Reconstructed Ancestral Myo-Inositol-3-Phosphate Synthases Indicate That Ancestors of the Thermococcales and Thermotoga Species Were More Thermophilic than Their Descendants

Nicholas C. Butzin*, Pascal Lapierre, Anna G. Green, Kristen S. Swithers, J. Peter Gogarten, Kenneth M. Noll

Department of Molecular and Cell Biology, University of Connecticut, Storrs, Connecticut, United States of America

Abstract

The bacterial genomes of Thermotoga species show evidence of significant interdomain horizontal gene transfer from the Archaea. Members of this genus acquired many genes from the Thermococcales, which grow at higher temperatures than Thermotoga species. In order to study the functional history of an interdomain horizontally acquired gene we used ancestral sequence reconstruction to examine the thermal characteristics of reconstructed ancestral proteins of the Thermotoga lineage and its archaeal donors. Several ancestral sequence reconstruction methods were used to determine the possible sequences of the ancestral Thermotoga and Archaea myo-inositol-3-phosphate synthase (MIPS). These sequences were predicted to be more thermostable than the extant proteins using an established sequence composition method. We verified these computational predictions by measuring the activities and thermostabilities of purified proteins from the Thermotoga and the Thermococcales species, and eight ancestral reconstructed proteins. We found that the ancestral proteins from both the archaeal donor and the Thermotoga most recent common ancestor recipient were more thermostable than their descendants. We show that there is a correlation between the thermostability of MIPS protein and the optimal growth temperature (OGT) of its host, which suggests that the OGT of the ancestors of these species of Archaea and the Thermotoga grew at higher OGTs than their descendants.

Introduction

From the publication of the first genome sequence of a member of the bacterial order Thermotogales, that of Thermotoga (Tt.) maritima, the importance of gene sharing with the Archaea was apparent for this lineage [1]. Subsequent examinations of the genomes from the other Thermotogales showed that these organisms have extensively shared genes with the Archaea through horizontal gene transfer (HGT) [2]. Many of these archaeal genes were derived from the Thermococcales, mostly represented in modern species as hyperthermophiles with optimal growth temperatures (OGTs) above those of Thermotogales species. Consequently, the genes inherited by the Thermotogales likely encoded proteins catalytically active at temperatures higher than that at which modern Thermotogales grow. The ancestral Thermotogales that inherited these genes likely grew at temperatures higher than modern species and so was suited to use them [2]. Genes acquired from the Thermococcales are largely found in the Thermotoga species, the species of Thermotogales with the highest OGTs. Thus, to examine the evolution of proteins acquired from the Archaea, we chose to reconstruct an ancestral protein, a myo-inositol-3-phosphate synthase (MIPS), shown to have been acquired by Thermotoga species from the Archaea (Nesbo et al. 2001).

MIPS is an essential enzyme in the Tt. maritima biosynthetic pathway for the compatible solute di-myositol-1,1-phosphate [3]. MIPS is common in both euryarchaeotes and crenarchaeotes, but is rarely found in Bacteria [4]. There is strong support that the MIPS gene and other genes, which allow for synthesis of myo-inositol phosphate, originated in an archael lineage [4–7]. Tt. maritima accumulates several types of inositols at superoptimal temperatures [8]. Recently, Gonçalves and co-workers showed that the Thermotoga ancestor acquired several genes for synthesis of various inositols [4]. They also determined that the enzymes used to synthesize myo-inositol-phosphate are limited to Bacteria and Archaea with OGTs above 55°C [4]. The importance of inositols for growth at high temperatures underscores the selective advantage that Thermotoga may have gained by acquiring a MIPS gene from the euryarchaeota.

To trace the functional history of the Thermotoga MIPS, we used ancestral sequence reconstruction (ASR) to reconstruct...
ancestral proteins at specific nodes of the MIPS phylogenetic tree. ASR predicts the sequences of ancestral proteins based on those of their modern homologs using phylogenetic relationships and statistical approaches [9–11]. Aligned sequences are used to predict the most probable amino acid residues in ancestral proteins, which can then be reconstructed by synthesizing the predicted genes and expressing recombinant proteins in a suitable host [9–11]. In previous studies ASR was used to predict properties of ancient life by analyzing reconstructed proteins from Eukaryotes [12–14], Archaea [15], and Bacteria [9,10,15,16].

Since ASR has an established record of predicting properties of ancient life, here we use it to test the hypothesis that the ancestral Thermotoga MIPS is more thermal stable and catalytically active at a higher temperature than extant Thermotoga MIPS proteins. In addition, we tested the related hypothesis that reconstructed archaeal MIPS proteins are also active and stable at higher temperatures than their modern descendants. To test these hypotheses we reconstructed ancestral proteins for both the Thermotoga and the Archaea nodes of the MIPS phylogenetic tree and examined their biochemical properties. We compared these with other MIPS proteins derived from extant species. The results of these tests are consistent with the hypothesis that the ancestors of the Thermotoga and the Thermococcales species were higher temperature hyperthermophiles.

Results and Discussion

Thermotoga Myo-inositol-3-phosphate Synthase (MIPS) was Inherited from Archaea

In order to examine the characteristics of ancestral proteins inherited by interdomain HGT, the Thermotoga genomes were screened for a gene of archaeal origin whose protein product met the following criteria. First, only soluble proteins were considered to avoid the potential difficulties associated with membrane protein expression, purification, and solubilization. Second, enzymes with easily measured activities were sought to allow facile functional characterization of the proteins’ activities. Third, only small, monomeric enzymes were acceptable so that only a single gene need be constructed. Fourth, proteins with known crystal structures were desirable to identify residues important for activity and to guide selection of amino acids important for catalysis. Finally, proteins were only considered if they had a clear phylogenetic signal indicating horizontal inheritance from Archaea. Based on these criteria, we chose to examine the thermal characteristics of Thermotoga myo-inositol-3-phosphate synthase (MIPS; EC 5.5.1.4) [17], a protein that shows a clear phylogenetic origin in the Archaea.

