Rampant horizontal gene transfer and phospho-donor change in the evolution of the phosphofructokinase

Eric Bapteste\textsuperscript{a,} *, David Moreira\textsuperscript{b}, Hervé Philippe\textsuperscript{a,1}

\textsuperscript{a}Equipe Phylogénie, Bioinformatique et Génome, UMR CNRS 7622, Université Pierre et Marie Curie, 9 quai St. Bernard, 75005 Paris, France
\textsuperscript{b}Unité d’Ecologie, Systématique et Evolution, UMR CNRS 8079, Université Paris-Sud, 91405 Orsay Cedex, France

Received 1 April 2003; received in revised form 2 June 2003; accepted 24 June 2003

Received by A. Roger

Abstract

Previous work on the evolution of the phosphofructokinase (PFK) has shown that this key regulatory enzyme of glycolysis has undergone an intricate evolutionary history. Here, we have used a comprehensive data set to address the taxonomic distribution of the different types of PFK (ATP-dependent and PPi-dependent ones) and to estimate the frequency of horizontal gene transfer (HGT) events. Numerous HGT events appear to have occurred. In addition, we focused on the analysis of sites 104 and 124 (usually Gly\textsubscript{104} + Gly\textsubscript{124} or Asp\textsubscript{104} + Lys\textsubscript{124}), known to be involved in catalysis (J. Biol. Chem. 275 (2000) 35677). It revealed the existence of numerous sequences from distantly related species carrying atypical combinations of amino acids. Several adaptive changes of phospho-donors, probably requiring a single mutation at position 104, have likely occurred independently in many lineages. The analysis of this gene suggests the existence of a high rate of both HGT and substitution in its active sites. These rampant HGT events and flexibility in phospho-donor use illustrate the importance of tinkering in molecular evolution.

© 2003 Published by Elsevier B.V.

Keywords: Phosphofructokinase; Phylogeny; Evolution; Horizontal gene transfer; Adaptability

1. Introduction

Energy metabolism of many organisms is based on glycolysis. The classical glycolytic pathway (the Embden–Meyerhof pathway) is regulated by the phosphofructokinase (PFK). However, at least two alternative pathways to degrade glucose also exist. One of them, the Entner-Douderoff pathway, does not require PFK and is broadly distributed in bacteria (Conway, 1992). The other one is present only in some Archaea and requires an enzyme called ADP-PFK, which does not seem to be homologous to PFK. It belongs instead to the glucokinase (GLK) family of kinases (Verhees et al., 2001).

PFK is central to the classical glycolytic pathway, which is present in all three domains of life (Siebers et al., 1998), so that one could assume that the evolution of this enzyme is very constrained. However, this does not seem to be the case, since various types of PFK exist. They use distinct energy phospho-donors, such as ATP and inorganic pyrophosphate (PPI), and fulfill different tasks in the cell. The widely distributed ATP-PFK catalyses an irreversible catalytic reaction, the phosphorylation of fructose-6-phosphate to fructose-1,6-bisphosphate, while the PPI-PFK, also called PPI-phosphofructokinase (PFP), catalyses the same reaction in a reversible way and can thus function both in glycolysis and gluconeogenesis. Furthermore, in many cases, the physiological role of PFP is not obvious (Siebers et al., 1998). More precisely, the simultaneous presence of PFP and ATP-PFK (i.e., of a reversible and an irreversible enzyme) in a single organism, even if rarely reported, has suggested that PFP may perform an alternative unknown function (Alves et al., 2001; Van Praag, 1997).
Biochemical approaches, including functional, structural, and mutational analyses, have led to the determination of several amino acids essential for ATP-PFK and PFP functions (Chi and Kemp, 2000; Moore et al., 2002), in particular, the positions 104 and 124 (according to the numbering of Escherichia coli ATP-PFK) (Chi and Kemp, 2000). First, it has been demonstrated that Gly at position 104, present in all ATP-PFK, is essential for the use of ATP (Chi and Kemp, 2000; Moore et al., 2002; Van Praag, 1997). By contrast, the Asp residue present at this position in PFP prevents ATP from binding by sterical hindrance (Moore et al., 2002). Second, when the position 124 in E. coli PFK is occupied by Gly, the absence of a lateral chain allows room for the alpha-phosphate of ATP. This is not the case when a larger amino acid, Lys, is present at this position, as it occurs in all characterised PFP (Hinds et al., 1998). In fact, this Lys plays an important role in the recognition of PPI, which has been biochemically tested in Entamoeba histolytica and in several other organisms (Chi and Kemp, 2000; Hinds et al., 1998; Lopez et al., 2002; Moore et al., 2002).

