Exploring an alternative metabolic pathway for production of adipic acid - a study on homocitrate synthase

Damaride Stallone

Table of contents

Glossary

acCoA - acetylCoA

AvHCS – *Azotobacter vinelandii* homocitrate synthase

GHG – greenhouse gasses

HCS – Homocitrate synthase

IPMS - isopropylmalate synthase

ScHCS – *Saccharomyces cerevisiae* homocitrate synthase

TtHCS – *Thermus thermophilus* homocitrate synthase

α-KA – alpha-ketoadipate

α-KG – alpha-ketoglutarate

Abstract

Adipic acid is considered as the most valuable dicarboxylic acid with industrial use; however, its production is heavily petroleum-based. This non-natural metabolite can be produced *de novo* through different metabolic pathways, one of which is the alpha-ketoacid elongation pathway. This investigation revolved around the study of the enzyme that catalyses the first +1C elongation reaction of the pathway: homocitrate synthase. Homocitrate synthases expressed in *Saccharomyces cerevisiae*, *Thermus thermophilus* and *Azotobacter vinelandii* were chosen as candidates. These candidates were evaluated on features such as: identification of the catalytic pocket residues; substrate specificity and possibility to widen such specificity though rational mutagenesis; literature available on the topic; reproducibility of the enzyme expression conditions. The fitting candidate would be able to employ C5 and C6 ketoacid as substrates - either naturally or through mutagenesis -, be fairly simple to express, isolate and test for enzyme activity using different substrates. Of these candidates, homocitrate synthase from *T. thermophilus* was successfully expressed and tested for enzymatic activity against alpha-ketoglutarate. *In silico* protein models for homocitrate synthase from *S. cerevisiae* and *A. vinelandii* were produced, along with the identification of the residues in the catalytic pocket of the two enzymes. This investigation is concluded by the choice of homocitrate synthase expressed from *A. vinelandii* as the best fitting candidate (between the one examined) for the +1C elongation step of the alpha-keto acid pathway finalized to the production of adipic acid.

Introduction

Changing the point of view: biomass for production of plastics. Plastic is a common global topic of discussion, primarily due to the enormous amounts of plastic waste generated, which lead to environmental damage caused by land and sea littering. Fossil-based plastics dominate the markets, to the price of great amount of greenhouse gasses (GHG) emissions. Nevertheless, plastic is a ubiquitous and necessary material, and the exponential growth of the world population will more likely determine a greater demand for its availability. Reduction of GHG emissions and human-induced climate change calls for plastic manufacturing to move away from fossil feedstock to search for valuable alternatives.

However, there is an alternative to petroleum-based plastics: biobased plastics. Biobased plastics date back to the 19th century with the production of parkesine, a cellulose derivate material (Parkes, 1865); industrial interest in such materials during the 70s-80s sparked for concerns about the reducing oilfields available. Between 90s and 00s, the concern finally becomes of environmental nature. The definition of biobased plastics (fully or partially), is of a plastic material whose base compounds are made using renewable carbons. Examples of partly biobased plastics are bio-polyethylene terephthalate (PET) or polytrimethylene terephthalate (PTT); fully biobased count bio-based polyethylene (PE), as well as polylactic acid (PLA) and polyhydroxyalkanaoates (PHA).

These biobased plastics, being biodegradable or non-biodegradable, are the product of great advancements in industrial biotechnology, which has played a prominent role in biobased production; starting from providing catalysts for hydrolysis of the biomass polymers to simple base compounds, which are then transformed to polymer building blocks or a whole polymers. The list of advantages for employing microorganism in this type of production is conspicuous. Thanks to the advancement into biotechnological methodologies, microorganisms can proliferate in bioreactors, are generally small but robust systems of expression, and are easily engineered. Some of the polymeric building blocks are simply natural metabolites of microorganisms, such as lactic acid (Juturu and Wu, 2016), succinic acid (Cao *et al.*, 2013), 1,3-propanediol (Ting *et al.*, 2014); increasing yield improving production can be reached by genetic and metabolic engineering of the organisms. However, this is also true for non-natural metabolites. Employment of microorganism for production of natural and non-natural building blocks designated to biobased plastics is therefore already a reality. Up to now, only building blocks with short carbon chain were successfully produced with these methods; nevertheless, the industry requires biobased long chain aliphatic and aromatic

building blocks to fulfill the demands for different varieties of plastics. In fact, biobased plastics constitute less than 1% of the 355 million tons of traditional plastic produced each year (Figure 1).
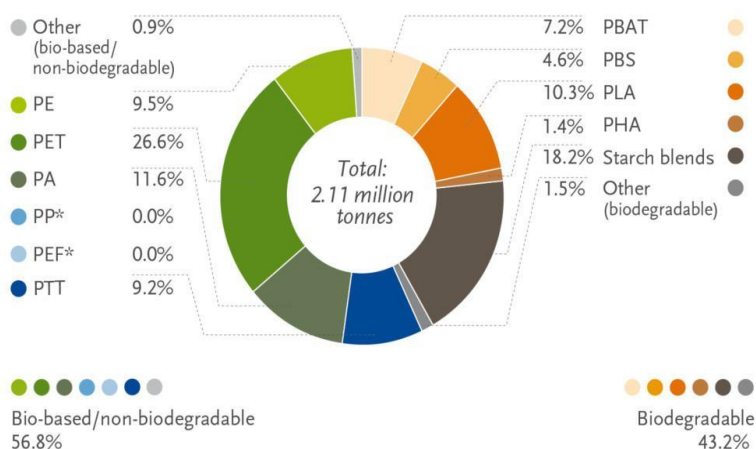

Figure 1. Percentages of global production capacities of bioplastics for the year 2018, divided by types. Reproduced from (European Bioplastics, nova-Institute, 2018; https://www.european-bioplastics.org/market/)

That is why the field of production of biopolymers is ever-growing and setting higher goals for itself, such as developing new properties and alternatives for high-performance traditional plastics. The scientific community is working to bring these polymers, used as building blocks for bioplastics, to an optimal production flow, competitiveness against the petroleum-based counterparts, and general wider use.

   Why adipic acid? Adipic acid (1,6 Hexanedioic acid) is a dicarboxylic acid with a market that is worth $7,539.2 million in 2019 (Adipic Acid Market, 2014), considered as the most valuable dicarboxylic acid with industrial use by the International Energy Agency (IEA) (de Jong *et al.*, 2012). This market concerns the employment of adipic acid as precursor of nylon 6,6 and other polyesters mostly in the automotive, electrical and electronics field. However, its petroleum-based production is the cause of a conspicuous amount of greenhouse gasses emissions (mostly nitrous oxide – US EPA, Reimer *et al.*, 1999). In fact, the oxidization of a cyclohexanol-cyclohexanone mixture (called KA, ketone-alcohol oil) using nitric acid (HNO_3) to produce adipic acid causes nitrous oxide to be coproduced (Figure 2) (Parmon *et al.*, 2004).

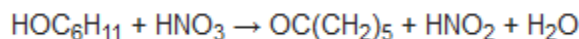$$HOC_6H_{11} + HNO_3 \rightarrow OC(CH_2)_5 + HNO_2 + H_2O$$

Figure 2. The oxidation of the cyclohexanol-cyclohexanone mixture caused by nitric acid results in the production of adipic acid, nitrous oxide and water. Nitrous oxide is a known greenhouse gas.

In recent years, a biodegradable polymer called PBAT (polybutylene adipate terephthalate) with adipic acid as one of the components has been introduced into the market (Jiang *et al.*, 2005). The long carbon chain of adipic acid allows easy access for microbial enzymes to break the ester bonds in the polymer, hence improving its biodegradability. PBAT is still fossil-based, however it can be potentially become biobased.

   During the 90s and 00s, different research teams have focused on production of *de novo* adipic acid in microorganisms such as *Escherichia coli* or *Saccharomyces cerevisiae* (Bart and Cavallaro, 2015; Boussie *et al.*, 2010; Draths and Frost, 1994; Vardon *et al.*, 2015) through different metabolic pathways. Literature on the topic is still fairly scarce, even after almost 20 years, however it points out a key approach to the issue. Since metabolic pathways often involve several intermediate compounds, there are many possibilities of diverting, fine-tuning or establishing the production of a specific molecule by intervening

on the pathway itself. This is the field of synthetic biology, which focuses onto the design of expression circuits that perform non-native tasks (Voigt, 2006). Consequently, the choice of a specific biosynthesis route lies in the hands of the research team. One discriminating feature is efficiency of carbon utilization, which determines how many carbon atoms from the original substrate are converted into the product, and how many will be destined to be the by-products. Adipic acid has been considered a valuable candidate for bio-based production (de Jong *et al.*, 2012), therefore is still gaining the interest of the scientific community. Yu *et al.* (2018) provided an overview of several means of bio-based *de novo* production of high straight-chain dicarboxylic acids; figure 3 summarizes which of these biosynthesis pathways result in adipic acid.
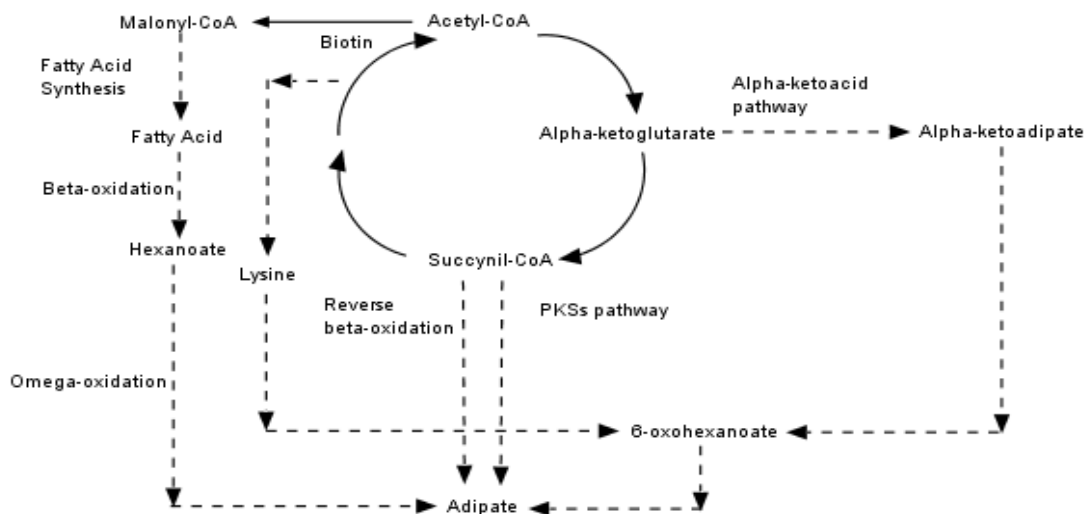

Figure 3. Overview of pathways that result in adipic acid, summarized. Original data (Yu *et al.*, 2018).
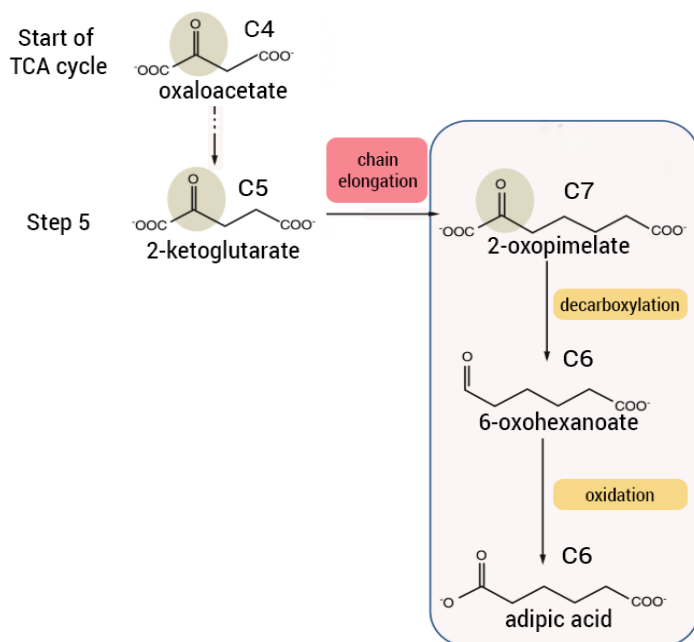

Figure 4. Elongation of 2-ketoglutarate produced in the TCA cycle, followed by decarboxylation and oxidation of the compound to result into adipic acid. Original graphics: Jambunathan and Zhang., 2014

One of the potential pathways involves elongation of alpha-keto acid formed during aerobic metabolism to a C7 keto acid followed by decarboxylation (performed by an alpha-ketoacid decarboxylase) and oxidation (performed by an aldehyde dehydrogenase) to adipic acid (Figure 4). The alpha-keto acid to be employed could be, for example, alpha-ketoglutarate (α-KG), a crucial C5 keto acid intermediate formed in Krebs cycle by chain elongation of oxaloacetate (C4). This choice would assure the presence of the needed intermediate in the host cells. Maximum yield of this theoretical pathway was determined to be 0.67 mol of adipate per mol of glucose in aerobic conditions; 0.45 mol of adipate per mol of glucose in anaerobic conditions (Burgard *et al.*, 2016). These numbers mean nothing when compared to palm oil and glycerol pathways that can produce respectively 50 g adipic acid/L and 68 g

adipic acid/L ([Beardslee and Picataggio, 2012](); [Zhao *et al.*, 2018]()). Nevertheless, biobased manufacturing of adipic acid that does not interfere with food-related production (palm oil and glycerol) is very much a need for a biobased economy ([Skoog *et al.*, 2018]()). For this reason, studies such as the present one are
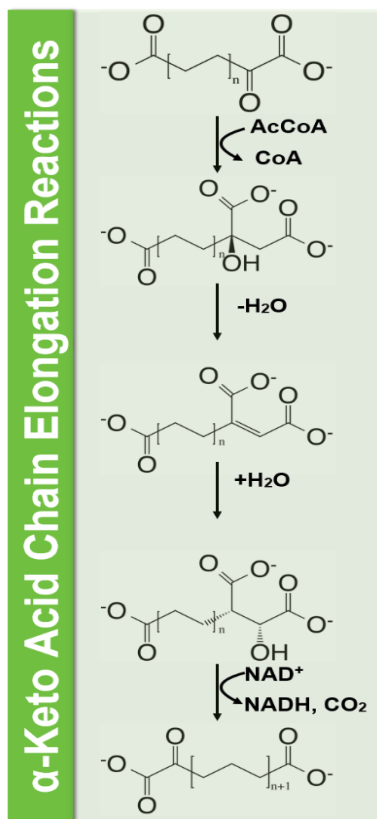

Figure 5. Overview of the metabolic flux for the overall project.

required to further characterize alternative, biobased pathways to adipic acid that might be employed in the future.

Aim of the project. The interest upon investigating this particular type of biosynthesis sparked when reading about the +1C elongation reaction of alpha-keto acids. Natural-occurring recursive chain elongation reaction occur in the biosynthesis of fatty acids, where the substrate undergoes several reactions involving addition of a C atom, while maintaining the functional group of the initial compound. Chain-elongation reactions are also found for keto-acids in the realm of the biosynthesis of branch-chained amino acids, with the difference of not being recursive ([Shen and Liao, 2011]()). Therefore, the aim of the overall project is to turn a single-direction elongation reaction into a flexible biosynthetic pathway for recursive '+1' elongation of the carbon chain of branched-chain, linear-chain, and even aromatic-chain keto acids (figure 5). The goal of the present investigation was focused on finding a fitting enzymatic candidate that catalyzes the +XC elongation, test its substrate specificity and the possibility of widening such specificity through active site mutagenesis. Since the elongation step must be performed two times to yield a C7 product, the chosen enzyme should be capable of catalyzing both C5 and C6 ketoacid substrates. Examination of the candidate´s catalytic pocket and overall role into native metabolic pathway would then qualify the enzyme as appropriate for follow-up studies concerning production of adipic acid through use of linear chain ketoacids.

Homocitrate synthase. In order to convert the keto-acid substrate, α-KG, into a carboxylic acid, the choice of the catalyst landed on Homocitrate synthase. Homocitrate synthase (acetyl-coenzyme A: 2-ketoglutarate C-transferase; HSC) is a metalloenzyme belonging to the family of acyltransferases, whose catalytic ability converts acyl groups into alkyl groups. HCS is also a wide-spread enzyme, being present in all domain: Archaea ([Howell *et al.*, 1998]()), Bacteria ([Zheng *et al.*, 1997]()), and Eukarya ([Chen *et al.*, 1997]()). The cellular pathways in which HCS is included are species-dependent: some concern pyruvate metabolism, others antibiotics biosynthesis or cofactor expression. However, the most remarkable role is undertaken by HCS is in the lysine biosynthesis, where it catalyzes the first reaction, condensing acetyl-CoA (acCoA) and α-KG into homocitrate (R)-2-Hydroxybutane-1,2,4-tricarboxylate). (Figure 1.1) ([Kegg, entry 2.3.3.14]())
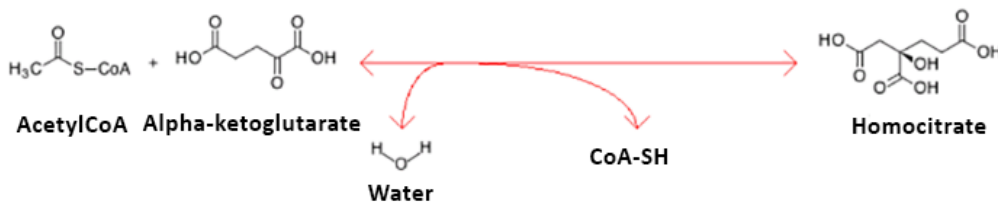

Figure 1.1. The reaction catalysed by HCS employs acetylCoA and alpha-ketogluarate to produce homocitrate with the release of CoA-SH and a water molecule ([Kegg, entry 2.3.3.14]()).
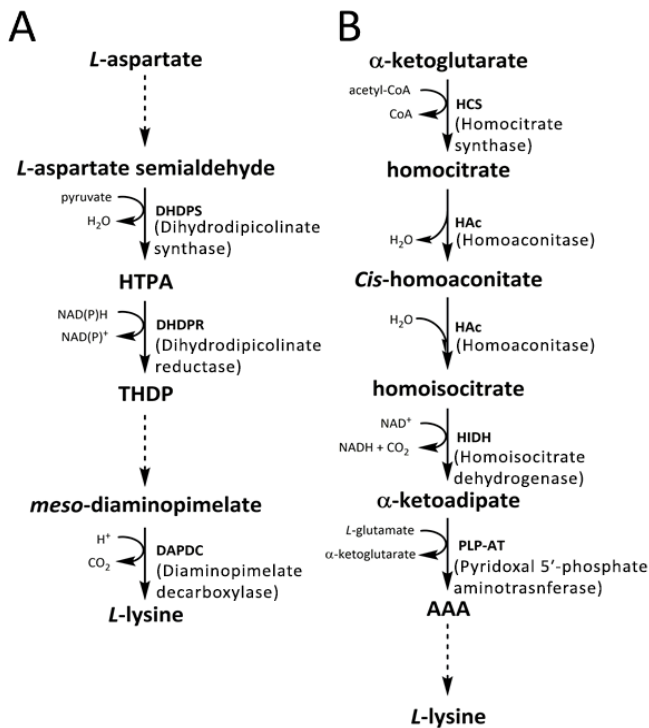
Figure 1.2. Pathway A, the diaminopimelate pathway, concerns lysine production in most bacteria and plants. Pathway B, the α-aminoadipate pathway, concerns higher fungi. Reproduced from: (Hall and Soares de Costa, 2018)

The lysine biosynthesis pathway is differentiated in two branches. In most bacteria and plants, lysine is produced through the diaminopimelate pathway (Vogel *et al.*, 1960). However, in higher fungi and some archaebacteria, lysine is produced through the α-aminoadipate (AAA) pathway (Garrad *et al.*, 1992) (Figure 1.2). Our target protein, HCS, is in fact the key enzyme of the AAA pathway, since it also the endpoint of a negative-feedback type of inhibition caused by lysine (Andi *et al.*, 2004). The reaction is a mixed aldol Claisen condensation that requires the presence of a general acid and a general base (Qian *et al.*, 2006). Kinetic studies on HCS identified its reaction to be a sequential Bi-Bi mechanism, as it deals with two substrates whose docking order is vital for the reaction to take place. Docking of the first substrate – α-KG – determines a structural change of the enzyme that is needed for acCoA to take contact with its active site. After the irreversible hydrolysis of the homocitryl-CoA intermediate, CoA is released first, then finally homocitrate (Figure 1.3). The reaction is irreversible (Andi *et al.*, 2004).

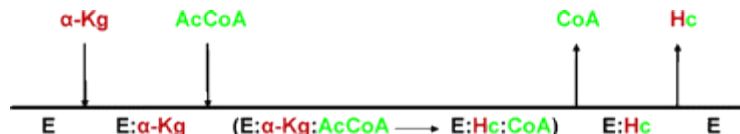

Figure 1.3. The Bi-Bi mechanism of HCS. α-KG will be the first substrate to be bound to the enzyme, triggering a morphological change of the protein, now open for acCoA docking. A homocitryl-CoA intermediate in formed, then CoA is released, followed by homocitrate. Original graphics: (Andi *et al.*, 2004)

The only HCSs to have been successfully crystallized are the ones in *Thermus thermophilus* and *Schizosaccharomyces pombe* (Okada *et al.*, 2008; Bulfer *et al.*, 2009). Both these structures are described as domain-swapped homodimers. Domain-swapping occurs when two monomers become entangled for dimerization, a rare feature – only 60 proteins exhibited this type of structure in 2012 (Rousseau *et al.*, 2012). The N-terminal – a TIM barrel domain - contains the active site of the enzyme, while the C-terminal has two subdomains (I and II). The C-terminal subdomain II of one single monomer stretches up to the active site of the other monomer, meaning the catalytic pocket is formed by the N-terminal of the first monomer and the ending C-terminal of the second monomer. Moreover, a lid motif has been identified in both *T. thermophilus* and *S. pombe* HCSs, a flexible portion of the C-terminal subdomain II which allows the availability of the active site (Okada *et al.*, 2008; Bulfer *et al.*, 2009).

