermann 1993; Sundström et al. 1993), and *C. plantagineum* (Bernacchia et al. 1995), and the possibility of two transketolase genes in humans has not been ruled out.

Comparisons of the aligned amino acid sequences with the secondary structure (Fig. 1) derived from the crystal structure of yeast transketolase (Sce1) (Nikkola et al. 1994) allowed us to introduce gaps without interruption of secondary-structure elements; all but one of the deletions lie within loop regions. Secondary-structure prediction programs were applied to the remaining sequences. Though the accuracy of such predictions is limited, all sequences have a similar predicted secondary structure, except for the apparent deletion of α-helix 11 from mammalian transketolases (Fig. 1).

The compositions of amino acids and nucleotides in all sequences are shown in Table 4. While the amino acid compositions are similar, the DNA sequences vary considerably in their base compositions. The gram-positive bacteria (*S. pneumoniae, M. leprae*) split into low and high G + C, respectively. *M. genitalium* expectedly has a very low G + C level (Fraser et al. 1995). The α- and β-subdivisions of the gram-negative proteobacteria have a very strong G + C bias (~70%). *E. coli* has a medium and *H. influenzae* Rd a low G + C content. Plant and mammalian sequences display a bias toward G + C (~60%) with the exception of *S. tuberosum* (~45%). *S. cerevisiae* and *P. stipitis* have a fairly low G + C content, whereas *H. polymorpha* has a bias similar to that of mammalian sequences.

Phylogenetic Analysis

Phylogenetic inference from sequence data is dependent on the model that is used. While various approaches implicitly assume stationarity and homogeneous base and amino acid compositions, these assumptions are often invalid in practice. Compositional biases may lead to erroneous tree topologies (Saccone et al. 1989). Some sites in the data set may be determined by processes of bias that may interfere with the true phylogenetic history of compared sequences (Lockhart et al. 1992). It is therefore necessary to check a given data set for homogeneity and stationarity prior to phylogenetic analysis.

The data of Tables 4A and 4B were subjected to contingency tests to determine whether the compositions are homogenous. For the DNA sequences we considered both the nucleotide and the GC/AT composition. The results are summarized in Table 5. The heterogeneity in the DNA sequences is very significant whereas the heterogeneity in the protein sequences is not so strong. Some of the amino acids are heterogeneously distributed and some homogeneously (Table 4A). To chech the data sets for stationarity, χ^2 values were calculated in pairwise comparisons assuming a multinomial distribution (Preparata and Saccone 1987; Saccone et al. 1990). A $\chi^2 \leqslant 1.5$ between two sequences was considered a satisfactory criterion to fulfill stationarity. For the DNA sequences we considered both the entire sequences and every codon position individually. Most of the bacterial sequences are nonstationary in all five comparisons (data not shown). The plant and yeast sequences do not meet the stationarity requirement, either, but the rejection of the null hypothesis (H_0) is in general less significant than in the bacterial sequences. Interestingly, for all eukaryotic sequences except for *P. stipitis*, stationarity seems to be fulfilled at the second codon position (Fig. 5).

Phylogenies derived from heterogeneous, nonstationary data sets must be evaluated with caution (Bull et al. 1993). We elected to apply a number of algorithms and compare the results. For distance-matrix-based methods, evolutionary distances were estimated in pairwise sequence comparisons using various distance measures (see above). Two of these approaches—method of Galtier and Gouy (1995) and the Log Det transformation (Lockhart et al. 1994)—take compositional biases into account. Additionally, we calculated the Euclidean distances between nucleotide and amino acid frequencies for each sequence pair in order to obtain phylogenies

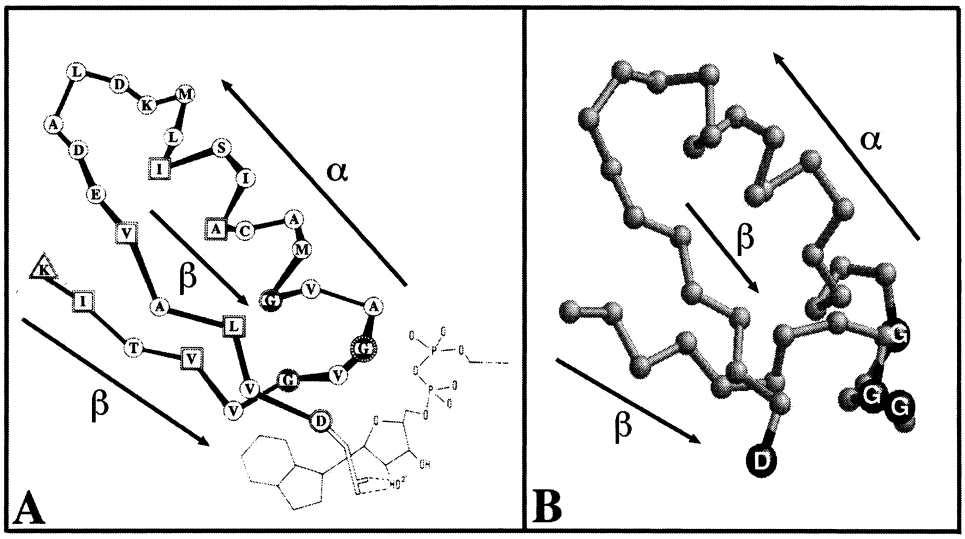

Fig. 2. Structure of the ''transketolase motif'' compared with the nucleotide-binding motif. **A** is adapted from Wierenga et al. (1985). The interaction between protein and the cofactor NADH is illustrated (in particular the hydrogen bond between the invariant aspartate and the 2′-hydroxyl group of the adenosine ribose). α and β indicate α-helix and β-sheet, respectively. **B** shows the fold of the sequence spanned by the ''transketolase motif'' (see text). The conserved residues of the motif reminiscent of the NADH-binding motif of various dehydrogenases are labeled.

Table 3. Residues predicted from the structure to be important for substrate binding (Lindqvist *et al.* 1992; Nikkola *et al.* 1994) to the *S. cerevisiae* (Sce 1) enzyme[a]

$His30^{A}$	$Arg359^{B}$	$Glu476^{B}$
$His69^{A}$	$Leu383^{B}$	$Asp477^{B}$
$Arg94^{A}$	$Ser386^{B}$	$His481^{B}$
$His103^{A}$	$Phe442^{B}$	$Asp503^{B}$
$Ile191^{A}$	$Phe445^{B}$	$Arg528^{B}$
$His263^{A}$	$His469^{B}$	

[a] An additional residue (Asp503), implicated by the findings of this paper, is listed. The character in superscript indicates the subunit to which the respective residues belong

based entirely on the compositions (Lockhart et al. 1994). All distance matrices were used as input files for the NJ and the ME programs. In addition, we used the ML method (Felsenstein 1981), which has been shown to be robust in cases where there is compositional heterogeneity (Galtier and Gouy 1995). We also derived MP phylogenies (Fitch 1971) using a heuristic search. Except for the trees based on Euclidean distances we obtained very similar results from each approach, indicative of a clear distinction between the phylogenetic and the compositional signals.

Figure 1 shows the occurrence of several large gaps, particularly in the aligned mammalian sequences. However, tree topologies were basically unaltered, regardless of whether gaps were excluded in pairwise comparisons or globally removed. Furthermore, the subset of sequential residues listed in the Methods section, which excludes all regions of major gaps, resulted in the same topology.

Distances calculated in pairwise comparisons of DNA and amino acid sequences were in most cases larger than 0.4 substitutions per site. For illustrative purposes one tree derived from the protein alignment (NJ, p-distance) and three trees from the DNA alignment (ME [Kimura distance—Kimura 1980], NJ [Galtier's and Gouy's method—Galtier and Gouy 1995], and ML [Felsenstein 1981]) are presented (Figs. 6–9). Phylogenies not illustrated generally agreed with those presented. In all trees the mammals, plants, and yeasts (with one exception, Hpo, which will be discussed later) form separate clades. The bacterial sequences appear to be polyphyletic.

As expected from the analysis of the alignment, the mammalian sequences differ distinctively from the remaining taxa. A lack of sequences from any animals other than mammals is the reason for the long internal branch. Nevertheless, it appears that the rate of evolution in this branch is slightly higher than that in most other branches. Except for the third codon position the mammalian sequences are significantly stationary, including the protein sequences. (Values for second codon positions are shown in Fig. 5.) Although only data from three mammalian species were available we estimated evolutionary rates and compared the results with those from other studies. First we considered the rates from mouse (Mmu) and rat (Rno) transketolase. A method for comparing evolutionary rates in homologous genes is a relative rate test that does not require the knowledge of divergence times (Sarich and Wilson 1973; Wu and Li 1985). Using the human gene Hsa(ad) as a reference sequence we calculated the differences in synonymous and nonsynonymous substitutions and obtained 4.5 ± 10.62 and -0.12 ± 0.97 substitutions per 100 sites, respectively, the negative sign indicating a higher rate in