In summary, PFK working with ATP harbours Gly at positions 104 and 124, while PFP has an Asp and a Lys (Chi and Kemp, 2000; Claustre et al., 2002; Moore et al., 2002). However, the atypical amino-acidic combination Gly$_{104}$ and Lys$_{124}$ has been reported in E. histolytica (Chi and Kemp, 2000), Trypanosoma brucei (Claustre et al., 2002), Leishmania donovani (Lopez et al., 2002), and Chlamydia trachomatis (Moore et al., 2002). For the three first cases, biochemical characterisation indicates that these PFK use ATP as a phospho-donor, and not PPI. Moore et al. (2002) argued that this same phospho-donor is likely used also in C. trachomatis.

ATP-PFK and PFP share a common ancestry, but phylogenies show a very complex evolutionary pattern (Müller et al., 2001; Siebers et al., 1998). An ancient duplication has been supposed to give birth to two groups of PFK, using PPI and ATP as phospho-donor, respectively (Alves et al., 1996). However, their monophyly is questioned (Müller et al., 2001; Siebers et al., 1998) as well as the validity of the amino acid content in sites 104 and 124 as a phylogenetic signature. In fact, the evolutionary history of PFK does not coincide in several points with accepted notions of organismic relationships. On the contrary, several duplications and horizontal gene transfer (HGT) blur the phylogeny of PFK (Müller et al., 2001). Some recent duplications were followed by the differentiation of catalytic and regulatory subunits in animals, fungi, and plants (Heinisch et al., 1989; Kemp and Gunasekera, 2002; Poorman et al., 1984; Van Praag, 1997). In addition, the fusion of the catalytic and regulatory subunits in fungi, animals, and Dictyostelium led to huge variation in sequence size of PFK. All these duplications and HGT events can eventually lead to multiple divergent copies of PFK in a single species.

In this work, we have analysed a large data set of ATP-PFK and PFP sequences to study the distribution of the different types of PFK (ATP-PFK, PFP, and those with atypical active sites) and to estimate the frequency of HGT events. We include a new sequence from the choanoflagellate Monosiga ovata providing an additional example of HGT affecting eukaryotic species. We also looked carefully at positions 104 and 124 and detected several atypical active sites (Gly$_{104}$ + Lys$_{124}$) in distantly related species. Hence, we suggest that adaptive changes of phospho-donors, requiring a mutation at position 104, have occurred independently in many lineages. The activity of these particular proteins should be biochemically tested, and if they possess the expected ATP-PFK activity, species owning simultaneously ATP-PFK and PFP in their genomes would be in fact quite common.

2. Materials and methods

2.1. Sequencing

The sequence of the pfk gene from M. ovata was obtained from random sequencing of a cDNA library (collaboration with Dr. P. Holland, to be published elsewhere). Both 5' and 3' ends of the pfk clone were sequenced, allowing and overlapping of around 500 nucleotides for a total cDNA length of 897 nucleotides. The sequence has been submitted to GenBank under the accession number AY291291.

2.2. Sequence recovery and alignment

Most PFK protein sequences were retrieved from GenBank using the program ALIBABA (Philippe Lopez, unpublished work). The sequences from Dictyostelium discoideum, fungi, and animals (formed by the fusion of the catalytic and regulatory subunits) were split in two parts to align separately the catalytic and the regulatory regions. The sequences were aligned with CLUSTAL W (Thompson et al., 1994) and the alignment was manually refined with the program ED of the MUST package (Philippe, 1993).

