The Chemistry and Evolution of Enzyme Function:

Isomerases as a Case Study

Sergio Martínez Cuesta

EMBL - European Bioinformatics Institute

Gonville and Caius College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

31st July 2014

This dissertation is the result of my own work and contains nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

No part of this dissertation has been submitted or is currently being submitted for any other degree or diploma or other qualification.

This thesis does not exceed the specified length limit of 60.000 words as defined by the Biology Degree Committee.

This thesis has been typeset in 12pt font using LaTeX according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

Cambridge, 31st July 2014 Sergio Martínez Cuesta

To my parents and my sister

Contents

Abstract

The study of the evolution of proteins has been traditionally undertaken from a sequence and structural point of view. However any attempt to understand how protein function changes during evolution benefits from consistent definitions of function and robust approaches to quantitatively compare them. The function of enzymes is described as their ability to catalyse biochemical reactions according to the Enzyme Commission (EC). This dissertation explores aspects of the chemistry and evolution of a small class of enzymes catalysing geometrical and structural rearrangements between isomers, the isomerases (EC 5).

A comprehensive analysis of the overall chemistry of isomerase reactions based on bond changes, reaction centres and substrates and products revealed that isomerase reactions are chemically diverse and difficult to classify using a hierarchical system. Although racemases and epimerases (EC 5.1) and cis-trans isomerases (EC 5.2) are sensibly grouped according to changes of stereochemistry, the overall chemistry of intramolecular oxidoreductases (EC 5.3), intramolecular transferases (EC 5.4) and intramolecular lyases (EC 5.5) is challenging. The subclass “other isomerases” (EC 5.99) sits apart from other subclasses and exhibits great diversity. The current classification of isomerases in six subclasses reduces to two subclasses if the type of isomerism is considered. In addition, the separation of groups of isomerases sharing similar chemistry such as oxidosqualene cyclases and pseudouridine synthases from chemically complex sub-subclasses like intramolecular transferases acting on “other groups” (EC 5.4.99) might also improve the classification.

An overview of the evolution of isomerase function in superfamilies revealed three main findings. First, isomerases are more likely to evolve new functions in different EC primary classes, especially lyases (EC 4), rather than evolve to perform different isomerase reactions. Second, isomerases change their overall chemistry and conserve the structure of their substrates and products more often than conserving the chemistry and changing substrates and products. Last, the relationship between sequence and functional similarity suggests that correlations should be investigated on the basis of closely related enzymes.

Although previous research assumes a one-to-one relationship between EC number and biochemical reaction, almost one-third of all known EC numbers are linked to more than one biochemical reaction. This complexity was characterised for isomerase reactions and used to develop an approach to automatically explore it across the entire EC classification. Remarkably, about 30% of the EC numbers bearing more than one reaction are linked to different types of reactions, bearing key differences in catalysed bond changes. Several recommendations to improve the description of complex biochemical reaction data in the EC classification were proposed.

This dissertation explores enzymes from a functional perspective as an alternative to classical studies based on homology. This standpoint might prove useful to help to search for sequence candidates for orphan enzymes and in the design of enzymes with novel activities.

Acknowledgements

Hello, my name is Bond, Change Bond.

My scientific endeavours during the last four years would not have arrived at this climax without the support of collaborators, friends and family. First, I would like to thank my supervisor Prof. Janet M Thornton for giving me the opportunity to do research in her group. Janet's guidance and encouragement have been a constant source of motivation for the delivery of this piece of work. After more than three years of being part of her research family, I am still day after day amazed by her ability to transfer energy to people in such an effective manner. Special thanks go to Dr. Syed Asad Rahman for allowing me to contribute to the EC-BLAST project and for overall support, limitless help and fruitful discussions on enzyme literature, which paved the way to build my research interests. Great thanks go to Dr. Nicholas Furnham for our scientific exchange while exploring the evolution of isomerase function. I would also like to acknowledge Handan Melike Dönertaş, who came all the way from Turkey as a visiting student to explore complexities in the description of enzyme function and whose contribution was fundamental to complete this dissertation. It has been a privilege to work with all of you.

During the course of this project, the interaction with some of you has helped to develop my need to manage and appreciate feedback and communication. I am grateful to the members of my Thesis Advisory Committee - Dr. Christoph Steinbeck, Dr. Anne-Claude Gavin and Prof. Robert Glen - for constructive critique, especially during our July 2012 meeting, which meant a turning point in the way I perform research. I am also thankful to Dr. Florian Hollfelder and Dr. John BO Mitchell for being my thesis examiners. Back to the Thornton group, I want to thank my colleagues Dr. Nidhi Tyagi for our ongoing search to understand human nature and Dr. Matthias Ziehm for limitless help and support. From the protein structure team, thanks to Dr. Roman A. Laskowski and Dr. Tjaart de Beer (now in Switzerland) for technical discussions. Greetings also go to Dr. Gemma L. Holliday and Dr. Julia D. Fischer, who introduced me to the world of enzymes when I first joined the group. In retrospect, I am also grateful to Prof. José Cristobal Martínez Herrerías for his help back during my chemistry studies at Universidad de Granada and his support, which guided me to apply to EMBL. Then, I would like to thank present and past members of the Thornton group and members of the Director's and Research office for contributing to my well-being, especially Dr. Dobril Ivanov for our hilarious conversations over lunch, Mrs. Deborah Howe and Mrs. Tracey Andrew for administrative support. Finally, I would like to acknowledge the EMBL International PhD Programme for accepting me as a predoc student and for providing financial support to cover my living expenses during the last four years. Last but not least, I am indebted to the University of Cambridge and Gonville and Caius College for allowing me to be a member of such great academic institutions.

Friends mean freedom. To Anna, Bruno and Claire for sharing my passion for cooking and being an unlimited source of social gatherings and walks alongside Paradise. To Matthias, Ulla and Nidhi for your everlasting bad jokes (see beginning and end of Acknowledgements), constant social exchange and tasty share of food. To Leonor, Mar and Rita for making the Iberian House a place to live despite the obvious differences that led us to move on. To Olya for your limitless curiosity and our exciting upcoming trip. To Chen for our adventurous exploration of Cambridge's swimming pools, restaurants and pubs. To Clem and Senay for enjoying social dancing as much as I do. To Anna and José for being such a nice neighbours during my Mill Road life. To Alberto, Isi and María José for our lifelong friendship and our yearly meetings in Poland, Germany and Spain. To predocs, especially Myrto and Nenad for useful help when I joined EMBL-EBI. To Audrey, John, Max and Sarah for being my current housemates. In many ways, you all have contributed to make my life worthwhile over the last few years.

Finally, my family deserves special recognition, especially my parents Pedro and Rosario and my sister Patricia for their constant love and support.

Take it EC!

List of Figures

List of Tables

List of Publications

Martinez Cuesta, S., Furnham, N., Rahman, S. A., Sillitoe, I. & Thornton, J. M. 2014 The evolution of enzyme function in the isomerases. *Curr. Opin. Struct. Biol.*, **26C**, 121–130. (doi: 10.1016/j.sbi.2014.06.002)

Rahman, S. A., Cuesta, S. M., Furnham, N., Holliday, G. L. & Thornton, J. M. 2014 EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nat. Methods*, **11**(2), 171–174. (doi: 10.1038/nmeth.2803)

Chapter 1

Introduction

Enzymes are the catalysts of life. They accelerate biochemical reactions up to the rates at which biological processes take place in living organisms. The first enzyme to be discovered was amylase in 1833 (Payen & Persoz, 1833), which was originally isolated from barley and catalyses the conversion of starch into sugars. This discovery triggered the development of commercial breadmaking and brewing techniques as well as the development of fermentative enzymes at the beginning of the 20th century. Sumner isolated and crystallised urease in 1926, an enzyme that catalyses the hydrolysis of urea into ammonia and carbamate (Sumner, 1926). His original findings led to the demonstration that enzymes are actually proteins, a topic of controversy at the time. As a result, he was awarded the Nobel Prize in Chemistry in 1946. Almost forty years elapsed until the first three-dimensional structure of an enzyme was determined using X-ray diffraction techniques. In 1965, Phillips and coworkers obtained the structure of the hen egg-white lysozyme, which hydrolyses peptidoglycan in bacterial cell walls (Blake *et al.*, 1965). During the 1960s, studies by Anfinsen on ribonuclease shed light on the dynamic nature of protein structure and served as a model to explore protein folding and the development of proteolytic methods. Subsequent research on the active site and catalytic mechanism of lysozyme, ribonuclease and other enzymes and further work on the stereochemistry of enzyme-catalyzed reactions by Cornforth and coworkers revealed mechanistic enzymology as a new scientific discipline.

In parallel, scientists started to capture enzymes performing function in their biochemical contexts, known as metabolic pathways. The idea of cataloguing and representing everything known about metabolism emerged from the original illustrations depicted by Krebs of the citric acid cycle (Krebs, 1940), which later translated into the elaboration of comprehensive wall charts of metabolic pathways (Reitz *et al.*, 2004) that still exist today in many biochemistry labs. Current efforts to make metabolic data publicly available in online repositories such as KEGG are prevalent (Kanehisa *et al.*, 2012). For instance, the ten enzymes present in the glycolysis pathway make the biochemical transformation of glucose into pyruvate possible. First elucidated by Meyerhof and coworkers in 1940 (Kresge *et al.*, 2005), this process meets essential demands such as the supply of adenosine triphosphate (ATP) and other key biosynthetic intermediates to many cellular activities (Bar-Even *et al.*, 2012).

The ability of enzymes to perform biochemical catalysis has been traditionally described in textbooks with two main characteristics: acceleration of reaction rate and specificity (Fersht, 1999; Silverman, 2002). Whereas overwhelming evidence from kinetics studies has extensively supported the former over many decades, exceptions to the latter formalised into the concept of enzyme promiscuity, also known as the ability of enzymes to catalyse more than one biochemical reaction. This discovery has radically changed the way we understand enzymes and has implications across a broad range of scientific disciplines, from the evolution of enzyme function (Copley, 2003; Khersonsky & Tawfik, 2010; O'Brien & Herschlag, 1999) to the root of biotechnology and biocatalysis (Hult & Berglund, 2007; Nobeli *et al.*, 2009).

In this chapter, basic facts about the chemistry and evolution of enzymes are first described. Then, existing approaches to explore the similarity of enzymes are reviewed. Next, our subject of study, the isomerases, which are a small class of enzymes catalysing geometrical and topological rearrangements between isomers, are introduced. Finally, the structure of the thesis is presented.

1.1 Chemistry of enzymes

As declared by Silverman, "enzymes are highly efficient organic chemists". Over many decades, enzymologists have collected data on different aspects of the chemistry of enzymes and reported their findings in the primary literature: biochemical reactions, usage of cofactors, kinetic data and mechanistic interpretations (Silverman, 2002). With the aid of structural studies, researchers revealed that enzyme catalysis takes place in a buried pocket within the enzyme structure known as the active site. Experimental evidence from X-ray crystallography suggested a model to explain how this process takes place: the enzyme and substrates initially form a complex in the active site which may induce large conformational changes in their structures. First suggested by Linus Pauling, substrate binding is followed by transition state stabilisation. Enzymes decrease the activation energy of the reaction because they are complementary in shape and electrostatic properties to the rate-limiting transition state, which explains the vast rate acceleration compared to uncatalysed reactions (Pauling, 1946). One or more reaction intermediates are generated which then turn into products and are finally released from the active site.

As the substrate approaches the enzyme, molecular recognition takes place and the complex enzyme-substrate arises from interactions of the substrate with various amino acids in the active site. There are two types of interactions driving the complex formation: covalent interactions, which involve the sharing of electrons; and non-covalent interactions comprising electrostatic, dipole, hydrogen bonding, hydrophobic and van der Waals forces. Non-covalent interactions also contribute to enzyme catalysis by stabilising the transition state and destabilising the ground state. Although individually weak, they collectively make a strong interaction that is maximum when the transition state is formed. In addition, their reversibility also allows the product to be released from the enzyme active site (Silverman, 2002). Despite the vast research aimed at building basic principles of enzyme catalysis, we still know very little about certain aspects of this phenomenon. One of the areas that has attracted considerable attention in recent years is the connection between enzyme motions and catalysis (Hammes-Schiffer, 2013).

1.1.1 Catalytic sites, mechanisms and cofactors

The active site of an enzyme comprises about ten to twelve amino acids. About three or four of them, known as catalytic amino acids, are directly involved in the catalysis of the biochemical reaction and form the catalytic site (Gutteridge & Thornton, 2005). Their role is defined according to the specific chemical function they performed in the mechanism. In general, catalytic amino acids tend to be conserved, they show slight preference to be located in coil regions of secondary structure and they exhibit limited solvent accessibility (Bartlett *et al.*, 2002). Histidine, cysteine and aspartate are the amino acids more often engaged in catalytic activities, whereas aliphatic amino acids such as alanine, leucine and glycine are rarely involved (Holliday *et al.*, 2009). Their most common functions are transition state stabilisation, general acid/base (proton donor and acceptor) and nucleophilic covalent catalysis (Bartlett *et al.*, 2002; Holliday *et al.*, 2009). In multiple occasions residues act in combination forming catalytic units, for instance some residues may polarise the substrates or orientate other residues in order to maximise the probability that catalysis occurs (Gutteridge & Thornton, 2005).

Information extracted from the literature about catalytic residues is available in public resources such as the Catalytic Site Atlas (CSA) (Furnham *et al.*, 2014). This repository provides curated annotations on the residues that are engaged in catalytic activity in enzyme structures deposited in the Protein Data Bank (PDB) (Berman *et al.*, 2013). Several computational strategies integrate these annotations for the inference of catalytic residues using sequence-based approaches that exploit sequence conservation or more sophisticated algorithms employing three-dimensional templates of the atoms present in the catalytic site (Barker & Thornton, 2003; Nosrati & Houk, 2012). These techniques proved to be useful in predicting enzyme function from sequence and structure data (Laskowski *et al.*, 2005*b*). Alternatively, researchers have also built repositories of catalytic site information for specific groups of enzymes (Gariev & Varfolomeev, 2006).

Enzymes use several types of chemical strategies to convert substrates into products - these are called mechanisms. Holliday and coworkers adapted an extended version of Ingold's classification of mechanisms (Holliday *et al.*, 2007*a*) and used it to build a database of enzyme mechanisms called MACiE (Holliday *et al.*, 2012). A comprehensive analysis of the database revealed that more than two-thirds of enzyme reactions rely on acid/base chemistry, especially proton transfer reactions (Holliday *et al.*, 2007*b*). The second most common mechanism is nucleophilic catalysis such as substitutions, additions and eliminations, whereas electrophilic reactions are rare. The chemical bonds O–H, N–H and C–H are commonly cleaved and formed in enzyme reactions.

To gain further insight into the enzyme mechanism, biochemists perform kinetics experiments to measure two key parameters: the Michaelis-Menten constant ($\mathrm{K_M}$) and the catalytic rate constant ($\mathrm{k_{cat}}$) (Stitt & Gibon, 2014; Wittig *et al.*, 2014). In general, enzyme reactions are described by Michaelis-Menten kinetics where $\mathrm{K_M}$ captures the substrate binding affinity to the enzyme and links the reaction rate with the substrate concentration. Secondly, the catalytic rate constant captures the speed whereby the substrate turns into product in the enzyme active site (Cornish-Bowden, 2014). Studies comparing $\mathrm{k_{cat}}$ with the rate constant of the uncatalysed reaction ($\mathrm{k_{uncat}}$) strongly support the idea of acceleration of reaction rate achieved in enzyme reactions. Some enzymologists prefer combining these two parameters in a ratio known as the specificity constant ($\mathrm{k_{cat}/K_M}$), which measures the enzyme's ability to discriminate among competing substrates. Nevertheless when $\mathrm{k_{uncat}}$ data are available, the so-called catalytic proficiency ($\mathrm{(k_{cat}/K_M)/k_{uncat}}$) is also common in the literature (O'Brien & Herschlag, 1999).

The enzymatic catalysis of many biochemical reactions requires the presence of cofactors. A cofactor is an organic molecule or metal ion that binds to the active site and is essential for catalysis. The use of cofactors allows enzymes to expand the catalytic abilities achieved by the 20 naturally occurring amino acids (Holliday *et al.*, 2007*b*), yet not all enzymes require a cofactor. Organic cofactors are mainly composed by nucleotides, amino acids and fatty acids substructures. For instance, glutathione consists of three amino acids glutamate, cystein and glycine. Although they chemically resemble other metabolites in the cell, cofactors are on average significantly more polar and slightly larger (Fischer *et al.*, 2010).

1.1.2 Enzyme classification

During the 20th century, the nomenclature and classification of enzymes has constantly being under debate. In the early days, enzymes were given trivial names in order to uniquely identify and distinguish them from the rest. Although trivial names were chosen by groups of biochemists, multiple trivial names were sometimes given to the same enzyme by different groups, likewise different enzymes were named the same way, which led to confusing and ambiguous discussions (Cornish-Bowden, 2014). For example, NADPH dehydrogenase was first known as "NADPH diaphorase" and "old yellow enzyme" due to its ability to reduce various dyes, both trivial names still persist today (Daugherty *et al.*, 2013; Savignon *et al.*, 2012). Soon after, D-amino acid oxidase was designated as "new yellow enzyme" and distinction between both enzymes became even more difficult. The remarkable increase in the number of newly discovered enzymes led to the development of a system to name and classify them in a consistent manner. Since 1956, the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) has been the committee responsible to name old enzymes according to which new enzymes could be classified. In an attempt to develop a unified system, the Enzyme Commission (EC) classification was created and enzymes are now named and identified systematically with an EC number; this code is a four-level description that is used to classify enzymes depending on the overall chemical transformation of substrates into products (Tipton & Boyce, 2000). The first level corresponds to six different classes according to the type of chemistry being carried out. Oxidoreductases catalyse oxidation/reduction reactions (EC 1), transferases transfer a chemical group (EC 2), for example, a methyl or glycosyl moiety from one compound to another; hydrolases perform hydrolysis of chemical bonds (EC 3), lyases also cleave chemical bonds by other means than by oxidation or hydrolysis (EC 4), isomerases catalyse geometric and structural changes between isomers (EC 5) and lastly, ligases join two compounds with associated hydrolysis of a nucleoside triphosphate molecule (EC 6). These EC classes are further divided in subclasses and sub-subclasses (second and third level, respectively) in line with a variety of criteria such as the chemical bond cleaved or formed, the reaction centre, the transferred chemical group and the cofactor used for catalysis. The final level of classification defines substrate specificity. For example, alanine racemase is identified as EC 5.1.1.1 and the four digits indicate the following: EC 5 refers to isomerases, EC 5.1 are racemases and epimerases, EC 5.1.1 are racemases and epimerases acting on amino acids or derivatives and finally, EC 5.1.1.1 indicates racemisation of alanine as substrate.

One of the goals of biochemistry and enzymology is to discover the molecular function of enzymes. Although strictly speaking EC numbers do not refer to enzymes but to reactions catalysed by enzymes, the assignment of EC numbers to enzymes is now a common routine in the functional annotation of proteins and protein-coding genes in databases such as UniprotKB (The Uniprot Consortium, 2013) and Ensembl (Kersey *et al.*, 2014). The EC classification serves as a bridge between biochemistry and genomics (Kotera *et al.*, 2004) and has proved to be very useful. In this dissertation, we have explored alternatives to improve the description of biochemical reactions in the EC classification following recent developments on the automatic search and comparison of enzyme reactions based on chemical attributes (Rahman *et al.*, 2014).

1.2 Evolution of enzyme function

The ability of organisms to adapt to the changing conditions of their habitat is crucial to guarantee their survival and reproduction. For example, the capability of bacteria to acquire resistance for drugs and pesticides. At the metabolic level, this process of adaptation is related to the ability of enzymes to evolve beneficial functions in an environment of changing chemical conditions (Copley, 2012). Multiple research studies have unveiled the genetic mechanisms whereby innovation in enzyme function emerges. In its simplest form, the genetic diversity driving adaptation relies on the accumulation of point mutations. Although the majority of them are neutral or deleterious, gain-of-function mutations may create a new promiscuous activity in an existing enzyme. Subsequently, beneficial mutations might either increase the level of the promiscuous activity or when occurring in regulatory regions, they might enhance gene expression and therefore boost the cellular concentration of the enzyme up to physiological levels (Schulenburg & Miller, 2014). If the new activity provides a selective advantage to the organism in its ability to perform a fundamental biological process such as the competition for resources, beneficial mutations that enhance this activity will gradually strengthen the fitness of the organism. Then the function might either remain as a promiscuous activity of a multifunctional enzyme or segregate as the primary activity of a new enzyme through evolutionary processes such as gene duplication and specialisation (Glasner *et al.*, 2006*b*). These observations are captured in a model known as Innovation-Amplification-Divergence (IAD) (Copley, 2012), which is considered to be iterative so the newly evolved enzyme may develop new promiscuous functions leading to further adaptive cycles (Kaltenbach & Tokuriki, 2014).

In this context, the evolution of new enzyme functions from the repertoire of existing activities plays a fundamental role on how organisms are able to perform adaptation. A model revealing promiscuous enzymes as evolutionary intermediates has been recently presented where the process of functional innovation is explained as a transition from specialised to promiscuous enzymes and *vice versa* (Khersonsky & Tawfik, 2010). Evidence from directed evolution experiments showed how mutations can enhance a promiscuous activity while maintaining the primary activity almost intact (Aharoni *et al.*, 2005).

Promiscuous activities are usually more plastic to mutations than primary ones, which are robust because they have been longer under selective pressure. From a structural perspective, the acquisition of new enzyme functions is currently explained by stability-function trade-offs (Socha & Tokuriki, 2013). A recent study showed that an enzyme having an active site hard-wired to the structural scaffold is less likely to evolve new functionality than enzyme bearing an active site with low density of contacts and good separation from the scaffold. The latter scenario facilitates the evolution of new enzyme functions (Dellus-Gur *et al.*, 2013).

Previous studies focusing on analysing enzyme superfamilies (Gerlt & Babbitt, 2001; Todd *et al.*, 2001) and directed evolution experiments (Khersonsky & Tawfik, 2010) discovered aspects of how the evolution of enzyme function is influenced by aspects of the chemistry of enzymes. The overall chemical reaction is often changed by recruiting different catalytic residues within an active site, whilst conserving a few residues required for the catalysis of at least one mechanistic step of the overall reaction (Bartlett *et al.*, 2003; Galperin & Koonin, 2012). Similarly, binding different substrates is commonly achieved by changing the residues involved in substrate binding and conserving residues involved in the overall reaction (Nobeli *et al.*, 2005). There is substantial evidence supporting changes of the overall chemical reaction (Gerlt *et al.*, 2012), as well as results reporting the importance of binding different substrates in the evolution of function in superfamilies (Furnham *et al.*, 2012*a*). Commonly, enzyme superfamilies evolve by a combination of these two strategies: chemistry-driven and substrate-binding-driven evolution (Babbitt, 2003; Chiang *et al.*, 2008). For instance, phosphate binding sites are often conserved, whilst the rest of the substrate can be changed during evolution (Amyes & Richard, 2013; Khersonsky *et al.*, 2011).

Other comprehensive studies on the variation of enzyme sequence and structure (Meng & Babbitt, 2011; Pandya *et al.*, 2013) and plasticity of active sites (Dessailly *et al.*, 2013; Todd *et al.*, 2002) have also been fundamental in understanding how homologous enzymes accommodate alternative chemistries. Similarly, research on the convergent evolution of enzyme mechanisms (Almonacid *et al.*, 2010) and active sites (Gherardini *et al.*, 2007) presented nature's strategies to evolve different structural solutions for the catalysis of similar reactions (Almonacid & Babbitt, 2011; Elias & Tawfik, 2012; Galperin & Koonin, 2012). The widespread interest in understanding the evolution and chemistry of enzymes has led to large-scale collaborative projects such as the Enzyme Function Initiative (EFI) (Gerlt *et al.*, 2011) which aims to determine enzyme function using both experimental and computational approaches. Starting from a comprehensive alignment of genomic regions, Zhao and co-workers from the EFI have identified the epimerase activity, pathway context and biological role in osmoprotection of a structurally characterised enzyme of unknown function from *Pelagibaca bermudensis* using a combination of virtual screening, metabolomics, transcriptomics and biochemical experiments (Zhao *et al.*, 2013).

1.3 Similarity between enzymes

Most studies exploring the similarity between enzymes share a common theme. They assume some form of correlation between protein sequence and function (Friedberg, 2006). Although the limits of this assumption have been explored extensively, our ability to investigate this correlation is limited by the available methods that measure similarity between enzymes and the quality of the experimental annotations in enzyme data. Researchers compare enzyme sequences and structures in order to extract relationships of common ancestry (homology), which arise due to speciation (orthology) or duplication events (paralogy) (Gabaldón & Koonin, 2013). Orthologous enzymes catalyse the same biochemical reaction in different species and their accurate identification is considered a crucial step in the automatic assignment of enzyme function in newly sequenced genomes. These strategies are useful to understand how enzymes evolve, however the advent of computational approaches to measure similarity between enzymes based on the genomic context, biochemical reactions and mechanisms adds a new dimension to existing evolutionary studies based on sequences and structures.

1.3.1 Comparing sequences and structures

In general, enzymes are regarded as similar if they are evolutionary related as measured by comparing their sequences and structures. During evolution, mutations occurring in enzyme sequences translate into changes in their three-dimensional structures. Early studies presented comparative analyses of the sequences and structures of homologous enzymes (Chothia & Lesk, 1986). The emergence of protein databases (Berman *et al.*, 2013; The Uniprot Consortium, 2013) and tools to perform sequence comparisons (Altschul *et al.*, 1990) allowed researchers to define “average” sequence similarity thresholds for the accurate transfer of function between enzymes using homology (Rost, 2002; Tian & Skolnick, 2003; Todd *et al.*, 2001) and subsequent efforts integrated structural data (Laskowski *et al.*, 2005*b*) and catalytic residue conservation (George *et al.*, 2005) to help to refine this assignment. Related enzymes are nowadays grouped in families (Punta *et al.*, 2012) and superfamilies (Andreeva *et al.*, 2014; Sillitoe *et al.*, 2013) according to sequence analyses and structural similarity, respectively. These resources rely on both the manual and automatic identification of domains, which are structurally and functionally conserved regions, also evolutionary building blocks within proteins. Whereas small enzymes contain only one domain, medium and large enzymes contain two or more domains, which are arranged in sequence defining a multidomain architecture (MDA) (Tamuri & Laskowski, 2010). Evolutionary processes such as gene duplication and divergence generate different combinations of domains allowing enzymes to acquire new functionality (Bashton & Chothia, 2007). The recent explosion of sequence information in biological databases triggered the improvement of techniques that use Hidden Markov Models (HMMs) to capture fine details of the conservation of sequence and structural elements in protein domains (Eddy, 2011; Remmert *et al.*, 2012). The availability of domain relationships in public resources has allowed researchers to explore the evolution of enzyme function on a superfamily basis (Baier & Tokuriki, 2014; Huang *et al.*, 2012; Voordeckers *et al.*, 2012) or across multiple superfamilies (Furnham *et al.*, 2012*a*). For example, Furnham and colleagues explored the evolution of function across 276 enzyme superfamilies using FunTree (Furnham *et al.*, 2012*b*), a resource containing phylogenetic trees decorated with structural, functional and mechanistic information representing the evolution of enzymes in a superfamily.

Multiple approaches aim to predict the likely function of enzymes of known sequence or structure but unknown function or mechanism. These employ several sources of evidence: sequence or structure similarity (Claudel-Renard *et al.*, 2003; Desai *et al.*, 2011; Quester & Schomburg, 2011; Yu *et al.*, 2009), integration of conserved motifs, domains, catalytic and function determining residues (De Ferrari *et al.*, 2012; De Ferrari & Mitchell, 2014; Kumar & Skolnick, 2012; Laskowski *et al.*, 2005*a*; Nagao *et al.*, 2014) and other features such as amino acid composition, secondary structure content and various physicochemical properties (Bray *et al.*, 2009; Dobson & Doig, 2005; Kumar & Choudhary, 2012).

1.3.2 Comparing genomic context

Although sequence and structure methods are prevalent in evolutionary studies, the assignment of function based solely on this information is often unreliable due to the functional divergence between distant homologues in the same superfamily and functional convergence between unrelated enzymes (Omelchenko *et al.*, 2010). In some cases, homology methods have been shown to produce incorrect functional annotations in protein databases (Hsiao *et al.*, 2010; Schnoes *et al.*, 2009) and do not always work well in the reconstruction of metabolic pathways (Karp, 2004). In order to overcome this, an array of strategies borrowed from comparative genomics integrate several sources of evidence from genomic context and provide an alternative to sequence and structure approaches (Plata *et al.*, 2012). These strategies are useful to improve genome-scale metabolic reconstructions by filling metabolic gaps or missing network content (Green & Karp, 2007; Hanson *et al.*, 2010) and to predict sequences for orphan metabolic activities (Kharchenko *et al.*, 2006; Shearer *et al.*, 2014; Smith *et al.*, 2012; Yamada *et al.*, 2012; Yamanishi *et al.*, 2007). Researchers have discovered four main types of relationships shared by genes that encode for enzymes belonging to the same metabolic pathway.

