

Calibrant Protein

The present invention concerns the provision of a molecular weight standard for various apparatus and techniques that determine, *inter alia*, molecular weight. In general, molecular weight standards contain a diverse range of unknown proteins, ranging in molecular weight from a few thousand to hundreds of thousands of Daltons. The present invention seeks to simplify the provision of molecular weight standards, by supplying a single protein species, that has the ability to form multiple distinct complexation states (i.e., dimers, trimers and/or tetramers), even at relatively low concentrations in solution. Such a simplified provision enables calibration of apparatus to be performed, reducing costs and provides a dedicated, appropriate and robust standard for newer techniques such as mass photometry.

The invention relates to ancestral citrate synthase proteins which are capable of forming at least 2, at least 3, at least 4 or at least 5 different complexation states simultaneously. The invention further relates to compositions comprising the ancestral citrate synthase proteins, use of the ancestral citrate synthase proteins or compositions, nucleic acids encoding the ancestral citrate synthase proteins, and methods of purifying the ancestral citrate synthase proteins.

Mass photometry is an example of an interferometric scattering microscopy technique (WO2018/011591) which can be used to determine the molecular mass of proteins, protein complexes and other biomolecules in solution in a single molecule approach; it is a novel technique that has only been recently commercially established. The measured signal is the interferometric contrast created from the light scattered by the biomolecule of interest and the light reflected by the measurement surface. The measured contrast directly correlates with the molecular mass (Young, G., et al.). Translation of the signal into the molecular weight of a biomolecule, however, requires a calibration step that has to be performed with samples of varying and known masses. These calibration measurements have to be performed regularly; usually at least once per day or measuring session and sometimes even more frequent since the interferometric contrast fluctuates depending on temperature, runtime of the laser and other changing parameters. The calibration samples, also called calibration standard, are therefore a necessary consumable to operate this technique.

The currently used calibration standard is not specifically dedicated or optimised for mass photometry, but reuses a commercially available protein mixture used for gel electrophoresis such as NativeMark™ Unstained Protein Standard (Invitrogen™ LC0725). This standard contains a mixture of multiple, non-disclosed proteins of different sizes at non-disclosed concentrations, and its use is fairly cost-intensive.

The present invention concerns a single protein that forms multiple distinct complexation states in solution. The resulting protein populations (populations of complexation states) are stable and have discrete molecular weight differences, which can be detected and distinguished by mass photometry. The invention enables reproducible and convenient mass calibration with a single protein amenable to large scale production in heterologous host bacteria. The protein itself is a citrate synthase but its amino acid sequence is not naturally found in an extant organism. It is an ancient representative of this enzyme family that is an amino acid sequence which has been inferred by ancestral sequence reconstruction – a method that uses related amino acid sequences to resurrect ancestral proteins.

Natural citrate synthases are known to form complexes, for example whilst eukaryotes (in their mitochondria) and Gram-positive bacteria use Type I citrate synthases which form dimers, Gram-negative bacteria use Type II citrate synthases which form hexamers. However, in extant citrate synthase proteins, the monomers form only their designated multimer, and not a variety of stable multimers. In a study by Schmidtman *et al*, *Arabidopsis* citrate synthase was analysed. When all mitochondrial proteins were studied together, Blue-Native PAGE noted the presence of a dimer (100 kDa) and a higher molecular weight complex of up to 1000 kDa. Given the presence of other proteins, the authors conclude these could be “multimeric aggregates”. The mitochondrial extract was then further analysed via gel-filtration chromatography, followed by Western blotting. Alternatively, recombinant CS4 was analysed. Whilst the concentration of the proteins was not confirmed, the amounts used were 100µg (recombinant CS4) and 180µg (mitochondrial extract) and the maximum loading volume for the column used in 500 µL, thus the concentrations far exceed the amounts used for mass standard purposes. Looking at the isolated protein, the authors concluded that the recombinant protein could form dimers (this is noted as the most abundant form) and oligomeric complexes.

In contrast, the citrate synthase proteins of the invention produce a series of defined and stable complexation states even at low protein concentrations, such as 10 nM. The individual complexation states each occur in adequate abundance and are sufficiently different in size to be detected and distinguished in mass photometry measurements. The difference in mass may be at least 25 kDa – the difference being measured between complexation states that differ by one monomeric unit (i.e., monomer to dimer).

The resulting mass calibrations reproducibly yield a mass error of about 1 %, which is at least as good as the commercial protein standard. The citrate synthase protein provided herein is also comparatively more stable for longer times at -20°C and after repeated freeze-thaw cycles. It has even shown to be sufficiently stable to be stored at 4 °C for several weeks, which is not the case for the

commercial protein mixture. The abundance of the different mass species and their mass differences are significantly better optimized for calibration of mass photometers than the commercial protein mixture. The protein can also be used as a molecular weight standard for other, non-denaturing techniques such as Native polyacrylamide gel electrophoresis (Native PAGE).

Additionally, the invention comprises a two-step purification protocol for the protein comprising heterologous production in *E. coli* which results in high yields, purity and batch-to-batch consistency.

Summary of the invention

The present invention provides a protein that can form multiple distinct complexation states simultaneously, and once formed, these complexation states are maintained, providing a heterologous mixture of complexes, each with a distinct molecular mass. Such is ideal for the calibration of instruments, apparatus, and techniques that determine mass of biomolecules and the like.

The present invention may be described in several ways:

The invention provided herein is a citrate synthase protein or functional fragment thereof, encoded by a nucleic acid sequence, wherein the nucleic acid is an ancestral gene and wherein the citrate synthase protein, or functional fragment thereof, is capable of forming at least 2, at least 3, at least 4 or at least 5 complexation states simultaneously.

The invention provided herein is a citrate synthase protein, or functional fragment thereof, having an amino acid sequence at least 75%, 80%, 85%, 90% or 95% identical to a sequence selected from SEQ ID NOs: 1-7 wherein the citrate synthase protein, or functional fragment thereof, is capable of forming at least 2, at least 3, at least 4 or at least 5 complexation states simultaneously.

The invention provided herein is a citrate synthase protein, or functional fragment thereof, having an amino acid sequence with at least 75%, 80%, 85%, 90% or 95% sequence similarity to a sequence selected from SEQ ID NOs: 1-7 wherein the citrate synthase protein, or functional fragment thereof, is capable of forming at least 2, at least 3, at least 4 or at least 5 complexation states simultaneously.

The invention provided herein is a citrate synthase protein, or functional fragment thereof, comprising a sequence selected from any of SEQ ID 1-7.

The invention provided herein is a citrate synthase protein comprising a sequence, or functional fragment thereof, comprising greater than 75%, 80%, 85%, 90% or 95% sequence identity to a sequence selected from SEQ ID No. 15-21 wherein the citrate synthase protein, or functional fragment

thereof, is capable of forming at least 2, at least 3, at least 4 or at least 5 complexation states simultaneously.

The invention provided herein is a citrate synthase protein, or functional fragment thereof, comprising a sequence comprising greater than 75%, 80%, 85%, 90% or 95% sequence similarity to a sequence selected from SEQ ID No. 15-21 wherein the citrate synthase protein, or functional fragment thereof, is capable of forming at least 2, at least 3, at least 4 or at least 5 complexation states simultaneously.

The invention provided herein, is a citrate synthase protein, or functional fragment thereof, comprising a sequence selected from any of SEQ ID 15-21.

The invention provided herein is a citrate synthase protein, or functional fragment thereof, comprising at least 1 ancestral mutation wherein the citrate synthase protein, or functional fragment thereof, is capable of forming at least 2, at least 3, at least 4 or at least 5 complexation states simultaneously.

The invention provided herein is a composition comprising a citrate synthase protein, or functional fragment thereof, according to any previous aspect wherein the concentration of the citrate synthase protein, or functional fragment thereof, is at least 5 nM, preferably the concentration is between 5 nM and 100 µM.

It may be preferred that the citrate synthase according to any definition of the invention is capable of forming at least 3 complexation states simultaneously. Each of these three complexation states are present in the solution in a sufficient amount to permit detection. Notably, each of the three complexation states may comprise at least 15% of the total citrate synthase protein complexes present.

It may be preferred that the citrate synthase according to any definition of the invention is capable of forming at least 4 complexation states simultaneously. Each of these four complexation states are present in the solution in a sufficient amount to permit detection. Notably, each of the four complexation states may comprise at least 5% of the total citrate synthase protein complexes present.

The most abundant complexation state may preferably comprise less than 50% of the total citrate synthase present. Optionally, the most abundant (most common, predominant) complexation states may preferably comprise less than 45%, less than 40% or less than 35% of the total citrate synthase present.

It is preferred that at least 3 distinct complexation states are formed in solution, and that these are sufficiently distinct to enable the use of the protein as a mass standard. This can be achieved with at least 25 kDa difference between each of the complexation states. Optionally, there is at least a 30 kDa,

35 kDa, 40 kDa, 45 kDa or 50 kDa difference between each of the complexation states. Thus, for example, a monomer could be 50 kDa, a dimer 100 kDa, a trimer 150 kDa and so on. The difference may be measured between complexation states that differ by one monomeric unit (i.e., monomer to dimer, tetramer to pentamer and so on).

The invention, provided herein is the use of a citrate synthase protein, or functional fragment thereof, or composition according to any previous definition of the invention as a molecular weight standard.

The invention provided herein is the use of a composition according to any previous definition of the invention in the calibration or standardisation of a biochemical technique.

The invention provided herein is a method of calibrating a mass photometry device comprising the step of detecting the mass of particles in the composition of any previous definition of the invention on a mass photometry device.