A previous study showed that the Thermotoga genus acquired a MIPS gene from an archaeal lineage [10]. The number of sequenced genomes has grown significantly since that work was published and more sequenced MIPS genes are available (Table S1), which provided an opportunity to reexamine these relationships. An updated phylogenetic analysis of Thermotoga MIPS revealed further support for its origin in the Archaea, specifically among the Euryarchaeota (Figure S1). The distribution of the MIPS protein within the Thermotogales is unique to the Thermotoga genus. Together the phylogeny and the distribution of the MIPS gene within the Thermotogales strongly support that the MIPS gene was transferred to the Thermotoga ancestor, rather than transferred to the common ancestor of all Thermotogales and then independently lost in the other lineages (Figure 1 and S1).

Prediction of Ancestral Thermotoga and Archaea MIPS Proteins

Ancestral sequence reconstruction (ASR) was used to reconstruct ancestral proteins at specific nodes of the MIPS phylogenetic tree. Several different methods of ASR were compared to determine which method resulted in minimal apparent biases for the MIPS dataset. The ancestral sequences were predicted using Ancecon [19] with a maximum likelihood (ML) tree calculated using PhyML as reference for the reconstruction. Ancestral proteins were predicted for three strongly supported nodes, the most recent common ancestor (MRCA) to the Thermotoga (node T), the MRCA to the Thermusococcus node (node C), and the node where nodes T and C and other Archaea (node A) branches join (Figure 1). One potential problem with Ancecon is that it assumes a homogeneous amino acid composition throughout the tree, which may produce ancestral proteins with amino acid compositions artificially biased to those of the extant proteins.

This bias was examined by generating the ancestral proteins using BppAncestor [20] and a non-homogeneous substitution model that allowed for a separate set of parameters for each clade. The predicted ancestral sequences using BppAncestor (non-homogeneous substitution model) and Ancecon (homogeneous) were quite similar and only varied slightly. There is no significant difference between the BppAncestor (non-homogeneous substitution model) and the final four reconstruct sequences (ATM_T1-4; explained below) IVWREL value (see next section for correlations) with means of 0.4083 and 0.4110, respectively (p = 0.998; t-test). This shows there is no composition bias toward the extant sequences when the Ancecon is used on our dataset; thus, the homogeneous model was used in further analyses.

Another potential problem stemming from a simple ML approach to ASR is it may lead to a bias towards more thermostable proteins for some datasets [21]. We employed a Bayesian method that samples from a posterior distribution to estimate the ancestral amino acid composition. Such a method is not prone to the tendency of ML methods to reconstruct ancestral proteins with a higher thermostability [21]. Ancestral sequences were predicted using BppAncestor with both the non-homogeneous model described above, and a homogeneous implementation of that same model. For each ancestral sequence, 1000 replicates were determined by sampling from the posterior distribution. As a quantitative marker of thermostability, we used the total fraction of seven amino acids (IVYWREL) for each sequence [22] (see next section for correlations). The mean and confidence interval of the IVYWREL values for the ancestral sequences were calculated from these replicates (Table S2). The IVYWREL values for the ML based ancestral reconstruction are within a 95% confidence interval determined through sampling from the posterior probabilities, indicating that our dataset is not prone to the bias mentioned by [21]. This method results in a large number of potential reconstructed proteins making it impractical to produce proteins for biochemical analysis.

However, our results suggest that the ML approach applied to this dataset does not have a strong basis towards a more thermostable ancestral reconstruction.

Taken together all these analyses show that Ancecon does not introduce an apparent bias towards generating thermostable proteins, nor does it introduce compositional biases to the extant sequences for our dataset. To account for variance in tree topology, (nodes with <70 bootstraps were considered to be weakly supported, Figure 1), Ancecon was used for ASR of sequences at nodes A, C, and T (Figure 1) generated from 1000 Bayesian trees (See Materials and Methods). These analyses resulted in four probable ancestor proteins predicted at node T,
called ancestral *Thermotoga* MIPS, ATM (ATM_T1-4). The reconstructed proteins for node C were called ancestral *Thermococcus* MIPS, ACM (ACM_C1-2), and the reconstructed proteins for node A were called ancestral *Archaea* MIPS, AAM (AAM_A1-2).

**In silico Analyses of MIPSs Show that the Ancestral Thermotoga MIPSs are more Thermostable**

A genome scale correlation has been shown between the OGT of an organism and the total fraction of seven amino acids (IVYWREL) in its soluble proteins [22]. To determine if this relationship holds for a single protein family a large sampling of MIPS proteins IVYWREL values were compared to their respective organisms' OGTs (Figure S1). The comparison of MIPS proteins from Figure S1 [Pearson correlation coefficient ($r_{pc}$) = 0.80; Figure S2] resulted in a similar pattern as observed previously at the genome scale. This suggested that IVYWREL values can be used as an indication of thermostability of MIPS proteins (Figure S2). Additionally the OGT of the organisms in Figure 1 are correlated to their MIPS IVYWREL ($R^2 = 0.85$; Figure S3A). This pattern suggests that the previous correlation observed between OGT and IVYWREL on a genome scale holds for the smaller set of MIPS proteins as well, and that IVYWREL can be used as a general indicator of protein thermostability.

IVYWREL values were calculated for all reconstructed ancestral proteins. ATM_T1-4 all have higher IVYWREL values than the average of the extant *Thermotoga* MIPSs ($p<0.001$, Table S3; Figure 1) suggesting greater thermostability of the ancestral *Thermotoga* MIPSs than modern *Thermotoga* MIPSs. Sequences ACM_C1-2 have similar IVYWREL values as the extant *Thermococcus* MIPSs of node C ($p>0.7$), AAM_A1-2 sequences are all higher in their IVYWREL values than that of sequences from Node T and Node C (Figure 1), and ATM_T1-4 ($p<0.001$; Table S3).