- They are more likely to be in close physical proximity on the chromosome. This phenomenon is more common in prokaryotes due to the existence of operon structures and regions transcribed by the same promoter (chromosomal gene clustering).
- They often share the same regulatory patterns, which lead to similar levels of gene expression (gene co-expression).
- They are either all present or all absent in an organism (phylogenetic profiles).
- They sometimes undergo fusion events, which translate into multifunctional enzymes or protein complexes acting on the same metabolic pathway (gene fusion).

The power of these approaches depends on the appropriate combination of these sources of evidence in order to find functional associations between the gene of unknown function and a database of genes. This strategy might help to prioritise the experimental tests aimed at discovering missing biological activities (El Yacoubi & de Crécy-Lagard, 2014). Multiple resources have been developed to accomplish this goal. For example, STRING is a database of functional associations between genes derived from genomic context, protein-protein interactions and literature. These are pre-computed and updated according to new releases of the source databases (Franceschini *et al.*, 2013). Several efforts to extend genomic context associations with chemical information are currently in progress (May *et al.*, 2013; Yamanishi *et al.*, 2005).

1.3.3 Comparing biochemical reactions and mechanisms

During the rise of chemoinformatics as a scientific discipline, the idea to quantitatively compare small molecules developed into the advent of metrics to estimate chemical similarity. This concept proved useful in research areas such as drug discovery and chemical retrieval systems and by extension, the first approaches to measure enzyme similarity on the basis of their catalysed reactions first relied upon comparing their ligands directly (Chiang *et al.*, 2008; Izrailev & Farnum, 2004; Nobeli *et al.*, 2003, 2005). However enzyme reactions involve several molecules: substrates and products (reactants), therefore methods had to be extended in order to account for the transformation between two or more molecules. In a similar way that methodologies to calculate similarity between protein sequences and structures are necessary to understand protein evolution, finding similar enzyme reactions is a valuable tool for biochemists working in areas as diverse as chemical synthesis, enzyme reaction databases, enzyme design, metabolic network reconstruction and overall, discerning evolutionary relationships between enzyme sequence, structure and function.

Methods to measure similarity between biochemical reactions first relied on the structure of the EC classification system (Tohsato *et al.*, 2000), however current approaches were extended to include chemical attributes (Bawden, 1991; Latino & Aires-de Sousa, 2011; Rose & Gasteiger, 1994). First, some strategies depend solely on the structure of the reactants and involve changes in the overall reaction, for instance bond changes and reaction centres. Second, alternative strategies based on catalytic mechanisms use mechanistic information derived from biochemical and structural studies. Although the mechanistic approach to compare biochemical reactions describes the process of enzyme catalysis in a more comprehensive manner (Almonacid & Babbitt, 2011), the amount of mechanistic information available in the literature and databases is not comparable to the extent of overall reaction data (Kraut *et al.*, 2013).

Common approaches based on changes in the overall reaction are divided in two classes depending on whether atom-atom mapping (AAM), the one-to-one correspondence between the substrate and product atoms, is applied (Chen *et al.*, 2013; Warr, 2014). First, approaches using AAM encode reactants in different ways. Gasteiger and coworkers calculated physicochemical and topological properties on the atoms and bonds present in manually-defined reaction centres in order to perform reaction similarity and classification (Hu *et al.*, 2010; Sacher *et al.*, 2009). A similar strategy was proposed whereby reaction centres are also manually-defined and classified in reaction classes depending on the functional groups undergoing transformation (Mu *et al.*, 2011, 2006). Alternatively, a strategy known as Condensed Graph of Reaction (CGR) uses AAM to superpose atoms in substrates and products into a single pseudomolecule, which is then described in terms of substructure descriptors (de Luca *et al.*, 2012). Some commercial algorithms are also common, for instance the CLASSIFY algorithm developed by InfoChem uses AAM to create circular environments around the atoms present in the reaction centre and serves as a search engine for similar reactions in organic chemistry resources such as SciFinder (Kraut *et al.*, 2013).

The second group of approaches do not rely on AAM, yet reactants are also represented in similar ways as in the AAM strategies. Some approaches use physicochemical and topological properties to capture differences between the structures of substrates and products (Latino & Aires-de Sousa, 2006; Latino *et al.*, 2008). Researchers associated with KEGG database (Kanehisa *et al.*, 2012) developed a method to automatically detect the reaction centre using structure comparisons between reactant pairs and define reaction patterns to measure chemical similarity between biochemical reactions (Kotera *et al.*, 2004). In a later study, they used these patterns to predict EC sub-subclasses (Yamanishi *et al.*, 2009). Other strategies encoded biochemical reactions as differences between the molecular signatures (Faulon *et al.*, 2008; Hu *et al.*, 2012) or ^{1}H NMR spectra

(Latino & Aires-de Sousa, 2007) of substrates and products.

The existing methods proved useful in the development of databases of metabolites and biochemical reactions (Kanehisa *et al.*, 2012), databases of substructures and functional groups (Kotera *et al.*, 2008), to evaluate the performance of AAM algorithms (Muller *et al.*, 2012), to discover inconsistencies in the EC classification (Apostolakis *et al.*, 2008; Egelhofer *et al.*, 2010; Kotera *et al.*, 2004; Latino & Aires-de Sousa, 2009), to search and predict metabolic pathways (Hatzimanikatis *et al.*, 2005; Moriya *et al.*, 2010; Oh *et al.*, 2007; Tohsato & Nishimura, 2009) and finally, to assign standard Gibbs energy changes (ΔG^0) to thermodynamically uncharacterised biochemical reactions (Rother *et al.*, 2010). However, the existing approaches have limitations. First, the high level of abstraction in the representation of reaction chemistry makes similarity results difficult to interpret. Second, the inability to detect changes of stereochemistry (chiral inversions and cis/trans isomerisations) in biochemical reactions was also a major disadvantage (Apostolakis *et al.*, 2008). The EC-BLAST algorithm (Rahman *et al.*, 2014) aims to overcome these drawbacks. It extracts bond changes and reaction centres using electron shift patterns as calculated using the Dugundji-Ugi model (Leber *et al.*, 2009) and it calculates changes in stereochemistry following recent developments in the Chemistry Development Kit (CDK) (Steinbeck *et al.*, 2003).

A strategy to measure enzyme similarity based on mechanistic data obtained from MACiE has also been presented (O'Boyle *et al.*, 2007). The procedure builds upon a measure of similarity between mechanistic steps based on bond changes and features of the Ingold classification of reaction mechanisms. The similarity score is derived from an alignment algorithm similar to the one used in classical methods to compare protein sequences. This strategy helped to identify cases of evolutionary convergence and divergence of catalytic mechanisms (Almonacid & Babbitt, 2011; Almonacid *et al.*, 2010), it has been recently employed to predict enzyme function from mechanism (Nath & Mitchell, 2012).

1.4 Isomerases

Scientists were already interested in isomerases back in 1970s when the three-dimensional structure of chicken triosephosphate isomerase (TIM) (EC 5.3.1.1) was first published (Banner *et al.*, 1975). This enzyme catalyses the interconversion of dihydroxyacetone phosphate (DHAP) and D-glyceraldehyde 3-phosphate (GAP), one of the three isomerisations among the total of ten biochemical reactions present in the glycolysis pathway (Bar-Even *et al.*, 2012). In the 1990s, mandelate racemase (EC 5.1.2.2) and muconate-lactonizing enzyme (EC 5.5.1.1), members of the enolase superfamily, were among the first enzymes reported to be highly structurally similar yet catalysing different overall reactions (Petsko *et al.*, 1993). Several isomerases belonging to this superfamily have been studied over the last two decades (Gerlt *et al.*, 2012). More recently, further structural and mechanistic studies on TIM (Aguirre *et al.*, 2014; Amyes & Richard, 2013) and two other isomerases, ketosteroid isomerase (EC 5.3.3.1) (Herschlag & Natarajan, 2013) and chorismate mutase (EC 5.4.99.5) (Kiss *et al.*, 2013; Silverman, 2002; Vamvaca *et al.*, 2004), have also been fundamental for understanding basic principles of enzyme catalysis. Here we review some biological relevance and applications of isomerases.

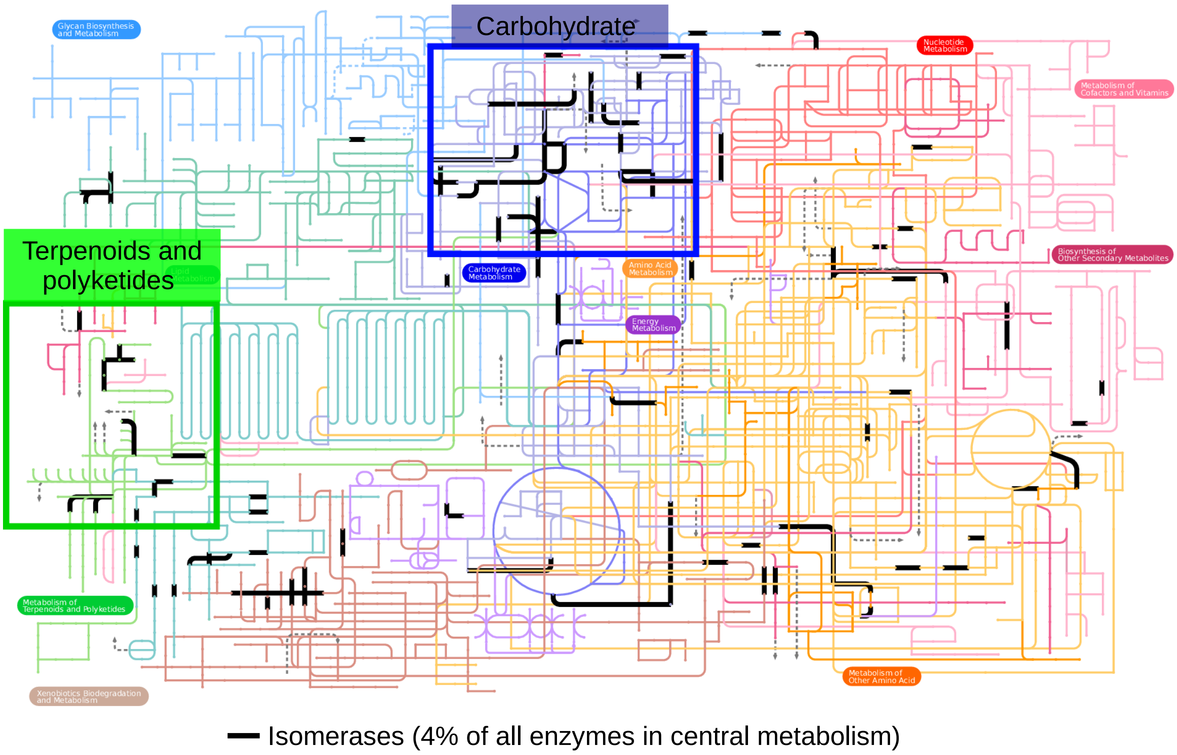

Glycan Biosynthesis and Metabolism
Carbohydrate
Nucleotide Metabolism
Metabolism of Cofactors and Vitamins
Terpenoids and polyketides
Carbohydrate Metabolism
Amino Acid Metabolism
Biosynthesis of Other Secondary Metabolites
Energy Metabolism
Metabolism of Terpenoids and Polyketides
Metabolism of Other Amino Acid
Xenobiotics Biodegradation and Metabolism
Isomerases (4% of all enzymes in central metabolism)

Figure 1.1: Metabolic pathways of the central metabolism (map01100), as in the 10th December 2012 version of KEGG database (Kanehisa *et al.*, 2012). A total of 146 pathways comprise 1590 compounds represented as nodes and 2255 reactions as edges. Different colours represent different metabolic pathways and black edges indicate isomerase reactions. Image generated using Interactive Pathways Explorer v2 (iPATH2) (Yamada *et al.*, 2011).

1.4.1 Metabolism

Isomerases are important biological components of the metabolism and genome of most living organisms. Most enzymes perform their catalytic function in the context of a metabolic pathway. Isomerases catalyse up to 4% of the biochemical reactions present in the central metabolism of most living organisms. In particular, isomerases are prevalent in carbohydrate metabolism and metabolism of terpenoids and polyketides (Figure 1.1). According to the total of 281 isomerase reactions present in the 10th December 2012 version of KEGG database (Kanehisa *et al.*, 2012), 95 (33.8%) belong to the central metabolism whereas the remaining 186 belong to other metabolic pathways that are specific to certain groups of species, such as the biosynthesis of secondary metabolites, especially in plants.

1.4.2 Genome

Genetic information is encoded as DNA sequences in the genome of living organisms. The regions of the genome that encode any kind of biological function are known as genes. These are classified depending on whether they are transcribed and/or translated, for instance, we may refer to non-transcribed regulatory genes, transcribed RNA-genes and translated protein-coding genes (Patthy, 1999). Since isomerases are enzymes and enzymes are proteins, genes with isomerase function are encoded in the genome as protein-coding genes, which then translate into proteins with isomerase function.

UniprotKB contains functional annotations of protein-coding genes, which are critically reviewed by a dedicated team of curators (The Uniprot Consortium, 2013). This resource defines "complete proteome" as the set of proteins that are thought to be expressed in fully-sequenced organisms. For a gold-standard set of five organisms, each reviewed entry corresponds to a canonical isoform of a known protein-coding gene, in addition unreviewed entries represent other isoforms of the same genes, which arise from biological processes such as alternative splicing (Light & Elofsson, 2013). These organisms are *Homo sapiens*, *Saccharomyces cerevisiae* (baker's yeast), *Schizosaccharomyces pombe*, *Escherichia coli* and *Bacillus subtilis*. For example, as in the 10th December 2012 version of UniprotKB, the complete proteome of *Homo sapiens* contains 20226 reviewed proteins and 47853 unreviewed proteins. UniprotKB assures that these many reviewed entries correspond to 20226 currently known human protein-coding genes whereas unreviewed entries equate to 47853 protein isoforms that are alternative products of these 20226 protein-coding genes.

In general, eukaryotes have more proteins than prokaryotes (Figure 1.2a). The relative proportion of enzymes encoding for isomerase activity depends on the species. Whereas 2.6% of the genes encoding for enzymatic activity correspond to isomerases in *Homo sapiens*, this proportion is higher in bacterial genomes such as *Escherichia coli* where they account for 6.2% (Figure 1.2b). These figures correlate well with the relative proportion of protein-coding genes encoding for enzymatic activity. Whereas in human, 20% of genes correspond to enzymes, this value increases to 37% in bacteria.

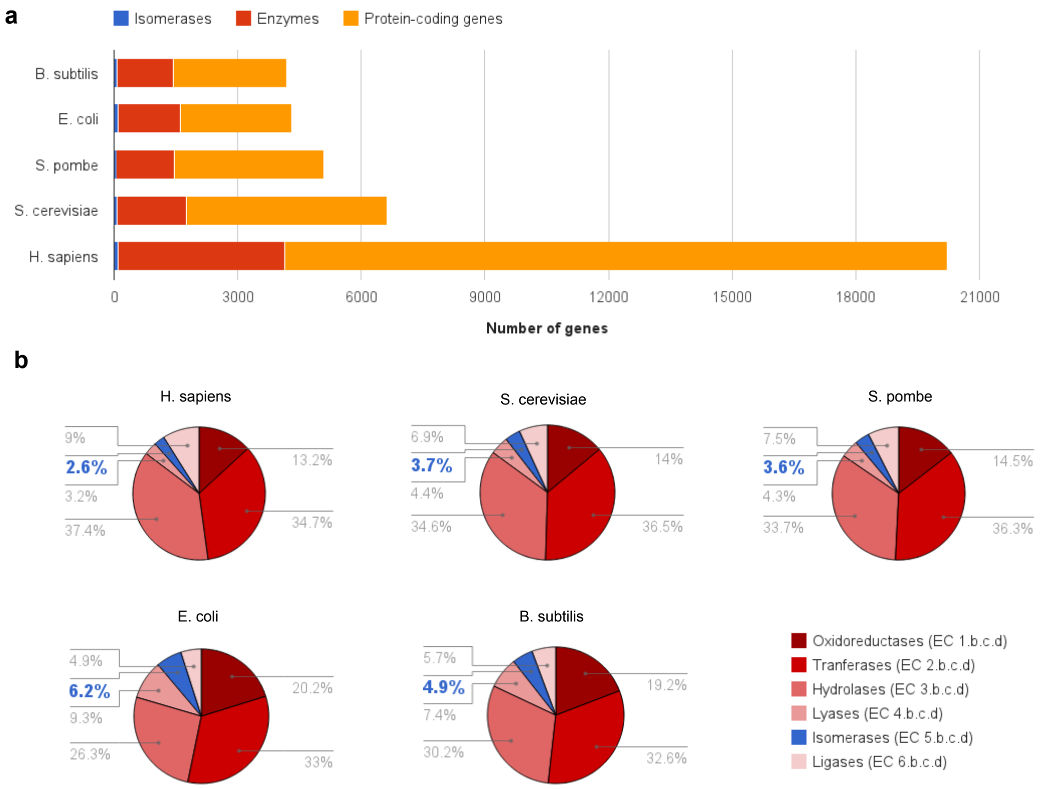


Figure 1.2: Distribution of isomerases in the genome. (a) Barplot representing the number of protein-coding genes (yellow), those annotated with enzymatic function (red) and particularly with isomerase function (blue). (b) Pie charts illustrating the relative amount of EC classes for each genome.

1.4.3 EC classification

The NC-IUBMB classifies isomerases as one of the six classes of enzymes belonging to the EC classification (McDonald & Tipton, 2014). They catalyse isomerisations, also known as geometrical and topological interconversions between isomers (Silverman, 2002). The International Union of Pure and Applied Chemistry (IUPAC) defines isomers as chemical species that share the same atomic composition but differ in their structural arrangement of atoms (IUPAC, 2014). As in the 1st November 2012, isomerases are identified as EC 5 class and classified according to three hierarchical levels: 6 subclasses (EC 5.b), 17 sub-subclasses (EC 5.b.c) and 245 EC numbers (EC 5.b.c.d) (Figure 1.3).

- **EC 5.1** consists of **racemases** and **epimerases**, which catalyse hydrogen shifts in molecules with one or more stereocentres, respectively.

- **EC 5.2** consists of **cis-trans isomerases** or enzymes catalysing geometry rearrangements in double bonds.

- **EC 5.3** consists of **intramolecular oxidoreductases**, which catalyse oxidation/reduction reactions between isomers.

- **EC 5.4** consists of **intramolecular transferases** or enzymes transferring acyl-, phospho-, amino-, hydroxy- or other groups within the substrate.
- **EC 5.5** consists of **intramolecular lyases**, which catalyse intramolecular eliminations between isomers.
- **EC 5.99** consists of **other isomerases** or enzymes catalysing different reactions to the rest of subclasses.

The distribution of isomerases into 6 subclasses depends upon two distinct criteria: type of isomerisation and overall chemistry of the reaction. Although subclasses EC 5.1 and 5.2 are defined according to the former criteria, subclasses EC 5.3, 5.4 and 5.5 are based on the latter. Lastly, EC 5.99 contains isomerases that do not fit any of the above. Three of the six isomerase EC subclasses are similar to three EC primary classes (intramolecular oxidoreductases - EC 5.3 are designated from oxidoreductases - EC 1; intramolecular transferases - EC 5.4 from transferases - EC 2; and intramolecular lyases - EC 5.5 from lyases - EC 4, but refer to intramolecular reactions). Only three subclasses are further divided in sub-subclasses depending on the chemical nature of the substrate and chemical groups involved in the reaction (Figure 1.3). While EC 5.1 was divided according to the type of substrate, EC 5.3 was split depending on the bond change and reaction centre and EC 5.4 according to the transferred chemical group.

Isomerase sub-subclasses are divided into 245 EC numbers based on the chemical nature of the substrate. For example, EC 5.1.1 comprises 18 EC numbers corresponding to enzymes that catalyse racemisations of different amino acids and derivatives. The isomerase EC numbers are associated with about 300 biochemical reactions - for example EC 5.1.1.9 describes the racemisation of arginine, lysine or ornithine and it is therefore linked to three distinct reactions. Consequently, the relationship between EC number and biochemical reaction is not one-to-one so EC numbers are associated to more than one reaction and *vice versa* (see *Chapter 5. Characterising Complex Biochemical Reaction Data*).

Isomerases were also selected as a case study due to several practical reasons. There has been already substantial research focused on other EC classes like oxidoreductases (EC 1) and hydrolases (EC 3) (Hu *et al.*, 2010; Mu *et al.*, 2006; Sacher *et al.*, 2009) and at the start of this study, other members of the Thornton group began a parallel study on ligases (EC 6), which was recently published (Holliday *et al.*, 2014). The total number of isomerase EC numbers is small compared to other EC classes, which makes them attractive for manual analysis. Most of the isomerase reactions are unimolecular (one substrate and

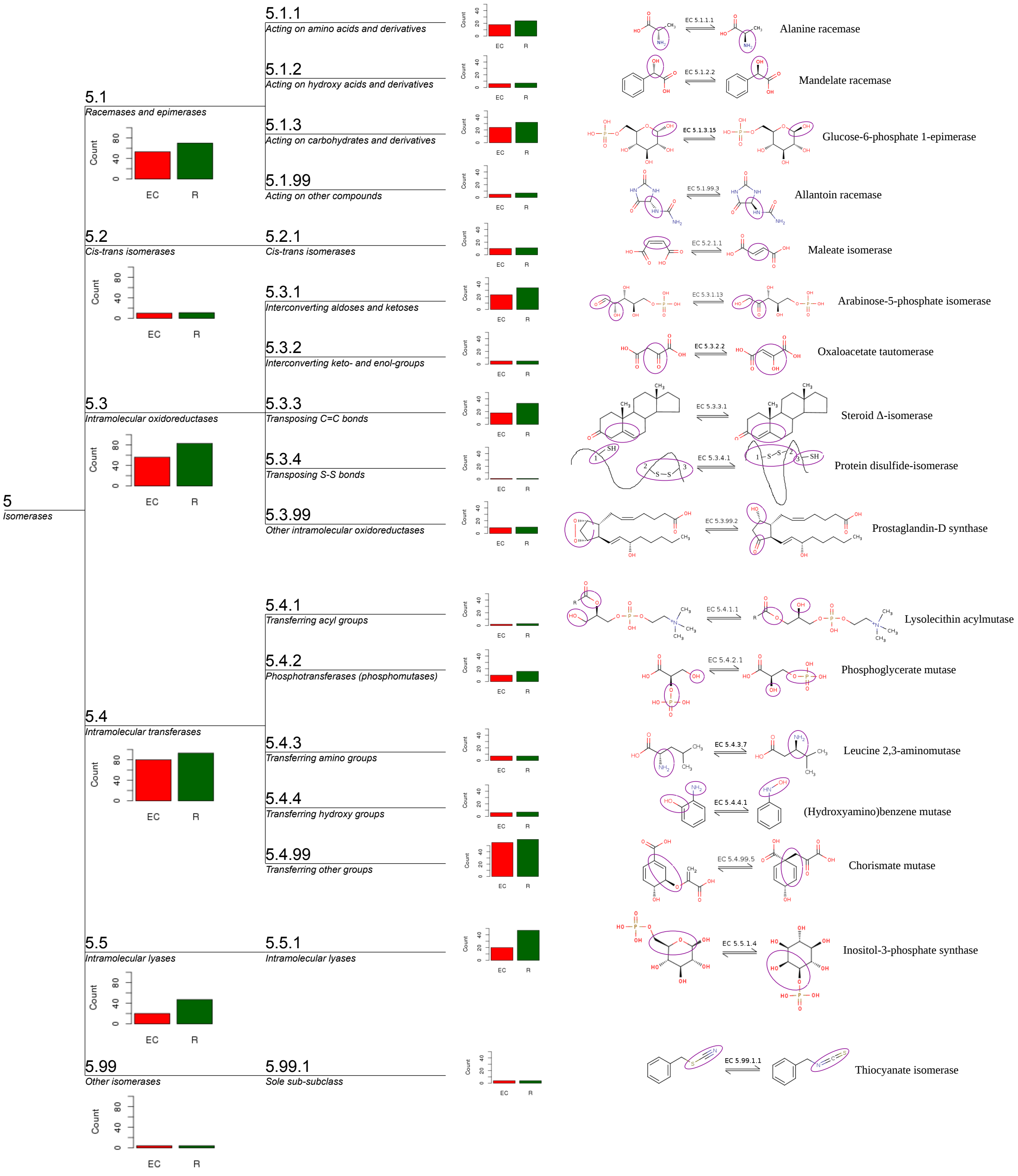


Figure 1.3: Tree representing the EC classification of isomerases. Barplots illustrate the total amount of EC numbers (red) and KEGG reactions (green) at the subclass and sub-subclass levels. An isomerase reaction is given as an example for each sub-subclass. Substructures undergoing the reaction are highlighted in purple ovals. For instance, alanine racemase (EC 5.1.1.1) interconverts L-alanine and D-alanine.

one product) and therefore relatively easy to compare. Also, the automatic detection of stereochemistry changes (chiral inversions and cis/trans isomerisations) in biochemical reactions has been a long-standing problem in chemoinformatics due to technical challenges (Apostolakis *et al.*, 2008) and problems with the stereochemical validity of metabolites in widely-used resources (Ott & Vriend, 2006). Since changes of stereochemistry are relevant in isomerase reactions and driven to overcome the existing technical limitations, isomerases were used as a test dataset in the development of EC-BLAST (Rahman *et al.*, 2014), where modules for the automatic detection of sterochemistry changes were implemented.

General strategies to assign isomerase specificity have also been recently presented (Bouvier *et al.*, 2014; Lukk *et al.*, 2012; Song *et al.*, 2007; Zhao *et al.*, 2013), as well as comparative genomic techniques to discover new isomerases in bacterial genomes (Rodionova *et al.*, 2012). Other investigations have partially explored isomerases in several superfamilies such as the haloacid dehalogenase, crotonase, vicinal oxygen chelate, amidohydrolase, alkaline phosphatase, cupin, short-chain dehydrogenase/reductase and PLP-dependent aspartate amino-transferase superfamilies (Galperin & Koonin, 2012; Glasner *et al.*, 2006*b*; Nguyen *et al.*, 2009; Singh *et al.*, 2012; Uberto & Moomaw, 2013).

1.4.4 Applications

Metabolic engineering

Metabolic engineering is the technology for the manipulation of organisms to synthesize high-value compounds of both natural and heterologous origin. Several research efforts engineered yeast and bacterial organisms for the synthesis of biofuels, an alternative to petroleum-based fuels, from cheap resources like lignocellulose, algal biomass, feedstocks and greenhouse gases (Kim *et al.*, 2012; Peralta-Yahya *et al.*, 2012). Seventy percent of lignocellulose's composition is made of sugars, particularly glucose, xylose and arabinose. Bacterial xylose isomerase (EC 5.3.1.5), which converts D-xylose into D-xylulose, was engineered using directed evolution to increase the yield of alcohol-based biofuels in yeast (Lee *et al.*, 2012; Young *et al.*, 2010). The additional ability of xylose isomerase to transform D-glucose into D-fructose was also used in the industrial production of high fructose corn syrup, a common food sweetener (Hilterhaus & Liese, 2012).

Another isomerase that has been further explored is the bacterial isopentenyl-diphosphate Δ-isomerase (IPPS, EC 5.3.3.2), which is present in the terpenoid biosynthesis pathway and catalyses the conversion of isopentenyl diphosphate (IPP) into dimethylallyl diphosphate (DMAPP). IPP and DMAPP are key precursors for the synthesis of terpenoids which are widely-used as cosmetics, pharmaceuticals such as levopimaradiene (Leonard *et al.*, 2010) and isoprenoid-based biofuels like isopentanol, farnesane, bisabolane and pinene (Peralta-Yahya *et al.*, 2012).

Organic synthesis

Asano and Hölsch thoroughly reviewed the multiple applications of isomerases to organic synthesis (Asano & Hölsch, 2012). A few examples are highlighted here. Enzymatic racemisation is considered as an alternative to existing methods of chemical racemisation, which tend to generate many undesired side products. Several racemases and epimerases (EC 5.1) are used to resolve racemic mixtures in mild chemical conditions and to synthesise stereochemically pure amino acids. For example, glutamate racemase (EC 5.1.1.3) was adopted for the large-scale production of D-glutamate, D-phenylalanine and D-tyrosine, where D-glutamate is an intermediate for the preparation of important pharmaceuticals such as penicillin derivatives (Schnell *et al.*, 2003). Amino acid racemase (EC 5.1.1.10) was used to synthesise L-tryptophan.

In an attempt to show how catalytic promiscuity can be applied to asymmetric catalysis, Vongvilai and coworkers coupled the native acylation activity of a lipase enzyme with its promiscuous racemase activity in order to develop a method for the synthesis of N-methyl α-aminonitriles, natural precursors of α-amino acids (Vongvilai *et al.*, 2011).