These structural features are closely linked to the catalytic process of HCS. As stated, α-KG is docked in the active site first, when the lid motif allows the substrates and water molecules to occupy the catalytic pocket. The presence of α-KG helps coordinate the neighbour amino acids through a network of hydrogen

bonds: this is a structural change that allows the active site to also accommodate AcCoA. The pantetheine arm of AcCoA, which is the portion of the molecule that is most important for the reaction, is oriented close to α-KG. Once all substrates are docked, the lid motif closes on itself, keeping the molecules tightly linked and oriented with the amino acid residues of the catalytic pocket. At this point, the enzyme is ready to perform its reaction. The acetyl portion of AcCoA is deprotonated by a residue acting as a general base; this newly formed nucleophile takes contact with the C2 of α-KG, an intermediate that is protonated by an amino acid residue acting as a general acid. This intermediate, homocitryl-CoA, is then divided by the hydrolysis of a water molecule on the thioester bond, allowing CoA to be first released, and the homocitrate (Bulfer *et al.*, 2009). However, it must be reminded that since HCS is a species-specific enzyme, structure and catalysis features might variate greatly from these two references.
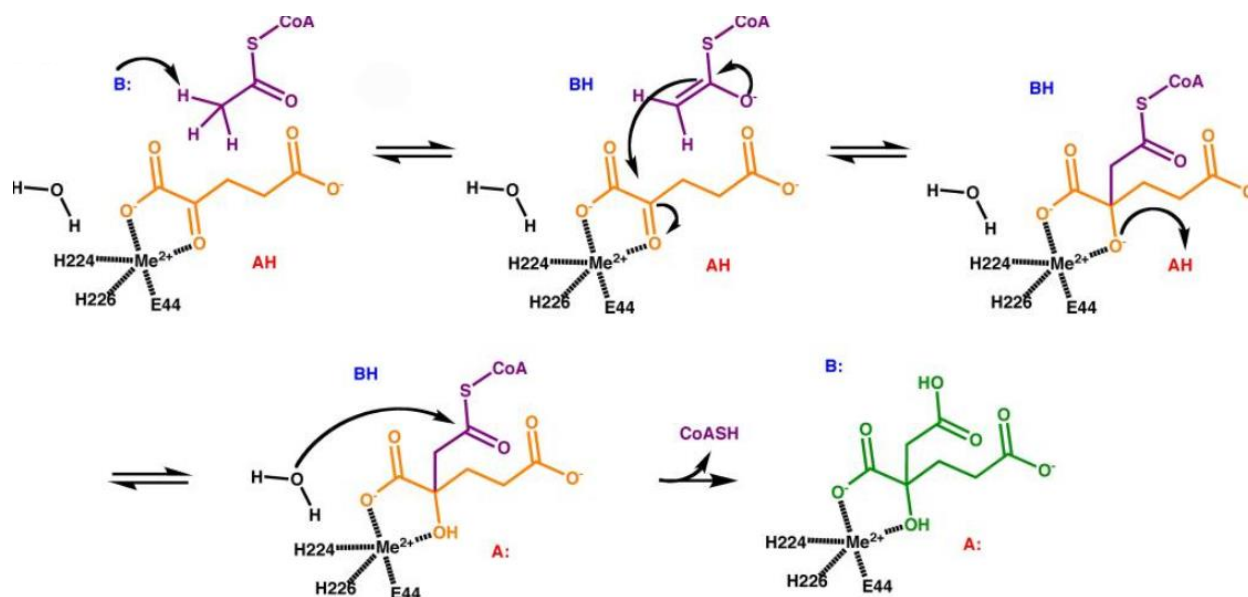

Figure 1.4. Proposed mechanism for HCS reaction in *S. pombe*. In yellow, α-KG; in purple, AcCoA; Me^{2+} defines a bivalent ion. The acyl group is deprotonated due to the basic residue, labelled B (in blue). The nucleophile attacks the C2 of α-KG, forming an intermediate called homocitryl-CoA, which is then protonated with the help of an acid residue, labelled AH (in red). A water molecule hydrolyses the thioester bond of the intermediate, allowing CoA-SH to be released and later homocitrate. Reproduced from (Bulfer *et al.*, 2009).

In silico studies

The popularity HCS allowed some freedom of choice when deciding which microorganism to employ. Consequently, a table collecting the 15 most pertinent HCS specimen (or similar) was designed. (Table 1.1)

Table 1.1. The 15 most pertinent HCS candidates or similar.

Entry name	Protein names	Gene names	Organism
HOSC_YEAST	Homocitrate synthase, cytosolic isoform	*LYS20 YDL182W, D1298*	*Saccharomyces cerevisiae*
HOSC_THET2	Homocitrate synthase	*LYS20 TT_C1550*	*Thermus thermophilus*
AKSA_METJA	Probable homocitrate synthase AksA	*aksA MJ0503*	*Methanocaldococcus jannaschii*
HOSM_YEAST	Homocitrate synthase, mitochondrial	*LYS21 YDL131W, D2195*	*Saccharomyces cerevisiae*
HOSM_SCHPO	Homocitrate synthase, mitochondrial	*lys4 SPBC1105.02c*	*Schizosaccharomyces pombe*
AKSA_METMA	Putative homocitrate synthase AksA	*aksA MM_2785*	*Methanosarcina mazei*

AKSA_METTH	Putative homocitrate synthase AksA	*aksA MTH_1630*	*Methanothermobacter thermautotrophicus*
AKSA_METAC	Putative homocitrate synthase AksA	*aksA MA_3342*	*Methanosarcina acetivorans*
AKSA_METKA	Putative homocitrate synthase AksA	*aksA MK1209*	*Methanopyrus kandleri*
NIFV_AZOVI	Homocitrate synthase	*nifV*	*Azotobacter vinelandii*
NIFVA_CLOPA	Homocitrate synthase subunit alpha	*nifV-ALPHA*	*Clostridium pasteurianum*
NIFV1_NOSS1	Homocitrate synthase 1	*nifV1 nifV, alr1407*	*Nostoc sp.*
NIFVO_CLOPA	Homocitrate synthase, omega subunit	*nifV-OMEGA*	*Clostridium pasteurianum*
NIFV2_NOSS1	Homocitrate synthase 2	*nifV2 alr2968*	*Nostoc sp.*
NIFV_RHOCA	Homocitrate synthase	*nifV*	*Rhodobacter capsulatus (Rhodopseudomonas capsulata)*
HOSM_PENRW	Homocitrate synthase, mitochondrial	*lys1 Pc22g13190*	*Penicillium rubens (Penicillium chrysogenum)*
NIFV_RHOSH	Homocitrate synthase	*nifV*	*Rhodobacter sphaeroides (Rhodopseudomonas sphaeroides)*

The final choice fell on the HCS expressed into three organisms, chosen because of their distribution in different domains: the higher fungus *Saccharomyces cerevisiae*, the archaebacterium *Thermus thermophilus*, and the nitrogen fixator bacterium *Azotobacter vinelandii*. This choice allowed us to investigate HCS homologues in the domain of Eukarya and Bacteria, and within the last domain, to confront a well-characterized form against a less characterized one. These three candidates were compared with respect to features such as substrate specificity, type of metalloenzyme, and compared on the matter of genetic and protein sequences.

HCS in Saccharomyces cerevisiae. The homocitrate synthase in *S. cerevisiae* (ScHCS) exists in two isoforms, expressed by the two nuclear paralog genes *LYS20* and *LYS21*; their origin is theorized to be the product of whole gene duplication. Being the first two genes of the lysine biosynthetic pathways, *LYS20* and *LYS21* are also involved in the inhibition and flux control of the pathway, as it often happens for amino acid synthesis. Lysine is the most known of the inhibitors of these genes, qualifying this type of inhibition as a negative feedback type (Quezada *et al.*, 2011). The structure of ScHCS has yet to be successfully crystallized. Structural studies often refer to the closest ScHCS homologue in the same organism, 2-isopropylmalate synthase (IPMS), the first enzyme involved in leucine biosynthesis. The active site of the *S. cerevisiae* HCS is thought to be composed of three key amino acids: Glu155, His309, and Tyr320, as proved by alignment of different HCS and IPMS homologues (Qian *et al.*, 2008) (Figure 1.5).

```
                                            320
Saccharomyces cerevisiae   HCS    ---------YEILDPHDFGMKRYIHFAN-RLTGWNAIKARVD
Thermus thermophilus       HCS    ---------YEPYPPEVFGVKRKLIIAS-RLTGRHAIKARAE
Pyrococcus abyssi          IPMS   --------YEHIDPELVGNRRKVVVS--ELSGRSNLIYKAK
Thermosynechococcus elongates IPMS --------YEIMDAQLIGLADNQIVLG-KLSGRNAFATRLR
Escherichia coli           IPMS   --------YEIMTPESIGLNQIQLNT-SRSGRAAVKHRMD
Methanococcus maripaludi   CMS    --------YEPIHPETVG-NRRKIILG-KHSGKAALKYKLE
Methanococcus jannaschii   CMS    --------YEPIKPEMVG-NRRRIILG-KHSGRKALKYKLD
Saccharomyces cerevisiae   IPMS   --ETQWRIPYLPLDPKDIGRDYEAVIRVNSQSGKGGAAWVIL
Mycobacterium tuberculosis IPMS   VDDMLWQVPYLPIDPRDVGRTYEAVIRVNSQSGKGGVAYIMK
                                    *       ** * *              :*
```

Figure 1.5. From Qian *et al.*, 2008. Multiple alignment of Homocitrate synthase (HCS), 2-isopropylmalate synthase (IPMS), and citramalate synthase (CMS) show Glu155, His309 and Tyr320 to be highly conserved.

Qian *et al.* (2008) confirmed these hypotheses by investigating the catalytic dyad of amino acids that seem to be the most critical for HCS functionality: Glu155 and His309. This dyad seems to be the reason for the deprotonation of the methyl group of acCoA, that leads to the formation of a hydrogen bond between E155 and H309. The newly formed bond generates a higher proton affinity on the imidazole ring of the histidine, making it an overall more efficient base – that might be the general base needed for the enzymatic catalysis. The role of the general base is still quite unclear; however, it might enhance stability of the enolization step of the condensation (Qian *et al.*, 2008).

A vital conformational difference exists between IPMS and HCS in higher fungi. Structural studies on the HCS expressed in *Schizosaccharomyces pombe* – of which the structure has been successfully crystallized - exhibited the presence of a lid motif that straddles the α-KG active site located inside of the TIM barrel. This lid motif acts as a "gatekeeper", allowing access to the active site of α-KG, however not taking any contact with the substrate; water molecules are also trapped in the catalytic pocket, providing distance between the lid and the substrate. As previously stated, the sequential catalytic mechanism of HCS makes the order of reactions crucial. (Bulfer *et al.*, 2009). As for its identity as a metalloenzyme, the protein was shown to utilize a divalent ion, Zn^{2+} or Cu^{2+}. One known notion about ScHCS concerns substrate specificity; the enzyme only employs α-KG and acCoA as substrates.

One unusual detail of ScHCS is its cellular location. First reports place ScHCS in mitochondria (Betterton *et al.*,1968; Tracy *et al.*, 1975), however, Chen *et al.* (1997) proved its nuclear localization through immunofluorescence assays. The enzyme is expressed into the cytosol, then transported into the nucleus thanks to what is theorized to be an atypical nuclear localization sequence (NLS). Scott (2010) proved that HCS is indeed needed in the fungal nucleus, revealing its supplementary role as a noncanonical histone acetytransferase (HAT). ScHCS will conditionally replace Esa1, a HAT whose role is acetylation of lysine residues on the four core histones. Mutation of this Esa1 results in sensitivity of DNA to double-stand breaks caused by camptothecin. The team proved that if *LYS20/21* is overexpressed into ESA1 mutant strains, it suppresses DNA sensitivity to damage (Scott, 2010).

HCS in Thermus thermophilus. *T. thermophilus*, a highly thermophilic Gram-negative bacterium, expresses its homocitrate synthase from the *LYS20* gene. The usage of the same gene name as ScHCS reinforces the similarities that these two genes and respective proteins exhibit.
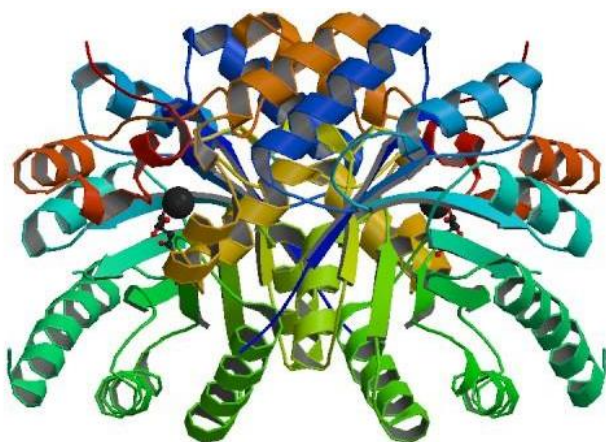
Figure 1.6. Structure of TtHCS complexed with α-KG. The black spheres represent Cu^{2+}, which was included in the crystallization buffer. Reproduced from (Okada *et al.*, 2009)

Contrary to general rule, which sees bacteria employing the diaminopimelate pathway for lysine production, *T. thermophilus* follows the α-aminoadipate pathway, along with higher fungi. The role of TtHCS is the same as ScHCS, both in being the first enzyme of the lysine biosynthesis pathway and being inhibited by overexpression of lysine. (Okada *et al.*, 2009)

TtHCS structure has been well characterized (PDB 2ZTK, 2ZTJ, 3A9I, 2ZYF). The team responsible for depositing the final crystallized structure (Okada *et al.*, 2009) describe it as a dimer formed by two symmetrical monomers (Figure 1.6). The catalytic pocket hosts a metal ion, most likely Mg^{2+} or Mn^{2+} (Kumar *et al.*, 2011). As for the active site, Glu-137, His-292 and Tyr-303 are confirmed as conserved when compared to ScHCS. Along with acCoA and α-KG as substrate, TtHCS was proved to employ oxaloacetate instead of α-KG, when in presence of KCl (Wulandari *et al.*, 2002) (Figure 1.7)
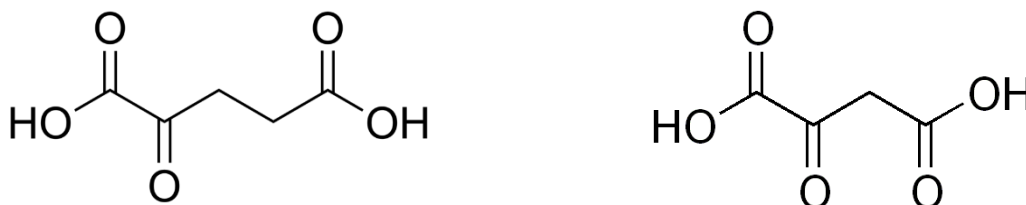
Figure 1.7. Alpha-ketoglutarate and oxaloacetate are both alpha-ketoacids, differentiated by the number of C in their chain.

Kinetic analysis of TtHCS with acCoA and oxaloacetate show K_M values and turnover values (table 1.2) that favour acCoA as the substrate.

Table 1.2. Kinetic analysis of TtHCS with different substrates. All assays were conducted in the presence of saturating amounts of acetyl-CoA and needed ions. Reproduced from (Wulandari *et al.*, 2002).

Substrate	K_m (µM)	k_{cat} (min^{-1})	k_{cat}/K_m (M^{-1} min^{-1})
Oxaloacetate (In presence of 10mM KCl)	255±60	58±4	2.3×10^5
Acetyl-CoA (In presence of 10mM KCl)	28±5	58±3	2.1×10^6
Oxaloacetate (In absence of KCl)	N.D.	N.D.	/

As for α-KG, comparison with oxaloacetate can be summarized in the $k_{cat}/K_m^{\text{2-oxoglutarate/oxaloacetate}}$ value, which determines that HCS catalyses a reaction with α-KG which is 13 times more efficient than the one with oxaloacetate (Wulandari *et al.*, 2002). Conclusively, α-KG and acCoA are once again the preferred
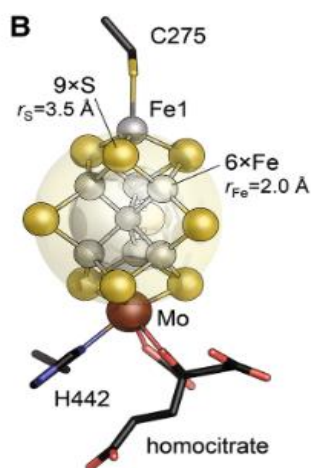
Figure1.8. Structure of the FeMo cofactor, considered to be "the largest single iron-sulfur cluster known to bioinorganic chemistry". The structure counts seven iron atoms, nine sulphur atoms, one molybdenum atom and one carbon atom in the very centre of the cofactor. Original graphics and quote from (Einsle, 2014).

substrates for HCS.

HCS in Azotobacter vinelandii. *A. vinelandii* is a Gram-negative bacterium known for the aerobic fixation of nitrogen. For this matter, AvHCS has small coverage as a homocitrate synthase, and more hits as part of the nitrogen fixation pathway. In fact, *NifV*, the gene expressing AvHCS, is contained in the gene cluster *Nif*, whose goal is the biogenesis of a FeMo cofactor for the bacterium three nitrogenases (Kennedy *et al.*, 1991) (Figure 1.8). The role of AvHCS in the pathway is the production of homocitrate, a vital component for a FeMo cofactor utilized by FeMo nitrogenases. (Madden *et al.*, 1991; Mayer *et al.*, 2002). AvHCS has not been crystallized for structure determination; the enzyme was theorized to be a homodimer-type metalloenzyme (Zheng *et al.*, 1997). Aside from the canonical employment of acCoA and α-KG as substrate, AvHCS was also tested for the employment of oxaloacetate and alpha-ketoadipate (α-KA) to replace α-KG (Zheng *et al.*, 1997) (Figure 1.9).
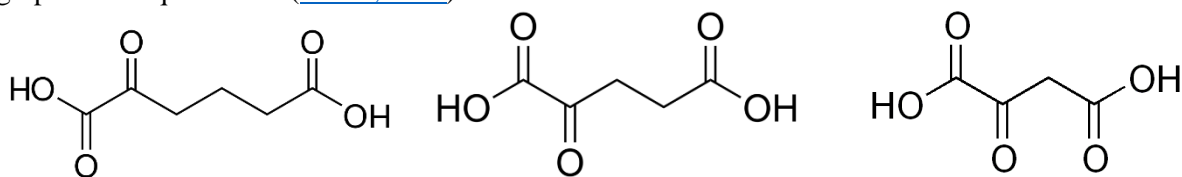

Figure 1.9. Alpha-ketoadipate (C6), alpha-ketoglutarate (C5) and oxaloacetate (C4) are all alpha-ketoacids, differentiated by the number of carbon atoms in their chain.

Kinetic analysis conducted with the same concentration of acCoA (the stable second substrate) revealed the apparent K_M to favour employment of alpha-KA, since its K_M value is the lowest recorded. α-KG only comes as second, while oxaloacetate is not a far third. However, since K_{cat} was not defined, no assurance can be determined by these values. Other substrates were tested, such as propionyl-CoA, pyruvate, or alpha-ketopimelate, finding no activity for any of them (Table 1.3) (Zheng *et al.*, 1997).

Table 1.3. Kinetic parameters of AvHCS with different substrates. These parameters for acCoA were determined using 10 mM of α-KG as the co-substrate. Parameters for all the other substrates were determined with 0.2 mM acCoA as second substrate.

Substrate	K_M (mM)	V_{max} (nmol/min/mg)
Acetyl-Coa	0.06	730
Oxaloacetate	2.83	350
Alpha-ketoglutarate	2.24	750
Alpha-ketoadipate	1.24	380

Summary of HCS candidate features. Table 1.4 and 1.5 summarize some of the most important features of the HCS candidates.

Table 1.4. Most notable features of candidate enzymes for this study.

	Metal ion	Substrates utilized	Structure type and mass (Da)	Protein sequence length (residues)
ScHCS (*LYS20*)	Zn^{2+} or Cu^{2+}	AcetylCoa; α-ketoglutarate	Homodimer 47,099 Da	428
ScHCS (*LYS21*)	Zn^{2+} or Cu^{2+}	AcetylCoa; α-ketoglutarate	Homodimer 48,594 Da	440
TtHCS	Mg^{2+} or Mn^{2+}	AcetylCoa; α-ketoglutarate or oxaloacetate	Homodimer 42,159 Da	376
AvHCS	Not reported	AcetylCoa; α-ketoglutarate, oxaloacetate or α-ketoadipate	Homodimer 41,653 Da	385

Table 1.5. Kinetic parameters listed by Bulfer *et al.* (2009), Wulandari *et al.* (2002) and Zheng *et al.* (1997)

	ScHCS (*LYS20*)	TtHCS	AvHCS
K_M acetylCoA (mM)	0.042	0.032	0.060
K_M α-ketoglutarate (mM)	0.14	0.05	2.2
K_M α-oxaloacetate (mM)	X	0.25	2.8
K_M α-ketoadipate (mM)	X	X	1.24
K_{cat} (min^{-1}) (AcCoA+α-KG)	63	92	Not reported

Part 2 - Materials and methods

Strains, media and buffers. Strains, media and buffers employed are listed in table 2.1.

Table 2.1. List of strains, media and buffers used for this study.