To construct a more comprehensive data set, we also included sequences obtained from ongoing expressed sequence tag (EST) and genome projects. PFK homologues were detected by TBLASTN search (Altschul et al., 1997). All of the high-scoring segments with a BLAST score below $10^{-10}$ were retained and incorporated into the alignment.

Only a few regions could be aligned without ambiguity for the entire data set, notably those around the active site of the enzyme. Our complete data set contained 227 sequences. The alignment is available upon request.

2.3. Phylogenetic analysis

The complete PFK data set was initially analysed by using the neighbour joining (NJ) method (Saitou and Nei, 1987). Partial and/or phylogenetically very closely related sequences were discarded, yielding a final alignment that
contained 152 sequences and 153 unambiguously aligned positions.

A phylogenetic tree with the 152 representative sequences was then reconstructed using TREE-PUZZLE 5.0 (Schmidt et al., 2002) and Neighbor (Felsenstein, 1999). To handle rate variation among sites, a maximum likelihood (ML) distance matrix with a $\Gamma$ law model (eight discrete classes) was computed and then used to reconstruct the tree by the NJ method. Bootstrap values were computed upon 1000 replicates using PUZZLEBOOT (www.tree-puzzle.de/puzzleboot.sh). An extended majority rule consensus tree from the replicates was inferred by CONSENSE from the PHYLIP package (Felsenstein, 1999). This NJ approach allows to work on a vast number of sequences using a complex model of sequence evolution, which is impossible by a standard maximum likelihood (ML) analysis.

Statistical comparisons of alternative tree topologies were carried out by applying the Shimodaira’s (2002) approximately unbiased (AU) test implemented in the program CONSEL (Shimodaira and Hasegawa, 2001).

3. Results

3.1. An odd taxonomic distribution

Our taxonomic sample contained almost exclusively bacteria and eukaryotes, with the exception of the sequence of the archaeon Thermoproteus tenax, most likely acquired by HGT (Siebers et al., 1998). All completely sequenced eukaryotic genomes contain a pfk gene, whereas it is missing in some bacterial and in all archaeal completely sequenced genomes. A BLAST search (using standard parameter values) shows that PFK is not present (with an expectation value threshold at $10^{-30}$) in the complete genome sequences of some alpha-proteobacteria (Brucella melitensis, Caulobacter crescentus, Rickettsia conorii, Rickettsia prowazekii), some beta-proteobacteria (Neisseria meningitidis, Ralstonia solanacearum), some gamma-proteobacteria (Pseudomonas aeruginosa, Xanthomonas campestris, Xanthomonas citri), epsilon-proteobacteria (Campylobacter jejuni, Helicobacter pylori), the Gram-positive Oceanobacillus iheyensis, the Green sulfur bacterium Chlorobium tepidum, the fusobacterium Fusobacterium nucleatum, and the cyanobacterium Thermosynechococcus elongatus. Several of these species lacking PFK (H. pylori, N. meningitidis, R. solanacearum, S. typhimurium, and X. citri) possess a KDPG aldolase, central in the Entner-Doudoroff pathway.

3.2. Numerous HGT events in the PFK tree

A comprehensive data set (152 sequences, 153 unambiguously aligned positions) was used to construct a phylogenetic tree (Fig. 1). It allowed the definition of eight monophyletic groups, five of them containing both bacterial and eukaryotic sequences. Five of these groups (X, P, SHORT, LONG, and III) have been previously named by Siebers et al. (1998) and Müller et al. (2001). In this work, we name three additional groups: B1 (containing only bacteria), B2 (containing Clostridium spp. and several proteobacteria), and E (containing mainly eukaryotes). In addition, four sequences (Aquifex aeolicus, Dictyoglosum thermophilum, T. tenax, and one sequence from Thermotoga maritima) do not belong to any of the eight monophyletic groups. The phylogenetic tree is very complex, suggesting an intricate evolutionary history involving duplications and HGT in addition to vertical inheritance.

