Evolution of Structure and Function in Phenylalanine Hydroxylase

With the Regulatory Properties in Sight

Jessica Siltberg-Liberles

Dissertation for the degree philosophiae doctor (PhD)
at the University of Bergen

2008

Scientific environment

This thesis was produced in

Department of Biomedicine, University of Bergen, Norway

The Computational Biology Unit, BCCS, University of Bergen, Norway

Department of Molecular Biology, University of Wyoming, USA

Acknowledgments

First and foremost, I want to thank my supervisor, Aurora Martinez, for always being supportive and encouraging. This thesis has taken us both to new grounds, and I have been given the confidence to explore unknown fields. Thanks for sharing this interesting protein family and your deep knowledge of it with me. Thank you also for giving me the opportunity to travel to conferences near and far, and for making it possible for me to experience another research environment and culture. For all of this I want to express my uttermost gratitude.

Thanks to Ida Steen for generating the *Dictyostelium discoideum* clone and thanks to Randi Svebak and Ali Javier Sepulveda for characterizing it. Thanks to everyone in the Martinez' group at BBB. Thanks to MajLill and Khahn for sharing "an *induced fit* office from time to time" and for bridging Swedish and bergensk by bokmål. Thanks to Andrea - I really miss our chats in the cafeteria – and you, of course.

Thanks to Inge Jonassen and Willie Taylor. Some interesting new lines of thought were triggered while working with you. I am highly grateful for the short time under your supervision.

Thanks to Chandra Thompson, Matthew Betts, and Katarina Dittmar De La Cruz, and everyone else who has rotated through the Liberles' research group at various points, both at CBU and at UW.

Thanks to Randy Lewis and everyone in the Spider silk - CDW group at UW for including me in the group. It has been interesting learning about the spiders (I still find them somewhat scary, though), the goats, and your cool projects.

Thanks to all my old science teachers throughout the years, especially Thorsten Dahlin and Bengt Pettersson, who both inspired me to go down this route.

Thanks to my friends, especially Cia –for being here, and Eva-Karin – for having extraordinary timing in arbitrary events. It is rather amazing.

Thanks to my brothers Henkan – for all your wise words, and Charlie – for your never ending enthusiasm and optimism. Thanks to my dear husband David for being my Socrates. You have helped my finding my way and to learn that there are many different ways to do things. There is a Swedish way, an American way, a Norwegian way, her way, his way, a fast way, a slow way, good way, a fun way, a long way, and a short way, and many, many other ways. Thanks to my parents for keeping a safe haven when the crossroads are too many – there is only one way back home.

I am forever grateful to my two boys, Benjamin and Nathan. I know it hasn't always been easy for you two, and I am very grateful that you have borne through it with me working long late hours. Thanks for providing your perspective, you make my heart and soul smile. I am so glad that I have you both, you enrich my life enormously.

Last, I want to thank all of the above for contributing to making these 5 years forever memorable and for participating in making me the scientist I am today.

Abstract

In the post-genomic era, an idea of how similar the genomes of different species actually are is on the horizon. Less than 10 years ago, the human genome was estimated to encode 100000 genes. That was an overestimation, as the real number of human genes is 20000-25000. Most genes are expressed as proteins. The 3D structure of a protein is more conserved than its sequence, and therefore the structural context of protein and gene evolution must not be forgotten. By its structure, the protein can propagate its function. In the early 90's the estimated number of different protein structure classes, so called folds, was predicted to be about 10000. Today there are slightly above 1000 folds and the discovery of new folds has leveled off, despite an increase in the number of protein structures that have been solved over the last few years. Indeed, some folds are used for more than one function, and found in various functional contexts. Then, if the many components are so similar, how is the biological species divergence from same component genomes achieved? One way to study biological diversity is by dividing it into its smaller components, e.g. by studying protein or gene family evolution. Here the evolution and regulation of the aromatic amino acid hydroxylase (AAAHs) have been under examination. This gene family encodes the proteins phenylalanine hydroxylase (PAH), tyrosine hydroxylase (TH), and tryptophan hydroxylase (TPH). These enzymes are highly physiologically important. PAH, expressed in liver, regulates the homeostasis of L-Phe by hydroxylating it into L-Tyr. TH, expressed in the central nervous system, hydroxylates L-Tyr into L-Dopa. L-Dopa is part of two important pathways i)

melanogenesis and ii) dopamine production. In humans, dysfunctions in PAH that cause elevated L-Phe concentration can result in phenylketonuria (PKU). Untreated PKU results in neurological damage. TPH produces a precursor of serotonin from L-Trp. The end products of these enzymes are neurotransmitters and hormones with increasingly important functions, from e.g. amoeba to nematode to man. As PAH has evolved in mammals its regulation has become increasingly sophisticated, e.g. homotropic positive cooperativity that shifts the conformational equilibrium from dimeric to tetrameric is seen in the mammalian lineage. Nematode PAH is devoid of positive cooperativity, but resembles the tetrameric high-affinity and high-activity mammalian PAH. TH and TPH are always tetrameric and not allosterically regulated. Each AAAH subunit has a regulatory domain, a catalytic domain, and an oligomerization domain. The promotion of positive cooperativity in PAH has been investigated by comparing mammalian PAH to nematode PAH. The low-affinity and low-activity dimer as well as the high-affinity and high-activity tetramer of PAH were modeled. Sequence analysis on a nematode sequence cluster and a mammalian sequence cluster identified sites with high probability of being involved in functional divergence, e.g. change in regulation. Residue specific electrostatic interaction energies were calculated for all ionizible residues in the models. In general, we note important differences in the substrate binding pocket that aids to explain why the active site in nematode PAH is less dynamic than in mammalian PAH. Our results suggest a pathway for the positive cooperativity from one active site to another, involving various predicted hinge regions from human PAH, where we find the nematode PAH more rigid.

The regulatory domain in PAH is part of the ACT domain family. The ACT domains are frequently found regulating metabolic enzymes in an allosteric manner. The allosteric effector is often an amino acid that binds to an interface formed by two ACT domains. No contacts are formed between two ACT domains and the stoichiometry of binding is 1:1 for L-Phe in PAH. Therefore the allosteric effect must originate in the active site when the substrate binds. An alternative pathway for aromatic amino acid biosynthesis is present in e.g. plants and bacteria. This pathway has an L-Phe binding ACT domain, which is homologous to the ACT domain in AAAH. The L-Phe binding motif in this domain is also conserved in PAH. A comparative structural analysis of this area shows why L-Phe may not bind in the AAAH regulatory domain and also indicates why it has remained.

The ACT domain has an abundant fold, a superfold. A structural approach was used to identify more potential ACT domains to gain further insights to the functional properties that this domain could perform in general, and in PAH in particular. Here we note e.g. two interesting potential domain families that could be homologous to the ACT domain, namely the GlnB-like domains and heavy metal binding domains.

The phylogeny of the AAAH family has not been resolved earlier given the lack of a suitable outgroup. As more genome sequences became available, we identified an outgroup candidate and had it experimentally characterized. The phylogeny was resolved, the ancestral function determined, and by comparing the chromosomal gene locations the order of events in AAAH evolution was envisioned.

List of publications

Siltberg-Liberles, J., Steen, I. H., Svebak, R. M. & Martinez, A. (2008), "The phylogeny of the aromatic amino acid hydroxylases revisited by characterizing phenylalanine hydroxylase from Dictyostelium discoideum.", GENE, doi:10.1016/j.gene.2008.09.005. *In press*

Siltberg-Liberles, J. & Martinez, A. (2008), "Structural determinants of the regulatory properties in phenylalanine hydroxylase," Manuscript. *To be submitted.*

Liberles, J. S.*, Thorolfsson, M, & Martinez, A. (2005): "Allosteric mechanisms in ACT domain containing enzymes involved in amino acid metabolism.", Amino Acids, 28:1-12.

Siltberg-Liberles, J. & Martinez, A. (2008): "Searching distant homologs of the regulatory ACT domain in phenylalanine hydroxylase.", Amino Acids, doi:10.1007/s00726-008-0057.2. *In press*

* S as is Siltberg

Contents

Abbreviations

AAAH	Aromatic Amino Acid Hydroxylase
AIC	Akaike statistical test
BH_4	Tetrahydrobiopterin
BLOSUM	BLOck SUbstitution Matrices
CE	Combinatorial Extension
DH_4	Tetrahydrodictyopterin
DictyoPAH	*Dictyostelium discoideum* PAH
FATCAT	Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists
h	Hill coefficient
IARS	Intrinsic Auto Regulatory Sequence
JTT	Jones, Taylor, and Thornton
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
MUSTANG	MUltiple STructural AligNment AlGorithm
NJ	Neighbor Joining
PAH	PhenylAlanine Hydroxylase
PAM	Point Accepted Mutations
PDB	Protein Data Bank
PKA	cAMP dependent Protein Kinase A
PKU	PhenylKetonUria
P-protein	bifunctional chorismate-prephenate dehydrogenase
RMSD	Root Mean Square Deviation
SSAP	Sequential Structure Alignment Program

TH	Tyrosine Hydroxylase
TPH	TryPtophan Hydroxylase
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
3PGDH	D-3-PhosphoGlycerate DeHydrogenase

1. General introduction

1.1. Evolution

The Earth is about 4.6 billion years old. The first forms of life appeared about 0.8 billion years later (Mojzsis et al. 1996; Holland 1997). Since then Nature has created an enormous biological diversity. The essence in this biological diversity originates in DNA. DNA carries all the essential information for a specific phenotype of a species in the genome. This information is carried on to the species offspring. However, genomes are always subjected to change and this brings on the evolution of new traits and, eventually, new species. Following is a brief introduction to how different life forms may have arisen and how that has affected e.g. the evolution of human kind.

1.1.1. Evolution of different life forms

Prokaryotes (Eubacteria and Archeabacteria) represent the first forms of life. Unicellular eukaryotes, also known as protozans, probably emerged from prokaryotic ancestry about 1.6 - 2.1 billion years ago (Knoll 1992), while multicellular eukaryotes, also known as metazoans (animals) and metaphytes (plants), originated

about 0.6 billion years ago. The period leading from unicellular to multicellular eukaryotes is often referred to as the Cambrian explosion, when extensive radiation took place (Lundin 1999). Many explanations have been put forward to account for this transition such as an increase in oxygen concentration in the atmosphere surrounding the Earth (Canfield and Teske 1996). As anaerobic metabolism has low energy yield it could only support low biomass, making aerobic metabolism a requirement for increasing biomass as seen in multicellular eukaryotes (Fenchel and Finlay 1994). Another explanation for the extensive radiation is cropping as a feedback cycle, involving the diversification of prey, giving rise to the diversification of predators. Such a scenario could make the rise of exoskeleton a potentially important event (Stanley 1973). Now in the post-genomic era we are starting to see the genomic effects of the burst in evolution taking place in the Cambrian era and, regardless of how and why this explosion took place, there were extensive gene duplications and domain rearrangements during that period, separating protozans from metazoans (Ekman et al. 2007). Another burst of radiation coupled to two rounds of whole genome duplication occurred during the chordate-vertebrate transition (McLysaght et al. 2002). These events are marked in Figure 1, which contains the species relevant to comprehend this thesis. The branches where duplication events and morphological innovation have taken place are marked. How gene duplications affect a genome are discussed below, but first a brief introduction to protein evolution.

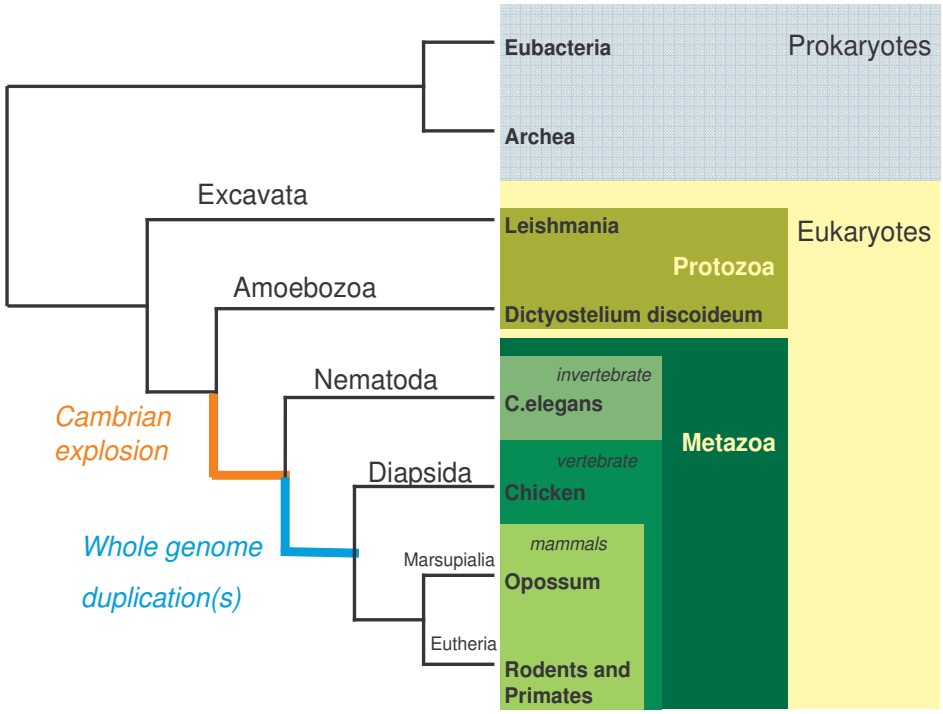

Figure 1. A simplified tree of life, representing the species and major gene duplication events discussed in this thesis.

1.1.2. Proteins and their evolution

Proteins are expressed and encoded by the genes in a species DNA. The protein alphabet consists of 20 amino acids. The amino acids have different physicochemical characteristics and are often divided into clusters given these. For an example of the complexity in these clusters see Figure 2.

As a consequence of their physicochemical characters the different amino acids have different propensities to form secondary structure elements. The two major secondary structure motifs are the α-helix and the β-sheet (Figure 3).

Given the properties of the different amino acids and their succession in a protein sequence, the protein can rapidly fold into a distinct, thermodynamically

stable structure. One intensely studied question is how a protein can fold so quickly and effectively. Since there are too many different conformations for a protein chain to be tested out by the protein one by one, how can the protein fold in just seconds or less, are there pathways for folding? This question was first asked by Levinthal in 1968 (Levinthal 1968), and since then different models have been put forward -e.g. the nucleation-condensation mechanism- where a protein folds by first forming important contacts between residues involved in the so called folding nucleus (Abkevich et al. 1994), the hydrophobic collapse (Rackovsky and Scheraga 1977), or the (rugged) folding funnel (Wolynes 2007). Importantly for these models is that they follow a two-state folding scenario, where the free energy difference between the unfolded state and the folded native state is referred to as the folding free energy change.

There has been however an increasing attention towards the existence of proteins with low-energy barriers for protein folding, which is associated to a down-hill folding scenario. During folding these proteins cross a thermodynamic barrier low enough to produce significant deviations from two-state folding behaviour (Garcia-Mira et al. 2002; Sadqi et al. 2006). Recently it has been shown that α-Lactalbumin that is a paradigm in protein folding presents a surprisingly low folding barrier, which appears to arise from the stabilization of partially unfolded conformations by electrostatic interactions (Halskau et al. 2008).

Many encoded proteins contain more than one foldable unit (Apic et al. 2001a). Protein structural units that independently of other such units can fold into a distinct, compact, and stable structure are referred to as protein domains.

Protein domains are the building blocks that in different combinations can perform highly diverse functions. The 3D structures of these domains are called folds. There are about 1000 folds, currently known. These folds are found in a highly skewed distribution, where i) *superfolds* are a small number of highly abundant folds found in many different proteins which are able to perform a vast variety of functions, ii) *common folds* are an increasing number of folds, shared by a few different proteins, and iii) *unifolds* are most folds but these are only found within one protein (Coulson and Moult 2002). Here protein actually means protein family, which will be further discussed below.

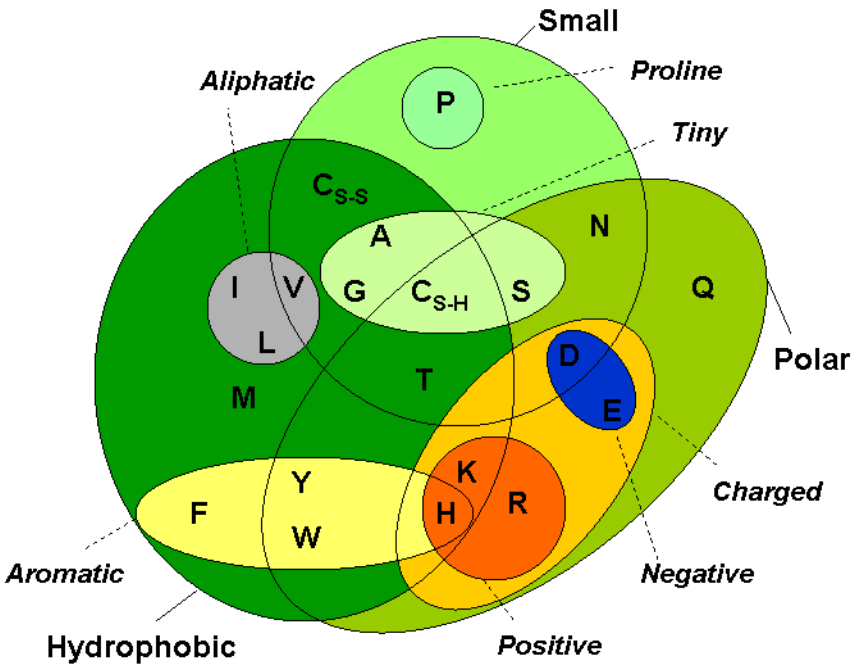

Figure 2. The 20 amino acids in the protein alphabet have different properties and can be divided into clusters according to different criteria, e.g. small, polar, hydrophobic, etc. The amino acids are named according to the 1-letter code. Figure inspired by (Livingstone and Barton 1993).