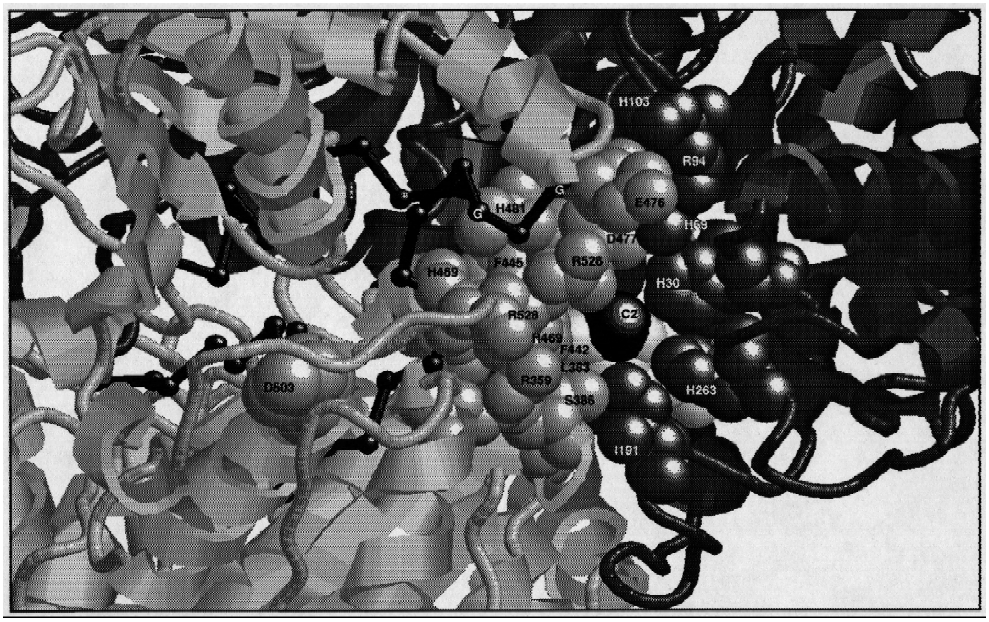

Fig. 3. Predicted substrate channel, based on the structure of yeast transketolase in absence of substrate. Residues in the two identical subunits A and B are shown *dark* and *light*, respectively. Residues predicted to be involved in substrate binding (Table 3) are shown as *space-filling spheres* and line a funnel: half are from each subunit. The C2 atom (*black*) of ThDP forms the bottom of the funnel. Residues belonging to the 'transketolase motif'' (except the ones involved in substrate binding) are indicated in *black*. Two of the conserved glycines of the motif reminiscent of the NADH-binding motif of various dehydrogenases are indicated. Only a major conformational change could bring the invariant D503 into proximity with the active site.

the rat gene. This result suggests a nearly equal rate for mouse and rat, in accordance with previous studies (Li et al. 1987; O'hUigin and Li 1992). Second, we analyzed the actual rate of the rodent genes by assuming a mouse/rat dichotomy of 44–50 Myr (Easteal et al. 1995). We obtained a rate estimate of 3.37 ± 0.41 synonymous substitutions per site per 10^9 years. This value is consistent with the result from Li et al. (1987), who obtained a rate of 3.9–11.8 synonymous substitutions per 10^9 years based on an assumed divergence time of 15 Myr. Third, actual rates based on the comparison between the adult human gene and the rodents were estimated. We used a divergence time of 115–129 Myr (Easteal et al. 1995). The rates of substitutions per site per 10^9 years for synonymous and nonsynonymous substitutions are 2.65 ± 0.28 and 0.12 ± 0.02, respectively. Li and Graur (1991) estimated average numbers for synonymous and nonsynonymous substitutions to be 4.61 ± 1.44 and 0.85 ± 0.73, respectively, per site per 10^9 years. Although they assumed a different rodent–primate divergence time (80 Myr), our results show that transketolase belongs to the more evolutionary conservative proteins. This result is not surprising considering that transketolase is catalyzing a reaction in a metabolic pathway found in most if not all living organisms. A slow rate of evolution might in this case simply reflect functional constraints.

It is interesting to note that the comparison between the rodent genes led to a slightly higher evolutionary rate than the comparison between the rodent and the human genes. Whether such an increase is consistent with the

```
H.sapiens (ad)          GSHCGVSIGEDGPSQMALEDLAMFRSVPTSTVFYPSD
H.sapiens (ft)          GSHCGVSIGEDGPSQMALEDLAMFRSVPTSTVFYPSD
R.norvegicus            GSHCGVSIGEDGPSQMALEDLAMFRSVPMSTVFYPSD
M.musculus              GSHCGVSIGEDGPSQMALEDLAMFRSVPMSTVFYPSD
C.plantagineum 7        MTHDSIGLGEDGPTHQPVEHLASFRAMPNILTLRPAD
C.plantagineum 10       MTHDSIGLGEDGPTHQPVEHLASFRAMPNILVLRPAD
C.plantagineum 3        MTHDSIGLGEDGPTHQPIEHLASFRAMPNILMLRPAD
S.tuberosum             MTHDSIGLGEDGPTHQPIEHLASFRAMPNILMFRPAD
H.polymorpha            GTHDSINEGENGPTHQPVESPALFRAYANIYYMRPVD
S.cerevisiae 1          ATHDSIGVGEDGPTHQPIETLAHFRSLPNIQVWRPAD
S.cerevisiae 2          ATHDSIGLGEDGPTHQPIETLAHLRAIPNMHVWRPAD
P.stipitis              ATHDSIGLGEDGPTHQPIETLAHFRATPNISVWRPAD
E.coli 1                YTHDSIGLGEDGPTHQAVEQLASLRLTPNFSTWRPCD
E.coli 2                YTHDSIGLGEDGPTHQPVEQVASLRVTPNMSTWRPCD
R.sphaeroides           MTHDSIGLGEDGPTHQPVEHLASLRAIPNLAVIRPAD
R.capsulatas            MTHDSIGLGEDGPTHQPVEHLTICRATPNTWTFRPAD
X.flavus                LTHDSIGLGEDGPTHQPVEHVESLRLIPNLDVWRPAD
A.eutrophus             LTHDSIGLGEDGPTHQPVEHAASLRLIPNNQVWRPCD
H.influenzaeRd          YTHDSIGLGEDGPTHQPVEQTASLRLIPNLETWRPCD
M.leprae                WTHDSVGLGEDGPTHQPIEHLAALRAIPRLSVVRPAD
M.genitalium            YTHDSYQVGGDGPTHQPYDQLPMLRAIENVCVFRPCD
S.pneumoniae            FTHDSIAVGEDGPTHEPVEHLAGLRAMPNLNVFRPSD

Consensus                THDS G GEDGPTH P E     R         RP D
                         S CG S  GN  SQ A D                Y
                              N
                              Q
                              A

NADH-binding                  GXGXXG----------X24----------D
like motif
```

Fig. 4. Transketolase motif. Residues belonging to the NADH-binding like motif are in **bold.** Positions that are invariant, conserved, or allow for only two different amino acids are in *italics*. The consensus and the motif reminiscent of the NADH-binding motif of various dehydrogenases (''NADH-binding like motif'') are shown below.

Table 4A. The amino acid compositions (in % values) of the 22 sequences compared are listed[a]

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Total
Hsa(ad)	11.7	1.9	5.6	5.6	3.9	7.2	2.7	7.5	6.7	7.4	2.1	3.2	4.5	4.2	4.7	6.7	5.1	5.8	0.6	2.7	623
Hsa(ft)	11.7	1.9	5.6	5.6	3.9	7.2	2.6	7.5	6.9	7.2	2.2	3.2	4.5	4.2	4.7	6.7	5.1	5.8	0.6	2.7	623
Rat	11.6	2.1	5.8	5.6	4.0	7.9	2.6	6.7	7.2	7.1	2.4	2.7	4.8	4.0	4.2	5.8	5.1	7.1	0.6	2.7	623
Mus	11.6	1.9	5.9	5.5	4.0	7.7	2.6	7.1	7.1	7.1	2.2	2.6	4.8	4.0	4.3	6.4	5.3	6.6	0.6	2.7	623
Cpl7	10.2	1.3	4.7	7.2	3.8	9.9	3.0	5.0	5.9	8.3	2.1	3.8	5.8	1.8	4.7	6.1	5.3	6.5	1.6	2.8	676
Cpl10	10.8	0.6	4.1	7.8	3.7	10.5	2.9	4.0	6.2	8.5	2.4	3.8	5.3	1.5	4.1	6.5	5.4	7.5	1.5	2.9	679
Cpl3	9.8	1.2	4.6	7.7	3.1	9.8	2.5	6.6	6.7	7.5	1.7	4.2	4.8	2.3	3.7	7.5	4.8	6.6	1.7	3.1	519
Stu	11.0	1.0	4.6	7.5	3.9	9.4	2.9	5.8	6.3	7.3	2.0	4.0	4.8	2.3	3.6	6.8	5.5	6.6	1.6	3.2	694
Hpo	10.6	2.3	4.6	6.3	4.1	7.6	2.7	5.8	5.6	8.2	2.4	5.1	4.6	3.8	4.8	5.5	3.5	6.5	0.8	5.2	710
Sce1	10.0	0.3	5.0	5.1	4.1	8.5	2.5	5.3	6.3	9.3	1.6	4.6	5.6	3.4	3.2	8.4	5.4	6.2	1.3	3.8	680
Sce2	7.8	0.4	6.0	6.0	4.4	9.0	2.6	5.1	5.9	8.4	1.8	3.8	5.0	4.0	4.1	8.7	5.0	6.5	1.3	4.3	681
Pst	10.4	0.7	4.9	5.2	3.7	8.1	2.4	5.3	6.0	9.8	1.4	3.6	5.3	3.9	3.2	8.1	6.3	6.9	1.3	3.5	695
Eco1	13.3	0.8	5.3	7.2	3.9	8.7	2.7	4.8	5.3	7.7	3.3	3.3	4.5	3.5	4.2	5.9	5.0	5.7	1.7	3.3	664
Eco2	11.4	0.6	4.9	7.3	3.4	9.4	3.9	5.4	5.7	8.5	2.4	3.7	5.2	3.4	4.0	4.6	5.2	5.4	1.8	3.3	667
Rsp	17.5	0.9	5.6	5.6	2.7	10.0	3.3	5.2	2.0	8.5	3.5	1.8	5.8	2.1	6.5	4.6	5.3	5.2	1.2	2.4	657
Rca	15.5	0.9	4.8	6.8	2.8	9.1	2.8	4.6	5.4	8.6	3.6	2.5	4.3	2.2	6.7	3.9	5.4	6.1	2.1	1.9	672
Xfl	16.4	0.9	5.4	6.3	3.6	9.5	3.1	4.4	2.2	9.0	2.3	1.9	6.6	1.7	7.1	4.2	5.8	6.0	1.9	1.7	687
Aeu	16.6	1.6	5.8	5.8	2.8	9.4	3.9	3.1	1.3	10.4	2.5	2.7	6.1	2.4	6.6	3.7	4.0	6.9	1.8	2.4	670
Hin	12.8	0.8	5.0	7.2	4.2	8.9	3.0	5.7	5.1	7.8	2.3	3.8	4.2	3.6	4.5	6.0	4.7	5.3	1.8	3.5	665
Mlp	11.3	0.4	6.4	6.9	2.6	9.3	3.0	5.6	2.4	9.0	1.1	2.0	5.3	3.3	5.7	6.4	7.2	6.9	2.0	3.1	699
Mge	5.9	1.1	6.5	4.6	6.0	6.2	3.4	6.6	8.0	10.2	2.0	6.8	3.2	5.2	2.9	6.2	4.6	5.4	0.9	4.2	648
Spn	12.2	0.3	6.3	5.9	3.9	8.5	2.5	4.5	4.6	8.3	2.0	4.8	4.6	2.9	4.0	6.2	6.0	8.0	0.9	3.7	649
	−	−	+	+	+	+	+	−	−	+	+	−	+	−	−	−	+	+	+	−	
Average	11.8	1.1	5.3	6.3	3.8	8.7	2.9	5.5	5.4	8.4	2.2	3.5	5.0	3.2	4.6	6.1	5.2	6.3	1.4	3.2	