Strains	*S. cerevisiae* CEN.PK2 *; E. coli* BL21 (DE3)
Media	YPD (yeast-peptone-dextrose): 10 g/L yeast extract, 20 g/L peptone, 20 g/L dextrose YPD-agar: 10 g/L yeast extract, 20 g/L peptone, 20 g/L dextrose, 20 g/L agar LB (lysogeny broth): 10 g/L tryptone, 10 g/L NaCl, 5 g/L yeast extract LB-agar: 10 g/L tryptone, 10 g/L NaCl, 5 g/L yeast extract, 20 g/L agar 2 x YT: 16 g/L tryptone, 10 g/L yeast extract, 5 g/L NaCl, adjust to pH 7.5
Buffers	Binding buffer: 10 mM Tris-HCl pH 7.5, 0.1% triton x-100, 0.5M NaCl, 10mM imidazole Elution buffer: 10 mM Tris-HCl pH 7.5, 0.1% triton x-100, 0.5M NaCl, 500mM imidazole

Unless stated, methods for production of ScHCS, TtHCS and AvHCS were identical. Genes fragments *LYS20* for TtHCS and *NifV* for AvHCS were synthetically produced by Twist Biosciences (USA), and subsequently inserted into pET28a (+) vector. Agarose gel (1,2%) electrophoresis was run after each PCR, enzymatic digestion or ligation step.

Cloning process. For ScHCS, strain CEN.PK2 of *S. cerevisiae* was first grown on a YPD-Agar plate; colonies were picked and inoculated into YPD liquid media. For the amplification of target genes ScHCS *LYS20* and *LYS21*, a PCR round was set up. Primers were designed *in silico* (Table 2.2), inserting two flanking restriction site sequences to allow enzymatic digestion (BamHI and XhoI), in accord to the restriction sites present on the receiving vector (Invitrogen by Thermo Fisher Scientific). PCR mixes were prepared according to table 2.3, with chemicals purchased from Thermo Scientific; thermocycler setup is shown in table 2.4 (Core life sciences).

Table 2.2. Primer design. The bold nucleotides in the sequence refer to the location of the restriction site sequence.

Primer name	ORF name	ORF length (bp)	Restriction site type	Sequence
BamHI-LYS20-F	*LYS20*	1272	BamHI	5´ **GGATCC**GATGACTGCTGCTAAACCA 3´
XhoI-LYS20-R	*LYS20*	1272	XhoI	5´ **GAGCTC**GGCGGATGGCTTAGTC 3´
BamHI-LYS21-F	*LYS21*	1323	BamHI	5´ **GGATCC**GATGTCTGAAAATAACGAATTC 3´
XhoI-LYS21-R	*LYS21*	1323	XhoI	5´ **GAGCTC**TTTGGTGACCTTTGCCTTTT 3´

Table 2.3. Composition of reaction mixes for PCR on *S. cerevisiae* genomic template.

Chemical	LYS20 samples	LYS21 samples
10X Taq Buffer	5 µL	5 µL
10 mM dNTPs (10X)	1 µL	1 µL
10 µM F primer	1 µL	1 µL
10 µM R primer	1 µL	1 µL
DNA template	1 µL	1 µL
Taq Polymerase	0,5 µL	0,5 µL
Nuclease free water	40,5 µL	40,5 µL
Total	50µL	50µL

Table 2.4. Thermocycler setup for PCR on *S. cerevisiae* genomic template.

	Temperature	Time
Initial denaturation	95°C	1 min
Cycles (X40)	95°C	30 sec
	63°C	30 sec
	72°C	1 min 19 sec
Final extension	72°C	5 min
Hold	4°C	∞

PCR products were purified using GeneJet PCR purification kit (Thermo Scientific). Expression vector for ScHCS genes was chosen to be pET28b(+); insert placement was in-between restriction sites BamHI and XhoI of the vector multiple cloning site, upstream of the 6xHisTag. The C-terminal 6xHisTag was used as purification tag.

Enzymatic double digestion of insert and vector was performed by employing the restriction enzymes BamHI and XhoI (ThermoScientific), following manufacturer instructions. Purification of digested inserts and vector was performed in two different ways: GeneJet PCR purification kit (Thermo Scientific) for the insert and QIAquick Gel Extraction Kit (Quiagen) for the vector. As for the ligation process, the insert:vector ratio was kept at 3:1. Negative control was set up as all other samples, however in absence of insert DNA. Ligation setup followed manufacture instruction (Thermo Scientific); incubation was conducted at 4°C O/N, and then at 16°C for two hours.

Transformation of *E.coli* BL21(DE3) competent cells followed heat-shock procedures of 40 seconds in a water bath at 42°C. LB medium at room temperature was used to restore cells, then plated on LB-Agar-Kanamycin plates (LB-agar-kan). Picked colonies were pre-inoculated in LB medium + 50 mg/ml kanamycin O/N, then centrifuged at 4500 rpm (Sorvall LYNX 400, rotor F21-8x50). Cell pellets were used for purification of final constructed plasmid with GeneJet Plasmid purification kit (Thermo Scientific). These extracted plasmids were used as templates for detection of the proper insert, which used two sets of primers. Internal primers for amplification of a section completely included in the target gene; external primers for amplified of part of the target gene and part of the plasmid backbone (Table 2.4).

Table 2.4. Internal and external primers used for target insert detection in transformed plasmids.

Primer name	Primer length (bp)	PCR product length	Sequence
Int_F	20	680 bp	5´ GTCGCCGTAGAGACTGGTGT 3´
Int_R	20		5´ TCGTGAGGGTCCAAGATTTC 3´
Ext_F	20	762 bp	5´ GAAATCTTGGACCCTCACGA 3´
Ext_R	20		5´ AAGGGCGAAAAACCGTCTAT 3´

***Expression and purification of target protein**.* Expression was performed as follows. For ScHCS, colonies were picked and pre-inoculated in LB medium containing 50 mg/ml kanamycin and grown O/N. One millilitre of the pre-inoculated samples was moved into flasks containing fresh LB-kan medium (kanamycin 50 mg/ml). For TtHCS and AvHCS, the method employed was the same, however, 2xYT medium was used instead of LB medium. After ~3 hours of growth, cultures reached OD 0.6, and were induced with IPTG (Sigma Aldrich) to a final concentration of 1 mM for ScHCS and 0.5mM for TtHCS and AvHCS. The flasks were put back into the incubator at different settings: for ScHCS and TtHCS, 25°C, 200 rpm, O/N; for AvHCS, 12°C, 200 rpm, 24 hours. Competent cells with plasmid including no insert were also induced, for negative control checks.

Cells were harvested by centrifuging samples at 6000 rpm (Sorvall RC-5C, rotor F14-6x250) for 10´ two times, then resuspending the resulting pellet with deionized water each time. For ScHCS, BugBuster® Protein Extraction Reagent (Millipore) x1 was added to each sample, along with Lysonase™ Bioprocessing Reagent (Millipore), to resuspend the pellet. The falcons were incubated with agitation at room temperature for 30 ´. For TtHCS and AvTHCS, pellets were resuspended with a resuspension buffer fit for purification (10 mM Tris-HCS pH 7.5, 0.1% triton x-100, 0.5M NaCl, 5mM imidazole). These samples were frozen and sonicated with 10 seconds on and 10 seconds off as cycle, for a total of 15 cycles (cycle 0.6, amplitude 60-80%). Afterwards, samples were added with BugBuster® and kept in agitation at room temperature for 20´. All three type of samples were centrifuged at 130000 rpm (Sorvall LYNX 400, rotor F21-8x50), 20´; the supernatant – containing the total extracted proteins - was labelled as soluble fraction, while the remaining pellet was labelled insoluble fraction.

Soluble fractions of total protein extract were applied on a HisTrap HP 1ml column (GE Healthcare) containing nickel-NTA resin for affinity chromatography. These columns were stripped and recharged of nickel ions with an $NiSO_4$ solution (according to manufacturer instruction) in-between purification rounds to assure each type of HCS could be purified with no interferences. Äkta Pure Protein Purification system (GE Healthcare) and UNICORN Pure Protein Purification software (GE Healthcare) were used for the purification process, following the manufacturer instructions and protocols. The choice of buffers comprised of a binding buffer (10 mM Tris-HCl pH 7.5, 0.1% triton x-100, 0.5M NaCl, 10mM imidazole) and an elution buffer (10 mM Tris-HCl pH 7.5, 0.1% triton x-100, 0.5M NaCl, 500mM imidazole). The purification rounds used a gradient-type elution (0-100%) to assure thorough elution. Subsequent to purification, an SDS-PAGE allowed to check for appropriate protein mass. Higher concentrated eluted fractions were acquired and pooled together, following the absorbance peaks in the resulting chromatograms. Protein quantification was performed with the Bicinchoninic Acid (BCA) protein assay (G Biosciences); manufacturer protocols were followed. Sephadex G-25 in PD-10 desalting columns (GE Healthcare) were used for the desalting process against distilled water.

***DCPIP activity assay**.* In order to detect HCS activity, redox dye 2,6-dichlorophenolindophenol (DCPIP) was employed. The dye presents itself with a dark blue tone, which turns gradually colourless once it is reduced by the CoA-SH produced by the enzymatic reaction (Figure 2.1).
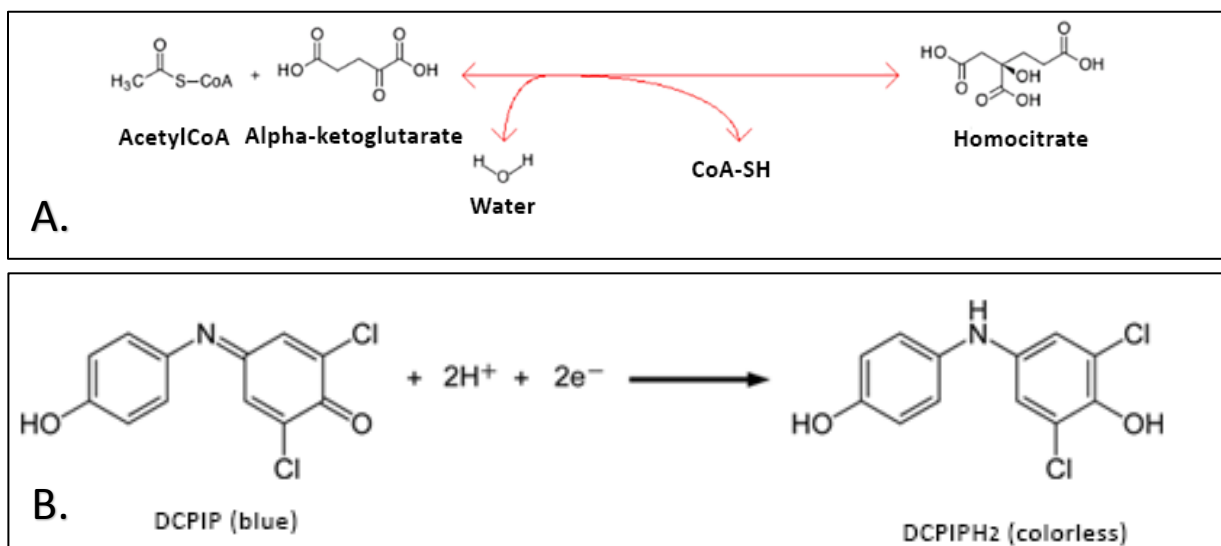
Figure 2.1. A. Enzymatic reaction of homocitrate synthase. B. Reduction reaction of DCPIP, which results in its loss of colour.

Optical density was monitored at 600nm in the UV-1650PC spectrophotometer (Shimadzu) for a total of 300 seconds, while temperature was kept stable at 30°C by a connected water bath (Julabo, by Skafte medlab). The reaction mix comprised of 0.5mM acCoA (Roche), 20mM alpha-ketoglutarate (Sigma-Aldrich), 50mM HEPES (Sigma-Aldrich) pH 7.5, 0.1mM DCPIP for a final volume of 1ml. This mixture was used as a cell blank for the spectrophotometer. Once 30-90ng of purified protein was added to the reaction mixture, the optical density was monitored and recorded.

Protein modelling. Preliminary studies for protein modelling concerned multiple alignments of protein sequences. BLAST (Basic Local Alignment Search Tool; Madden, 2002) was employed for alignments; T-Coffee (Notredame *et al.*, 2000) and BOXSHADE version 3.21 for alignment visualization. Protein models were produced for AvHCS and ScHCS from their two paralogs. All models were built with the help of YASARA (Yet Another Scientific Artificial Reality Application) (Krieger, 2003). Models were built if at least one deposited PDB structure had an identity score above 30% with the target sequence. Overall quality of the structures was tested through ERRAT, PROCHECK and ProSAWeb. ERRAT determines the quality of a protein structure by plotting error values based on atom-atom interaction statistics, which are compared to high-resolution structures in database (Colovos and Yeates, 1993). As for the stereochemical quality of the protein, PROCHECK was employed to obtain Ramachandan plots of each template, homology model and refined structure (Laskowski *et al.*, 1993). Finally, ProSAWeb checked the model quality by comparing it to the other approved structures in the Protein Data Bank, allowing to determine how close the model is to actual deposited structures (Wiederstein and Sippl, 2007). For final pictures, UCSF Chimera was employed (Pettersen *et al.*, 2004).

Part 3 – Results

Cloning process. Target genes *LYS20* and *LYS21* of *S. cerevisiae* were first amplified with the use of primers containing restriction sites on the 5′ ends. PCR products resulting from the amplification of genes *LYS20* (1272 bp) and *LYS21* (1323 bp) are shown in figure 3.1, both before and after purification. *S. cerevisiae* genomic DNA templates used for the amplification were extracted from two different colonies, A and B, however amplification from template A was used for the following experiments.
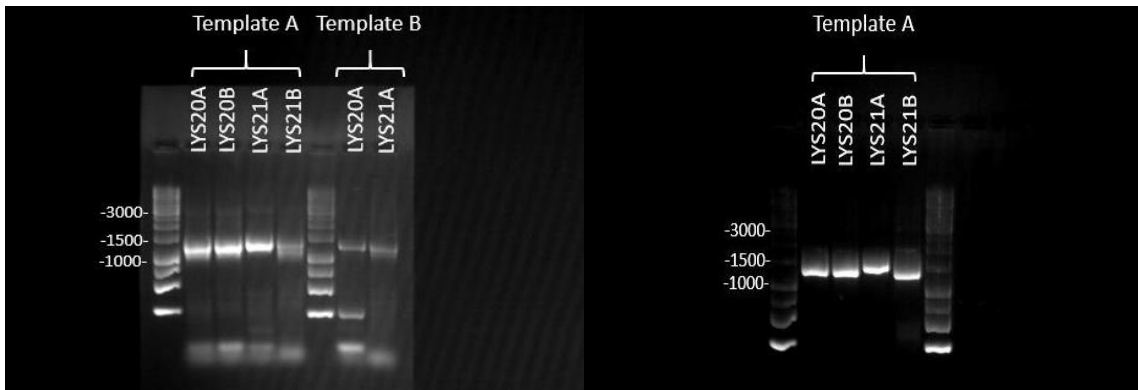
Figure 3.1. Gel electrophoresis of PCR products *LYS20* and *LYS21*. Left, PCR products *LYS20* and *LYS21* from two different *S. cerevisiae* genomic templates (A and B). Right, purified PCR products.

Following the purification of the PCR products, the enzymatic double digestion of both PCR products (*LYS20, LYS21*) and vector (pET28b(+)) was performed with the chosen restriction enzymes, BamHI and XhoI, as shown in Figure 3.2. Different incubation intervals for the enzymatic digestion of the pET28b(+) vector (5368 bp) were tested, in which the optimized 15 minutes incubation proved to be successful. As for the inserts, the optimal incubation time of 15 minutes was also applied to this assay, showing a defined band. This step provided both vector and insert with compatible sticky ends, allowing the construction of a plasmid carrying the target genes.
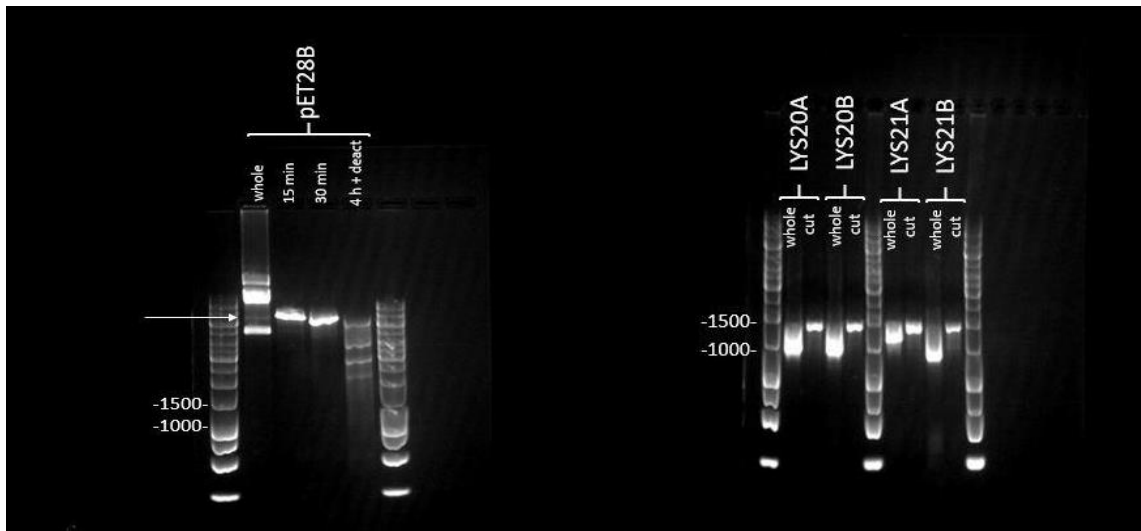
Figure 3.2. Gel electrophoresis of double digested vector and inserts. On the left, different incubation times for the double enzymatic digestion of the vector pET28b (+), in which 15 minutes seemed to be the most optimal. The white arrow displays the linearized plasmid in the undigested sample of the vector, since it is also a natural-occurring event. On the right is displayed the double digested PCR products. The size difference between digested and not digested PCR products is due to the state of coiling the plasmid. Intact plasmids (labelled "whole") are usually supercoiled, therefore migrate faster during the electrophoresis run. Linearized plasmids (labelled "cut") migrate slower, due to their size.

Linking of vector and insert was performed with the ligation step. The outcome of the assay is exhibited in Figure 3.3; some faint bands appear just above the vector backbones in the lanes where the ligated construct was expected. There is no similar band in the negative control lane.
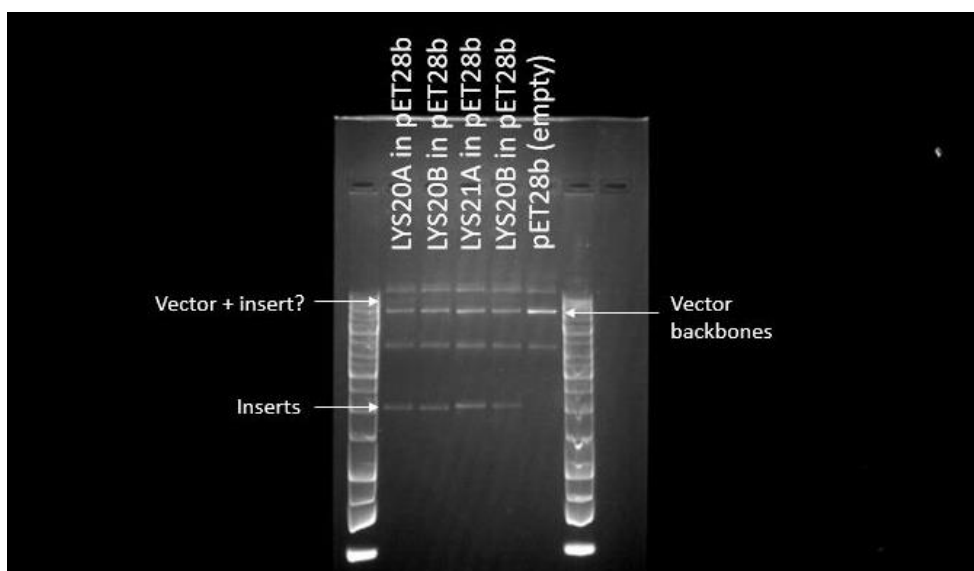
Figure 3.3. The arrow labelled "insert" shows the appropriate ORFs sizes (around 1300 bp). The arrow labelled "vector + insert" refers to the first four lanes, in which the construct formed by the vector and the insert should appear.

Competent cells *E. coli* BL21 (DE3) were transformed with the constructed plasmids carrying genes for the expression of ScHCS (both *LYS20* and *LYS21*), TtHCS (*LYS20*) and AvHCS (*NifV*). Plasmids for ScHCS expression were constructed during the previous steps, as stated; plasmids for the expression of TtHCS and AvHCS were purchased as carrying the target genes. After transformation of the competent cells, transformed colonies were grown into liquid media, induced to express the target protein, then harvested, lysed and finally purified using affinity chromatography.