The invention provided herein is a nucleic acid sequence encoding a citrate synthase, or functional fragment thereof, of any of the previous definitions of the invention. The invention, provided herein is a vector comprising the nucleic acid of a previous definitions of the invention.

The invention provided herein is a cell transformed with a vector comprising a nucleic acid according to a previous definitions of the invention.

The invention provided herein is a method of producing a citrate synthase protein, or functional fragment thereof, according to a previous definition of the invention comprising culturing a cell transformed with a vector according to a previous definition of the invention and carrying out at least one purification step.

Description of figures

Figure 1 shows the overall profiles of Phylogenetic Trees 1 and 2 with general taxonomic descriptors for each branch.

Figures 2a-f show expanded views of Tree 1 used for ancestral sequence reconstruction. Amino acid sequences of 418 extant citrate synthase genes (from bacteria, archaea and eukaryotes) were collected from the NCBI Reference Sequence Database and aligned via MUSCLE v3.8.31 28. The maximum likelihood phylogeny was inferred from the multiple sequence alignment using raxML v8.2.10 29. Nodes corresponding to resurrected proteins are labelled CS1, CS5, CS6, and CS7. Labels a-j indicate how the branches of the phylogeny connect across the Figure.

Figures 3a-g expanded views of Tree 2 produced used for ancestral sequence reconstruction. Amino acid sequences of 418 extant citrate synthase genes (from bacteria, archaea and eukaryotes) were collected from the NCBI Reference Sequence Database and aligned via MUSCLE v3.8.31 28. The maximum likelihood phylogeny was inferred from the multiple sequence alignment using raxML v8.2.10 29. Nodes corresponding to resurrected proteins are labelled CS2, CS3, and CS4. Labels a-i indicate how the branches of the phylogeny connect across the Figure.

Figure 4 shows an image of an SDS-PAGE gel of CS1-His to CS7-His proteins.

Figure 5 shows an image of a Native PAGE gel of CS1-His, CS2-His and CS3-His proteins.

Figures 6a to d shows mass photometry histograms taken for samples of CS1-His from four different purifications showing batch-to batch reproducibility.

Figure 7 shows a comparison of proteins of known molecular weight against the mass obtained through mass photometry when using CS1-His as a calibrant.

Figures 8 a and b shows a comparison of a mass photometry histogram using NativeMark™ versus CS1-His HMW.

Detailed description of the invention

Citrate synthase is a ubiquitous enzyme involved in the citric acid cycle where it catalyses the condensation of acetate (from acetyl-CoA) and oxaloacetate to form citrate. All citrate synthases likely originated from the same common ancestral protein, however whilst eukaryotes (in their mitochondria) and Gram-positive bacteria use Type I citrate synthases which form dimers, Gram-negative bacteria use Type II citrate synthases which form hexamers. To better understand the factors leading to this divergence in structure, ancestral protein sequences were inferred using phylogenetic methods and these proteins were expressed and purified from *E. coli*. To great surprise, instead of forming only dimers or hexamers, each of the resulting “ancestral” citrate synthase proteins exhibited a range of complexation states simultaneously in solution. Different ancestral citrate proteins exhibit different distributions of complexation states, with most forming dimers, tetramers, hexamers, octamers and decamers, providing a set of detection targets spanning a range of molecular weights from about 85 kDa to about 430 kDa with regular molecular weight spacing between each state. The ancestral citrate synthase proteins are also highly stable, being able to survive multiple rounds of freezing and thawing and extended incubation under refrigeration with minimal change to their complexation properties. As a result of these characteristics, the ancestral citrate synthase proteins provide an ideal solution to the problem of mass photometry calibration.

According to one definition of the invention provided herein is citrate synthase protein, or functional fragment thereof, encoded by a nucleic acid sequence wherein the nucleic acid is an ancestral gene and wherein the citrate synthase protein, or functional fragment thereof, is capable of forming at least 2, at least 3, at least 4 or at least 5 complexation states simultaneously.

The invention may relate simply to the citrate synthase protein as defined herein.

The term "functional fragment" as used herein means a portion or part of the protein sequence which retains the capability of forming multiple complexation states. The functional fragment may be produced in any appropriate way but can involve the deletion of sequences not critical for interaction with other protein sequences. At least 3%, 5%, 7%, 10%, 12%, 15% or 20%, or more of the protein sequence may be removed in the functional fragment. For the present invention, for the avoidance of doubt, the term "functional fragment" does not refer to any enzymatic function in this instance.

As used herein "ancestral gene" refers to a common gene from which a family of genes is proposed to have descended. An ancestral gene may be derived from phylogenetic techniques such as ancestral gene restoration wherein an ancestral protein sequence is inferred (Harms & Thornton 2010, Harms & Thornton 2013, Selberg et al. 2021) as described above and a DNA molecule coding for that protein is synthesized. Thus the ancestral gene is typically an artificial DNA sequence inferred from an artificial amino acid sequence. The amino acid sequence may be inferred using ancestral sequence reconstruction as described below.

The citrate synthase protein originating from an ancestral gene may be referred to as "ancestral citrate synthase" for brevity.

The present invention may utilise ancestral sequence reconstruction as starting point for engineering citrate synthase proteins that could be used as mass calibration standards. Thus, the ancestral gene may be inferred by ancestral sequence reconstruction. Insofar as this invention relates, ancestral sequence reconstruction uses the immense and ever-expanding amount of sequence data available in sequence databases to create an alignment of present-day amino acid sequences of a citrate synthase protein. Phylogenetic and statistical analyses under appropriate models of evolution are then used to define amino acid sequences at the branch points or "nodes" of the phylogenetic tree generated in the ancestral sequence reconstruction. The amino acid sequences at the tree nodes are candidates for ancestral amino acid sequences of the citrate synthase, which have given rise to the amino acid sequences of the extant citrate synthase proteins. In brief, a phylogenetic tree, also known as a phylogeny, is a diagram that depicts the lines of evolutionary descent of different genes from a common ancestor.

The citrate synthase proteins resulting from production of the ancestral sequences reconstructed according to the invention are found to have unique properties not observed in extant citrate synthase proteins. Accordingly, these ancestral sequences benefit from an inherently different self-interaction specificities. The ancestral sequences have been found to form different multimeric forms simultaneously in a solution. Thus, the ancestral sequences may simultaneously form two or more, preferably three or more of dimers, trimers, tetramers, pentamers, hexamers, and even higher multimeric forms, including octamers and decamers.

This approach of ancestral sequence reconstruction has been applied to the citrate synthase protein of bacteria, archaea and eukaryotes. Specifically, as described in the Examples, extant citrate synthase proteins were used to construct Phylogenetic Trees 1 and 2 as shown in Figures 1 to 3. Briefly, the extant protein sequences have been aligned. Comparisons are made by aligning the sequences so that amino acid differences can be scored. This is not straightforward if the sequences are relatively dissimilar and/or have diverged by the accumulation of insertions and deletions as well as point mutations. Therefore, for this reason and for aligning multiple sequences, it is more common to rely upon a relevant computer program to perform the alignment. In the Examples, MUSCLE (Multiple Sequence Comparison by Log-Expectation as described in Edgar, R.C. *BMC Bioinformatics* 5, 113 (2004). doi.org/10.1186/1471-2105-5-113) was used for sequence alignment. Other available alignment software includes Clustal, T-Coffee, Bali-Phy or Phylo.

Once the alignment has been completed, any one of the various appropriate models for tree reconstruction can be utilised. There are several widely used methods for estimating phylogenetic trees (Neighbour Joining, Maximum Parsimony, Bayesian Inference, and Maximum Likelihood). In the Examples, RAxML (Randomized Axelerated Maximum Likelihood -as described in Stamatakis A., *Bioinformatics*) was used, this is a program for sequential and parallel Maximum Likelihood based inference of large phylogenetic trees. Other available phylogeny generating software includes PhyML (Guindon S. & Gascuel O. *Syst. Biol*). Other available software to infer phylogenetic trees includes PhyML, MrBayes, FastTree or IQ-Tree.

As used herein "complexation state" refers to a protein complex comprising a specific number of monomeric units of the same protein. When more than one monomeric unit is present, a complexation state may also be described as a macromolecular complex or multimer, these are generally formed of the monomers non-covalently bound to each other. In general, non-covalent interactions between the hydrophobic and hydrophilic regions on the monomer units help to stabilise the quaternary structure of the complexation state. Thus, in multimers, it is usual that non-covalent interactions between the monomeric units arrange the monomers into a particular structure. Usually,

the monomeric units are therefore assembled in a regular, predictable structure. . The number of monomeric units in a complexation state may be 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or greater than 10. Thus, the complexation state may be a monomer (1), dimer (2), trimer (3), tetramer (4), pentamer (5), hexamer (6) and larger multimers (usually 7+). The term may therefore be considered to comprise both monomers and homo-oligomer, wherein a monomer is a single protein with the capability of forming complexes and a homo-oligomer is a complex comprising a specific number of the same monomer.

The monomeric units are present in solution, and form various complexation states, which are discrete from each other and stable such that environmental conditions such as freeze/thaw does not impact the proportion of each multimer present. The solution may be any appropriate solution for a protein. The solution may contain any appropriate buffers and/or salts, for example. The concentration of the protein in solution may be 5 nM to 1000 nM (1 μ M).

In contrast, an agglomeration refers to the clustering of proteins that are not coherently structured and are based on random interactions.