[a] The ten frame shifts denoted by an X in the alignment in Fig. 1 (Sundström et al. 1993) are not listed. The last column indicates the total number of amino acids per sequence. Homogeneous and heterogeneous distribution are indicated by + and −, respectively

finding of a higher rate in rodents than in humans, as reported elsewhere (Wu and Li 1985), cannot be ascertained from this limited data set. However, accumulation of other mammalian sequences (e.g., artiodactyls, lagomorphs, marsupials) will shed light on the uniformity of the mammalian rate. The ubiquity and the slow evolution might make transketolase a suitable model for a molecular clock, particularly in the case of mammals in which the amino acid and nucleotide distributions are stationary.

Although transketolase genes from only two different plants have been sequenced so far (Bernacchia et al. 1995; Teige et al. 1995: GenBank) the branching order of the three Cpl genes is ambiguous. While NJ (p-distance) and ME trees suggest that the constitutively expressed Cpl3 is more closely related to the gene of *S. tuberosum* (Stu) than to the stress-induced Cpl7 and Cpl10 genes (Figs. 6 and 7), the ML and the NJ (Galtier-Gouy distance) trees suggest the opposite (Figs. 8 and 9). Although the differences in base composition in the three Cpl genes are not very large, Cpl3 shows clearly the highest degree of similarity to Stu (Table 4B). Additionally, a slight bias in the transition/transversion ratio between Cpl3 and Cpl7/Cpl10 was observed. (The results for pairwise comparisons were 0.977, 0.959, and 0.771 for Cpl3 vs Cpl7, Cpl3 vs Cpl10, and Cpl7 vs Cpl10, respectively.) Consistently, in all inferred phylogenies the plants form a monophyletic group.

In contrast the yeast group dooes not form a cluster. The branching order that suggests a gene duplication in *S. cerevisiae* after divergence from *P. stipitis* is supported by all trees (Figs. 6–9). Unusual was the branching of Hpo, which consistently clustered with the mammalian sequences. However, pairwise comparisons of the distances between amino acid and nucleotide sequences both indicated that Hpo is slightly more distant from mammalian sequences than from the other yeasts. The G + C content in Hpo, however, is elevated, similar to the mammalian sequences (Table 4B). In an attempt to clarify the branching of Hpo an ML tree was constructed, using only the mammalian and yeast sequences (Fig. 10). The resulting phylogeny is trifurcated; although the Hpo transketolase gene sequence seemed to cluster closer to the yeast than to the mammalian genes, no clear branching pattern was obvious. Implications of these results for the molecular evolution of the Hpo gene are discussed below.

The branching topology of the bacterial sequences is complex. In all phylogenetic trees (Figs. 6–9) the gram-negative bacteria form three distinct groups: Aeu and Xfl share the same ancestral gene. The same is true for Rca and Rsp. *E. coli* contains two different transketolase genes, one of which is responsible for the major activity (Eco1). Eco2, in contrast, is responsible for only minor activity (Iida et al. 1993) but appears to be the ancestral gene. The G + C content for both genes is ~55% (Table 4B). The entire genome of *H. influenzae* Rd has been sequenced recently (Fleischmann et al. 1995) and only

Table 4B. Base compositions (in % values), G + C contents, and length of sequences (in bp)

	A	T	C	G	G + C	Total
Hsa(ad)	23.2	18.1	30.4	28.3	58.7	1869
Hsa(ft)	23.6	17.9	30.3	28.3	58.6	1869
Rat	24.0	19.1	28.9	27.9	56.8	1869
Mus	24.0	18.9	29.3	27.8	57.1	1869
Cpl7	22.1	18.0	28.3	31.7	60.0	2028
Cpl10	21.8	18.6	27.4	32.1	59.5	2037
Cpl3	24.1	20.0	26.7	29.3	56.0	1557
Stu	26.6	28.9	19.9	24.6	44.5	2082
Hpo	23.6	20.0	27.4	29.1	56.5	2130
Sce1	27.5	28.0	23.0	21.5	44.5	2040
Sce2	27.9	27.1	20.8	24.2	45.0	2043
Pst	24.5	27.2	25.6	22.7	48.3	2085
Eco1	22.7	21.4	27.9	28.0	55.9	1992
Eco2	24.2	21.1	25.0	29.7	54.7	2001
Rsp	14.8	15.1	34.5	35.6	70.1	1971
Rca	17.8	15.3	32.0	34.9	66.9	2016
Xfl	14.5	14.7	38.7	32.2	70.9	2061
Aeu	14.5	14.7	35.9	35.0	70.9	2010
Hin	29.3	27.6	20.7	22.4	43.1	1995
Mlp	20.8	18.4	31.0	29.7	60.7	2097
Mge	34.6	32.4	15.1	17.9	33.0	1944
Spn	28.0	27.6	21.7	22.7	44.4	1948
	–	–	–	–	–	
Average	23.4	21.4	27.2	28.0	55.5	

Table 5. Analysis of compositions[a]

Data set	Degrees of freedom	α	χ^2 (P = 5%)	χ^2 (calc.)	H_0
GC/AT	21	0.05	32.67	1698.68	Rejected
Nucleotides	63	0.05	82.53	1821.62	Rejected
Amino acids	399	0.05	446.29	634.37	Rejected

[a] Contingency tests for homogeneity of the 22 sequences were applied to the nucleotide pair GC (or AT), all four nucleotides, and all 20 amino acids. The null hypothesis (H_0: homogeneity) was rejected at the 5% level (α). χ^2 values expected for a homogeneous distribution (χ^2 [P = 5%]) and the values obtained from our data set (χ^2 [calc.]) are indicated

one gene coding for transketolase was discovered. The G + C content is low (~43%). Consistently, *H. influenzae* Rd transketolase branches with Eco1.

The branching order between these three groups of the gram-negative bacteria was incongruent in the different topologies. ME and ML trees (Figs. 7 and 8) suggest a common ancestor for *A. eutrophus, X. flavus, R. sphaeroides,* and *R. capsulatus.* All four organisms belong to the α and β subdivisions of the proteobacteria. The same result has been reported using the sequences of the class II fructosebisphosphate (FBP) aldolase (Van Den Bergh et al. 1996). In contrast, the two NJ trees (Figs. 6 and 9) suggest a common ancestor for *A. eutrophus, X. flavus,* and the ancestor of *E. coli* and *H. influenzae* Rd. While the ML and the two NJ trees (Figs. 6, 8, and 9) suggest an ancestor common to all gram-negative bacteria in this comparison, the ME tree (Fig. 7) clustered the γ subdivision with the plants and yeasts. A similar result was reported for the class II FBP aldolase (Van Den Bergh et al. 1996).

The closest relative for *M. genitalium* was expected to be a low G + C gram-positive bacterium (Fraser et al. 1995). The only representative of this group in our analysis is *S. pneumoniae,* which contains the putative transketolase recP. In all trees shown they share a common ancestor; however, the bootstrap support in the NJ trees is fairly low (Figs. 6 and 9).

NJ (p-distance) and ME trees (Figs. 6 and 7) suggest that *M. leprae* might share an ancestral gene with the plant and yeast groups, whereas ML and NJ (Galtier-Gouy method) trees cluster this sequence with the gram-negative bacteria (Figs. 8 and 9). Although the paucity of data makes it impossible to resolve the phylogenetic relationship between bacterial sequences, it seems that the bacteria form a polyphyletic group.

Discussion

Sequence Comparison and Secondary Structure

Alignment of the 22 transketolase sequences derived from four different phyla (Fig. 1) illustrates the high level of conservation that has been maintained in this enzyme throughout evolution. For the purpose of the studies reported here we have used our own human transketolase sequence, since it differs from the sequence previously available in the GenBank database. McCool et al. have published the sequences encoding human transketolase from five individuals (McCool et al. 1993). With the exception of a single conservative polymorphism, four of these sequences were identical and could be termed the consensus sequence. The fifth sequence differed at nine bases; hence, it is perhaps unfortunate that this sequence is the one deposited in the database. We have independently cloned and sequenced the cDNA encoding human transketolase, and it is entirely in agreement with the consensus sequence mentioned above. Since this sequence appears to be the common one, we have deposited this, also, into the database.