Enzyme design

Advances on enzyme design have been fundamental for the application of enzymes in organic chemistry, protein thermostability and the interconversion of catalytic activities (Bornscheuer *et al.*, 2012; Kiss *et al.*, 2013). At the root of this discipline lies the development of experimental techniques of directed evolution such as random mutagenesis and rational design of mutational libraries. These approaches broadly consist of introducing mutations in the enzyme followed by a selection strategy based on the activity of interest. Most engineered enzymes are hydrolases, however isomerases and ligases are under-represented (Kaltenbach & Tokuriki, 2014). Over the past few years, several attempts have anyway managed to successfully interconvert the activity of some isomerases. For example, triosephosphate isomerase (EC 5.3.1.1) and phosphoribosylanthranilate isomerase (EC 5.3.1.24) have a $(\beta\alpha)_8$-barrel fold, which was shown to be useful starting scaffold for enzyme design (Höcker *et al.*, 2001). More recently, researchers transformed racemases and epimerases acting on amino acids and derivatives (EC 5.1.1) into enzymes with lyase activity (EC 4) (Seebeck & Hilvert, 2003; Vick & Gerlt, 2007). In addition, Glasner and coworkers presented structural evidence of how evolution turned a lyase, namely o-succinylbenzoate synthases (EC 4.2.1.113), into an isomerase, N-succinylamino acid racemase (EC 5.1.1.-) in the enolase superfamily (Glasner *et al.*, 2006*a*).

Isomerases in diseases and drug discovery

Ultimately, some racemases and epimerases acting on amino acids and derivatives (EC 5.1.1) are also targets for the development of antimicrobial drugs and the treatment of neuropathological disorders (Conti *et al.*, 2011). For instance, glutamate racemase (EC 5.1.1.3) plays an essential role in the biosynthesis of peptidoglycan, a fundamental component of the bacterial cell wall. Therefore this racemase has been considered as a target for the development of antibacterial drugs (Lundqvist *et al.*, 2007).

1.5 Structure of the thesis

The following chapters present an approach to understand the link between the chemistry and evolution of isomerases and a focused study aiming to dissect the diversity in the relationship between EC number and biochemical reaction.

Chapter 2 is a description of the data resources and methods used in this thesis. It first covers how biochemical reaction data and enzyme information were collected and curated. It also describes the technical tests performed on the EC-BLAST tool to compare enzyme reactions during its development (Rahman *et al.*, 2014). Next, it presents a strategy to explore the evolution of isomerases using FunTree (Furnham *et al.*, 2012*b*) and other enzyme resources. The various statistical methods used to analyse data are discussed in the order they appear in later chapters.

Chapters 3 and 4 focus on the chemistry and evolution of isomerase function, respectively. The first presents a broad overview of the overall chemistry of isomerases followed by a discussion about existing challenges in the EC classification of isomerases. The latter explores how isomerase function evolves in enzyme superfamilies.

The relationship between EC number and biochemical reaction is complex. Chapter 5 explores the principles and challenges of this relationship and proposes recommendations for the useful description of intricate EC numbers.

Finally, chapter 6 summarises and critiques the main results and presents some ideas about future research.

Some of the contents of chapters 1, 2 and 4 have been published in Martinez Cuesta *et al.* (2014) and Rahman *et al.* (2014).

Chapter 2

Data Resources and Methods

2.1 Introduction

In this chapter, the approach used to collect and curate information about the molecular function and evolution of isomerases is presented. Also, the various methods and statistical techniques employed to analyse the data are introduced.

Publications in the scientific literature and textbooks are the most reliable means by which biochemists, enzymologists and molecular biologists have traditionally shared enzyme knowledge. In recent years, the need to make this vast wealth of information publicly available and easily accessible to the scientific community and society led to the development of resources and databases that capture data such as biochemical assays, kinetics and enzyme sequences. Even more recently, with the advent of standards for the reporting of enzyme data (Apweiler *et al.*, 2010; Tipton *et al.*, 2014), many journals now demand electronic submission of enzyme data as a routine publication practice, which facilitates integration. Alternatively, curators from resources like UniprotKB (The Uniprot Consortium, 2013) and KEGG (Kanehisa *et al.*, 2012) manually scrutinise the literature in order to extract the information that populates databases. From the perspective of a user aiming to perform a medium to large-scale analysis on a group of enzymes of interest, the starting point is usually the development of a workflow to retrieve and filter data in a semi-automatic manner. A technical description follows concerning the extraction and curation of isomerase data.

2.2 Biochemical reactions

2.2.1 Resources and dataset

The molecular function of enzymes corresponds to their ability to catalyse biochemical reactions as identified by the EC number (Tipton & Boyce, 2000) (see *1.1.2 Enzyme classi-*

fication). Biochemical reactions are available from databases such as KEGG, BRENDA and MetaCyc (Caspi *et al.*, 2014; Kanehisa *et al.*, 2012; Schomburg *et al.*, 2013*a*). In April 2014, the NC-IUBMB listed 5385 active four-digit EC numbers in the classification. Structural information about the substrates and products of the reactions was accessed on the 9th April 2014 from the 70.0+ release of KEGG using the Advanced Programming Interface (API) (Kawashima *et al.*, 2003). The data includes 6494 unique biochemical reactions that are linked to 4237 EC numbers, comprising almost 80% of all EC numbers (Figure 2.1).

As a way to summarise the diversity existing in a multi-reaction EC number, biological databases such as KEGG rely on the so-called "IUBMB reaction". This is the reaction assigned to the EC number by the NC-IUBMB in the first place, which is chosen by KEGG as the representative reaction for the group of reactions associated with the same EC number (Figure 5.7). Whereas this assignment is useful when selecting an example reaction from an EC number and it was adopted as a principle in the development of other reaction databases such as Rhea (Alcántara *et al.*, 2012), it is sometimes missing or conflicting and it also overlooks the existing diversity present in the rest of reactions. Similarly, some EC numbers are not associated to any IUBMB reaction and also, EC numbers are sometimes linked to the same IUBMB reaction, 2,3-diphosphoglycerate-dependent and independent phosphoglycerate mutases (EC 5.4.2.11 and EC 5.4.2.12) are both assigned the same IUBMB reaction comprising the isomerisation of 2-phospho-D-glycerate to 3-phospho-D-glycerate. Taken together, a more robust and consistent approach to describe multi-reaction EC numbers is necessary (see *Chapter 5. Characterising Complex Biochemical Reaction Data*). In order to facilitate the interpretation of results, our analysis of the chemistry and evolution of isomerase function (*Chapters 3* and *4*, respectively) assumes a one-to-one relationship whereby a single representative reaction uniquely designates any given isomerase EC number.

The structures of substrates and products were downloaded in MDL MOL format, these are connection tables listing atoms, their 2D coordinates and bonds in a tabular format (Warr, 2014). Explicit hydrogens were manipulated using the Molecule File Converter of Marvin version 5.9.4 (http://www.chemaxon.com) and reaction files were built and stored in Rxnfile format. Among other reaction formats available such as RDfile and RInChI (Grethe *et al.*, 2013), Rxnfile was chosen as it is the preferred input format for EC-BLAST. The Rxnfile format is an extension of the MDL MOL format. At the top of the file, a header indicates the number of substrates and products of the reaction followed by their structures.

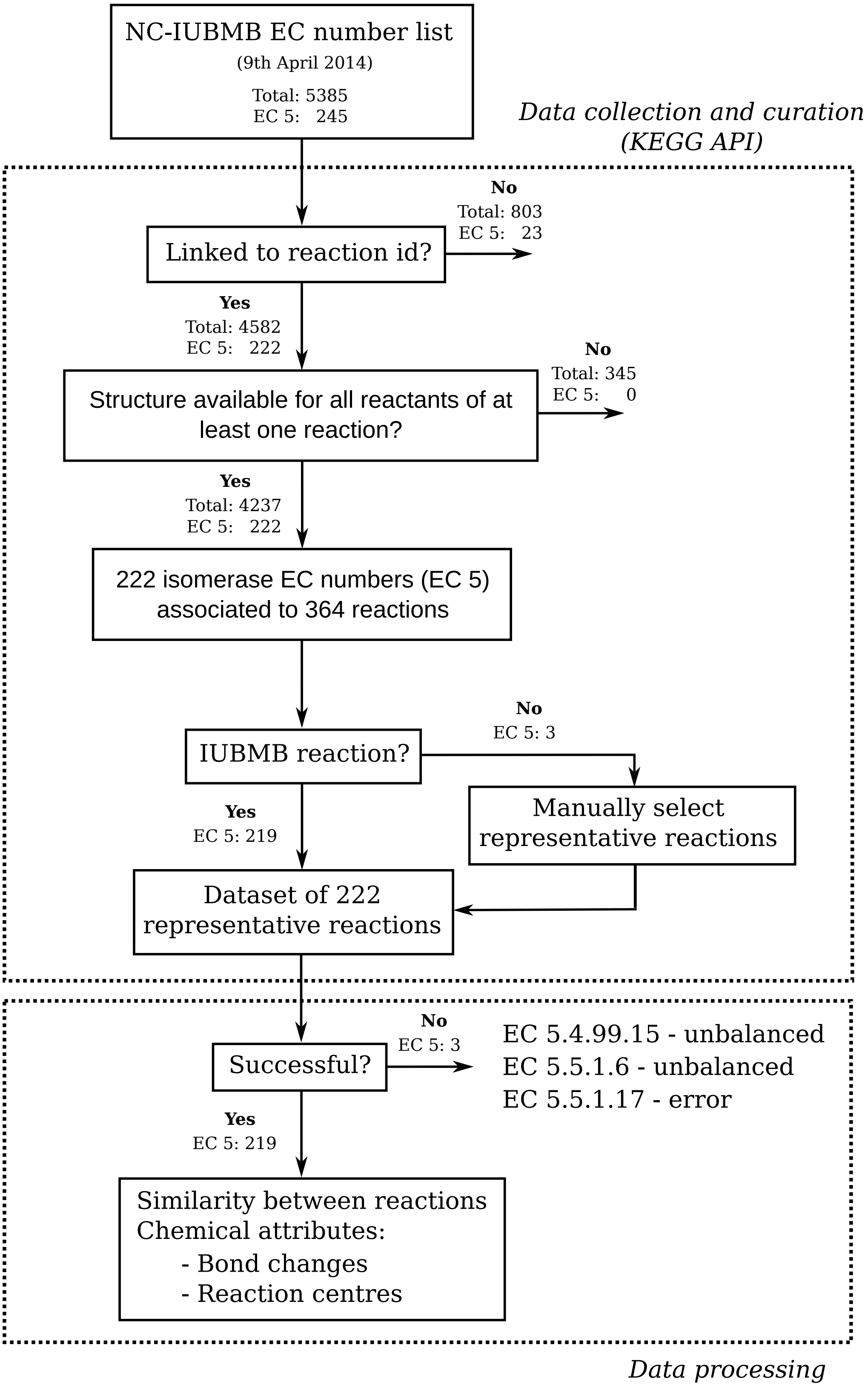


Figure 2.1: A flow diagram illustrating the collection of isomerase EC numbers (EC 5) and associated biochemical reactions from NC-IUBMB and KEGG.

2.2.2 Methods and tools

One of the multiple approaches to obtain functional similarity between enzymes is to compare their catalysed biochemical reactions. As presented in *Chapter 1. Introduction*, comparisons can be performed according to different criteria. EC-BLAST introduces three measures of functional similarity based on chemical attributes such as bond changes, reaction centres and the structures of substrates and products derived from the reaction (Rahman *et al.*, 2014) (Figure 2.2). Chemical structures were also analysed using RDKit, a widely-used chemoinformatics software (Landrum, 2013) and visualised using Marvin version 5.9.4 (http://www.chemaxon.com).

At the core of the EC-BLAST tool, four different algorithms are built upon the concept of maximum common substructure (MCS) to perform atom-atom mapping (Chen *et al.*, 2013; Rahman *et al.*, 2009). Bond changes and reaction centres are derived automatically using the Dugundji-Ugi model. The best solution is chosen on the basis of the number of bond changes and fragments generated and bond energies according to the principle of minimum chemical distance (Jochum *et al.*, 1980). In this model, substrates and products are represented using bond-electron matrices where diagonal elements correspond to the number of free valence electrons and off-diagonal entries give the bond orders between atoms. The reaction matrix is obtained when the substrates matrix is subtracted from the products matrix. A positive element on this matrix indicates bond formation, whereas a negative element corresponds to bond cleavage. Changes in the stereochemistry of atoms and bonds, also known as stereochanges, are coded in parallel.

In order to allow comparisons between the chemistry of enzymes, EC-BLAST encodes reactions using three distinct fingerprints based on bond changes, reaction centres and structures of substrates and products, respectively. Bond changes refer to the cleavage and formation of chemical bonds, changes in bond order and stereochanges, which are due to chemical processes such as chiral inversions or cis-trans isomerisations. Cleaved and formed bonds are indicated as lines connecting atoms, for instance C–C means a single carbon-carbon bond that is cleaved or formed in the reaction. Bond order changes are represented as double arrows connecting bonds, for example C–C ↔ C=C means a single carbon-carbon bond turning into double carbon-carbon bond or *vice versa*. Stereochanges are represented as atoms that change their absolute configuration, for instance C(R/S) means a carbon atom that changes from R to S configuration.

A reaction centre is the collection of atoms and bonds that are changed during the reaction (Warr, 2014), also known as the local atomic environment around the atoms involved in bond changes. EC-BLAST calculates circular fingerprints around the non-hydrogen

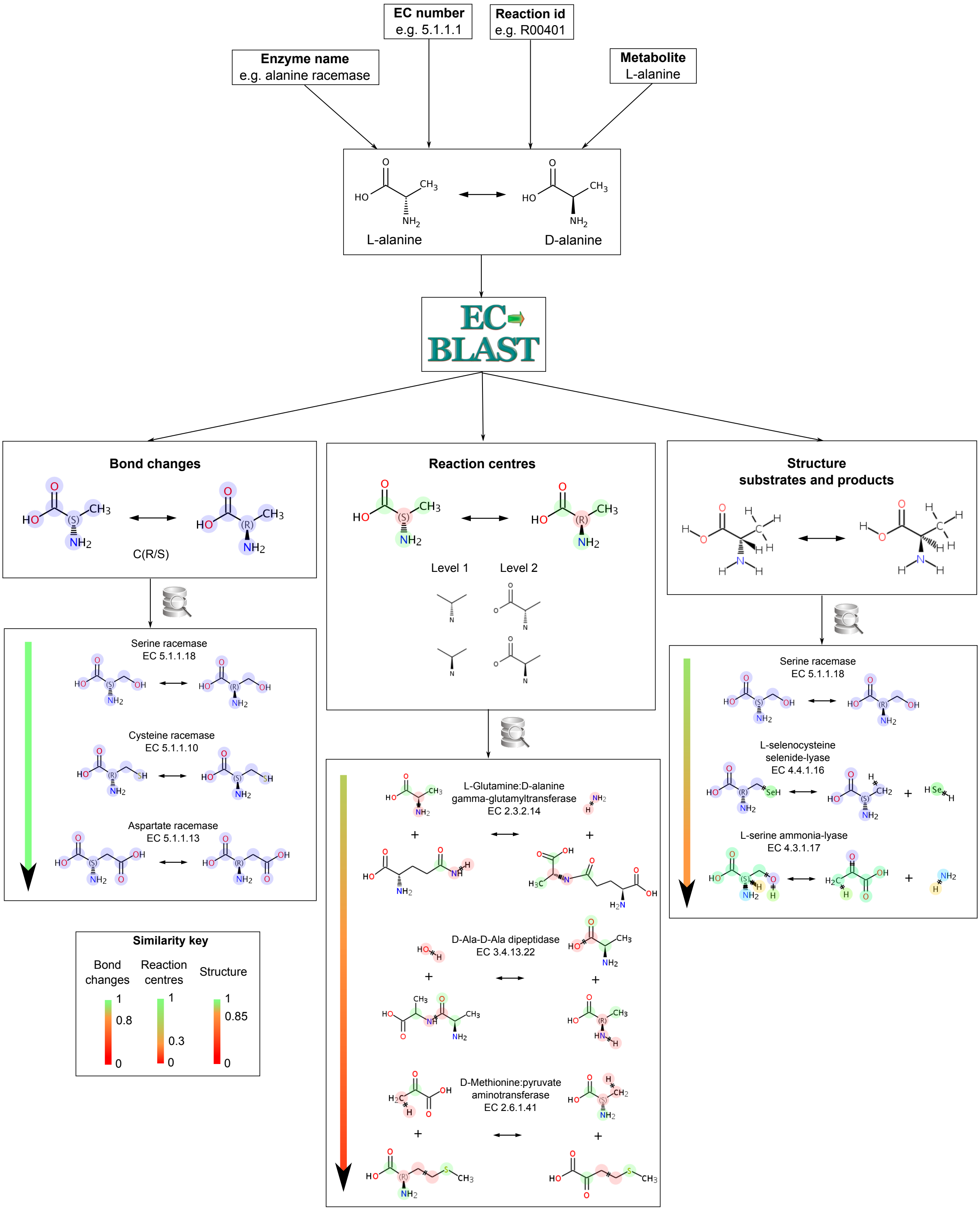


Figure 2.2: Flow diagram illustrating how EC-BLAST works. Alanine racemase (EC 5.1.1.1) is used as a query example.

atoms present in bond changes at three different levels (0, 1 and 2) (Rahman *et al.*, 2014). Reaction centres at level 0 simply comprise the atoms involved in bond changes, level 1 extends by one covalent bond from level 0. Finally, level 2 extends by two covalent bonds from level 0 (Figure 2.3).

In order to measure the specificity of bond changes or reaction centres in each EC subclass or sub-subclass, the Shannon-Wiener (SW) information statistic was used (Roberts, 2012). This is an index which is commonly used to measure species diversity in ecology. However, for this purpose it allows to quantify the uncertainty associated to predict the subclass from a bond change or reaction centre. It is calculated as $SW = -\sum_{i=1}^{6} p_i \ln p_i$ where p_i is the proportion of bond changes or reaction centres belonging to the ith subclass summed up across all 6 subclasses. Then it is scaled in the interval 0 to 1. In general, rare bond changes or reaction centres tend to be only present in one subclass and have a SW statistic equal to 1. On the other hand, common bond changes, such as C(R/S) are distributed across more than one subclass so they have a SW statistic considerably lower than 1 (see *Chapter 3. The Chemistry of Isomerases*).

EC-BLAST calculates similarity between reaction fingerprints using a Tanimoto score ranging between 0 (no similarity) and 1 (identical reactions) (Rahman *et al.*, 2014). The results obtained with different chemical attributes were stored in similarity matrices, which were then compared using two strategies. First, two-sample Kolmogorov-Smirnov (KS) tests were used to explore whether similarity distributions are significantly different from one another. Specifically, KS tests evaluate the null hypothesis that similarity distributions are drawn from the same continuous distribution (Crawley, 2007). Second, Mantel tests were adopted to obtain correlations between similarity matrices using the Pearson's product-moment correlation as the method. Significance was assessed by permuting rows and columns of one of the similarity matrices under comparison (Legendre & Legendre, 2012; Oksanen, 2011; R Core Team, 2012).

To find groups of similar reactions according to the chemical attributes mentioned above, hierarchical clustering was performed using the R environment for statistical computing (R Core Team, 2012). Three approaches were employed to select the best clustering algorithm and to choose the optimal number of clusters. First, using external evaluation, clustering algorithms were compared on their ability to obtain greatest purity in EC subclasses or sub-subclasses using the F-measure as implemented in in-house scripts provided by Dr. Syed Asad Rahman. This is the harmonic mean of precision (p) and recall (r) defined as $F - measure = \frac{2pr}{p+r}$. In the context of EC subclasses, precision is the fraction of isomerase reactions classified in the correct subclass ($p = \frac{TP}{TP+FP}$) and captures the subclass purity of clusters. Recall refers to the fraction of isomerase reactions

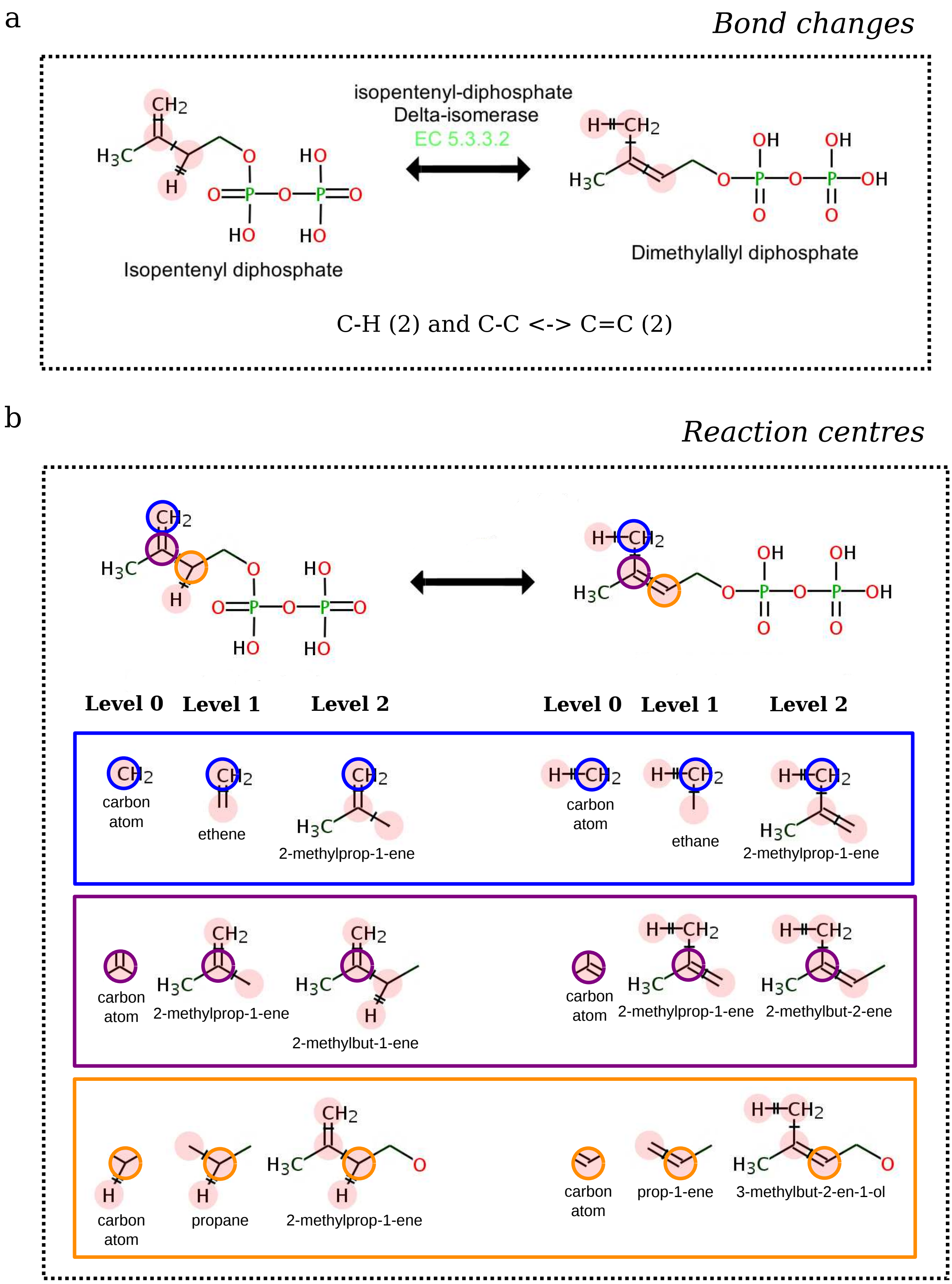


Figure 2.3: The reaction catalysed by isopentenyl-diphosphate Δ-isomerase (EC 5.3.3.2) has 6 carbon atoms involved in bond changes (3 in the substrate and 3 in the product). (a) Bond changes as calculated by EC-BLAST (b) Reaction centres obtained based on the atoms involved in bond changes.

belonging to the same subclass grouped in the same cluster ($r = \frac{TP}{TP+FN}$) and represents how spread subclasses are across different clusters. TP (true positives) are the number of pairs of reactions in the same EC subclass and the same cluster, FP (false positives) are the number of pairs in different EC subclass but located in the same cluster, finally FN (false negatives) are the number of pairs in the same EC subclass but different clusters. Second, as internal evaluation, hierarchical trees were pruned at the height that simultaneously minimises the number of clusters and the spread within each cluster as discussed in Kelley *et al.* (1996) and Fischer *et al.* (2010), and implemented in White & Gramacy (2012). Alternatively, Rand Index and Silhouette width (Mavridis *et al.*, 2013) were also investigated as external and internal evaluation metrics respectively, however they did not perform better than the F-measure and Kelley-Gardner-Sutcliffe penalty function. Third, the best correspondence between clusters and EC classification was explored using the mclust package (Fraley *et al.*, 2012), which helped to determine the extent to which subclasses and sub-subclasses are prevalent in clusters.

The hierarchical clustering solutions computed for each chemical attribute were contrasted using three methods. First, cross tabulations of EC numbers clustered according to two chemical attributes, for example bond changes *versus* reaction centres, are captured in contingency tables, also known as matching matrices. Second, topological distances between clustering trees were calculated, which are defined as twice the number of internal branches representing different bipartitions of the tips (Paradis, 2012; Paradis *et al.*, 2004; Penny & Hendy, 1985) are also useful. Third, tanglegrams involve drawing two clustering trees opposite to each other, which help to visualise comparisons between different clustering solutions (Scornavacca *et al.*, 2011). Clustering evaluation metrics such as F-measure and Rand Index were also explored here.

Other techniques to explore and visualise the similarity and clustering of biochemical reactions were investigated. For instance, dimensionality reduction techniques such as principal component analysis and multi-dimensional scaling (Everitt & Hothorn, 2011), correspondence analysis (Greenacre, 2007; Husson *et al.*, 2011) and self-organising maps (Chen & Gasteiger, 1997; Wehrens & Buydens, 2007). The overall interpretation of results using these alternative approaches was similar to the analysis using hierarchical clustering, however the visualisation was not as intuitive, therefore they were not shown in this dissertation.

2.2.3 Testing EC-BLAST with isomerase reactions

During the period 2012-2014, isomerase reactions were used as a test set in the development of EC-BLAST (Rahman *et al.*, 2014). At the start of the test in April 2012, the first challenge involved a critical assessment of the chemical validity of reaction data contained in an internal EC-BLAST database built upon the release 58.1 of KEGG (accessed on 1st June 2011). Although approaches to perform automatic checks of reaction data quality had already been presented (Ott & Vriend, 2006), a manual workflow aiming to report reaction data inconsistencies was designed for our specific purposes. The second challenge was to test EC-BLAST's ability to extract chemical attributes such as bond changes from biochemical reactions. Many isomerase reactions involve changes of stereochemistry caused by chemical processes such as chiral inversion and cis-trans isomerisation. To certain extent, this test helped Dr. Syed Asad Rahman to implement modules dealing with the automatic detection of stereochemistry changes in EC-BLAST.

The reference list of isomerase EC numbers used in this test was obtained from the 21st March 2012 release of ENZYME database (Bairoch, 2000). This resource actively follows the recommendations of the NC-IUBMB concerning standard nomenclature and classification of enzymes. The test set comprised 202 four-digit isomerase EC numbers linked to 289 unique biochemical reactions, which were obtained from release 62.0 of KEGG (accessed using KEGG API). The manual curation of isomerase reactions comprised three major tasks: detection of KEGG errors, outdated data and software errors. KEGG errors and outdated data were easy to discover because they originate from data quality problems or mismatch between the NC-IUBMB, KEGG and internal EC-BLAST databases. However, discovery of software errors required a more thorough process of manual inspection and validation of isomerase reactions. The types of errors encountered in the initial stages of the curation process are illustrated in Figure 2.4 and Table 2.1.

The test covered eight rounds of manual checks and seven rounds of code optimisations, which ultimately led to a significant decrease in the number of software errors. The first manual check resulted in a total of 312 errors. A third of them pointed to inconsistencies with the atom-atom mapping performed by EC-BLAST as indicated by their bond changes. Another third consisted of outdated data issues and the remaining third comprised KEGG errors and other software errors. This test took almost a year to complete and by the eighth manual check, the total number of errors considerably reduced to 200 (Figure 2.6). Although only three stereochemistry errors were resolved during this set of optimisations (pink, black and gray bars), this situation improved as the R/S perception library was optimised and modules to detect E/Z stereochemistry were implemented in later stages of software development (Rahman *et al.*, 2014).