The ancestral citrate synthase sequences identified by the present inventors has the surprising ability to form more than one complexation state when present in solution. Indeed, the inventors have observed that the ancestral citrate synthase proteins were able to form more than 3 different complexation states simultaneously in solution. Indeed, in Figures 6a to d, where the number of each citrate synthase protein complex was counted via mass photometry, at least 5 discretely identifiable complexation states were identifiable. Figure 6d, for example, has 7 identifiable complexation states. Further, the monomeric unit appears to be capable of forming multiple complexation states, without one complexation state being overly predominant (greater than 50% (by number) of the total citrate synthase protein complexes present). For example, in Figures 6a – 6d, the maximum amount of the most common complexation state in the total solution was 33% (172kDa) of total citrate synthase protein complexes measured. In contrast, the least favoured complexation state – present on Figures 6a and 6b made up 2% of the total. However, many complexation states are present in relatively good proportions – between 10 and 35% (by number) of the total citrate synthase protein complexes.

Mass photometry has the ability to detect the mass of single particles in solution without the need for labels. In general, it is usual to detect the mass of all particles present (for example on the coverslip) and count the number of particles assigned to each mass – for example see Figure 6a-6d where the “count” of each particle is given. Thus, the total number of particles present is known, as well as the number of counts for each species of particle – in this case each protein complex. It will be understood that the monomer is included in the term “protein complex”, as it is one possible species formed by

the citrate synthase. Thus, to determine the percentage of each citrate synthase protein complex, the number of that complex present is determined, together with the number of total. Thus, the percentages listed here are based on number.

Thus, it may be preferred that the ancestral citrate synthase can form at least 3 complexation states simultaneously, and that each of those three complexation states provides at least 10%, 15%, 20%, 21%, 22%, 23%, 24% or 25% (by number) of the total citrate synthase protein complexes present in the solution. Alternatively put, the 3 major (or most common) complexation states are each present between 15 and 40% (by number), optionally 18% to 35% (by number) of the total citrate synthase protein complexes in solution.

As used herein “simultaneously” means at the same time. Thus, in the solution of the citrate synthase of the invention, there are present different complexations states which form at the same time – for example, monomers, dimers, trimers and tetramers. Once formed, these complexation states are stable and are discrete, i.e., there is little, if any, dissociation and reassociation.

As with most standardisation techniques it is theoretically possible to provide measurements of just two known points which define a gradient, in the case of mass photometry this is a gradient of molecular weight against radiometric contrast. However, in practice, at least three known points are generally required as this allows for confirmation of the gradient and the definition of the end points of the detection range for the instrument being used. Thus, it is preferred that at least 3 discrete complexation states are formed by the citrate synthase of the present invention. This permits use of the citrate synthase as a molecular mass marker for mass photometry. As discussed above, at least 3 complexation states of the citrate synthase of the invention are individually present at least 10%, 15%, 20%, 21%, 22%, 23%, 24% or 25% (by number) of the total citrate synthase protein complexes present in the solution. Thus, each of the three most common/most abundant complexation states at present at a sufficient level to permit accurate detection and thus calibration of the apparatus.

The citrate synthase protein, or functional fragment thereof, of the present invention may preferably have an amino acid sequence at least 75%, 80%, 85%, 90% or 95% identical to a sequence selected from SEQ ID NOs: 1-7 wherein the citrate synthase protein, or functional fragment, thereof is also capable of forming at least 3 complexation states simultaneously.

The citrate synthase protein, or functional fragment thereof of the present invention may preferably have an amino acid sequence with at least 75%, 80%, 85%, 90% or 95% sequence similarity to a sequence selected from SEQ ID NOs: 1-7 wherein the citrate synthase protein, or functional fragment, thereof is also capable of forming at least 3 complexation states simultaneously.

Sequence identity refers to the comparison of aligned sequence. If both sequences contain the same amino acid in the same position after alignment then the sequences are considered identical at this position. Tools to calculate sequence identity are readily available and widely utilised in many life disciplines of life sciences. For the purposes of this application sequence identity refers to the identity output from the publicly available Emboss Needle program https://www.ebi.ac.uk/Tools/psa/emboss_needle/ (Needleman and Wunsch (<https://www.sciencedirect.com/science/article/abs/pii/0022283670900574?via%3Dihub>)).

Sequence similarity is related to sequence identity, however it makes allowance for the similar properties of some amino acids. For instance both glutamate and aspartate have similar chemical, structural and electrostatic properties. Therefore, if one sequence contains glutamate in a position where the other contains aspartate after alignment the sequences are considered similar at this position, but not identical. Similar relationships between other amino acids are taken into account. For the purposes of this application sequence similarity refers to the similarity output from the publicly available Emboss Needle program https://www.ebi.ac.uk/Tools/psa/emboss_needle/ (Needleman and Wunsch (<https://www.sciencedirect.com/science/article/abs/pii/0022283670900574?via%3Dihub>)).

It may be preferred that the present invention relates to a citrate synthase protein, or functional fragment thereof, comprising a sequence selected from any of SEQ ID 1-7. Further, the citrate synthase protein may comprise or consist of the sequence of SEQ ID No. 1.

The citrate synthase protein, or functional fragment thereof, of the present invention may comprise at least 1 ancestral mutation. The citrate synthase protein, or functional fragment thereof, of the present invention may comprise at least 2, 3, 4, 5, 6, 7, 8, 9, or 10 ancestral mutations.

As used herein “ancestral mutation” refers to an alteration of a first amino acid sequence of a protein such as citrate synthase, wherein the alteration to the first amino acid sequence increases the similarity and/or identity of the first amino acid sequence to a second amino acid sequence of the same protein wherein the second amino acid sequence is an ancestral sequence which has been generated by phylogenetic methods such as ancestral sequence reconstruction. In such techniques an ancestral sequence is inferred by comparison of the known sequences of extant organisms together with the knowledge of their evolutionary history as described above.

By way of example only, the sequence for *E. coli* citrate synthase (Accession number WP_166726827) is given below (SEQ ID NO. 29) as aligned to SEQ ID No. 1 which is an ancestral sequence inferred by comparison of the known sequences of extant organisms (including *E. coli* though this is not a

SEQ ID NO 29 1 MADTKAKLTNGNTAVELDVLKGTGLGQDVIDIRTLGSKGVFTFDPGFTST 50
|.....

51 ASCESKITFIDGDEGILLHRGFPIDOLATDSNYLEVICYILLNGEKPTOEO 100

[illegible]

101 YDEFKTTVTRHTMIHEOITRLFHAFRRDSHPMAVMCGITGALAAFYHDSL 150

$\cdot$ | | | | $\cdot \cdot$: $\cdot \cdot$: $\cdot \cdot$ | $\cdot$: | $\cdot \cdot$ | $\cdot \cdot \cdot \cdot$: : | | | | $\cdot$: $\cdot \cdot \cdot \cdot$ | | $\cdot \cdot$ | $\cdot \cdot$:

60 LDEFKEELAENRAIPEEIIDLLRTLPKSAHPMAALRTAVSALGMFDPDAD 109

151 DVNNPRHREIAAFRLLSKMPTMAAMCYKYSIGOPFVYPRNDLSYAGNFLN 200

| | : | | . | : : : | : | . | . . : : . . . | | . . | . | | . | | : | . | | .

110 DVOSPEANYRKAIRLIAKIPTIVAAFHRIROGOEPVAPRPDLSHAANFLY 159

201 MMFSTPCEPYEVNPILERAMDRILILHADHEONASTSTVRTAGSSSGANPF 250

| : . . . | : : | : . . : : | | . . | | | | : | | . | | | | . . . | . . . | : : : : :

160 ML-----NGEESPVOAKVMDVALILHAEHEMNASTFAARVVASTLSDMY 204

251 ACIAAGIASLWGPAHGGANEAAALKMLEEISSV-KHIPEFVRRRAKDKNDSF 299

[illegible]

205 SAITAAIGALKGPLHGGANEQVMKMLQEIGSPDKAEPWVQEKLANKR--- 251

300 RLMGFGHRVYKNYDPRATVMRETCHEVLKELG-TKDDLLEVAMELENIAL 348

| : | | | | | | | | | . | | | | | | . : : : . . . : : . : : . | | | | . : | . : | . : . :

252 RIMGFGHRVYKTYDPRAKILKKMARQLAEKHGDTK--LYEIAEAVEKVVV 299

349 N--DPYFIEKKLYPNVDFYSGIILKAMGIPSSMFTVIFAMARTVGWIAHW 396

[illegible]

300 ERLGP----KGIYPNVDFYSGLVYHALGIPTDLFTPIFAMARVSGWTAHV 345

397 SEMHSDGMKIARPROLYTGYEKRFKSDIKR 427

. | . . . | . : : . | | | . : | . | . . . | . : . . . : |
 346 LEQLEDN-RLIRPRAVYVGPTDRKYVPIDQR 375

Examples of ancestral mutations are:

T50V mutation in SEQ ID No. 29 which increases the identity with SEQ ID No. 1;
 W396I mutation in SEQ ID No. 29 which increases the similarity with SEQ ID No. 1;
 removal of any/all of residues 297-299 from SEQ ID No. 29 to increase the overlap with SEQ ID No. 1;
 addition of suitable residues between N349 and D350 of SEQ ID No. 29 to increase the overlap, and
 optionally the identity or similarity, with SEQ ID No. 1.

The citrate synthase protein, or functional fragment thereof of the present invention may preferably have at least 2, at least 3, at least 4, at least 5, at least 10, at least 15 or at least 20 ancestral mutations. The skilled person can readily determine by trial and error the type, position and number of ancestral mutations required to achieve a citrate synthase protein which is capable of forming at least 2, at least 3, at least 4 or at least 5 complexation states simultaneously.