The crystal structure of transketolase from *S. cerevisiae* (Sce1 in Fig. 1) has allowed the identification of critical residues required for cofactor and substrate binding, plus subunit interactions (Lindqvist et al. 1992; Nikkola et al. 1994). The alignment presented in Fig. 1 enables a comparison to be made between these residues and the corresponding residues in the transketolases from other sources, identifying those that are totally invariant across species (see Table 2). Totally conserved residues number 50 and include, among others, His30 and His69 which, along with His103 and His263, form part of a cluster of conserved histidines predicted to be involved in substrate binding (Lindqvist et al. 1992; Nikkola et al. 1994). Although His481 has also been included in this group, our alignment shows that there is a substitution His481Gln in all of the studied mammalian sequences

	Hsa(ad)	Hsa(ft)	Rno	Mmu	Cpl7	Cpl10	Cpl3	Stu	Hpo	Sce1	Sce2
Hsa(ft)	0.00	*									
Rno	0.02	0.03	*								
Mmu	0.01	0.02	0.02	*							
Cpl7	0.35	0.34	0.51	0.32	*						
Cpl10	0.31	0.31	0.42	0.26	0.17	*					
Cpl3	0.14	0.11	0.26	0.20	0.29	0.36	*				
Stu	0.23	0.20	0.32	0.25	0.48	0.20	0.15	*			
Hpo	0.62	0.60	0.54	0.75	**1.95**	**1.80**	0.67	1.05	*		
Sce1	1.50	1.49	1.46	1.47	**2.42**	1.36	1.43	0.91	**2.20**	*	
Sce2	0.65	0.62	0.63	0.74	**1.72**	1.16	0.60	0.45	0.50	0.65	*
Pst	**2.39**	**2.41**	**2.32**	**2.29**	**3.36**	**2.01**	**2.32**	**1.76**	**3.60**	0.21	**1.55**

Fig. 5. Stationarity check of eukaryotic sequences at the second codon position. χ^2 values were calculated according to the method of Saccone et al. (1990). Values that do not fulfill the adopted criterion for stationarity ($\chi^2 \leq 1.5$) are in *bold.*

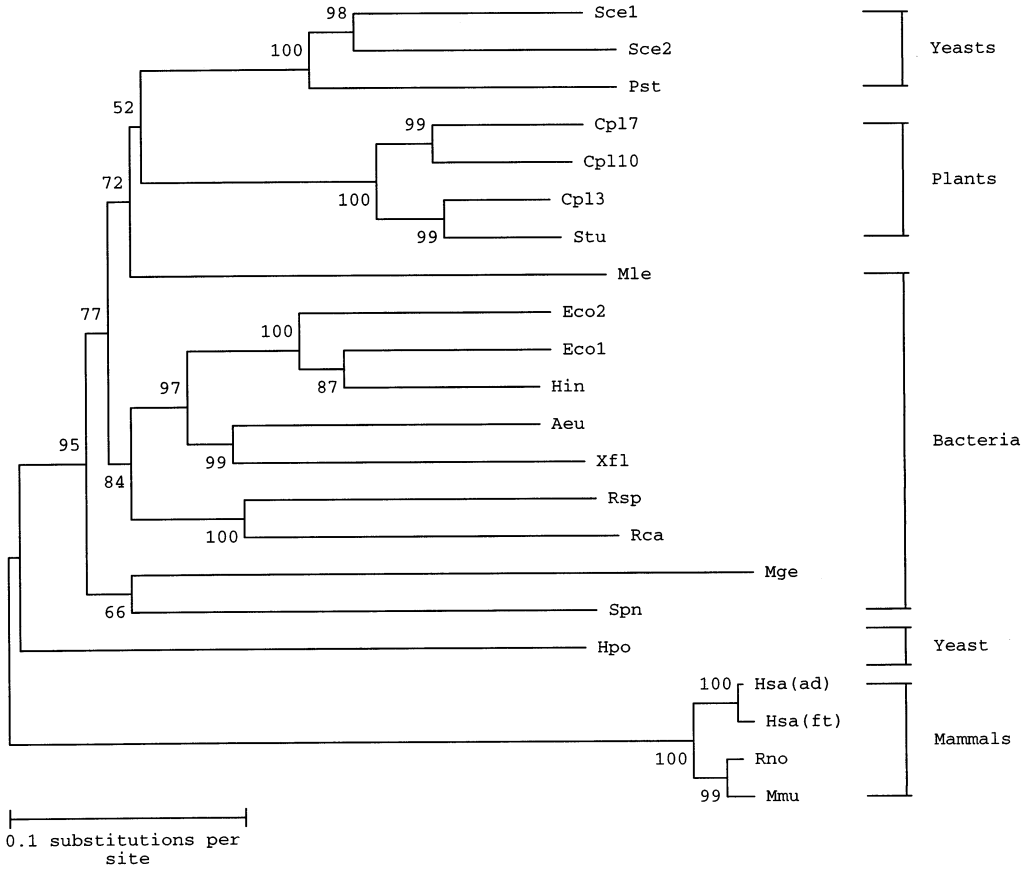

Fig. 6. Unrooted neighbor-joining tree based on the amino acid alignment in Fig. 1. Percentage confidence levels from 1,000 replicates are indicated at the nodes of the branches. Evolutionary distance (p-distance) is indicated by a *scale bar* below the tree.

(Figs. 1 and 4). From the alignment we were able to identify another invariant residue, Asp503, which could possibly interact with the substrate (Fig. 2A,B). Although analysis of the structure of Sce1 transketolase free of substrate showed that this residue is not in the immediate proximity of the substrate channel (Fig. 3), it is possible that conformational changes during binding of the substrate might close the gap between Asp503 and the substrate. Until knowledge of the crystallographic structure, with substrate bound to enzyme, and site-directed mutagenesis studies become available, this issue remains open. Interestingly, a further 24 residues are completely invariant in all but the mammalian enzymes (see Table 2B), indicating that these residues may be involved in some aspect of transketolase function. For example, it is known that transketolases from plants and yeasts display a wide range of substrate specificity (Kochetov 1986; Villafranca and Axelrod 1971). Dihydroxyacetone synthase seems to have the largest spectrum of possible substrates (Kato et al. 1982), whereas mammalian transketolases are more selective (Paoletti 1983; Masri et al. 1988; Waltham 1990). Recently transketo-

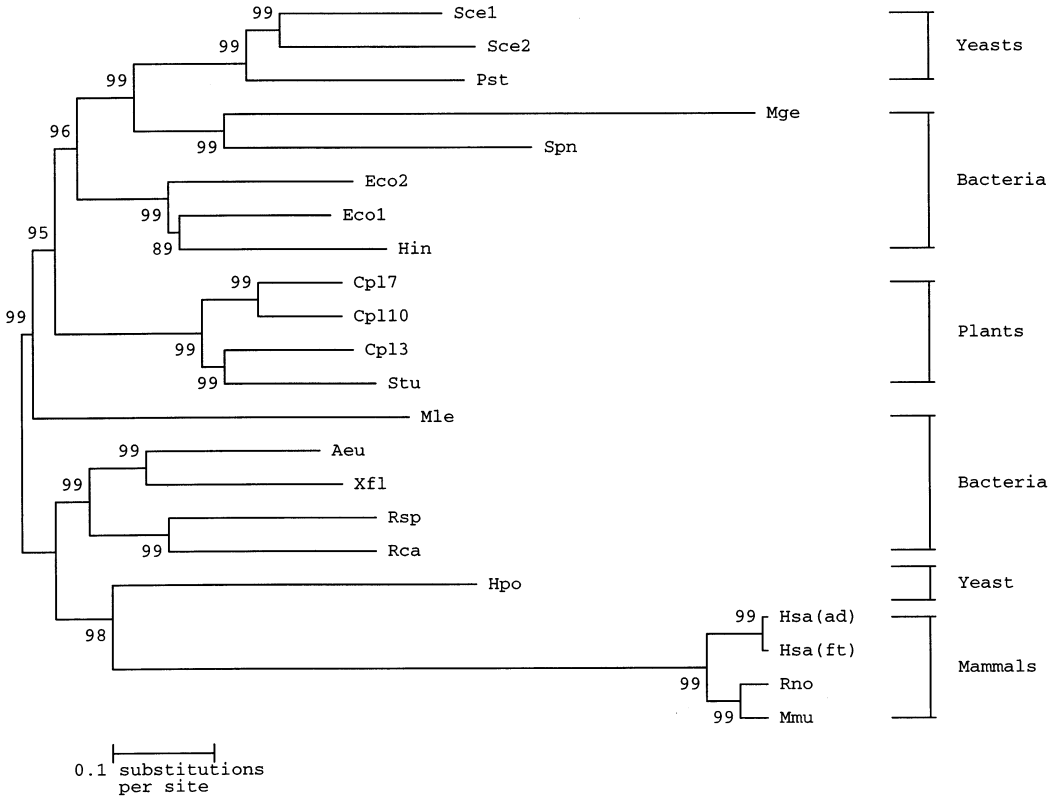

Fig. 7. Unrooted minimum evolution tree based on the DNA alignment. Percentage confidence levels from 1,000 replicates are indicated at the nodes of the branches. Evolutionary distance (estimated according to the method of Kimura, 1980) is indicated by a *scale bar* below the tree.

lase from *E. coli* has been shown to display a similar substrate specificity to yeast and plant transketolase (Sprenger et al. 1995). Table 3 lists residues identified as possibly being involved in substrate binding (Lindqvist et al. 1992; Nikkola et al. 1994]. Although those residues are highly conserved among all species in our comparison (Fig. 1), there are a few distinct differences between the mammalian and the remaining transketolases (Ile191Gln, Leu383Thr, and His481Gln, respectively), as well as between the remainder and that of Hpo (Arg94His, Ile191Cys, and Asp477Asn, respectively). Table 2B lists sequence variations between mammalian and other transketolases and, as mentioned above, some of these variations may reflect variations in functions. His481Gln falls into this category as it may reflect different substrates specificities between the two groups. Currently only a small number of transketolase enzymes have been well characterized in respect to turnover number. Calculations based on studies of transketolases from *E. coli* (Sprenger et al. 1995), *S. cerevisiae* (Sundström et al. 1993), and human (Waltham 1990) resulted in k_{cat} values of $60s^{-1}$, $45s^{-1}$, and $20s^{-1}$, respectively. These modest turnover numbers are consistent with the assumption that transketolase is adapted better for breadth of substrate than for efficiency.