  ***Expression and purification of target protein*.** Unfortunately, the expression and ScHCS from the two paralogs *LYS20* and *LYS21* proved to be unsuccessful, as the resulting chromatograms showed no elution peaks, as well as the SDS-PAGE exhibited no target protein bands. AvHCS was expressed, but not purified: an appropriate band size for AvHCS was found in the SDS-PAGE, however the enzyme was trapped in the insoluble fraction (Figure 5.1, supplementary data). TtHCS, however, was expressed and purified successfully. From the plate containing the competent cells transformed with the plasmid carrying the TtHCS gene, three different colonies were chosen and underwent the purification step. The choice of different colonies from which to isolate TtHCS allowed to determine if the resulting target proteins performed consistently throughout the following experiments. Chromatograms from the purification rounds of TtHCS show a distinctive peak during the elution portion of the process, when the gradient hits the 40-60% mark of elution buffer composition (~200-300mM imidazole) (Figure 3.4). Total protein extract of the negative control sample was also applied to the purification column, giving no major elution peak. Superimposed chromatogram of the three TtHCS purified samples are exhibited in figure 3.4. An SDS-PAGE run further confirmed the mass and purification of the target protein, whose size is around 43 kDa including the HisTag (Figure 3.5).
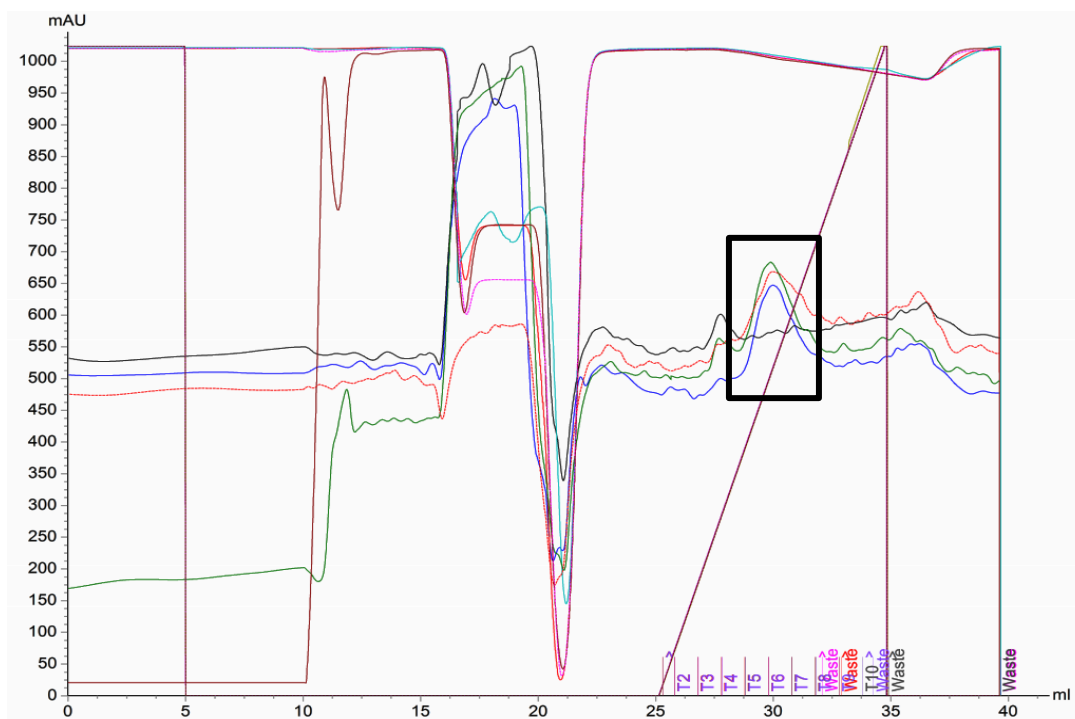
Figure 3.4. Superimposed chromatograms for four samples. UV signals marking optical density are in blue, green, and red for the soluble fraction of cell lysate of TtHCS expressing colonies; the UV signal in black is from the soluble fraction of cell lysate of non-transformed colony (negative control). Highlighted in black, the elution for peaks for all samples, confirming the presence of the same type of enzyme.
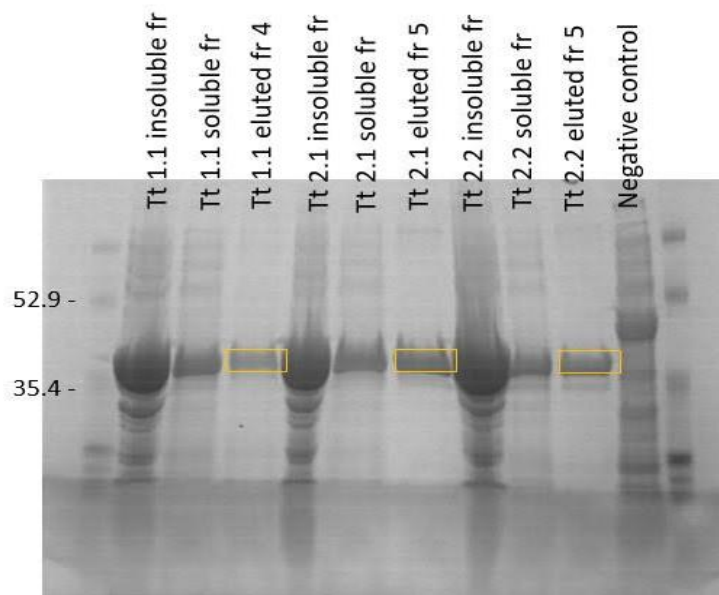

Figure 3.5. SDS-PAGE of samples from purification of TtHCS from *Escherichia coli* BL21 (DE3), confirming the success of the purification process. Insoluble and soluble fractions were produced during cell lysis. Yellow shapes highlight the purified protein band. Negative control lane was occupied by its soluble fraction.

Eluted fractions T4 to T7 of each TtHCS clone were pooled together and checked for concentration. Protein concentration for the three samples were defined at 438,125 ug/ml, 374.375 ug/ml and 268.125 ug/ml. (Figure 3.6)
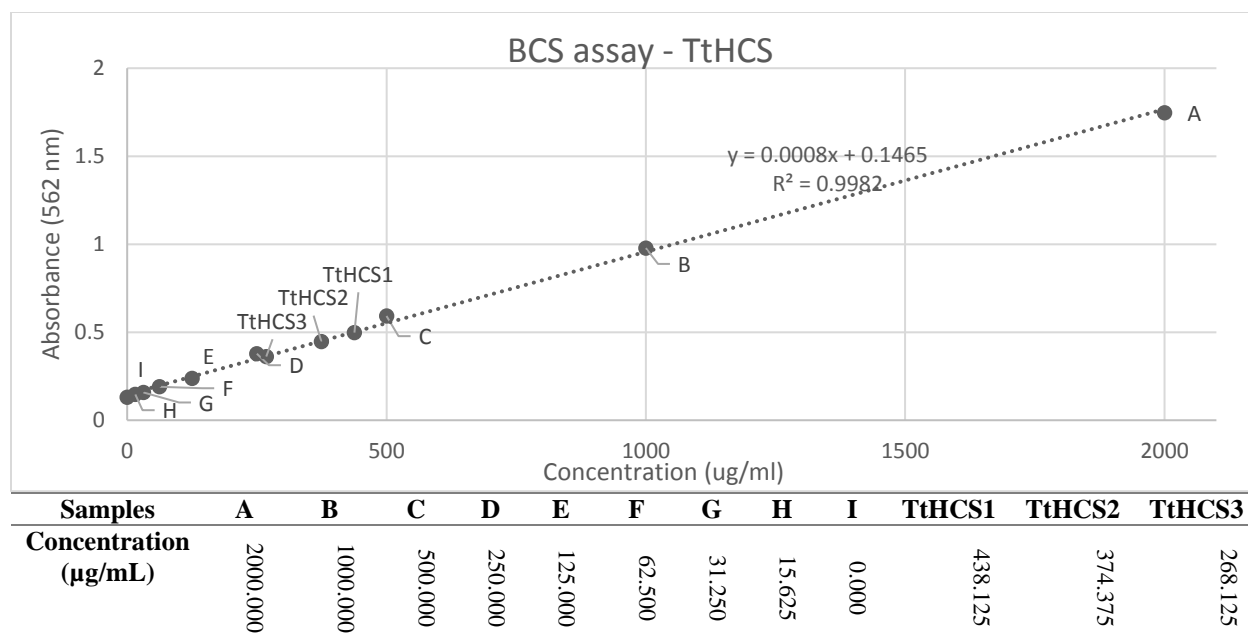
Figure 3.6. BCA assay graph, determining concentration for the three TtHCS purified samples, along with the concentration values for all the plot points.

DCPIP enzyme assay. The DCPIP enzymatic assay was performed with the intent of testing if TtHCS samples purified from three different transformed colonies behaved consistently and had retained their catalytic activity; therefore, it must be considered a qualitative assay. TtHCS activity was confirmed through the reducing OD signal (figure 3.7). Protein amount used for the assay was purposefully variated: 87,6 ng TtHCS 1, 27 ng TtHCS 2 and 37,5 ng TtHCS 3. As for the desalted sample (87,6 ng), it appeared to have lost all activity, possibly due to the lack of optimal condition for protein storage (lack of salts and buffer solution). The difference between cell blank and samples starting point at t = 0 is attributed to the interval occurring between protein placement into the cuvette and manual start of the monitoring.
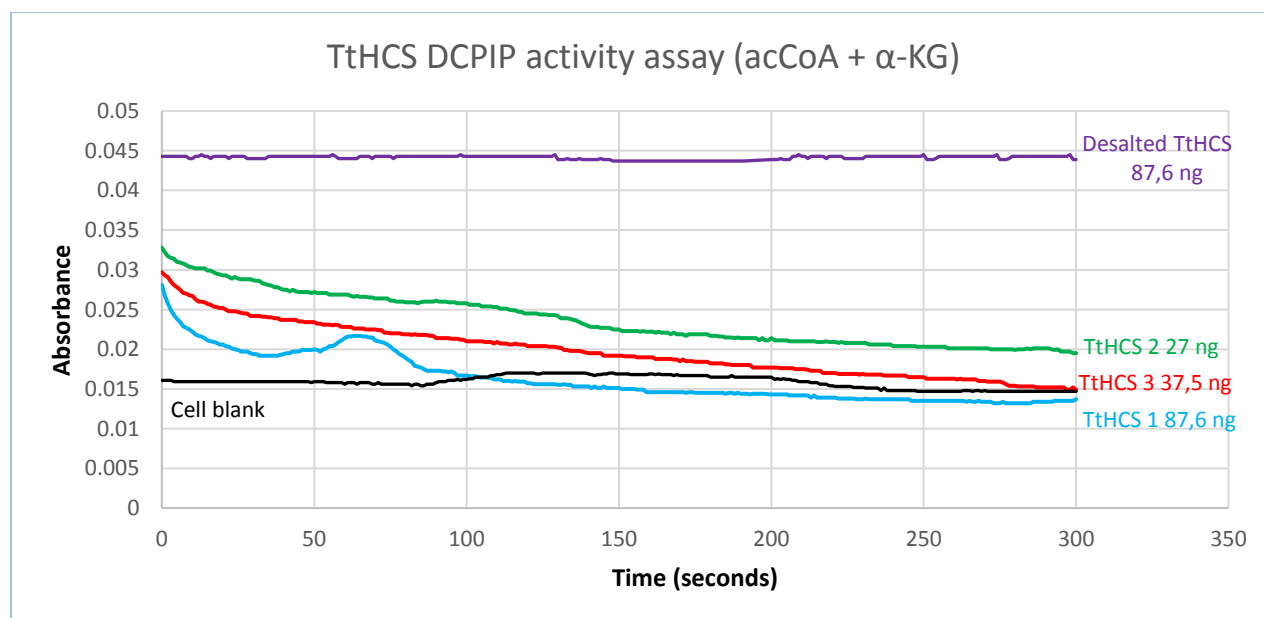
Figure 3.7. DCPIP activity assay for TtHCS purified samples with acCoA and α-KG as substrates. OD signals for the three TtHCS purified samples are highlighted in green, red and clear blue, which gave

confirmation of appropriate protein activity by OD reduction. Cell blank is in black, and TtHCS desalted sample in purple; both these samples showed no major change in OD signals overtime.

Since the activity assay follows the decrease of absorbance values for DCPIP, its decrease in concentration was calculated. The extinction coefficient for DCPIP used was determined by Kumar *et al.* ([2011](#)), DCPIP $\varepsilon_{600} = 19,100M^{-1}cm^{-1}$ (Figure 3.8).
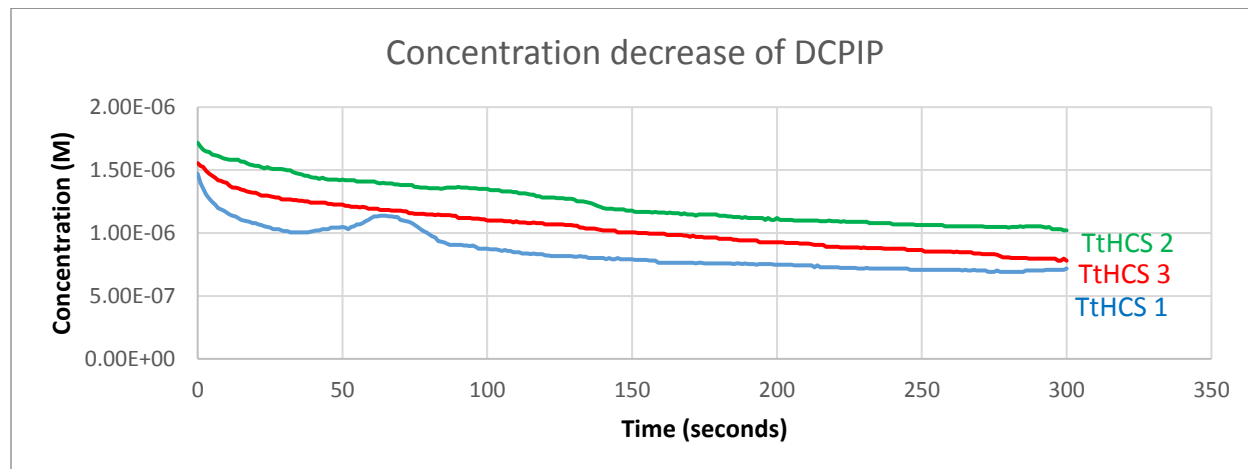

Figure 3.8. DCPIP concentration decrease due to its reduction, caused by CoA-SH. Plot lines in green, red and clear blue relate to the three different monitoring rounds of the activity test. Total amount of reduced DCPIP is summarized in table 3.3.

By plotting 1/DCPIP concentration decrease values against time, the reaction can be qualified as a second order reaction, as it results in a linear, positively sloped plot (Figure 3.9). The type of plot, 1/DCPIP concentration decrease values against time, is used to distinguish different order reactions, according to their integrated rate laws. Second order reactions result in linear plots in 1/[reactant] versus time graphs, first order reactions show linear plots in ln[reactant] versus time graphs and zero order reactions exhibit linear plots in [reactant] versus time graphs. In this case, the highest linear regression value (R^2) was recorded in the 1/[reactant] versus time graph, determining the plots as close to a straight line as possible, therefore determining the reaction as belonging to a second order reaction.

The order of reaction determines the dependence of the reaction rate on the concentration of a substrate. Second order reactions proceed at rates that depend on the concentration of two reactants: for this assay, the two reactants are DCPIP and the CoA-SH resulting from the reaction catalyzed by HCS. This notion can determine k, the reaction rate constant, by finding the slope of the trendline of the 1/concentration plot, as useful for rate calculations. Larger k values determine faster reactions (Figure 3.10).

Initial reaction rates were calculated by determining the slope of linear portion of the DCPIP concentration decrease against time plot, around the t=0 mark (Figure 3.8). The linear portion of the plot was determined when the value of its linear regression (R^2) was above 0.980, around the t=0 mark.

The initial reaction rate represents the instantaneous rate at the beginning of the reaction, during which the reaction proceeds at constant velocity. However, it must be noted that initial rates are not accurate values due to the interval between placement of the enzyme and start of the manual monitoring. Enzyme assay calculations are summarized in table 3.3.
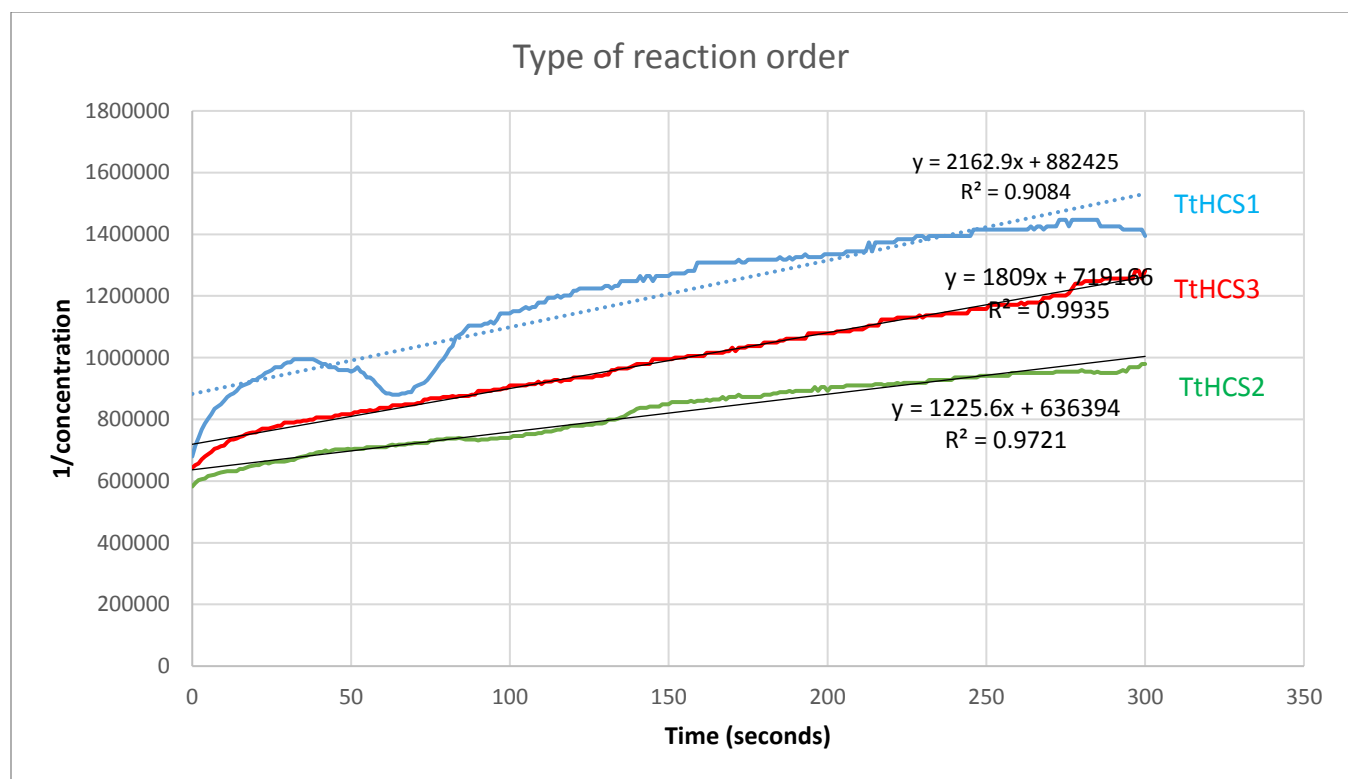
Figure 3.9. 1/DCPIP concentration decrease plot against time for each reaction. The plots seem to be the mostly linear and positively sloped, therefore it is a second order reaction.

Table 3.3. Enzyme assay calculations summarized.

Sample	Amount of HCS (ng)	Total DCPIP reduced (nM)	k, reaction rate constant (M^{-1} sec^{-1})	Initial reaction rate (M/sec)
TtHCS 1	87,6	0.754	2162.9	$5.50*10^{-8}$
TtHCS 2	27	0.696	1225.6	$1.86*10^{-8}$
TtHCS 3	37,5	0.775	1809.0	$2.88*10^{-8}$
Av±SD		0.742±0.04	1732.5±473.3 $*10^{-9}$	3.41±1.88 $*10^{-8}$

Protein modelling. AvHCS. The choice of the template landed on 6E1J, a methylthioalkymalate synthase from *Brassica juncea* (PDB, 6E1J), as it scored the highest identity values when compared to the protein sequence of AvHCS, above the needed threshold of 30% identity score. The alignment scored 38% identities and 53% similarities when aligned with the target sequence, the highest in the PDB database (Figure 3.10).

```
AvHCS_protein_s     1  MASV--------------------------------------------------------
Template_chosen     1  MASSLLTSSGIIPTTGSNMVVRSFLPFGSVRLTRPYNNSSLLISCCSSVSKNAETSGTDL

AvHCS_protein_s     5  ---------------------ILDDTTLRDGEQSAGVAFNADEKTAIARALAELGVPELE
Template_chosen    61  KTIVERWPEYIPNKLPDKNYVRVFDTTLRDGEQSPGAALTPPQKIEIARQLAKLRVDIME

AvHCS_protein_s    44  IGIPSMGEEEREVVHAIAGL----------GLSSRILAWCRLCDVDIAAARST----GVTM
Template_chosen   121  VGFPVSSEEEFETIQTIAKTVGNEVDEETGYIPVICVIARSKERDIKAAWESVKYAKRPR

AvHCS_protein_s    91  VDISLPVSDIMLHHKINRDRIWAIREVARLVGEARMAGLE-VCLGCEDASRADLEFVVQV
Template_chosen   181  IVIFTSTSDIHKYKLKMTREEVVDMVASSIRFAKSLGFEDIEFGCEDGGRSDKDYICTV

AvHCS_protein_s   150  GEVAQAAGARRLRFADTVGVMEPF--GMLDRERFLS-R-RIDMELEVHAHDDFGLATANT
Template_chosen   241  FEEAIKAGATTLACPDTVGINMPHEYGKLVRYIKANTPGIDDVIFSAHCHNDLGVATANT

AvHCS_protein_s   206  LAAVMGGATHINTTVNGLGERAGNAALEECVIALKN-----LHGIDTGIDTRGIPAISAI
Template_chosen   301  IAGICAGARQVEVTINGIGERSGNAPLEEVVMALKCRGAFVMGGVYTRIDTRQIMATSKV

AvHCS_protein_s   261  VERASGRQVAWQKSVVGAGVFTHEAGIHVDGLLKHRRNYEGINPDELGRS----HSLVLG
Template_chosen   361  VQEYTGLYVQPHKPIVGANCFVHESGIHQDGILKNRSTYEIISPEDVGVVKSQNSGIVLG

AvHCS_protein_s   317  KHSGAHMVRNTYRDLGIEIADWQSCALLGRIRAFSTRTKRRSPQPAELQDFYRQICE---
Template_chosen   421  KLSGRHAVKGRLKELGYEISDEKLNEVFSRFRDLIKQKKRVID------DDLKAIVTCGD

AvHCS_protein_s   374  ----------QGNPEIAAGGMA----------
Template_chosen   475  EIFSSDKLNGTDDNEINSNGYVPAPQISSVV
```

Figure 3.10. Alignment of target sequence for AvHCS with chosen template 6EIJ, a methylthioalkymalate synthase from *Brassica juncea.*
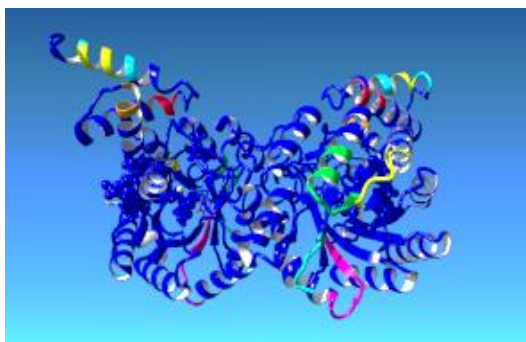

Figure 3.11. Model produced for AvHCS using YASARA. The different coloration represents different PDB portion of proteins used to build the final model.