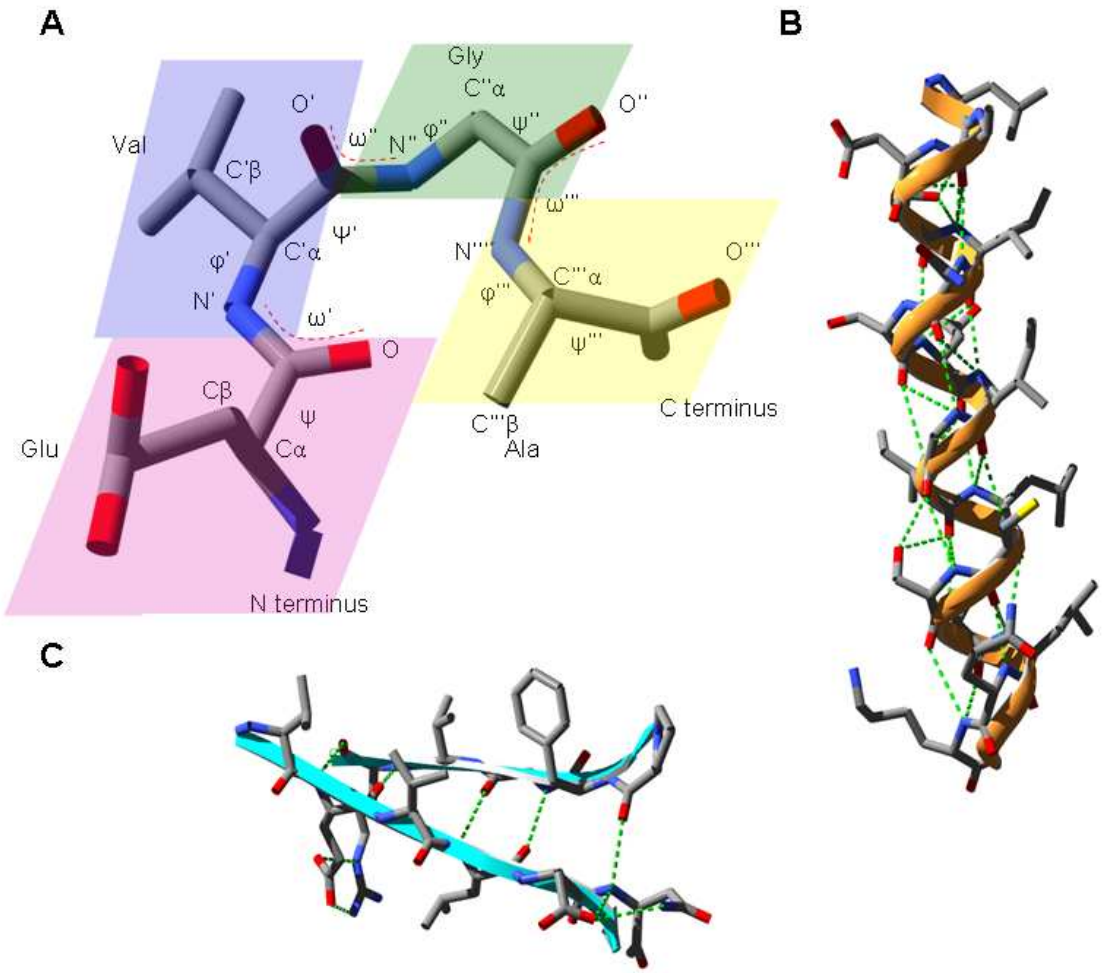

Figure 3. Introduction to protein structure. A short peptide of four amino acids (A). From the N terminus: Glu (purple background), Val (blue background), Gly (green background), and Ala (yellow background). Each amino acid has an amino group (N terminus) and a carboxyl group (C terminus) (C). The amino acids are connected to each other by forming a peptide bond. The atoms forming the peptide bond are the carbonyl group (red) from aa_i and the amino group (blue) from aa_{i+1}. The peptide bond (red dotted line in (A)) has double bond character and is strong and almost planar. The backbone of a protein also includes the Cα-carbons, to which the different side chains are connected by the Cβ-carbon (A). By combining the amino acids in different ways the φ and ψ angles will vary (A). Depending on φ and ψ, higher order structures, so called secondary structure elements, α-helix (B) and β-sheet (C), can form. Figure generated in DeepView (Guex and Peitsch 1997) and rendered by POV-Ray (POV-Ray).

The natural question to ask here is how can the superfolds be explained? Are all proteins that display the same fold related, or are they a random sampling of folds?

Has the same fold evolved multiple times? Why are just these folds so abundant? This is a topic of debate, where some say that these energetically stable superfolds can harbor high sequence divergence while maintaining the fold (Shakhnovich et al. 2003), while others suggest that evolution has converged on these folds many times (Marsden et al. 2006). However, many shared physicochemical interactions and smaller substructural and sequence motifs as well as similarities in functions indicate that the domains sharing a fold also have common ancestry (Saraste et al. 1990; Koonin 1993; Kiel and Serrano 2006). Two important mechanisms that enable sequences to diversify and proteins to attain new functions are gene duplication and different multidomain combinations.

1.1.3. Gene duplication – a driving force for new protein functions

Sasumo Ohno postulated in 1970 that gene duplication is required for the evolution of new gene/genome functions (Ohno 1970). Gene duplications may be small scale events, e.g. duplicating a single gene or chromosome, but also whole genome events have occurred. After a complete gene has been duplicated, there is gene redundancy in the genome. Given the gene redundancy, the functional constraint on that gene is relieved. Hence, the gene duplicates can explore more of sequence space as there now is a back up for its function. This effect is even larger in the case of whole genome duplications which let entire pathways evolve with relieved functional constraints (Roth et al. 2007). The possible scenarios for the duplicated gene copies include pseudogenization, neofunctionalization, subfunctionalization (Ohno 1970; Hughes 1994; Force et al. 1999; Ohno 1999), dosage compensation (Birchler et al. 2005), and genetic robustness (Gu et al. 2003).

In eukaryotes the complex gene structure with introns and exons also constitutes a mechanism for altering function by e.g. changes in gene expression and alternative splicing (Lynch 2006).

Pseudogenization

Pseudogenization is the most common fate for one of the gene copies after gene duplication. By fixation of a null mutation one copy is non-functionalized. Eventually a pseudogenized gene will no longer be recognizable in the genome.

Neofunctionalization

The gene redundancy following a gene duplication allows the two gene copies to explore new functions. However, as one copy is devoid of the original function while attaining a new function, the functional redundancy is lost and the remaining copy must maintain the original function. The neofunctionalization mechanism is a way for both copies -with new, old, or slightly modified functions- to be fixed within the genome.

Subfunctionalization

Subfunctionalization also provides a mechanism for both duplicated gene copies to be retained in the genome. As many genes encode multidomain proteins it is intuitive to imagine that different domains from the two duplicate genes, or proteins, may interact to perform the original function. It has recently been shown that subfunctionalization, as a mechanism to retain both duplicates, can be followed by neofunctionalization, where the redundant *domains*, instead of the redundant *gene* can explore more of sequence space.

Dosage compensation

If a gene and its regulatory region are duplicated, the expression of that gene is doubled. Gene duplication may also play a role in increasing the expression of genes where there is a selective advantage to increased expression (Wagner 2005). Dosage compensation is a mechanism to maintain a similar ratio of the different expressed genes although their copy number has been altered, as after a whole genome duplication. This effect is greater for regulatory genes, and can also be referred to as *hierarchical regulatory balance*. As selection occurs on one dosage-dependent regulator, other regulators in a balanced relationship can co-evolve. Conflicts among components of regulatory complexes within the genome could accelerate evolutionary change (Birchler et al. 2005).

Genetic robustness

Genetic robustness is another mechanism to retain both duplicated copies of a gene in a genome as a backup (Gu et al. 2003). Under this mechanism both copies are thought to maintain their original function and expression profile. Theoretical work has suggested that this mechanism is mostly applicable in species with high mutation rates and large effective population sizes (Elena et al. 2007).

1.2. Homology

Homology is one of the cornerstones in comparative biology and molecular evolution and, therefore, a brief introduction to homology and other related concepts will be given here in the light of gene duplications and speciation events. A

speciation event occurs when the genetic boundary between two sub-species becomes large enough to prevent reproduction. A *gene duplication event* occurs when there is an additional gene copy in the genome. This additional gene copy may result from an entire genome duplication or a smaller scale duplication, e.g. a chromosomal duplication.

1.2.1. Orthologs and paralogs

When a gene duplication is followed by neofunctionalization, the result is two paralogs, *A* and *B*. Upon a speciation event, gene *A* is present in the two new species, 1 and 2. Genes 1*A* and 2*A* are orthologs and have the same function. If this gene is further duplicated in both lineages these new duplicates will be paralogs to each other within each species, but orthologs to both copies in the other species (Figure 4). This can be further complicated if e.g. one of the genes is lost from one of the species, or if its sequence has diverged beyond recognition. In order to recognize protein sequences as homologs using only basic sequence analysis their pairwise sequence identity ought to be > 35 %. Pairwise protein sequence alignments in the so called twilight zone (20-35%) include a large fraction of non homologous proteins, as far as different structures are found among the pairs (Rost 1999).

Remote homologs

As sequence divergence becomes large enough for making homology identification with sequence based methods uncertain, there is another dimension of proteins to utilize. It has long been recognized that protein structure is more conserved than sequence (Zuckerkandl and Pauling 1965; Chothia and Lesk 1986).

Therefore, using protein structures aids to the identification of homologous proteins with highly diverged sequences. Different reasons for the extensive sequence divergence are given in the following sections.

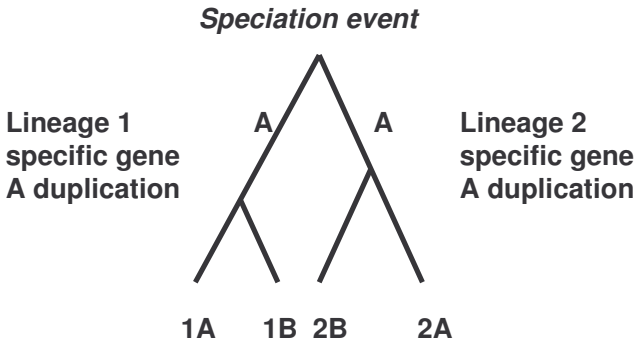

Figure 4. Paralogs are separated by a gene duplication event and orthologs by a speciation event. At the time of speciation between lineage 1 and lineage 2, both genomes had one copy of gene A. In lineage 1, gene A was duplicated. The resulting genes were 1A, which maintains the original function, and 1B that has another function. In lineage 2, gene A was also duplicated. The resulting genes were 2A, which maintains the original function, and 2B that has another function. Genes 1A and 1B are paralogs; similarly genes 2A and 2B are also paralogs. In all other comparisons within this group of genes are the genes orthologs.

1.2.2. Protein structure comparisons

As remote homology detection can gain from comparing structures, homologous proteins can benefit from having their structures compared. Although it is generally true that structure is more conserved than sequence (Zuckerkandl and Pauling 1965; Chothia and Lesk 1986), there are exceptions. Not all homologous proteins have the same structure (Grishin 2001). The process by which homologous

proteins acquire significantly different structures is called neostructuralization (Liberles 2005), but how often it occurs remains to be investigated.

Structural similarity

As discussed above many folds, with the same order of secondary structure elements and the same connectivity, are reoccurring in different proteins, with different functions, and with very different sequences. Most methods have their own scheme for scoring the obtained similarities for two aligned structures. These scores are often based upon sequence identity, Root Mean Square Deviation (RMSD), alignment length, and number of gaps in the alignment. RMSD is a very important factor for the comparison. It is calculated from the coordinates of different equivalent atoms in the two structures and indicates how similar the structures are, regardless of homology. The RMSD can be calculated on just the Cα carbons, the peptide backbone, and, in some cases, the side chains are included in the calculation.

1.2.3. Protein fold distribution

Some authors estimate that most of protein fold space is covered by the currently known 1000 folds (Taylor 2007; Goldstein 2008), while others estimate that there are many unique folds that remain to be detected (Coulson and Moult 2002). Over the last few years, although the number of experimentally determined structures has rapidly increased, the number of folds is almost constant (PDB statistics). The distribution of folds among protein families is highly skewed. Some folds are very abundant and found in many different contexts with a variety of functions, and with sequences that have diverged beyond recognition of each other. Other folds are

unique to one particular family, and some folds are something in between (Coulson and Moult 2002).

Different explanations to the fold distribution have been put forward. One explanation is designability. Designability is a measurement for how many sequences that can fold into a certain structure (Buchler and Goldstein 1999). Some folds have high designability which means that their sequences are robust to substitutions while still maintaining their folds (Melin et al. 1999; Zeldovich et al. 2006). The reason for the highly designable folds is very frequent and long range contacts (England and Shakhnovich 2003). An alternative or perhaps complementary view was asked; are some folds overrepresented due to the functions they can provide (Goldstein 2008)? Further, the role of evolutionary dynamics in the fold distribution may also be important. It has been shown, using lattice-models, that for a large protein population with high mutation rate, the evolving population is polymorphic in stability and subjected to frequent mutations, so the more stable and thus more mutationally tolerant proteins will produce more folded offspring. For a small population with low mutation rate, the evolving population is monomorphic in stability so all members of the population are equally likely to produce foldable offspring (Bloom et al. 2007). It seems likely that one or various combinations of designability, functionality, and evolutionary dynamics are needed to explain the fold distribution.

1.2.4. Multidomain proteins

As mentioned above, protein domains are the evolutionary units, which in different arrangements can perform different functions. However, it must be noted that many protein domains are found as stand-alone one domain proteins.

Evolution of multidomain proteins

By combining domains in different genes to be expressed as proteins, domains are the building blocks of diversity. Multidomain proteins are the result of different domain combinations. In metazoans > 80% of all proteins are multidomain proteins, while about two-thirds of proteins in unicellular organisms contain more than one domain (Apic et al. 2001b; a). Multidomain proteins are formed by exon shuffling, recombination, fusion, and fission of genes (Long et al. 2003). One of the major contributions to the larger fraction of multidomain proteins in metazoans is exon shuffling of exon bordering domains (Ekman et al. 2007). Regardless of mechanism, a domain is frequently added at the N-terminus, or at the C-terminus, of a current protein that already has one or several domains to create a new multidomain protein (Ekman et al. 2007). For the aromatic amino acid hydroxylase (AAAH) family, its different domains are all present in some, but not all, bacteria. For many bacteria and all archea, the catalytic domain is not found within their genomes. Most protozoan (unicellular eukaryotes), and all plantae (non-animalia multicellular eukaryotes) lack the catalytic AAAH domain as well. However, as more protozoan genomes sequencing projects are finished, species with solely one full-length AAAH are identified, e.g. *Dictyostelium diccoideum*, the *Leishmania* species, and *Monosiga brevicollis*. Metazoans (multicellular animals) have at least three different full-length AAAH genes.

1.2.5. Protein domain organization in databases

Regardless of how the distribution of folds is explained, there is a need to classify the protein domains found with different or similar folds. If there are about

1000 or even 10000 different structural domains (also known as topologies or folds), comparing the number of folds to the number of genes in the human genome, which has about 20 000 -25 000 genes - not even counting the number of different domains, indeed reveals that some folds must be represented more than once. However, as seen above, many proteins are multidomain proteins and therefore a first step is to identify the domain boundaries in different protein structures. Once domains are determined the classification can begin. There is a hierarchical organization of folds and domains, and to further group them into families based upon their sequence similarities has proven very useful, as in e.g. protein structure classification databases like CATH (Pearl et al. 2003; Pearl et al. 2005) and SCOP (Murzin et al. 1995). Following the CATH hierarchy, proteins are classified into *Class* - based upon its secondary structure elements, *Architecture* – based upon the direction of the secondary elements, *Topology* – based upon the direction and connectivity of the secondary elements, and *Homologous superfamily* – based upon proven common ancestry. The homologous superfamily is further divided into sequence families (Pearl et al. 2003; Pearl et al. 2005). The number of homologous superfamilies found for a particular fold is what first led to the identification of the superfolds. In 1994, the CATH database contained 9 superfolds found in more that 10 different homologous superfamilies (Orengo et al. 1994). The three most common superfolds, the Rossman fold, the ferredoxin-like fold, and the immunoglobulin-like fold (Figure 5A-5C) are today present in over 110, 80, and almost 70 different homologous superfamilies, respectively. The TIM barrel (Figure 5D) is another superfold, commonly used fold for different enzymatic functions.

A prevalent database for protein sequence families is Pfam. Pfam includes the entire domain sequence and not just conserved motifs. It is therefore representing biologically meaningful sequence annotation instead of just short sequence motifs alone (Sonnhammer et al. 1998). A recent feature of Pfam is Pfam clans. In a Pfam clan, two or more Pfam families that have arisen from a single evolutionary origin can be found (Finn et al. 2006). The latest version of Pfam had more than 9300 protein families (Finn et al. 2008). This discrepancy between the 1000 known folds and almost 10 times as many protein families indicates that sequence divergence is a major factor for protein family distinction, but what drives it beyond recognition?

1.2.6. Functional divergence on a sequence level

As protein sequences start to diverge there are a few requirements that must be fulfilled. For instance, the original function must be maintained or improved for a substitution to be fixed with high probability. This function may be catalytic activity, agonist binding, or interactions with other proteins. Further, to maintain its structural integrity, the protein must also maintain its ability to fold. As described above high designability of a fold allows its protein sequences a higher potential to evolve beyond sequence recognition than for the sequences of a fold with low designability.