Release of cofactor from holoenzyme requires completely different conditions for the two groups, and this property may depend on interactions involving residues known to participate in closure of the cofactor binding pocket at the subunit interface and that vary between mammalian and nonmammalian transketolases, such as Tyr448Arg and Ala381Gly. Additionally, the substitutions Tyr448Arg, together with Arg417Ala, might be responsible for different subunit interactions observed between the two groups. The highly conserved ThDP-binding motif first identified by Hawkins et al. (1989) is present in the N-terminal domain, but, surprisingly, is not the most highly conserved stretch of sequence. Crystallographic studies have shown that the ThDP-binding site is accessible from the solvent through a deep cleft between the two subunits. The cleft is lined with conserved residues located on loop regions (Lindqvist et al. 1992; Nikkola et al. 1994), among which is one of the longest stretches and most highly conserved regions of sequence in the whole polypeptide chain (Figs. 2B and 4). This region has been identified as reminiscent of a nucleotide-binding motif (Abedinia et al. 1992). It has been suggested that other invariant residues within this motif are involved in subunit dimerization and substrate binding (Nikkola et al. 1994). Part of the subunit dimerization interface forms the site for binding of the thiazolium and pyrimidine rings of the cofactor, and since the substrate has to react with the C2 of the thiazolium ring, some of these residues could be involved in both substrate binding and subunit interactions. It seems that this highly conserved stretch between Thr468 and Asp503 (Figs. 2B

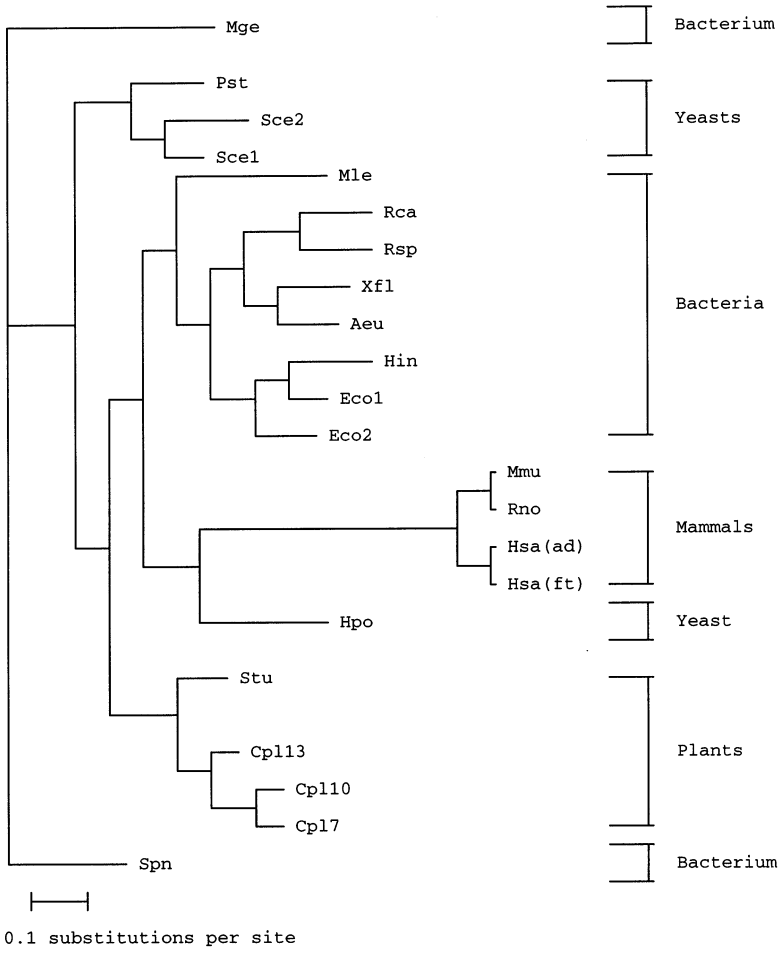

Fig. 8. Unrooted maximum likelihood tree based on the DNA alignment. Evolutionary distance is indicated by a *scale bar* below the tree.

and 4) cannot be assigned to one specific function in the protein, such as binding of cofactors or substrates. However, residues crucial for structure and function of the enzyme are conserved in this segment. Through comparison with the nucleotide-dependent dehydrogenase enzymes, particularly D-lactate dehydrogenase (Bernard et al. 1995) and formate dehydrogenase (Lamzin et al. 1992), we have shown that this region has a structure that is reminiscent of the nucleotide-binding region of various dehydrogenases. The term ''transketolase motif'' seems therefore an appropriate description because this motif is conserved in all transketolases so far sequenced and is found in no other sequence in the various databases.

Phylogenetic Inference

Phylogenies derived from species that are very divergent often result in trees with deep branches. Such trees may be unreliable because of stochastic errors due to random substitutions and compositional biases (e.g., Hashimoto et al. 1995). Phylogenetic trees constructed from a single gene or protein may differ from species phylogeny. Analysis of different sequences led to different rootings for the ''tree of life'' (Saccone et al. 1995). Events such as gene duplication, horizontal gene transfer, and endosymbiosis are normally considered responsible for such discrepancies. Nonetheless, homologous sequences have a natural tendency to diverge over time. It is believed that with a large data set comprising many different sequences, discrepancies are cancelled so that the evolutionary process behaves stochastically (e.g., Doolittle et al. 1996).

We used amino acid and the corresponding DNA sequences for the enzyme transketolase from two of the Ur-kingdoms, Eukarya and Bacteria (Woese et al. 1990). All 19 phylogenetic trees generated by various methods (e.g., Figs. 6–9) indicate that the branching order obtained from transketolase sequences is, in general, conventional, as has been reported elsewhere (Van Den Bergh et al. 1996). There is a distinct subdivision into four groups representing mammals, yeasts, bacteria, and plants, as would be expected from their taxonomic relationship. However, since we have no outgroup, all trees shown are unrooted. Hence, it is not possible to determine with any accuracy the position of the ancestral gene in the trees.

It is clear that mammalian and plant enzymes form monophyletic groups, but the branching order of the three Cpl sequences within the plant phyla is ambiguous.

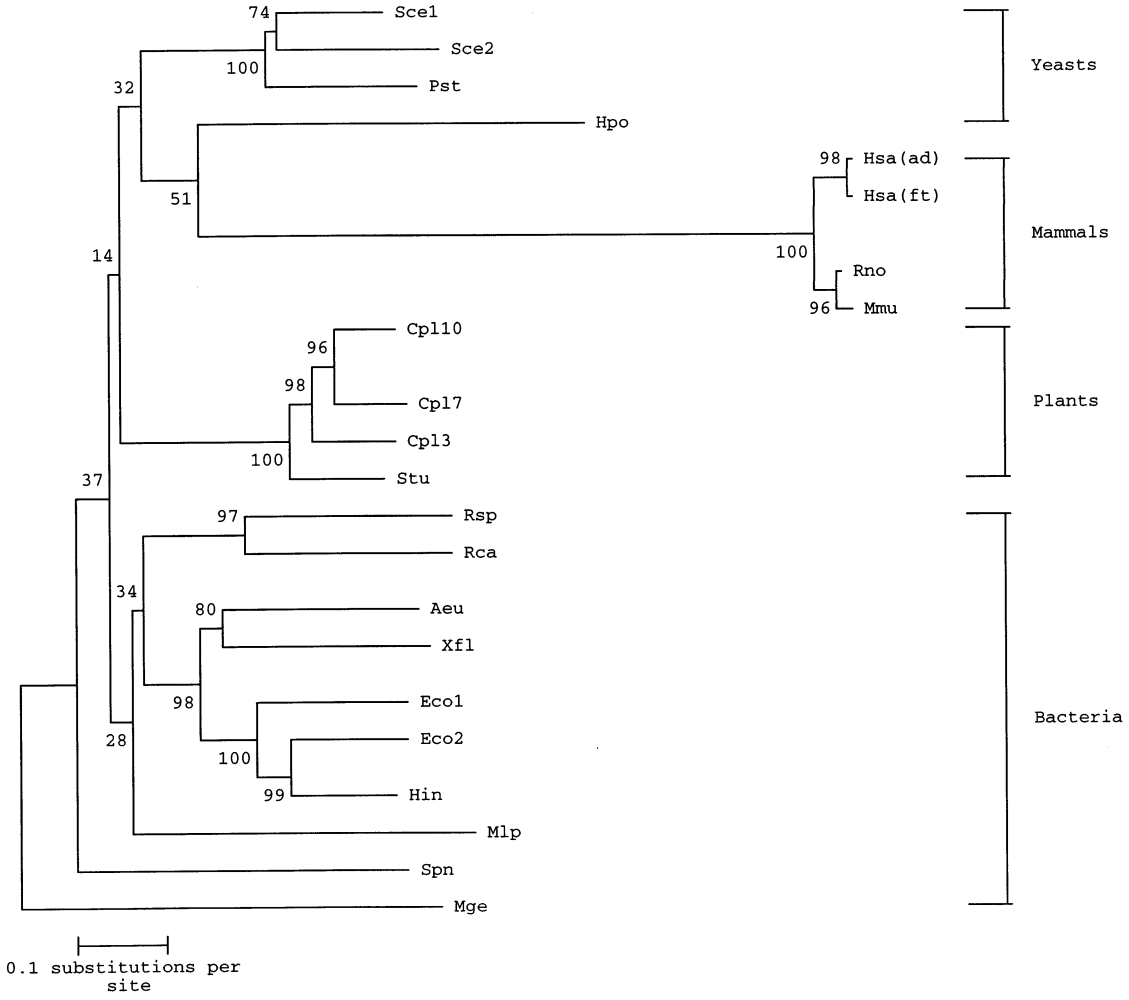

Fig. 9. Unrooted neighbor-joining tree based on the DNA alignment. Percentage confidence levels from 1,000 replicates are indicated at the nodes of the branches. Evolutionary distance (as estimated according to the method of Galtier and Gouy, 1995) is indicated by a *scale bar* below the tree.