After the eighth round of manual checks, most software errors were resolved (Table 2.2). KEGG errors were fixed in parallel to this test by manually curating reactions from release 58.1 of KEGG, storing them in a database of reactions within EC-BLAST and ac-

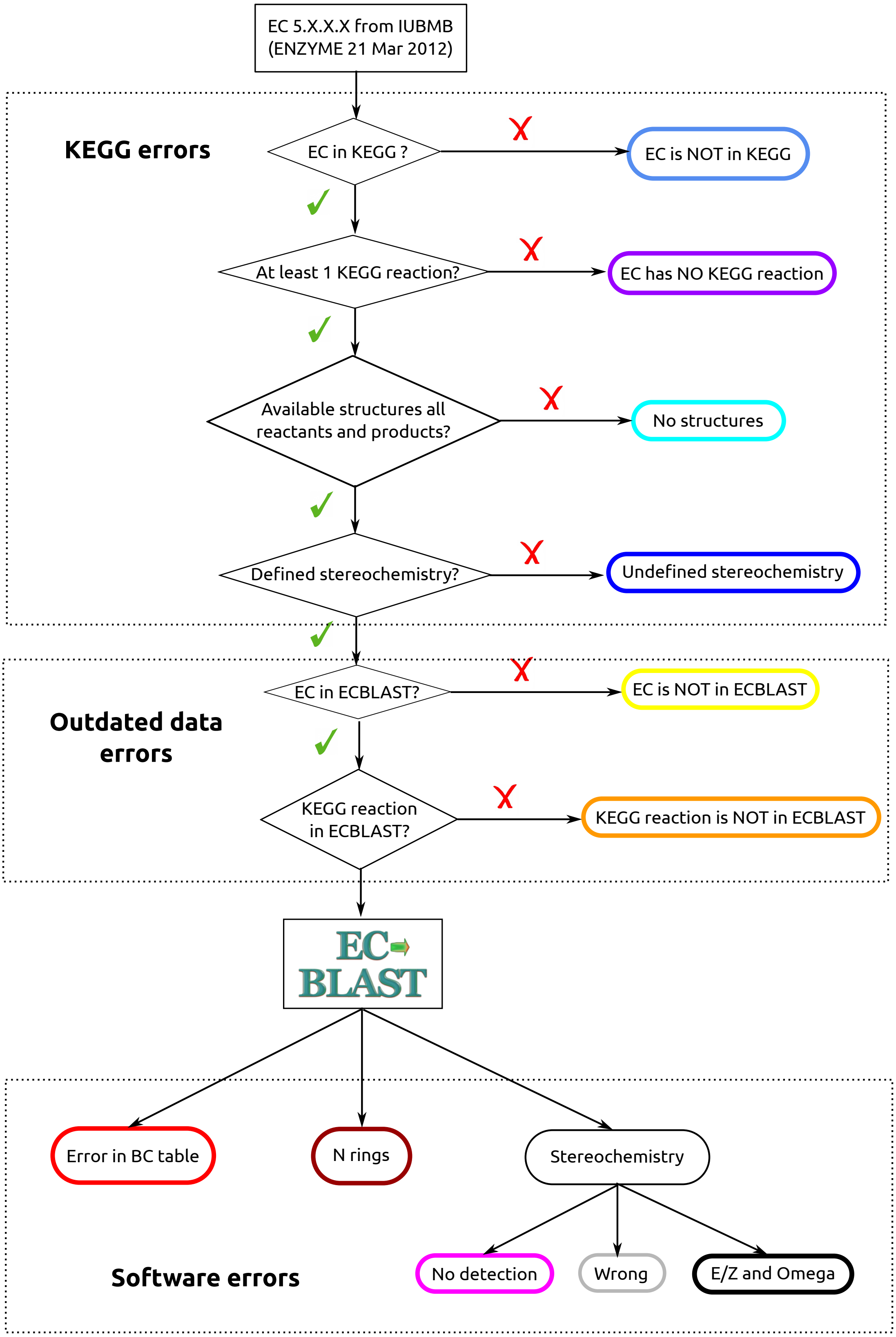


Figure 2.4: Flow diagram representing the manual process of error-checking in EC-BLAST using isomerase reactions.

Table 2.1: Summary of errors detected at the beginning of the manual curation of isomerase reactions using EC-BLAST.

Type of error	Error name	Description
KEGG errors	EC number is not in KEGG	Mismatch between the NC-IUBMB and KEGG classification of enzymes.
	EC number has no KEGG reaction	KEGG does not link the EC number to reaction data. Example: protein disulfide-isomerase (EC 5.3.4.1) catalyses the rearrangement of S–S bonds in proteins.
	No structures	KEGG does not provide structural information for all substrates and products of a reaction. Example: cellobiose epimerase (EC 5.1.3.11).
	Undefined stereochemistry	The stereochemistry of some chiral atoms in KEGG is unknown (depicted as flat bond) or any (wiggly bond). Defined stereochemistry is represented as wedge or hatch bonds, also known as up or down, respectively (Brecher, 2008) (http://www.chemaxon.com) (Figure 2.5a).
Outdated data errors	EC is not in EC-BLAST database	Example: 2,3-diacetamido-2,3-dideoxy-α-D-glucuronate 2-epimerase (EC 5.1.3.23).
	KEGG reaction is not in EC-BLAST database	Example: R09600 involves the conversion of UDP-α-D-GlcNAc3NAcA into UDP-α-D-ManNAc3NAcA.
Software errors	Error in BC table	Bond changes extracted from EC-BLAST do not match bond changes derived manually due to problems in the atom-atom mapping. Example: styrene-oxide isomerase (EC 5.3.99.7) (Figure 2.5b).
	N rings	EC-BLAST incorrectly cleaves all N-H bonds. Double bonds in five and six member heterocycles are represented incorrectly. Example: L-dopachrome isomerase (EC 5.3.3.12).
	Stereochemistry	(a) Some cases of R/S stereochemistry are not detected. Example: amino acid racemase (EC 5.1.1.10) (Figure 2.5c). (b) Wrong R/S stereochemistry is detected. Example: proline racemase (EC 5.1.1.4) (Figure 2.5d). (c) E/Z (also known as cis/trans) and omega stereochemistry is not detected. At this early stage of software development, EC-BLAST was not able to account for this type of stereochemistry. Example: maleate isomerase (EC 5.2.1.1) (Figure 2.5e).

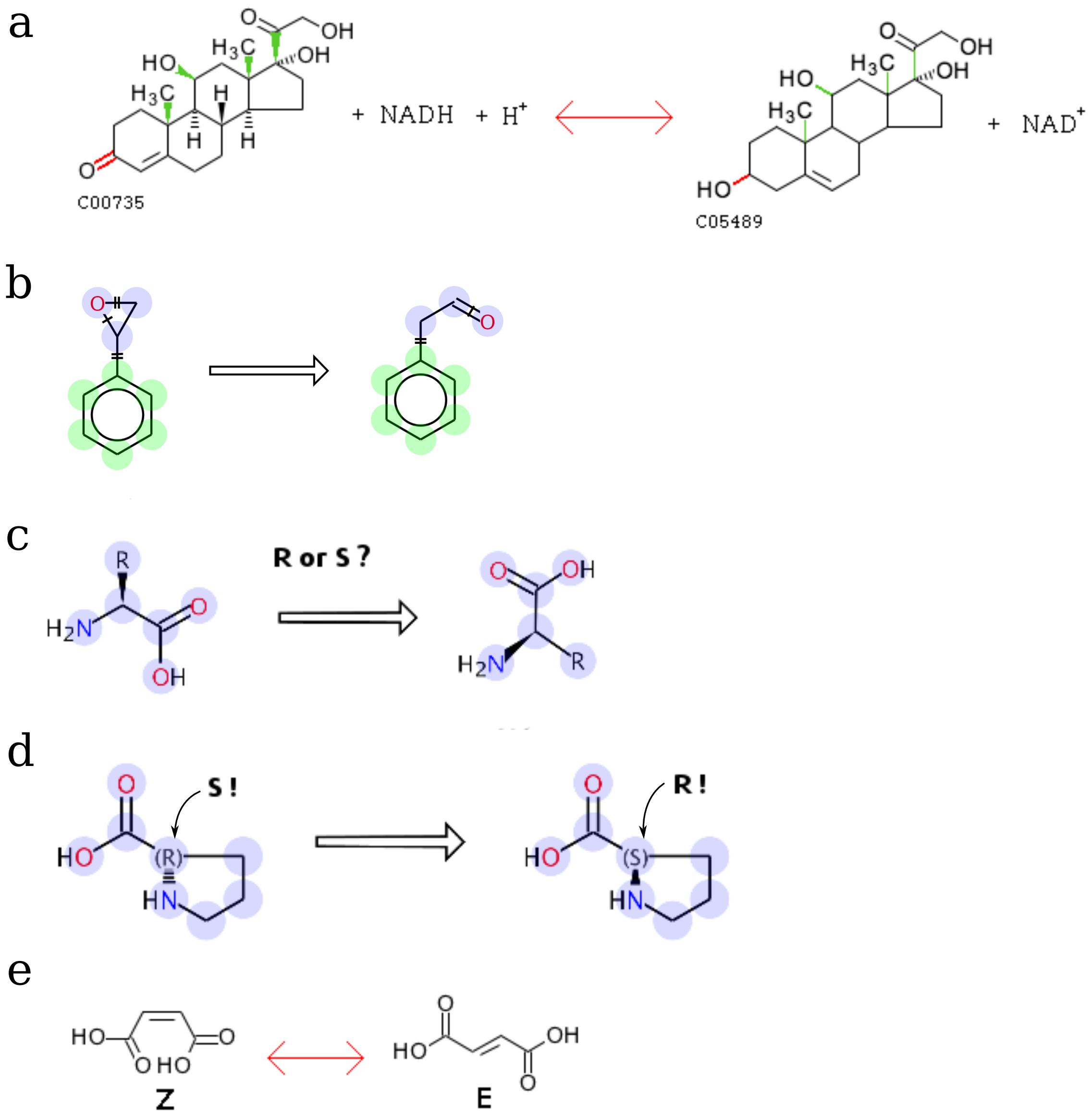


Figure 2.5: Examples of errors discovered while testing EC-BLAST. (a) Undefined stereochemistry in 3-oxosteroid Δ5-Δ4-isomerase (EC 5.3.3.1). In red, an example of unknown stereochemistry in a reacting atom (see product). Although there are four cases of defined stereochemistry in non-reacting atoms shown as green wedge bonds in the substrate, stereochemistry is depicted as unknown in the product. (b) Although EC-BLAST reports the cleavage of a C–C bond in styrene-oxide isomerase (EC 5.3.99.7), according to mechanistic studies the substrate undergoes an epoxide ring opening via the cleavage of a C–O only (Hartmans *et al.*, 1989) and C–C cleavage does not take place. (c) R/S stereochemistry is not detected in amino acid racemase (EC 5.1.1.10). (d) Wrong R/S stereochemistry is detected in proline racemase (EC 5.1.1.4). (e) E/Z stereochemistry is not detected in maleate isomerase (EC 5.2.1.1).

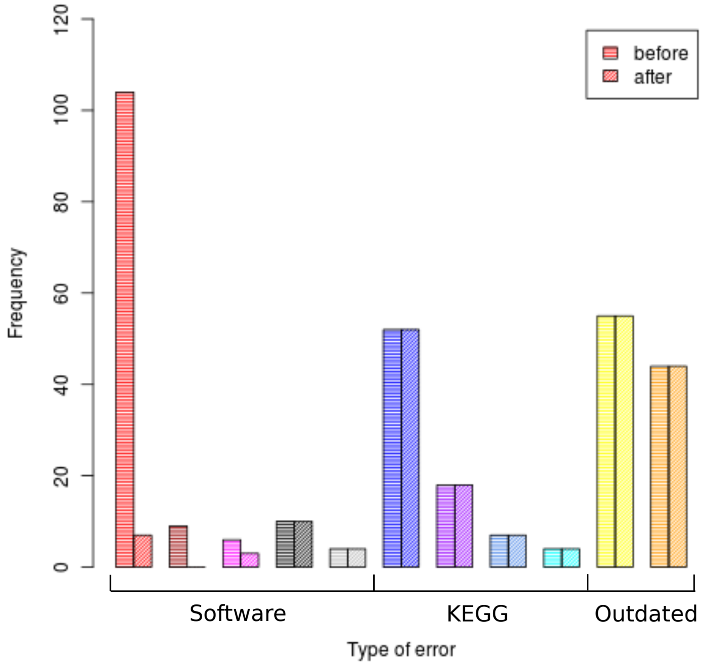


Figure 2.6: Distribution of errors found in the first (before) and eighth (after) manual checks of the EC-BLAST test. Colouring of bars indicates the error type as in Figure 2.4.

tively reporting KEGG about data inconsistencies found during this process. As a result, KEGG corrected many of the inaccurate entries in following releases of their database. Resolving outdated data issues first involved the exploration of other reaction databases such as BKM-React (Lang *et al.*, 2011) and Rhea (Alcántara *et al.*, 2012). Second, in April 2014 release 58.1 of KEGG was upgraded to release 70.0+ (Figure 2.1) so to represent a more recent version of the space of biochemical reactions in this dissertation.

Only the initial stages of the process of testing EC-BLAST have been described here, however manual validation of the chemical content of EC-BLAST and KEGG still continues today covering a broader spectrum of biochemical reactions from different EC classes. Also, when newer versions of the algorithm are developed, manual analysis against a challenging test set of reactions is undertaken. This evaluation set has continuously been updated during EC-BLAST development. There are several reactions though that remain challenging due to theoretical difficulties while performing the atom-atom mapping and extracting chemical attributes (Figure 2.7). These cases represent a motivation for further improvement of existing algorithms.

Table 2.2: Distribution of software errors and consequent actions of code optimisation during the EC-BLAST test.

Round	Software error			Total	Code optimisation strategy
	BC	N rings	Stereo		
1	104	9	20	133	-
2	66	9	20	95	Addition of a missing piece of code.
3	31	9	20	60	Optimisation of the Maximum Common Subgraph algorithm.
4	21	0	17	38	Resolving conflicts between ChemAxon and CDK libraries.
5	16	0	17	33	Resolving wrong C–O cleavage in R06989-like reactions.
6	17	0	17	34	Fixing hybridisation issues in R01819-like reactions.
7	7	0	17	24	Exception handling to resolve wrong C–O cleavage in R01577-like reactions.
8	3	0	17	20	Exception handling rest of cases.

2.3 Evolution

2.3.1 Resources and dataset

Protein similarity networks have been used very successfully to map biological information to large sets of proteins (Brown & Babbitt, 2012; Uberto & Moomaw, 2013). However, it is also necessary to include associated changes of catalytic function during evolution preferably in an automated fashion. FunTree is a resource developed to accomplish that goal (Furnham *et al.*, 2012*b*) and it is maintained in collaboration with the CATH classification of protein structures (Sillitoe *et al.*, 2013). By combining sequence, structure, phylogenetic, chemical and mechanistic information, it allows one to answer fundamental questions about the link between enzyme activities and their evolutionary history in the context of superfamilies (Figure 2.8a).

First, FunTree clusters the structures of the CATH domains involved in enzyme function as defined by MACiE and CSA, which are both resources that manually annotate enzyme mechanisms and catalytic sites, respectively (Furnham *et al.*, 2014; Holliday *et al.*, 2012).

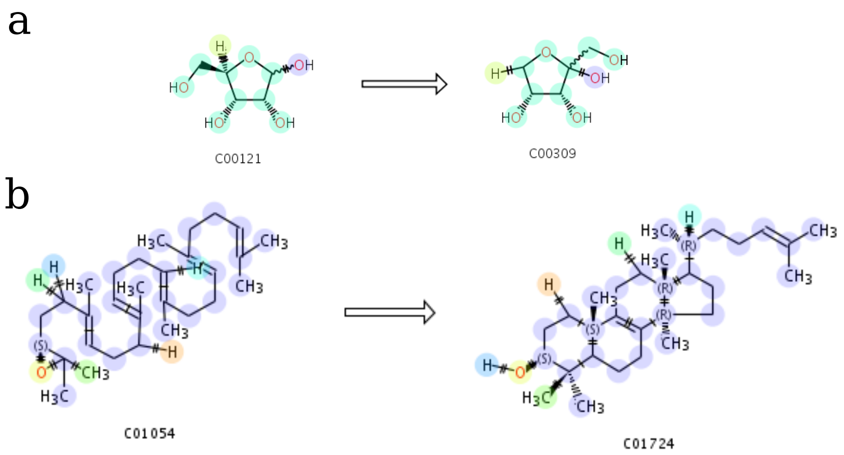


Figure 2.7: Challenging reactions in EC-BLAST. (a) The reaction catalysed by ribose isomerase (EC 5.3.1.20) exhibits a ring opening and closure, rather than a C–O cleavage. (b) EC-BLAST performs a non-optimal atom-atom mapping of the reaction catalysed by lanosterol synthase (EC 5.4.99.7). In general, reactions catalysed by oxidosqualene cyclases (EC 5.4.99.-) proved difficult for accurate atom-atom mapping (see *Chapter 3. The Chemistry of Isomerases*).

Structural clusters are subsequently populated with sequence relatives from Gene3D (Lees *et al.*, 2010) using BLASTp (Altschul *et al.*, 1997) and structurally-informed multiple sequence alignments are then generated with FUGUE (Shi *et al.*, 2001). Second, alignments are the starting point to create species-guided phylogenetic trees generated with the maximum likelihood (ML) method and the WAG substitution model using PHYML as implemented in TreeBest (Heng, 2006; Ruan *et al.*, 2008). For tree visualisation purposes, sequences were filtered by taxonomic lineage while maintaining functional diversity, sequences with known structure and multidomain architecture diversity. Last, functional annotations are retrieved from the reviewed section of UniprotKB, PDBSum and KEGG (de Beer *et al.*, 2014; Kanehisa *et al.*, 2012; The Uniprot Consortium, 2013).

2.3.2 Methods and tools

FunTree uses phylogenetic methods to infer ancestral enzymes in superfamilies and estimate their most likely functions (Paradis, 2012). By systematically traversing the phylogenetic tree from ancestor to modern enzymes in a superfamily, explicit changes of function are identified between groups of enzymes. Ultimately, each functional change is represented by two sets of enzymes catalysing two distinct functions (Figure 2.8b). Pairs of functions and their corresponding sets of enzymes are comparatively analysed using functional and all-against-all sequence similarity. Functional similarity was obtained using EC-BLAST (see *2.2 Biochemical reactions*) and results were interpreted in the light of mechanistic data extracted from MACiE and extensive literature searches, which com-

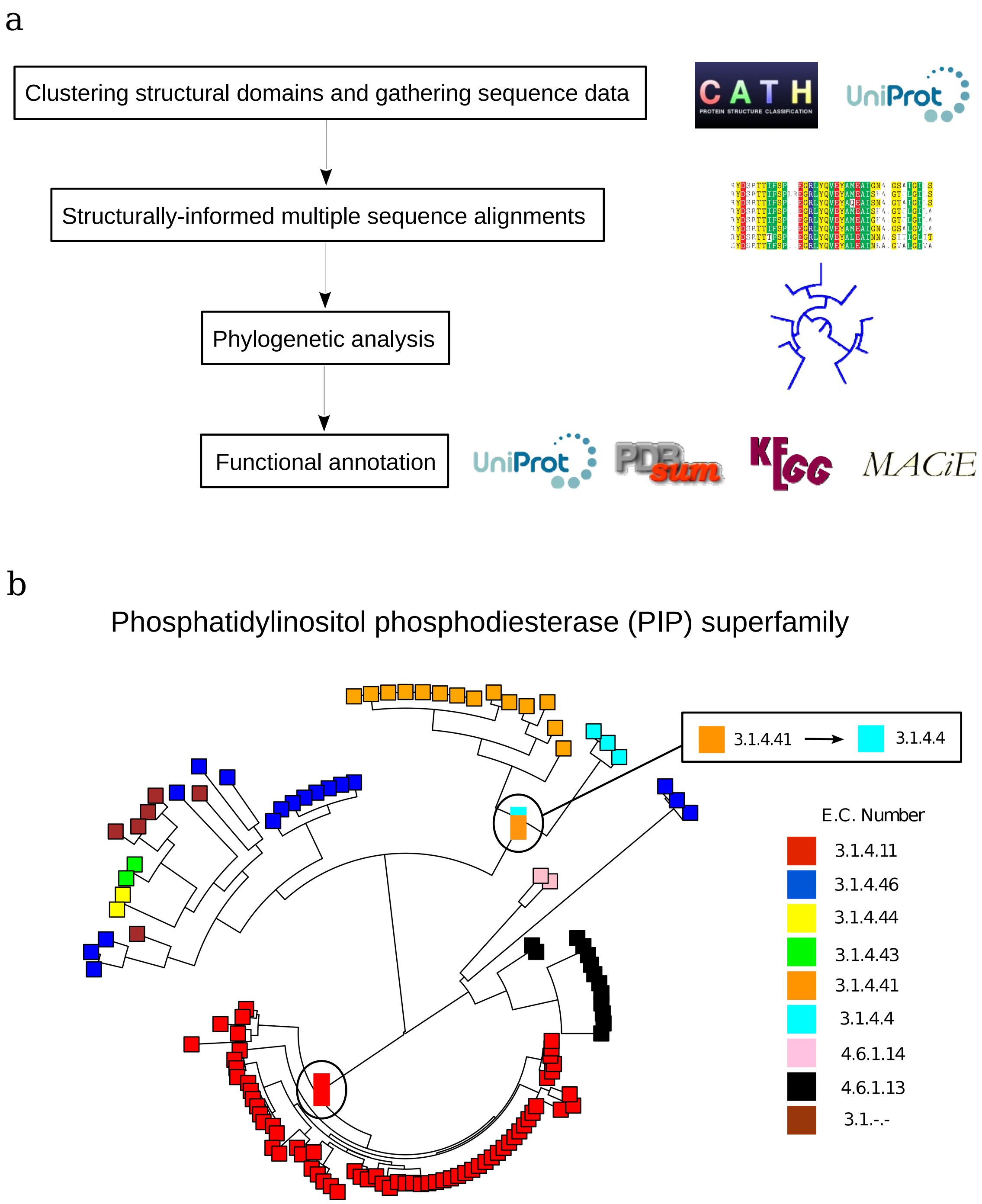


Figure 2.8: (a) Illustration of the FunTree pipeline (b) Example of the estimation of a functional change during the evolution of PIP superfamily. Phospholipase D activity (EC 3.1.4.4, cyan) was inferred to have evolved from sphingomyelin phosphodiesterase D activity (EC 3.1.4.41, orange). Image courtesy of Dr. Nicholas Furnham.

prehensively informed the analyses. Similarity between homologous enzyme sequences was calculated as sequence identity derived from the multiple sequence alignments used to generate the phylogenetic trees.

Chapter 3

The Chemistry of Isomerases

This chapter describes an analysis of isomerase reactions on the basis of three chemical attributes: bond changes, reaction centres and structures of reactants (substrates and products). Next, the quality of the EC classification of isomerases is discussed as well as ways to improve it. Last, three isomerase EC subclasses are further explored: racemases and epimerases (EC 5.1), intramolecular oxidoreductases (EC 5.3) and intramolecular transferases (EC 5.4). The discussion section addresses the ability of the EC classification to represent isomerase function as measured by the overall reaction.

3.1 Analysis of bond changes

3.1.1 Distribution

The six isomerase subclasses have different numbers of representative reactions and bond changes associated with them (Figure 3.1). EC 5.4 (intramolecular transferases) is the most abundant subclass with 71 reactions, closely followed by EC 5.1 (racemases and epimerases) and EC 5.3 (intramolecular oxidoreductases) with 58 and 57 reactions, respectively. In terms of bond changes EC 5.4, followed by EC 5.3 and EC 5.5 (intramolecular lyases) are the subclasses with largest number of bond changes. The overall distribution of the number of reactions with a given number of bond changes indicates diversity across different subclasses (Figure 3.2a). The average number of bond changes per reaction is 7.3. There are several extreme cases with more than 20 bond changes, which correspond to complex reactions belonging to sub-subclass EC 5.4.99. EC 5.1, 5.2 and 5.99 catalyse only a small number of bond changes (1 to 8). However EC 5.3, 5.4 and 5.5 correspond to reactions with more bond changes on average (Figure 3.2b).

A total of 30 different types of bond changes were found in our dataset of isomerase reactions. The most common bond changes are R/S stereochange (C(R/S)), cleavage and formation of carbon-hydrogen bond (C–H) and cleavage and formation of oxygen-

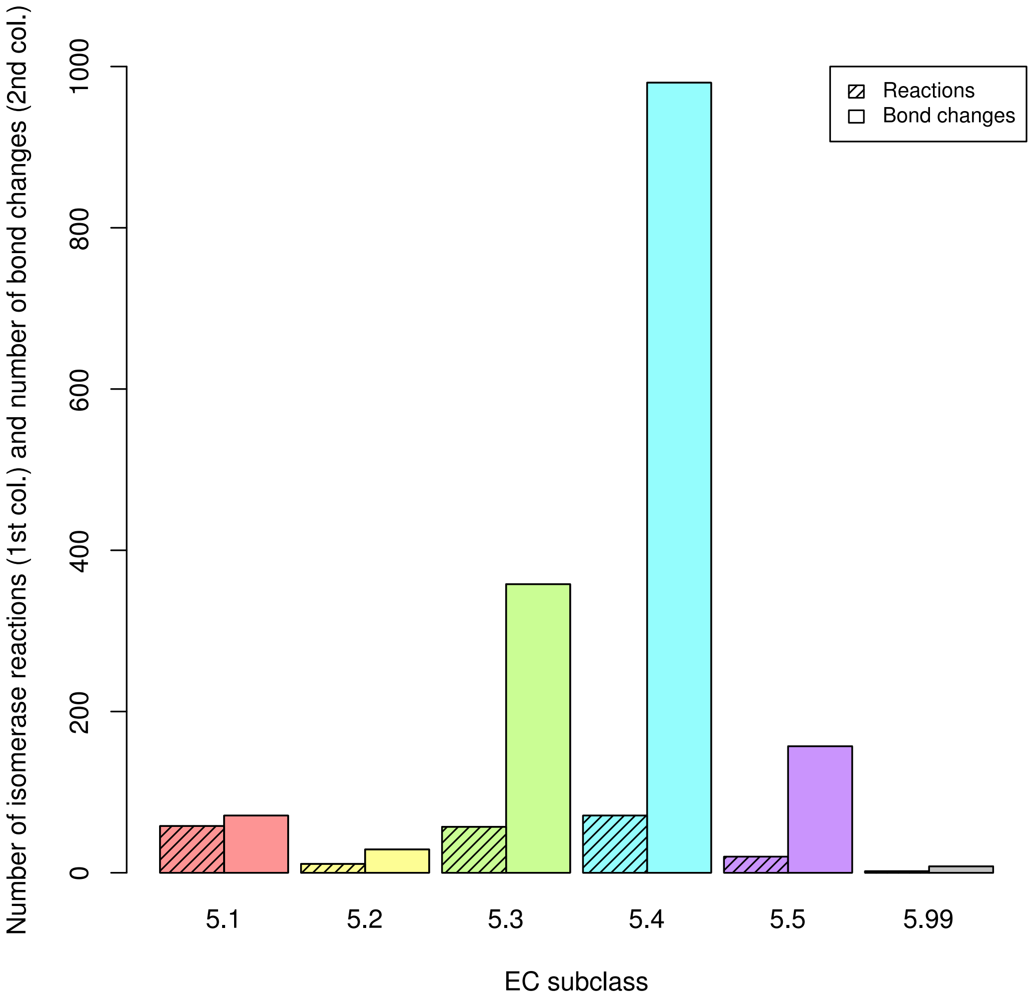


Figure 3.1: Distribution of 219 representative isomerase reactions (1^{st} column) and the 1603 ocurrences of bond changes (2^{nd} column) associated with them across EC subclasses. Bars are coloured according to subclass. This colouring scheme of the different subclasses is adopted throughout the thesis.

hydrogen bond (O–H) (Figure 3.3). Some types of bond changes occur more often in certain subclasses. Stereochange C(R/S) occurs in all the subclasses except EC 5.2 and 5.99 (Figure 3.3). It is also the only type of bond change observed in subclass EC 5.1 with the exception of one reaction: the conversion of L-phenylalanine into D-phenylalanine where L-phenylalanine racemase (EC 5.1.1.11) catalyses a single stereochange C(R/S) and unusually for a racemase, it also involves the cleavage and formation of two O–P bonds and two O–H bonds from ATP and water. This is the only isomerase reaction that is not unimolecular. Bond change C–H occurs in four of the six subclasses with the exception of EC 5.1 and EC 5. 99, whereas bond change H–O occurs in all the subclasses except EC 5.2. From the point of view of the subclass, EC 5.1 and 5.2 are represented by a small subset of three bond changes: EC 5.1 is mainly C(R/S) whereas O–P and H–O are rare, EC 5.2 is C(E/Z), C–H and C–C ↔ C=C are rare. On the other hand, EC 5.3,

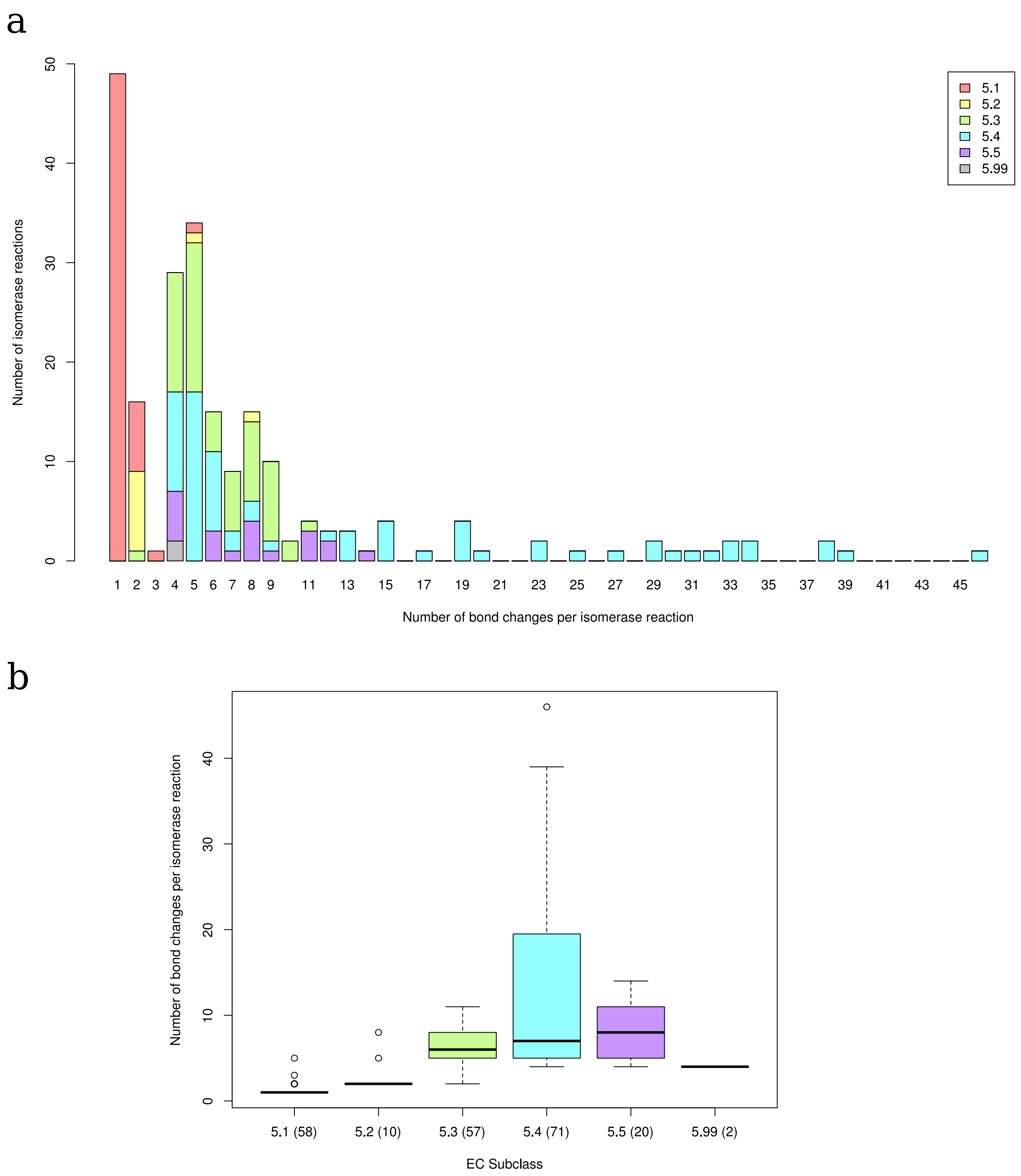


Figure 3.2: Distribution of the number of bond changes per isomerase reaction by EC subclass. (a) Bar plot showing the number of isomerase reactions against number of bond changes. (b) Box and whisker plot representing the distribution of the number of bond changes per isomerase reaction by EC subclass. The number of reactions belonging to each subclass is shown in brackets. Coloured boxes contain the interquartile range of each subclass. Whiskers represent 1.5 times the size of the coloured boxes. Single points beyond the whiskers represent outliers (Crawley, 2007).