Early models describing sequence divergence simply divided amino acid sites into two classes, one with the potential to change and one that could not change. When different substitutions became fixed the sites in the class with the potentially varying sites would differ. As sequence space starts to be explored one substitution changes the probabilities for substitutions in the other positions (Fitch and Markowitz 1970) This is called covarion behavior. The group of H. Philippe described

heterotachy as an important factor for sequence divergence (Lopez et al. 2002) . Heterotachy means that the evolutionary rates for all sites in the protein can change over time, where the covarion behavior represents a subclass of heterotachy (Lopez et al. 2002).

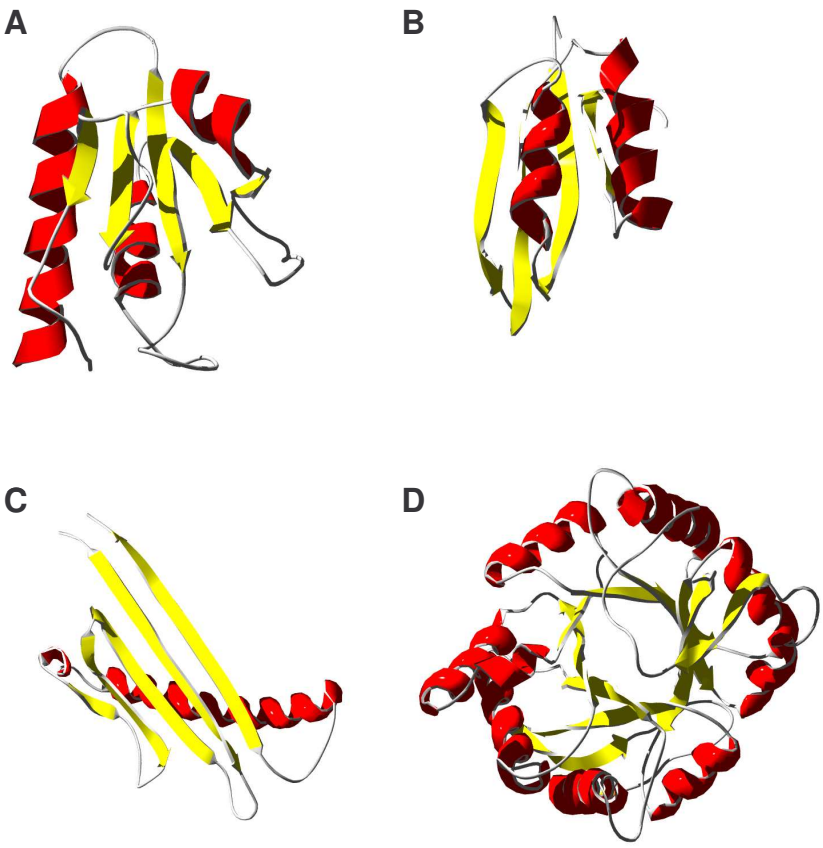

Figure 5. Cartoon representation of the three most common superfolds, the Rossman fold (A), the ferredoxin-like fold (B), and the immunoglobulin-like fold (C), together with a superfold found in many catalytic domains, the TIM barrel fold (D). The secondary structure elements are α-helix (red), β-strand (yellow), loops and unstructured regions (grey). The N-terminus to C-terminus direction is given by the arrows at the C-terminus end of the β-strands. PDB id's used 1DLJ (Campbell et al. 2000)(A), 1PHZ (Kobe et al. 1999) (B), 1K5N (Hulsmeyer et al. 2002) (C), and 1XX1 (Murakami et al. 2005) (D). Figure generated in DeepView (Guex and Peitsch 1997) and rendered by POV-Ray (POV-Ray).

Further, it should also be mentioned that sequence divergence is highly context dependent as interacting proteins and domains are evolving together (Lopez et al. 2002).

Superfold domains evolve in different environments (e.g. the surrounding domains) and this makes their sequences diverge further. As the same domain re-occurs within the same genome or even within the same protein, the different copies of it will start to diversify. It has been found that two similar domains located next to each other in a protein sequence will rapidly diversify to avoid unwanted contacts or aggregation (Wright et al. 2005). A common way to form aggregates is by runaway swaps (Bennett et al. 2006). Runaway swaps are different than domain-domain swapping, which is a common mechanism for multidomains to dimerize (Kinch and Grishin 2002). Runaway swaps are frequently found in deposition diseases, as e.g. in prion disease (Bennett et al. 2006). Runaway swaps are the result of e.g. a two domain protein where the two domains, *A* and *B*, are connected by a hinge region. If the coordination of the hinge region changes, this opens up for dimer formation, with domain *A* from one chain and domain *B* from the other chain. If domain *B* from the first chain and domain *A* from the second chain also form the same contact it can prevent a deposit to form (the classic domain-domain swap). If not, runaway swaps may be formed, as domain *B* from the first chain and domain *A* from the second chain now are open binding sites where another similar chain can attach and form higher order structures (Bennett et al. 2006). A similar effect is likely to occur for the protein folds that are found in many different domains, as e.g. the superfolds. Given their abundance, the superfold domains may further drive their sequence divergence to

similar superfold domains as they have a tendency to be promiscuous and form unwanted contacts by e.g. inter domain-domain swapping.

The driving force for protein designabilty is the stability of its folded state relative to the unfolded state. Proteins have been found to be marginally stable. How does marginal stability correspond to sequence divergence?

1.3. Marginally stable proteins

1.3.1. Proteins are marginally stable

Proteins are not rigid molecules. They are marginally stable and Fersht showed that their folding free energy (ΔG_f) is between -5 and -15 kcal/mol (Fersht 1999). The stability (or instability) of folding is due to the contributions from the hydrophobic effect, hydrogen bonding, packing interactions between buried residues, and also electrostatic interaction energies between surface accessible residues, both in the unfolded and the folded states. This indicates that there is a delicate balance of stabilizing and destabilizing interactions which may be important for the protein's biological function, but is this balance driven by nature's selection for marginally stable proteins, or are proteins marginally stable as an artifact of their design?

Why are proteins marginally stable?

The classic view of protein evolution is that proteins are marginally stable because there is a tradeoff between stability and flexibility. Therefore evolution selects for proteins that are stable, but not too stable. The explanation is often that

proteins need their flexibility to perform a certain function. Most mutations will have large effects on the structure and/or dynamics of the protein. Deleterious effects of mutations will be compensated for by conditionally beneficial mutations, as part of adaptive evolution (see (DePristo et al. 2005) and refs. therein). An alternative view of why proteins are marginally stable is given from a more evolutionarily neutral point of view. Here the view is that most substitutions are neutral and do not alter structure, function or stability. There is no selection or thermodynamic reason that proteins cannot be more stable; that they are not is an artifact of the evolutionary process, since destabilizing mutations are more common than stabilizing ones in the folded protein (Taverna and Goldstein 2002). This is related to designablity and evolutionary dynamics as discussed above.

Proteins are dynamic

The marginal stability in (non-fibrous) proteins means that the energy gap to slightly altered conformations is small. All together the result is that many proteins are highly dynamic. These dynamic properties can result in equilibrium of different conformations for many proteins. Dynamic proteins can explore a wide range of different conformations. For many dynamic proteins there is coupling between at least two different conformations, e.g. ligand-bound and ligand-unbound conformations, which in many cases are provided by widely distributed binding sites.

1.3.2. Allostery

Allostery is a term borrowed from Greek, where *allos* means other and *stereos* means solid or space. The central dogma of allosterism is that as an effector binds at

one site and it causes a conformational change at another site. Allostery was first defined as the regulation of a protein by a small molecule that differs in shape from the substrate, but was later redefined to include regulation of a protein through a change in its quarternary structure as induced by a small molecule, including the substrate (Monod et al. 1963). In enzyme terminology allostery is often associated with oligomeric enzymes showing variable ligand binding affinity that enhances or depresses catalytic activity. This is referred to as cooperativity (Fersht 1984). When the ligand binding affinity increases, the allosteric behavior is called positive cooperativity, and similarly, negative cooperativity is when the affinity decreases. The effect of cooperativity is measured by a Hill plot. The Hill plot describes the binding of ligands to allosteric proteins in the region of 50 % saturation, where the value of the Hill coefficient (h) is given from the slop of the Hill plot. If h=1 there is no cooperativity; if h>1, there is positive cooperativity; and if h<1, there is negative cooperativity (Fersht 1984). The substrates and other effectors that cause the cooperativity by binding at the substrate binding site are homotropic allosteric modulators. Homotropic modulators typically activate the enzyme, as in e.g. positive cooperativity, where they can bind the active site of one subunit in an oligomer increasing the affinity and activity at the other active sites. Heterotropic allosteric modulators are effectors of the allosteric response, which bind to a site different to the active site where they modulate the substrate affinity and activity. Heterotropic modulators can be inhibitors or activators. The two classical mechanisms commonly used to explain allostery are the concerted mechanism (Monod et al. 1965) and the

sequential mechanism (Koshland et al. 1966). It is becoming evident that these two models are not able to explain all the different types of allosteric behavior seen today.

The concerted mechanism

The concerted mechanism involves a preservation of the symmetry in the quaternary structure of the protein. This mechanism is appropriate for oligomeric proteins that are present in two different conformations: (1) the unactivated conformation of the protein, without the allosteric effector bound -called the T (tense) state- and (2) the ligand-activated conformation -called the R (relaxed) state. The T state must have lower ligand affinity than the R state. Last, if all binding sites in the T and R states, respectively, are equivalent, then this is a concerted mechanism (Monod et al. 1965).

The sequential mechanism

The sequential mechanism for allosteric behavior is based on a gradual change from the T state to the R state. The two underlying assumptions for the sequential model are that the protein exists in one conformation, prior to ligand binding. Upon binding the ligand a local change occurs in the subunit where the ligand is bound. However, the effect can be transmitted to other binding sites at the adjacent subunits by changes at the subunit interfaces (Koshland et al. 1966).

The expanded view

The thermodynamic understanding of proteins is consistent with an equilibrium model of different conformations and that allosterism can drive the increase of one conformation at the cost of another (Tsai et al. 2008). However, it is

necessary to note that conformations seen in allosteric proteins are present as populations which coexist in equilibrium. When a ligand binds to a ligand binding site it can redistribute the population (Gunasekaran et al. 2004). Allosterism is not exclusive of multidomain proteins, but also single domain proteins can show allosteric behavior (Gunasekaran et al. 2004; Leiros et al. 2007). It is also becoming evident that allostery does not need to involve a major conformational change (Tsai et al. 2008).

1.4. Phenylalanine hydroxylase and its homologs

Phenylalanine hydroxylase (PAH, EC 1.14.16.1) is a member of the tetrahydrobiopterin dependent AAAHs, together with its homologs, tyrosine hydroxylase (TH, EC 1.14.16.2) and trypthophan hydroxylase (TPH, EC 1.14.16.4). These enzymes have high physiological importance in mammals. PAH, which is mainly found in liver, catalyzes the first step in catabolism of L-Phe, by hydroxylating L-Phe into L-Tyr. Dysfunctions of PAH that cause elevated L-Phe concentrations result in hyperphenylalaninemia and phenylketonuria (PKU). TH, which is found in the central nervous system and the adrenal gland, catalyzes the first step in the formation of catecholamines by converting L-Tyr into L-Dopa. L-Dopa is decarboxylated by the aromatic L-amino acid decarboxylase (EC 4.1.1.28) and dopamine is the product. Dopamine can further be converted into norepinephrine and epinephrine, and these are all important hormones and/or neurotransmitters. Together with tyrosinase, PAH and TH are also important for melanin formation (Schallreuter

et al. 2008). The reaction catalyzed by TPH is the first and rate-limiting step in the biosynthesis of serotonin and the initial and uncommitted step in the synthesis of melatonin, by hydroxylating L-Trp into 5-hydroxytrypthophan (5-HTP). Serotonin has many functions as e.g. it is involved in smooth muscle contraction in intestine (Erspamer and Asero 1952) and in controlling mood changes such as depression and impaired cognitive function (Owens and Nemeroff 1994). Melatonin is important in regulating the circadian rhythm. Dysfunctions in melatonin production have been shown in sleep disorders, Alzheimer's and Parkinson's disease, depressive disorders, and various cancers (Pandi-Perumal et al. 2008). There are two tissue specific forms of TPH; TPH1 is found in the gut, pineal gland, spleen, and thymus, while TPH2 is found in the brain stem (Walther and Bader 2003). Most aspects of the AAAHs are conserved among them, with the main exception being their different substrate specificities and regulation.

1.4.1. Domain composition

The AAAHs include three domains. Starting from the N-terminus these domain are, a regulatory ACT domain (in PAH residues 30-111), a central catalytic domain (in PAH residues 112-408), and a C-terminus oligomerization domain (in PAH residues 409-452) (Figure 6). In addition, the first sequence stretch in the N-terminus, prior to the regulatory domain is different among the AAAHs. For PAH it is called the intrinsic autoregulatory sequence (IARS) (Teigen and Martinez 2003). The IARS has a phosphorylation site and phosphorylation is part of the regulatory mechanism for mammalian PAH (Wretborn et al. 1980; Kaufman 1993). Also TH and TPH have phosphorylation sites in their N-terminus sequence stretches. In TH,

phosphorylation decreases its inhibition by catecholamines (Ramsey and Fitzpatrick 1998), and for both TH and TPH phosphorylation regulates their interaction with 14-3-3 proteins (Kleppe et al. 2001; Winge et al. 2008) .

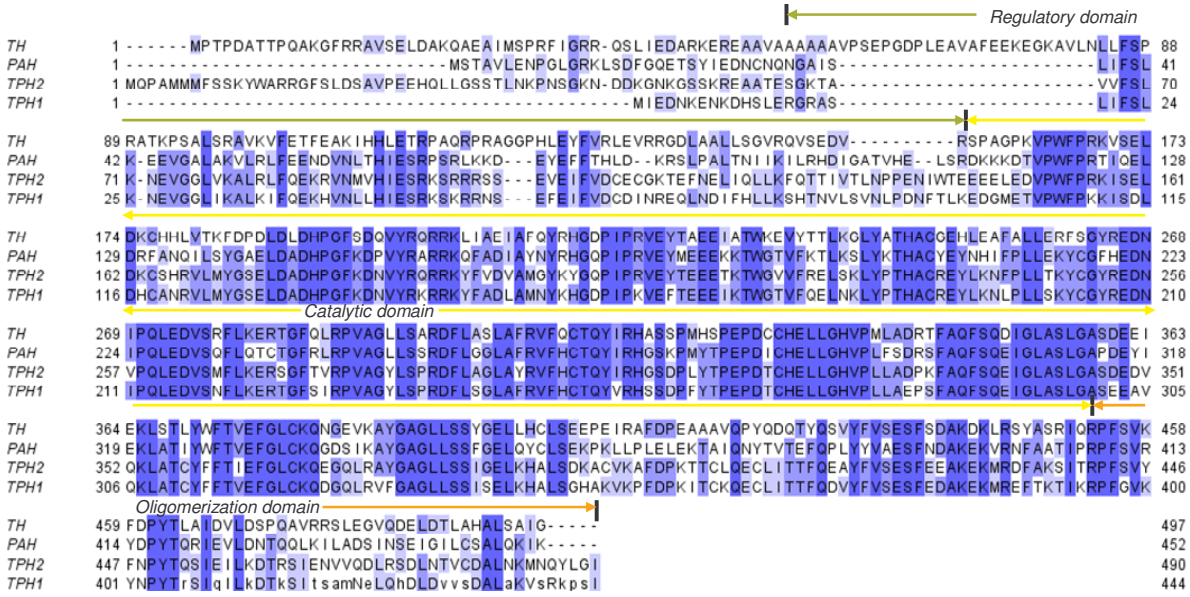

Figure 6. Multiple sequence alignment of human AAAHs. The following sequences were used and are numbered accordingly: TH (NP00531), PAH (NP000268), TPH2 (NP775489), and TPH1 (NP004170). Bars (black) denote domain boundaries and the regulatory domain (Melin et al.), the catalytic domain (yellow), and the oligomerization domain (orange) are shown. The sequences are colored by conservation.

The regulatory domain

The regulatory domain in PAH is classified as an ACT domain (Kobe et al. 1999). The initial classification of the ACT domain included the regulatory domain of PAH and TPH, but not the regulatory domain of TH (Aravind and Koonin 1999). The sequence divergence of this domain in the different AAAHs is very high (Figure 6) and this is probably related to the superfold topology of this domain. The ACT domains have highly abundant 2-layer α/β plaits topology (Figure 5, 7A) (also called

the ferredoxin-like fold), with an anti-parallell β-sheet with two α-helices on top. There are more than 70 different homologous superfamilies with this fold in the CATH database. ACT domains are, according to their definition, regulatory domains that form dimers and bind amino acids at the dimer interface. They are often found to modulate the activity of other proteins or domains e.g. in allosteric enzymes. Now, this definition does not entirely apply to the ACT domain in mammalian AAAH. No ligands are known to bind to the regulatory domains of these enzymes. For PAH, it has been speculated that L-Phe could bind to an additional effector site, possibly located in the regulatory domain (Tourian 1971; Shiman 1980; Parniak and Kaufman 1981; Kaufman 1993; Kappock and Caradonna 1996) but it has been shown that the stoichiometry of L-Phe to PAH subunit is 1:1, which does not support an effector binding site in addition to the active site (Thorolfsson et al. 2002). However, L-Phe allosterically modulates the activity of PAH by positive cooperativity (Wretborn et al. 1980; Kaufman 1993; Kappock and Caradonna 1996; Knappskog et al. 1996), but the mechanism is still unclear. The regulatory ACT domain is connected to the catalytic domain by a long flexible linker sequence, but the relative orientation of these two domains is not certain since no full-length AAAH structure is available. The regulatory ACT domain is also in contact with the catalytic domain of the adjacent chain in the dimer and with the oligomerization domain of its own chain (Figure 6).