Once the template was chosen, 19 different iterations of homology analysis were combined, in order to enhance accuracy beyond the contributions of the single homology models. In figure 3.11, the blue colour of the protein represents the structure of the chosen template, while the other different colours represent other protein portions that were composed together to produce the final hybrid structure of AvHCS. AvHCS final model presents itself as a homodimer, with a Mn^{2+} ion in its catalytic pocket; there is no literature confirming or denying the identity of this ion. Z-score was determined to be -0.627. Unfortunately, the final hybrid structure had two ligands attached: acCoA and ketomethiobutirate (KMT). Both these ligands had to be removed in order to conduct further refinements. First refinement saw a slight decrease into the overall quality of the model, as expected after removal of ligands. Second refinement resulted in a slightly increased model quality. Quality checks for the template, the first hybrid model and the refinements are summarized in table 3.4 and figure 3.12. Figure 3.13 shows the finalized protein structure. The finalized structure of AvHCS produced in this study resulted satisfactory even if being an *in-silico* simulation of the structure of the enzyme. In fact, the simulated structure passed all quality checks, being above the needed thresholds established by ERRAT (>95%) and PROCHECK (>90%). For the quality check performed by ProSAWEB, the AvHCS model appeared to be in range with similar deposited structures in the Protein Data Bank.

Table 3.4. ERRAT, PROCHECK, Ramachandran plot and ProSAWeb results for construction of AvHCS model.

|	Template 6E1J	Hybrid Model AvHCS	First refinement AvHCS	Second refinement AvHCS
ERRAT	(must be >95%)			
Dimer1	97,51 %	97,04 %	96,22 %	96,76 %
Dimer2	97,75 %	96,50 %	95,03 %	95,15 %
PROCHECK	(must be >90%)			
Residues in most favoured regions	91,6 %	92,3 %	92,2 %	93,2 %
Residues in additional allowed regions	7,5 %	6,6 %	6,8 %	5,7 %
Residues in generously allowed regions	0,1 %	0,6 %	0,8 %	0,8 %
Residues in disallowed regions	0,7 %	0,5 %	0,3 %	0,3 %
Total	100%	100%	100%	100%
ProSAWeb Z-score (In range)	-10,19	-7,86	-7,69	-7,37
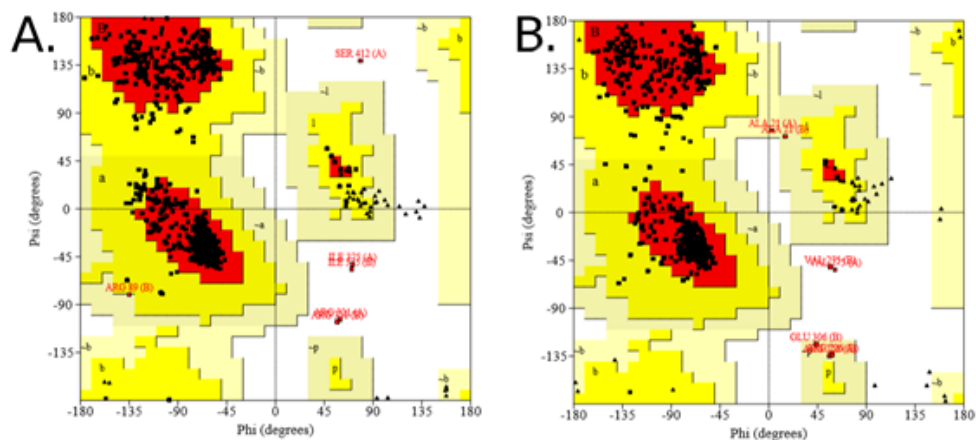
Figure 3.12. A. Ramachandran plot for chosen template 6E1J. B. Ramachandran plot for hybrid model of AvHCS after first and second refinement.
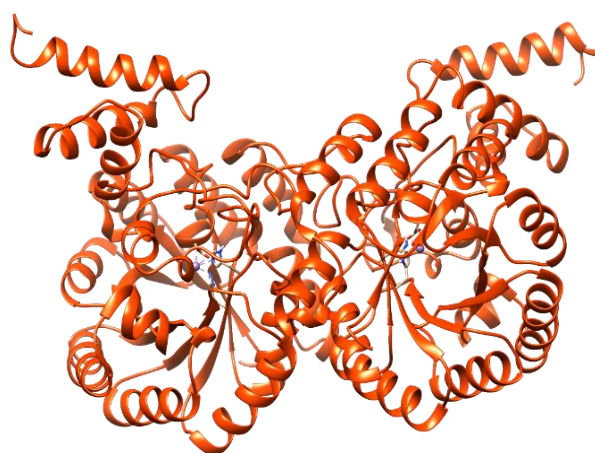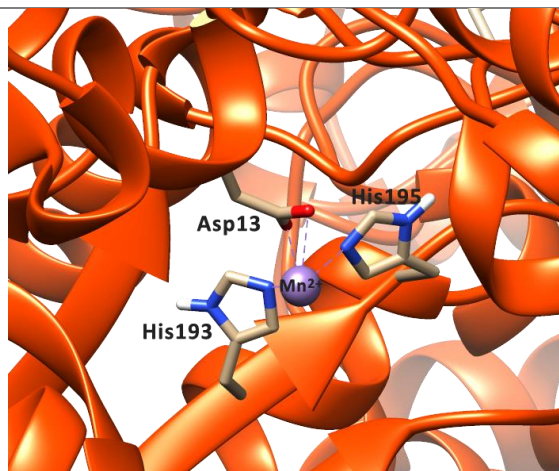
A. Final protein model of AvHCS   B. Active site of the final protein model of AvHCS
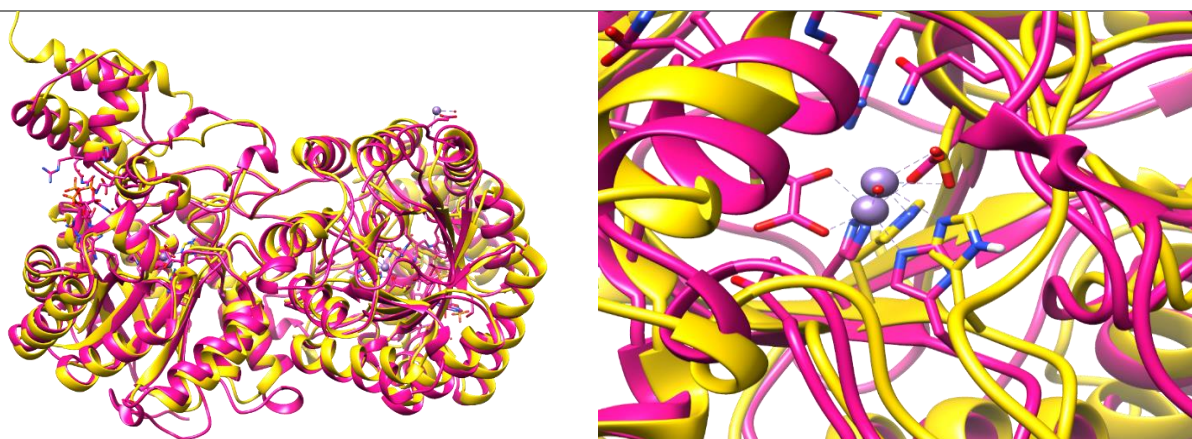
| C. Final protein model of AvHCS superimposed on chosen template 6E1J | D. Active site of final protein model of AvHCS superimposed on the active site of chosen template 3IVT |

Figure 3.13. AvHCS protein model produced in this study. In all images, the lilac sphere represents the Mn^{2+} ion. A. In orange, the final structure of AvHCS after refinement rounds. B. Active site of the final model of AvHCS, with labels for the residues interacting with the Mn^{2+} ion. C. In yellow, the final structure of AvHCS superimposed with the chosen template (6E1J) in pink. D. Active sites of AvHCS (yellow) superimposed with template 6E1J (pink).

The crystallized structure of TtHCS was used as reference (Okada *et al.*, 2010) to determine the fold of the simulated structure for AvHCS produced in this investigation. This choice was made as TtHCS represented the only bacterial candidate with a deposited structure in the PDB archive. AvHCS resulted to be a domain-swapped homodimer, as often reported for other HCS candidates (Bulfer, 2010; Okada *et al.*, 2010). Each monomer is composed of: a N-terminal domain (TIM barrel domain), and a C-terminal domain divided into two subdomains of mixed topology. Since the structure is domain-swapped, the C-terminal subdomain I of one single monomer reaches put to cover the active site of the other monomer, therefore each catalytic pocket is placed between the N-terminal of the first monomer and the ending C-terminal of the second monomer (Figure 5.3, supplementary data).

The N-terminal TIM barrel has a $(\beta/\alpha)_8$ fold, named after triose-phosphate isomerase; it hosts the catalytic site. $(\beta/\alpha)_8$ fold alternates β-strands and α-helices, in a unit that is strand-loop-helix-turn repeated eight times. The β-strands are located in the inner portion of the domain, forming a "barrel"; the α-helices, on the contrary, surround the barrel (Silverman *et al.*, 2001). The only discrepancy between reference and AvHCS simulated structure appears between the β4 strand and the α4 helix, which in the reference is labelled as a disordered region. In the produced structure, on the contrary, an α-helix structure is present. The C-terminal subdomain I is instead of mixed topology, as the TtHCS reference reports (α9- α10- β9- α11- β10), however the parallel β sheet formed by β9 and β10 could not be found in the produced structure, probably because of the shortness of the strands. Finally, the C-terminal is composed of three final α-helices that are only theorized in the TtHCS references; instead, in the simulated AvHCS, these three final α-helices structures are visible (α12-α13-α14). A summary of the two compared structures is contained in table 3.5, which highlights how the reference and simulated protein structures are mostly similar in fold.

Table 3.5. Structural comparison of ThHCS (reference) with the simulated structure produced in this study for AvHCS. Nomenclature of each secondary structure is derived by Okada *et al.*, 2010. Blue cells represent the reference structure; green cells represent conserved secondary structures; yellow cells mean divergent structures.

N- terminal domain (TIM barrel domain)													
	N- terminal												
TtHCS	β1	α1	β2	α2	β3	α3	β4	disordered region	α4	β5	α5	β6	α6
AvHCS	β1	α1	β2	α2	β3	α3	β4	α helix	α4	β5	α5	β6	α6

N- terminal domain			C- terminal subdomain I						C- terminal subdomain II	
										C- terminal
TtHCS β7	α7	β8	α8	α9	α10	β9	α11	β10	α12	α13
AvHCS β7	α7	β8	α8	/	α10	/	α11	/	α12	α13

The active site of each monomer of AvHCS is located between the TIM barrel (towards the C-terminal portion of each (β/α) unit) and the C-terminal subdomain I of the other monomer (α11 helix). The binding site is individuated by the presence of a divalent metal ion that determines α-KG binding, the first substrate that causes the morphological change of the enzyme allowing for AcCoA docking (figure 1.4). The divalent metal ion is bound to each monomer by six ligands in an octahedral geometry: Asp13, His193, His195, a water molecule, C1 and C2 groups of α-KG. The last three ligands were not found in the produced structure; however the three amino acid residues were visualized (Supplementary data, figure 5.4).

ScHCS. *LYS20* ScHCS and *LYS21* ScHCS were both produced as hybrid structures. As expected, the template for both was chosen to be the homocitrate synthase from *S. pombe* (PDB, 3IVT), as it was the highest-scoring PDB entry when compared to the target protein sequences. When aligned with the chosen template, both *LYS20* ScHCS and *LYS21* ScHCS scored 78% identities and 88% similarities (Figure 3.14).

```
LYS20             1 MTA--------------AKPNPYAAKPGDYLSNVNNFQLIDSTLREGEQFANAFFDTEKK
LYS21             1 MSENNEFQSVTESTTAPTTSNPYGPNPADYLSNVRNFQLIDSTLREGEQFANAFFDTEKK
Template_chosen   1 MSVSEA--NGTETIKPPMNGNPYGPNPSDFLSRVNNFSIIESTLREGEQFANAFFDTEKK

LYS20            47 IEIARALDDFGVDYIELTSPVASEQSRKDCEAICKLGLKAKILTHIRCHMDDAKVAVETG
LYS21            61 IEIARALDDFGVDYIELTSPVASEQSRKDCEAICKLGLKAKILTHIRCHMDDARVAVETG
Template_chosen  59 IQIAKALDNFGVDYIELTSPVASEQSRQDCEAICKLGLKCKILTHIRCHMDDARVAVETG

LYS20           107 VDGVDVVIGTSKFLRQYSHGKDMNYIAKSAVEVIEFVKSKGIEIRFSSEDSFRSDLVDLL
LYS21           121 VDGVDVVIGTSKFLRQYSHGKDMNYIAKSAVEVIEFVKSKGIEIRFSSEDSFRSDLVDLL
Template_chosen 119 VDGVDVVIGTSQYLRKYSHGKDMTYIIDSATEVINFVKSKGIEVRFSSEDSFRSDLVDLL

LYS20           167 NIYKTVDKIGVNRVGIADTVGCANPRQVYELIRTLKSVVSCDIECHFHNDTGCAIANAYT
LYS21           181 NIYKTVDKIGVNRVGIADTVGCANPRQVYELIRTLKSVVSCDIECHFHNDTGCAIANAYT
Template_chosen 179 SLYKAVDKIGVNRVGIADTVGCATPRQVYDLIRTLRGVVSCDIECHFHNDTGMAIANAYC

LYS20           227 ALEGGARLIDVSVLGIGERNGITPLGGLMARMIVAAPDYVKSKYKLHKIRDIENLVADAV
LYS21           241 ALEGGARLIDVSVLGIGERNGITPLGGLMARMIVAAPDYVRSKYKLHKIRDIENLVADAV
Template_chosen 239 ALEAGATHIDTSILGIGERNGITPLGALLARMYVTDREYITHKYKLNQLRELENLVADAV

LYS20           287 EVNIPFNNPITGFCAFTHKAGIHAKAILANPSTYEILDPHDFGMKRYIHFANRLTGWNAI
LYS21           301 EVNIPFNNPITGFCAFTHKAGIHAKAILANPSTYEILDPHDFGMKRYIHFANRLTGWNAI
Template_chosen 299 EVQIPFNNYITGMCAFTHKAGIHAKAILANPSTYEILKPEDFGMSRYVHVGSRLTGWNAI

LYS20           347 KARVDQLNLNLTDDQIKEVTAKIKKLGDVRSLNIDDVDSIIKNFHAEVSTPQVLSAKKNK
LYS21           361 KSRVDQLNLNLTDDQIKEVTAKIKKLGDVRELNIDDVDSIIKDFHAELSTPLLKPV--NK
Template_chosen 359 KSRAEQLNLHLTDAQAKELTVRIKKLADVRTLAMDDVDRVLREYHADLSDADRIT-----

LYS20           407 KNDSDVPELATIPAAKRTKPSA
LYS21           419 GTDDDNIDISNGHVSKKAKVTK
Template_chosen 414 --------------KE--ASA
```

Figure 3.14. Alignment of *LYS20* ScHCS, *LYS21* ScHCS, and the chosen template, the homocitrate synthase expressed in *S.pombe*.

The HCS expressed from *LYS20* only needed 14 homology iterations to be combined to produce the final hybid ScHCS. As for the HCS expressed from *LYS21*, 17 homology iterations were combined (Figure 3.15).
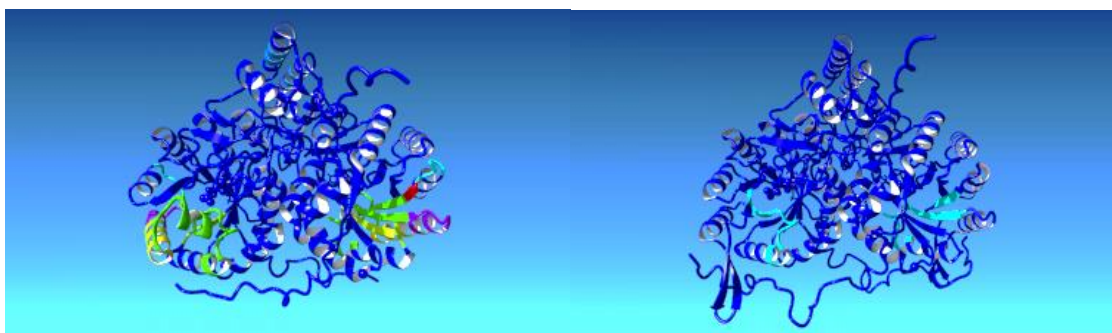
Figure 3.15. Model produced for ScHCS from *LYS20* (left) and from *LYS20* (right).

For both structures, the final model is a homodimer which contains a Zn^{2+} ion, pertinent to the literature (Kumar *et al.*, 2011). Z-score was determined to be -0.521 for *LYS20* ScHCS and -0,638 for *LYS21* ScHCS. Contrarily to the previous case, the ligands placed into the first hybrid structure were maintained, as α-KG is the first complexed substrate of the reaction flow. Refinement of the hybrid model proved to be beneficial, as shown in the overall quality check in table 3.6 and 3.7, figures 3.16 and 3.18. Final structures are shown in figures 3.17 and 3.19. Once again, the finalized structure of *LYS20* and *LYS21* ScHCS produced in this study resulted satisfactory even if being *in-silico* models. In fact, the simulated structure passed all quality checks; moreover, in this case, the ligand maintained from the chosen template (α-KG) only added to the plausibility of these structures.

Table 3.6. ERRAT, PROCHECK, Ramachandran plot and ProSAWeb results for construction of *LYS20* ScHCS model.

|	Template 3IVT	Hybrid Model ScHCS *LYS20*	First refinement ScHCS *LYS20*
ERRAT	(must be >95%)		
Dimer1	97,55 %	96,67 %	99,23 %
Dimer2	98,64 %	96,95 %	98,46 %
PROCHECK	(must be >90%)		
Residues in most favoured regions	91,0 %	92,3 %	92,20 %
Residues in additional allowed regions	8,5 %	7,4 %	6,6 %
Residues in generously allowed regions	0,3 %	0,4 %	0,4 %
Residues in disallowed regions	0,3 %	0,4 %	0,3 %
Total	100 %	100 %	100 %
ProSAWeb Z-score (In range)	-9,22	-9,75	-10,2
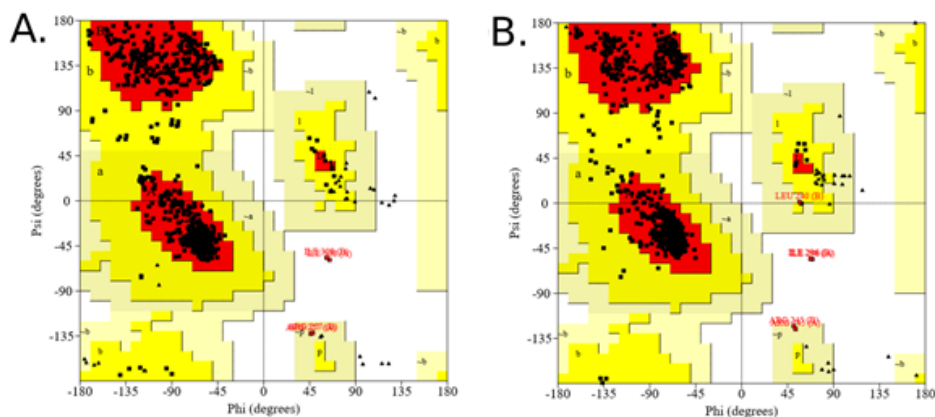
Figure 3.16. A. Ramachandran plot for chosen template 3IVT. B. Ramachandran plot for final *LYS20* ScHCS model after refinement.