Any of the above referenced citrate synthase proteins, or functional fragments thereof, may be joined to another amino acid sequence and thus may form part of a fusion protein. For example, any of the citrate synthase proteins provided herein may further comprise a purification tag which may be a His-tag, Strep-tag or any other suitable affinity tag. The purification tag may be positioned at the N terminus or the C terminus of the citrate synthase protein. Such purification tags simplify the purification of the target protein. There may be a spacer between the citrate synthase protein and the purification tag. The spacer may be 1 to 5 amino acids. The spacer may increase the flexibility of the purification tag in solution and thus increase its availability for binding its affinity partner. The spacer may comprise a target for cleavage such that the purification tag can be removed in a post-processing step.

Additionally, or alternatively, provided herein is a citrate synthase protein, or functional fragment thereof comprising a sequence comprising greater than 75%, 80%, 85%, 90% or 95% sequence identity to a sequence selected from SEQ ID No. 15-21 wherein the citrate synthase protein, or functional fragment thereof is capable of forming at least 2, at least 3, at least 4 or at least 5 complexation states simultaneously.

Alternatively, provided herein is a citrate synthase protein, or functional fragment thereof, comprising a sequence comprising greater than 75%, 80%, 85%, 90% or 95% sequence similarity to a sequence selected from SEQ ID No. 15-21 wherein the citrate synthase protein, or functional fragment thereof, is capable of forming at least 2, at least 3, at least 4 or at least 5 complexation states simultaneously.

Alternatively, provided herein is a citrate synthase protein, or functional fragment thereof, comprising a sequence selected from any of SEQ ID 15-21. In an embodiment the citrate synthase protein, or functional fragment thereof, comprises the sequence SEQ ID No. 15.

According to any description of the invention, the citrate synthase protein, or functional fragment thereof, is not expressed by an extant organism. "Extant organisms" means organisms living today and includes any known living organism but does not include organisms that have been genetically modified/transformed/edited for the purpose of expressing the citrate synthase proteins, or functional fragments thereof, described herein. The citrate synthase proteins described herein are inferred ancestral proteins and are not believed to be produced by extant organisms. Thus, the citrate synthase proteins provided herein have an artificially constructed amino acid sequence, they are not known to currently exist in nature.

In any embodiment provided herein the citrate synthase protein, or functional fragment thereof, may be an artificial citrate synthase protein, or functional fragment thereof. In this context, artificial means produced from a non-natural gene sequence rather than from a gene sequence found in an extant organism.

The citrate synthase proteins, or functional fragments thereof, of the present invention can be readily screened for their capacity to form multiple complexation states using mass photometry or other native mass analysis techniques.

Additionally, provided herein is a composition comprising a citrate synthase protein, or functional fragment thereof, according to description of the invention wherein the concentration of the citrate synthase protein, or functional fragment thereof is at least 5 nM, preferably the concentration is between 5 nM and 100 μ M. Thus, the concentration of the protein in solution may be 5 nM to 1000 nM (1 μ M), or 1 μ M to 100 μ M, and any value between these ranges. The concentration of the proteins in the solution may be varied according to the parameters of the apparatus or technique it is used to calibrate.

The composition may further comprise:

buffering reagents, suitable to maintain a pH between 5 and 10 during routine use; and/or

any appropriate salts which may optionally be selected from a list including sodium chloride and potassium chloride;

wherein the total salt concentration is at least 50 mM, preferably the salt concentration is in the range 50 to 500 mM.

Further provided herein according to the present invention is the use of a citrate synthase protein, or functional fragment thereof, or composition according to any previous description of the invention as a molecular weight standard. Several biochemical techniques rely on the use of standard proteins with known molecular weights. The citrate synthase proteins of the current invention have potential use in any non-denaturing technique including but not limited to mass photometry, native PAGE and gel filtration.

Mass photometers require regular calibration; usually at least once per day or measuring session and sometimes even more frequent since the interferometric contrast fluctuates depending on temperature, runtime of the laser and other changing parameters.

The complexation states of the citrate synthase protein, or functional fragment thereof, are dependent on the structure of the protein which determines their protein-protein interactions. The citrate synthase protein is therefore not suitable as a molecular weight standard in denaturing techniques which reduce proteins to their primary structure such as sodium dodecyl sulphate PAGE (SDS-PAGE). If the complexation states of the protein were fixed by, for example, chemical cross linking then they would be resistant to disassociation and may have wider applications for more techniques.

In another aspect, provided herein is the use of a composition according to any description of the invention in the calibration or standardisation of apparatus and/or a biochemical technique. Preferably, the composition may be used in the calibration of a mass photometry device. Optionally, the composition may be used as a molecular weight standard in Native PAGE.

In another aspect, further provided herein is a method of calibrating a mass photometry device comprising detecting the mass of particles in the composition as described herein.

For calibration purposes, the detected contrast values are converted to molecular masses using the instrument's software, which is calibrated using molecular weight standards (as described herein). Calibration should be regularly validated by measuring the composition of the present invention and verifying the molecular masses the protein complexes, which are known. Thus, the molecular masses of the calibrant/molecular weight standard are known, and the instrument's software can be adjusted based on the detected masses of the calibrant/molecular weight standard. As good laboratory practice, it is recommended to include the calibration measurement with every series of experiments.

Additionally, provided herein is a nucleic acid sequence encoding a citrate synthase of any of the aspects above.

The citrate synthase protein may be encoded by an ancestral gene.

Optionally, the nucleic acid is codon optimised for production of the protein in an organism, preferably the sequence has been codon optimised for expression in *E. coli*.

Whilst the genetic code has some redundancy, meaning different codons encode for the same amino acid, organisms typically do not produce tRNA for each codon at the same levels and may have preferred codons for which they produce higher amounts of tRNA. Codon optimization refers to the redundant alteration of the nucleic acid encoding a protein to take advantage of the known tRNA distribution of an organism and thereby enhance expression of the protein.

Additionally, provided herein is a vector comprising the nucleic acid as described here.

A vector as used herein refers to any particle used as a vehicle to artificially carry a foreign nucleic sequence into a cell, where it can be replicated and/or expressed. Vectors include but are not limited to plasmids, cosmids, viruses and phages.

Optionally, the vector is a plasmid, preferably the vector is a pET plasmid.

Further, provided herein is a cell transformed with a vector comprising a nucleic acid as described herein. Preferably, the cell is an *E. coli* cell. Optionally, the transformed *E. coli* cell expresses a citrate synthase as described herein.

Further, the invention comprises a method of producing a citrate synthase protein as described herein comprising culturing a cell as described herein and carrying out at least one purification step.

Optionally, the purification step comprises affinity chromatography. Alternatively, the purification step comprises size exclusion chromatography (SEC).

Whilst the aspects above use the term "citrate synthase protein" to describe the proteins of the invention there is no requirement that the protein have any enzymatic activity in order to meet this definition. The protein merely needs to comprise a sequence originally deriving from, or be similar to, a citrate synthase protein as provided herein; and may also have been modified in any of the above indicated ways. The proteins provided herein may be isolated or purified.

It should be noted that ancestral proteins and ancestral genes are a product of algorithmic analysis which determines the probability of each amino acid being in any position within the protein sequence for a particular node of the phylogenetic tree. The ancestral sequences disclosed herein are high probability solutions to the ancestral sequence reconstruction algorithm. We will likely never know for certain the actual sequence of the ancestor proteins or the genes encoding them since too much time has passed for biological samples to survive intact. This in no way diminishes the utility of the invention.

Examples:**Example 1 – Ancestral protein/gene reconstruction of citrate synthase proteins:**

Two sets of amino acid sequences of 418 extant citrate synthase genes from bacteria, archaea and eukaryotes (Table 2) were collected from the NCBI Reference Sequence Database and each set was aligned via MUSCLE v3.8.31. Maximum likelihood phylogenies were inferred from the multiple sequence alignment using raxML v8.2.10. The LG substitution matrix was used as determined by automatic best-fit model selection as well as fixed base frequencies and a gamma-model of rate heterogeneity. The two trees are largely similar, differing mainly in the location of peroxisomal citrate synthases. The location of this family in Tree 2 better reflects the established evolutionary history.

Based on the citrate synthase trees and the multiple sequence alignments, ancestral sequences (CS1-CS7) were inferred using the codeML package within PAML v4.9. To adjust for gaps and the different length of the N-termini of the citrate synthase sequences, their ancestral state was determined using parsimony inference in PAUP 4.0a based on a binary version of the multiple sequence alignment (1=amino acid, 0=gap, no residue). The state assignment for each node in the tree (amino acid or gap) was then applied to the inferred ancestral sequences.