The catalytic domain

The catalytic domain has a unique fold, only found in the AAAH's (Figure 7B). As the name implies this is where the active site is located. The catalytic domains in two adjacent chains do not interact with each other. Only mammalian

PAH has an allosteric activation mechanism to our current knowledge, but for TH negative cooperativity for the binding of the cofactor (6*R*)-tetrahydrobiopterin (BH_4) has also been shown (Flatmark et al. 1999). The sequences of the different catalytic AAAH domains are highly conserved (Figure 6).

The oligomerization domain

The oligomerization domain consists of two structural motifs. From the N-terminus there is a small β-hairpin, called the β-ribbon (Figure 7C), with highly conserved sequence (Figure 6). The β-ribbon is followed by a long α-helix with rather divergent sequence (Figure 6, 7C). The β-ribbon is involved in dimer formation and the long α-helix is involved in tetramerization by forming a leucine zipper (Fusetti et al. 1998).

Domain assembly

TH and TPH form homotetramers. Mammalian PAH has been found to exist *in vitro* as a dimer-tetramer equilibrium, where the dimer represents a low-affinity and low-activity state while the tetramer represents a high-affinity and high-activity state. As the L-Phe concentration increases the equilibrium is shifted towards the tetrameric form (Doskeland et al. 1982; Martinez et al. 1995). A composite model of the full length dimer (Fusetti et al. 1998; Kobe et al. 1999) is shown in Figure 7D and of the tetramer, which is a dimer of dimers (Goodwill et al. 1997; Fusetti et al. 1998), is seen in 7E.

1.4.2. The catalyzed reaction

The enzymatic reaction is highly similar in all AAAHs, with the substrate being the only difference (Fitzpatrick 2003; Teigen et al. 2007). The reactions performed by the AAAHs are dependent upon the different substrates, the cofactor BH_4, non-heme iron, and molecular oxygen. In PAH, L-Phe is hydroxylated in the *para* position and converted to the product L-Tyr (Figure 8). In TH, L-Tyr is hydroxylated into the product L-Dopa (Figure 8). In TPH, L-Trp is hydroxylated into the product 5-OH-TRP (Figure 8). Regarding the order of reactant binding, it has been shown that the pterin cofactor binds first both in the case of TH (Fitzpatrick 1991) and bacterial PAH (Volner et al. 2003). The mechanism of reaction and order of substrate binding is however not clear for mammalian PAH and TPH. A requirement for the cofactor to bind before the amino acid substrate might explain why the AAAH are inhibited by high substrate concentrations (Fitzpatrick 2003).

1.4.3. The PKU phenotype

Elevated concentration of L-Phe causes neurological damage, if left untreated (Scriver and Kaufman 2001). The reasons behind hyperphenylalaninemia can be a dysfunction in PAH, as in phenylketonuria (PKU), or in the supply of the cofactor BH_4 (Thony et al. 2000). PKU was discovered in 1934, by the Norweigan Asbjörn Föllingsen (Folling 1934). It is as an autosomal recessive metabolic disorder. The typical PKU phenotype is often consistent with growth failure, light pigmentation, microcephaly, seizures, global development delay and severe intellectual impairment (Williams et al. 2008). Many different mutations can cause PKU and in general the mutations in human PAH result in decreased stability and misfolding of the enzyme

(Pey et al. 2007). The known mutations associated with PKU can be found on the PAH knowledge based website (http://www.pahdb.mcgill.ca/).

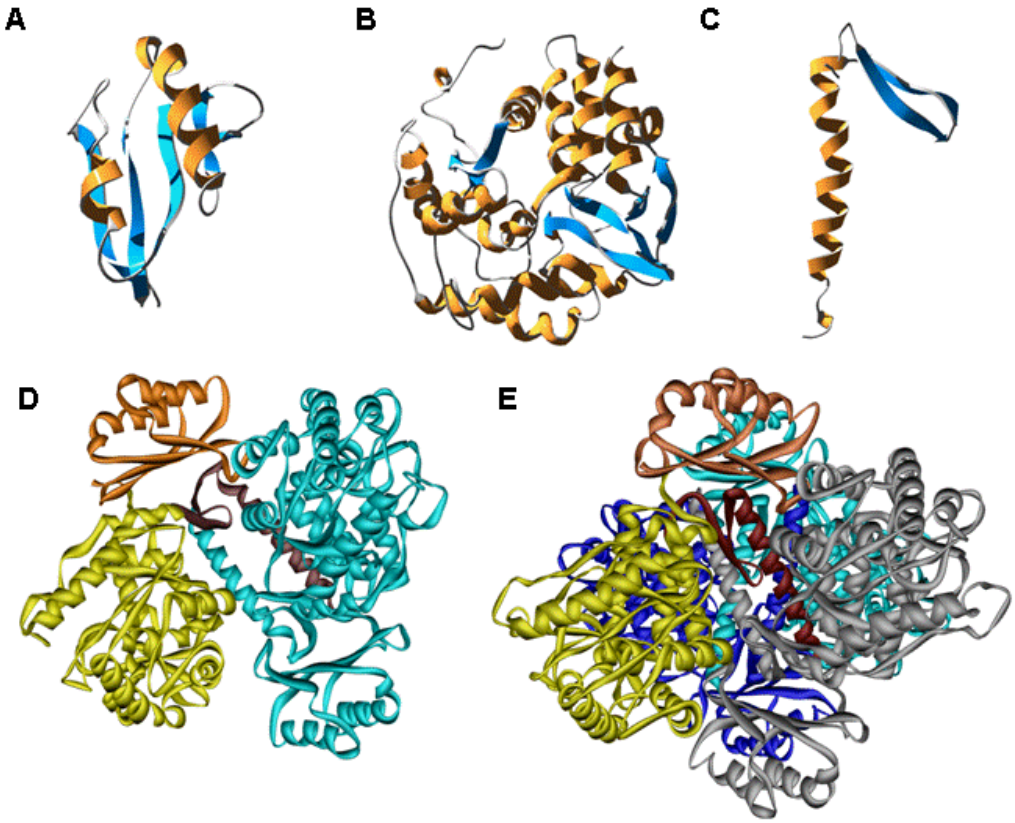

Figure 7. The domain organization in PAH; the regulatory domain (A), the catalytic domain (B), and the oligomerization domain (C), colored according to secondary structure (α-helix; orange, and β-strand: blue). The arrangement of these domains in one chain, as part of the modeled full-length dimer (D) and tetramer (E). The dimer is colored as follows; chain A: regulatory domain (orange), catalytic domain (yellow), and oligomerization domain (brown), and chain B (blue). The tetramer is colored as follows; chain A: regulatory domain (orange), catalytic domain (yellow), and oligomerization domain (brown), and chain B (grey), chain C (light blue), and chain D (dark blue). Figure generated in DeepView (Guex and Peitsch 1997) and rendered by POV-Ray (POV-Ray)

Deficient BH_4 production affects the activity of all AAAHs and can result in Parkinson's disease, autism, depression, and Alzheimer's disease (Thony et al. 2000). While BH_4 deficiency may be treatable by supplying BH_4 or derivatives thereof ((Thony et al. 2000) and refs. therein), the classical treatment for PKU patients is to exclude or minimize L-Phe and protein in their diets, and supplement the other essential amino acids. However, it has recently been shown that some forms of PKU can respond to pharmalogical doses of BH_4 by a multifactorial therapeutic effect, including a chaperone-like effect of the cofactor (Erlandsen et al. 2004; Scriver 2007).

1.4.4. PAH regulation

Positive cooperativity

As mentioned above, mammalian PAH, *in vitro*, is found as dimers and tetramers in equilibrium, an equilibrium which is driven towards the tetrameric form in the presence of L-Phe (Knappskog et al. 1996). Mammalian PAH is activated by i) its substrate, L-Phe, and ii) phosphorylation at Ser16, with these two activation mechanisms operating synergistically (Kaufman 1993; Kappock and Caradonna 1996). L-Phe binds with positive cooperativity to mammalian PAH. The *hill*-coefficient (*h*) for the cooperative binding of L-Phe to human and rat PAH is about 2.0 (Kaufman 1993; Knappskog et al. 1996; Thorolfsson et al. 2002).

However, the molecular basis for the positive cooperativity and substrate activation of this enzyme remains unresolved, partly because of the so far unsuccessful task to crystallize the tetrameric full-length enzyme in the presence and the absence of L-Phe (Flatmark and Stevens 1999). Nevertheless, site directed

mutagenesis and molecular dynamics simulations have been used to partly investigate the molecular basis for the activation of human PAH by L-Phe, and the homotropic positive cooperativity mechanism for the substrate (Thorolfsson et al. 2003).

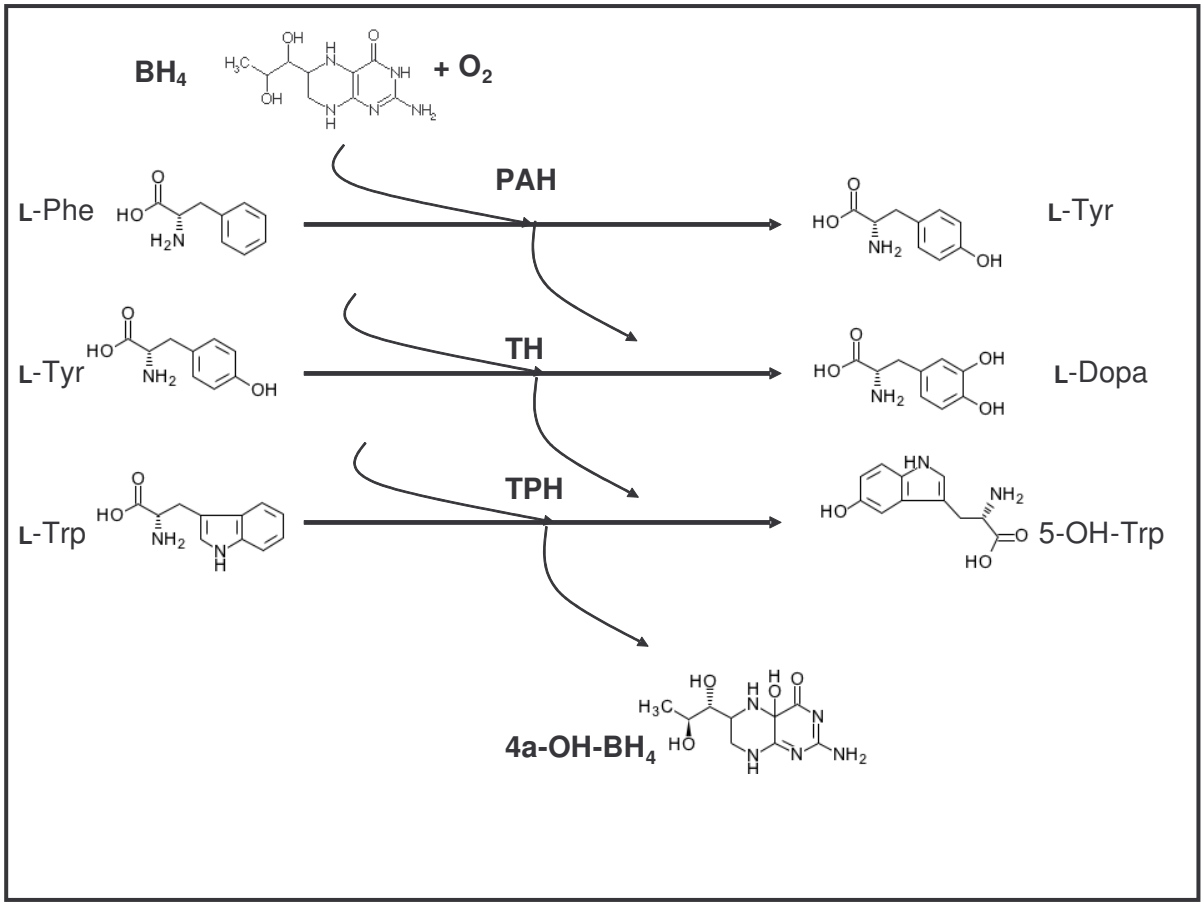

Figure 8. The reactions catalyzed by the AAAHs. PAH converts L-Phe to L-Tyr, TH converts L-Tyr to L-Dopa, and TPH converts L-Trp into 5-OH-Trp using molecular oxygen as cosubstrate. The cofactor, BH_4, is utilized by all mammalian AAAH. In addition, Fe^{2+} at the active site of the enzymes is necessary for the reaction.

Activation by phosphorylation

As mentioned above, the IARS contains a regulatory phosphorylation site for mammalian PAH. The cAMP dependent kinase A (PKA) phosphorylates PAH at Ser16. Upon phosphorylation PAH is activated and the activation mechanisms by

phosphorylation and L-Phe activation are interdependent and synergistic (Phillips and Kaufman 1984; Doskeland et al. 1996). As supported by recent site-directed mutagenesis and computational structural biology studies, the IARS covers the active site in the unphosphorylated state and phosphorylation at Ser16 induces local structural changes that result in a higher accessibility of the L-Phe binding site (Miranda et al. 2002; Miranda et al. 2004). PAH has at least one more regulatory mechanism, provided by its cofactor.

Inhibition by BH_4

BH_4 exerts a regulatory inhibiting mechanism and is believed to interact with residues in the IARS, which causes the entire regulatory domain to move closer to the catalytic domain, hence hindering PKA access to the phosphorylation site, barricading L-Phe binding at the active site and stabilizing the enzyme (Mitnaul and Shiman 1995; Miranda et al. 2002; Teigen and Martinez 2003).

2. Methods and Theoretical considerations

2.1. Protein sequence alignments

All comparisons of protein sequences require the amino acid positions to be aligned in columns, in a so called sequence alignment. The main goal of a sequence alignment is to group homologous positions in horizontal sequences into vertical columns. The information content is higher in protein sequences than in DNA sequences, given the number of character states and physicochemical properties. For this thesis all sequence analysis have been performed on protein sequences, although many of the methods presented below are applicable to nucleotide data as well. When the sequence divergence is too high to generate a good sequence alignment, structural alignments can improve the results.

2.1.1. Sequence alignments

Sequence alignments can be pairwise of just two sequences or multiple if they contain more sequences. Sequence alignments that span the entirety or most of the length of sequences to be aligned are called global alignments. When searching for re-occuring motifs, e.g. in a BLAST search (Altschul et al. 1990), local alignments

that match a smaller more conserved sequence section are used. Dynamic programming was for long the foundation for all sequence alignments. The two most widely used dynamic programming algorithms are Needleman-Wunsch (Needleman and Wunsch 1970) for global alignments and Smith-Waterman (Smith and Waterman 1981) for local alignments. These algorithms search for the best path between two sequences by using an amino acid substitution matrix for matching positions in the different sequences and a scoring function for indel events (also known as gaps). Today more sophisticated multiple sequence alignment programs that use e.g. the phylogeny of the sequences (i.e. Muscle (Edgar 2004)) or structural information (i.e. Expresso (Armougom et al. 2006)) are also available. These are also dependent upon amino acid substitution matrices, which are further discussed below.

2.1.2. Structural alignments

Comparing different proteins with similar structure can provide important information. This requires that the protein structures are aligned or superimposed. The easiest and fastest way is to align the sequences and then superimpose the structures based upon the sequence alignment. This method is however not recommended for less similar sequences. Instead, methods that compare the local environment e.g. distances and angles of nearby residues in the 3D protein space have been developed.

Aligned fragments pairs

A common way of performing structural alignments is to initially identify the smaller fragments that show high similarities in their local environments surrounding

the Cα-carbons in the two structures to be aligned. The structural alignment is formed by connecting the fragments together. This has been the underlying procedure for different methods, e.g. Dali (Holm and Sander 1993), CE (Combinatorial Extension) (Shindyalov and Bourne 1998) and FATCAT (Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists) (Ye and Godzik 2003). However, FATCAT is different than the other two methods in the way it lets the user decide if the structures are to be considered rigid bodies or not. By allowing twists, FATCAT permits more flexibility between the two structures in the alignment, which can be very important in e.g. homology searches. A comparison of FATCAT to two different rigid structure alignment programs, DALI and CE, showed that FATCAT performed better alignments. DALI stopped at hinge positions, where FATCAT inserted a twist, while CE continued and thereby aligned non-homologous residues (Ye and Godzik 2003).

SSAP

The Sequential Structure Alignment Program (SSAP) is used to align protein structures in the CATH database. The structural alignment is generated by double dynamic programming. It compares the surrounding environment (distances to the neighboring residues) of each Cβ-carbon in both structures and uses this information to create the sequence alignment based on structural criteria (Taylor et al. 1994; Orengo and Taylor 1996). Besides differences in the implementation of e.g. SSAP and FATCAT, an additional important difference is the starting point for SSAP, the Cβ-carbon, instead of the commonly used Cα-carbon (Figure 3). A dummy Cβ-carbon is used for Gly.

MUSTANG

Similarly, to generating multiple sequence alignments, also multiple structure alignments can be constructed. One algorithm that constructs these is MUSTANG (MUltiple STructural AligNment AlGorithm). By aligning residues based upon similarities in patterns of both residue-residue contacts and local topology, MUSTANG compares the structures to be aligned in a pairwise manner. The pairwise alignments are scored after similarity. Finally, the multiple structural alignment is constructed from the pairwise alignments, based upon their score (Konagurthu et al. 2006).

2.2. Tree-building

In order to reconstruct the different relationships between genes or proteins, a phylogenetic reconstruction is performed. The most commonly used methods can be divided into 2 categories: distance methods and discrete methods. Among the distance methods, Neighbor-Joining (NJ) and Unweighted pair group method with arithmetic mean (UPGMA) are prevalent. The discrete methods are parsimony and maximum likelihood (ML) or Bayesian based methods, which both use a model of evolution to find the optimal solution.

2.2.1. Distance methods

The distance methods reconstruct trees from the pairwise distances (e.g. the sequence identities) in a matrix between the different genes or proteins. These are not

recommended methods for larger phylogentic analysis, since sequence evolution is oversimplified when reduced to a number. NJ is more a clustering method than a tree building procedure and UPGMA applies a constant rate of evolution, according to a molecular clock, which is rarely true. However, these methods are fast and easy to use, e.g. as implemented in the Phylip package (Felsenstein 1996).