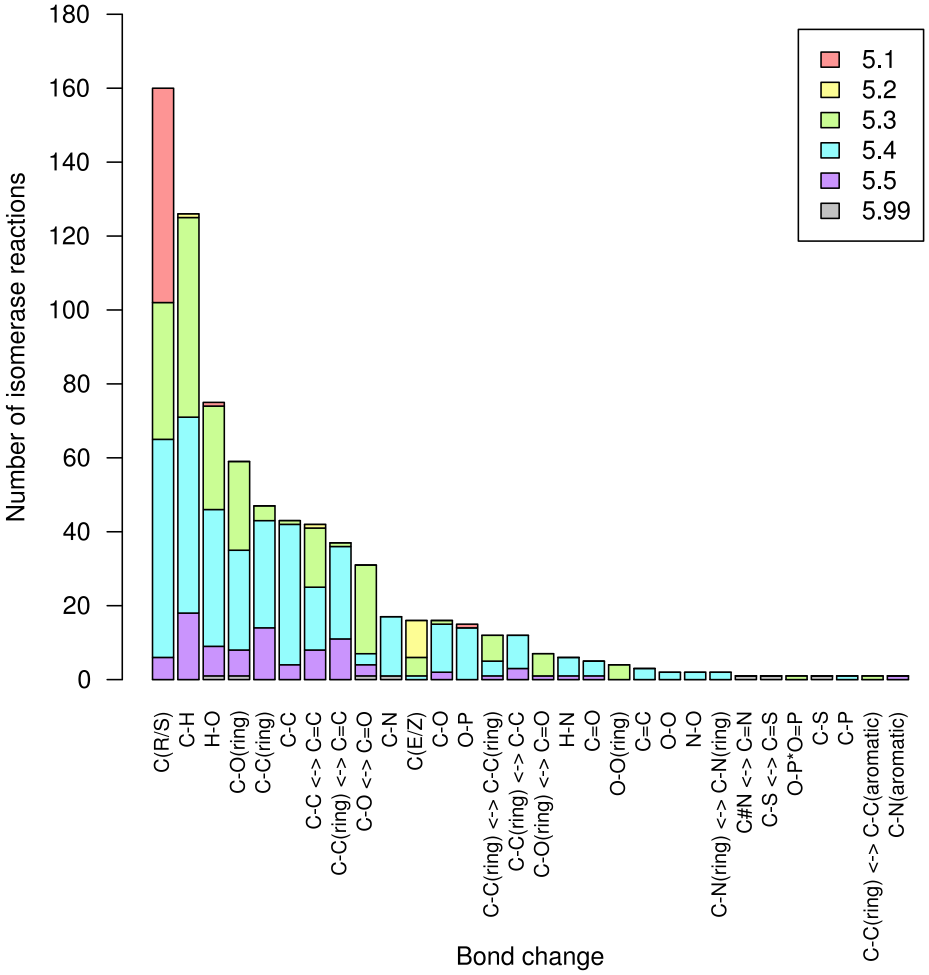


Figure 3.3: Distribution of the number of isomerase reactions by bond change and EC subclass.

5.4, 5.5 and 5.99 exhibit a much broader variety of bond changes. Overall, subclasses EC 5.1, 5.2 and 5.99 have less diversity of bond changes than the rest, which seem more complex.

How much does each bond change contribute to a subclass? Do some bond changes discriminate subclasses more specifically? The relative frequency of occurrence of each bond change in subclasses was used to calculate the degree whereby bond changes are specific to certain subclasses using the Shannon-Wiener (SW) information statistic (Roberts, 2012) (see *Chapter 2. Data Resources and Methods*). In general, rare bond changes, e.g. C≡N ↔ C=N (Figure 3.3) are catalysed by enzymes from only one subclass and have a SW statistic equal to 1 (Table 3.1). On the other hand, abundant bond changes, such as C(R/S) are distributed across more than one subclass so they have a SW statistic lower than 1. Twelve bond changes (40% of all) are perfect dicriminators of subclasses. For instance, C≡N ↔ C=N is specific for the EC 5.99. However the relative frequency of specific bond changes is generally low - they are rare. O–O(ring) is specific for EC 5.3 and it is a chemical attribute of four different isomerase reactions catalysed by enzymes present in the arachidonic acid metabolism: prostaglandin synthases D, E and I (EC 5.3.99.2, 5.3.99.3 and 5.3.99.4) and thromboxane-A synthase (EC 5.3.99.5). These enzymes catalyse the opening of epidioxy bridges in prostaglandins.

Table 3.1: How specific are bond changes to isomerase subclasses?

Bond change	**SW**	**EC subclasses**	**Number of EC subclasses**	**Frequency**
O–P ↔ O=P	1.00	5.3	1	2
C≡N ↔ C=N	1.00	5.99	1	1
N–O	1.00	5.4	1	2
C=C	1.00	5.4	1	4
C–N(aromatic)	1.00	5.5	1	2
O–O(ring)	1.00	5.3	1	4
O–O	1.00	5.4	1	2
C–N(ring) ↔ C–N(ring)	1.00	5.4	1	3
C–P	1.00	5.4	1	1
C–S	1.00	5.99	1	1
C–C(ring) ↔ C–C(aromatic)	1.00	5.3	1	2
C–S ↔ C=S	1.00	5.99	1	1
H–N	0.69	5.4, 5.5	2	8
C–N	0.64	5.4, 5.99	2	31
C=O	0.62	5.4, 5.5	2	6
C–O(ring) ↔ C=O	0.61	5.3, 5.5	2	8
O–P	0.61	5.1, 5.4	2	29
C–C(ring) ↔ C=C	0.61	5.4, 5.5	2	117
C–C(ring) ↔ C–C	0.45	5.3, 5.4, 5.5	3	12
C–C(ring)	0.44	5.3, 5.4, 5.5	3	152
C–C(ring) ↔ C–C(ring)	0.44	5.3, 5.4, 5.5	3	31
C(E/Z)	0.40	5.2, 5.3, 5.4	3	34
C–C	0.40	5.3, 5.4, 5.5	3	158
C–O	0.39	5.3, 5.4, 5.5	3	26
C–H	0.28	5.2, 5.3, 5.4, 5.5	4	413
C–C ↔ C=C	0.25	5.2, 5.3, 5.4, 5.5	4	57
C(R/S)	0.24	5.1, 5.3, 5.4, 5.5	4	251
C–O ↔ C=O	0.23	5.3, 5.4, 5.5, 5.99	4	45
C–O(ring)	0.23	5.3, 5.4, 5.5, 5.99	4	81
H–O	0.11	5.1, 5.3, 5.4, 5.5, 5.99	5	119

3.1.2 Similarity and clustering

Given this distribution of bond changes the optimal subdivision of isomerase reactions into subclasses was explored using hierarchical clustering. A two step procedure was performed, first pairwise similarities between isomerase reactions were computed based on the frequency of bond changes using EC-BLAST (Figure 3.4a). The comparisons were stored in a similarity matrix. Secondly, a clustering algorithm was applied to the similarity matrix in order to group similar reactions (Figure 3.4b). After testing various clustering procedures by scanning how different cuts along the trees lead to greatest purity in EC subclasses, the Ward algorithm was selected to produce an optimal number of 6 chemically sensible clusters (see *Chapter 2. Data Resources and Methods*). Clustering results are illustrated in Figure 3.5 and summarised in Table 3.2 according to three main descriptors:

- The **main subclass** of a cluster was defined as the subclass with the highest number of reactions. For instance, EC 5.1 was considered the main EC subclass of cluster F.

- The **main bond changes** of a cluster were considered as those present in at least 80% of the reactions belonging to that cluster. For example, H–O and O–P were identified as main bond changes in cluster A.

- We also defined an **outlier** in a cluster as a reaction belonging to a different EC subclass than adjacent reactions in the clustering. Two main types of outliers were identified: first, outliers were defined as *distinct* if at least one of its bond changes is unique among the rest of bond changes from surrounding reactions. Conversely, an outlier was defined as *misannotated* if its bond changes are shared by at least two other neighboring reactions of a different subclass. As an example, in cluster D we identified EC 5.5.1.3 as a *distinct* reaction because among other bond changes, H–O is present and C–H is absent in reactions EC 5.3.2.5, 5.3.1.13 and 5.3.1.27. On the other hand, EC 5.2.1.5 in cluster C was considered as a *misannotated* reaction because it shares the presence of all its bond changes C(E/Z), C–C ↔ C=C and C–H with reactions EC 5.3.3.8, 5.3.3.13 and 5.3.3.14.

Table 3.2 explores the six clusters depicted in Figure 3.5. Only one cluster is pure in subclasses: F in EC 5.1. Three clusters are more than 75% pure: A in EC 5.4, B in EC 5.2 and D in EC 5.3. Finally, two clusters are mixed: C (mainly EC 5.3 and 5.4) and E (EC 5.4 and 5.5). From the subclass point of view, almost all racemase and epimerase reactions (EC 5.1) group in cluster F. The only exception is EC 5.1.1.11, which belongs to cluster A and it is defined as *misannotated* because of sharing identical bond changes to five intramolecular transferase reactions (EC 5.4). Cis-trans isomerase reactions (EC 5.2)

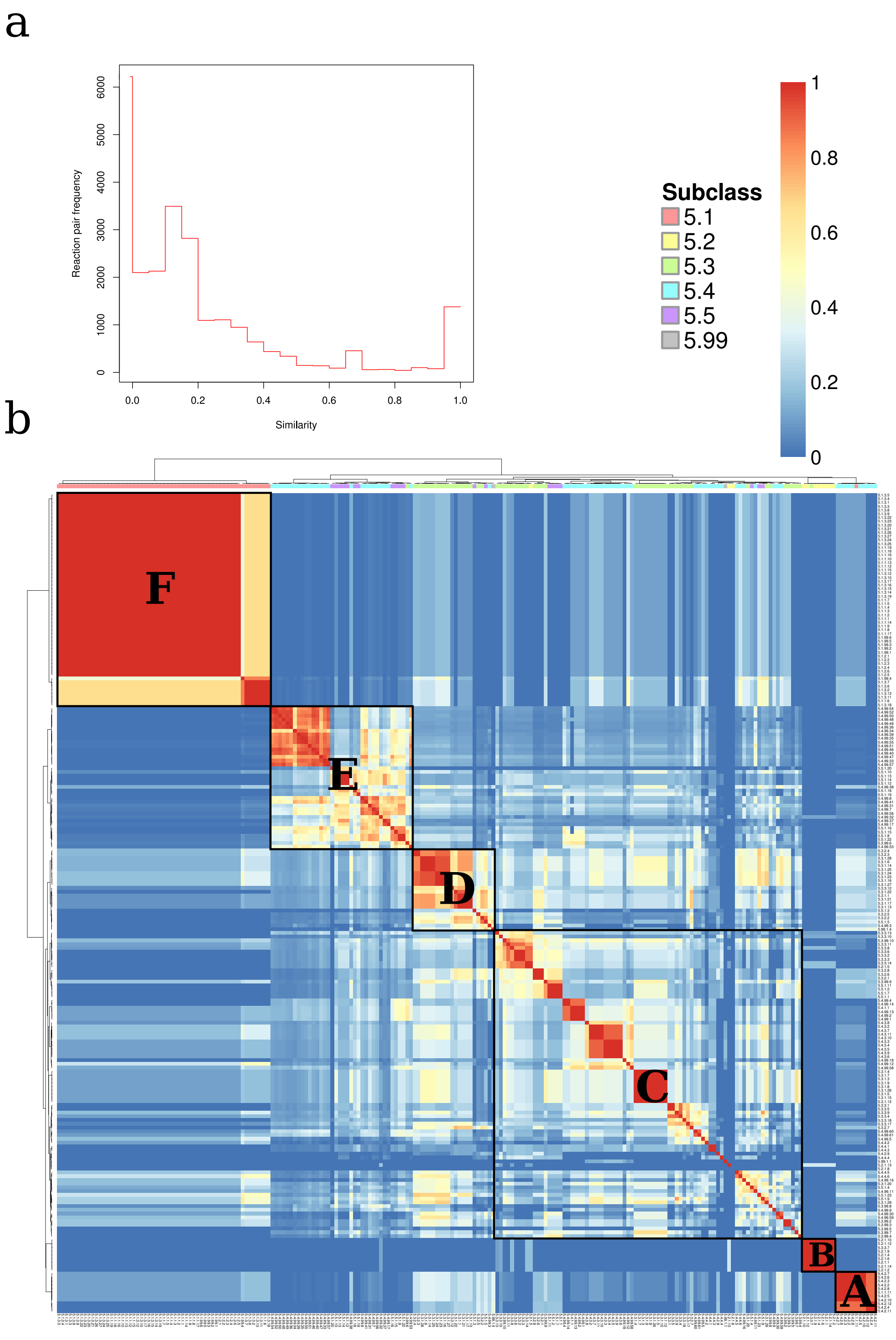


Figure 3.4: (a) Distribution of pairwise similarities of isomerase reactions by bond changes. (b) Bond change similarity matrix. Blue-to-red scale represents increasing similarity, identical reactions have a similarity of 1 (red). Reactions are annotated in colours according to their EC subclass (top row). Six clusters were identified (A to F) (see Figure 3.5).

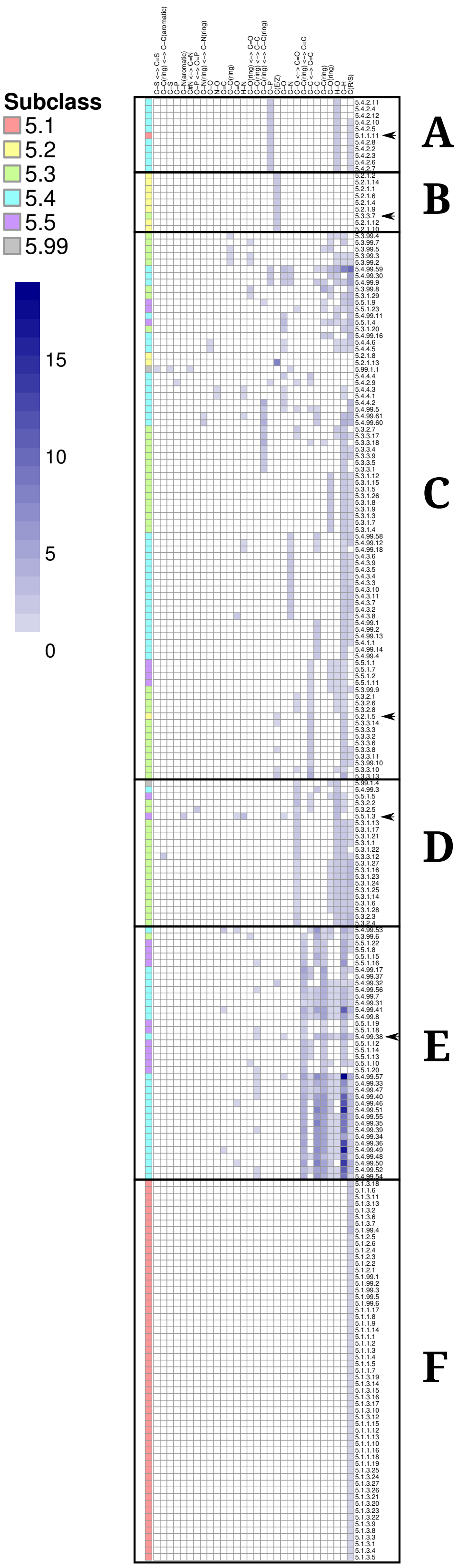


Figure 3.5: Hierarchical clustering of isomerase reactions (rows) by bond changes (columns). The blue scale represents frequency of bond changes in reactions. Reactions are annotated in colours according to their EC subclass (left-hand column). Six clusters were identified (A to F) (see Figure 3.4b). Bond changes were ordered left-to-right according to increasing frequency (Figure 3.3). Outliers are annotated with an arrow (see main text).

Table 3.2: Analysis of the clusters obtained in Figure 3.5. The columns Cluster, Number of reactions and EC subclasses are self-explanatory. The main subclass is underlined and the number of reactions belonging to each subclass is shown in brackets. The columns Main bond changes and Outliers and comments are defined in the main text.

Cluster	**Number of reactions**	**EC subclasses**	**Main bond changes**	**Outliers and comments**
A	11	EC 5.1 (1) and <u>5.4</u> (10)	H–O and O–P	EC 5.1.1.11 is a *misannotated* reaction. Pure cluster
B	9	EC <u>5.2</u> (8) and 5.3 (1)	C(E/Z)	EC 5.3.3.7 is a *misannotated* reaction. Pure cluster
C	82	EC 5.2 (3), <u>5.3</u> (37), 5.4 (34), 5.5 (7) and 5.99 (1)	C–H	EC 5.2.1.5 is a *misannotated* reaction. Mixed cluster
D	22	EC <u>5.3</u> (18), 5.4 (1), 5.5 (2) and 5.99 (1)	C–H, O–H and C–O ↔ C=O	EC 5.5.1.3 is a *distinct* reaction. Almost pure cluster
E	38	EC 5.3 (1), <u>5.4</u> (26) and 5.5 (11)	C–H, C–C(ring), C–C and C–C(ring) ↔ C=C	EC 5.4.99.38 is a *distinct* reaction. Mixed cluster
F	57	EC <u>5.1</u> (57)	C(R/S)	No outliers. Pure cluster

are mostly grouped in cluster B. Exceptions are EC 5.2.1.5 (*misannotated*), 5.2.1.8 and 5.2.1.13 in cluster C. The rest of subclasses are mixed across several clusters, intramolecular oxidoreductases (EC 5.3) are present in four clusters: B, C, D and E; intramolecular transferases (EC 5.4) are in four clusters: A, C, D and E; intramolecular lyases (EC 5.5) are in three clusters: C, D and E; and other isomerases (EC 5.99) are in two clusters: C and D.

Out of the total of 219 representative isomerase reactions considered in this study, our bond change analysis questioned the EC subclass classification of 3 isomerase reactions:

- Located in cluster A, EC 5.1.1.11 corresponds to phenylalanine racemase, an enzyme catalysing the interconversion of L-phenylalanine and D-phenylalanine using ATP as a cofactor. As usual for a racemase (see cluster F), this reaction has a C(R/S) stereochange as an attribute, however it also shares the presence of O–P and H–O with ten intramolecular transferase reactions (EC 5.4). Although the usage of ATP makes this enzyme unique among isomerases, its bond changes place it between two subclasses: racemases and epimerases (EC 5.1) and intramolecular transferases

(EC 5.4).

- EC 5.3.3.7 is located in cluster B and corresponds to aconitate isomerase, which catalyses the transformation of trans-aconitate and cis-aconitate. It shares the presence of C(E/Z) with eight cis-trans isomerase reactions (EC 5.2). According to bond changes only, this biochemical transformation is closer to a cis-trans isomerase reaction than to an intramolecular oxidoreductase.

- Finally, EC 5.2.1.5 is in cluster C and corresponds to linoleate isomerase and catalyses the interconversion of linoleate and 9-cis,11-trans-octadecadienoate. As most cis-trans isomerases (EC 5.2) (cluster B), it has a bond change C(E/Z) as chemical attribute, however it also shares identical bond changes to three intramolecular oxidoreductase reactions (EC 5.3), namely EC 5.3.3.8, 5.3.3.13 and 5.3.3.14.

These results suggest that only 6 clusters are enough to separate all isomerase reactions. However only clusters B and F purely correspond to subclasses EC 5.1 and 5.2. The other clusters are mixtures. Although the choice of clustering algorithm and number of clusters was optimal in terms of fitting EC subclasses, this choice was verified by visual inspection. A lower or higher number of clusters or different clustering parameters at the expense of a more coarse or fine-grained description of the clusters could have been selected. However it might not have aligned to the EC classification of isomerases as closely as this procedure.

To summarise, from the similarity and clustering by bond changes subclasses EC 5.1 and 5.2 stand out from the rest as having distinct distinct bond change attributes - only C(R/S) or C(E/Z) stereochanges, respectively - which distinguish them from the rest. The other subclasses involve more complex combinations of bond changes.

3.2 Analysis of reaction centres

Using a similar strategy as in the analysis of bond changes, this section explores the distribution, similarity and clustering of isomerase reactions on the basis of reaction centres. Results on whether isomerase EC subclasses are better described by reaction centres rather than just by bond changes is presented in *3.4 Discussion*.

3.2.1 Distribution

There are 6354 occurrences of 595 distinct reaction centres (levels 0, 1 and 2) in isomerase reactions (for details about calculating reaction centres see *2.2.2 Methods and tools*). As observed in the distribution of bond changes, the six isomerase subclasses have different numbers of reactions centres (Figure 3.6a). Intramolecular transferases (EC 5.4), followed by intramolecular oxidoreductases (EC 5.3) are the most populous subclasses by means of reaction centres. Then, medium-size subclasses are intramolecular lyases (EC 5.5) accounting for 666 reaction centres and racemases and epimerases (EC 5.1) with 420 reaction centres. Ultimately, cis-trans isomerases (EC 5.2) and other isomerases (EC 5.99) have a smaller number of reaction centres.

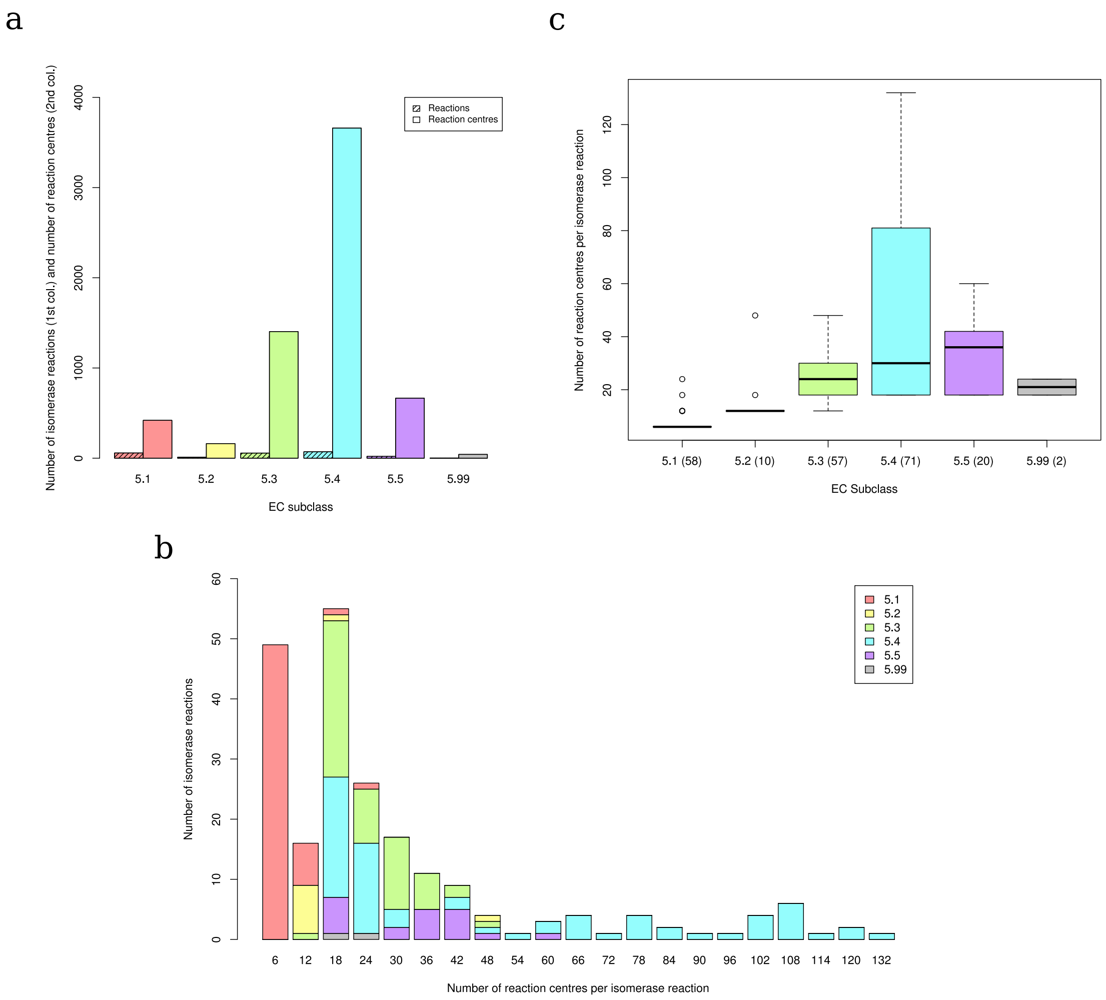


Figure 3.6: Distribution of reaction centres in isomerase reactions. (a) Distribution of the 6354 occurrences of reaction centres in 219 representative isomerase reactions across EC subclasses. (b) Bar plot showing the number of isomerase reactions against number of reaction centres. (c) Box and whisker plot representing the distribution of the number of reaction centres per isomerase reaction by EC subclass.

There is a clear distinction between EC 5.1, EC 5.2 and the rest (Figure 3.6b), as observed for bond changes. The number of reaction centres per isomerase reaction ranges from 6 to 132. It is by definition always an even number due to the atom-atom mapping between the substrates and products of the reaction. The average number of reaction centres per isomerase reaction is 29 and there are a few extreme intramolecular transferases (EC 5.4) with 114, 120 and 132 reaction centres. In general, racemases and epimerases (EC 5.1) and cis-trans isomerases (EC 5.2) catalyse either 6 or 12 reaction centres, whereas the rest of subclasses range from 18 to 132 (see Figure 3.6c).