The finding of several monophyletic groups with a wide and phylogenetically coherent taxonomic composition (e.g., the bacterial clade B1 and the eukaryotic clade E) is in agreement with an extensive vertical inheritance. Moreover, phylogenetic relationships within these groups are coherent with accepted phylogenetic groups based on ribosomal RNA and protein markers (Baldauf et al., 2000; Brochier et al., 2002; Embley et al., 1992; Van de Peer et al., 2000). For instance, metazoa are sister group to fungi in group E, and gamma proteobacteria emerge within a well-supported clade. Within group B1, we found both the monophyly of Thermus + Deinococcus and that of Cytophaga + Bacteroides.

Yet, other regions of the tree are more puzzling. Some species that are thought to be closely related are separated among several groups (e.g., the alpha-proteobacteria in groups B1, B2, P, and III). In other cases, some groupings were recovered with unusual relationships inside a group (e.g., spirochetes forming a monophyletic group with the choanoflagellate M. ovata and plants in group X). This may have resulted from ancient duplication events followed by differential gene losses. However, the number of duplications and losses needed to explain that the complete phylogeny would probably be very large, so that at least for a number of cases, independent HGT events seem a parsimonious explanation. One of the clearest examples of HGT concerns the choanoflagellate M. ovata, expected to branch in clade E as sister of metazoa (Lang et al., 2002) but which emerges as relative of spirochetes in clade X (Fig. 1).

HGT may also concern species with several pfk copies (see numbers on the right of the species names in Fig. 1). For instance, a delta proteobacterium Desulfitobacterium hafniense and a low GC Gram-positive bacterium Clostridium perfringens harbours three gene copies, branching in groups B1, B2, and III. Some eukaryotes also have multiple pfk gene copies. However, in some cases, they arose from duplication events, such as in Dictyostelium, fungi, and metazoa forming two paralogous clusters within the group E, as supported by the tree topology. In other cases, one of the copies has probably been acquired by HGT (e.g., E. histolytica, present in groups X and LONG). More complex scenarios simultaneously involving HGT and gene duplications are also found. For instance, in group LONG, plants, apicomplexa (Plasmodium and Cryptosporidium), and chlamydiales have at least two copies. One possible interpreta-
tion is to invoke an ancient duplication in the ancestor of plants and apicomplexa (or two independent duplications in these two lineages) and the subsequent acquisition of the two paralogous copies by the chlamydialae. Finally, species for which a single copy has been identified are also involved in HGT, such as the cases of *T. tenax* and *Mastigamoeba balamuthi* (Müller et al., 2001; Siebers et al., 1998).

Our phylogenetic tree suggests several HGT events. However, some of them have to be tested because of the lack of resolution of certain regions in the tree. We have therefore analysed alternative tree topologies that minimise the number of HGT events. These topologies concerned nodes supported by bootstrap values <90%, since we have assumed that nodes with superior bootstrap values were correctly inferred. In particular, we have analysed three alternative topologies affecting the groups that show a mixture of eukaryotic and prokaryotic sequences, namely, groups E, LONG, and X. In group E, we have constrained the position of *Chloroflexus aurantiacus* at the base of the group; in group LONG, we have constrained the monophyly of the eukaryotic sequences, and in group X, we have constrained the monophyly of the bacterial sequences. Shimodaira’s approximately unbiased test confirmed that these constrained topologies minimising HGT were significantly worse than the preferred tree. Therefore, it appears that HGT events are a likely explanation for the complex phylogeny of PFK.

### 3.3. A complex evolution of PFK function

This complicated PFK phylogeny that seems largely blurred by HGT events, gene duplications, and gene losses provides the framework to understand the relationships between ATP-PFK and PFP as well as phospho-donor transitions in PFK. Sensu stricto ATP-PFK, classically defined on the base of their sequence conservation, notably by the possession of two glycines in the active site, is monophyletic. They were all included into the groups B1 and E. Similarly, four other monophyletic groups in the tree (B2, P, SHORT, and X) showed a homogeneous amino acid mixture of eukaryotic and prokaryotic sequences, namely, groups E, LONG, and X. In group E, we have constrained the position of *Chloroflexus aurantiacus* at the base of the group; in group LONG, we have constrained the monophyly of the eukaryotic sequences, and in group X, we have constrained the monophyly of the bacterial sequences. Shimodaira’s approximately unbiased test confirmed that these constrained topologies minimising HGT were significantly worse than the preferred tree. Therefore, it appears that HGT events are a likely explanation for the complex phylogeny of PFK.