2.2.2. Discrete methods

The discrete methods use character data, gene or protein sequences, to resolve the relationships between the homologous sequences in an alignment. As noted above, an alignment of amino acid sequences contains more information than a non codon-based nucleotide alignment. The same is true for tree building purposes.

Parsimony

Parsimony tries to find the right tree from the sequence data that has the fewest number of substitutions. However, parsimony thereby ignores branch length differences -such as multiple events along a long branch being more likely than on a short branch- for the data and it has problems finding only one solution (Page and Holmes 1998).

Maximum likelihood and Bayesian methods

These are computationally expensive methods, but they often outperform all other tree building methods. Much of what applies to maximum likelihood also applies to the Bayesian method. The major difference between the two is the use of prior probability and the characterization of a distribution. For ML, the following equation is used;

$$L_D = f(X \mid \tau, \nu, \theta) \qquad \text{(Eq. 1)}$$

This equation is to be interpreted as follows: if given some data (X) and tree topology (τ), branch length (ν), and model of evolution (θ), then the likelihood (L_D) is the probability of obtaining X given τ, ν, and θ. This is also called the likelihood function (Page and Holmes 1998). The likelihood function is also found in Bayes's rule, which is used for the Bayesian phylogenetic inference:

$$f(\tau, \nu, \theta \mid X) = [f(\tau, \nu, \theta)\ f(X \mid \tau, \nu, \theta)] / f(X) \qquad \text{(Eq. 2)}$$

where $f(\tau, \nu, \theta \mid X)$ is the *posterior probability*, $f(\tau, \nu, \theta)$ is the *prior probability*, $f(X \mid \tau, \nu, \theta)$ is the *likelihood function*, which describes the probability of the data under different parameter values, and $f(X)$ is the total probability of the data summed and integrated over the parameter space (Ronquist and Huelsenbeck 2003). In MrBayes, this is coupled to a Markov Chain Monte Carlo (MCMC) technique. By applying a Metropolis-Hastings sampler, MrBayes updates single parameters or blocks of related parameters in each generation. If for the current generation the Markov chain has the parameters τ, ν, θ, then a new value for θ (θ^*), picked from a proposal distribution $q(\theta^* \mid \theta)$, will be accepted with the following probability (r):

$$r = \min (1, f(\theta^*)/f(\theta)\ f(X \mid \tau, \nu, \theta^*)/f(X \mid \tau, \nu, \theta)\ q(\theta \mid \theta^*)/q(\theta^* \mid \theta)) \qquad \text{(Eq. 3)}$$

The MCMC chains will run until the chain converges, or for as many generations that have been specified by the user. Further, the user can specify how often to keep a sample (a tree topology with its corresponding probability). These are used to create a consensus tree for the data when finished. As the initial chains are a bit unstable, it is

wise to disregard the first samples, called the *burnin* phase. The consensus tree will have posterior probabilities of each node (Ronquist and Huelsenbeck 2003). However, both ML and Bayesian phylogenetic inference require a specified model of evolution.

2.3. Models of evolution

An evolutionary model usually contains a substitution matrix and the substitution rates for the data. An amino acid substitution matrix gives the probability of going from one amino acid to another and how these probabilities are estimated designates the model. For an amino acid substitution matrix the dimensions are 20x20 and each transition is given a probability (P). The rate matrix (Q) consists of two factors, the observed substitutions (S_{ij}) multiplied by the frequency of the resulting amino acid (π_j), hence

$$(Q_{ij}=S_{ij} \times \pi_j) \qquad \text{(Eq. 4)}$$

and

$$P(\upsilon)=e^{(\upsilon Q)} \qquad \text{(Eq. 5)}$$

where υ represents the branch length. The amino acid frequencies, π, are often estimated form a large dataset and not recalculated. Margret Dayhoff was a pioneer in the field of estimating empirical substitution matrices from homologous global alignments in a phylogenetic (parsimony based) context in the 1960's and 1970's.

This procedure was repeated in 1978 and the resulting substitution matrices are generally referred to as Dayhoff or PAM (Point accepted mutations per 100 residues) matrices, e.g. PAM250 equals a 20 % sequence identity while PAM80 equals 50 % sequence identity (Dayhoff et al. 1978). A more recent but similarly constructed substitution matrix is the JTT matrix, named by its authors Jones, Taylor, and Thornton (Jones et al. 1992). Another commonly used substitution matrix also derived from real data is BLOSUM (BLOck SUbstitution Matrices) (Henikoff and Henikoff 1992). These are estimated from local alignments of similar proteins at different thresholds, but not in a phylogenetic context. These matrices are referred to as e.g. BLOSUM62 when derived for a threshold of 62 % sequence identity. It is worth noting that these empirically derived matrices have been found to outperform matrices based upon physicochemical properties. Whelan and Goldman proposed a model based upon maximum likelihood in combination with a counting approach to make more accurate substitution matrices without the computational cost associated with a pure ML approach (Whelan and Goldman 2001). This matrix is referred to as WAG.

Here we have presented a few of the commonly used substitution matrices, but there is an enormous variety of substitution matrices to choose from. The appropriate model for the data can be tested by performing a model test (Posada and Crandall 1998). For proteins ProtTest is the state of the art (Abascal et al. 2005). ProtTest suggests a model based upon an Akaike statistical test (AIC) utilizing the relationship between the likelihood (L) and the number of parameters (K) for some of the most common substitution matrices:

$$AIC = -2\ln L + 2K \qquad \text{(Eq. 6)}$$

This is also done with the different combinations of: *frequency*, the model will fit the data better by re-estimating amino acid frequency from the actual data; *invariant*, the model will fit the data better by specifying a proportion of invariant sites; and *gamma*, the model fit the data better by being divided into different rate categories. In order to reduce computational complexity four discrete rate categories are often used, instead of a continuous gamma distribution (Yang 1994). The shape parameter of the gamma distribution is called α and provides information on the evolutionary rate distribution in the data.

2.4. Modeling protein structure

The information in a linear protein sequence provides much less information that in its 3D form. There is a major effort undergoing right now to determine a large number of protein structures, carried out by Structural Genomics Consortiums all over the world. Today there are about 50000 experimentally determined protein structures compared to 5800 in 1998 according to the Protein Data Bank (www.pdb.org) (Berman et al. 2000). In 1998 there were almost 500 different folds classified. Today there are about 1000 different folds. Since 2005 very few new folds have been discovered, while the number of protein structures has increased with almost 20000 structures. This tells us that there is much to gain from comparative modeling or homology modeling, since its rationale is to identify a homolog with

known structure upon which the sequences without experimentally determined structures can be modeled. Given that the extent of neostructuralization still needs to be determined some caution should be used when applying homology modeling procedures.

2.4.1. Homology modeling

Swiss-Model is a homology modeling server (Schwede et al. 2003). It allows the user to perform a fully automated step where only the sequence, the so called target, needs to be provided. A structural homolog, the so called template, if available, will be identified by the server. The modeling procedure is rather simple as the coordinates from the templates Cα atoms are transferred to the targets Cα atoms. Thereafter the side chains are added from a rotamer library, which has the most common conformations for the amino acids. If there are any insertions in the target, these are modeled as loops, by searching a database of experimentally determined loop structures. Finally, when all side chains are in place the model is energy minimized. However, the automated mode has limited use since it can not build biological units in the case of oligomers. In order to build biological units, the *project* mode must be used and the user must build the modeling project in Swiss-Modeller's own protein structure viewer and builder, DeepView (Guex and Peitsch 1997). This mode leaves lots of interesting features for the user to work with. First a structural homolog must be identified, from here on referred to as the template, and its sequence must be aligned with the protein to be modeled, the so called target.

The alignment is very important in homology modeling. If the alignment is wrong, e.g. if non-homologues residues are aligned, it will result in an inaccurate

model. Therefore, it is sometimes important to use more sequences and to build a multiple sequence alignment instead of a pairwise sequence alignment since the former provides more information. The template is loaded into DeepView (Guex and Peitsch 1997). The target is loaded onto the template and the user can modify the alignment in DeepView accordingly to the pre-generated alignment before submitting it to the Swiss-Modeler server. Once the model is done it can be analyzed in DeepView. Perhaps it needs to be further optimized, in the way of changing the alignment or template, until a reasonable model is achieved. In DeepView many different operations such as fusing different templates and building biological units can be done. These are important steps in protein structure modeling since it adds to the functionality of the resulting models.

PAH structures

No full-length structure, with all three domains, is available for PAH or any AAAH in the PDB. Nevertheless, the available truncated structures have allowed the preparation of composite full-length models, as e.g. in (Flatmark and Stevens 1999; Thorolfsson et al. 2002; Kim et al. 2006). Our models from Paper II are shown in Figure 7D and 7E. There are various structures of catalytic domains (Figure 7B) with different combinations of pterin cofactors, presence and oxidation state of the Fe, and substrate analogs, mostly from human PAH. There are two structures of the regulatory domain (Figure 7A) in conjunction with the catalytic domain from rat PAH (Kobe et al. 1999). Only one structure of human PAH with the catalytic domain and the entire oligomerization domain is available (Fusetti et al. 1998) (Figure 7C). This structure shows two dimers, which each are internally symmetrical, but the

tetramer is asymmetrical. The biggest difference of the two different dimers is the positioning of the tetramerization α-helix in the oligomerization domain (Fusetti et al. 1998). In one of these dimers the oligomerization helices are kinked at position T427 and G442. The first kink leads to a rotation of the catalytic domain of 22˚ relative to the oligomerization domain and the second kink in the helices themselves leads to a 20˚ rotation which may be important for the interactions between the helices in the tetramer. Both kinks contribute to make the helices reach further away from the catalytic domain as to interact with the corresponding helices in a tetrameric conformation similar to what is seen for the structure of the catalytic and oligomerization domains in TH (Goodwill et al. 1997). The straight tetramerization α-helices in one dimer are on the other hand too close to the catalytic domain to be involved in forming a tetrahelical bundle as seen for TH. By combining these structures in various ways we constructed our working models of full-length tetrameric PAH at different conformational –and probably functional- states (see paper II, below).

Protein docking

When combining various domains together to form full-length protein chains, it is sometimes necessary to find alternative ways of docking those domains together. In addition, when composing dimers or tetramers of full-length subunits, some domains may clash in the higher order structure. For PAH all available structures are truncated at one end or another and therefore we used ZDOCK to explore alternative quaternary structures for our models. ZDOCK utilizes an unbound docking approach. Unbound docking is for protein structures, models, and domains, which were not

crystallized together. This leaves many surface side chains to optimize in the search for probable interactions, if they are allowed to be flexible. ZDOCK instead uses a soft docking approach which allows six rotational/translational degrees of freedom in addition to a less strict target function between the two domains or proteins to be docked. The target function for ZDOCK considers shape complementarity (the rotational and translational conformations), desolvation free energy for atoms at the interface, and electrostatics, and is based of a Fast Fourier Transform algorithm (Chen and Weng 2002; Chen et al. 2003).

2.5. Diverge analysis – prediction of residues involved in functional change

When comparing sequences, conserved sites are often taken as functionally important, while varying sites are not. When the sequence family is large enough, the sequences can often be divided into subgroups or clusters based upon their evolutionary and functional relationship. By forming clusters we can ask what is different among the sequences in the different clusters as a lead to explain the experimentally seen functional changes. The Diverge software estimates the probability of different sites in an alignment under a certain tree topology being involved in functional divergence (Gu et al. 2003; Gu 2006). Functional divergence in protein sequences can be divided into two main categories, types I and II functional divergence. Type I functional divergence, corresponds to site specific rate changes and type II functional divergence to site specific physicochemical property

changes when comparing different sequence clusters (Gu 2006). Both the change from being a conserved site in one cluster to being a variable site in another cluster (Type I) and the change from a conserved small apolar amino acid to a big charged amino acid (Type II) indicate that the functionality of these sites has changed between the two different clusters. We used Diverge 2.0 (Gu 2006) to detect type II functional divergence in PAH between the nematode cluster and the mammalian cluster. The different amino acid classes used in Diverge are preset and contain the following groups: charge positive (K, R, and H), charge negative (D and E), hydrophilic (S, T, N, Q, C, G, and P), and hydrophobic ((A, I, L, M, F, W, V, and Y) (Gu 2006). A cluster specific change from one amino acid group to another is called a radical change. The sites that differ between the clusters are given a score (S) that can be converted to the posterior probability (PII) of that site being involved in a functional change by PII = S/1+S (Gu 2006). In our case a limitation for the application of this method was the number of nematode PAH sequences as we only had four of these. However, by combining our sites into 3 different classes the type I divergence was represented as well. Class 1 were the sites with cluster-conserved radical changes, class 2 were sites with radical changes but not strictly conserved within each cluster, and class 3 were sites with conserved physicochemical properties but with different cluster-specific residues. Class 2 therefore covers many of the type I sites.

2.6. Electrostatic interaction energies

The residue specific electrostatic interaction energies for accessible ionizable residues has been calculated in this work using an implementation of Tanford-Kirkwood model (Tanford and Kirkwood 1957) together with the Bashford-Karplus (reduced-set-of-sites) approximation (Bashford and Karplus 1991) as described (Ibarra-Molero et al. 1999), using a program kindly provided by Jose Manuel Sanchez-Ruiz (University of Granada). This approach calculates the interaction energies between unit charges placed on the protonation sites of the ionizable groups using the Tanford-Kirkwood model, obtaining an average charge on each ionizable group, the effective p*K* values and the pairwise interaction energies using the Tanford-Roxby mean field procedure. The following parameters were used; atomic van der Waals radius according to Chothia (Chothia 1976), solvent molecule radius 1.4 Å, temperature 298 K, pH 7, ionic strength 0.2, and dielectric constants for the protein and solvent 4 and 78.5, respectively. Gurd's accessibility correction (Matthew and Gurd 1986) was applied. The energy of charge–charge interaction of group i with the rest of the ionizable groups in the protein $\langle W_i \rangle$ can be used to estimate the total charge–charge interaction energy in the proteins ($\langle W_{q-q} \rangle$):

$$\langle W_{q-q} \rangle = \tfrac{1}{2} \sum_{i=1}^{n} \langle W_i \rangle \qquad \text{(Eq. 7)}$$

3. Aims

3.1. The evolution of the AAAHs

Earlier attempts at resolving the phylogeny of the AAAH family have suffered from the lack of an appropriate outgroup, as they have focused on metazoans with three or more full-length AAAH genes and sometimes included the bacterial one domain PAH gene (Grenett et al. 1987; Patton et al. 1998; Wiens et al. 1998). The suggested ancestral substrate specificities were inconsistent. Hence, the determination of the genome sequencing of the social amoeba *Dictyostelium discoideum,* which only contains one AAAH gene, prompted us to resolve the phylogeny and the ancestral specificity of this enzyme family (Paper I).

3.2. The function of the regulatory ACT domain in PAH

The essence of this thesis was to generate further insights regarding the regulation, in particular the cooperative response, of PAH provided by its regulatory domain by using a comparative approach. The regulatory domain has been classified as an ACT domain homolog, which certainly is an adequate classification from the

structural point of view. However, from a functional view, the ACT domain of PAH does not appear to fit the strict description of this domain. There is a debate in the field regarding the ability of PAH to bind L-Phe to an effector binding site at the regulatory ACT domain. Various authors claim that there is a L-Phe binding site in the regulatory ACT domain (Tourian 1971; Shiman 1980; Parniak and Kaufman 1981; Kaufman 1993; Kappock and Caradonna 1996; Gjetting et al. 2001) and the assumption that PAH could bind L-Phe to the regulatory domain seemed correct based upon the ligand-binding properties of the ACT domain in other proteins. However, previous experimental results from our group (Thorolfsson et al. 2002) did not fit into this context and we set out to investigate this domain further. First, we performed a comparative study of the ACT domain from PAH and from two other allosterically regulated ACT domains, D-3-phosphoglycerate dehydrogenase (3PGDH) and the bifunctional chorismate-prephenate dehydrogenase (P-protein), which highlighted significant motifs in the sequences and corroborated important evolutionary relationships among these proteins (Paper III). Further, the ACT domain has a superfold topology, the so called ferredoxin-like fold, and domains with this fold are present in many different protein families and also as standalone domains. We explored more distantly related potential ACT domain homologs to gain further insights to the possible (present and ancient) functions of this domain in PAH (Paper IV). In addition, it was discovered that PAH in *Caenorhabditis elegans* was devoid of positive cooperativity and it was always found as a tetramer (Calvo et al. 2008). This provided an excellent system to study the differences between an allosterically regulated PAH in mammals and a non cooperative PAH with high activity in

nematodes. Being in the post-genomic era we could gather enough sequences of PAH from both mammals and nematodes to predict residues involved in generating the mammal and nematode specific phenotypes of PAH. As a variety of different structures of the different domains in PAH are available, we aimed at generating full-length models to study the 3D location of the residues predicted to have changed their function between nematodes and mammals. From the structural models, the residue specific electrostatics interaction energies were calculated to add another comparable dimension to our analysis of these two groups in order to get further insights to the promotion of the allosteric response in mammalian PAH that involves all three domains (Paper II).

4. Contributions

4.1. List of papers

Paper I:

Siltberg-Liberles, J., Steen, I. H., Svebak, R. M. & Martinez, A. (2008), "The phylogeny of the aromatic amino acid hydroxylases revisited by characterizing phenylalanine hydroxylase from Dictyostelium discoideum.", GENE, doi:10.1016/j.gene.2008.09.005. *In press*

Paper II:

Siltberg-Liberles, J. & Martinez, A. (2008), "Structural determinants of the regulatory properties in phenylalanine hydroxylase," Manuscript. *To be submitted.*

Paper III:

Liberles, J. S.*, Thorolfsson, M, & Martinez, A. (2005): "Allosteric mechanisms in ACT domain containing enzymes involved in amino acid metabolism.", Amino Acids, 28:1-12.