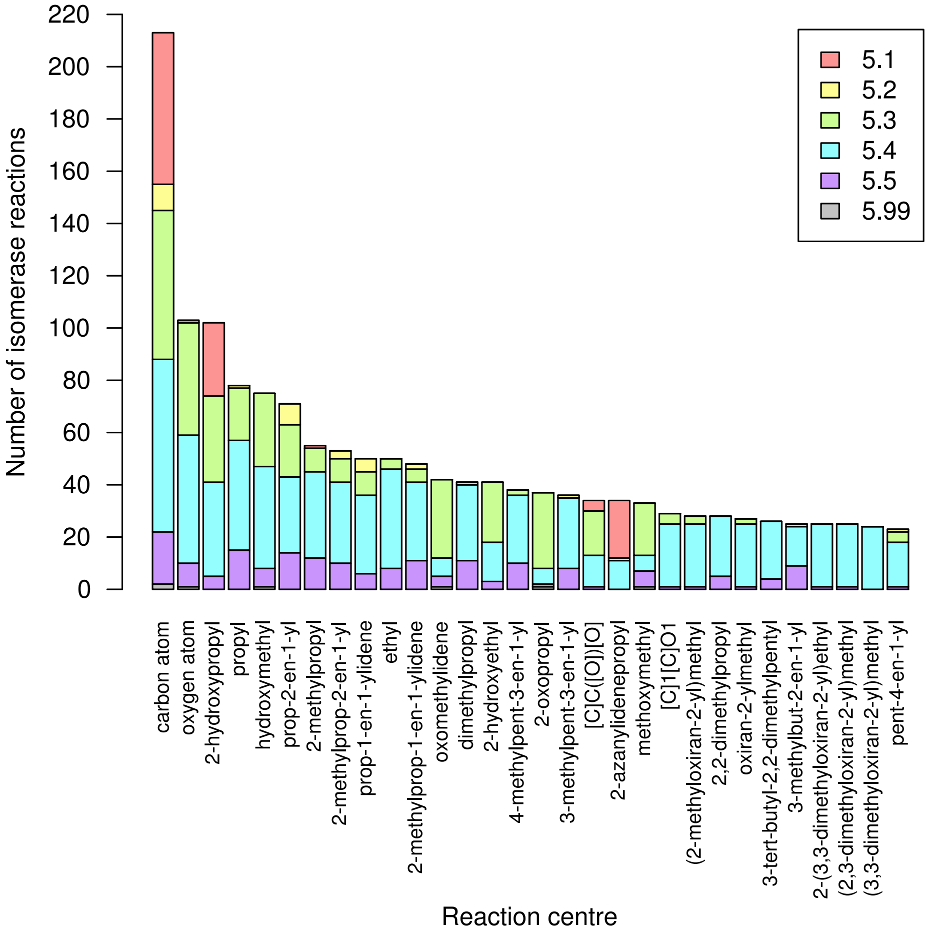


Figure 3.7: Distribution of isomerase reactions according to the 30 most common reaction centres.

Reaction centres were named using IUPAC nomenclature (IUPAC, 2014) and the most common are carbon atom, oxygen atom and 2-hydroxypropyl (Figure 3.7). Subclasses EC 5.1, 5.2 and 5.99 are connected with up to six distinct reaction centres among the most common ones. For example, EC 5.2 relates to six of them: carbon atom, propyl, prop-2-en-1-yl, 2-methylprop-2-en-1-yl, prop-1-en-1-ylidene and 2-methylprop-1-en-1-ylidene. The rest of subclasses have many of the most common reaction centres, which underlines the broad chemical diversity of these classes. For instance, all of the top 30 reaction centres are present in EC 5.4. On the whole, the top 22 reaction centres (up to 2,2-dimethylpropyl in Figure 3.7) are present in at least three subclasses, highlighting how enzymes from different subclasses can perform catalysis on the same reaction centre. As in the analysis of bond changes, subclasses EC 5.1 and 5.2 have less diversity and frequency of reaction centres compared to the other subclasses.

When calculating the degree to which reaction centres are specific to certain subclasses, we discovered that rare reaction centres, for instance (3,3-dimethyloxiran-2-yl)methyl (Figure 3.7), are only present in one subclass (SW=1). On the other hand, abundant reaction centres, such as 2-hydroxypropyl are distributed across more than one subclass (SW<1). In the former, there are 25 enzymes classified as intramolecular transferases acting on squalene derivatives (EC 5.4.99), which are active on (3,3-dimethyloxiran-2-yl)methyl as a reaction centre, for example, lanosterol synthase (EC 5.4.99.7). However in the latter, there are 102 isomerases from four different subclasses: EC 5.1, 5.3, 5.4, 5.5 catalysing isomerase reactions where 2-hydroxypropyl is a reaction centre. Overall, 468 reaction centres (79% of the total) are perfect discriminators of subclasses and some of them are present in multiple isomerase reactions of the same subclass.

3.2.2 Similarity and clustering

The distribution of pairwise similarities of isomerase reactions based on reaction centres has a different shape to the bond change distribution. Whereas bond change similarity scores adopt a reverse J-shaped distribution with three peaks at similarities 0, 0.15 and 1 (Figure 3.4a), reaction centres follow a normal distribution with positive skew (Figure 3.8a) and mean similarity of 0.22 (bond changes mean similarity is 0.20). The complete-linkage algorithm was selected to produce an optimal number of 10 clusters (Figure 3.8b and 3.9), which are summarised in Table 3.3.

Seven clusters are pure in subclasses (B', C', D', G', H', I' and J'), although three of them are singletons (D', H' and J'). The three remaining clusters are mixed. All subclasses are spread across more than one cluster. Racemases and epimerases (EC 5.1) is the subclass adopting the best clustering where 98% of reactions are placed in pure clusters whereas the rest are in mixed clusters. However, intramolecular oxidoreductases (EC 5.3) have most reactions assigned to mixed clusters.

Almost all the reactions (216 out of 219) have at least one carbon atom involved in bond changes. The three exceptions are the intramolecular transfer reactions catalysed by bisphosphoglycerate mutase (EC 5.4.2.4) and 2,3-diphosphoglycerate-dependent and independent phosphoglycerate mutase (EC 5.4.2.11 and 5.4.2.12) (see cluster G' in Figure 3.9), which catalyse the cleavage of O–P and O–H bonds only. This analysis discovered 11 reactions with *distinct* profiles of reaction centres (Table 3.3). For instance,

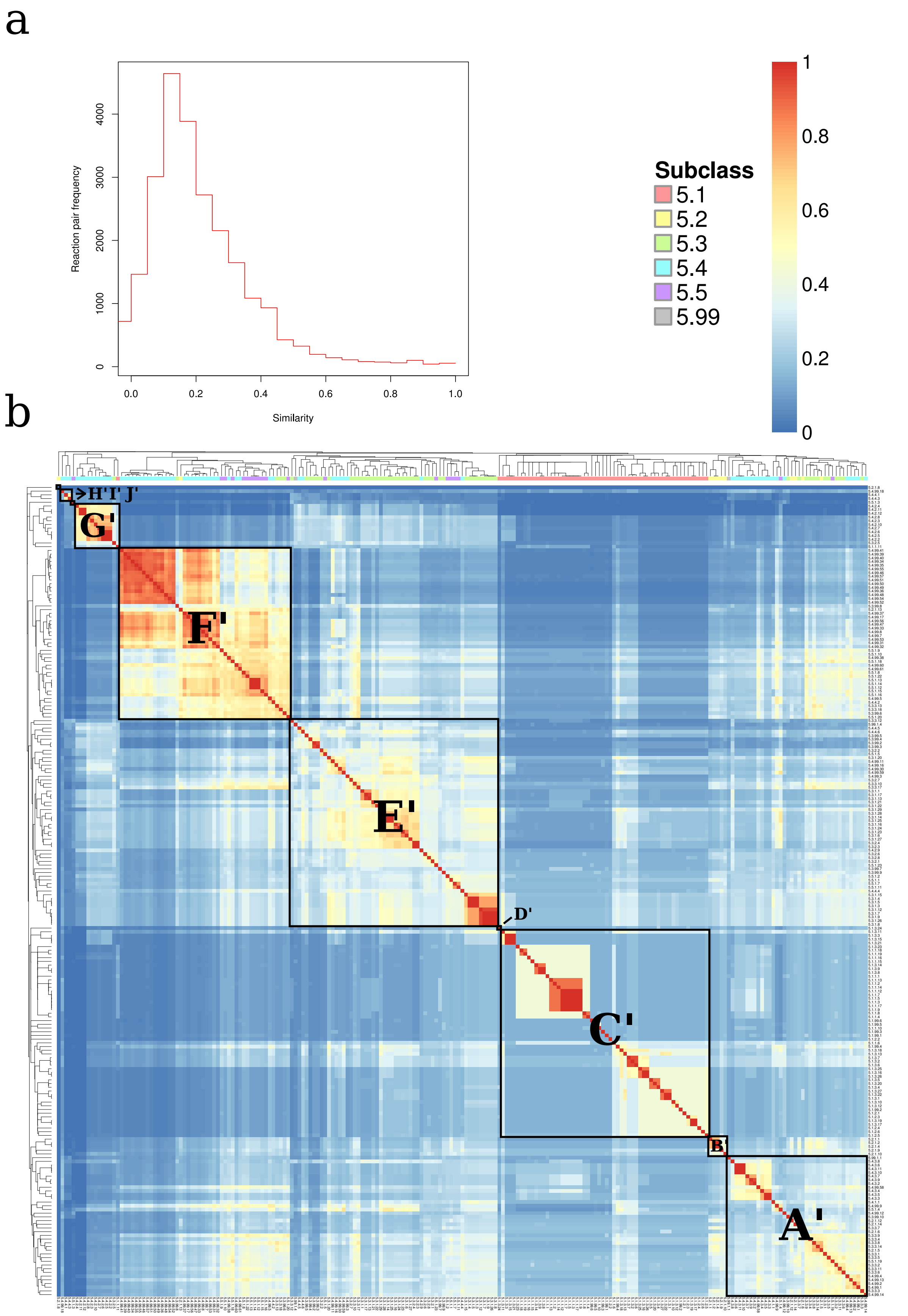


Figure 3.8: (a) Distribution of pairwise similarities of isomerase reactions by reaction centres. (b) Reaction centre similarity matrix.

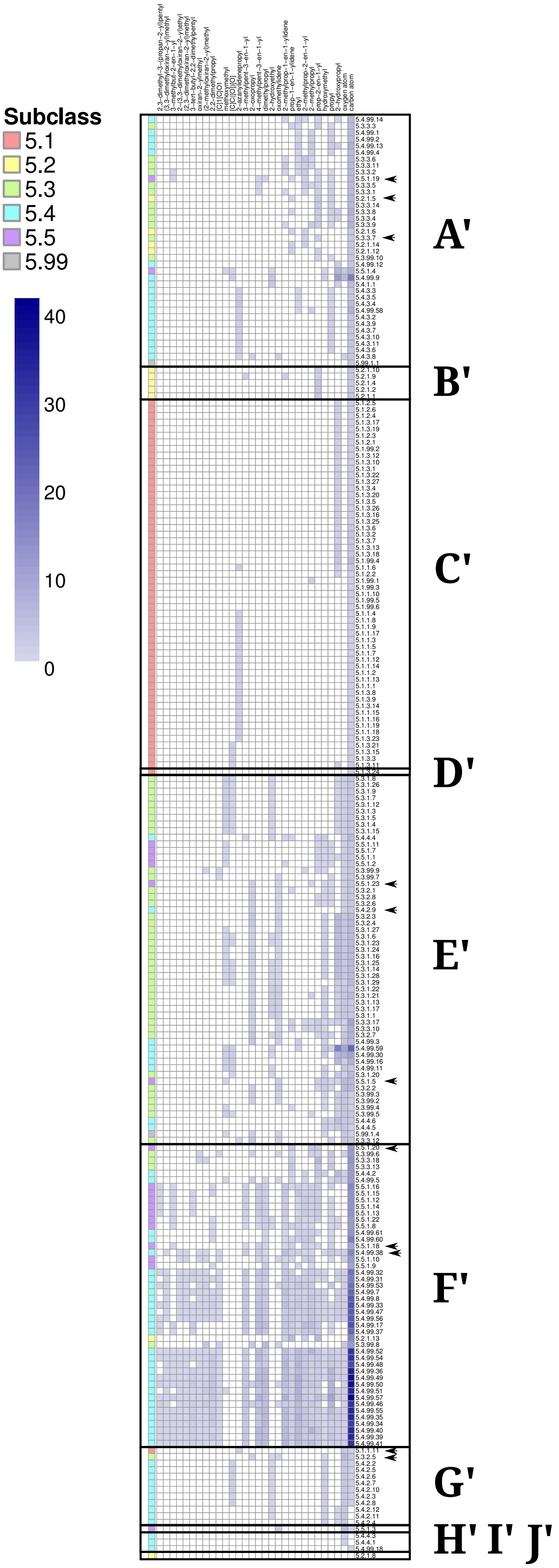


Figure 3.9: Hierarchical clustering of isomerase reactions (rows) by reaction centres (columns). Ten clusters were identified (A' to J'). Only the 30 most common reaction centres are shown.

Table 3.3: Analysis of the clusters from Figure 3.9. Columns are defined as in the bond changes analysis (Table 3.2). Cluster names are arbitrary. Clusters A' to J' obtained by reaction centres do not necessarily correspond to clusters A to F obtained by bond changes.

Cluster	**Number of reactions**	**EC subclasses**	**Most common reaction centres**	**Outliers and comments**
A'	38	EC 5.2 (4), 5.3 (12), 5.4 (19), 5.5 (2) and 5.99 (1)	carbon atom and propyl	EC 5.5.1.19, 5.2.1.5 and 5.3.3.7 are *distinct* reactions. Mixed cluster
B'	5	EC 5.2 (5)	carbon atom and prop-2-en-1-yl	No outliers. Pure cluster
C'	56	EC 5.1 (56)	carbon atom, 2-hydroxypropyl and 2-azanylidenepropyl	No outliers. Pure cluster
D'	1	EC 5.1 (1)	carbon atom	No outliers. EC 5.1.3.24 is a singleton
E'	56	EC 5.3 (40), 5.4 (9), 5.5 (6) and 5.99 (1)	carbon atom, oxygen atom, 2-hydroxypropyl, hydroxymethyl, oxomethylidene and 2-oxopropyl	EC 5.5.1.23, 5.4.2.9 and 5.5.1.5 are *distinct* reactions. Mixed cluster
F'	46	EC 5.2 (1), 5.3 (4), 5.4 (30) and 5.5 (11)	Most of the top 30 reaction centres	EC 5.5.1.20, 5.5.1.18 and 5.4.99.38 are *distinct* reactions. Mixed cluster
G'	12	EC 5.1 (1), 5.3 (1) and 5.4 (10)	Oxygen atom, hydroxymethyl and 2-hydroxyethyl	EC 5.1.1.11 and 5.3.2.5 are *distinct* reactions. Almost pure cluster.
H'	1	EC 5.5 (1)	Carbon atom, oxygen atom and oxomethylidene	No outliers. EC 5.5.1.3 is a singleton
I'	3	EC 5.4 (3)	-	No outliers. Pure cluster
J'	1	EC 5.2 (1)	-	No outliers. EC 5.2.1.8 is a singleton

situated in cluster A' aconitate isomerase (EC 5.3.3.7) shares at least two common reaction centres (carbon atom and 2-methylprop-2-en-1-yl) with furylfuramide isomerase (EC 5.2.1.6), beta-carotene isomerase (EC 5.2.1.14) and other cis-trans isomerases (EC 5.2) (see cluster B' in Figure 3.9) while remains *distinct* due to the presence of prop-2-en-1-yl and absence of prop-1-en-1-ylidene.

Although the analysis by bond changes identified three reactions that appear to be more similar to other isomerase subclasses and were labeled as *misannotated* (EC 5.1.1.11, 5.2.1.5 and 5.3.3.7), these have *distinct* profiles of reaction centres. Similarly, the two reactions annotated as *distinct* by bond changes remain *distinct* (EC 5.4.99.38) and turned into a singleton (EC 5.5.1.3) by reaction centres. The clustering by reaction centres provides a more detailed view of the overall chemistry of isomerases and helps to resolve most of the discrepancies found in bond changes.

Overall, this analysis of reaction centres supports the drastic difference in overall chemistry between EC 5.1 and 5.2 and the rest of subclasses as presented in the analysis of bond changes. Whereas reactions belonging to EC 5.1 and 5.2 contain only a few simple reaction centres such as 2-hydroxypropyl, 2-azanylidenepropyl and prop-2-en-1-yl, the rest of subclasses undergo more complex combinations of reaction centres highlighting the chemical diversity catalysed by these enzymes.

3.3 Analysis of substrates and products

The third chemical attribute used to investigate the functional similarity between isomerases is the structures of the substrates and products of the reactions. KEGG assumes all isomerase reactions to be reversible. Although this might not be true in all cases, reversibility was also accepted in this analysis so both substrates and products were broadly designated as reactants.

3.3.1 Distribution

A total of 442 occurrences of 370 unique substrates and products are present in isomerase reactions. Almost all reactions are unimolecular (a single substrate leads to a single product), the only exception is the interconversion between L-phenylalanine and D-phenylalanine catalysed by L-phenylalanine racemase (EC 5.1.1.11), which is an ATP-hydrolysing isomerase and involves three substrates (L-phenylalanine, ATP and water) and three products (D-phenylalanine, AMP and diphosphate). As a result, the number of substrates and products in an isomerase subclass is about double the number of reactions. The distributions of substrates and products per reaction are proportionally

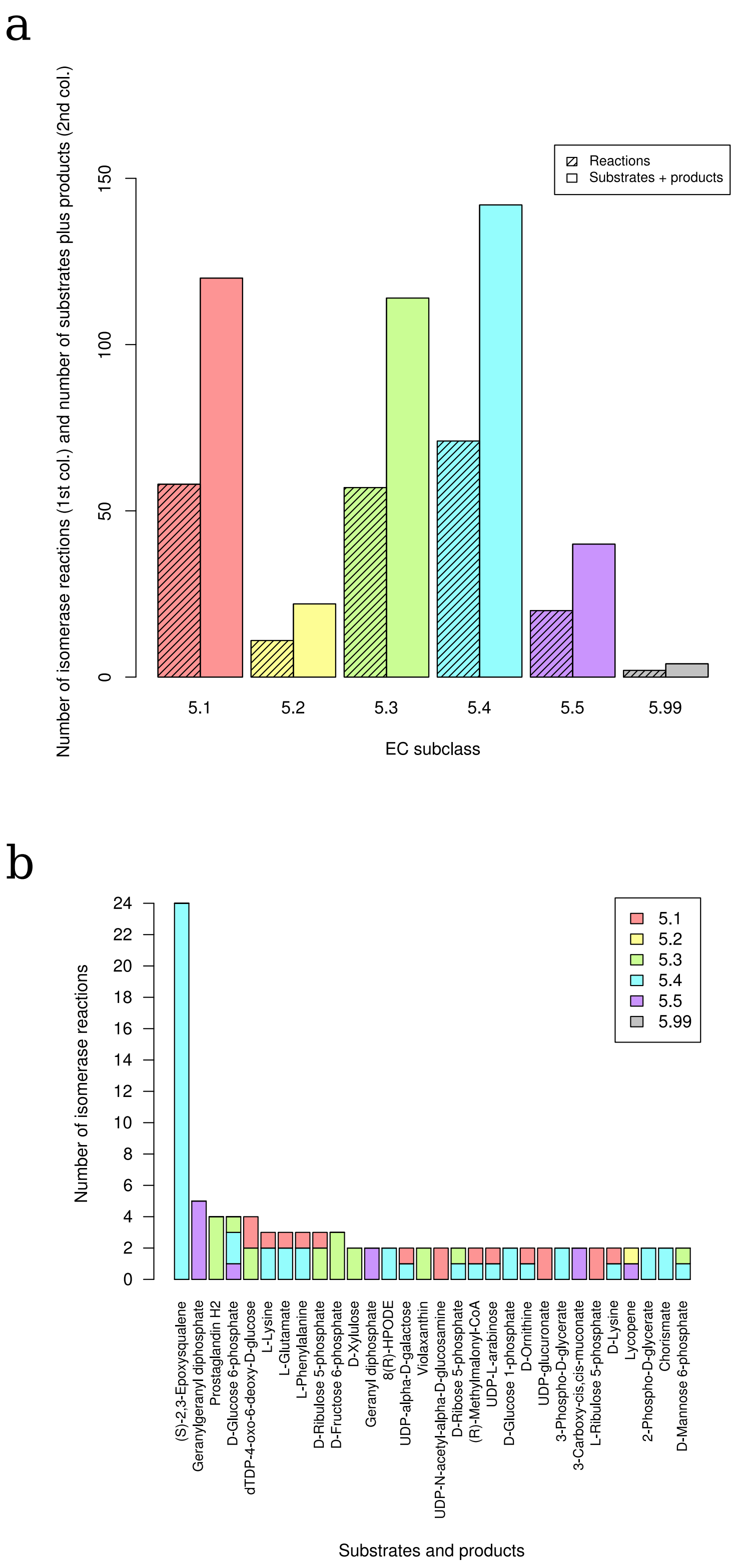


Figure 3.10: Distribution of substrates and products in isomerase reactions. (a) Distribution of the 442 occurrences of substrates and products in 219 representative isomerase reactions across EC subclasses (b) Bar plot showing the distribution of the 30 most common compounds present as substrates or products in isomerase reactions.

similar across subclasses (Figure 3.10a). Intramolecular transferases (EC 5.4), followed by racemases and epimerases (EC 5.1) and intramolecular oxidoreductases (EC 5.3) are the subclasses containing more substrates and products. With a much smaller number of reactions, intramolecular lyases (EC 5.5), cis-trans isomerases (EC 5.2) and other isomerases (EC 5.99) involve few substrates and products.

Almost 10% of the substrates and products are present in more than one reaction. The three most common substrates and products are: (S)-2,3-epoxysqualene, geranylgeranyl diphosphate and prostaglandin H2 and each exists in only one subclass: EC 5.4, 5.5 and 5.3, respectively (Figure 3.10b). Remarkably, (S)-2,3-epoxysqualene, an intermediate in the biosynthesis of terpenoids in plants, animals and fungi, is the substrate of 25 different oxidosqualene cyclases (EC 5.4.99.-), which catalyse diverse cyclisation/rearrangement reactions to produce cyclic sterols and triterpenes products (Abe, 2014). Particularly, these intramolecular transferases differ minimally in the structure of their active sites to generate structurally diverse cyclisation products. Geranylgeranyl diphosphate is also involved in cyclisation reactions undertaken by 5 different intramolecular lyases (EC 5.5.1.-) present in the mevalonate pathway of higher eukaryotes and bacteria. As for most prostaglandins, prostaglandin H2 is a lipid metabolite functioning as an important regulatory molecule in animals. It is the substrate of 4 different intramolecular oxidoreductases (EC 5.3.99.-), which share similar patterns of bond changes (Figure 3.5) and reaction centres (Figure 3.9). In contrast, the other most common substrates and products are reactants of enzymes from different subclasses. For instance, D-glucose 6-phosphate, the second metabolite of the glycolysis pathway, is the substrate of 4 isomerases from 3 different subclasses: phosphoglucose isomerase (EC 5.3.1.9), phosphoglucomutase (EC 5.4.2.2 and 5.4.2.5) and 1D-myo-inositol-3-phosphate lyase (EC 5.5.1.4).

3.3.2 Similarity and clustering

As in the analysis of reaction centres, the distribution of reaction similarities based on the structure of substrates and products adopts a normal distribution with positive skew (Figure 3.11a). Structural similarities are greater than similarities based on reaction centres with a mean similarity of 0.32. The best fit to isomerase subclasses was obtained using the complete-linkage algorithm in the form of five optimal clusters (Figures 3.11b and 3.12) summarised in Table 3.4.

All five clusters are mixed containing reactions from two or more subclasses. Racemases and epimerases (EC 5.1) and intramolecular transferases (EC 5.4) are present in all clusters. Intramolecular oxidoreductases (EC 5.3), cis-trans isomerases (EC 5.2) and

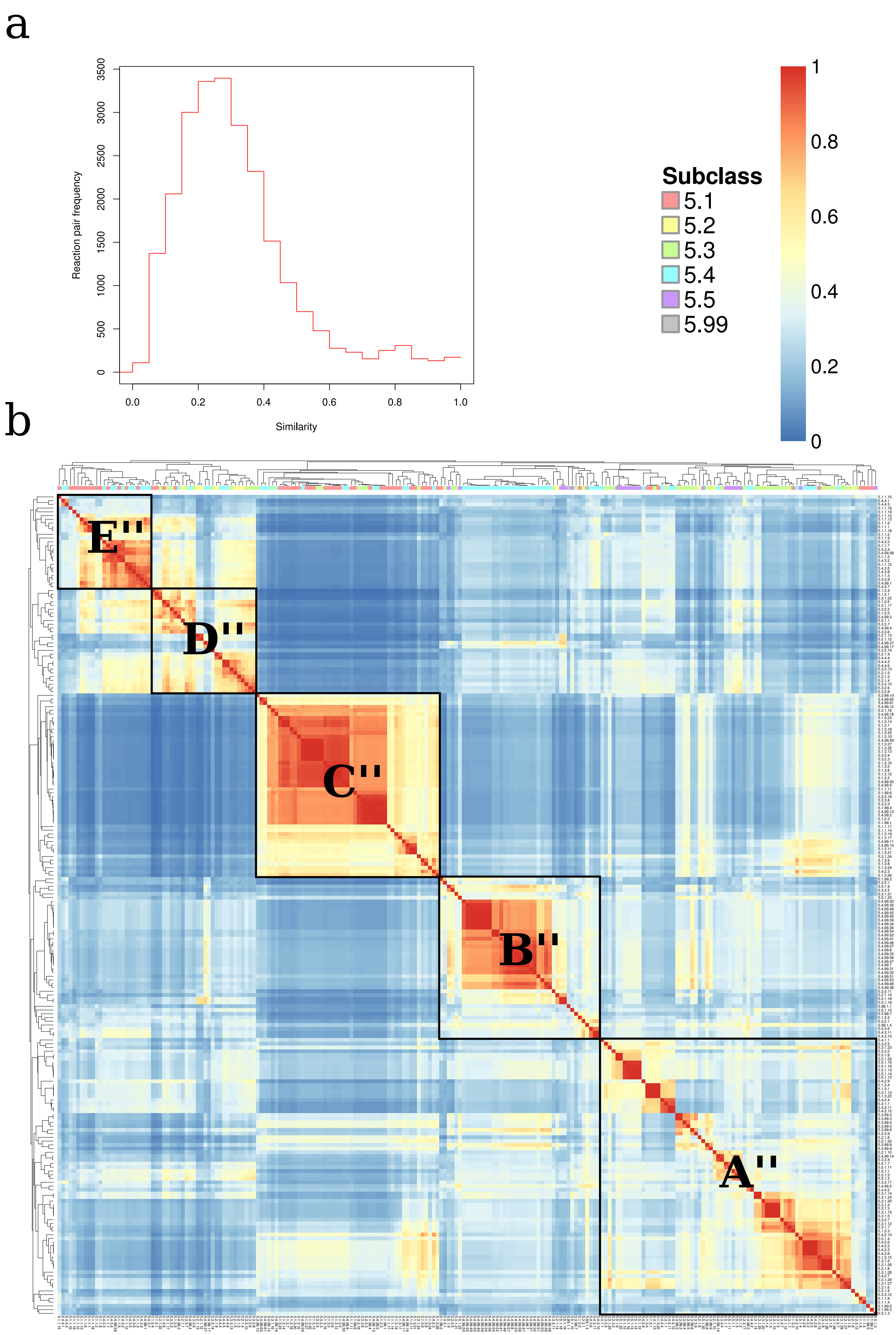


Figure 3.11: (a) Distribution of pairwise similarities of isomerase reactions by the structure of substrates and products. (b) Structure of substrates and products similarity matrix.

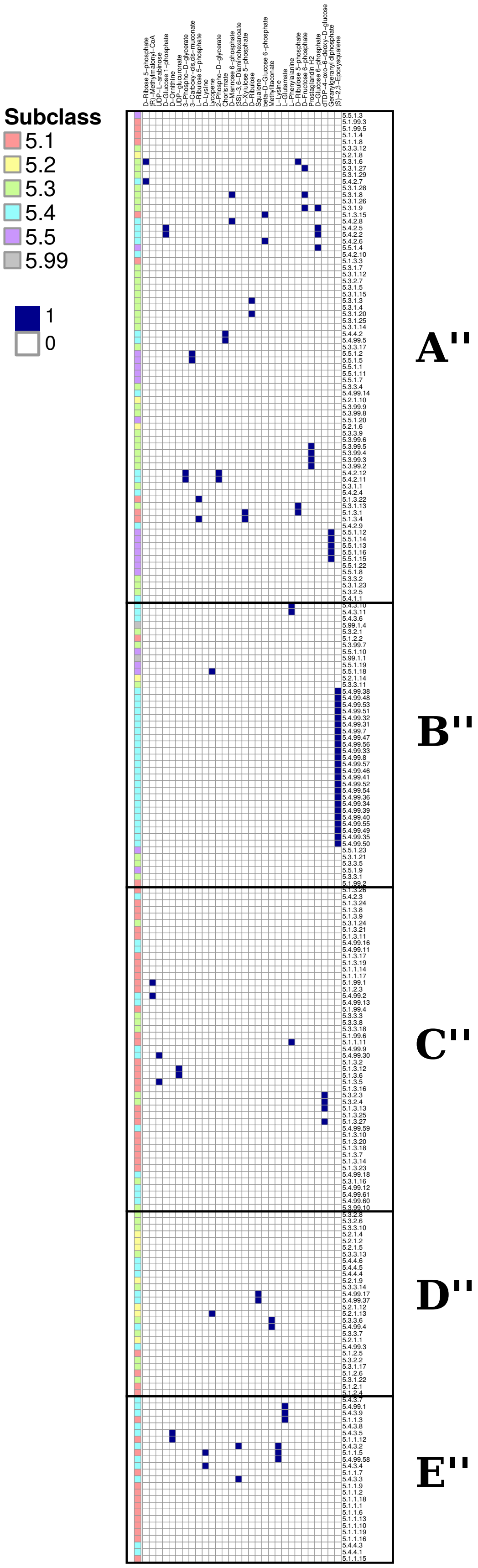


Figure 3.12: Hierarchical clustering of similarities of isomerase reactions (rows) based on the structures of substrates and products (columns). Five clusters were identified (A” to E”). Only the 30 most common reaction substrates and products are shown.

intramolecular lyases (EC 5.5) exist in four, three and two clusters, respectively. Finally, other isomerases (EC 5.99) are only present in cluster B”. More than 90% of the substrates and products are just involved in one isomerase reaction, therefore analysing outliers (*distinct* and *misannotated* reactions) is not helpful here.