**G*K** in Fig. 1) and PFK using PPI probably appeared several times in evolution. Consequently, these amino acid combinations and, likewise, the use of ATP or PPI as phospho-donor do not define reliable phylogenetic signatures. It is also possible to envisage that the transition from one form to the other has occurred several times.

Nevertheless, amino acids of these two sites display very little polymorphism, and their evolutionary conservation, covariation, and involvement in phospho-donor recognition have been discussed in several works (Chi and Kemp, 2000; Claustre et al., 2002; Lopez et al., 2002; Moore et al., 2002). In most catalytic PFK sequences, the position 104 is either a Gly or an Asp, while position 124 is either a Gly or a Lys. The exceptions were rare (see Fig. 1). In fact, in our data set, the three sequences of the regulatory alpha subunits of plants (Van Praag, 1997) and the two closely related sequences from apicomplexa contained other amino acids, which would be coherent with the fact that the alpha-subunits do not link to phospho-donors. The other exceptions were the sequences from *Chlamydomonada psitacica*, *C. pneumoniae*, and *Chlamydia muridarum* (Asn at position 104), from *A. aeolicus* (Cys at position 124), and from *T. maritima* (Ala at position 124).

### 4. Discussion

Despite being a key enzyme involved in glycolysis regulation, PFK has a very complex evolutionary history. The presence of this protein in several bacterial species whose genomes have been completely sequenced but its absence in closely related species (e.g., the alpha-proteobacteria) strongly suggests that PFK can be lost. However, the loss of PFK is not synonymous of the loss of glycolysis (for instance, in the species harbouring a KDPG aldolase, central in the Entner-Douderoff pathway). Hence, since PFK can be lost, it is not surprising to find a phylogenetic tree with odd relationships. Such independent losses make the task of identifying paralogous more difficult and could allow subsequent LGT of PFK in the species secondarily devoid of PFK.

Moreover, flexibility in the use of PFK also resulted from independent acquisitions of new PFKs by HGT. Our tree illustrates that a key enzyme can be obtained from another species. Indeed, HGT led to a complex phylogeny with groups at odds with the classical relationships (Müller et al., 2001). This could occur if the native PFK is replaced by a new one or if the original copy is conserved together with

---

**Fig. 1.** Unrooted NJ tree for 152 PFK sequences and 153 amino acid positions based on distances calculated with a J law correction. Eukaryotic species are in bold, while prokaryotic species are in italic. Numbers at nodes are bootstrap values (only values >50% are shown). Monophyletic groups are named according to Müller et al. (2001) and this work. Solid circles indicate the most parsimonious distribution of Gly104 to Lys124 mutations on the tree. Numbers after species names indicate species with multiple PFK copies in distantly related groups. The various amino acid combinations at positions 104 and 124 are reported after Müller et al. (2001; Siebers et al., 1998).
the new copy, leading to species harbouring distantly related PFK (e.g., in *Amycolatopsis methanolica*). In many cases, conservation of two distantly related PFK enzymes in a single species could be further explained in terms of adaptability, each copy allowing the use of either a phospho-donor or the other, potentially enhancing the fitness of the species. Such species have a PFK using ATP and a PFK using PPI, one of which may have been obtained by HGT. In fact, species with both enzymes would be or would have been able to initiate glycolysis either with ATP or with PPI, which may be adaptive, according to the relative concentrations of these metabolites in the cell and its environment. For example, parasites may be advantaged by a PFK using PPI to economise their limited stock of ATP (Moore et al., 2002). Moreover, supplementary PFK copies may be recruited in alternative metabolic pathways. For instance, ATP-PFK participate in the RuMP cycle in *Amycolatopsis* (Alves et al., 2001).