Paper IV:

Siltberg-Liberles, J. & Martinez, A. (2008): "Searching distant homologs of the regulatory ACT domain in phenylalanine hydroxylase.", Amino Acids, doi:10.1007/s00726-008-0057.2. *In press*

*** S as is Siltberg**

4.1.1. Residue denomination

For this and following sections, various positions in the PAH sequence will be discussed. All numbering refers to the human PAH sequence, and 1-letter codes for amino acid residues are used. When amino acid differences are present at a position compared between species or by a mutation, the amino acid in human or mammalian PAH is given first and the other species amino acid is given after the position, e.g. A322S refers to position 322 in human PAH, where human has an A and the other species has a S.

4.2. The phylogeny of the aromatic amino acid hydroxylases (Paper I)

Previous attempts to reconstruct the AAAH phylogeny suffered from not having an appropriate outgroup. TH has often been assigned as the ancestral function, but this selection lacks experimental evidence. Further, there have been different explanations put forward to the series of events leading to the set of AAAH genes in the mammalian genomes seen today. By identifying and characterizing the first AAAH gene form a eukaryotic genome with only one AAAH gene, *D. discoideum*, we have found an outgroup to root the AAAH tree. We believe this is close to the ancestor of the metazoan AAAH because i) it is only present in one gene copy and ii) *D. discoideum* is a close outgroup to metazoans. This gene was experimentally characterized and the phylogeny of the AAAHs completed. Our MrBayes tree places *D. discoideum* PAH (DictyoPAH) as the natural outgroup to all metazoan AAAHs,

with the PAH branch being the closest, allowing us to resolve the AAAH phylogeny. The most parsimonious explanation is to assume that the last common ancestor of DictyoPAH and the AAAH genes in metazoans had PAH activity and that the TH and TPH functions are derived from it. By analyzing only complete eukaryotic genomes and the chromosomal location of the AAAH genes we could envision how the AAAH genes have evolved. In short, the ancestral AAAH gene was duplicated twice during the Cambrian explosion, separating protozoans and metazoans. By a neofunctionalization mechanism, these duplications gave rise to TH and TPH. Hence, e.g. nematodes have one PAH, one TH, and one TPH gene, located on the same chromosome. A whole genome duplication took place in the chordate-vertebrate lineage (McLysaght et al. 2002). In mammals, this eventually led to pseudogenization and loss of the initial TH copy and the duplicated PAH copy. The only gene that was successful in maintaining both its gene copies in the genome was TPH by the two copies having different tissue specificity. Birds and probably marsupials have kept the initial TH copy as well.

It should be noted that PAH from *D. discoideum* is devoid of positive cooperativity and the cofactor specificity is different from other eukaryotic AAAHs. *D. discoideum* synthesizes another variant of biopterin, i.e. DH_4, not found in other eukaryotes (Klein et al. 1990) and DictyoPAH shows higher activity with DH_4 than with BH_4. We identified interesting differences between the PAH from human and *D. discoideum*: i) the S251A and A322S substitutions, which seem responsible for the change in cofactor preference, and ii) an additional disulfide bridge in DictyoPAH located close to 322S, namely 324C. In DictyoPAH, 324C (corresponding to I234 in

human PAH) forms a disulfide bridge to 373C and it seems possible that this covalent interaction reduces the dynamics of DictyoPAH compared to human PAH and also contributes to cofactor selectivity. This may partly explain the lack of conformational flexibility in DictyoPAH. Finally, iii) in close vicinity to the substrate binding pocket, DictyoPAH has a CC motif, where human PAH has an IC motif (positions 283, 284) (Paper I). We characterized a human PAH I283C mutant. It had strict PAH activity, but was devoid of positive cooperativity. It was tetrameric with similar stability as seen for the wild-type human PAH tetramer, but with lower affinity for both substrate and cofactor. Residue I283 in human PAH therefore seems important for the substrate induced activation. If it is replaced with a rigid residue like C the flexibility in this area adjacent to the active site is reduced and the substrate induced activation is prohibited, as positive cooperativity is lost. However, the I283C mutant is a tetramer. This indicates that there may be structural changes at the domain-domain interfaces so that the tetrameric conformation is preferred, but the active site changes associated with increased affinity are missing. Many other mutations have been shown to affect the positive cooperativity in human PAH, located throughout the enzyme, e.g. I65, E178, A300S (Erlandsen et al. 2004), C237 (Thorolfsson et al. 2003), P244, R261, A309, V388 (Pey et al. 2004), Y325 (Miranda et al. 2005), R408 and T427 (Bjorgo et al. 2001). All these residues are conserved in DictyoPAH (Paper I). Many of these residues appear to be essential for maintaining the structural integrity since they are involved in misfolding mutations associated with PKU (http://www.pahdb.mcgill.ca/).

4.4. The archetypical ACT domain (Paper III)

The regulatory domain of the AAAH has been classified as an ACT domain. In this comparative review we consider three different ACT domains to the depth of their allosteric regulation in three different enzymes, namely D-3-phosphoglycerate dehydrogenase (3PGDH), the bifunctional chorismate-prephenate dehydrogenase (P-protein), and PAH. L-Ser is the enzyme product of 3PGDH, which is allosterically feedback inhibited by L-Ser binding to the interface between two ACT domains. This ACT domain dimer with L-Ser bound has served as the paradigm for the domain function in ACT-domain containing proteins. The ACT regulatory domain in the P-protein has also been found to coordinate L-Phe to a sequence motif (GALV-ESRP). The G from the GALV motif is found in most ACT domains (Aravind and Koonin 1999) and the entire GALV – ESRP motif is found in the regulatory domain of PAH. The P-protein is involved in the two first enzymatic steps in the shikimate pathway, used by plants and microorganisms to biosynthesize L-Phe and L-Tyr (Herrmann and Weaver 1999). The P-proteins are allosterically feedback inhibited by L-Phe. Further, when L-Phe binds to the GALV-ESRP motif (with positive cooperativity) the protein shifts from being an active dimer to becoming a less active tetramer. The homology between the regulatory domains in the P-protein and in PAH is thus obvious. But while the dimeric regulatory ACT domain in the P-protein has been shown to bind L-Phe (Pohnert et al. 1999) and by so doing it fits the description of the archetypical ACT domain, the only available structure of the ACT domain in PAH (Kobe et al. 1999) reveals that no contacts form between different ACT domains. Hence, there is no ACT domain dimer interface for L-Phe to bind to in mammalian PAH, and

experimental data suggests that L-Phe does not bind to any other site other than the active site (Thorolfsson et al. 2002). We reconstructed the phylogeny of the ACT domain in the P-protein and in the AAAHs. The phylogenetic tree shows that the AAAHs are a monophyletic group as compared to the ACT domain in the P-protein, which certainly may be the ancestor of all AAAH ACT domains, as previously suggested (Gjetting et al. 2001). The known function of the ACT domain in PAH today however indicates that the oligomeric state and the ligand binding functionality have changed compared to the P-protein, but the allosteric properties remain. It appears that the ACT domain itself is a highly flexible module (Paper III).

4.5. Distant homologs of the ACT domain (Paper IV)

The high sequence divergence, evolutionary mobility, and superfold topology of the ACT domain led us to believe that there may be more homologous relationships for this domain to be found. As we concluded in paper III, the ACT domain is a flexible module included in several proteins with various domain combinations, and although they are described as binding amino acids, this does not seem to be a strict requirement. We were especially interested in finding other functionalities that could be extrapolated to the AAAH. As we continued to search for the purpose of the ACT domain, it became evident that the ACT domain classification was inconsistent when comparing e.g. SCOP, CATH, and Pfam. Therefore, we decided to search for remote potential homologs of the ACT domain in the PDB, using the regulatory ACT domain in PAH as a query. Using a consensus

approach of FATCAT and SSAP led us to 18 potential ACT domains, many devoted to regulation. To avoid redundancy in our data, no pair had a sequence identity > 30 %. Using the inverted SSAP scores as distances we built a tree of these results (Paper IV). This method clustered the known Pfam families well. PAH clustered next to another known ACT domain that is not binding amino acids, but Ni-ions. This cluster is located in between the main ACT domain cluster and a cluster of metallobinding domains. Yet another cluster away, we find the GlnB-like domains. These domains are also involved in amino acid metabolism, of e.g. L-His and L-Gln.

A recent structure of prephenate dehydratase with L-Phe bound at the dimer interface (Tan et al. 2008), when superimposed to the structure of the regulatory domain and catalytic domain of PAH, reveals that the interface that is used by L-Phe to bind is not present in PAH, although the residues are conserved. The two binding-sites in prephenate dehydratase are formed by the region GALV from one chain and ESRP from another chain, and *vice versa*. This leaves one the GALV binding-site too open as the catalytic domain is too distant to interact. The G-L motif is found in many ACT domains and in many of the potential ACT domain, which indicates that it may have a direct structural importance (Paper IV).

5. General discussion

The AAAHs have provided an interesting system to study the interplay of protein structure and function. As phenylalanine hydroxylase evolved it has gone through various transitions. One gene encoding the catalytic domain in PAH is present in a seemingly random distribution of bacterial species. It is absent from archea and from some eukaryotic lineages, e.g. plants and fungi. As more protozoan genomes are being sequenced we notice that many of them have one copy of a PAH. Compared to bacterial PAH, the protozoan PAH has had two domain fusion events, aquiring an N-terminal domain as well as a C-terminal domain, consistent with domain fusion being more common than fission in the evolution of multidomain proteins (Kummerfeld and Teichmann 2005). As we proceed down the eukaryotic lineage into the metazoans, we note that all metazoans have at least three copies of the ancestral PAH gene. This is in accordance with the extensive radiation resulting in major gene duplications in early metazoan evolution (Lundin 1999). By experimentally characterizing DictyoPAH we can confirm that PAH was the original function. This means that two gene copies have been subject to neofunctionalization, TH and TPH. For TH a forward grade evolution as part of pathway evolution can be

depicted since its substrate is the product of PAH. The pathway in which PAH and TH are found, also include dopamine and adrenaline production and is highly important for various neurological process. TPH is found in a parallel pathway leading to e.g. serotonin production. Therefore, the evolution of the enzyme family from one PAH to a family of aromatic amino acid hydroxlases involved in various neurotransmitter production pathways were likely very important for the evolution of the animal kingdom. The triplication (or more) of PAH in early metazoans, was later followed by at least one round of genome duplication in the chordates to vertebrate transition (McLysaght et al. 2002). Initially, this ought to have meant at least 6 AAAH genes, but in the eutherian mammalian lineage two genes were pseudogenized as there are 4 AAAH genes found in their genomes today. An additional copy of TPH was added to the AAAH family in eutherian mammals as compared to invertebrates. This last genome duplication also resulted in a division of the AAAH genes onto two chromosomes, with PAH and TPH on one and TH and TPH on the other. Other vertebrates may have additional copies of the AAAH genes as a result of the genome duplication mentioned above or by later lineage specific genome duplications. In e.g. chicken, we noted that it had kept an additional TH gene, since it is located on the same chromosome as PAH and one TPH. This is similar to the chromosomal location of the three AAAH genes in non-vertebrate metazoans. One interesting feature of DictyoPAH is its clear preference for its species specific biopterin cofactor (Paper I). *D. discoideum* is one of seemingly few protozans that can synthesis a close analog of biopterin, using the same pathway as higher eukaryotes (Werner-Felmayer et al. 2002). Some bacteria, e.g. *Clorobium*

tepidum, has the enzymes needed for biopterin synthesis, while most bacteria cannot synthesize biopterin (Supangat et al. 2008). Interestingly, the bacterial biopterin is yet another conformation of the biopterin found in higher eukaryotes (Supangat et al. 2008).

The domains that were added to the catalytic domain sometime in the early eukaryotic lineage added regulatory properties to the enzyme. The N-terminal domain is an ACT domain, which are often found in allosterically regulated metabolic enzymes, while the C-terminal domain has two structural motifs involved in the oligomerization of the enzymes. TH and TPH are functional as homotetramers. Mammalian PAH is sophistically regulated by its substrate, cofactor, and phosphorylation, all together altering the equilibrium between a low-affinity and low-activity state to a high-affinity and high-activity state. This allosterically controlled regulation has not been detected for TH and TPH. Human PAH is the malfunctioning enzyme in PKU. More than 500 PKU causing mutations have been found (http://www.pahdb.mcgill.ca/). In a recent study of 10 PKU mutations located through-out the PAH sequences, several were found to impact the folding of the regulatory domain (Gersting et al. 2008). Our investigation of the allostery in PAH led us to the important domain junction of the regulatory and oligomerization domain from one subunit to the catalytic domain of the other subunit (Paper II), similar to another study by Thorolfsson et al (Thorolfsson et al. 2003). In addition we note a couple of radical amino acid substitution changes in the immediate substrate binding pocket that could account for stabilizing the high-affinity and high-activity state in nematode PAH, e.g. G272H and M276K (Paper II). Similarly, we note that also

DictyoPAH has two additional Cys residues lining the active site, e.g. I283C and I324C (Paper I). When the Dictyostelium discoideum specific biopterin cofactor is used, this PAH has high affinity for L-Phe (Paper I). These PAH are all devoid of the positive cooperativity seen in mammalian PAH. This indicates that the acquisition of a dynamic binding site is coupled to the acquisition of allostery, similar to what is seen for the allosterically activated PAH from the cold-adapted bacterium *Colwellia psychrerythraea* (Leiros et al. 2007).

One of the conformational changes that we note between the low-affinity and low-activity models of human PAH seems to originate in the substrate binding pocket and lead over to the adjacent regulatory domain. Thus the human enzyme is characterized by a high mobility abd a large number of hinges (Stokka et al. 2004). Further, both the N-terminal and the C-terminal have been shown to be very flexible and susceptible to degradation, even for some PKU associated mutations that are not located in the N-terminal regulatory domain itself (Gersting et al. 2008).

PKU is today regarded as a misfolding disease with many destabilizing mutations as the main pathogenic mechanism (Pey et al. 2007). For several PKU mutations treatment with the biopterin cofactor can reduce the L-Phe concentration by 20-30 % depending on the severity of the mutation ((Fiege and Blau 2007) and refs therein). The mechanism for the rescued PAH activity may relay in a chaperone-like effect of biopterin that can stablize PAH (Scavelli et al. 2005) and also prevent its degradation (Erlandsen et al. 2004).

The regulatory domains of the different AAAH have highly divergent sequences. This is in accordance with its superfold topology. However, the

oligomeric state of the ACT domain has changed compared to other known ACT domains (Paper III and IV). The ACT domains are found in various combinations, forming dimers and tetramers face to face or side by side (Paper IV). In most cases ACT domains interact with other ACT domains, but this is not the case for the ACT domain in PAH. Most known ACT domains bind various amino acids, which is logical given their involvement in amino acid metabolism. However, there is at least one example of an ACT domain that in addition to its amino acid binding specificity also has acquired an additional ligand binding site, namely aspartate kinase which binds L-Lys and S-adenosylmethionine (Curien et al. 2008). However, while L-Lys binds to the dimer interface between two ACT domains, S-adenosylmethionine binds to only one ACT domain (Curien et al. 2008). This adds proof of principle to the ACT domains versatility. Truncating the PAH enzyme to only include the ACT domain results in dimer formation and L-Phe binding at the interface (Gjetting et al. 2001). This is likely the ancestral function of this domain, similar to what is seen for the P-protein (Paper III and IV). Perhaps the loss of the ACT domains ability to dimerize or form higher order structures as seen for other potential ACT domain homologs (Paper III) may be involved in rendering PAH highly susceptible to degradation for many PKU mutations. PAH has kept the L-Phe binding motif, but the structural constraints on how the domains assemble abolish its ancestral binding functionality (Paper IV).

The ESRP part of the L-Phe binding motif in the P-protein is interacting with α-helix 5 from the catalytic domain; an α-helix which we predicted had important interactions for promoting the allosteric response (Paper II). In addition, the R in

ESRP is R68, has been found to be important for the allosteric response by its interactions to C237 (Thorolfsson et al. 2003). In paper II, we found these residues at a domain junction formed by the catalytic domain from one chain and the regulatory domain as well as the oligomerization domain from an adjacent chain. The difference between mammalian PAH and nematode PAH in our models is that nematode PAH can form more inter domain interactions in this area. The regulatory domain has been shown to have large impact on human PAH stability and active site binding (Erlandsen et al. 2004; Gersting et al. 2008). Therefore, stabilizing the interactions between the regulatory domain and the rest of the enzyme may be beneficiary for correction of PKU mutations. Decreasing the pKa of the residues in the vicinity of R68 from the ESRP motif and the area around C237 in particular seems to be important for high activity ((Thorolfsson et al. 2003) and Paper II). This region may be a good target for stabilizing destabilized PAH in order to control the PKU phenotype by a small molecule drug approach.

6. Conclusions and future directions

6.1. Conclusions

Human PAH has become a fragile enzyme. This is highly noticeable given how frequent destabilizing mutations are. When comparing human PAH to PAH from both *D. discoideum* and *C. elegans* it becomes evident that human PAH must be more dynamic than these two, and it is indeed the case. The two additional Cys residues in close vicinity to both cofactor and substrate binding pockets in DictyoPAH are providing additional stability; one of these was confirmed to abolish cooperativity in human PAH as well. For PAH from *C. elegans* we have identified two strong candidates for making the active site less dynamic compared to human PAH. Both these residues, G272H and M276H, are located in close vicinity to the substrate binding pocket and probably involved in stabilizing the high-affinity and high-affinity state.