CS1 Seq ID No. 1

MVAKGLEGVVAAESSISYIDGQEGRLYYRGYPIDELAEHSSFEVAYLLWHGRLPTREELDEFKEELAENRAIPEDIDL
LRTLPSAHPMAALRTAVSALGMFDPDADDVQSPEANYRKAIRLIAKIPTIVAAFHRRIRGQEPVAPRPDLSHAANF
LYMLNGEESPVPQAKVMDVALILHAEHEMNASTFAARVVASTLSDMYSITAAGALKGPLHGGANEQVMKML
QEIGSPDKAEPWVQEKLANRRIMGFGHRVYKTYDPRAKILKKMARQLAEKHGDTKLYEIAEAVEKVVVERLGPKG
IYPNVDFYSGLVYHALGIPTDLFTPIFAMARVSGWTAHVLEQLEDNRLIRPRAVYVGPTDRKYVPIDQR

CS2 Seq ID No. 2

MATMEIKKLEGVVVAETKISYIDGQEGRLYYRGYPIDELAEHSTFEVAYLLYGRLPTRDELDEFKEELAENRALPE
QIIDLLKNLPKDAHPMAALRTAVSALGMFDPDADDTSPARYRKAIRLIAKIPTIVAAFHRRIRGQDPVAPRPDLSH
AANFLYMLNGEESPVEAKVFDVALILHADHEMNASTFAALVVASTLSDMYSITAAGALKGPLHGGANEVVMK
MLQEIGSPDKAEPWVQEKLANRIMGFGHRVYKTYDPRARILKKYAKQLAEKHGDSKLYEIAEAVEKVVVERLGP
KGIYPNVDFYSGIVYYSMGIPTDLFTPIFAMARIAGWTAHILEYLDNRLIRPRAVYVGPTDRKYVPIDQR

CS3 Seq ID No. 3

MSTTEIAKLEGVVFETKLSFIDGQEGRLYYLGYPIDELAEHSTFEVVSFLLHGRLPTRDELEAFKEELAAANRALPEEL
IDALRAYPKDAHPMSALRTAVSELGMFDPDAEDTSPEGRYQKSVRLIAKFATIVAAIKRIREGQDPVAPRPDLSHAA
NFLYMLNGEESPQAKLFDVALILHADHGMNASTFTALAVASTLSDMYSSITAAGALKGPLHGGANEAVMKML
QEIGSPDKAEAWVQEKLANRIMGMGHRVYKAFDPRARILKKYAEQVAEKHGKSKYYEILETVEKEVVKRLGPKG
IYPNVDFYSGVVYSDLGIPTEFFTPFAVARISGWTAHILEYTRDNRLLRPKAVYVGELDRKYVPIDQR

CS4 Seq ID No. 4

MASMEYTPGLAGVVAAESSISYIDGQEGILRYRGYPIDELAEHSTFEVAYLLFGELPTRDELEEDHELKHHRALPE
RIIDLLKNLPKSAHPMAALQSAVAALGMFYDPADVTDPENYEAARVLIAPLPTIVAAFHRRIRGQDPPIAPRDDLGH
AANFLYMLNGEEDPLAARVFDVCLILHAEHSMNASTFTARVVGSTLADPYSIAAAAIGSLSGPLHGGANEVLMQ

LEEIGSPDNVEPWLEEKLARKEKIMGFGHRVYKVKDPRATILQKMAEQLFEKHGSTPLYDIALELEKVAERLGPCKGI
YPNVDFYSGIVYQKMGIPTDLFTPIFAIARVAGWTAHWLEQLEDNRIFRPSQIYVGPTDRSYVPIDER

CS5 Seq ID No. 5

MYTPGLAGVVAEAESSISYIDGQEGILRYRGYPIEELAEHSSFEVAYLLLYGRLPTREELEEFDELKHHRAIPEGIIDL
KTLPKSAHPMAALQSABAALGMFYPEADDVEDPEANYKKAVRLIAKLPTIVAFAHRIRRGKEPVAPRSDLSHAANF
LYMLNGEEDPLAARVMDVCLILHAEHEMNASTFTARVVGSTLADPYSIAAAAIGALSGPLHGGANEVLMKMLQEI
GSADNVEPWLEEKLATKQKIMGFGHRVYKTYDPRAKILKKMAEQLFEKHGSTPLYDIAVELEKVAERLGPCKGIYPN
VDFYSGLVYQALGIPTDLFTPIFAIARVAGWTAHWLEQLEDNRIFRPRQVYVGPTDRKYVPIDQR

CS6 Seq ID No. 6

MIAGLEGVVAEAESSISYIDGQEGRLYYRGYPIEELAEHSSFEVAYLLWHGRLPTREELEEFKEELAKNRAIPEEIDL
RTLPKSAHPMAALRTAVSALGMFDPDADDTQSPEARYKAIKLIKIPTIVAFAHRIRRGQEPVAPRPDLSHAANFL
YMLNGEEDPSVQAKVMDVALILHAEHEMNASTFAALVVASTLSDMYSITAAGALKGPLHGGANEQVMKMLQ
EIGSPDKAEPWVQEKLANRRIMGFGRVYKTYDPRAKILKKYARQLAEKQGDITLYEIAEAVEKVVVERLGPCKGIY
PNVDFYSGLVYHALGIPTDLFTPIFAMARVSGWTAHVLEYLKDNRILIRPRAVYVGPTDRKYVPIDQR

CS7 Seq ID No. 7

MYTPGLAGVVAEAESSISYIDGQEGILRYRGYPIEELAEHSSFEVAYLLFGKLPTRREELEEFDELKHHRAIPEGIIDL
KTLPKSAHPMAALQSABAALGMFYPADDDVEDPEANYEAAVRLIAKLPTIVAFAHRIRRGKDPIAPRSDLGHAANFL
YMLNGEEDPLAARVMDVCLILHAEHSMNASTFTARVVGSTLADPYSIAAAAIGSLSGPLHGGANEVLMKMLQEI
GSAENVEPWLEEKLATKQKIMGFGHRVYKVKDPRATILQKMAEQLFEKHGSTPLYDIAVELEKVAERLGPCKGIYP
NVDFYSGLVYQKLGIPDLFTPIFAIARVAGWTAHWLEQLEDNRIFRPTQIYVGPTDRSYVPIDQR

CS1-His Seq ID No. 15

MVAKLEGVVAEAESSISYIDGQEGRLYYRGYPIEELAEHSSFEVAYLLWHGRLPTREELDEFKEELAENRAIPEEIDL
LRTLPSAHPMAALRTAVSALGMFDPDADDVQSPEANYRKAIKLIKIPTIVAFAHRIRRGQEPVAPRPDLSHAANF
LYMLNGEEDPSVQAKVMDVALILHAEHEMNASTFAARVVASTLSDMYSITAAGALKGPLHGGANEQVMKML
QEIGSPDKAEPWVQEKLANRRIMGFGRVYKTYDPRAKILKKMARQLAEKHGDTKLYEIAEAVEKVVVERLGPCKG
IYPNVDFYSGLVYHALGIPTDLFTPIFAMARVSGWTAHVLEQLEDNRILIRPRAVYVGPTDRKYVPIDQRLEHHHHH
H

CS2-His Seq ID No. 16

MATMEIKKLEGVVAETKISYIDGQEGRLYYRGYPIQELAEHSTFEVAYLLLYGRLPTRDELDEFKEELAEHRALPE
QIIDLLKNLPKDAHPMAALRTAVSALGMFDPDADDTQSPEARYKAIKLIKIPTIVAFAHRIRRGQDPVAPRPDLSH
AANFLYMLNGEEDSPVEAKVFDVALILHADHEMNASTFAALVVASTLSDMYSITAAGALKGPLHGGANEVMMK
MLQEIGSPDKAEPWVQEKLANRIMGFGRVYKTYDPRARILKKYAKQLAEKHGDSKLYEIAEAVEKVVVERLGP
CKGIYPNVDFYSGIVYYSMGIPTDLFTPIFAMARIAGWTAHILEYLDNRILIRPRAVYVGPTDRKYVPIDQRLEHHHHH
H

CS3-His Seq ID No. 17

MSTTEIAKLEGVVFETETKLSFIDGQEGRLYYLGYPIQELAEHSTFEVVSFLLHGRLPTRREELEAFKEELANRALPEEL
IDALRAYPKDAHPMSALRTAVSELGMFDPDAEDTSPEGRYQKSVRLIAKFATIVAIAKRIREGQDPVAPRPDLSHAA
NFLYMLNGEEDSPEAKLFDVALILHADHGMNASTFTALAVASTLSDMYSSITAAGALKGPLHGGANEAVMMKML
QEIGSPDKAEAWVQEKLANRIMGMGHRVYKAFDPRARILKKYAEQVAEKHGKSKYIELETVEKEVVKRLGPCKG
IYPNVDFYSGVVYSDLGIPTEFFTPIFAVARISGWTAHILEYTRDNRLLRPKAVYVGELDRKYVPIDQRLEHHHHHH

CS4-His Seq ID No. 18

MASMEYTPGLAGVVAEASSISYIDGQEGILRYRGYPIEELAEHSTFEVAYLLFGELPTRDELEEFDELKHHRALPE
RIIDLKLNLPKSAHPMAALQSAVAALGMFYDPADVTDPEGNYEAAVRLIAKLPTIVAFAHRIRRGQDPIAPRDDLGH
AANFLYMLNGEEDPLAARVFDVCLILHAEHSMNASTFTARVVGSTLADPYSIAAAAIGSLSGPLHGGANEVLQM
LEEIGSPDNVEPWLEEKLARKEKIMGFGHRVYKVKDPRATILQKMAEQLFEKHGSTPLYDIALELEKVAERLGPKEI
YPNVDFYSIGVYQKMGIPDFTLFTPIFAIARVAGWTAHWLEQLEDNRIFRPSQIYVGPTDRSYVPIDERLEHHHHHH

CS5-His Seq ID No. 19

MYTPGLAGVVAEASSISYIDGQEGILRYRGYPIEELAEHSSFEVAYLLYGRLPTRREEEFDELKHHRAIPEGIIDL
KTLPKSAHPMAALQSAVAALGMFYPEADDVEDPEANYKKAVRLIAKLPTIVAFAHRIRRGKEPVAPRSDLSHAANF
LYMLNGEEDPLAARVMDVCLILHAEHEMNASTFTARVVGSTLADPYSIAAAAIGSLSGPLHGGANEVLKMLQEI
GSADNVEPWLEEKLATKQKIMGFGHRVYKTYDPRAKILKKMAEQLFEKHGSTPLYDIAVELEKVAERLGPKEIYPN
VDFYSGLVYQALGIPTDFTLFTPIFAIARVAGWTAHWLEQLEDNRIFRPRQVYVGPTDRKYVPIDQRLEHHHHHH