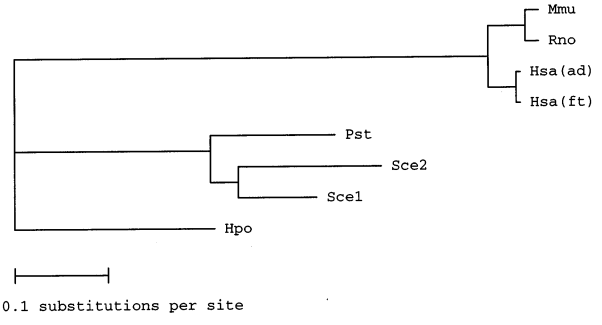

Fig. 10. Unrooted maximum likelihood tree based on the DNA sequences of mammalian and yeast sequences. Evolutionary distance is indicated by a *scale bar* below the tree.

While the NJ (based on p-distance) and ME (Kimura distance) trees suggest a duplication event prior to the split of *C. plantagineum* and *S. tuberosum* (Figs. 6 and 7), the ML and the Galtier and Gouy distance NJ tree suggest a duplication after the divergence (Figs. 8 and 9). Analysis of base compositions showed that Cpl7 and Cpl10 have a fairly high G + C content (~60%). In Cpl3 it is ~56% and in Stu ~45% (Table 4B). In cases of compositional biases the method of Galtier and Gouy and the ML algorithm have been shown to perform more reliably than methods that implicitly assume homogeneous and stationary compositions (Galtier and Gouy 1995). Therefore, a speciation of *S. tuberosum* and *C. plantagineum* prior to gene duplications in *C. plantagineum* is more likely. This is in agreement with the observation that Cpl7 and Cpl10 are expressed only after relief of dessication, while Cpl3 is constitutively expressed. The occurrence of the two inducible genes might simply reflect an adaptation to alteration of the environment.

An unexpected observation was the grouping of *H. polymorpha* with the mammalian sequences rather than with those of the other yeasts. This is in contrast to catalase of *H. polymorpha*, which has been shown to group with the other yeasts (Von Ossowski et al. 1993). Obviously, the topology is not simply due to compositional bias, because all applied algorithms (including Log Det) show the same branching order. The bootstrap, where applied, varies between 51% and almost 100%

(Log Det: 80%). Using only the yeast and mammalian sequences did not clarify the phylogeny (Fig. 10).

The phylogenetic trees in our analysis reveal that the internal branches are relatively short in comparison with the branches leading to extant species (Figs. 6–10). This increases the probability of obtaining an erroneous tree topology due to stochastic errors caused by random substitutions of bases or amino acids (Nei 1987). Long branches signify a large number of changes. In accordance with increasing numbers of total change the number of homoplastic changes may increase, which can lead to an underestimation of evolutionary distances and thus a sequence convergence that does not reflect the true phylogenetic relationship. In general, small data sets are more sensitive to errors caused by stochastic and homoplastic processes. Often, addition of more taxa (ingroups and outgroups) leads to discovery of such processes (Sanderson 1990). We attempted to resolve the effect of homoplasy by use of the bacterial sequences as an outgroup but *H. polymorpha* still branched with the mammals. We argue that sequence convergence due to homoplastic changes does not account for the unusual branching of *H. polymorpha* but, without knowledge of the root of the tree or a proper outgroup, we cannot validate this assumption.

Unexpected branching of one member of a tree which otherwise displays conventional topology might be evidence for a horizontal gene transfer (Smith et al. 1992), especially when sequences of other molecules from the same member result in conventional phylogenies. However, such a transfer is in general very difficult to prove. An argument against a horizontal gene transfer between *H. polymorpha* and the ancestor of the mammalian clade is that the evolutionary distance between the transketolases of *H. polymorpha* and mammals is rather large; in fact the number of substitutions per site is smaller between *H. polymorpha* and the other yeasts than between *H. polymorpha* and the mammalian sequences.

Assuming a gene duplication prior to the split of yeasts and animals, the unusual branching pattern of *H. polymorpha* can be rationalized. While the majority of the yeasts retained one of the paralogous genes, *H. polymorpha* kept the other one, as did the mammals. Selective pressure or advantage may be the reason. *H. polymorpha* can utilize methanol as a carbon source, and its transketolase displays a very wide range of substrate specificity (Kato et al. 1982). However, the mammalian enzymes have a very different and limited substrate specificity, as mentioned above. Therefore, selective pressure related to the enzymatic properties does not explain why *H. polymorpha* and mammals retained the same ancestral gene, and a more likely reason is that environmental influences favored one gene over another. In the course of evolution the mammalian sequences may have lost parts of their DNA through deletions of exons and/or alternative splicing. Though all these deletions occurred in functionally less significant regions (Fig. 1) the selectivity for substrates was increased, reflecting a more complex organization of the evolving organisms. These results suggest that the ancestral genes of *H. polymorpha* and the other yeasts in our comparison were paralogous and that the mammalian and *H. polymorpha* genes are orthologous.

Eubacterial radiation is reported to be bush-like (e.g., Galtier and Gouy 1994). Expectedly, we were unable to resolve the bacterial phylogeny unambiguously. All trees except the ME tree (Fig. 7) suggest monophyly for the α, β, and γ subdivisions of the proteobacteria, as expected (e.g., Galtier and Gouy 1994). The discrepancy in the ME tree may be due to compositional biases. This tree was generated with a distance matrix based on Kimura's two-parameter model (Kimura 1980), which does not account for compositonal heterogeneities. Table 4B shows, however, that the γ-subdivision has a much lower G + C content than the α- and β-subdivisions. In contrast, the NJ tree in Fig. 6 has been inferred by estimating the p-distance between the protein sequences and, although this measure does not account for heterogeneities either, the expected tree topology was obtained, in agreement with the observation that the heterogeneity within the protein sequences is less significant than that within the DNA sequences (Table 5).

The represented gram-positive bacteria belong to the high G + C (*M. leprae*) and the low G + C (*S. pneumoniae* and *M. genitalium*) subdivisions, respectively. *S. pneumoniae* is closest to *M. genitalium*, a result that could be expected. It has been reported that 50% of all identified translation products in *M. genitalium* have great similarity with low G + C gram-positive bacteria such as *Bacillus subtilis* and other mycoplasma species (Fraser et al. 1995).

The monophyly of gram-positive bacteria is controversial (Galtier and Gouy 1994). Nonetheless, many studies report monphyly for gram-positive and gram-negative bacteria (e.g., Doolittle et al. 1996). Furthermore, the gram-positive bacteria seem to be divided into two subdivisions, the high G + C and low G + C (Galtier and Gouy 1994). While the low G + C sequences branched together in our analysis, the relative position of the gram-positive sequences varied strongly between the different trees (Figs. 6–9). Bootstrap support was often low. However, the ML tree and the NJ tree based on the algorithm of Galtier and Gouy (1995) suggest an ancestral gene common to both the ancestor of the gram-negative bacteria and *M. leprae*. This could suggest a polyphyly for the gram-positive sequences in our comparison. Addition of more sequences to the phylogeny in the future will probably clarify the branching of the bacterial clade.

At least three organisms have more than one gene coding for a transketolase. For example, there are two genes in *E. coli* (Sprenger 1993; Iida et al. 1993), two in *S. cerevisiae* (Sundström et al. 1993; Schaaf-Gerstenschläger and Zimmerman 1993), and three in *C.*

plantagineum (Bernacchia et al. 1995). This observation suggests yet other events of gene duplication.

In *E. coli* Eco2 is responsible for minor activity, apparently as a backup for Eco1 (Iida et al. 1993). Phylogenetic analysis shows that Eco2 is the ancestral gene. Transketolase from *H. influenzae* Rd is more closely related to Eco1 than to Eco2. The whole genome of *H. influenzae* Rd has been sequenced with high reliability (Fleischmann et al. 1995) and no second transketolase gene has been found. This suggests that an initial duplication of Eco2 occurred prior to the speciation of *E. coli* and *H. influenzae* Rd. After speciation, *E. coli* retained the ancestral gene while *H. influenzae* Rd excluded it. An alternative explanation involves a gene transfer from *H. influenzae* Rd to *E. coli*. Such a transfer is not unusual for *E. coli*. One of the few firm cases of horizontal gene transfer is the transfer of a second gene coding for glyceraldehyde 3-phosphate dehydrogenase from a eukaryote to *E. coli* (Smith et al. 1992). The relatively short distance between Eco1 and Hin favors such a hypothesis. However, analysis of the base compositions (Table 4B) shows that Eco1 and Eco2 have a similar but much higher G + C contents than Hin. Therefore, a horizontal gene transfer is not very likely in this case.

In our comparisons the phylogeny depicted in Fig. 9 best reflects the expected ''tree of life'' (Doolittle et al. 1996). Even though bootstrap support is far from convincing, Fig. 9 suggests that yeasts and mammals are more closely related to each other than to plants, in agreement with recent reports (Baldauf and Palmer 1993; Wainright et al. 1993).