The *in silico* analyses suggests that the *Thermotoga* and *Thermococcus* ancestors may have been capable of living in hotter environments than their descendants do now. This contention is based on the fact that the reconstructed proteins at each node are predicted to be more thermostable than their descendant proteins. Since the thermal characteristics of extant MIPS proteins show a correlation with the OGTs of their hosts, we surmise that the hosts of our reconstructed ancestral proteins had OGTs higher than their descendants.

One might argue that our reconstructed *Thermotoga* MIPS has characteristics of its archaeal donor, not its new *Thermotoga* host. That would be the case if the extant genes descended from the newly arrived gene before it ameliorated to its new *Thermotoga* host. To support our argument that this is not the case, we compared the sequences of our reconstructed MIPS proteins with those of extant *Thermotoga* and *Thermococcus* MIPS sequences to look for evidence of the bacterial rather than archaeal nature of our reconstructed *Thermotoga* MIPSs. When we compared the amino acid compositions of the extant *Thermococcus* and *Thermotoga* sequences we found that amino acids H, E, D, N, T, Y, A, and G showed significantly different abundances ($p<0.01$; Figure 2 and Table S4). Using these differences, we can distinguish an archaeal from a bacterial origin. This demonstrates that our reconstructed *Thermotoga* MIPSs...
The MIPS proteins.

significantly different from that of ATM_T1-4 proteins values (0.4110)
with the with the ‘other’ most likely residues does not cause a
ATM_T1-4
are representative of the Thermotoga host-ameliorated MIPS
sequence and so reflect the thermal characteristics of that host
rather than the donor.

We used similar comparisons to assess the changes that have
taken place in each lineage. The average Thermococcus MIPS
sequence is not significantly different from any of the reconstructed
Thermococcus ACM MIPSs (Table S4). This suggests that the
extant Thermococcales MIPSs have not changed significantly
from their ancestral state. By contrast, the average Thermotoga
MIPS showed significant differences in abundances of amino acids
R, F, E, and G, and sometimes in I and T when compared with the
reconstructed Thermotoga ATM MIPSs (Table S4). The composition of these proteins have apparently changed from their
ancestral state, but not in the same manner as the Thermococcus
sequences, so the ATM proteins are not descended directly from a
Thermococcus donor sequence. It is not surprising that there have
been minor changes in the descendants because they grow at
temperatures slightly lower than their common ancestral
state, but not in the same manner as the
composition of these proteins have apparently changed from their
ancestral state. By contrast, the average
Thermotoga MIPSs showed significant differences in abundances of amino acids
from their ancestral state. By contrast, the average
Thermotoga MIPSs showed significant differences in abundances of amino acids
Thermotoga
Thermococcus
ATM_T1-4
Thermotoga
Thermococcus

IVYWREL values as an indicator of thermostability. The MIPSs

Table 1. Measures of thermostability of extant and ancestral MIPSs.

| Organism          | Protein | OGT (°C) | $T_{opt}$ (°C) | $T_m$ (°C) | pH 7.0 | pH 4.2 | pH 3.5 | IVYWREL value |
|-------------------|---------|----------|----------------|------------|--------|--------|--------|---------------|
| *T. sibiricus* MM 739 | TSIB_1788 | 78       | ND             | 81.1±0.06  | 63.5±0.20 | ND     | 0.4184 |
| *Tt. maritima* MSB8 | TM1419  | 80       | 75             | 81.0±0.16  | 64.5±0.17 | ND     | 0.4031 |
| Thermotoga sp. str. RQ2 | TRQ2_1313 | 80      | 80             | 85.2±0.07  | 63.6±0.14 | ND     | 0.4005 |
| ATM_T1       | 90      |          | 88.6±0.05      | 66.7±0.09  | 44.7±0.07 | 0.4084 |
| ATM_T2       | 83      |          | 88.8±0.02      | 67.5±0.09  | 47.9±0.07 | 0.4110 |
| ATM_T3       | 85      |          | 88.9±0.08      | 68.1±0.17  | 47.4±0.13 | 0.4110 |
| ATM_T4       | 83      |          | 88.8±0.01      | 68.3±0.39  | 49.2±0.13 | 0.4136 |
| ACM_C1       | 95      |          | >99            | 76.6±0.11  | 55.2±0.11 | 0.4188 |
| ACM_C2       | 95      |          | >99            | 76.5±0.04  | 55.3±0.03 | 0.4162 |
| *T. kodakarensis* KOD1 | TK2278  | 85       | 95             | >99        | 80.2±0.12 | 57.7±0.23 | 0.4215 |
| *P. furiosus* DSM 3638 | PF1616  | 100     | 99             | >99        | 81.6±0.08 | 69.1±0.15 | 0.4334 |
| AAM_A1        | 99      |          | >99            | >99        | 89.7±0.25 | 0.4491 |
| AAM_A2        | 99      |          | >99            | >99        | 90.6±0.17 | 0.4465 |

The $T_m$ values of MIPSs were determined using DSF. The $T_{opt}$ values were determined using a MIPS/malachite green assay. Standard deviations for MIPS $T_m$ values were
determined from at least four replicates. $T_{opt}$ values were determined from at least four replicates. ND, not determined.

Experimental Evidence Validates that Ancestral Thermotoga MIPSs are More Thermostable than Extant Thermotoga MIPSs

We showed in our computational analyses that the IVYWREL values of the MIPS proteins can be used to indicate thermostability, but high IVYWREL values can be related to factors other than thermostability [22,23]. To verify that the reconstructed MIPSs are more thermostable than the extant proteins, extant MIPS homologs from species of Thermotoga, Thermococcus (*Tc*), and Pyrococcus were tested in vitro for activity and thermostability, and their properties compared to those of the reconstructed proteins.