26

A. Final protein model of *LYS20* ScHCS

B. Active site of the final protein model of *LYS20* ScHCS

C. Final protein model of *LYS20* ScHCS superimposed on chosen template 3IVT

D. Active site of final protein model of *LYS20* ScHCS superimposed on the active site of chosen template 3IVT
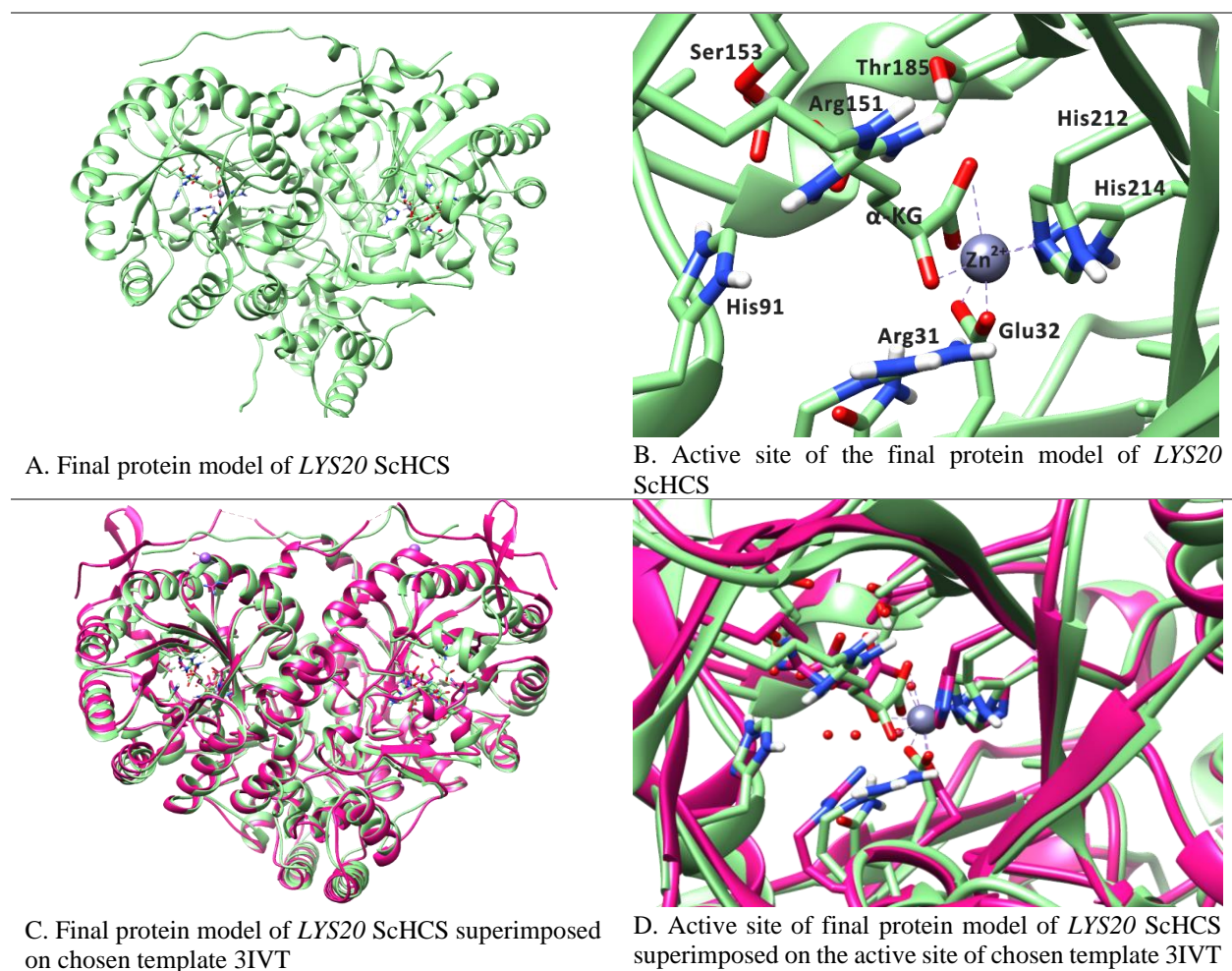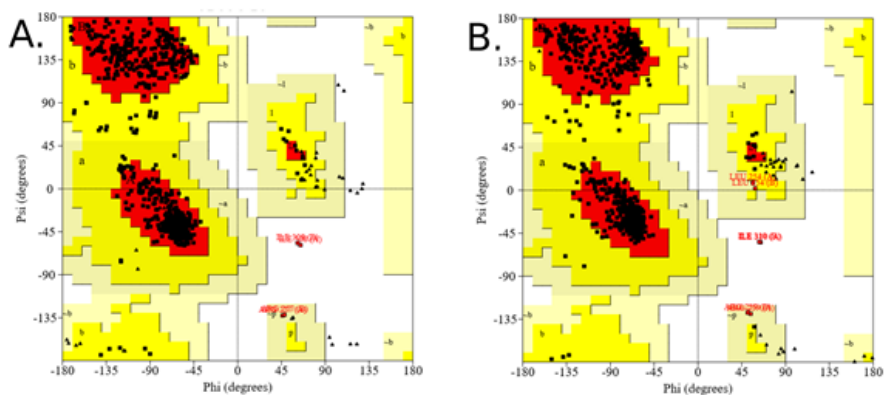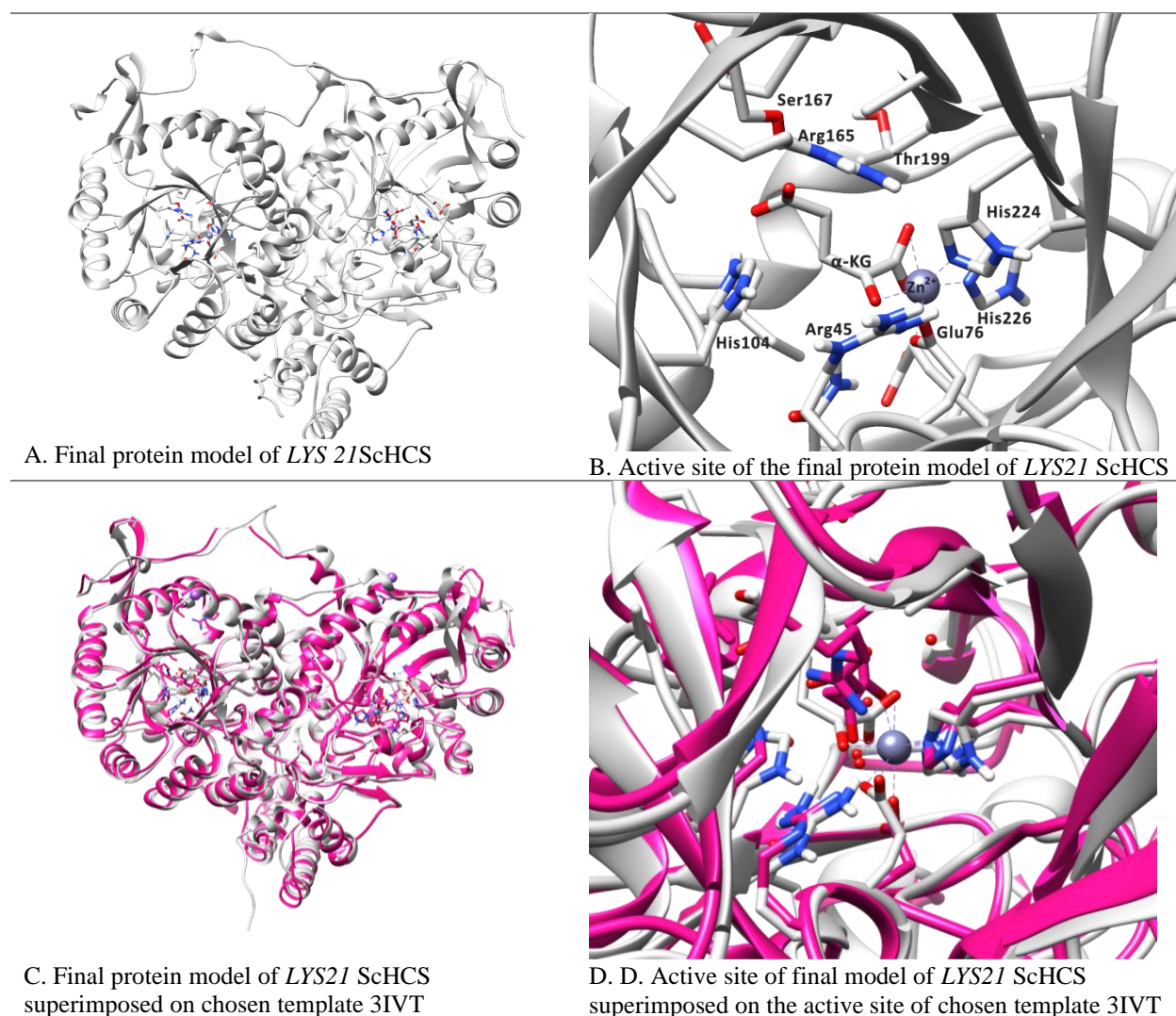
Figure 3.17. *LYS20* ScHCS protein model produced in this study. In all images, the lilac sphere represents the Zn^{2+} ion. A. In green, the final structure of *LYS20* ScHCS after refinement rounds. B. Active site of the final model of *LYS20*ScHCS, with labels for the residues interacting with the Zn^{2+} ion and α-KG. C. In green, the final structure of *LYS20* ScHCS superimposed with the chosen template (3IVT) in pink. D. Active sites *LYS20* ScHCS (green) superimposed with template 3IVT (pink).

Table 3.7. ERRAT, PROCHECK, Ramachandran plot and ProSAWeb results for construction of *LYS21* ScHCS model.

	Template 3IVT	Hybrid Model ScHCS *LYS21*	First refinement ScHCS *LYS21*
ERRAT	(must be >95%)		
Dimer1	97,55 %	95,07 %	96,54 %
Dimer2	98,64 %	95,57 %	97,48 %
PROCHECK	(must be >90%)		
Residues in most favoured regions	91,0 %	91,5 %	93,10 %
Residues in additional allowed regions	8,5 %	7,5 %	6,1 %
Residues in generously allowed regions	0,3 %	0,7 %	0,5 %
Residues in disallowed regions	0,3 %	0,3 %	0,3 %
Total	100 %	100 %	100 %
ProSAWeb Z-score (in range)	-9,22	-8,94	-10,04

Figure 3.18. A. Ramachandran plot for chosen template 3IVT. B. Ramachandran plot for final *LYS21* ScHCS model after refinement.

A. Final protein model of *LYS 21*ScHCS

B. Active site of the final protein model of *LYS21* ScHCS

C. Final protein model of *LYS21* ScHCS superimposed on chosen template 3IVT

D. D. Active site of final model of *LYS21* ScHCS superimposed on the active site of chosen template 3IVT

Figure 3.19. *LYS21* ScHCS protein model produced in this study. In all images, the lilac sphere represents the Zn^{2+} ion. A. In silver, the final structure of *LYS21* ScHCS after refinement rounds. B. Active site of the final model of *LYS21*ScHCS, with labels for the residues interacting with the Zn^{2+} ion and α-KG. C. In silver, the final structure of *LYS21* ScHCS superimposed with the chosen template (3IVT) in pink. D. Active sites *LYS21* ScHCS (silver) superimposed with template 3IVT (pink).

LYS20 ScHCS and *LYS21* ScHCS final structures were superimposed to visualize conserved regions in the active site (Figure 3.20). The active sites of the two enzymes appear to be fully conserved.
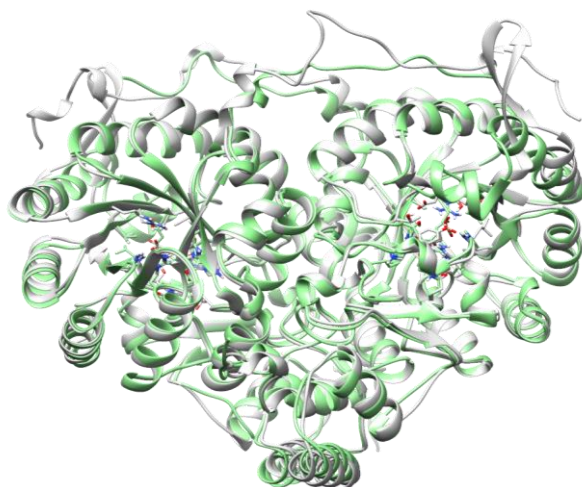
A. Final model of ScHCS *LYS20* superimposed on ScHCS *LYS21*
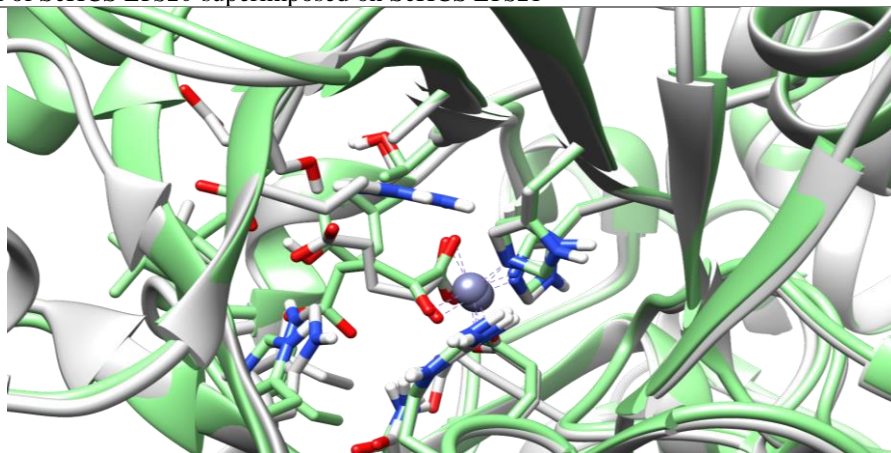
B. Active site of final protein model of ScHCS *LYS21* superimposed on the active site of final protein model of ScHCS *LYS21*

Figure 3.20. A. Protein structure of *LYS20* ScHCS (green) superimposed on *LYS21* ScHCS (silver). B. Active sites protein model of ScHCS *LYS21* (silver) superimposed on the active sites protein model of ScHCS *LYS20* (green).

Finally, using *S. pombe* HCS (SpHCS) as reference (Bulfer, 2010), the fold of the simulated structures for *LYS20* and *LYS21* ScHCS could be analysed. *LYS20* and *LYS21* ScHCS simulated structures are composed of: an N-terminal TIM barrel domain and a C-terminal domain subdivided in two subdomains, of mixed topology. The produced structures are domain-swapped homodimers, as reported by Bulfer (2010) for SpHCS. The C-terminal subdomain II of one single monomer covers the active site of the other monomer, as the two monomers are largely entangled (Figure 5.2, supplementary data).

The N-terminal TIM barrel has a $(\beta/\alpha)_8$ fold, named after triose-phosphate isomerase; it hosts the catalytic site. C-terminal subdomain I is instead of mixed topology, with two β-sheets (one anti-parallel, one parallel) interspersed with three 3_{10} helices. Two of these helices, $3_{10}5$ and $3_{10}6$ form the lid motif of the protein. Finally, the C-terminal is composed of three final α-helices.

The reference structure N-terminal domain starts with a β-strand (β0), followed by a disoriented region, a 3_{10} helix ($3_{10}1$), and finally the first β-strand (β1) involved in the TIM barrel structure. In the simulated structure for *LYS20* HCS, the TIM barrel structure starts with one supplementary 3_{10} helix right between

the $3_{10}1$ helix and the β1 strand. Instead, the simulated structure of for *LYS21* ScHCS starts with three 3_{10} helices before the $3_{10}1$ helix noted in the reference; another 3_{10} helix is also found before the β1 strand. This incongruence at the N-terminal of the enzymes might be caused by the difference in the organisms from which the reference and the simulated structures are expressed from; or, the inaccuracies could be due to the software employed for structure simulation.

From the β1 strand on, *LYS20* ScHCS and *LYS20* ScHCS structures are mostly identical to the SpHCS reference, except for the hairpin structure formed by β4.1 and β4.1, in-between the β4 strand and the α4 helix, which was missing into the *LYS20* ScHCS model and present in the *LYS21* ScHCS model.

As for the C-terminal subdomains: subdomains I of both simulated structures were mostly similar to the reference. The small anti-parallel β-sheet structure (β9-β10) was not individuated, possibly because of its shortness; on the contrary, the parallel β-sheet motif (β11- β12) was found in both produced structures. The parallel β-sheet motif is spaced by the presence of three 3_{10} helices ($3_{10}5$ to $3_{10}7$): as per reference, $3_{10}5$ and $3_{10}6$ represent the lid motif of the enzyme, which is also true for *LYS20* ScHCS and *LYS21* ScHCS (Figure 3.26). Finally, C-terminal subdomains II of both simulated ScHCS were mostly similar to the reference as well, with three α-helices (α11- α12- α13); however, after these α-helices, the reference signals a disordered region, which is instead a 3_{10} helix for the simulated structure of *LYS20* ScHCS and *LYS21* ScHCS. Once again, ending portions of simulated structures usually have less accuracy and are most variable among proteins belonging to the same family.

A summary of the three compared structures is contained in table 3.8, which highlights how these structures are mostly similar in fold, however diverge mostly in the N-terminal and C-terminal portions.

Table 3.8. Structural comparison of *Lys4* SpHCS (reference) with the simulated structures produced in this study for *LYS20* ScHCS and *LYS21* ScHCS. Nomenclature of each secondary structure is derived by Bulfer, 2010. Blue cells represent the reference structure; green cells are conserved secondary structures; yellow cells mean divergent structures. Aside from the initial and final portion, the two simulated structures for *LYS20* and *LYS21* ScHCS are mostly similar to the *Lys4* SpHCS reference.

N- terminal domain (TIM barrel domain)															
	N- terminal		TIM barrel											Hairpin	
Lys4 SpHCS	β0	disordered region	$3_{10}1$	/	β1	$3_{10}2$	α1	β2	α2	β3	α3	β4	β4.1	β4.2	
LYS20 ScHCS	/	/	$3_{10}1$	3_{10}	β1	$3_{10}2$	α1	β2	α2	β3	α3	β4	3_{10}		
LYS21 ScHCS	3_{10}	3_{10}	3_{10}	$3_{10}1$	3_{10}	β1	$3_{10}2$	α1	β2	α2	β3	α3	β4	β4.1	β4.2

N- terminal domain (TIM barrel domain)												
Lys4 SpHCS	α4	β5	$3_{10}3$	α5	β6	α6	β7	α7	β8	$3_{10}4$	α8	α9	α10
LYS20 ScHCS	α4	β5	$3_{10}3$	α5	β6	α6	β7	α7	β8	$3_{10}4$	α8	α9	α10
LYS21 ScHCS	α4	β5	$3_{10}3$	α5	β6	α6	β7	α7	β8	$3_{10}4$	α8	α9	α10

C-terminal domain subunit I							C-terminal domain subunit I				
											C-terminal
Lys4 SpHCS	β9	β10	β11	$3_{10}5$	$3_{10}6$	$3_{10}7$	β12	α11	α12	α13	disordered region
LYS20 ScHCS	/	/	β11	$3_{10}5$	$3_{10}6$	$3_{10}7$	β12	α11	α12	α13	3_{10}
LYS21 ScHCS	/	/	β11	$3_{10}5$	$3_{10}6$	$3_{10}7$	β12	α11	α12	α13	3_{10}

The active site of each monomer is located between the TIM barrel (towards the C-terminal portion of each (β/α) unit) and the C-terminal subdomain I of the other monomer ($3_{10}5$ and $3_{10}6$ helices, which form the lid motif). The Zn^{2+} divalent metal ion is hosted in the binding site of each monomer and takes contact with six ligands in an octahedral geometry: Glu32, His212, His214, a water molecule, C1 and C2 groups of α-KG for *LYS20* HCS and Glu46, His226, His228, a water molecule, C1 and C2 groups of α-KG for *LYS21* HCS. The only ligand not visualized was the water molecule. (Supplementary data, figure 5.4).

Figure 3.21. Amino acids involved in the docking of α-KG for TtHCS.
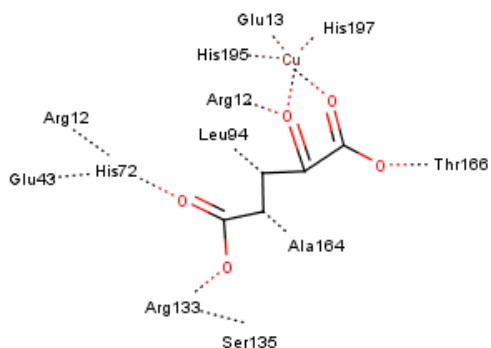
In order to determine how much the active site pocket was conserved in the three target proteins, the amino acids taking part in docking of the primary substrate (α-KG) were noted, based on the deposited structure of TtHCS (Okada *et al.*, 2009), literature from ScHCS (Qian *et al.*, 2008) (Figure 3.21) and the simulated structures produced in the present investigation. Alignment of the four target HCS showed these interacting residues to be mostly conserved. (Figure 3.22).

```
LYS20       223  NAYTALEGGARLIDVSVLGIGERNGITPLGGLMARMIVAAPDYVKSKYKLHKIRDIENLV
LYS21       237  NAYTALEGGARLIDVSVLGIGERNGITPLGGLMARMIVAAPDYVRSKYKLHKIRDIENLV
Thermus     206  NAYEAIEAGATHVDTTILGIGERNGITPLGGFLARMYTLQPEYVRRKYKLEMLPELDRMV
Azotobacter 204  NTLAAVMGGATHINTTVNGLGERAGNAALEECVLALKNL--HGIDTGIDTRGIPAISALV

LYS20       283  ADAVEVNIPFNNPITGFCAFTHKAGIHAKAILANPSTYEILDPHDFGMKRYIHFANRLTG
LYS21       297  ADAVEVNIPFNNPITGFCAFTHKAGIHAKAILANPSTYEILDPHDFGMKRYIHFANRLTG
Thermus     266  ARMVGVEIPFNNYITGETAFSHKAGMHLKAIYINPEAYPPYPPEVFGVKRKLIIASRLTG
Azotobacter 262  ERASGRQVAWQKSVVGAGVFTHEAGIHVDGLLKHRRNYPGLNPDELGRSHSLVL-GKHSG

LYS20       343  WNAIKARVDQLNLNLTDDQIKEVTAKIKKLGDVRSLNIDDVDSIIKNFHAEVSTPQVLSA
LYS21       357  WNAIKSRVDQLNLNLTDDQIKEVTAKIKKLGDVRPLNIDDVDSIIKDFHAELSTPLLK--
Thermus     326  RHAIKARAFELGLHYGEEEHRVTQHIKALADRGQLTLEELDRILREWIT----------
Azotobacter 321  AHMVRNTYRDLGIELADWQSQALLGRIRAFSTRT-KRRSPQPAELQDFYRQLC-------

LYS20       403  KKNKKNDSDVPELATIPAAKFTKPSA
LYS21       415  PVNKGTDDDNIDISNGHVSKKAKVTK
Thermus     376  ------------------------A
Azotobacter 373  ------EQGNPELAAGGM-------A
```

Figure 3.22. Alignment of target HCS proteins. Colour code for interacting residues: green, catalytic triad; red, conserved or mostly conserved; pink, similar in properties; blue, divergent.

The catalytic triad was confirmed to be glutamate, histidine and tyrosine. As for the rest of the residues concerned with docking, seven out of eleven were completely conserved, or replaced with an amino acid with similar properties. Four additional residues were mostly conserved between ScHCS and TtHCS, while diverging for AvHCS (Table 3.9). Ordered interacting residues were summarized in figure 3.23.

Table 3.9. Table A, catalytic triad residues. Table B, conserved interacting residues. Table C, diverging interacting residues.