Table 3.4: Analysis of the clusters from Figure 3.12. Columns are defined as in the bond changes and reaction centres analyses (Tables 3.2 and 3.3).

Cluster	**Number of reactions**	**EC subclasses**	**Common substrates and products**
A”	74	EC 5.1 (9), 5.2 (3), 5.3 (33), 5.4 (14) and 5.5 (15)	Geranylgeranyl diphosphate, D-glucose 6-phosphate and prostaglandin H2
B”	43	EC 5.1 (2), 5.2 (1), 5.3 (6), 5.4 (27), 5.5 (5) and 5.99 (2)	(S)-2,3-epoxysqualene
C”	49	EC 5.1 (29), 5.3 (8) and 5.4 (12)	dTDP-4-oxo-6-deoxy-D-glucose
D”	28	EC 5.1 (4), 5.2 (7), 5.3 (10) and 5.4 (7)	Methylitaconate and squalene
E”	25	EC 5.1 (14) and 5.4 (11)	L-glutamate and L-lysine

Although the analysis of bond changes and reaction centres revealed a clear difference in the overall chemistry of EC 5.1, 5.2 and the rest of subclasses, this distinction is not observed in the analysis of substrates and products. However this study is useful to discover similar isomerase reactions from different subclasses based on the structure of their reactants. For instance, in cluster E”, lysine racemase (EC 5.1.1.5) is adjacent to lysine 2,3-mutase (EC 5.4.3.2) and methylornithine synthase (EC 5.4.99.58). The first isomerase catalyses the racemisation of L-lysine to D-lysine. The second is a radical S-adenosyl-L-methionine (SAM) enzyme and transfers an amino group from C2 to C3 in L-lysine to produce (3S)-3,6-diaminohexanoate (Frey *et al.*, 2008). Finally, the third is also a SAM enzyme catalysing a mutase reaction that uses L-lysine to generate 3-methylornithine, a key precursor in the biosynthesis of pyrrolysine, the twenty-second proteinogenic amino acid encoded as the UAG codon in the genetic code of methanogenic archaea and bacteria (Gaston *et al.*, 2011). Although UAG is the stop codon in the standard genetic code, it is an exception in these organisms. Although the three enzymes use L-lysine as a substrate and also, the products are structurally similar, there are differences in their overall chemistry. The two SAM enzymes share similar chemistry as evidenced by bond changes and reaction centres, however the chemistry of lysine racemase is different.

3.4 Discussion

3.4.1 Overall

This chapter explores chemical attributes (bond changes, reaction centres and structures of substrates and products) of all known isomerase reactions, which were then utilised to calculate similarity between reactions using EC-BLAST and clustering of reactions into groups of similar chemistry. Although previous studies found overall agreement between clustering of biochemical reactions and the EC classification, especially in oxidoreductase (EC 1), hydrolase (EC 3) and ligase (EC 6) reactions (Egelhofer *et al.*, 2010; Holliday *et al.*, 2014; Hu *et al.*, 2010; Sacher *et al.*, 2009), this analysis demonstrates that isomerase reactions are chemically diverse and challenging to classify using a hierarchical system. Other approaches found difficulties to handle isomerase reactions because they often involve stereochanges in the structures of substrates and products only, so no bonds are formed, cleaved or order changed from an overall perspective (Apostolakis *et al.*, 2008; Chen *et al.*, 2013; Körner & Apostolakis, 2008; Latino *et al.*, 2008).

Reaction similarity and clusters obtained using bond changes are more similar to results obtained using reaction centres than those obtained using substrates and products. First, although the similarity distributions (Figures 3.4a, 3.8a and 3.11a) are all significantly different according to the Kolmogorov-Smirnov test ($p<0.001$), there is higher correlation between bond changes and reaction centres (Pearson's correlation coefficients (r) = 0.51, $p<0.001$) than between bond changes and substrates and products ($r = 0.21$, $p<0.001$) (Figure 3.13) (Oksanen, 2011; R Core Team, 2012). Second, clusters of isomerase reactions by bond changes are more similar to clusters by reaction centres than to clusters by substrates and products (Figure 3.14). Topological distances between clustering trees also confirm this observation. For example, bond change clusters A (11 reactions), D (20 reactions), E (37 reactions) and F (56 reactions) directly correspond to reaction centre clusters G', E', F' and C', respectively. However there is not a clear correspondence between clusters by bond changes and substrates and products.

This analysis presents however some clear caveats. Although mechanistic steps and cofactors are essential to understand the chemistry of isomerases and enzymes in general (Holliday *et al.*, 2009), this study ignores these aspects by relying upon the overall chemistry of reactions only. There are three main reasons for this. First, mechanistic components are not captured in reaction files (except cofactors such as ATP and NADH, which are sometimes included). Second, mechanistic details are difficult to retrieve from literature and are scarce in databases. For instance, at the time of writing only one-fifth (43) of the isomerase reactions have mechanistic data in the development version of MA-

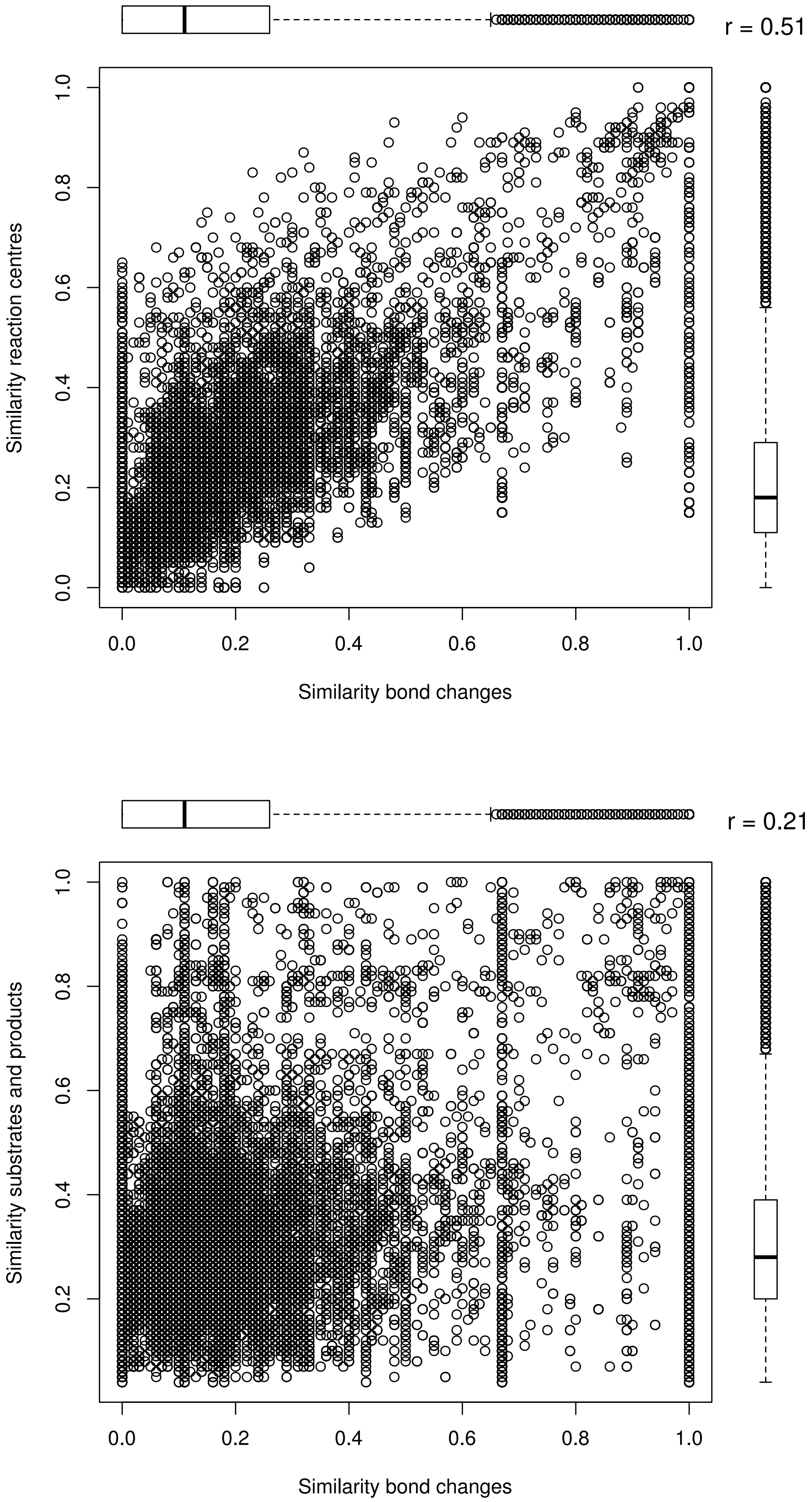
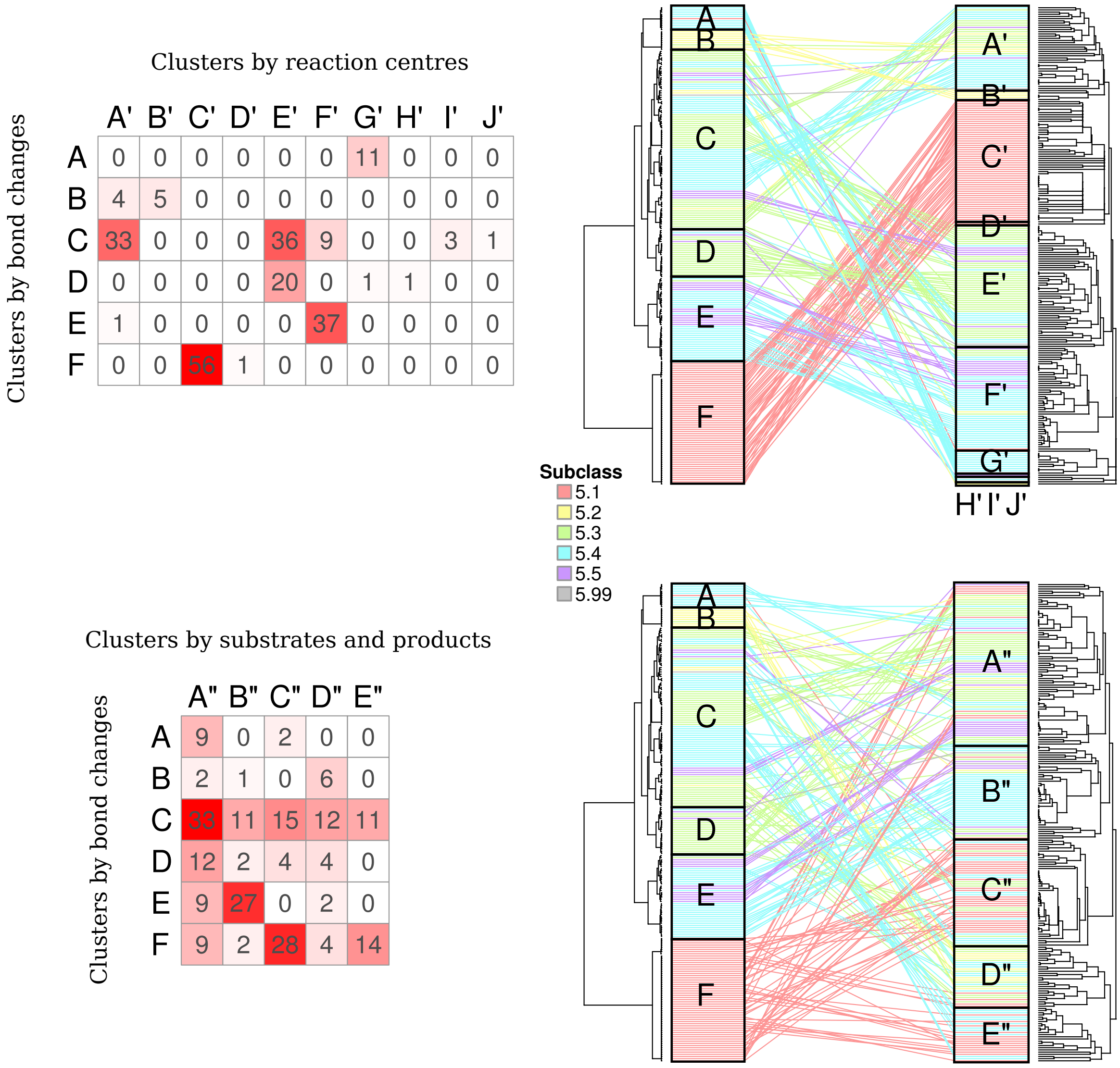


Figure 3.13: Comparison of similarity distributions of isomerase reactions as calculated by bond changes, reaction centres and substrates and products. Each point represents a pair of isomerase reactions.



Clusters by reaction centres

Clusters by bond changes

	A'	B'	C'	D'	E'	F'	G'	H'	I'	J'
A	0	0	0	0	0	0	11	0	0	0
B	4	5	0	0	0	0	0	0	0	0
C	33	0	0	0	36	9	0	0	3	1
D	0	0	0	0	20	0	1	1	0	0
E	1	0	0	0	0	37	0	0	0	0
F	0	0	56	1	0	0	0	0	0	0

Clusters by substrates and products

Clusters by bond changes

	A"	B"	C"	D"	E"
A	9	0	2	0	0
B	2	1	0	6	0
C	33	11	15	12	11
D	12	2	4	4	0
E	9	27	0	2	0
F	9	2	28	4	14

Figure 3.14: Comparison of clusters of isomerase reactions as calculated by bond changes, reaction centres and substrates and products using contingency tables (left) and tanglegrams (right) (Scornavacca *et al.*, 2011).

CiE. Third, a recent study showed that overall reaction information performs better at predicting isomerase EC numbers than mechanistic descriptors (Nath & Mitchell, 2012). Nevertheless there are still some isomerase reactions involving cyclisations of (S)-2,3-epoxysqualene and derivatives (EC 5.4.99) which are fairly complex even at the overall level. As a result, these reactions are challenging for atom-atom mapping and difficult to handle in EC-BLAST. This led to a likely overestimation of their number of bond changes (Figure 3.2a), which will be addressed in future releases of EC-BLAST.

3.4.2 Correspondence with isomerase EC subclass

The overall statistics of bond changes, reaction centres and substrates and products found in isomerase reactions are displayed in Table 3.5. Bond changes are best for partially recreating the current EC classification of isomerase reactions into six subclasses. Nevertheless, the bond change distribution is not pure, which reflects the complex, surely non-hierarchical, nature of the chemistry of isomerases. Reaction centres generate a more complex classification but support the results obtained by bond changes. Structures of substrates and products are not helpful for the EC classification but are useful to find enzymes which work on the same reactants.

Table 3.5: Table displaying statistics of the chemical attributes found in isomerase reactions. Column Types refers to the number of distinct types of chemical attributes. Columns Pairs (total, observed and frequency) represent the total number of possible pairs (Types × 219), observed pairs and frequency of observed pairs of attributes and reactions, respectively. Column specificity indicates the number and proportion of attributes specific to an isomerase subclass (see main text).

Chemical attribute	**Types**	**Pairs (total)**	**Pairs (observed)**	**Pairs (frequency)**	**Specificity**
Bond changes	30	6570	746 (11.3%)	1603	12 (40%)
Reaction centres	595	139065	3067 (2.2%)	6354	468 (79%)
Substrates and products	370	81030	442 (0.5%)	442	354 (96%)

The specificity of chemical attributes to isomerase EC subclasses (Table 3.5) is not a fundamental property of the subclass but depends on the chemical nature of reactions and their subclass association as established by the NC-IUBMB. Although specificity increases from bond changes (40%) to reaction centres (79%) to substrates and products (96%), the latter are rare given the small proportion (0.5%) of possible unique combinations that are observed. The ability of the NC-IUBMB to manually update the EC classification in the form of new, transferred and deleted reactions as new enzyme data becomes available will change the specificity of chemical attributes for subclasses. In line with this, the analysis of bond changes first highlighted potential inconsistencies for three isomerase EC numbers, which were later recognised as *distinct* reactions in the analysis of reaction centres.

Isomerases are a rare class of enzymes. Unlike other EC classes such as the ligases (EC 6) (Holliday *et al.*, 2014), their functional classification is rather complex. First, the frequency of bond changes and reaction centres is drastically different between subclasses. Racemases and epimerases and cis-trans isomerases (EC 5.1 and 5.2) have relatively few bond changes and reaction centres compared to intramolecular oxidoreductases, intramolecular transferases and intramolecular lyases (EC 5.3, 5.4 and 5.5). Second, the overall chemistry differs significantly between subclasses as well. While EC 5.1 and 5.2 are sensibly grouped according to changes of stereochemistry (clusters B and F in Figure 3.5 and clusters A', B' and C' in Figure 3.9), EC 5.3, 5.4 and 5.5 are present in mixed clusters and are also very similar in overall chemistry to other EC primary classes, oxidoreductases (EC 1), transferases (EC 2) and lyases (EC 4), respectively. The subclass "other isomerases" (EC 5.99) sits apart from the rest and exhibits great diversity, as evidenced by the distinct chemistry of winding DNA catalysed by topoisomerases (O'Brien, 2006).

The overall chemistry of isomerases does not always involve stereochanges (C(R/S) or C(E/Z)), there are 45 isomerase reactions that do not have any stereochange (Figure 3.5). According to the NC-IUBMB, isomerases catalyse structural rearrangements between isomers, namely the substrate and product have the same molecular formula but different chemical structures. This definition obviates the need to know whether reactions undergo stereochanges. For example, isopentenyl-diphosphate Δ-isomerase (EC 5.3.3.2) interconverts two isomers: isopentenyl diphosphate and dimethylallyl diphosphate, but no stereochange takes place (see Figure 2.3 from *Chapter 2. Data Resources and Methods*). The first step of the reaction is protonation of a C=C bond and formation of a carbocation, and then reformation of the C=C bond by deprotonation of an adjacent carbon (MACiE mechanism M0190 (Holliday *et al.*, 2012)).

The EC classification of isomerases in six subclasses can effectively be simplified according to the IUPAC definitions of isomers (IUPAC, 2014). There are two types:

(a) **Stereoisomers:** molecules only differ on the spatial location of atoms without any differences in atom connectivity. The substrate and product of enzymes belonging to subclasses EC 5.1 and 5.2 are stereoisomers. As a result, the only chemical attributes are stereochanges such as R/S and E/Z isomerisations.

(b) **Structural isomers:** molecules differ in atom connectivity, which as a result changes the spatial location of atoms as well. The substrate and product of enzymes belonging to subclasses EC 5.3, 5.4 and 5.5 are structural isomers. Chemical attributes are not only stereochanges, but also formed/cleaved bonds and bond order changes.

The clear separation between EC 5.1 and 5.2 and the rest of isomerase subclasses observed in the analysis of bond changes and reaction centres suggests that classifying isomerases in two subclasses according to the type of isomerism between substrate and product is chemically more sensible than the current classification in six subclasses. This recommendation involves the reorganisation of isomerases in two groups: stereoisomerases (current

EC 5.1 and 5.2) and structural isomerases (current EC 5.3, 5.4 and 5.5). Regarding EC 5.99, topoisomerases (EC 5.99.1.2 and 5.99.1.3) change the topology of DNA. Although the overall chemistry of these enzymes is not described by any descriptor, they qualify as stereoisomerases because they change the spatial location of atoms while mantaining the atom connectivity. In addition, their catalytic mechanism involve catalysis of O–P bonds in the phosphate backbone of DNA (MACiE entries M0064, M0232 and M0366). Finally, thiocyanate isomerase (EC 5.99.1.1) and 2-hydroxychromene-2-carboxylate isomerase (EC 5.99.1.4) qualify as structural isomerases.

3.4.3 Correspondence with isomerase EC sub-subclass

Only the subclasses racemases and epimerases (EC 5.1), intramolecular oxidoreductases (EC 5.3) and intramolecular transferases (EC 5.4) are further divided into sub-subclasses (Figure 3.15 and Table 3.6).

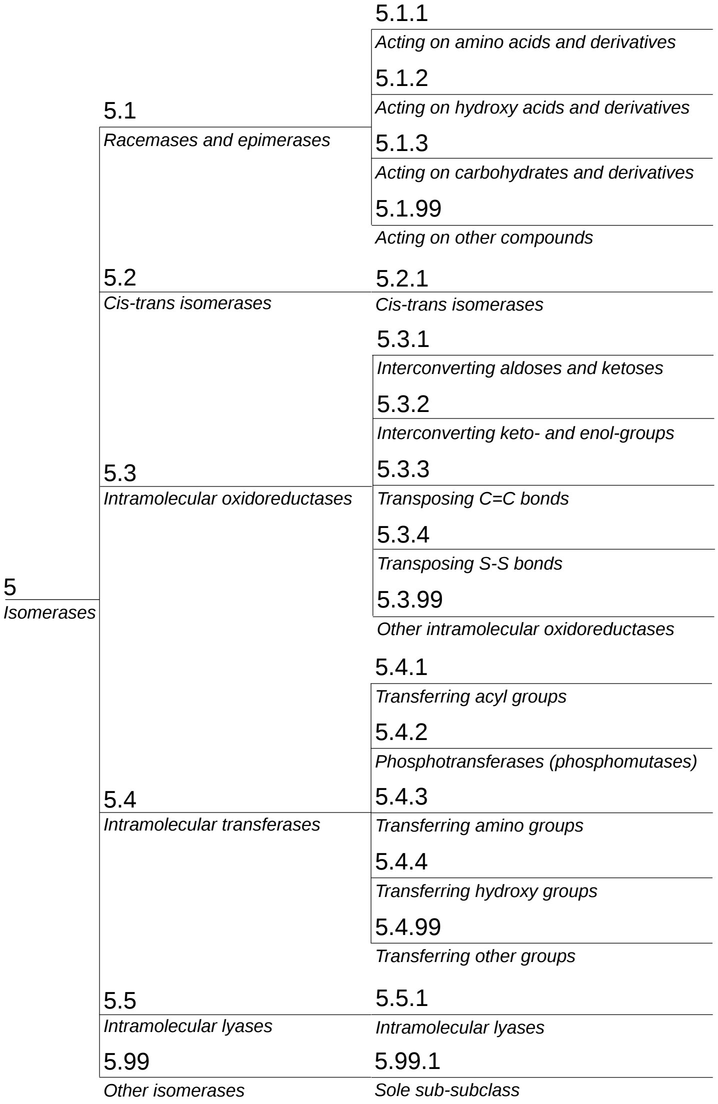
Table 3.6: Statistics of the chemical attributes of the isomerase subclasses that split into sub-subclasses: racemase and epimerase (EC 5.1), intramolecular oxidoreductase (EC 5.3) and intramolecular transferase (EC 5.4) reactions.

	EC 5.1	**EC 5.3**	**EC 5.4**
Reactions	58	57	71
Bond changes	3	16	22
Reaction centres	61	233	309
Susbtrates and products	115	104	108

Racemase and epimerase reactions (EC 5.1)

Fifty-eight racemases and epimerases catalyse chiral inversions between stereoisomers, also known as changes in absolute stereochemistry (R or S) of asymmetric carbon atoms. This subclass is divided into 4 sub-subclasses depending on the nature of the substrate. Racemases from EC 5.1.1 act on amino acids and derivatives, EC 5.1.2 on hydroxy acids and derivatives, EC 5.1.3 on carbohydrates and derivatives and EC 5.1.99 on “other compounds”. The analysis of bond changes is simple, only three bond changes are involved, all EC 5.1 reactions have only C(R/S) with the exception of phenylalanine racemase (EC 5.1.1.11), which also catalyses O–P and O–H. The similarity and clustering analyses based on the 61 unique reaction centres and the 115 substrates and products suggest two complementary ways to explore these sub-subclasses (Figure 3.16).

Both approaches set EC 5.1.1 and 5.1.3 reactions apart from each other in two different clusters. EC 5.1.1 reactions are characterised by 2-azanylidenepropyl as the most common reaction centre in amino acids as reactants. In contrast, EC 5.1.3 reactions are characterised by the reaction centre 2-hydroxypropyl in sugars and derivatives. However,



5
Isomerases
5.1
Racemases and epimerases
5.1.1
Acting on amino acids and derivatives
5.1.2
Acting on hydroxy acids and derivatives
5.1.3
Acting on carbohydrates and derivatives
5.1.99
Acting on other compounds
5.2
Cis-trans isomerases
5.2.1
Cis-trans isomerases
5.3
Intramolecular oxidoreductases
5.3.1
Interconverting aldoses and ketoses
5.3.2
Interconverting keto- and enol-groups
5.3.3
Transposing C=C bonds
5.3.4
Transposing S-S bonds
5.3.99
Other intramolecular oxidoreductases
5.4
Intramolecular transferases
5.4.1
Transferring acyl groups
5.4.2
Phosphotransferases (phosphomutases)
5.4.3
Transferring amino groups
5.4.4
Transferring hydroxy groups
5.4.99
Transferring other groups
5.5
Intramolecular lyases
5.5.1
Intramolecular lyases
5.99
Other isomerases
5.99.1
Sole sub-subclass

Figure 3.15: EC classification of isomerases in subclasses and sub-subclasses.

there are also reactions sharing characteristic attributes of the two sub-subclasses. For instance, epimerases acting on glucosamine and derivatives (EC 5.1.3.8, 5.1.3.9, 5.1.3.14 and 5.1.3.23) catalyse transformations of the reaction centre 2-azanylidenepropyl not in amino acids but in sugar molecules. Alternatively, threonine racemase (EC 5.1.1.6) acts on the amino acid threonine even though 2-hydroxypropyl is also a reaction centre. In addition, amino acid racemase (EC 5.1.1.10) and hydantoin racemase (EC 5.1.99.5) form a separate cluster in Figure 3.16a because they catalyse reactions containing a generic substructure or R-group in the reaction centre (for a detailed explanation of R-groups see *Chapter 5. Characterising Complex Biochemical Reaction Data*).

Although EC 5.1.2 reactions group with EC 5.1.3 due to the presence of the 2-hydroxypropyl as reaction centre (Figure 3.16a), they cluster with EC 5.1.1 reactions in substrates and products because hydroxy acids resemble amino acids more than carbohydrates at

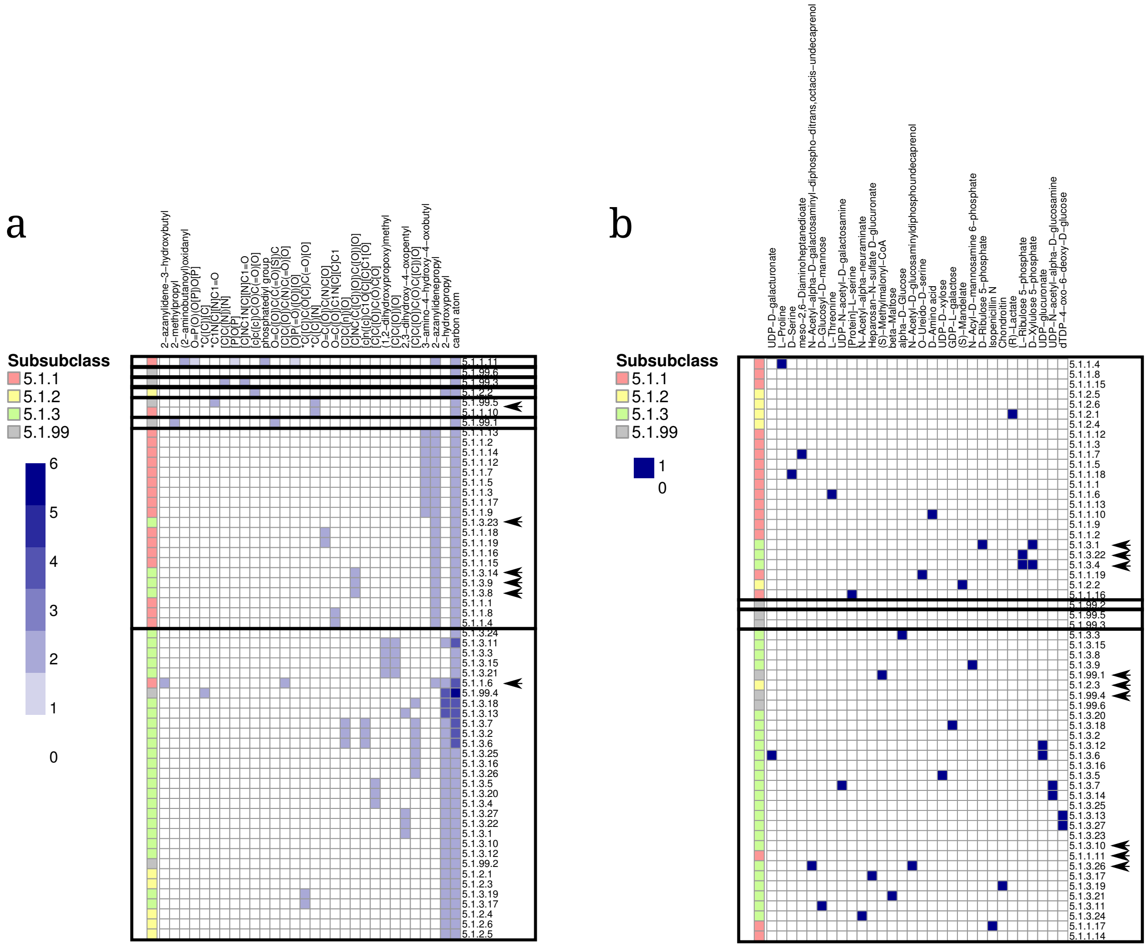


Figure 3.16: Hierarchical clustering of racemase and epimerase reactions (rows) based on (a) reaction centres and (b) structures of substrates and products (columns). Only the thirty most common chemical attributes are shown. Clusters are indicated in boxes and reactions discussed in the main text are marked with arrows.

the overall structure level (Figure 3.16b). Exceptionally, the reaction catalysed by 3-hydroxybutyryl coenzyme A (CoA) epimerase (EC 5.1.2.3) groups with EC 5.1.99 reactions because they share CoA as a substructure. Lastly, the reaction catalysed by phenylalanine racemase (EC 5.1.1.11) shares the diphosphate substructure with the substrates of some sugar epimerases (EC 5.1.3.10 and 5.1.3.26).