MIPS converts glucose-6-phosphate to myo-inositol-3-phosphate [4,6,7,17]. The crystal structure of *Tt. maritima* MIPSs has been determined (TM1419; Joint Center for Structural Genomics, PDB #3C1N, unpublished), but its catalytic activity has not been described. Only one thermophilic MIPS, from Archaeoglobus fulgidus, has had its activity characterized in vitro [24]. *A. fulgidus* MIPS is a class II aldolase [24]. Phylogenetically the Thermotoga MIPSs are more closely related to class I than class II enzymes. Class I enzymes use NH$_4^+$ as an allosteric activator while class II enzymes, like that from *A. fulgidus*, use metals.

The extant and reconstructed ancestral MIPSs have similar kinetic profiles (Table 2). All MIPSs tested used NH$_4^+$ as an allosteric activator, which is a characteristic of class I aldolases [25–27]. The temperature optima ($T_{opt}$) of the extant MIPSs were found to be near those of the OGTs of the source organisms (Table 1). The IVYWREL values of the extant proteins are correlated to the experimentally tested $T_{opt}$ ($R^2 = 0.91$; Figure S3B) of the MIPS proteins. This, in conjunction with the above computational correlations, provides further support for using the IVYWREL values as an indicator of thermostability. The MIPSs from the archaeal branches have the highest $T_{opt}$, which similar to the OGT trend (Table 1). As predicted, the four ancestral MIPSs from node T have higher $T_{opt}$ than those of *Tt. maritima* and Thermotoga sp. str. RQ2 (Table 1). The higher $T_{opt}$ of the ancestral Thermotoga MIPSs indicate that this protein may have originated from an organism that lived at a higher temperature.
The melting temperatures ($T_m$) of several MIPSs were determined using differential scanning fluorimetry (DSF) (Table 1). The $T_m$s were determined at different pH values, because some proteins unfold at pH 7.0 at a temperature higher than 99°C, which is the maximum temperature of the thermocycler. $T_m$s for such extremely thermostable proteins have been determined by DSF at low pH [28]. The $T_m$ (pH 3.6–4.2) was compared to the MIPS IVYWREL and $T_{opt}$ values of extant MIPS and they are all correlated to each other ($R^2 = 0.95$; Figure S3D–E). This shows that there is a correlation between $T_m$ and IVYWREL values, and $T_m$ and $T_{opt}$ values for the MIPSs dataset. Furthermore, a correlation was shown when the OGT was compared to $T_m$ (pH 3.6–4.2) ($R^2 = 0.64–0.88$; Figure S3C).

The extant archaeal MIPSs (except *Tc. sibiricus* MM 739) have the highest $T_m$ values, which correlate with the OGTs of the source organisms (Table 1). The MIPS from *Tc. sibiricus* MM 739, the tested archaeon with the lowest OGT, also had the lowest $T_m$ (Table 1). This suggests that in *Tc. sibiricus* MM 739 the composition of the MIPS protein has adapted to the lower temperature environment in which this microbe lives. An analogous adaptation to lower temperature may have happened to the MIPS that the *Thermotoga* genus acquired through HGT.

*Thermotoga* sp. str. RQ2 MIPS has a higher $T_m$ than that of *Tt. maritima* MIPS, which corresponds to their IVYWREL values. The $T_m$ values at pH 7.0 of the reconstructed ancestral MIPSs ATM_T1-4 are at least 3.4°C higher than that of *Thermotoga* sp. str. RQ2 MIPS, which is the most thermostable *Thermotoga* protein tested (Table 1). The average of the ATM_T1-4 values is 5.7°C at pH 7.0, which is higher than that of the average of the extant *Thermotoga* species’ proteins (Table 1). There is a significant difference between the $T_m$ values of the extant MIPSs and reconstructed ATM_T1-4 at pH 7.0 ($p<0.01$; Table S3A). The difference in $T_m$ values between these two groups is significant at

![Figure 2. Comparison of amino acid compositions between *Thermococcus* sequences and *Thermotoga* sequences.](#)

The y-axis is amino acid (AA) counts and the x-axis is the amino acid. Red bars are the mean of the extant *Thermotoga* sequences; white, yellow, green and orange bars are the *Thermotoga* ancestral reconstructed sequences ATM_T1–T4, respectively. Dark blue bars are the mean archaeal sequences, and pink and light blue bars are the *Thermococcus* ancestral reconstructed sequences ACM_C1–C2, respectively. A (*) marks where there is a significant difference ($p<0.01$) in the AA counts between the Archaea and the *Thermotoga*.

A (####) marks where there is a significant difference between the extant *Thermotoga* sequences and the reconstructed *Thermotoga* sequences. This shows there is no significant difference between the *Thermococcus* extant sequences and the *Thermococcus* ASR sequences, see Table S4 for p-values.

doi:10.1371/journal.pone.0084300.g002
pH ranging from 3.6 to 7.0 (Table S3A). The ASR of Thermotoga MIPSs at node T supports the hypothesis that the OGT of the last common ancestor of the Thermotoga genus grew at a higher OGT. Both the higher \( T_m \) and higher \( T_{opt} \) of the reconstructed Thermotoga MIPSs support this hypothesis (Table 1).

\( Tc. \) kodakarensis KOD1 and Pyrococcus furiosus DSM 3638 MIPSs are extremely thermostable. Using pH 4.2, the \( T_{opt} \) of \( Tc. \) kodakarensis and \( P. \) furiosus MIPSs were determined to be greater than those of the other extant MIPSs (Table 1). ACM_C1-2 IVYWREL values (0.4188 and 0.4162, respectively) are similar to the average of the extant Thermococcus MIPSs at node C (0.4162). \( Tc. \) kodakarensis MIPS was tested because it had the highest IVYWREL value of node C, 0.4215. \( Tc. \) kodakarensis KOD1 MIPS has a higher \( T_m \) than that of ACM_C1 and ACM_C2 when measured at both pH 3.5 and 4.2 (Table 1).