The ACT domain is a functionally highly versatile domain, which structure is one of the most common in the current fold pool. As a superfold domain, the same secondary structure elements and connectivity can be achieved with highly divergent sequences. As structure is more conserved than sequence, we set out to identify more

potential homologs. Using only one of the current ACT domain, we idenitified 18 new ACT domain candidates. Some of these where structural genomics targets with unknown function. By using a combination approach like this, insights to unknown domain functions can be gained. In addition, several non-redundant candidates were found from the same Pfam domains, the GlnB-like and heavy metal associated domains. There is probably a homologous connection between these domains and we are working on a way to show that (see future directions).

The L-Phe binding motif found in the homologous P-protein ACT domain is conserved in PAH. In PAH, this motif can no longer bind L-Phe, due to the quaternary assembly of PAH. The catalytic domain is today surrounded by two different oligomerization domains, one at the N-terminus and one at the C-terminus. We have seen that the ACT domain at the N-terminus is frequently involved in dimerization (Paper III and IV), and the leucine zipper motif in the very C-terminus is frequently involved in forming tetramers. In the quarternary assembly of the full-length multidomain AAAH protein, no contacts are formed between the different ACT domains. This may have resulted in the instability of the regulatory domain since it cannot interact with the other ACT domain from the adjacent subunit.

The AAAH family does not seem to have evolved from an enzyme with broad substrate specificity, but from a PAH specific enzyme. This is also what we see in bacteria. However, it should be noted that PAH in bacteria may be the result of a horizontal gene transfer, but distribution in the bacterial kingdom is scattered.

As we have shown for AAAH, the evolution of this multidomain protein in eukaryotes could be resolved with a suitable outgroup. This is probably the case for

many other eukaryotic multidomain proteins as well. By experimentally confirming the outgroup specificity and characteristics we have lots to learn about protein evolution of structure and function.

In addition, we need to resolve the superfold domain ancestry in order to know how multidomain proteins really evolve, e.g. in TH the ACT domain has diverged beyond sequence recognition. If this and other domains are not classified in accordance to their homologs in various databases, then many studies based on various databases will be wrong in establishing various trends for e.g. multidomain evolution.

Another application that can follow the understanding of superfold ancestry is protein structure prediction.

6.2. Future directions

(1) To get a better understanding of the different changes at the different interfaces, molecular dynamics simulations can be helpful. We have provided various interfaces and residues that are predicted to be important using bioinformatic statistical methods, electrostatic interaction energies, and structural models. Further, site directed mutagenesis can help test some of our hypothesis.

(2) Now is the time to start revisiting phylogenies of the multidomain proteins in order to get insights into their ancestral functions. The AAAHs are only one

of many other enzymes where copy number has increased in higher eukaryotes as compared to *D. discoideum*. Testing activities and substrate specificities will aid the understanding of domain and pathway evolution. This is crucial to our understanding of biology.

(3) Ultimately, we would like to be able to prove that the known and potential ACT domains identified in this thesis are homologs, but how does one do that? There is no precise way to demonstrate homology between domains that have a sequence identity of 0-10 %, although the conserved features are unlikely to be the result of convergent evolution. We would like to mention that we have tried to prove homology using various ideas. Our first attempt was to identify the residues involved in forming the folding nucleus, believed to be essential for the folding process. There are at least two known sets of overlapping folding nucleons in ribosomal protein S6 (Olofsson et al. 2007). We found these patterns to be conserved throughout the set of our known and potential ACT domains, but this was not enough to claim homology. Further, we tried to use ancestral sequence reconstruction using a very basic idea that as the node of each known and potential ACT domain was reconstructed, the sequence identity and alignment score should increase; with the underlying hypothesis that if the domains are truly homologous their ancestors ought to show higher sequence similarity as present day sequences continue to evolve and diversify. This almost took us all the way, but we did not see improvement for all known ACT domains. Now, after giving up on this approach, I have come to realize that the fatal flaw was to not create

homology models of the reconstructed sequences and to compare the models, not the sequences, to each other. Further, we could have extended the method to include more sequences since the top scoring sequence may not be an exact replica of the ancestral sequence. Collaboration with Prof. Lesley Collins at Massey University, New Zealand, has been initiated to continue developing the method of modeling ancestral sequences and structures in order to prove homology between superfold domains.

7. References

Abascal, F., Zardoya, R., and Posada, D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21:** 2104-2105.

Abkevich, V.I., Gutin, A.M., and Shakhnovich, E.I. 1994. Specific Nucleus as the Transition-State for Protein-Folding - Evidence from the Lattice Model. *Biochemistry* **33:** 10026-10036.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Andersen, O.A., Stokka, A.J., Flatmark, T., and Hough, E. 2003. 2.0A resolution crystal structures of the ternary complexes of human phenylalanine hydroxylase catalytic domain with tetrahydrobiopterin and 3-(2-thienyl)-L-alanine or L-norleucine: substrate specificity and molecular motions related to substrate binding. *J Mol Biol* **333:** 747-757.

Apic, G., Gough, J., and Teichmann, S.A. 2001a. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* **310:** 311-325.

Apic, G., Gough, J., and Teichmann, S.A. 2001b. An insight into domain combinations. *Bioinformatics* **17 Suppl 1:** S83-89.

Aravind, L., and Koonin, E.V. 1999. Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J Mol Biol* **287:** 1023-1040.

Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V., and Notredame, C. 2006. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucl Acids Res* **34:** W604-608.

Bashford, D., and Karplus, M. 1991. Multiple-Site Titration Curves of Proteins - an Analysis of Exact and Approximate Methods for Their Calculation. *J Phys Chem* **95:** 9556-9561.

Bennett, M.J., Sawaya, M.R., and Eisenberg, D. 2006. Deposition diseases and 3D domain swapping. *Structure* **14:** 811-824.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucl Acids Res* **28:** 235-242.

Birchler, J.A., Riddle, N.C., Auger, D.L., and Veitia, R.A. 2005. Dosage balance in gene regulation: biological implications. *Trends Genet* **21:** 219-226.

Bjorgo, E., de Carvalho, R.M., and Flatmark, T. 2001. A comparison of kinetic and regulatory properties of the tetrameric and dimeric forms of wild-type and Thr427-->Pro mutant human phenylalanine hydroxylase: contribution of the flexible hinge region Asp425-Gln429 to the tetramerization and cooperative substrate binding. *Eur J Biochem* **268:** 997-1005.

Bloom, J.D., Raval, A., and Wilke, C.O. 2007. Thermodynamics of neutral protein evolution. *Genetics* **175:** 255-266.

Buchler, N.E.G., and Goldstein, R.A. 1999. Effect of alphabet size and foldability requirements on protein structure designability. *Proteins* **34:** 113-124.

Calvo, A.C., Pey, A.L., Ying, M., Loer, C.M., and Martinez, A. 2008. Anabolic function of phenylalanine hydroxylase in Caenorhabditis elegans. *Faseb J*.

Campbell, R.E., Mosimann, S.C., van De Rijn, I., Tanner, M.E., and Strynadka, N.C. 2000. The first structure of UDP-glucose dehydrogenase reveals the catalytic residues necessary for the two-fold oxidation. *Biochemistry* **39:** 7012-7023.

Canfield, D.E., and Teske, A. 1996. Late Proterozoic rise in atmospheric oxygen concentration inferred from phylogenetic and sulphur-isotope studies. *Nature* **382:** 127-132.

Chen, R., Li, L., and Weng, Z. 2003. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **52:** 80-87.

Chen, R., and Weng, Z. 2002. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins* **47:** 281-294.

Chothia, C. 1976. Nature of Accessible and Buried Surfaces in Proteins. *J Mol Biol* **105:** 1-14.

Chothia, C., and Lesk, A.M. 1986. The relation between the divergence of sequence and structure in proteins. *Embo J* **5:** 823-826.

Coulson, A.F., and Moult, J. 2002. A unifold, mesofold, and superfold model of protein fold use. *Proteins* **46:** 61-71.

Curien, G., Biou, V., Mas-Droux, C., Robert-Genthon, M., Ferrer, J.L., and Dumas, R. 2008. Amino acid biosynthesis: new architectures in allosteric enzymes. *Plant Physiol Biochem* **46:** 325-339.

Dayhoff, M.O.E., Schwartz, R.M., and Orcutt, B.C. 1978. *Atlas of Protein Sequence and Structure*

National Biomedical Research Foundation, Washington, DC.

DePristo, M.A., Weinreich, D.M., and Hartl, D.L. 2005. Missense meanderings in sequence space: A biophysical view of protein evolution. *Nat Rev Genet* **6:** 678-687.

Doskeland, A., Ljones, T., Skotland, T., and Flatmark, T. 1982. Phenylalanine 4-monooxygenase from bovine and rat liver: some physical and chemical properties. *Neurochem Res* **7:** 407-421.

Doskeland, A.P., Martinez, A., Knappskog, P.M., and Flatmark, T. 1996. Phosphorylation of recombinant human phenylalanine hydroxylase: effect on catalytic activity, substrate activation and protection against non-specific cleavage of the fusion protein by restriction protease. *Biochem J* **313 (Pt 2):** 409-414.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acids Res* **32:** 1792-1797.

Ekman, D., Bjorklund, A.K., and Elofsson, A. 2007. Quantification of the elevated rate of domain rearrangements in metazoa. *J Mol Biol* **372:** 1337-1348.

Elena, S.F., Wilke, C.O., Ofria, C., and Lenski, R.E. 2007. Effects of population size and mutation rate on the evolution of mutational robustness. *Evolution* **61:** 666-674.

England, J.L., and Shakhnovich, E.I. 2003. Structural determinant of protein designability. *Phys Rev Lett* **90:** -.

Erlandsen, H., Pey, A.L., Gamez, A., Perez, B., Desviat, L.R., Aguado, C., Koch, R., Surendran, S., Tyring, S., Matalon, R., et al. 2004. Correction of kinetic and stability defects by tetrahydrobiopterin in phenylketonuria patients with certain phenylalanine hydroxylase mutations. *Proc Natl Acad Sci USA* **101:** 16903-16908.

Erspamer, V., and Asero, B. 1952. Identification of enteramine, the specific hormone of the enterochromaffin cell system, as 5-hydroxytryptamine. *Nature* **169:** 800-801.

Felsenstein, J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Computer Methods for Macromolecular Sequence Analysis* **266:** 418-427.

Fenchel, T., and Finlay, B.J. 1994. The Evolution of Life without Oxygen. *Am Sci* **82:** 22-29.

Fersht, A. 1984. *Enzyme structure and mechanism*, 2nd ed. W. H. Freeman and Company, New York.

Fersht, A.R. 1999. *Structure and Mechanism in Protein Science*. W. H. Freeman, New York.

Fiege, B., and Blau, N. 2007. Assessment of tetrahydrobiopterin (BH4) responsiveness in phenylketonuria. *J pediatr* **150:** 627-630.

Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., et al. 2006. Pfam: clans, web tools and services. *Nucl Acids Res* **34:** D247-251.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L.L., et al. 2008. The Pfam protein families database. *Nucl Acids Res* **36:** D281-D288.

Fitch, W.M., and Markowitz, E. 1970. An Improved Method for Determining Codon Variability in a Gene and Its Application to Rate of Fixation of Mutations in Evolution. *Biochem Genet* **4:** 579-&.

Fitzpatrick, P.F. 1991. Steady-state kinetic mechanism of rat tyrosine hydroxylase. *Biochemistry* **30:** 3658-3662.

Fitzpatrick, P.F. 2003. Mechanism of aromatic amino acid hydroxylation. *Biochemistry* **42:** 14083-14091.

Flatmark, T., Almas, B., Knappskog, P.M., Berge, S.V., Svebak, R.M., Chehin, R., Muga, A., and Martinez, A. 1999. Tyrosine hydroxylase binds tetrahydrobiopterin cofactor with negative cooperativity, as shown by kinetic analyses and surface plasmon resonance detection. *Eur J Biochem* **262:** 840-849.

Flatmark, T., and Stevens, R.C. 1999. Structural Insight into the Aromatic Amino Acid Hydroxylases and Their Disease-Related Mutant Forms. *Chem Rev* **99:** 2137-2160.

Folling, A. 1934. The separation of phenylpyruvic acid in urine as a metabolism anomaly in connection with imbecillitate. *Hoppe-Seylers Zeitschrift Fur Physiologische Chemie* **227:** 169-176.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531-1545.

Fusetti, F., Erlandsen, H., Flatmark, T., and Stevens, R.C. 1998. Structure of tetrameric human phenylalanine hydroxylase and its implications for phenylketonuria. *J Biol Chem* **273:** 16962-16967.

Garcia-Mira, M.M., Sadqi, M., Fischer, N., Sanchez-Ruiz, J.M., and Munoz, V. 2002. Experimental identification of downhill protein folding. *Science* **298:** 2191-2195.

Gersting, S.W., Kemter, K.F., Staudigl, M., Messing, D.D., Danecka, M.K., Lagler, F.B., Sommerhoff, C.P., Roscher, A.A., and Muntau, A.C. 2008. Loss of function in phenylketonuria is caused by impaired molecular motions and conformational instability. *Am J Hum Genet* **83:** 5-17.

Gjetting, T., Petersen, M., Guldberg, P., and Guttler, F. 2001. Missense mutations in the N-terminal domain of human phenylalanine hydroxylase interfere with binding of regulatory phenylalanine. *Am J Hum Genet* **68:** 1353-1360.

Goldstein, R.A. 2008. The structure of protein evolution and the evolution of protein structure. *Curr Opin Struct Biol* **18:** 170-177.

Goodwill, K.E., Sabatier, C., Marks, C., Raag, R., Fitzpatrick, P.F., and Stevens, R.C. 1997. Crystal structure of tyrosine hydroxylase at 2.3 A and its implications for inherited neurodegenerative diseases. *Nat Struct Biol* **4:** 578-585.

Grenett, H.E., Ledley, F.D., Reed, L.L., and Woo, S.L. 1987. Full-length cDNA for rabbit tryptophan hydroxylase: functional domains and evolution of aromatic amino acid hydroxylases. *Proc Natl Acad Sci USA* **84:** 5530-5534.

Grishin, N.V. 2001. Fold change in evolution of protein structures. *J Struct Biol* **134:** 167-185.

Gu, X. 2006. A simple statistical method for estimating Type-II (Cluster-Specific) functional divergence of protein sequences. *Mol Biol Evol* **23:** 1937-1945.

Gu, Z., Steinmetz, L.M., Gu, X., Scharfe, C., Davis, R.W., and Li, W.H. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* **421:** 63-66.

Guex, N., and Peitsch, M.C. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18:** 2714-2723.

Gunasekaran, K., Ma, B.Y., and Nussinov, R. 2004. Is allostery an intrinsic property of all dynamic proteins? *Proteins* **57:** 433-443.

Halskau, O., Jr., Perez-Jimenez, R., Ibarra-Molero, B., Underhaug, J., Munoz, V., Martinez, A., and Sanchez-Ruiz, J.M. 2008. Large-scale modulation of thermodynamic protein folding barriers linked to electrostatics. *Proc Natl Acad Sci USA* **105:** 8625-8630.

Henikoff, S., and Henikoff, J.G. 1992. Amino-Acid Substitution Matrices from Protein Blocks. *Proc Natl Acad Sci USA* **89:** 10915-10919.

Herrmann, K.M., and Weaver, L.M. 1999. The Shikimate Pathway. *Annu Rev Plant Physiol Plant Mol Biol* **50:** 473-503.

Holland, H.D. 1997. Geochemistry - Evidence for life on Earth more than 3850 million years ago. *Science* **275:** 38-39.

Holm, L., and Sander, C. 1993. Protein structure comparison by alignment of distance matrices. *J Mol Biol* **233:** 123-138.

Hughes, A.L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* **256:** 119-124.

Hulsmeyer, M., Hillig, R.C., Volz, A., Ruhl, M., Schroder, W., Saenger, W., Ziegler, A., and Uchanska-Ziegler, B. 2002. HLA-B27 subtypes differentially associated with disease exhibit subtle structural alterations. *J Biol Chem* **277:** 47844-47853.

Ibarra-Molero, B., Loladze, V.V., Makhatadze, G.I., and Sanchez-Ruiz, J.M. 1999. Thermal versus guanidine-induced unfolding of ubiquitin. An analysis in terms of the contributions from charge-charge interactions to protein stability. *Biochemistry* **38:** 8138-8149.

Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. The Rapid Generation of Mutation Data Matrices from Protein Sequences. *Computer Applications in the Biosciences* **8:** 275-282.

Kappock, T.J., and Caradonna, J.P. 1996. Pterin-Dependent Amino Acid Hydroxylases. *Chem Rev* **96:** 2659-2756.

Kaufman, S. 1993. The phenylalanine hydroxylating system. *Adv Enzymol Relat Areas Mol Biol* **67:** 77-264.

Kiel, C., and Serrano, L. 2006. The ubiquitin domain superfold: structure-based sequence alignments and characterization of binding epitopes. *J Mol Biol* **355:** 821-844.

Kim, S.W., Jung, J.S., Oh, H.J., Kim, J., Lee, K.S., Lee, D.H., Park, C., Kimm, K., Koo, S.K., and Jung, S.C. 2006. Structural and functional analyses of mutations of the human phenylalanine hydroxylase gene. *Clin Chim Acta* **365:** 279-287.

Kinch, L.N., and Grishin, N.V. 2002. Evolution of protein structures and functions. *Curr Opin Struct Biol* **12:** 400-408.

Klein, R., Thiery, R., and Tatischeff, I. 1990. Dictyopterin, 6-(D-threo-1,2-dihydroxypropyl)-pterin, a new natural isomer of L-biopterin. Isolation from vegetative cells of Dictyostelium discoideum and identification. *Eur J Biochem* **187:** 665-669.

Kleppe, R., Toska, K., and Haavik, J. 2001. Interaction of phosphorylated tyrosine hydroxylase with 14-3-3 proteins: evidence for a phosphoserine 40-dependent association. *J Neurochem* **77:** 1097-1107.

Knappskog, P.M., Flatmark, T., Aarden, J.M., Haavik, J., and Martinez, A. 1996. Structure/function relationships in human phenylalanine hydroxylase. Effect of terminal deletions on the oligomerization, activation and cooperativity of substrate binding to the enzyme. *Eur J Biochem* **242:** 813-821.