CS6-His Seq ID No. 20

MIAKGLEGVVIAETSISYIDGQEGRLYYRGYPIEELAEHSSFEVAYLLWHGRLPTREEEFKEELAKNRAIPEEIDL
RTLPKSAHPMAALRTAVSALGMFDPDADDTQSPEARYRKAIRLIAKIPTIVAFAHRIRRGQEPVAPRPDLSHAANFL
YMLNGEESPVPQAKVMDVALILHAEHEMNASTFAALVVASTLSDMYSAITAAIGALKGPLHGGANEQVMKMLQ
EIGSPDKAEPWWQEKLANKRRIMGFGHRVYKTYDPRAKILKKYARQLAEKQGDITLYEIAEAVEKVVVERLGPKEI
PNVDFYSGLVYQALGIPTDFTLFTPIFAMARVSGWTAHVLEYLKDNRILIRPRAVYVGPTDRKYVPIDQRLEHHHHHH

CS7-His Seq ID No. 21

MYTPGLAGVVAEASSISYIDGQEGILRYRGYPIEELAEHSSFEVAYLLFGKLPTRREEEFDELKHHRAIPEGIIDL
KTLPKSAHPMAALQSAVAALGMFYPADDDVEDPEANYEAAVRLIAKLPTIVAFAHRIRRGKDPIAPRSDLGHAANFL
YMLNGEEDPLAARVMDVCLILHAEHSMNASTFTARVVGSTLADPYSIAAAAIGSLSGPLHGGANEVLQMLQEI
GSAENVEPWLEEKLATKQKIMGFGHRVYKVKDPRATILQKMAEQLFEKHGSTPLYDIAVELEKVAERLGPKEIYP
NVDFYSGLVYQKLGIPDFTLFTPIFAIARVAGWTAHWLEQLEDNRIFRPTQIYVGPTDRSYVPIDQRLEHHHHHH

Sequence Name	SEQ ID No	Sequence Name	SEQ ID No
CS1	1	<i>cs1</i>	8
CS2	2	<i>cs2</i>	9
CS3	3	<i>cs3</i>	10
CS4	4	<i>cs4</i>	11
CS5	5	<i>cs5</i>	12
CS6	6	<i>cs6</i>	13
CS7	7	<i>cs7</i>	14
CS1-His	15	<i>cs1-His</i>	22
CS2-His	16	<i>cs2-His</i>	23
CS3-His	17	<i>cs3-His</i>	24
CS4-His	18	<i>cs4-His</i>	25
CS5-His	19	<i>cs5-His</i>	26
CS6-His	20	<i>cs6-His</i>	27

CS7-His	21	<i>cs7-His</i>	28
---------	----	----------------	----

Table 1 – Sequence names and corresponding SEQ ID No. *italics* indicates the nucleic acid sequence encoding the indicated protein

XP_021771303	WP_075632642	PZM92907	XP_034701501
XP_011657098	WP_062140045	WP_100278409	PON693981
XP_031258027	EGE48820	WP_038250090	XP_010087849
XP_007217210	AGG68320	WP_018293817	PON965231
XP_030930192	WP_025286398	WP_138238250	XP_030515497
XP_006448012	WP_048862517	WP_100264645	XP_030462884
XP_021843771	WP_027284291	MAQ19445	JAT552231
XP_020089322	WP_073132632	MAT30100	NUU834511
XP_031391603	WP_043834052	NLY94321	JAU500961
WP_100623629	WP_014746543	KPK15571	MPA322281
WP_009496991	WP_074768451	MBV70632	XP_010053061
WP_039888516	WP_008617533	MAD59944	XP_020249621
WP_025346249	WP_041041603	MSP74006	XP_015880346
WP_011373714	WP_045584945	WP_121990435	WP_041963167
WP_021286800	WP_029014739	WP_080887793	WP_014770600
WP_104412023	WP_004274395	WP_053379665	KIM10656
WP_014468117	WP_009540255	WP_090750018	AFV97513
WP_013554211	WP_033069782	WP_050726599	WP_013460302
WP_046551885	WP_064781486	WP_015165717	WP_066167812
WP_067332396	WP_018632472	WP_010873488	WP_014660224
WP_011139296	WP_038143270	WP_015192519	WP_006563885
WP_060472753	WP_038072739	QFZ93205	WP_089729779
WP_053940702	WP_071673378	BAD79102	WP_141608633
WP_013890021	WP_058291545	ABB56644	KPV52016
WP_034353392	WP_061049479	WP_036531233	WP_120553418
WP_023928952	WP_027259026	WP_017713050	WP_044185363
WP_077388477	WP_004600962	WP_071108047	WP_084527404
WP_006562450	WP_011075192	WP_045054973	WP_155379654
WP_012119141	WP_074025403	WP_104546235	WP_008456139
WP_015942419	ANR61903	WP_073550812	WP_122715784
WP_066788679	WP_018581733	WP_016860693	WP_026873888
WP_014432505	WP_085550151	WP_006199069	CRH91988
WP_059139550	WP_081357969	WP_169363481	WP_034999430
WP_095981408	WP_076761629	WP_175355653	WP_048643927
WP_110022383	WP_082080390	WP_103670708	WP_011738392
WP_011357099	WP_086535247	WP_019499953	WP_011930330
WP_011745897	WP_005198206	WP_150975134	WP_053951376

WP_006365367	WP_074850854	WP_015230097	WP_066045757
WP_012509079	WP_070894613	WP_124978658	WP_138146120
WP_071903157	WP_085156365	WP_013322203	WP_139475463
WP_043408006	WP_083051482	WP_012594711	WP_071564556
WP_047857464	WP_040855128	WP_073600083	WP_065412805
WP_013940115	WP_084536288	WP_015144810	RLS211811
WP_002640596	WP_084523192	WP_006277808	RLS59731
WP_093515602	WP_007529626	WP_072619680	RLS90155
WP_013862092	WP_040274279	WP_015212760	NLE49304
WP_067859713	WP_015888812	WP_017306814	WP_158023502
WP_084653578	WP_012828685	WP_017319174	WP_171574076
WP_022926837	WP_014066315	WP_109010387	WP_017311993
WP_084269874	WP_146647715	WP_100898806	WP_028948348
WP_093115919	WP_146734513	WP_017652774	WP_144865333
WP_183342938	KYF62269	WP_124145770	WP_088890046
WP_052666979	WP_044239490	WP_015204448	WP_164929714
WP_018636778	WP_050429001	WP_023068476	WP_012185374
WP_085003341	WP_027001279	WP_046278649	WP_013776897
WP_131764724	WP_070745390	WP_035758783	WP_039685122
WP_101831117	WP_013687488	WP_083623165	WP_022797330
WP_051323755	WP_010854200	WP_026785464	WP_018110572
WP_130649287	WP_092168647	WP_072718234	WP_054392095
WP_011601376	WP_014022688	WP_027255342	SHE09884
WP_091012498	WP_020891317	WP_083627063	<i>WP_012215482</i>
WP_061176888	WP_092897243	WP_073072652	<i>WP_048115900</i>
WP_019141269	WP_008199985	WP_080809297	<i>WP_015017867</i>
WP_016833002	WP_026952098	WP_169615075	<i>WP_075054685</i>
WP_058896046	WP_073573029	WP_073606441	<i>WP_148701439</i>
WP_145603639	WP_087939755	WP_172354665	<i>WP_147147647</i>
WP_005666166	WP_024283207	WP_011433503	<i>WP_013703859</i>
WP_020652857	WP_009186329	WP_023172997	<i>WP_010878837</i>
WP_091874894	WP_084123068	BAJ48827	<i>WP_148208684</i>
WP_154359079	WP_026999684	WP_068576448	<i>WP_193807863</i>
WP_070251567	WP_009284278	WP_011011316	<i>WP_015591338</i>
WP_076591519	WP_093825590	WP_010978565	<i>WP_156786146</i>
WP_038489891	WP_073478803	WP_148690635	<i>WP_107583781</i>
WP_061539808	WP_054030308	pdb_6ABX	<i>WP_028962312</i>
WP_202636164	WP_054693210	WP_013737319	<i>WP_013076317</i>
WP_011239598	WP_084553375	WP_009074395	<i>WP_100668299</i>
WP_026687301	WP_020877706	WP_027481744	<i>WP_155863975</i>
WP_066612634	WP_028319996	WP_013557451	<i>RLS21181</i>
WP_050894920	WP_013258312	WP_013614211	<i>RLS59736</i>
WP_037395186	WP_092207573	WP_019007989	<i>RLS9015</i>