The reconstructed ancestral proteins from node A (AAM_A1-2) have higher IVYWREL values relative to the average values for all three nodes and the reconstructed ancestral proteins (Figure 1). Thus, IVYWREL values for the ancestral proteins at nodes T, C, and A were generally higher than the averages of the extant proteins at each node and the more deeply located nodes had higher IVYWREL values than the more recent nodes (A>C>T). All three nodes have statistically different IVYWREL values (Table S3). AAM_A1 and AAM_A2 were the most thermostable of all the MIPS proteins tested, including \( P. \) furiosus MIPS (Table 1, pH 3.5). \( P. \) furiosus grows at the highest OGT and contains the most thermostable extant MIPS tested. There is a significant difference between the \( T_m \) values of AAM_A1-A2 group compared to PF1616, and TK2278, and ACM_C1-2 group (Table S5B). These results suggest that the ancestral \( P. \) furiosus may have grown at a higher OGT than the extant species.

The data gathered here leads to a general trend, the higher the OGT of the organism the higher the MIPS protein IVYWREL, \( T_{opt} \), and \( T_m \) values. Several comparisons were made with these variables and they all support this correlation. These results are based upon a few data points and the pattern observed between OGT and IVYWREL, \( T_{opt} \) and \( T_m \) values has not been shown with other proteins or organism. Overall, these results indicate a correlation between MIPSs IVYWREL value, \( T_{opt} \) and \( T_m \), and organism’s OGT, and the correlation can be used to predict the OGT of an organism or an ancestral organism given IVYWREL values.

**Conclusion**

This study suggests that the Thermotoga and Thermococcus ancestors may have been capable of living in hotter environments than their descendants do now. This contention is based on the fact that the reconstructed proteins at each node are more thermostable and have higher optimal catalytic temperatures than their descendant proteins. Since the thermal characteristics of extant MIPS proteins show a correlation with the OGTs of their hosts, we surmise that the hosts of our reconstructed ancestral proteins had OGTs higher than their descendants. Additionally, we showed that the MIPS sequences have ameliorated to the Thermotoga genome based on reconstruction of proteins from the most recent common ancestor of Thermotoga, which did not have the signature of the donor (Archaea) and were more similar to the extant Thermotoga proteins.

**Methods**

**Reagents**

All reagents were reagent grade and purchased from either Sigma-Aldrich Co. or Fisher Scientific, Inc., unless otherwise stated.

**Statistical Analysis**

\( P \) values were determined either by \( t \)-test or \( z \)-test unless otherwise stated. An \( f \)-test was used to determine the appropriate \( f \)-test for a given data set. \( R^2 \) and Pearson correlation coefficient (\( P_{cc} \)) were determined from a linear regression line. Standard deviations were calculated from at least three replicates. 95% confidence intervals were calculated from replicates by finding the 2.5% and 97.5% quantiles.

**Phylogenetic Analyses**

Phylogenetic analyses were used to identify genes recently acquired from Archaea in the Thermotoga. The species of Thermotoga screened were \( Tt. \) maritima MSB6 [29], \( Tt. \) kodakarensis KOD1 [30], \( Tt. \) maritima RQ2 [29], \( Tt. \) neapolitana RKU-10 [31], and Thermotoga sp. str. RQ2 [29], Thermotoga sp. str. RQ2 [29], Thermotoga sp. str. RQ2 [29].
and *Tt. petrophila* RKU-1 [30]. In this work, the genus *Thermotoga* only refers to the above species and their close relatives as measured by 16S rRNA gene sequence comparisons. *Thermotoga* lettingae and *Thermotoga* thermarum were not considered as true members of the genus *Thermotoga* since their 16S rRNA gene sequence identities are too dissimilar from the true *Thermotoga* species. Genes unique to sequenced *Thermotoga* and *Thermosphaera* species were identified using bidirectional BLAST and Branch-Cut [32]. Protein sequences were aligned using default parameters in MUSCLE v3.8.31 [33]. After visual inspection of the alignments, maximum likelihood trees were made using PhyML 3.01 with Gamma+I WAG substitution model [34]. These ML trees were used to identify genes acquired from Archaea through HGT. One of the proteins identified with a clear phylogenetic history of transfer from the Archaea to the *Thermotoga* genus was myo-inositol-3-phosphate synthase (MIPS).

Bayesian analyses of protein sequences were also used to screen for Archaea-derived genes using MrBayes [35]. The best model for amino acid substitution was determined using ProtTest [36]. IVYWREL values were calculated using modified scripts initially developed by Olga Zhaxybayeva (Dartmouth College).

**Tests of Models of Ancestral Reconstruction**

The sequences of ancestor proteins at nodes of the MIPS phylogenetic tree were derived from analyses of extant MIPS protein sequences. The complete MIPS sequences were obtained from published genome sequences at NCBI. The partial *Thermotoga* MIPS sequences indicated below were obtained from published data [18]. *Tt. maritima* MSB8 and *Thermotoga* sp. str. ROQ2 MIPS peptide sequences are 382 amino acids in length. The MIPSs partial peptide sequences from *Thermotoga* sp. str. KOL6, *Thermotoga* sp. str. ROQ7, *Tt. neapolitana* LA10, *T. maritima* SL7, and *Tt. maritima* FjSS3B1 are missing 13 residues at their N-termini, and 37 residues at their C-termini. PCR amplified MIPS genes from other *Thermotoga* strains were also included in this analysis (Table S1). The full peptide sequences for these MIPSs were determined, except for that of *Thermotoga* sp. NTLA3, which is missing 67 residues at its N-terminus and 31 residues at its C-terminus. The accession numbers for these sequences are in Table S1. PhyML 3.01 with Gamma+I WAG substitution model [34] was used to construct a maximum likelihood tree using these sequences (Figure 1).

The best model for amino acid substitution was determined using ProtTest [36]. Ancescon [19] and MrBayes [37] were used to reconstruct the ancestral *Thermotoga* MIPS sequences *in silico*. Four different ASR methods were tested. First, a homogeneous substitution model was tested using Ancescon. Ancescon was run with the default parameters with optimize alpha (O) and reconstruct sequence for biological root and all internal nodes (R).