Table A

TtHCS	Glu 137	His 292	Tyr 303
AvHCS	Glu 136	His 288	Tyr 299
LYS20 ScHCS	Glu 154	His 309	Tyr 320
LYS21 ScHCS	Glu 169	His 323	Tyr 334

Table B

TtHCS	Arg 12	Glu 43	Ala 164	Thr 166	His 195	His 197	Glu 13
AvHCS	Arg 12	Glu 43	Ala 164	Thr 166	His 193	His 193	Asp 13
LYS20 ScHCS	Arg 31	Glu 62	Ala 183	Thr 185	His 212	His 214	Glu 32
LYS21 ScHCS	Arg 45	Glu 76	Ala 197	Thr 199	His 226	His 228	Glu 46

Table C

TtHCS	His 72	Leu 94	Arg 133	Ser 135
AvHCS	Trp X	Ser X	Cis X	Gly X
LYS20 ScHCS	His 91	Val 112	Arg 151	Ser 153
LYS21 ScHCS	His 104	Val 127	Arg 167	Ser 167

TtHCS

Glu13 His197
His195 Cu
Arg12
Leu94
Glu43 ···· His72
Thr166
Ala164
Arg133
Ser135

AvHCS

Asp13 His195
His193 ··· Cu
Arg12
X
Glu43 ···· X
Thr166
Ala164
X
X

Lys20
ScHCS

Glu32 His214
His212 ··· Cu
Arg31
Val112
Glu62 ··· His91
Thr185
Ala183
Arg151
Ser153

LYS21
ScHCS

Glu46 His228
His226 ··· Cu
Arg45
Val127
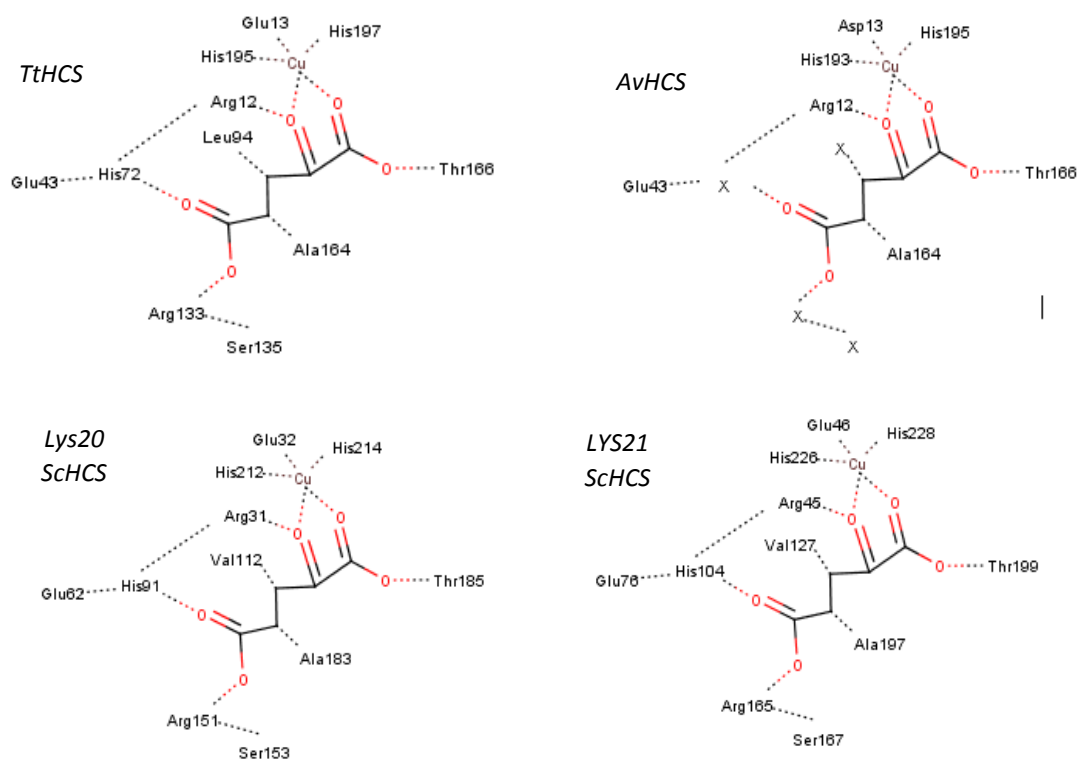Glu76 ···· His104
Thr199
Ala197
Arg165
Ser167

Figure 3.23. Ordered residues in the active site of HCS from *T. thermophilus* (TtHCS), *A. vinelandii* (AvHCS) and *S. cerevisiae* (*LYS20* ScHCS, *LYS21* ScHCS) interacting with α-KG. X defines a non-conserved residue.

However, one great difference between TtHCS and ScHCS is the domain the expressing organism belong to. *S. cerevisiae*, as previously stated, is adorned of a peculiar lid motif that encloses the catalytic pocket. Firstly discovered in *S. pombe*, this structure was detected to be formed by residues 320-330 in helices $3_{10}5$ and $3_{10}6$ of C-terminal subdomain I (Bulfer, 2010). Alignment of the target proteins along with *S. pombe* homocitrate synthase (SpHCS) in this investigation showed the residues pertaining to the lid motif to be completely conserved in both yeasts: for paralog *LYS20* it is residues 309-320, while for paralog *LYS21* residues 323-334, both located in helices $3_{10}5$ and $3_{10}6$ of C-terminal subdomain I. The lid structure is also present in TtHCS, even if diverging for five residues out of twelve when compared to its yeast relatives; presence of this gatekeeping motif was also confirmed by the team responsible for the crystallized structure of TtHCS (Okada *et al.*, 2009). TtHCS lid motif residues are 292-303 (Figure 3.24). Lid residues have been highlighted in the ScHCS structures produced during this investigation (Figure 3.25).

```
ScHCS Lys20   309 HAKAILANPSTYEILDPHDFGMK---------------------RYIHFANRLTGWNAIK
ScHCS Lys21   323 HAKAILANPSTYEILDPHDFGMK---------------------RYIHFANRLTGWNAIK
SpHCS Lys4    321 HAKAILANPSTYEILKPEDFGMS---------------------RYVHVGSRLTGWNAIK
TtHCS Lys20   292 HLKAIYINPEAYEPYPPEVFGVKRKLIIASRLTGRHAIKARAEELGLHYGEEELHRVTQHI
AvHCS NifV    301 G------------LNPDELGRSHSLVLGKHSGAHMVRNTYRDLGIELADWQSQALLGRIR
```

Figure 3.24. Aligned target sequences with SpHCS show TtHCS and ScHCS to exhibit the residues needed to form the lid motif, gatekeeper of the α-KG catalytic site.
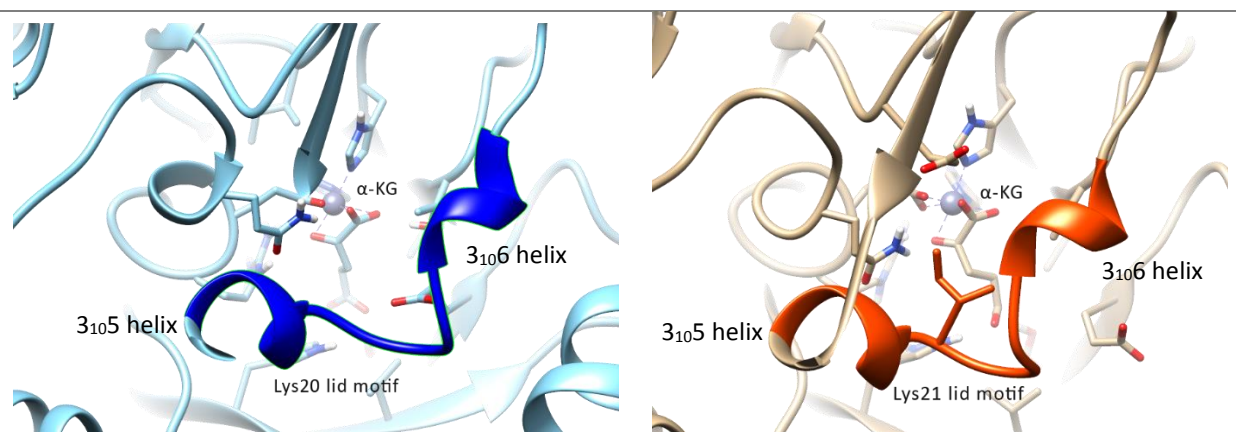
Figure 3.25. Lid motif highlighted within *LYS20* ScHCS (blue) and *LYS21* ScHCS (orange). In the background, the substrate (α-KG) is visible.

In order to find the missing interacting residues in the AvHCS catalytic pocket, a DALI search (Protein Structure Comparison Server) was performed. Highest identity value was scored by 2-Isopropymalate synthase (IPMS) expressed in *Mycobacterium tuberculosis*, an enzyme involved in the leucine pathway biosynthesis. However, the crystallized ligand, alpha-ketoisovalerate (KIV) is a short-chain keto acid derivative with a hydrophobic moiety that replaces the characteristic C5 carboxyl group of α-KG. Once again, the conserved residues pertain the C1-C3 portion of the substrate (Figure 3.26).
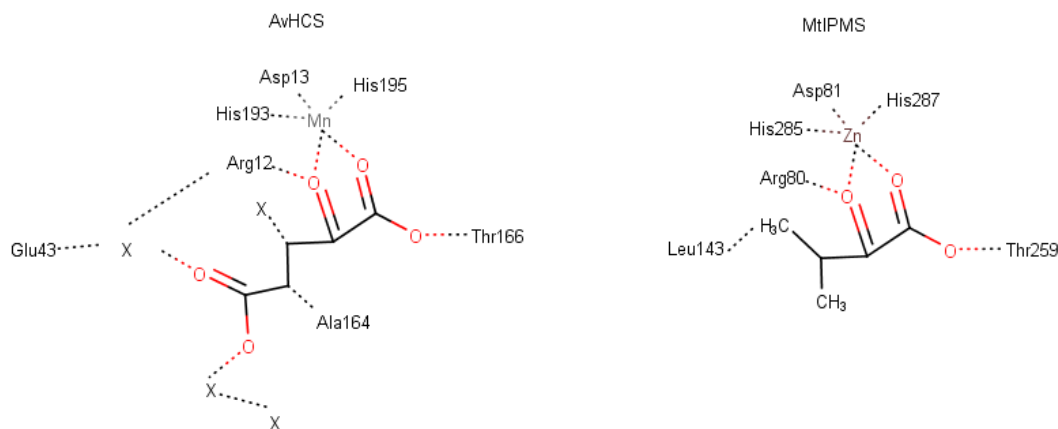

Figure 3.26. Catalytic pockets of AvHCS and 2-Isopropymalate synthase from *Mycobacterium tuberculosis* (MtIPMS).

Part 4 – Discussion

Since the investigation revolved around HCS substrate specificity, supplementary substrates were of great interest: the question concerned the reason these shorter-chain (oxaloacetate) or longer-chain keto-acids (alpha-ketoapidate) could be employed as substrates by different HCS homologues. For example, the use of alpha-ketoadipate as substrate replacing α-KG in AvHCS always stuck out as the most unique ability of this enzyme. The reason might be purely logical: in TtHCS and ScHCS, alpha-ketoadipate appears further down the alpha-aminoadipate (AAA) metabolic pathway, as taken care by a 2-aminoadipate transaminase using L-glutamate as donor, as shown in figure 8. This implies that HCS in the AAA pathway can leave the use of alpha-KA to a more specific enzyme further down the line.

On the contrary, lysine biosynthesis in prokaryotes such as *A. vinelandii* follows the diaminopimelate pathway (DAP), that originates with aspartate ([Kegg, Lysine biosynthesis in *A. vinelandii*](#)). Consequently, alpha-ketoadipate usage can be also taken care of by HCS, and specifically for AvHCS, funneling as many ketoacids into the production of homocitrate, a vital part of the nitrogenase cofactor shown in figure 14.

Whether this ability to use this C6 ketoacid was lost in the AAA pathway HCS or gained in the DAP pathway HCS is another notion that would require more in-depth evolutionary studies.

As for oxaloacetate, it is the preferred substrate of a closely related enzyme, citrate synthase (CS), whose reaction is a Claisen condensation of oxaloacetate and acCoA to form citrate. However, from a structural, sequence and catalytic point of view, fungal and bacterial CS and HCS have very little in common, as proved by Bulfer *et al.* (2009). The similarity in the type of reaction opens the possibility for a common enzymatic ancestor from which homocitrate synthase, citrate synthase, alpha-isopropylmalate synthase and malate synthase originated. In fact, these four enzymes all catalyze an aldol-type condensation reaction that employs acCoA as donor of the acetyl group, even if differing greatly on many other features. More in-depth studies would be needed for this topic as well.

Finally, the three HCS candidates are evaluated through the investigation results.

ScHCS. Expression of ScHCS for this study was not successful. The reason is largely unknown; the bottleneck appeared mostly between the ligation of vector and insert and expression of target protein. User error, inhibitors in the solution and general lack of optimal conditions could only be a few of the reasons. It must also be noted that ScHCS is regarded as a quite unstable enzyme, as the crude extract from *S. cerevisiae* was tested to be stable for only 4 days at 0°C (Gray and Bhattacharjee, 1976).

The present study produced protein models of both ScHCS paralogs (figures 3.18 and 3.20) whose quality checks confirmed these models as satisfactory for *in silico* structures (Tables 3.6 and 3.7, summary in table 4.1). The produced structures are domain-swapped homodimers, as literature reports (Bulfer, 2010). The overall structure of each monomer was subdivided according to literature on SpHCS: the N-terminal – a TIM barrel domain - containing the active site, then a C-terminal with two subdomains (I and II) (Bulfer, 2010). The C-terminal subdomain I of one single monomer stretches up to the active site of the other monomer, meaning the catalytic pocket is formed by the N-terminal of the first monomer and the ending C-terminal of the second monomer (Figure 5.2, supplementary data).

The N-terminal of both structures has a $(\alpha/\beta)_8$ fold (TIM barrel), while C-terminal domain is instead of mixed topology. As also proved in this investigation, two of the 3_{10} helices in the C-terminal subdomain I, $3_{10}5$ and $3_{10}6$, form the lid motif of the protein (figure 3.26). When compared to the deposited crystallized structure of *Lys4* SpHCS (Bulfer, 2010), the simulated structures produced in this study for *LYS20* ScHCS and *LYS21* ScHCS result mostly similar (table 3.8). Diverging features are represented by a β-hairpin structure between β4 strand and α4 helix, which was only missing in the produced structure for *LYS20* ScHCS. The reason for the lack of this structural motif in *LYS20* ScHCS is unknown, possibly due to the need for supplementary refinement of the produced model. According to Bulfer *et al.* (2010), in SpHCS this β4-α4 portion of the enzyme is a disordered region in the free enzyme, which turns ordered in a β-hairpin once α-KG is bound; this is further proof of the conformational change that the enzyme undergoes once docking of α-KG takes place.

Other diverging features when compared to the reference structure are present in the early portion of the N-terminal or the last portion of the C-terminal, as reported in table 3.8. This discrepancy might be caused by the different organisms from which the reference and the target protein derive from. In fact, as previously stated, *LYS20* ScHCS and *LYS21* ScHCS are transported into the nucleus by a non-identified mechanism that could be an atypical nuclear localization sequence at one of the two terminals. Another possibility could be the inaccuracy of the software employed to produce the simulated structures. No matter the cause, the only solution to dissipate these doubts would be to crystallize *LYS20* ScHCS and *LYS21* ScHCS.
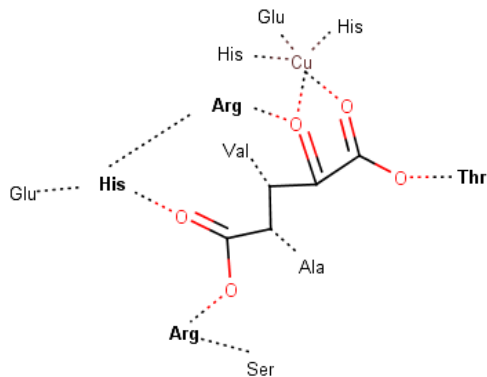
Figure 4.2. *In vivo* rational mutagenesis of the residues in bold resulted in decreased or inexistent activity in SpHCS, as reported by Qian *et al.* (2008).

Simulating these protein structures resulted helpful in finding the identity and location of ScHCS catalytic residues, as summarized in figure 3.24, in order to value this protein as a possible target of active site mutagenesis to allow employment of diverse keto acids as substrates.

Literature reports that attempted to perform rational mutagenesis on ScHCS and SpHCS active sites highlight how strictly conserved the interacting aminoacidic residues are (Figure 4.1) (Qian *et al.*, 2008) (Bulfer *et al.*, 2009). Moreover, the presence of a lid motif – whose residues were also first reported and visualized for ScHCS in this study (figure 3.24 and 3.25) - greatly increases the difficulty of this operation, along with its nuclear localization and the involvement in a vital second function (Scott, 2010). Consequently, the delicate structure of fungal HCS leaves small hope to widening substrate specificity through rational mutagenesis. For these reasons, ScHCS strict substrate specificity and highly conserved catalytic residues do not qualify ScHCS as a fitting candidate to be the elongation catalyst for a recursive +1C system.

TtHCS. The availability of literature upon TtHCS and the present study prove how the expression of this enzyme was fairly simple to reproduce. This feature qualifies TtHCS as a better candidate than ScHCS. The enzymatic assays performed (figures 3.7, 3.8 and 3.9) proved how the tested reaction is a second order reaction, therefore is dependent on the concentration of two reactants (DCPIP and CoA-SH). In table 3.3, the amount of protein placed into the reaction mix always correlated to a faster reaction (greater k constant). The total amount of reduced DCPIP is consistent between the samples as it defined the total amount of the substrate used, meaning all reactions exhausted all the substrate available. However, this exhaustion of substrate happened at different points in time: TtHCS1, having the greater k value among the samples, depleted the substrate long before the other two samples. TtHCS2 is the sample with the slowest employ of substrate. This notion can also be confirmed by the initial rates, even if not accurately, due to the interval between placement of the enzyme and start of the manual monitoring. Initial rates are higher for samples TtHCS1 and TtHCS3, which define the reactions as employing the substrate faster and depleting it quickly. However, once the substrate was finished, all the plots reached the stationary state of the reaction. Kinetic parameters could not be calculated as multi-substrate kinetics are evaluated by fixing the concentration of one of the two substrates and variating the concentration of the second substrate. Moreover, use of different ketoacids – in this case, oxaloacetate – is also needed for further testing. However, this assay proved that the samples of TtHCS extracted and isolated from different colonies behaved consistently and did not lose their catalytic ability.

As for substrate choice, TtHCS use of C4 and C5 ketoacids (Figure 1.7) still places this candidate below AvHCS, since the +1C recursive system will have to deal with C5 and C6 substrates. Once again, rational mutagenesis could be considered to push the enzyme into accepting C6 ketoacids, though it would require additional analysis. There are no hints indicating which residues should undergo rational mutagenesis to allow this change in substrate specificity. Nevertheless, one interesting mutation could pertain His72 of TtHCS, which is one of the key residues that causes TtHCS and ScHCS to be feedback-inhibited by lysine (Kumar *et al.*, 2011). AvHCS, on the contrary, exhibits a tryptophan residue instead of histidine (as confirmed during this investigation), and is not inhibited by lysine, as the HCS in *A. vinelandii* is not involved in lysine biosynthesis (Table 3.9). Histidine to tryptophan substitution in TtHCS might avoid inhibition issues once the protein will be part of the +1C alpha-ketoacid elongation pathway. For these

36

reasons, TtHCS can still be considered a viable candidate to be the elongation catalyst for a recursive +1C system, if it undergoes mutagenesis tests.

     AvHCS. Purification of AvHCS was unsuccessful, though the reasons may lie in expression conditions, which were not optimized due to time restraints. However, *in silico* studies of this investigation proved some interesting features. The 3D model produced in this report was the first ever presented for AvHCS (Figure 3.13), scoring satisfactory quality control values for being an *in-silico* simulation of the protein structure (Table 3.4 and figure 3.12, summary in table 4.1).

Table 4.1. Summary of final protein model quality checks.

	AvHCS final model	*LYS20* ScHCS final model	*LYS21* ScHCS final model
ERRAT (>95%)	96,76 %	99,23 %	96,54 %
	95,15 %	98,46 %	97,48 %
PROCHECK (>90%)	93,2 %	92,20 %	93,10 %
ProSAWeb (in range)	-7,37	-10,2	-10,04

On account of the comparison with the crystallized structure of TtHCS (Okada *et al*., 2010) employed as a reference, it was determined that the AvHCS structure is qualified once again as a domain-swapped homodimer, with N-terminals hosting the active site and C-terminal of the first monomer stretching out to the catalytic pocket (Figure 5.3, supplementary data). The N-terminal TIM barrel (named after triose-phosphate isomerase) has a $(\beta/\alpha)_8$ fold. The C-terminal domain fold is more variated; in subdomain I it was not possible to individuate the two β9 and β10 strands of the α9-α10-β9- α11-β10 reported in the reference. This might be due to the shortness of the β-strands. In C-terminal subdomain II, the bundle of three α-helices proposed by the references were instead individuated (table 3.5). Further analysis in the direction of crystallization is certainly needed for this enzyme candidate, as it undertakes a metabolic role which is different from the deposited references of SpHCS and TtHCS.

Structure modelling facilitated finding the identity and location of some of AvHCS catalytic residues, a notion first reported in this study (figure 3.23). These amino acid residues involved in docking of α-KG might be the answer to AvHCS ability to employ diverse ketoacids as substrates: α-KG C1-C3 docking residues were successfully identified; α-KG C4-C5 docking residues, however, remain a mystery (figure 3.24). Crystallization of the enzyme would finally unveil this notion. In conclusion AvHCS, if capable of using C4 to C6 ketoacids as literature reports, is therefore the most valid candidate as elongation catalyst for a recursive +1C system, if further characterized. Employment of this candidate eliminates the need to perform rational mutagenesis on other HCS isoforms.