Knoll, A.H. 1992. The early evolution of eukaryotes: a geological perspective. *Science* **256:** 622-627.

Kobe, B., Jennings, I.G., House, C.M., Michell, B.J., Goodwill, K.E., Santarsiero, B.D., Stevens, R.C., Cotton, R.G.H., and Kemp, B.E. 1999. Structural basis of autoregulation of phenylalanine hydroxylase. *Nat Struct Biol* **6:** 442-448.

Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J., and Lesk, A.M. 2006. MUSTANG: A multiple structural alignment algorithm. *Proteins* **64:** 559-574.

Koonin, E.V. 1993. A superfamily of ATPases with diverse functions containing either classical or deviant ATP-binding motif. *J Mol Biol* **229:** 1165-1174.

Koshland, D.E., Nemethy, G., and Filmer, D. 1966. Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits. *Biochemistry* **5:** 365-385.

Kummerfeld, S.K., and Teichmann, S.A. 2005. Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet* **21:** 25-30.

Leiros, H.K., Pey, A.L., Innselset, M., Moe, E., Leiros, I., Steen, I.H., and Martinez, A. 2007. Structure of phenylalanine hydroxylase from Colwellia psychrerythraea 34H, a monomeric cold active enzyme with local flexibility around the active site and high overall stability. *J Biol Chem* **282:** 21973-21986.

Levinthal, C. 1968. Are There Pathways for Protein Folding. *Journal De Chimie Physique Et De Physico-Chimie Biologique* **65:** 44-&.

Liberles, D.A. 2005. Datasets for evolutionary comparative genomics. *Genome Biol* **6:** 117.

Livingstone, C.D., and Barton, G.J. 1993. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci* **9:** 745-756.

Long, M., Betran, E., Thornton, K., and Wang, W. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4:** 865-875.

Lopez, P., Casane, D., and Philippe, H. 2002. Heterotachy, an important process of protein evolution. *Mol Biol Evol* **19:** 1-7.

Lundin, L.G. 1999. Gene duplications in early metazoan evolution. *Semin Cell Dev Biol* **10:** 523-530.

Lynch, M. 2006. The origins of eukaryotic gene structure. *Mol Biol Evol* **23:** 450-468.

Marsden, R.L., Ranea, J.A., Sillero, A., Redfern, O., Yeats, C., Maibaum, M., Lee, D., Addou, S., Reeves, G.A., Dallman, T.J., et al. 2006. Exploiting protein structure data to explore the evolution of protein function and biological complexity. *Philos Trans R Soc Lond B Biol Sci* **361:** 425-440.

Martinez, A., Knappskog, P.M., Olafsdottir, S., Doskeland, A.P., Eiken, H.G., Svebak, R.M., Bozzini, M., Apold, J., and Flatmark, T. 1995. Expression of recombinant human phenylalanine hydroxylase as fusion protein in Escherichia coli circumvents proteolytic degradation by host cell proteases. Isolation and characterization of the wild-type enzyme. *Biochem J* **306 (Pt 2):** 589-597.

Matthew, J.B., and Gurd, F.R.N. 1986. Calculation of Electrostatic Interactions in Proteins. *Methods Enzymol* **130:** 413-436.

McLysaght, A., Hokamp, K., and Wolfe, K.H. 2002. Extensive genomic duplication during early chordate evolution. *Nat Genet* **31:** 200-204.

Melin, R., Li, H., Wingreen, N.S., and Tang, C. 1999. Designability, thermodynamic stability, and dynamics in protein folding: A lattice model study. *J Chem Phys* **110:** 1252-1262.

Miranda, F.F., Kolberg, M., Andersson, K.K., Geraldes, C.F., and Martinez, A. 2005. The active site residue tyrosine 325 influences iron binding and coupling efficiency in human phenylalanine hydroxylase. *J Inorg Biochem* **99:** 1320-1328.

Miranda, F.F., Teigen, K., Thorolfsson, M., Svebak, R.M., Knappskog, P.M., Flatmark, T., and Martinez, A. 2002. Phosphorylation and mutations of Ser(16) in human phenylalanine hydroxylase. Kinetic and structural effects. *J Biol Chem* **277:** 40937-40943.

Miranda, F.F., Thorolfsson, M., Teigen, K., Sanchez-Ruiz, J.M., and Martinez, A. 2004. Structural and stability effects of phosphorylation: Localized structural changes in phenylalanine hydroxylase. *Protein Sci* **13:** 1219-1226.

Mitnaul, L.J., and Shiman, R. 1995. Coordinate regulation of tetrahydrobiopterin turnover and phenylalanine hydroxylase activity in rat liver cells. *Proc Natl Acad Sci USA* **92:** 885-889.

Mojzsis, S.J., Arrhenius, G., McKeegan, K.D., Harrison, T.M., Nutman, A.P., and Friend, C.R.L. 1996. Evidence for life on Earth before 3,800 million years ago. *Nature* **384:** 55-59.

Monod, J., Changeux, J.P., and Jacob, F. 1963. Allosteric Proteins and Cellular Control Systems. *J Mol Biol* **6:** 306-&.

Monod, J., Wyman, J., and Changeux, J.P. 1965. On Nature of Allosteric Transitions - a Plausible Model. *J Mol Biol* **12:** 88-&.

Murakami, M.T., Fernandes-Pedrosa, M.F., Tambourgi, D.V., and Arni, R.K. 2005. Structural basis for metal ion coordination and the catalytic mechanism of sphingomyelinases D. *J Biol Chem* **280:** 13658-13664.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. Scop - a Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J Mol Biol* **247:** 536-540.

Needleman, S.B., and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48:** 443-453.

Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, Berlin, Germany.

Ohno, S. 1999. Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. *Semin Cell Dev Biol* **10:** 517-522.

Olofsson, M., Hansson, S., Hedberg, L., Logan, D.T., and Oliveberg, M. 2007. Folding of S6 structures with divergent amino acid composition: pathway flexibility within partly overlapping foldons. *J Mol Biol* **365:** 237-248.

Orengo, C.A., Jones, D.T., and Thornton, J.M. 1994. Protein superfamilies and domain superfolds. *Nature* **372:** 631-634.

Orengo, C.A., and Taylor, W.R. 1996. SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol* **266:** 617-635.

Owens, M.J., and Nemeroff, C.B. 1994. Role of serotonin in the pathophysiology of depression: focus on the serotonin transporter. *Clin Chem* **40:** 288-295.

Page, R.D.M., and Holmes, E.C. 1998. *Molecular Evolution: a phylogenetic approach*. Blackwell Science Ltd, Oxford, UK.

Pandi-Perumal, S.R., Trakht, I., Srinivasan, V., Spence, D.W., Maestroni, G.J., Zisapel, N., and Cardinali, D.P. 2008. Physiological effects of melatonin: Role of melatonin receptors and signal transduction pathways. *Prog Neurobiol* **85:** 335-353.

Parniak, M.A., and Kaufman, S. 1981. Rat liver phenylalanine hydroxylase. Activation by sulfhydryl modification. *J Biol Chem* **256:** 6876-6882.

Patton, S.J., Luke, G.N., and Holland, P.W. 1998. Complex history of a chromosomal paralogy region: insights from amphioxus aromatic amino acid hydroxylase genes and insulin-related genes. *Mol Biol Evol* **15:** 1373-1380.

Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., et al. 2005. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucl Acids Res* **33:** D247-D251.

Pearl, F.M.G., Bennett, C.F., Bray, J.E., Harrison, A.P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J., and Orengo, C.A. 2003. The CATH database: an extended protein family resource for structural and functional genomics. *Nucl Acids Res* **31:** 452-455.

Pey, A.L., Perez, B., Desviat, L.R., Martinez, M.A., Aguado, C., Erlandsen, H., Gamez, A., Stevens, R.C., Thorolfsson, M., Ugarte, M., et al. 2004. Mechanisms underlying responsiveness to tetrahydrobiopterin in mild phenylketonuria mutations. *Hum Mutat* **24:** 388-399.

Pey, A.L., Stricher, F., Serrano, L., and Martinez, A. 2007. Predicted effects of missense mutations on native-state stability account for phenotypic outcome in phenylketonuria, a paradigm of misfolding diseases. *Am J Hum Genet* **81:** 1006-1024.

Phillips, R.S., and Kaufman, S. 1984. Ligand effects on the phosphorylation state of hepatic phenylalanine hydroxylase. *J Biol Chem* **259:** 2474-2479.

Pohnert, G., Zhang, S., Husain, A., Wilson, D.B., and Ganem, B. 1999. Regulation of phenylalanine biosynthesis. Studies on the mechanism of phenylalanine binding and feedback inhibition in the Escherichia coli P-protein. *Biochemistry* **38:** 12212-12217.

Posada, D., and Crandall, K.A. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14:** 817-818.

POV-Ray.

Rackovsky, S., and Scheraga, H.A. 1977. Hydrophobicity, Hydrophilicity, and Radial and Orientational Distributions of Residues in Native Proteins. *Proc Natl Acad Sci USA* **74:** 5248-5251.

Ramsey, A.J., and Fitzpatrick, P.F. 1998. Effects of phosphorylation of serine 40 of tyrosine hydroxylase on binding of catecholamines: Evidence for a novel regulatory mechanism. *Biochemistry* **37:** 8980-8986.

Ronquist, F., and Huelsenbeck, J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572-1574.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng* **12:** 85-94.

Roth, C., Rastogi, S., Arvestad, L., Dittmar, K., Light, S., Ekman, D., and Liberles, D.A. 2007. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J Exp Zoolog B Mol Dev Evol* **308:** 58-73.

Sadqi, M., Fushman, D., and Munoz, V. 2006. Atom-by-atom analysis of global downhill protein folding. *Nature* **442:** 317-321.

Saraste, M., Sibbald, P.R., and Wittinghofer, A. 1990. The P-loop--a common motif in ATP- and GTP-binding proteins. *Trends Biochem Sci* **15:** 430-434.

Scavelli, R., Ding, Z., Blau, N., Haavik, J., Martinez, A., and Thony, B. 2005. Stimulation of hepatic phenylalanine hydroxylase activity but not Pah-mRNA expression upon oral loading of tetrahydrobiopterin in normal mice. *Mol genet metab* **86 Suppl 1:** S153-155.

Schallreuter, K.U., Kothari, S., Chavan, B., and Spencer, J.D. 2008. Regulation of melanogenesis--controversies and new concepts. *Exp Dermatol* **17:** 395-404.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C. 2003. SWISS-MODEL: An automated protein homology-modeling server. *Nucl Acids Res* **31:** 3381-3385.

Scriver, C.R. 2007. The PAH gene, phenylketonuria, and a paradigm shift. *Hum Mutat* **28:** 831-845.

Scriver, C.R., and Kaufman, S. 2001. Hyperphenylalaninemia:phenylalanine hydroxylase deficiency. . In *The Metabolic and Molecular bases of Inherited Disease.* , 8 ed. (ed. C.R. Scriver, Beaudet, A.L., Valle, D. and Sly, W.S.), pp. 1667-1724. McGraw-Hill, New York.

Shakhnovich, B.E., Dokholyan, N.V., DeLisi, C., and Shakhnovich, E.I. 2003. Functional fingerprints of folds: evidence for correlated structure-function evolution. *J Mol Biol* **326:** 1-9.

Shiman, R. 1980. Relationship between the substrate activation site and catalytic site of phenylalanine hydroxylase. *J Biol Chem* **255:** 10029-10032.

Shindyalov, I.N., and Bourne, P.E. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* **11:** 739-747.

Smith, T.F., and Waterman, M.S. 1981. Identification of common molecular subsequences. *J Mol Biol* **147:** 195-197.

Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. 1998. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucl Acids Res* **26:** 320-322.

Stanley, S.M. 1973. Ecological Theory for Sudden Origin of Multicellular Life in Late Precambrian - (Adaptive Radiation-Cambrian-Evolution-Paleontology-Predation). *Proc Natl Acad Sci USA* **70:** 1486-1489.

Stokka, A.J., Carvalho, R.N., Barroso, J.F., and Flatmark, T. 2004. Probing the role of crystallographically defined/predicted hinge-bending regions in the substrate-induced global conformational transition and catalytic activation of human phenylalanine hydroxylase by single-site mutagenesis. *J Biol Chem* **279:** 26571-26580.

Supangat, S., Park, S.O., Seo, K.H., Lee, S.Y., Park, Y.S., and Lee, K.H. 2008. Role of Phe-99 and Trp-196 of sepiapterin reductase from Chlorobium tepidum in the production of L-threo-tetrahydrobiopterin. *Acta biochim biophys Sin* **40:** 513-518.

Tan, K., Li, H., Zhang, R., Gu, M., Clancy, S.T., and Joachimiak, A. 2008. Structures of open (R) and close (T) states of prephenate dehydratase (PDT)--implication of allosteric regulation by L-phenylalanine. *J Struct Biol* **162:** 94-107.

Tanford, C., and Kirkwood, J.G. 1957. Theory of Protein Titration Curves .1. General Equations for Impenettrable Spheres. *J Am Chem Soc* **79:** 5333-5339.

Taverna, D.M., and Goldstein, R.A. 2002. Why are proteins marginally stable? *Proteins* **46:** 105-109.

Taylor, W.R. 2007. Evolutionary transitions in protein fold space. *Curr Opin Struct Biol* **17:** 354-361.

Taylor, W.R., Flores, T.P., and Orengo, C.A. 1994. Multiple protein structure alignment. *Protein Sci* **3:** 1858-1870.

Teigen, K., and Martinez, A. 2003. Probing cofactor specificity in phenylalanine hydroxylase by molecular dynamics simulations. *J Biomol Struct Dyn* **20:** 733-740.

Teigen, K., McKinney, J.A., Haavik, J., and Martinez, A. 2007. Selectivity and affinity determinants for ligand binding to the aromatic amino acid hydroxylases. *Curr Med Chem* **14:** 455-467.

Thony, B., Auerbach, G., and Blau, N. 2000. Tetrahydrobiopterin biosynthesis, regeneration and functions. *Biochem J* **347 Pt 1:** 1-16.

Thorolfsson, M., Ibarra-Molero, B., Fojan, P., Petersen, S.B., Sanchez-Ruiz, J.M., and Martinez, A. 2002. L-phenylalanine binding and domain organization in human phenylalanine hydroxylase: a differential scanning calorimetry study. *Biochemistry* **41:** 7573-7585.

Thorolfsson, M., Teigen, K., and Martinez, A. 2003. Activation of phenylalanine hydroxylase: Effect of substitutions at Arg68 and Cys237. *Biochemistry* **42:** 3419-3428.

Tourian, A. 1971. Activation of phenylalanine hydroxylase by phenylalanine. *Biochim Biophys Acta* **242:** 345-354.

Tsai, C.J., del Sol, A., and Nussinov, R. 2008. Allostery: absence of a change in shape does not imply that allostery is not at play. *J Mol Biol* **378:** 1-11.

Volner, A., Zoidakis, J., and Abu-Omar, M.M. 2003. Order of substrate binding in bacterial phenylalanine hydroxylase and its mechanistic implication for pterin-dependent oxygenases. *J Biol Inorg Chem* **8:** 121-128.

Wagner, A. 2005. Energy constraints on the evolution of gene expression. *Mol Biol Evol* **22:** 1365-1374.

Walther, D.J., and Bader, M. 2003. A unique central tryptophan hydroxylase isoform. *Biochem Pharmacol* **66:** 1673-1680.

Werner-Felmayer, G., Golderer, G., and Werner, E.R. 2002. Tetrahydrobiopterin biosynthesis, utilization and pharmacological effects. *Curr Drug Metab* **3:** 159-173.

Whelan, S., and Goldman, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18:** 691-699.

Wiens, M., Koziol, C., Batel, R., and Muller, W.E. 1998. Phenylalanine hydroxylase from the sponge Geodia cydonium: implication for allorecognition and evolution of aromatic amino acid hydroxylases. *Dev Comp Immunol* **22:** 469-478.

Williams, R.A., Mamotte, C.D., and Burnett, J.R. 2008. Phenylketonuria: an inborn error of phenylalanine metabolism. *Clin Biochem Rev* **29:** 31-41.

Winge, I., McKinney, J.A., Ying, M., D'Santos, C.S., Kleppe, R., Knappskog, P.M., and Haavik, J. 2008. Activation and stabilization of human tryptophan hydroxylase 2 by phosphorylation and 14-3-3 binding. *Biochem J* **410:** 195-204.

Wolynes, P.G. 2007. Lectures on biomolecular energy landscapes. In *Protein Folding and Drug Design*. (ed. R.A. Broglia, Serrano, L., and Tiana, G.). IOS Press, Amsterdam, Oxford, Tokyo, Washinton DC.

Wretborn, M., Humble, E., Ragnarsson, U., and Engstrom, L. 1980. Amino-Acid-Sequence at the Phosphorylated Site of Rat-Liver Phenylalanine-Hydroxylase and Phosphorylation of a Corresponding Synthetic Peptide. *Biochem Biophys Res Commun* **93:** 403-408.

Wright, C.F., Teichmann, S.A., Clarke, J., and Dobson, C.M. 2005. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature* **438:** 878-881.

Yang, Z.H. 1994. Maximum-Likelihood Phylogenetic Estimation from DNA-Sequences with Variable Rates over Sites - Approximate Methods. *J Mol Evol* **39:** 306-314.

Ye, Y., and Godzik, A. 2003. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* **19 Suppl 2:** ii246-255.

Zeldovich, K.B., Berezovsky, I.N., and Shakhnovich, E.I. 2006. Physical origins of protein superfamilies. *J Mol Biol* **357:** 1335-1343.

Zuckerkandl, E., and Pauling, L. 1965. Molecules as documents of evolutionary history. *J Theor Biol* **8:** 357-366.