Second, a non-homogeneous substitution model was implemented to predict the ancestral proteins. ProtTest [36] was used to determine which of 112 homogeneous substitution models best fit the data. LG+G best described our dataset based on a Bayesian Information Criterion score. The Bio++ suite of programs was then used with the best model to implement the model non-homogeneously [20] by dividing the dataset a priori into two clades: and the *Thermotoga* species, defined as node T and its descendants (Figure 1), and the *Thermococcus* and *Pyrococcus* species, defined as all other nodes on the tree. Each clade was described by a separate set of equilibrium frequencies, while the parameters of the gamma distribution remained constant throughout the tree. Using this non-homogeneous implementation of LG+G model, the branch length and parameter values were optimized in BppML [20] using four rate categories and an initial alpha of 1. The root was placed in the *Pyrococcus* group basal to the gene transfer from *Thermococcus* to *Thermotoga*. Ancestral reconstruction was run in BppAncestor [20] using the optimized tree and parameters. The positions of gaps were inferred for reconstructed sequences using the method detailed in [30]. Briefly, to calculate the position of gaps in the ancestral sequences, we changed all of the gaps in existing sequences to Cs, and all of the amino acids to As. We then used the F84 substitution model in BppAncestor to determine ancestral ‘sequences,’ which represent the position of gaps as Cs. We used BppAncestor to calculate the position of gaps (Cs) in the output ancestral sequences and inserted the gaps into the actual reconstructed ancestral sequences using in-house PERL scripts (Figure S4). Alternative rootings of the tree within the *Methanococcus*, *Pyrococcus*, *Thermococcus* or *Thermotoga* did not affect ancestral sequence composition at the ancestral *Thermotoga* node.

Third, three ancestral nodes with strong bootstrap support from the PhyML tree were used for ASR of the node closest to the *Thermotoga* (node T), a *Thermococcus* node (node C) and ancestral to both node T and C and to other Archaea. MrBayes was used to construct a million trees, Burnin was used to remove the first 250,000 trees, and one tree for every 750 trees (total of 1,000 trees) was used to construct the ancestral protein for nodes A, C, and T using Ancescon. Ancescon was run with the default parameters with optimize alpha (O) and reconstruct sequence for biological root and all internal nodes (R). The probability of each residue was calculated from the 1,000 trees and averaged.

Fourth, it has been suggested that ancestral reconstruction using maximum likelihood may lead to a bias towards more thermostable protein for some datasets [21]. The Bayesian approach employed here samples from the posterior distribution to estimate the ancestral amino acid composition. Such a method is not prone to the tendency of ML methods to reconstruct ancestral proteins with a higher thermostability than they actually had [21]. This was implemented in BppAncestor, and was done using both the non-homogeneous model described above, and a homogeneous implementation of that same model. We created 1000 replicates for each ancestral sequence, determined by sampling from the posterior distribution. The mean and confidence interval of the IVYWREL biases for the ancestral sequences were calculated from these replicates (Table S2). The values within the 95% confidence interval are never more than a 5% deviation from the mean value, indicating that our dataset is not prone to the bias mentioned by [21].

**Reconstruction of Ancestral Sequences**

The ancestor reconstruction methods did not assign each residue unambiguously. At positions for which more than one amino acid could be used, we used the following methods to decide which amino acid to incorporate into our reconstructed proteins. The decision process is described in detail for determining the amino acid sequence of the node T proteins and this same process was used for the proteins at nodes A and C.

The ancestor protein sequences for node T derived from the phylogenetic analyses contained a small number of gaps. Gaps that occurred in the ancestral sequence caused by a residue from the deep branching *Pyrococcus* species (the N-terminus *Pyrococcus* species had up to three residues before the methionine in the *Thermotoga* sequences) and gaps that occurred due to a residue found in only one peptide sequence were removed from the ancestral sequence. No other gaps were present for the sequence for node T. Residues assigned a probability score of 0.9000 or above were considered strongly supported and were used in the construction of the ancestral proteins. For node T, 89 residues had values below 0.9000 and were examined by other criteria to decide which
amino acid should be at those positions. If two possible residues at a position had the same charge and polarity and their predictions added up to at least 0.9000 with at least one of them at or above 0.7000, the residue with the highest probability was chosen.

Of the remaining 43 residues, Ancecon predicted a probability of ≥0.7000 for 39 of them. Thirty-eight of these positions had a single amino acid in all or all but one of the extant Thermotoga MIPSs. That amino acid was assigned to those positions, leaving only five ambiguous residues at positions 119, 125, 126, 142, and 191.

All Thermotoga MIPS have Arg or Thr at position 119. When there is an Arg at 119 there is a Glu at 118, but when there is a Thr at 119 there is an Asp at 118 (Figure S5). The only Thermotoga to have an Arg at 119 and an Asp at 118 is Thermotoga sp. str. KOL6. Since the support for residue 118 being Glu is high (0.9606), Arg was placed at residue 119 in the ancestral protein.

Amino acid 125 was predicted to be either Ser or Thr, while the residue at 126 was predicted to be either Glu or Lys. In extant MIPS proteins, a Ser at 125 is associated with Lys at 126, but when there is a Thr at 125 there is a Glu at 126 (Figure S5). The only exception to this pattern is Thermotoga sp. str. KOL6 that has a Thr at position 125 and a Lys at 126. This pattern suggests a relationship between amino acids at positions 125 and 126 that is related to the structure or function of MIPS. Consequently, two variations of the ancestral MIPS were constructed with Ser/Lys or Thr/Glu at positions 125/126, respectively.

At position 142, Lys (polar, positive; higher support) and Ile (nonpolar, neutral; lower support) were predicted. The most likely residue at this position could not be resolved, so both possibilities were constructed.

The residue at position 191 (Figure S5) was predicted to be Ile, Phe, or Tyr. Ile and Phe are both nonpolar and neutral while Tyr is polar and neutral. When the Thermotoga MIPSs have Ile or Phe at residue 191, they have Asn (polar and neutral) at residue 177 (Figure S5). When the Thermotoga MIPSs have a Tyr (polar, neutral residue) at 191, all have Lys at 177 residue (polar, positive), except one has Ser (polar and neutral). The support for residue 177 being Lys is high (0.9986), so Tyr was placed at residue 191 in the ancestral protein.