WP_037588602	WP_084070770	WP_064014971	WP_062485514
WP_009118270	WP_014959233	WP_029477527	WP_011143005
WP_047760537	WP_022666926	WP_041221228	WP_193386667
WP_034296229	WP_013163875	WP_019585398	WP_179228407
WP_042924152	WP_015404178	WP_062157472	WP_186375862
WP_027008816	WP_028318877	WP_027479989	WP_140862691
WP_063143154	WP_035274005	WP_014685674	WP_081691039
WP_059178627	WP_074201701	WP_013457601	QDZ25721
PNP04344	WP_127470136	WP_013158070	PON69398
WP_166726827	WP_128384312	WP_038045386	PTQ33131
WP_000785805	WP_011369771	WP_038056021	XP_024362303
WP_192477935	WP_083917336	WP_038043130	PRW58871
WP_042320646	WP_029935991	WP_028494325	XP_001695571
WP_045110225	WP_044411821	WP_018465754	XP_007512659
WP_017222226	PCJ84499	WP_027893481	XP_002502663
WP_042028895	WP_123102589	WP_027883229	XP_001421921
WP_042023908	WP_014002210	WP_027878340	OUS45491
WP_067416364	WP_012537576	WP_012659955	AND95600
WP_062662721	WP_126456647	WP_076430839	AND95703
WP_004413186	WP_134083971	WP_049986530	CDW76733
WP_012549628	WP_014703737	WP_089694032	XP_002668999
CED70878	WP_008291662	WP_009366796	KRX04979
WP_018691731	WP_007146507	WP_089870224	XP_001431395
WP_013346000	WP_091715369	WP_123164019	BAD05063
WP_028108839	WP_093428830	WP_063177825	XP_001014802
WP_090364586	WP_002813885	WP_092097629	KXN72308
WP_028112303	WP_013219562	WP_013489793	XP_004355114
WP_068858516	WP_134356310	WP_091659977	XP_020438535
WP_032027738	WP_013031659	WP_029097729	KYR02419
WP_027848412	WP_125181773	WP_188067419	XP_647596
WP_060394219	WP_127475799	WP_089522205	XP_003291998
WP_196180145	WP_092992342	PLT46992	KJE93333
WP_043200037	WP_149108826	WP_043072376	ORZ33874
WP_017902731	WP_088251988	WP_093227907	PKK68967
CCK75855	WP_010038645	WP_092038800	KAF0555079
WP_028292555	WP_162669086	WP_089966156	ORX43392
WP_027313361	WP_029630008	WP_068135191	ORX76294
WP_094061195	WP_162658391	WP_013495717	ORY64656
WP_091823387	WP_171474460		OON03039
WP_027857214	MSR49827		XP_016608261

Table 2: RefSeq Accession codes for extant citrate synthase proteins used to construct Trees 1 and 2 as shown in Figures 1-3. Plain text accession codes (first three columns) are common to both Trees.

Accession codes in bold are found only in Tree 1 (Figure 2). Accession codes in italics are found only in Tree 2 (Figure 3).

Sequence [Seq ID No.]	Sequence Identity to CS1 (Seq ID No. 1)	Sequence Similarity to CS1 (Seq ID No. 1)
CS2 [2]	88	95
CS3 [3]	78	87
CS4 [4]	75	84
CS5 [5]	82	90
CS6 [6]	96	98
CS7 [7]	78	86

Table 3: Sequence identity and similarity of CS2-7 when compared to CS1. Results generated using EMBOSS NEEDLE program https://www.ebi.ac.uk/Tools/psa/emboss_needle/

Example 2 – Expression and purification of citrate synthase proteins

pET vectors comprising Seq ID Nos. 22-28 were transformed into BL21(DE3) cells which were cultured at 30 °C with shaking in 2 L baffled flasks containing 500 mL LB medium supplemented with 6.25 g lactose and 100 µg/mL carbenicillin. Cells were cultured for approximately 16 hours after which time the cells were harvested by centrifugation for 15 mins at 4500 ×g.

Cells were resuspended in 20 mM Tris:HCl, 300 mM NaCl, 20 mM Imidazole pH 8.0 buffer (Buffer A) at a ratio of 30 mL per litre of original cell culture. Cells were lysed using a microfluidizer with 3 cycles at 15000 psi. Lysed cells were ultracentrifuged at 30000 ×g for 30 mins to remove membranes and cell debris. The clarified lysate was then filtered through a 0.45 micron syringe filter.

The lysate was loaded onto a Nickel-NTA preppacked column pre-equilibrated with Buffer A. The bound protein was washed with 7 column volumes of Buffer A and 7 column volumes of 20 mM Tris:HCl, 300 mM NaCl, 68 mM Imidazole pH 8.0 (Buffer B). Bound protein was then eluted in 20 mM Tris:HCl, 300 mM NaCl, 500 mM Imidazole pH 8.0 (Buffer C). Eluted protein was exchanged into 20 mM Tris:HCl, 200 mM NaCl, pH 7.5 (Buffer X) or Phosphate buffered saline (PBS) using a PD-10 column.

Protein is sufficiently pure to use as a molecular weight standard/mass photometry calibrant and can be frozen in liquid nitrogen and stored for extended periods at -20 °C. SDS-PAGE (Fig. 4) and Native PAGE (Fig. 5) were performed on the samples.

The resulting protein can be enriched for higher molecular weight complexation states using SEC. This step can be performed on the protein after elution from the Nickel-NTA column, or after the step of exchanging into Buffer X.

An analytical SEC-column Enrich 650 (Biorad) was equilibrated with Buffer X. Protein sample was concentrated to 5-25 mg/mL protein and 250 µL was loaded onto the column. The sample was eluted at 1 mL/min. The sample eluted as a first, broader peak encompassing higher molecular weight (HMW) complexation states (trimers and above) and a second sharp peak attributed to lower molecular weight (LMW) complexation states (dimers and below) however the two peaks could not be completely resolved from each other.

Example 3 – Mass photometry measurements of citrate synthase proteins

Reusable silicone gaskets (CultureWellTM, CW-50R-1.0, 50-3mm diameter x 1 mm depth) were set up on a cleaned microscopic cover slip (1.5 H, 24 x 60 mm, Carl Roth) and mounted on the stage of the mass photometer (Refeyn Ltd., UK) using immersion oil (Immersion Oil 518F, Zeiss). The gasket was filled with 19 µl buffer (PBS or Buffer C) to focus the instrument. The protein was prediluted to a concentration of approx. 400nM in the same buffer. Then 1 µl of prediluted protein solution was added to the buffer droplet and mixed thoroughly. Final concentration of the proteins during measurement was 20 nM. Data was acquired for 60 s at 100 frames per second.

Nucleic acid Seq ID No.	Citrate synthase Seq ID No.	Monomeric molecular weight (kDa)	Radiometric contrast at peak	Relative peak area (%)	Corresponding mass (kDa)	Mass (kDa)/contrast ($\times 10^{-5}$) [Maximum error (%)]
CS1-His	15	42.9	-0.00295	60	86	-3.406 [0.9]
			-0.00586	10	172	
			-0.00875	6	258	
			-0.01174	4	344	
CS1-His HMW	15	42.9	-0.00315	10	86	-3.730 [1.0]
			-0.00643	19	172	
			-0.00961	37	258	
			-0.01279	15	344	
			-0.01601	4	430	
CS2-His	16	43.5	-0.00302	75	86	-3.500 [0.5%]
			-0.00604	10	172	
			-0.00899	4	258	
			-0.01207	3	344	
CS3-His	17	43.3	-0.00326	60	86	-3.840 [0.5]
			-0.00653	16	172	
			-0.00991	7	258	
			-0.01314	3	344	
CS4-His	18	43.1	-0.00165	22	43	-3.623 [1.1]
			-0.00316	30	86	
			-0.00635	10	172	
			-0.00941	7	258	
CS4-His HMW	18	43.1	-0.00332	15	86	-3.737 [0.8]
			-0.00489	17	129	
			-0.00655	36	172	
			-0.00806	6	215	
			-0.00977	6	258	
CS5-His	19	42.8	-0.00288	50	86	-3.040 [1.1]
			-0.00541	23	172	
			-0.00812	12	258	
			-0.01069	9	344	
CS6-His	20	43.1	-0.00305	61	86	-3.401 [0.2]
			-0.00596	28	172	
			-0.00890	5	258	
CS7-His	21	42.4	-0.00294	72	85	-3.376 [1.2]
			-0.00571	8	170	
			-0.00868	4	255	

Table 4: Mass photometry data for CS1-His to CS7-His.

Example 4 – Reproducibility across batches

To ensure the batch-to-batch consistency of the mass photometry data several further purifications of CS1-His were prepared and tested as described above. Mass photometry (MP) experiments were performed using a Two^{MP} mass photometer (Refeyn Ltd, UK) using Acquire^{MP} software. Mass photometry plots for 20 nM samples of purified CS1-His from two separate purifications from two cell lysates are shown in Figure 6a - d, all showing the same peaks with only small differences in the ratio of peak areas between each batch. All of these samples are suitable for use as a mass photometry standard.

Example 5 – Use of CS1-His as a mass photometry standard for proteins of known molecular weight

CS1-His was used to calibrate a Two^{MP} mass photometer as described above and proteins of known molecular weight were analysed by mass photometry. Sample concentration was 10-40 nM depending on the protein being analysed. Figure 7 shows the proteins' masses as determined by mass photometry when calibrated using CS1-His are shown next to the masses expected from the proteins' primary sequences. The results show that calibration with CS1-His leads to calculated masses that are within 2% of the expected masses meaning the calibration is effective.

Example 6 – Comparison with NativeMarkTM as a calibration standard

A commonly used calibration standard for mass photometry, prior to the development of the citrate synthase proteins of the invention, was NativeMarkTM unstained protein standard (InvitrogenTM) which is designed for use as a standard for native PAGE. NativeMarkTM is advertised as containing 8 proteins (1236, 1048, 720, 480, 242, 146, 66 and 20 kDa).

NativeMarkTM was compared against CS1-His HMW for the usability and accuracy of the calibration. NativeMarkTM was diluted 250-fold in PBS whilst CS1-His HMW was diluted to 40 nM (monomer concentration). After setting up the mass photometry device, 10 µL of the diluted protein was added to the well and a 1 minute movie was recorded. The data was then analysed using Discover^{MP} software (RefeynTM). The comparison between Native MarkTM and CS1-His HMW was performed on the same day.