Applications of methods that are known to be robust in cases of compositional heterogeneities (Log Det [Lockhart et al. 1994] and Galtier's and Gouy's method [1995]) assisted in resolving some of the ambiguities in the phylogeny of transketolase sequences, but in general, variations between the trees (including the ones inferred by algorithms that assume homogeneous compositions) were small. This is in agreement with the proposition that the phylogenetic signal is significantly stronger than the compositional signal. As mentioned above, phylogenies constructed from a single gene or protein may differ from the actual species phylogeny. However, with the exception of the anomalous behaviour of Hpo, the proposal that transketolase sequences seem to reflect ''true'' phylogenies is quite reasonable.

The ThDP-binding motif shares some sequence similarities with the transketolase motif, and it has been noted previously (Lindqvist et al. 1992) that the N-terminal and middle domains have similar topologies. It is possible that these two domains are related, having arisen by partial gene duplication, and it is of interest that the two motifs occupy similar sequence positions within their respective domain, each commencing approximately 150 residues from the beginning of the domain. We therefore considered the possibility that these two motifs might have evolved from a common ancestor. Detailed examination of the secondary structure argues against this proposition. The ThDP-binding motif begins immediately after β2 of the N-terminal domain and extends through α7, α8, and β3; the equivalent secondary structures in the middle domain are β13, α17, α18, and β14, but the transketolase motif begins only after β14. Thus, it appears that the two motifs have arisen independently. Analysis of the ThDP and transketolase motif showed that the number of nonsynonymous substitutions in both motifs is only about half as many as in the entire sequence. The number of synonymous substitutions, however, seems to be slightly higher than in the whole gene (data not shown). This result is consistent with the imposition of functional constraints upon these two conserved motifs.

We have tried to analyze evolutionary rates as far as possible within this small set of data. Such calculations are always limited by the accuracy of the divergence times assumed. One of the major controversies of molecular evolution is the regularity of the molecular clock. Kimura predicted a uniform rate of evolution (Kimura 1968, 1983). Generation time and uncertainty about divergence times were some of the factors used to explain deviations from this uniformity. Recently, it has been suggested that a mutation rate of $2–2.25 \times 10^{-9}$ substitutions per site per year blends molecular and fossil data best, at least for mammals (Easteal et al. 1995). Our result for the human rate is comparable (see Results). However, the rate obtained for the rodents is faster and is in agreement with previous studies (Li et al. 1987).

Although the evolutionary rate appears to have increased in some branches of the phylogenetic trees, the variation seems to be rather small. The fairly constant and conserved evolution of transketolase, together with its position in an ancient metabolic pathway whose function has remained conserved throughout evolution, leads us to suggest that transketolase might be a good model for a molecular clock, in particular for mammals.

Factors such as gene duplication (both in eukaryotes and prokaryotes), exon shuffling, transposition, nucleotide mutations, random genetic drift, and others (Langridge 1991) are all thought to be responsible for most present-day molecular diversities. Transketolase has evolved through the combined effects of at least three of these mechanisms, i.e., gene duplication, insertions/deletions, and accumulated point mutations, and maybe horizontal gene transfer. It has been postulated (Keese and Gibbs 1992) that two classes of genes might exist; those representing ''ancient'' housekeeping genes, and ''younger'' genes encoding proteins with specialized functions acquired through environmental pressure. Transketolase undoubtedly ranks with the ''ancient'' housekeeping gene classification since it is common to all cellular organisms, as is the pentose-phosphate metabolic pathway.

Acknowledgments. We are very grateful to Drs. Chris Collett, Robert Slade, Steve Barker, Lars Jermiin, Don Maclean, and Lindsay Sly for their supportive suggestions and constructive comments, particularly in the phylogenetic section.

References

Abad-Zapatero C, Griffith JP, Sussman JL, Rossman MG (1987) Refined crystal structure of dogfish M4 apo-lactate dehydrogenase. J Mol Biol 198:445–467

Abedinia M, Layfield R, Jones SM, Nixon PF, Mattick JS (1992) Nucleotide and predicted amino acid sequence of a cDNA clone encoding part of human transketolase. Biochem Biophys Res Commun 183:1159–1166

Baldauf SL, Palmer JD (1993) Animals and fungi are each other's closest relative: congruent evidence from multiple proteins. Proc Natl Acad Sci USA 90:11558–11562

Bernacchia G, Schwall G, Lottspeich F, Salamini F, Bartels D (1995) The transketolase gene family of the resurrection plant *Craterostigma plantagineum:* differential expression during the rehydration phase. EMBO J 14:610–618

Bernard N, Johnsen K, Holbrook JJ, Delcour J (1995) D175 dicriminates between NADH and NADPH in the coenzyme binding site of *Lactobacillus delbrueckii* subsp. *bulgaricus* D-lactate dehydrogenase. Biochem Biophys Res Commun 208:895–900

Biellmann J-F, Samama J-P, Bränden CI, Eklund H (1979) X-ray studies of the binding of Cibacron Blue F3GA to liver alcohol dehydrogenase. Eur J Biochem 102:107–110

Birktoft JJ, Rhodes G, Banaszak LJ (1989) Refined crystal structure of cytoplasmic malate dehydrogenase at 2.5 Å resolution. Biochemistry 28:6065–6081

Booth CK (1991) Studies on vitamin K and thiamin. PhD Thesis, The University of Queensland, Brisbane, Australia

Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford DL, Waddell PJ (1993) Partitioning and combining data in phylogenetic analysis. Syst Biol 42:384–397

Chen JH, Gibson JL, McCue LA, Tabita FR (1991) Identification, expression, and deduced primary structure of transketolase and other enzymes encoded within the form II CO_2 fixation operon of *Rhodobacter sphaeroides.* J Biol Chem 266:20447–20452

Datta AG, Racker E (1961) Mechanism of action of transketolase. J Biol Chem 236:617–623

de la Haba G, Leder IG, Racker E (1955) Crystalline transketolase from baker's yeast. Isolation and properties. J Biol Chem 214:409–426

de Sury d'Aspremont R, Toussaint B, Vignais PM (1996) Isolation of *Rhodobacter capsulatas* transketolase: cloning and sequencing of its structural tktA gene. Gene 169:81–84

Doolittle RF, Feng DF, Tsang S, Cho G, Little E (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. Science 271:470–477

Easteal S, Collet CC, Betty D (1995) The mammalian molecular clock. Austin, RG Landes, New York, Springer-Verlag, p 126

Eklund H, Nordström B, Zeppezauer E, Söderlund G, Ohlsson I, Boiwe T, Söderberg BO, Tapia O, Bränden CI, Åkeson A (1976) Three-dimensional structure of horse liver alcohol dehydrogenase at 2.4 Å resolution. J Mol Biol 102:27–59

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Felsenstein J (1989) PHYLIP 3.4 user manual. University of Washington, Seattle, USA

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topoogy. Syst Zool 20:406–416

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J, Dougherty BA, Merrick JM, McKenny D, Sutton G, FitzHugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu L, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, O'Smith H, Venter JC (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Delley JM, Fritchman JL, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb J, Doherty BA, Bott KF, Hu P, Lucier TS, Peterson SN, O'Smith Hamilton, Hutchison CA, Venter JC (1995) The minimal gene complement of *Mycoplasma genitalium.* Science 270:397–403

Galtier N, Gouy M (1994) Molecular phylogeny of eubacteria: a new multiple tree analysis method applied to 15 sequences data sets questions the monophyly of gram-positive bacteria. Res Microbiol 145:531–541

Galtier N, Gouy M (1995) Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci USA 92:11317–11321

Hashimoto T, Nakamura Y, Kamaishi T, Nakamura F, Adachi J, Okamoto K, Hasegawa M (1995) Phylogenetic place of mitochondrion-lacking protozoan, *Giardia lamblia*, inferred from amino acid sequences of elongation factor 2. Mol Biol Evol 12:782–793

Hawkins CF, Borges A, Perham RN (1989) A common structural motif in thiamin pyrophosphate-binding enzymes. FEBS Lett 255:77–82

Heinrich PC, Wiss O (1971) Transketolase from human erythrocytes: purification and properties. Helv Chim Acta 54:2658–2668

Horecker BL, Smyrniotis PZ, Klenow HJ (1953) The formation of sedoheptulose phosphate from pentose phosphate. J Biol Chem 205:661–682

Hosomi S, Tara H, Terada T, Mizoguchi T (1989) Inhibitory effect of 5-phosphoribosyl 1-pyrophosphate and ADP on the nonoxidative pentose phosphate pathway activity. Biochem Med Metab Biol 42: 52–59

Iida A, Teshiba S, Mizobuchi K (1993) Identification and characterization of the *tktB* gene encoding a second transketolase in *Escherichia coli* K-12. J Bacteriol 175:5375–5383

Jameson BA, Wolf H (1988) The antigenic index: a novel algorithm for predicting antigenic determinants. Comput Appl Biosci 4:181–186

Janowicz ZA, Eckart MR, Drewke C, Roggenkamp RO, Hollenberg CP (1985) Cloning and characterisation of DAS gene encoding the major methanol assimilatory enzyme from the methylotrophic yeast *Hansenula polymorpha.* Nucleic Acids Res 13:3043–3062

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–123

Jung EH, Sheu KFRE, Szabo P, Blass JP (1993) Molecular cloning, sequence and chromosome localization of human transketolase. GenBank Accession No. L12711, unpublished

Kato N, Higuchi T, Sakazawa C, Nishizawa T, Tani Y, Yamada H (1982) Purification and properties of a transketolase responsible for formaldehyde fixation in a methanol-utilising yeast, *Candida boidinii* (Kloeckera sp) No. 2201. Biochim Biophys Acta 715:143–150