Based on the above analyses, four probable ancestor proteins were predicted at node T. These reconstructed proteins were called ancestral Thermotoga MIPS, ATM, and labeled with their respective node and number, ATM_T1-4.

The same strategy used to predict the ancestor proteins at node T was done for those at nodes C and A. After using this analytical approach, only one residue remained questionable at node C and one at node A. At node C, position 365, Arg (polar, positive; higher support) and Gln (polar, neutral; lower support) were predicted. At node A, position 76, Glu (polar, positive) and Lys (polar, positive) were predicted. The most likely residue at these positions for nodes C and A could not be resolved, so both possibilities were constructed at each node. The reconstructed proteins for node C were called ancestral Thermococcus MIPS, ACM (ACM_C1-2). The reconstructed proteins for node A were called ancestral Archaea MIPS, AAM (AAM_A1-2).

Amino Acid Composition Comparison of Archaea and Bacteria MIPS

It has been predicted that the archaean and bacterial kingdoms use different amino acids for thermal adaptation of proteins; specifically, Gln, Ile, and positively charged amino acids [39]. In order to compare amino acid compositions between the Thermotoga and archaean sequences amino acid counts were preformed for each amino acid in each sequence. All Thermus and Methanococcus MIPS sequences were used to determine the average Archaea counts (Figure 1). The average amino acid compositions for the extant archaean sequences and the extant Thermotoga sequences were compared to determine if there were significant differences using a z-test (Table S4). The within group comparisons of the extant Archaea sequences and ASR ACM_C1–C2 sequences, and the extant Thermotoga sequences and ASR ATM_T1–4 sequences were tested for significant differences between the mean extant amino acid composition and the reconstructed composition using a t-test (Table S4).

Cloning and Sequencing of MIPS-encoding Genes

MIPS encoding genes were amplified using polymerase chain reaction (PCR) (Tables S1 and S6) using FailSafe enzyme mix (Epicentre), and were cloned into pGEM-T using Easy Vector System (Promega) or directly cloned into pBAD TOPO® (Invitrogen). The genes were excised from the pGEM-T vector using restriction enzyme sites incorporated during PCR, and cloned into pBAD TOPO® and transformed into Escherichia coli TOP10 cells (Invitrogen). The DNA for the ATM_T1-4, ACM_C1-C2, and AAM_A1-A2 genes (Table S1) were synthesized by GenScript USA Inc., and were cloned like the other genes. The sequence of each gene was determined at the University of Connecticut DNA Biotechnology Facility.

Expression and Purification of Ancestral and Extant MIPS Proteins

MIPS genes were expressed and purified from the pBAD TOPO® vector in E. coli TOP10 cells as described by the manufacturer (Invitrogen) with a few modifications. The cultures were grown in Bertani’s Lysogeny Broth (LB)/ampicillin (100 µg/ml) and induced on arabinose for 3–4 h at 37°C or overnight at 18°C.

The cultures were pelleted at 4,955 g for 30 min and washed with 500 mM NaCl, 0.5 mM dithiothreitol (DTT), and 20 mM imidazole-HCl, pH 7.8. Bacterial Protein Extraction Reagent (B-PER; Thermo Scientific) containing lysozyme and DNase I was used to resuspend the pellets as described by the manufacturer with a few modifications. The reaction mixture contained HalotM Protease Inhibitor Cocktail EDTA-Free (10 µl/ml B-PER; Thermo Scientific) and RNase A (Qiagen; 1 ng/ml of B-PER reagent). The reaction mixture was incubated for at least 15 min at room temperature, and heat-treated at 70°C for 10 min to denature non-thermophilic proteins. The cell extracts were pelleted and the supernatants filtered through a 0.2 µm filter (Thermo Scientific). Immobilized metal ion affinity chromatography was performed with a His SpinTrap (GE Healthcare) or Ni Sepharose High Performance Resin (GE Healthcare) using the protocol of the manufacturer with the following modifications: 500 mM NaCl, 0.5 mM DTT, and 20 mM imidazole-HCl, pH 7.8 was used as wash buffer, and 500 mM NaCl, 0.5 mM DTT, and 300 mM imidazole-HCl, pH 7.8 was used as elution buffer. The protein was concentrated with an Ultra-4 10K Centrifugal Filter (Amicon), washed 3–5 times and resuspended in 500 mM NaCl, 0.5 mM DTT, and 50 mM imidazole-HCl, pH 7.8, and stored at 4°C. Protein concentrations were determined using the Bradford reagent and bovine serum albumin as the standard following the manufacturer’s instructions (Thermo Scientific).

Differential Scanning Fluorimetry (DSF)

Differential scanning fluorimetry (DSF) was done as previously described with modifications [28]. The proteins were assayed in a
myo-inositol-3-phosphate Synthase (MIPS) Assay

The activities of MIPSs were determined using a MIPS/malachite green assay as previously described with modifications [25–27]. The reaction conditions were optimized for Thermotoga sp. strain RK2 MIPS and ATM_T1. These enzymes were shown to be in their initial velocities during the first 2.5 min. NH₄⁺ increased the activity of all MIPSs tested. A typical MIPS reaction contained 15 mM NH₄Cl, 20 mM D-glucose-6-phosphate, and 10 mM imidazole-HCl, pH 7.8 in a 50 mM volume. The reaction was preheated for 2.5 min in 0.2 ml Thermowell Gold Flat cap PCR tubes. The tubes were carried out in a heating block. To minimize evaporation, the temperature at half the maximal fluorescence and was determined using Gnuplot with curve fitting to the Boltzmann equation with an in-house script [40].