Future prospects of this project would tie the elongation process presently studied to the remaining enzymatic steps in the direction of adipic acid production through use of linear-chained ketoacid substrates (figure 5). Of course, optimization with the successive enzymatic steps is needed to create a seamless and highly functioning metabolic flow that could lead to an efficient production, hoping one day to reach satisfying yields. Many more obstacles are still on the way of this production to prove successful, and even then, adipic acid is still only a building block, a small brick. Nevertheless, brick by brick, the bridge connecting sustainable bio-based production of plastic and environment care can be built, allowing us to further enrich our knowledge about the biggest toolbox at your disposal: nature.

References

Andi, B., West, A.H. and Cook and P.F. 2004. Kinetic mechanism of histidine-tagged homocitrate synthase from *Saccharomyces cerevisiae*. Biochemistry 43(37):11790-5.

Bart J.C.J. and Cavallaro S. 2015. Transiting from adipic acid to bioadipic acid. 1, Petroleum-Based Processes. Industrial and Engineering Chemistry Resesearch 54:1–46.

Beardslee, T. and Picataggio, S. 2012. Bio-based adipic acid from renewable oils. Lipid Technology 24(10).

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. The Protein Data Bank. 2000. PDB IDs: 2ZTK, 2ZTJ, 3A9I, 2ZYF. 5 May 2019.

Boussie, T.R., Dias, E.L., Fresco, Z.M., Murphy, V.J., Shoemaker, J., Archer, R., Jiang, H. 2010. Production of Adipic Acid and Derivatives from Carbohydrate-Containing Materials. US20100317823 A1.

Bulfer, S.L., Scott, E.M., Couture, J.F., Pillus, L., Trievel, R.C. 2009. Crystal structure and functional analysis of homocitrate synthase, an essential enzyme in lysine biosynthesis. Journal of Biological Chemistry 284(51):35769-80.

Bulfer, S.L. 2010. Structure, Mechanism and Regulation of Homocitrate Synthase. Doctoral thesis. University of California, San Francisco.

Bulfer, S.L., Scott, E.M., Pillus, L., Trievel, R.C. 2010. Structural basis for L-lysine feedback inhibition of homocitrate synthase. Journal of Biological Chemistry 285(14):10446-53.

Cao, Y., Zhang, R., Sun, C., Cheng, T., Liu, Y. and Xian, M. 2013. Fermentative succinate production: an emerging technology to replace the traditional petrochemical processes. Biomed Research International 723412.

Chen, S., Brockenbrough, J.S., Dove, J.E. and Aris, J.P. 1997 Homocitrate synthase is located in the nucleus in the yeast *Saccharomyces cerevisiae*. Journal of Biological Chemistry 272(16):10839–10846.

Colovos, C. and Yeates, T.O. 1993. Verification of protein structures: Patterns of nonbonded atomic interactions. Protein Sciences 2:1511-1519.

de Jong, E., Higson, A., Walsh, P., and Wellisch, M. 2012. Bio-based chemicals value added products from biorefineries. IEA Bioenergy, Task42 Biorefinery, 34.

Draths K.M., Frost J.W. 1994. Environmentally compatible synthesis of adipic acid from D-Glucose. Journal of the American Chemical Society 116:399–400.

European Bioplastics. 2018. Bioplastics market data. [https://www.european-bioplastics.org/market/]. 13 August 2019.

Garrad, R.C. and Bhattacharjee, J.K. 1992. Lysine biosynthesis in selected pathogenic fungi: characterization of lysine auxotrophs and the cloned LYS1 gene of *Candida albicans*. Journal of Bacteriology 174(22):7379–7384.

Gray, G.S., and Bhattarcharjee, J.K. 1976. Biosynthesis of lysine in *Saccharomyces cerevisiae*: properties and spectrophotometric determination of homocitrate synthase activity. Canadian Journal of Microbiology 22(11): 1664-1667.

Hall, C.J. and Soares de Costa, T.P. 2018. Lysine: biosynthesis, catabolism and roles. Wikijournal of Science 1 (1):4.

Howell, DM., Harich, K., Xu, H., White, RH. 1998. Alpha-keto acid chain elongation reactions involved in the biosynthesis of coenzyme B (7-mercaptoheptanoyl threonine phosphate) in methanogenic Archaea. Biochemistry 37(28):10108-17.

Juturu, V. and Wu, J.C. 2016. Microbial production of lactic acid: the latest development. Critical Reviews in Biotechnology 36(6):967-977.

Kegg pathway, Kaneisha Laboratories. 23 June 2017. Lysine biosynthesis - *Azotobacter vinelandii* DJ. [https://www.genome.jp/kegg-bin/show_pathway?avn00300]. 14 August 2019.

Kegg. Entry 2.3.3.14, [https://www.genome.jp/dbget-bin/www_bget?ec:2.3.3.14]. 14 August 2019.

Kennedy, C., and Dean, D. 1992. The nifU, nifS and nifV gene products are required for activity of all three nitrogenases of *Azotobacter vinelandii*. Molecular Genetics and Genomics 231(3):494-8.

Kindler, S.H. and Gilvarg, C. 1960. N-Succinyl-L-2,6-diaminopimelic acid deacylase. Journal of Biological Chemistry 235:3532-5.

Kumar, V.P., West, A.H. and Cook, P.F. 2011. Kinetic and chemical mechanisms of homocitrate synthase from *Thermus thermophilus*. Journal of Biological Chemistry 286(33):29428–29439.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. 1993 PROCHECK: A program to check the stereochemical quality of protein structures. Journal of Applied Crystallography 26:283–291.

Madden, T. 2002. Chapter sixteen - The BLAST Sequence Analysis Tool. In: McEntyre, J. and Ostell, J., editors. The NCBI Handbook [Internet].

MarketsandMarkets Analysis. April 2014. Adipic Acid Market by Application (Polyamide 66, Polyurethane, Adipic Esters, & Others), by Material Type (Engineering Plastics & Fibers) & by Geography (North America, Europe, Asia-Pacific, & ROW) - Global Trends & Forecast to 2019 [https://www.marketsandmarkets.com/Market-Reports/adipic-acid-269.html]. 10 January 2019.

Nostredame, C., Higgins, D.C. and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology 302(1);205-17.

Okada, T., Tomita, T., Wulandari, A.P., Kuzuyama, T. and Nishiyama, M. 2010. Mechanism of substrate recognition and insight into feedback inhibition of homocitrate synthase from *Thermus thermophilus*. Journal of Biological Chemistry 285(6):4195-205.

Parmon, V.N., Panov. G.I., Uriarte, A., Noskov, A.S. 2004. Nitrous oxide in oxidation chemistry and catalysis: application and production. Catalysis Today 100: 115-131.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. Journal of Computational Chemistry. 25(13):1605-12.

Philip, J.C. 2013. Policies for Bioplastics in the Context of a Bioeconomy. OECD Science, Technology and Industry Policy Papers. 10:1089.

Qian, J., West, A.H. and Cook, P.F. 2006. Acid-base chemical mechanism of homocitrate synthase from *Saccharomyces cerevisiae*. Biochemistry 45(39): 12136-12143.

Qian, J., Khandogin, J., West, A.H. and Cook, P.F. 2008. Evidence for a catalytic dyad in the active site of homocitrate synthase from *Saccharomyces cerevisiae*. Biochemistry 47(26):6851-8.

Quezada, H., Marín-Hernández, A., Aguilar, D., López, G., Gallardo-Pérez, J.C., Jasso-Chávez, R., González, A., Saavedra, E. and Moreno-Sánchez, R. 2011. The *LYS20* homocitrate synthase isoform exerts most of the flux control over the lysine synthesis pathway in *Saccharomyces cerevisiae*. Molecular Microbiology 82(3): 578-90.

Reimer, R. A., Slaten, C. S., Seapan, M., Koch, T. A., Seapan, M. and Triner, V. G. 1999. Adipic Acid Industry — N2O Abatement. In: Non-CO2 Greenhouse Gases: Scientific Understanding, Control and Implementation. van Ham, J., Baede, A. P. M., Meyer, L. A., Ybema, R. Springer Science+Business Media, pp 347-358.

Rousseau, F., Schymkowitz, J. and Itzhaki, LS. 2012. Implication of 3D domain swapping for protein folding, misfolding and function. Advances in Experimental Medicine and Biology 747:137-52.

Scott, E.M. 2010. Characterization of the *esa1* suppressor *LYS20* and discovery of its novel nuclear functions. UC San Diego.

Shen, C.R. and Liao, J.C. 2011. Chapter twenty - A Synthetic Iterative Pathway for Ketoacid Elongation. In: Methods in Enzymology.

Silverman, J.A., Balakrishnan, R. And Harbury, P.B. 2001. Reverse engineering the (β/α)8 barrel fold. Proceedings of the National Academy of Sciences 98(6) 3092-3097.

Skoog, E., Shin, J. and Saez Jimenez, V. 2018. Biobased adipic acid – the challenge of developing the production host. Biotechnology Advances 36(8): 2248-2263.

Tang, T., Qi, F., Hongjuan, L., Dehua, L. 2013. Recent developments in the microbial production of 1,3-propanediol. Biofuels 4: 651-667.

Vardon, D.R., Franden, M.A., Johnson, C.W., Karp, E.M., Guarnieri, M.T., Linger, J.G., Salm, M.J., Strathmann, T.J. and Beckham, G.T. 2015. Adipic acid production from lignin. Energy and Environmental Science 8:617–628.

Vogel, H.J., 1960. On biochemical evolution: Lysine formation in higher plants. Proceedings of the National Academy of Sciences. 45(12):1717-21.

Voigt, C.A. 2008. Genetic parts to program bacteria. Current opinion in biotechnology 17: 548-557.

Wiederstein, M. and Sippl, M.J. 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Research 35, W407-W410.

Wulandari, A,P., Miyazaki, J., Kobashi, N., Nishiyama, M., Hoshino, T. and Yamane, H. 2002. Characterization of bacterial homocitrate synthase involved in lysine biosynthesis. FEBS Letters 522(1-3):35-40.

Yu, J., Qian, Z. and Zhong, J. 2018. Advances in bio-based production of dicarboxylic acids longer than C4. Engineering in Life Sciences 18:9.

Zhao, M., Huang, D., Zhang, X., Koffas, M.A.G., Zhou, J., and Deng, Y. 2018. Metabolic engineering of *Escherichia coli* for producing adipic acid through the reverse adipate-degradation patheay. Metabolic Engineering 47:254-262.

Zheng, L., White, RH., Dean, DR. 1997. Purification of the *Azotobacter vinelandii* nifV-encoded homocitrate synthase. Journal of Bacteriology 179(18):5963-6.

Exploring an alternative metabolic pathway for production of adipic acid
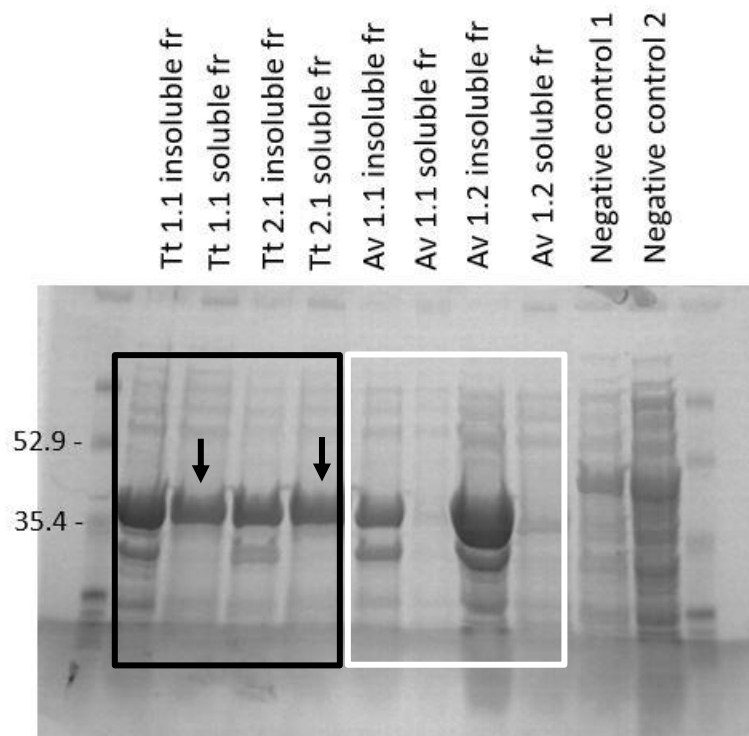## Supplementary data

Figure 5.1. SDS-PAGE exhibiting insoluble and soluble fractions produced during cell lysis of colonies transformed with plasmids carrying TtHCS or AvHCS genes. In the black square, the colonies transformed with the plasmid carrying the TtHCS gene showed bands of the appropriate size (~43 kDa) in both insoluble and soluble fractions. The soluble fractions of these samples (indicated with black arrows) were then employed for purification of the target protein. In the white square, the colonies transformed with the plasmid carrying the AvHCS gene showed bands of the appropriate size (~42 kDa) only in the insoluble fractions, meaning the enzyme precipitated with the cell debris. For this reason, AvHCS was not purified.
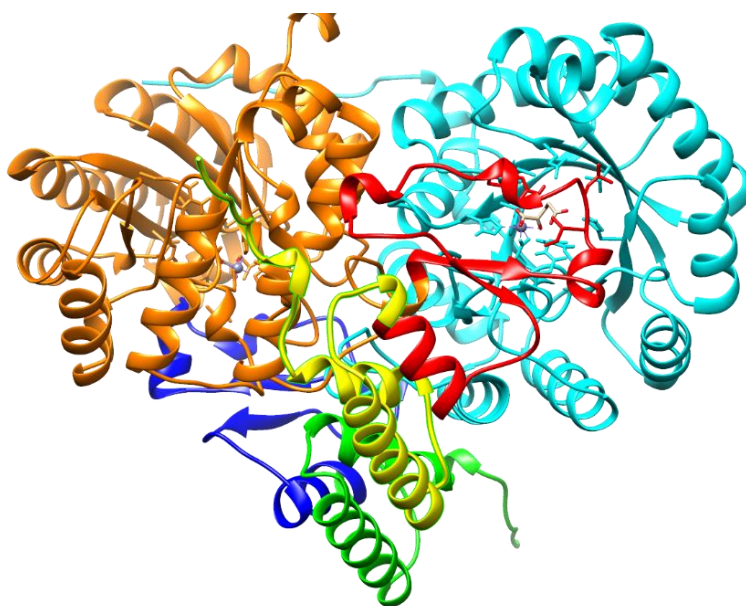
Figure 5.2. A, colour-coded domains of *LYS20* ScHCS *in-silico* structure produced in this study. *S. pombe* HCS used as reference to mark each domain (Bulfer, 2010). The homodimer is formed by two entangled monomers; each monomer has N-terminal domain, and a C-terminal domain divided in two subdomains: I and II. For monomer one, the N-terminal domain and the C-terminal subdomains I and II are respectively coloured in orange, red and yellow; monomer two has the same type of domains coloured in cyan, blue and green. The domain-swapped nature of this protein can be easily observed where the N-terminal of monomer one, in cyan, takes contact with the C-terminal of monomer two, in red. The interaction of these two portions of the two monomers creates the catalytic pocket of the enzyme.
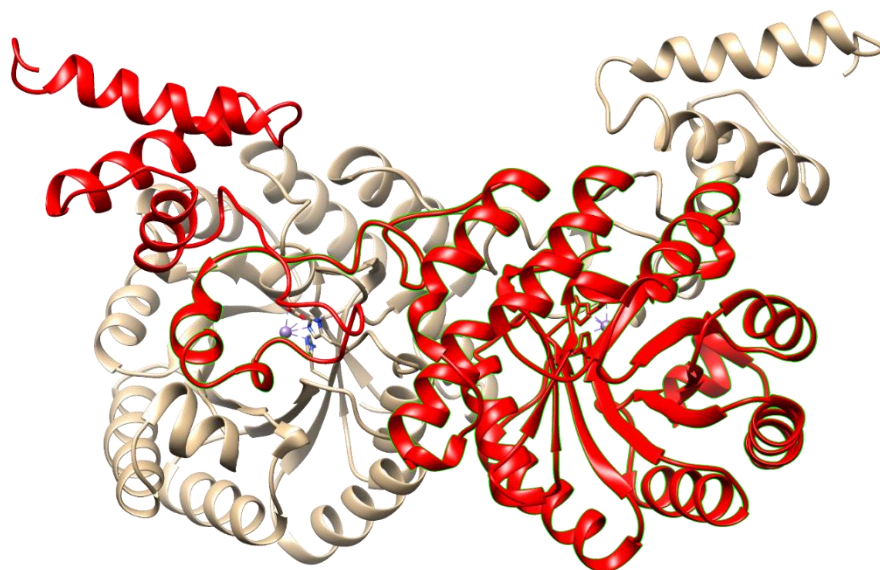
Figure 5.3. Colour-coded monomers of AvHCS *in-silico* structure produced in this study. The homodimer is formed by two entangled monomers, monomer one in white and monomer two in red; each monomer has N-terminal domain, and a C-terminal domain divided in two subdomains: I and II. Due to the lack of literature on AvHCS, specific domains residues could not be identified. However, the domain-swapped nature of this protein can be easily observed where the active site of the N-terminal (identified by the presence of a bivalent metal ion, represented as a blue sphere) of each monomer is covered by the C-terminal subdomain II of the other monomer. The interaction of these two portions of the two monomers creates the catalytic pocket of the enzyme.
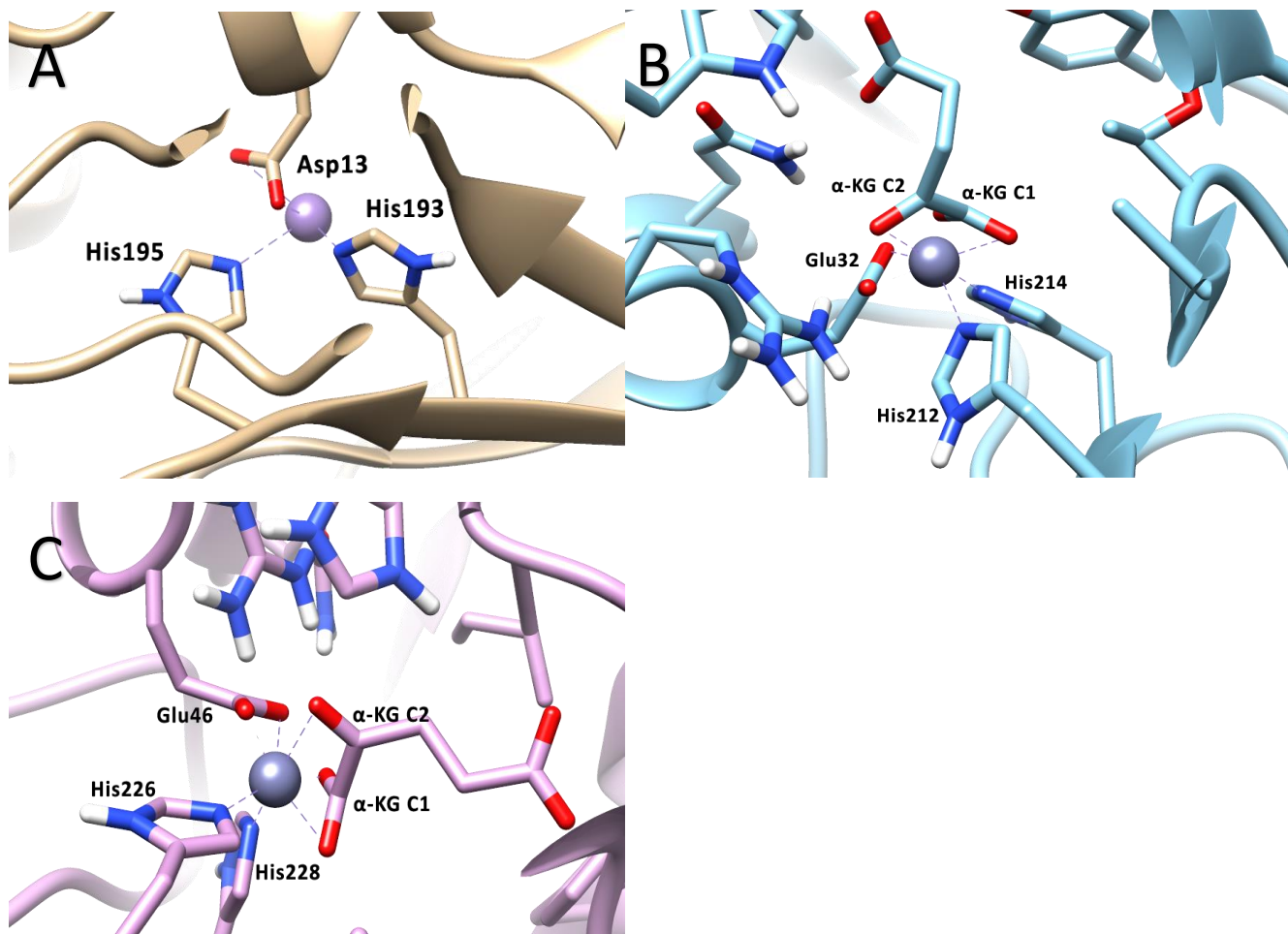
Figure 5.4. Metal ion binding site in AvHCS (A), *LYS20* ScHCS (B) and *LYS21* ScHCS (C). In all cases, the purple sphere represents a divalent ion. The metal ion is bound to the enzyme by six ligands in an octahedral geometry. Of these ligands, three are amino acid residues, one in a water molecule, and the last two are the C1-C2 groups of α-KG. A, AvHCS metal binding site lacks interaction with a water molecule and C1-C2 groups of α-KG. The three amino acid ligands are Asp13, His193 and His195. B, *LYS20* ScHCS metal binding site lacks interaction with a water molecule, however the metal ion takes contact with the C1 and C2 groups of α-KG. The three amino acid ligands are Glu32, His212 and His214. C, *LYS21* ScHCS metal binding site lacks interaction with a water molecule, however the metal ion takes contact with the C1 and C2 groups of α-KG. The three amino acid ligands are Glu46, His226 and His228.