As a single-point mutation can induce a change of the phospho-donor (Chi and Kemp, 2000), one can imagine that the opposite mutation should revert it, allowing the PFK to switch from the use of PPI to the use of ATP and vice versa. Our phylogenetic tree suggests that these changes of phospho-donor concerned several species. One example can be found within group III, containing a monophyletic group of Gram-positive bacteria showing a mixture of sequences with either Gly104 or Asp104. Moreover, the direction of these changes can be suggested. It is most parsimonious to postulate that Gly$_{104}$ + Lys$_{124}$ PFK using ATP was ancestral and that PFK using PPI evolved more recently, in seven independent occasions (i.e., once in *T. tenax*, once in group LONG before the emergence of *Borrelia burgdorferi*, once in the ancestor of group SHORT, once in the ancestor of groups B2 and P, and three times in the group III, in the ancestor of *Thermobifida fusca*, *Streptomyces coelicolor*, and *Amycolatopsis mediterranei*). This suggests that at least seven species have evolved from an irreversible glycolysis towards a reversible one, whereas the opposite appears to have never occurred in nature (Fig. 1). Yet, this result is very sensitive to our still limited taxonomic sample and would need additional confirmation.

We show here that the presence of multiple PFK copies with different active sites is a common situation in nature. It is notably the case for all the plants, some apicomplexa (*Cryptosporidium parvum*, *Plasmodium falciparum*), some anoeba (*E. histolytica*), some high GC Gram-positive bacteria (*S. coelicolor*), some alpha-proteobacteria (*Magnetococcus* sp., *Magnetospirillum magnetotacticum*), some low GC Gram-positive bacteria (*D. hafniense*), some spirochetes (*B. burgdorferi*, *Treponema pallidium*), and green sulfur bacteria (*C. aurantiacus*). Thus, in these species, multiple PFK copies may be involved in different metabolic functions.

The complex phylogeny of PFK illustrates that tinkering in evolution can concern fundamental molecular processes. Our study emphasises the potential adaptability of this key enzyme—lost, gained, and modifiable by single-point mutations—in living beings. Finally, our results complicates the debate on the nature of the ancestral phospho-donor used by the PFK, which opposes the advocates of the emergence of metabolism in an organic or an inorganic context (see Siebers et al., 1998; Chi and Kemp, 2000, for two opposed opinions). We suggest that in several cases, the use of PPI as phospho-donor may be derived. However, in addition to possible convergent adaptive mutations, HGT events, including those of mutated sequences, further complicate the distribution of these enzymes. Hence, it is not possible with this tree to conclude which phospho-donor, ATP, or PPI was ancestrally used by PFK.

**Acknowledgements**

We thank Simonetta Gribaldo and Miklós Müller for critical reading of the manuscript, and Céline Brochier for advice on bacterial phylogeny. We thank Peter Holland and Elizabeth Snell for the *M. ovata* PFK sequence. Genome and EST sequence data of the species *E. histolytica*, *Oryza sativa*, *C. parvum*, and *P. falciparum* were obtained from databases deposited at GenBank (www.ncbi.nlm.nih.gov/dbGSS/index.html and www.ncbi.nlm.nih.gov/dbEST/index.html). *Neurospora crassa* EST data were obtained from the Center for Genome Research (www-genome.wi.mit.edu/annotation/fungi/neurospora). *Bacteroides fragilis*, *Clostridium difficile*, *Clostridium botulinum*, *Staphylococcus aureus*, *Streptococcus equisimilis*, *Yersinia enterocolitica*, and *Corynebacterium diphtheriae* sequence data were obtained from the Sanger Institute (www.sanger.ac.uk/DataSearch/). *Treponema denticola* and *Chlamydomphila psittaci* sequence data were obtained from the Institute for Genomic Research (www.tigr.org/tdb/mdb/mdbcomplete.html). *Cytophaga hutchinsonii*, *D. hafniense*, *C. aurantiacus*, *M. magnetotacticum*, *Nostoc punctiforme*, and *T. fusca* sequence data were obtained from the DOE Joint Genome Institute (www.science.doc.gov/ober/EPR/mig_cont.html).

**References**