Supporting Information

Figure S1 Maximum likelihood (PhyML) trees of related MIPS protein sequences. The tree was constructed using PhyML. Support values were calculated from 100 bootstrap samples. Phylogenetically related microbes were grouped together and the major groups are shown: Thermotoga (red), Euryarchaeota (blue), Crenarchaeota (green), Thaumarchaeota and Korarchaeota (yellow), Aquifex (gray), and other Bacteria species (purple). Select bootstrap values with ≥70 bootstrap are indicated by black filled circles (•). Related groups were collapsed using FigTree (http://tree.bio.ed.ac.uk). Methanogenic Archaea are indicated with an asterisk. On the tree, two MIPS from methanogenic Archaea species are near the Thermotoga species, while 17 are located near the bottom of the tree. Homologs were gathered from the NCBI non-redundant database using T. petrophila RKU-1 peptide as query sequence. An E-value 1E-20 and Bit score of 100 were used as cut offs.

Figure S3 Comparison of extant organisms OGT and their MIPS proteins IVYWREL bias values (IVYWREL) were compared for several MIPS proteins. All organisms with known OGT from Figure S1 were used. In cases where a range of OGT was reported, the average of the range was used. The average IVYWREL bias values for each OGT was calculated and the mean standard error is shown. Linear regression yields a Pearson correlation coefficient (Pcc) of 0.80.

Table S1 Amplified MIPS-encoding genes. Accession numbers are shown for genes sequenced in this study. (DOC)

Table S2 A comparison of in silico prediction of protein thermostability, IVYWREL values, by posterior sam-
plunging prediction to that of Acescon/MrBayes prediction.

(DOC)

Table S3 Statistical analysis of IVYWREL values of extant and reconstructed MIPS proteins.

(DOC)

Table S4 Statistical analysis of the amino acid compositions of extant and reconstructed Thermotoga and Thermococcus MIPS proteins.

(DOC)

Table S5 Statistical analysis of $T_{\text{m}}$ (°C) values of extant and reconstructed MIPS proteins.

(DOC)

References

1. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, et al. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 399: 323–329.

2. Zhaxybayeva O, Swithers KS, Lapierre P, Fournier GP, Bickhart DM, et al. (2009) On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermogasales. Proc Natl Acad Sci U S A 106: 3065–3070.

3. Rodionov DA, Kurnasov OV, See C, Wang Y, Roberts MF, et al. (2007) Genomic identification and in vitro reconstruction of a complete biosynthetic pathway for the osmolyte di-nmyo-inositol-phosphate. Proc Natl Acad Sci U S A 104: 4279–4284.

4. Goncalves LG, Borges N, Serra F, Fernandes PL, Dospazo H, et al. (2011) Evolution of the biosynthesis of di-nmyo-inositol phosphate, a marker of adaptation to hot marine environments. Environ Microbiol.

5. Maddhison WP.(2011) Mesquite: a modular system for evolutionary analysis. Version 2.75 http://mesquiteproject.org.

6. Michell RH (2006) Inositol derivatives: evolution and functions. Nat Rev Mol Cell Biol 9: 151–161.

7. Michell RH (2011) Inositol and its derivatives: their evolution and functions. Adv Enzyme Regul 51: 84–90.

8. Sanyo J, Costa MS (2002) Comparable solutes of organisms that live in hot saline environments. Environ Microbiol 4: 501–509.

9. Gaucher EA, Govindarajan S, Ganesh OK (2008) Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451: 704–707.

10. Holms J, Shepherd C, Saj D, Demetras NJ, Hanning S, et al. (2012) On the origin and evolution of thermophily: reconstruction of functional precambrian enzymes from ancestors of bacteria. Mol Biol Evol 29: 625–635.

11. Harms MJ, Thornton JW (2010) Analyzing protein structure and function using ancestral gene reconstruction. Curr Opin Struct Biol 20: 360–366.

12. Chang BS, Jonsson K, Kazma MJ, Donoghue MK, Sakmar TP (2002) Recreating a functional ancient archaosaur visual pigment. Mol Biol Evol 19: 1483–1489.

13. Thornton JW, Need E, Crews D (2003) Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. Sci ence 301: 1714–1717.

14. Voordeckers K, Brown CA, Vanneste K, van der Zande E, Voet A, et al. (2012) Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. PLoS Biol 10: e1001446.

15. Akamana S, Nakajima Y, Yokobori S, Kimura M, Nemoto N, et al. (2013) Experimental evidence for the thermophilicity of ancestral life. Proc Natl Acad Sci U S A 110: 11067–11072.

16. Gaucher EA, Thomson JM, Burgan MF, Benner SA (2005) Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature 435: 285–288.

17. Majumder AL, Biswas BB (2006) Biology of inositols and phosphoinositides. Adv Enzyme Regul 46: 1–31.

18. Nesbo CL, L'Haridon S, Stetter KO, Doolittle WF (2001) Phylogenetic analyses of two "archaeal" genes in Thermotoga maritima reveal multiple transfers between archaea and bacteria. Mol Biol Evol 18: 362–375.

19. Cai W, Pei J, Grishin NV (2004) Reconstruction of ancestral protein sequence and its applications. BMC Evol Biol 4: 33.

20. Dhut J, Roos U (2008) Non-homogeneous models of sequence evolution in the Bc++ suite of libraries and programs. BMC Evol Biol 8: 255.

Table S6 Primers used for amplification of MIPS-encoding genes.

(DOC)

Acknowledgments

We thank Olga Zhaxybayeva of Dartmouth College for the scripts for IVYWREL calculation, Takaaki Sato from Kyoto University, Japan, for genomic DNA from *Tc. kodakarensis* KOD1, and Greg Fournier from Massachusetts Institute of Technology for his help with ASR.

Author Contributions

Conceived and designed the experiments: NCB PL AGG KSS JPG KMN. Performed the experiments: NCB PL AGG KSS. Analyzed the data: NCB PL AGG JPG KMN. Contributed reagents/materials/analysis tools: NCB JPG KMN. Wrote the paper: NCB KSS MN.


doc/2212-2221