Keese PK, Gibbs A (1992) Origins of genes: ''Big bang'' or continuous creation? Proc Natl Acad Sci USA 89:9489–9493

Kiely ME, Tan EL, Wood T (1969) The purification of transketolase from *Candida utilis.* Can J Biochem 47:455–460

Kim S, Kim B, Jeng J, Song BJ (1994) Characterisation of a DNA clone for rat transketolase: evidence for tissue-specific pretranslational activation in neonatal rat liver. GenBank Accession No. U09256, unpublished

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kimura M (1980) A simple method for estimating evolutionary rate of

base substitution through comparative studies of nucleotide sequences. J Mol Evol 16:111–120
Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England
Klein H, Brand K (1977) Purification and properties of transketolase from *Candida utilis.* Hoppe-Seyler's Z Physiol Chem 358:1325–1337
Kochetov GA (1982) Transketolase from yeast, rat liver and pig liver. Methods Enzymol 90:209–223
Kochetov GA (1986) Structure and mechanism of action of transketolase. Biokhimiya 51:2020–2029
Kochhar S, Hunziker PE, Leong-Morgenthaler P, Hottinger H (1992) Evolutionary relationship of NAD^+-dependent D-lactate dehydrogenase: comparison of primary structure of 2-hydroxy acid dehydrogenases. Biochem Biophys Res Commun 184:60–66
Kumar S, Tamura K, Nei M (1993) MEGA: molecular evolutionary genetics analysis, version 1.0. The Pennsylvania State University, University Park, PA 16802, USA
Lamzin VS, Aleshin EA, Strokopytov BV, Yukhnevich MG, Popov VO, Harutyunyan EH, Wilson KS (1992) Crystal structure of NAD-dependent formate dehydrogenase. Eur J Biochem 206:441–452
Langridge J (1991) Molecular genetics and comparative evolution. Research Studies Press, Taunton, England, p 216
Li W-H, Graur D (1991) Fundamentals of molecular evolution. Sinauer, Sunderland, MA, p 69
Li W-H, Tanimura M, Sharp PM (1987) An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol 25:330–342
Lindqvist Y, Schneider G, Ermler U, Sundström M (1992) Three-dimensional structure of transketolase, a thiamine diphosphate dependent enzyme, at 2.5 Å resolution. EMBO J 11:2373–2379
Lockhart PJ, Howe CJ, Bryant DA, Beanland TJ, Larkum AWD (1992) Substitutional bias confounds inference of cyanelle origins from sequence data. J Mol Evol 34:153–162
Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11:605–612
Masri SW, Ali M, Gubler CJ (1988) Isolation of transketolase from rabbit liver and comparison of some of its kinetic properties with transketolase from other sources. Comp Biochem Physiol [B] 90: 167–172
McCool BA, Plonk SG, Martin PR, Singleton CK (1993) Cloning of human transketolase cDNAs and comparison of the nucleotide sequence of the coding region in Wernicke-Korsakoff and non-Wernicke-Korsakoff individuals. J Biol Chem 268:1397–1404
Metzger MH, Hollenberg CP (1994) Isolation and characterization of the *Pichia stipitis* transketolase gene and expression in a xylose-utilising *Saccharomyces cerevisiae* transformant. Appl Microbiol Biotechnol 42:319–325
Murthy MRN, Garavito RM, Johnson JE, Rossmann MG (1980) The structure of lobster apo-D-glyceraldehyde-3-phosphate dehydrogenase at 3.0 Å resolution. J Mol Biol 138:859–872
Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York
Nikkola M, Lindqvist Y, Schneider G (1994) Refined structure of transketolase from *Saccharomyces cerevisiae* at 2.0Å resolution. J Mol Biol 238:387–404
O'hUigin C, Li W-H (1992) The molecular clock ticks regularly in muroid rodents and hamsters. J Mol Evol 35:377–384
Paoletti F (1983) Purification and properties of transketolase from fresh rat liver. Arch Biochem Biophys 222:489–496
Preparata G, Saccone C (1987) A simple quantitative model of the molecular clock. J Mol Evol 26:7–15
Philippov PP, Shestakova IK, Tikhomirova NK, Kochetov GA (1980) Characterisation and properties of pig liver transketolase. Biochim Biophys Acta 613:359–369
Radnis BA, Rhee DK, Morrison DA (1990) Genetic transformation in *Streptococcus pneumoniae:* Nucleotide sequence and predicted amino acid sequence of recP. J Bacteriol 172:3669–3674
Reizer J, Reizer A, Bairoch A, Saier MH Jr (1993) A diverse transketolase family that includes the RecP protein of *Streptococcus pneumoniae*, a protein implicated in genetic recombination. Res Microbiol 144:341–347
Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 232:584–599
Rzhetsky A, Nei M (1992) A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol 9:945–967
Rzhetsky A, Nei M (1994) METREE: a program package for inferring and testing minimum-evolution trees. Comput Appl Biosci 10:409–412
Saccone C, Pesole G, Preparata G (1989) DNA microenvironments and the molecular clock. J Mol Evol 29:407–411
Saccone C, Lanave C, Pesole G, Preparata G (1990) Influence of base composition on quantitative estimates of gene evolution. Methods Enzymol 183:570–583
Saccone C, Gissi C, Lanave C, Pesole G (1995) Molecular classification of living organisms. J Mol Evol 40:273–279
Saitou N, Nei M (1987) The neighbour-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425
Sanderson MJ (1990) Estimating rates of speciation and evolution: a bias due to homoplasy. Cladistics 6:387–391
Sarich VM, Wilson AC (1973) Generation time and genomic evolution in primates. Science 179:1144–1147
Schaaf-Gerstenschläger I, Zimmermann FK (1993) Pentose-phosphate pathway in *Saccharomyces cerevisiae:* analysis of deletion mutants for transketolase, transaldolase, and glucose 6-phosphate dehydrogenase. Curr Genet 24:373–376
Schäferjohann J, Yoo JG, Kusian B, Bowien B (1993) The cbb operons of the facultative chemoautotroph *Alcaligenes eutrophus* encode phosphoglycolate phosphatase. J Bacteriol 175:7329–7340
Schenk G (1996) Studies on the thiamin diphosphate-dependent enzymes transketolase and pyruvate decarboxylase. PhD Thesis, The University of Queensland, Brisbane, Australia
Schimmer BP, Tsao J, Czerwinski W (1996) Amplification of the transketolase gene in desensitization-resistant mutant Y1 mouse adrenocortical tumor cells. J Biol Chem 271:4993–4998
Skarzynski T, Moody PCE, Wonacott AJ (1987) Structure of hologlyceraldehyde-3-phosphate dehydrogenase from *Bacillus stearothermophilus* at 1.8 Å resolution. J Mol Biol 193:171–187
Smith DR (1994) Sequence of a cDNA clone from *Mycobacteriuim leprae* encoding transketolase. GenBank Accession No. U00013, unpublished
Smith MW, Feng DF, Doolittle RF (1992) Evolution by acquisition: the case for horizontal gene transfers. Trends Biochem Sci 17:489–493
Sprenger GA (1993) Nucleotide sequence of the *Escherichia coli* K-12 transketolase (*tkt*) gene. Biochim Biophys Acta 1216:307–310
Sprenger GA, Schörken U, Sprenger G, Sahm H (1995) Transketolase A of *Escherichia coli* K12. Purification and properties of the enzyme from recombinant strains. Eur J Biochem 230:525–532
Srere P, Cooper JR, Tabachnick M, Racker E (1958) The oxidative pentose-phosphate cycle. I. Preparation of substrates and enzymes. Arch Biochem Biophys 74:295–305
Sundström M, Lindqvist Y, Schneider G, Hellman U, Ronne H (1993) Yeast TKL1 gene encodes a transketolase that is required for efficient glycolysis and biosynthesis of aromatic amino acids. J Biol Chem 268:24346–24352
Teige M, Kopriva S, Bauwe H, Suess KH (1996) Primary structure of chloroplast transketolase from potato. GenBank Accession No. Z50099, unpublished
Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680
Van Den Bergh ERE, Baker SC, Raggers RJ, Terpstra P, Woudstra EC, Dijkhuizen L, Meijer WG (1996) Primary structure and phylogeny

of the Calvin cycle enzymes transketolase and fructosebisphosphate aldolase of *Xanthobacter flavus*. J Bacteriol 178:888–893

Villafranca JJ, Axelrod B (1971) Heptulose synthesis from nonphosphorylated aldoses and ketoses by spinach transketolase. J Biol Chem 246:3126–3131

Von Ossowski I, Hausner G, Loewen PC (1993) Molecular evolutionary analysis based on the amino acid sequence of catalase. J Mol Evol 37:71–76

Voskoboev AI, Gritsenko EA (1981) Nature of bond between coenzyme and protein in transketolase from porcine liver. Biokhimiya 46:1383–1388

Wainright PO, Hinkle G, Sogin ML, Stickel SK (1993) Monophyletic origins of the metazoa: an evolutionary link with fungi. Science 260:340–342

Waites MJ, Quayle JR (1981) The interrelation between transketolase and dihydroxyacetone synthase activities in the methylotrophic yeast *Candida boidinii*. J Gen Microbiol 124:309–316

Waltham M (1990) Studies on dihydrofolate reductase and transketolase. PhD Thesis, The University of Queensland, Brisbane, Australia

Wierenga RK, Terpstra P, Hol WGJ (1986) Prediction of the occurrence of the ADP-binding βαβ-fold in proteins, using an amino acid sequence fingerprint. J Mol Biol 187:101–107

Woese CR, Kandler O, Wheelis ML (1990) Toward a natural system of organisms: proposal for the domains Archae, Bacteria and Eukarya. Proc Natl Acad Sci USA 87:4576–4579

Wu C-I, Li W-H (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci USA 82:1741–1745