Figures 8 a and b shows a comparison of the results for the two standards. As can be clearly seen, the peaks obtained with NativeMarkTM are heavily weighted towards a single species which accounts for 67% of the binding events. As a result, the low intensity peaks have a lower signal to noise ratio and are less accurate. The error for the mass/radiometric contrast gradient for NativeMarkTM is 5.8%, whereas for the data set shown for CS1-His HMW the error is only 0.7%. A more concentrated sample

of NativeMark™ could be used to achieve higher intensity peaks, and thus more accurate data, for the higher molecular weight species but this adds complexity, time and cost.

These results are not overly surprising as NativeMark™ is not designed and formulated for use as a mass photometry standard, instead being designed to achieve roughly similar staining intensity when stained after running in native PAGE.

References

Guindon S. & Gascuel O. PhyML : "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." *Syst. Biol.* **52(5)**:696-704 (2003) DOI: 10.1080/10635150390235520).

Harms, M., Thornton, J. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat Rev Genet* **14**:559–571 (2013). <https://doi.org/10.1038/nrg3540>

Harms, M.J., Thornton, J.W. Analyzing protein structure and function using ancestral gene reconstruction. *Curr. Opin. Struct. Biol.*:**20**, 360–366. (2010). <https://doi.org/10.1016/j.sbi.2010.03.005>.

Needleman, S.B., Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *JMB* **48**:443-453 (1970). [https://doi.org/10.1016/0022-2836\(70\)90057-4](https://doi.org/10.1016/0022-2836(70)90057-4)

Selberg, A.G.A., Gaucher, E.A. & Liberles, D.A. Ancestral Sequence Reconstruction: From Chemical Paleogenetics to Maximum Likelihood Algorithms and Beyond. *J Mol Evol* **89**:157–164 (2021). <https://doi.org/10.1007/s00239-021-09993-1>

Schmidtman et al., Redox Regulation of Arabidopsis Mitochondrial Citrate Synthase. *Mol. Plant* **7**:156-159 (2014) doi.org/10.1093/mp/sst144.

Stamatakis A. Randomized Accelerated Maximum Likelihood *Bioinformatics*. **30(9)**:1312-3 (2014). doi: 10.1093/bioinformatics/btu033.

Young G, Hundt N, Cole D, et al., Quantitative mass imaging of single biological macromolecules. *Science*. **360**:423-427 (2018). doi:10.1126/science.aar5839

The above description provides several methods, proteins and nucleic acids of the invention. This invention is susceptible to modifications in the methods and materials. Such modifications will become apparent to those skilled in the art from a consideration of this application or practice of the invention provided herein. Consequently, it is not intended that this invention be limited to the specific embodiments provided herein, but that it cover all modifications and alternatives coming within the true scope and spirit of the invention as embodied in the attached claims. All patents, applications, and other references cited herein are hereby incorporated by reference in their entirety.

Clauses:

- A. A citrate synthase protein, or functional fragment thereof, comprising at least 1 ancestral mutations wherein the protein is capable of forming at least 2, at least 3, at least 4 or at least 5 complexation states simultaneously.
- B. A citrate synthase protein according to clause A, wherein the protein comprises at least 2, at least 3, at least 4, at least 5, at least 10, at least 15 or at least 20 ancestral mutations.
- C. A citrate synthase protein, or functional fragment thereof, having an amino acid sequence at least 75%, 80%, 85%, 90% or 95% identical to a sequence selected from SEQ ID NOs: 1-7 wherein the protein is capable of forming at least 2, at least 3, at least 4 or at least 5 complexation states simultaneously.
- D. A citrate synthase protein, or functional fragment thereof, having an amino acid sequence with at least 75%, 80%, 85%, 90% or 95% sequence similarity to a sequence selected from SEQ ID NOs: 1-7 wherein the protein is capable of forming at least 2, at least 3, at least 4 or at least 5 complexation states simultaneously.
- E. A citrate synthase protein, or functional fragment thereof, comprising or consisting of an amino acid sequence selected from any of SEQ ID NOs: 1-7.
- F. A citrate synthase protein according to clause E wherein the citrate synthase protein comprises the sequence SEQ ID NO: 1.
- G. A citrate synthase protein of any of clauses A-F wherein the protein further comprises a tag to facilitate purification.
- H. A citrate synthase protein of clause G wherein the purification tag is a His-tag.
- I. A citrate synthase protein of any of clauses A-H wherein the protein is not expressed by an extant organism.
- J. A nucleic acid sequence encoding a citrate synthase protein according to clauses A-I.
- K. A nucleic acid sequence encoding a citrate synthase protein, or functional fragment thereof, wherein the nucleic acid sequence is an ancestral gene and wherein the citrate synthase protein encoded by the nucleic acid sequence is capable of forming at least 2, at least 3, at least 4 or at least 5 complexation states simultaneously.
- L. The citrate synthase protein, or functional fragment thereof, encoded by the nucleic acid sequence of clause K.
- N. The nucleic acid sequence according to clause J or clause K wherein the sequence has been codon optimised for expression in an organism, preferably the sequence has been codon optimised for expression in E.coli.
- O. A vector comprising the nucleic acid sequence of any of clauses J, K or N.
- P. The vector of clause O wherein the vector is a plasmid, preferably the vector is a pET plasmid
- Q. A cell comprising the nucleic acid sequence or vector according to any of clauses K, clause N or clause O.
- R. The cell according to clause Q wherein the cell is an E. coli cell.

- S. The E. coli cell according to clause R, wherein the cell expresses a citrate synthase protein according to any of clauses A-I or clause L.
- T. A method of producing a citrate synthase protein of any of clauses A-I or clause L comprising culturing a bacterium according to clause R and carrying out at least one purification step.
- U. A composition comprising a citrate synthase protein according to any of clauses A-I or clause L wherein the concentration of the citrate synthase protein is at least 5 nM, preferably the concentration is between 5 nM and 100 μ M.
- V. The composition according to clause U wherein the composition further comprises at least one component selected from a buffering reagents and a salt.
- W. Use of a citrate synthase protein according to clauses A-I or clause L, or a composition according to clauses U or V as a molecular weight standard.
- X. The use of a citrate synthase protein or a composition as a molecular weight standard according to clause W wherein the use is as a calibrant for mass photometry.
- Y. The use of a citrate synthase protein or a composition as a molecular weight standard according to clause W wherein the use is as a molecular weight standard native polyacrylamide gel electrophoresis.
- Z. A method of calibrating a mass photometry device comprising the step of assaying the composition of clause W or clause X.

Claims:

1. A citrate synthase protein, or functional fragment thereof, encoded by a nucleic acid sequence wherein the nucleic acid is an ancestral gene and wherein the citrate synthase protein, or functional fragment thereof, is capable of forming at least 3 complexation states simultaneously.
2. A citrate synthase protein, or functional fragment thereof, which is capable of forming at least 3 complexation states simultaneously, and has an amino acid sequence selected from:
 - i) an amino acid sequence at least 75%, 80%, 85%, 90% or 95% identical to a sequence selected from SEQ ID NOs: 1-7;
 - ii) an amino acid sequence with at least 75%, 80%, 85%, 90% or 95% sequence similarity to a sequence selected from SEQ ID NOs: 1-7;
 - iii) an amino acid sequence selected from any of SEQ ID NOs: 1-7; and/or
 - iv) an amino acid sequence comprising at least 1 ancestral mutation.
3. A citrate synthase protein, or functional fragment thereof as claimed in claim 1 or claim 2, wherein said citrate synthase protein, or functional fragment thereof is capable of forming at least 4 complexation states simultaneously.
4. A citrate synthase protein, or functional fragment thereof as claimed in any one of claims 1 to 3, wherein each of the three most abundant complexation states comprise at least 15% of the total number of citrate synthase protein complexes present.
5. A citrate synthase protein, or functional fragment thereof, according to any one of claims 1-4 wherein the citrate synthase protein, or functional fragment thereof, further comprises a tag to facilitate purification, preferably a His-Tag
6. A citrate synthase protein, or functional fragment thereof, according to any one of claims 1-5 wherein the citrate synthase protein, or functional fragment thereof, is not expressed by an extant organism.
7. A composition comprising a citrate synthase protein, or functional fragment thereof, according to any one of claims 1-6 wherein the concentration of the citrate synthase protein, or functional fragment thereof, is at least 5 nM, optionally the concentration is between 5 nM and 100 μ M.

8. Use of a citrate synthase protein, or functional fragment thereof, according to any one of claims 1-6, or a composition according to claim 7 as a molecular weight standard.
9. The use of a citrate synthase protein, function fragment thereof, or a composition as a molecular weight standard according to claim 8 wherein the use is as a calibrant for mass photometry.
10. A method of calibrating a mass photometry device comprising detecting the mass of particles in the composition of claim 7 on a mass photometry device.
11. A nucleic acid sequence encoding a citrate synthase protein, or functional fragment thereof, according to any one of claims 1-6.
12. The nucleic acid sequence according to claim 11 wherein the sequence has been codon optimised for expression in an organism, preferably the sequence has been codon optimised for expression in *E.coli*.
13. A vector comprising the nucleic acid sequence of claim 11 or claim 12, preferably a pET plasmid
14. A cell comprising the nucleic acid sequence of claim 11 or claim 12 or the vector according to claim 13, preferably an *E. coli* cell.
15. The cell according to claim 14, wherein the cell expresses a citrate synthase protein, or functional fragment thereof, according to any one of claims 1-6.
16. A method of producing a citrate synthase protein, or functional fragment thereof, according to any one of claims 1-6 comprising culturing a cell according to claim 14 or 15 and carrying out at least one purification step.