To summarise, this analysis shows a clear separation between the overall chemistry of EC 5.1.1 and 5.1.3 reactions. The results are consistent with previous investigations of the overall chemistry of EC 5.1 reactions using a different strategy based on chirality codes and self-organising maps (Latino *et al.*, 2008). EC 5.1.2 and 5.1.99 reactions are diverse and group alternately with EC 5.1.1 and 5.1.3.

Intramolecular oxidoreductase reactions (EC 5.3)

The larger number of chemical attributes of the 57 intramolecular oxidoreductases (EC 5.3) indicates a more diverse chemistry in comparison with EC 5.1 reactions. This subclass is divided in 5 sub-subclasses. EC 5.3.1 interconvert aldoses and ketoses, EC 5.3.2 transform keto and enol groups, EC 5.3.3 and 5.3.4 transpose C=C and S–S bonds, respectively, and EC 5.3.99 act on other substrates. This subclass is characterised by 16 bond changes, 233 reaction centres and 104 substrates and products (Table 3.6). Sub-subclass EC 5.3.4 contains only one reaction catalysed by protein disulfide-isomerase (EC 5.3.4.1). The substrate of this enzyme is a protein, which does not have a structure file available, therefore it is not considered in this study.

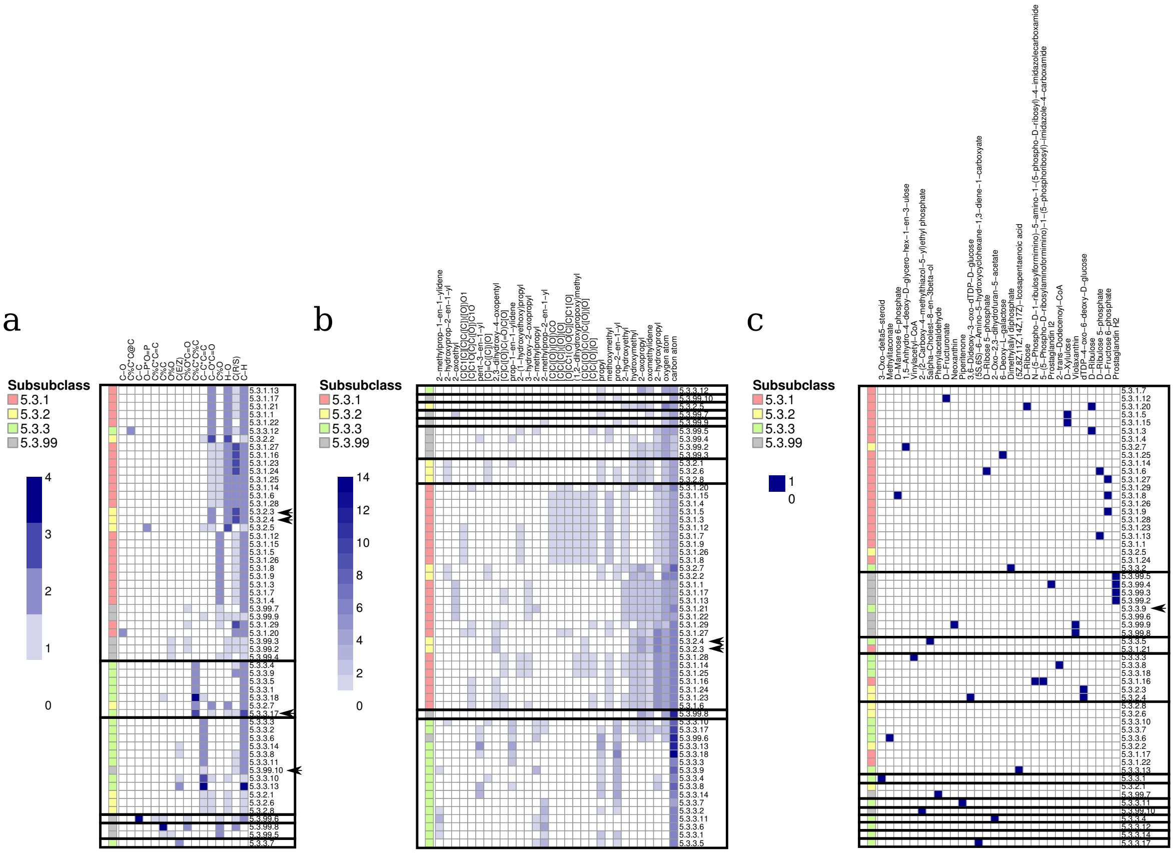


Figure 3.17: Hierarchical clustering of intramolecular oxidoreductase reactions (rows) based on (a) bond changes, (b) reaction centres and (c) structures of substrates and products (columns). Only the thirty most common chemical attributes are shown in (b) and (c). Clusters are indicated in boxes and reactions discussed in the main text are marked with arrows.

Overall, cluster analysis indicates a clear distinction between the overall chemistry of EC 5.3.1 and 5.3.3 reactions (Figure 3.17). Despite EC 5.3.2 reactions clustering with EC 5.3.1 and 5.3.3 in the analysis of bond changes, they solely group with EC 5.3.1 according to reaction centres but with EC 5.3.3 in substrates and products. For example, the reactions catalysed by TDP-6-deoxy-hex-4-ulose isomerase (EC 5.3.2.3) and TDP-4-oxo-6-deoxy-alpha-D-glucose-3,4-oxoisomerase (EC 5.3.2.4) share similar bond changes and reaction centres with EC 5.3.1 reactions, however the overall structure of their reactants resemble EC 5.3.3 reactions more closely. Similarly, the reaction catalysed by trans-2,3-dihydro-3-hydroxyanthranilate isomerase (EC 5.3.3.17) shares H–O and C–O ↔ C=O with EC 5.3.2 reactions. Finally, EC 5.3.99 reactions group with 5.3.1 and 5.3.3 in bond changes but form separate clusters in reaction centres and substrates and products. For instance, thiazole tautomerase (EC 5.3.99.10) catalyses a C=C transposition, which is the most common bond change in EC 5.3.3 reactions. However, prostaglandin A isomerase (EC 5.3.3.9) also catalyses a C=C transposition, but its substrate is (13E)-(15S)-15-Hydroxy-9-oxoprosta-10,13-dienoate, which is structurally similar to prostaglandin H2, a common substrate in EC 5.3.99.

Intramolecular transferase reactions (EC 5.4)

The 71 intramolecular transferase reactions (EC 5.4) are grouped into 5 sub-subclasses depending on the nature of the transferred chemical group. EC 5.4.1 involve acyl groups, EC 5.4.2 transfer phospho groups, EC 5.4.3 shift amino groups, EC 5.4.4 involve hydroxy groups and last but not least, EC 5.4.99 transfer other groups. For instance, isochorismate synthase (EC 5.4.4.2) and chorismate mutase (EC 5.4.99.5) isomerise the substrate chorismate into isochorismate and prephenate, respectively. Although both reactions share the same substrate and similar bond changes and reaction centres, the former involves the transfer of a hydroxy group whereas the latter converts a 2-hydroxyprop-2-enoic acid group. The subclass EC 5.4 is dominated by the sub-subclass EC 5.4.99 where 44 reactions account for 62% of all EC 5.4 reactions and one-fifth of all EC 5 reactions, therefore defining EC 5.4.99 as the most populated sub-subclass in isomerases. Chemically speaking, different types of bond changes and reaction centres result in a complex and diverse overall chemistry, especially the cyclisation reactions catalysed by oxidosqualene cyclases (Figure 3.18).

EC 5.4.1 contains only one reaction catalysed by lysolecithin acylmutase (EC 5.4.1.1). EC 5.4.2 reactions involve O–P bonds and group together according to bond changes and reactions centres (Figure 3.18a,b), however in the analysis of substrates and products they split into two groups according to acyclic and cyclic reactants, especially sugar phosphates (Figure 3.18c). Interestingly, the reaction catalysed by phosphoenolpyruvate mutase (EC 5.4.2.9) sits apart from the rest of EC 5.4.2 reactions because it involves the formation of C–P bonds. This extra chemical ability has been extensively explored since it allows the biosynthesis of phosphonates in nature (Yu *et al.*, 2013). EC 5.4.3 reactions

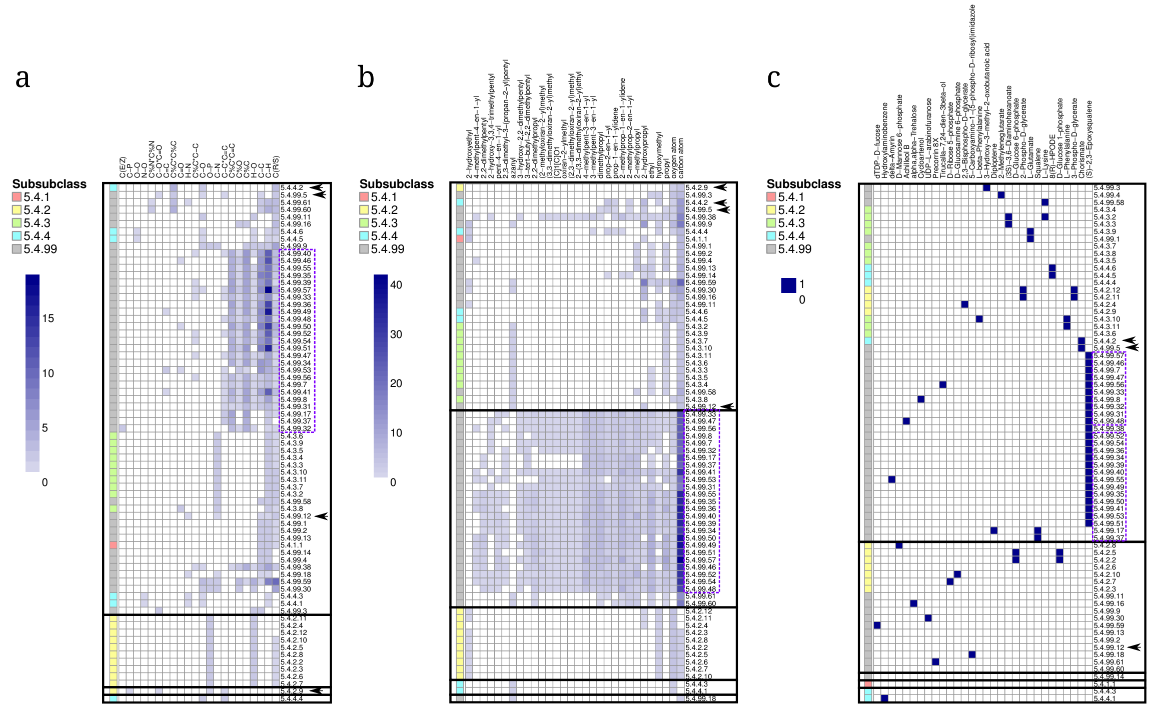


Figure 3.18: Hierarchical clustering of intramolecular transferase reactions (rows) based on (a) bond changes, (b) reaction centres and (c) structures of substrates and products (columns). Only the thirty most common chemical attributes are shown in (b) and (c). Clusters are indicated in black boxes, oxidosqualene cyclisations are highlighted in purple boxes and reactions discussed in the main text are marked with arrows.

involve C–N bonds and cluster by bond changes and reaction centres in a similar way that EC 5.4.2 reactions do, however the analysis of their substrates and products segregates them into acyclic and aromatic reactants. EC 5.4.4 are diverse and spread across multiple clusters. Finally, in order to dissect the complexity of EC 5.4.99 reactions, the clustering results suggest that splitting EC 5.4.99 into additional sub-subclasses would be sensible according to an overall chemistry point of view. First, the 25 oxidosqualene cyclisations (EC 5.4.99.X, X = 7, 8, 17, 31-37, 39-41 and 46-57) act on (S)-2,3-epoxysqualene and derivatives and their bond changes, reaction centres and substrates and products are different from EC 5.4.99 and the rest of the isomerase reactions. Second, the chemistry of the 16 RNA pseudouridine synthases (EC 5.4.99.X, X = 12, 19-30 and 42-45) also sets apart from the rest of the EC 5.4.99 reactions. These enzymes post-transcriptionally isomerise specific uridine residues to pseudouridine in RNA. They all share the same overall chemistry but differ in the type of RNA (tRNA or rRNA) and the sequence sites where modifications take place (Hamma & Ferré-D'Amaré, 2006). For example, tRNA pseudouridine38-40 synthase (EC 5.4.99.12) isomerises the uridine residues at positions 38, 39 and 40 of nearly all tRNAs and it is the only reaction with structural data available in this analysis (Figure 3.18). Third, there are 6 mutases catalysing carbon rearrangements (EC 5.4.99.X, X = 1, 2, 4, 13, 14 and 18), which also share similar overall chemistry. Overall, these three distinct groups of reactions within EC 5.4.99 could sensibly be considered as three new sub-subclasses.

3.5 Conclusion

The EC classification of isomerases is not easy. The various criteria used to define EC subclasses and sub-subclasses do not entirely follow a sensible classification based on chemical attributes of the overall reaction such as bond changes, reaction centres and substrates and products. This analysis suggests two main ways to improve the EC classification. First, the current classification of isomerases into six subclasses can be reduced to two subclasses according to the type of isomerism shared between substrate and product. Racemases and epimerases (EC 5.1) and cis-trans isomerases (EC 5.2) group under a “metaclass” entitled stereoisomerism whereas intramolecular oxidoreductases (EC 5.3), intramolecular transferases (EC 5.4), intramolecular lyases (EC 5.5) and “other isomerases” (EC 5.99) represent structural isomerism. Second, complex sub-subclasses such as intramolecular transferases acting on “other groups” (EC 5.4.99) can be further split into groups of similar chemistry such as oxidosqualene cyclases and pseudouridine synthases in order to make the EC classification more useful.

Chapter 4

The Evolution of Isomerase Function

Almost thirty years ago, Chothia and Lesk started investigating the relationship between sequence and structural divergence in related proteins (Chothia & Lesk, 1986). A few years later, scientists began to explore the origins of the evolution of enzyme function using techniques from molecular biology, crystallography and enzymology (Petsko *et al.*, 1993). Today, evidence suggests that the major route for creating new enzyme functions is gene duplication and subsequent evolution of one enzyme to another with a novel, though usually related, function (Copley, 2012). New computational approaches to measure functional similarity between enzymes extend the existing evolutionary studies based on sequence and structure. In this chapter our knowledge of the evolution of the isomerase class of reactions in enzyme superfamilies is reviewed, using newly developed tools to compare enzyme reactions (Rahman *et al.*, 2014) and their evolution (Furnham *et al.*, 2012*b*).

The study of protein superfamilies has revealed how enzymes evolve their overall chemistry and mechanism during evolution (Bartlett *et al.*, 2003). Several groups have already attempted to explore the evolution of function in specific isomerases. Over the last two decades, comprehensive analyses of the enolase superfamily showed that mandelate racemase (EC 5.1.2.2) and muconate-lactonizing enzyme (EC 5.5.1.1) are related by divergent evolution from a common ancestor and they catalyse similar mechanisms but different overall chemistry (Gerlt *et al.*, 2012; Petsko *et al.*, 1993). Subsequent studies on sugar isomerases found that common ancestry, similar overall chemistry but different mechanisms is also an alternative evolutionary pathway (Banerjee *et al.*, 1995). Recent experimental analyses presented evidence of how isomerase function exchanges with other EC classes (Table 4.1).

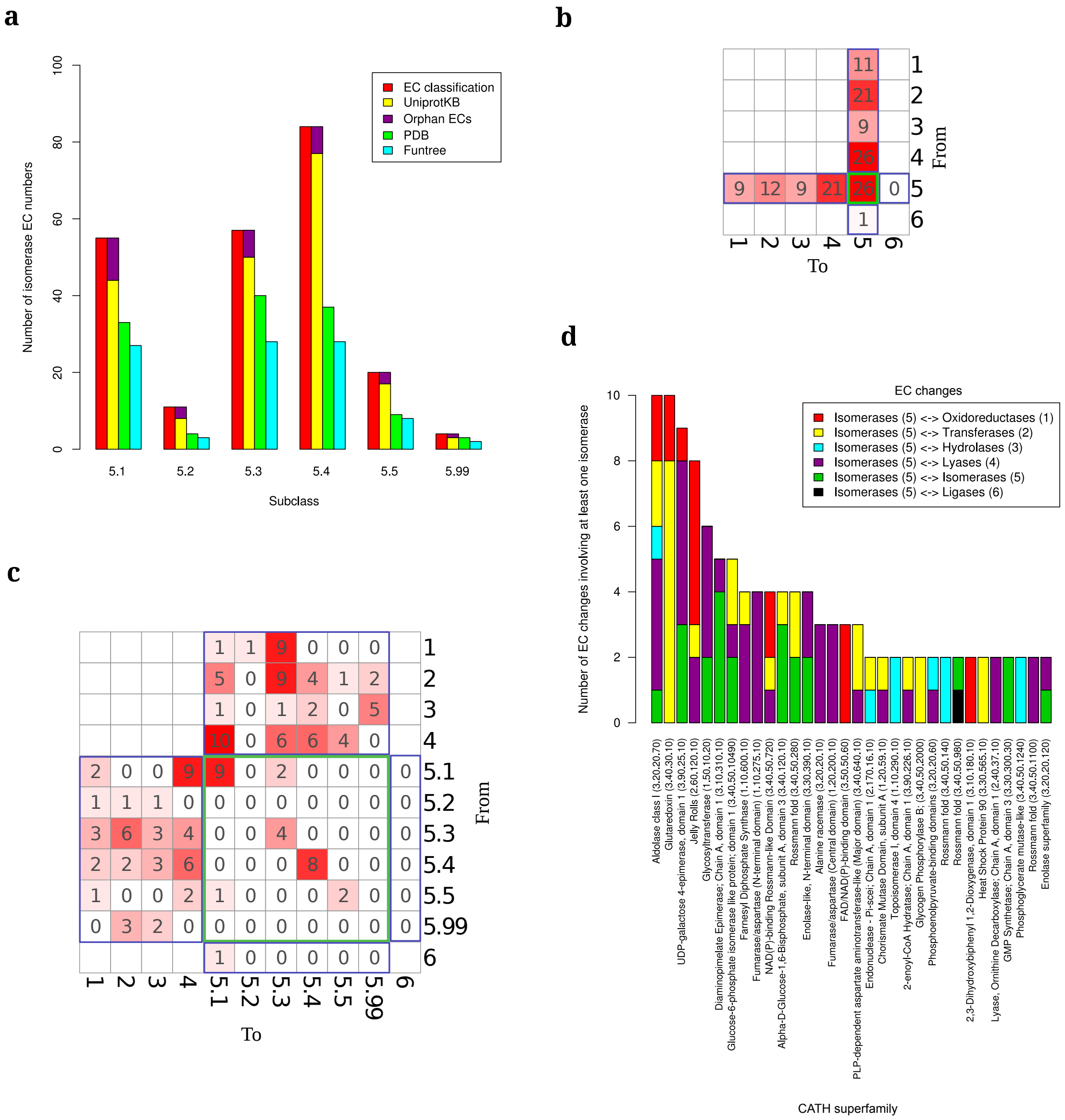
Table 4.1: Summary table of experimental studies exploring the evolution of isomerase function.

Evolved from	Evolved to	Superfamily	Species	References
O-succinylbenzoate synthases (EC 4.2.1.113)	N-succinylamino acid racemase (EC 5.1.1.-)	Enolase	*E. coli*, *B. subtilis*, *G. stearothermophilus* and *S.* sp. PCC 6803	Glasner *et al.* (2006*a*)
Aspartate aminotransferase (EC 2.6.1.1)	Alanine, glutamate and aspartate racemases (EC 5.1.1.1, 5.1.1.3 and 5.1.1.13)	Pyridoxal 5'-phosphate-dependent enzymes	*E. coli*	O'Brien & Herschlag (1999); Vacca *et al.* (1997)
Isochorismate pyruvate-lyase (EC 4.2.99.21)	Chorismate mutase (EC 5.4.99.5)	Chorismate mutase domain	*P. aeruginosa*	Künzler *et al.* (2005); Schulenburg & Miller (2014)
Various hydrolases and lyases (EC 3 and 4)	Uronate isomerase (EC 5.3.1.12)	Amidohydrolase	*E. coli* and *B. halodurans*	Nguyen *et al.* (2008, 2009)
Maleylacetoacetate isomerase (EC 5.2.1.2)	TCHQ dehalogenase and 2,5-DCHQ dehalogenase (EC 1.97.1.-)	Glutathione S-transferase	*S. chlorophenolica*	Anandarajah *et al.* (2000)

4.1 Availability of data on isomerase sequences and structures

In 24th July 2013, the NC-IUBMB listed 231 active four-digit isomerase EC numbers in the classification. Information about the enzyme sequences is easily accessible in UniprotKB (The Uniprot Consortium, 2013), 199 of them have sequence information and 32 are orphan isomerase EC numbers, also known as orphan enzymes (Lespinet & Labedan, 2005; Pouliot & Karp, 2007), a term given to EC numbers where no gene has been associated with these reactions and no sequence information is available in protein sequence repositories. Almost half of the isomerase EC numbers with sequence information (96) are present in FunTree and Figure 4.1a shows the distribution by EC 5 subclass.

Protein structural data are available for 126 isomerase EC numbers, which have at least



b

From \ To	1	2	3	4	5	6
1					11	
2					21	
3					9	
4					26	
5	9	12	9	21	26	0
6					1	

c

From \ To	1	2	3	4	5.1	5.2	5.3	5.4	5.5	5.99	6
1					1	1	9	0	0	0	
2					5	0	9	4	1	2	
3					1	0	1	2	0	5	
4					10	0	6	6	4	0	
5.1	2	0	0	9	9	0	2	0	0	0	0
5.2	1	1	1	0	0	0	0	0	0	0	0
5.3	3	6	3	4	0	0	4	0	0	0	0
5.4	2	2	3	6	0	0	0	8	0	0	0
5.5	1	0	0	2	1	0	0	0	2	0	0
5.99	0	3	2	0	0	0	0	0	0	0	0
6					1	0	0	0	0	0	

Figure 4.1: Distribution of isomerase data in sequence, structure, function and evolution resources. (a) Distribution of isomerases in EC classification, UniprotKB, PDB and FunTree. EC exchange matrices representing the changes in function during evolution of isomerases at the EC (b) class and (c) subclass levels. More frequent changes of isomerase function are highlighted in red. Green and blue boxes represent changes within isomerases and with other EC classes, respectively. (d) Frequency of EC changes involving isomerases by superfamily. The 32 superfamilies bearing multiple changes are illustrated.

one entry in the PDB (Berman *et al.*, 2013). The 96 isomerases currently present in FunTree include domains, which are distributed across 81 CATH superfamilies: 17 are mostly alpha, 5 mostly beta and 59 mixed alpha/beta. Some superfamilies include more isomerases than others, for example, the superfamily UDP-galactose 4-epimerase, domain 1 (CATH 3.90.25.10) includes 7 racemases and epimerases (EC 5.1). In FunTree, one-third of the 96 isomerases include more than one domain superfamily (multidomain), with most of them including two or three superfamilies, but rarely more. Exceptionally, the subclass "other isomerases" (EC 5.99), which has two EC numbers (EC 5.99.1.2 and 5.99.1.3) is distributed across seven and eight superfamilies, respectively. These are types I and II DNA topoisomerases, which are characterised by multiple domains required for the complex process of winding DNA (O'Brien, 2006).

4.2 Observed changes of isomerase function

4.2.1 Change in EC number

Analysis of FunTree data on 58 domain superfamilies identified a total of 145 unique changes of isomerase activity that occurred during evolution (for details on how changes of function are calculated see *2.3.2 Methods and tools* and Figure 2.8b). Only one-fifth of the changes occur between isomerases whereas the rest involve changing from isomerases to perform reactions in other EC primary classes (Figure 4.1b). This is strikingly different from enzymes in other EC classes where changes in lower levels of the EC classification are more common than changes in the primary classification (Furnham *et al.*, 2012*a*). Among the 26 changes within isomerases, only 3 change the EC subclass and 23 change the EC serial number, indicating a change in substrate (Figure 4.1c). A previous limited study of 24 pairs of enzymes reported that changes involving isomerases and lyases (EC 5 $\leftrightarrow$ EC 4) occur more often than changes to other EC classes (Bartlett *et al.*, 2003). Other analyses provided further evidence of these changes by revealing the structural insights of the evolution of an isomerase from a family of lyases, namely N-succinylamino acid racemase (EC 5.1.1.-) from o-succinylbenzoate synthases (EC 4.2.1.113) in the enolase superfamily (Glasner *et al.*, 2006*a*). This comprehensive analysis confirms that such changes are indeed prevalent, with 39% of the 119 changes in primary classification involving lyases.

Most domain superfamilies show multiple changes of reaction chemistry involving different EC classes (Figure 4.1d). The most adaptable superfamily domains are aldolase class I (CATH 3.20.20.70) and glutaredoxin (CATH 3.40.30.10), each of them exhibiting 10 changes of isomerase function. Whereas the glutaredoxin "isomerase" domain only exhibits changes of isomerase, oxidoreductase and transferase reactions, the aldolase class

I domain has also evolved to become a hydrolase and lyase (Figure 4.1d).

4.2.2 Correlation of sequence and function evolution

To gain an overview of the relationship between sequence and functional divergence, an overall representation of the sequence and functional similarity between the homologous enzymes that perform different catalytic reactions is presented in Figure 4.2. This illustrates that most sequences have diverged considerably, with sequence identities in the range lower than 40%. The three measures of functional similarity (Figure 4.2a-c) capture different properties of the change in function, but none of the plots show any linear relationship between sequence and functional divergence. In addition, the distributions for each of these measures look quite different. In Figure 4.2a, which assesses the overall bond changes, there are two clusters, one consists of changes exhibiting bond change conservation when the isomerase EC subclass is maintained, and in the second changes at the isomerase EC subclass or EC primary class do not exhibit bond change conservation. This partition is not observed in the comparisons by reaction centres and structures of substrates and products and in overall, the similarities tend to be more uniformly spread (Figure 4.2b,c). Remarkably, there are only a few changes in which enzymes retain a relatively high degree of sequence and functional similarity. For instance, the glycosyltransferase superfamily (CATH 1.50.10.20) exhibits a change of arabidiol synthase (EC 4.2.1.124) into thalianol synthase (EC 5.4.99.31) (circled in red in Figure 4.2a-c). This change involves different enzyme sequences from the terpenoid biosynthesis pathway of *Arabidopsis thaliana* that share high sequence identity (79%) and high reaction similarity (48% - bond change, 72% - reaction centre and 84% - structure similarity). They both act on (S)-2,3-epoxysqualene as the main substrate to synthesize a different product, which explains why the structure similarity is high.

Previous research presented how enzyme superfamilies evolve by a combination of chemistry-driven and substrate-binding-driven evolution (see *1.2 Evolution of enzyme function*). In an attempt to analyse the chemical diversity of the domain superfamilies performing changes of function in isomerases, the functional similarity space was divided into four quadrants as depicted in Figure 4.2d. Each point represents a superfamily whose changes of isomerase function were averaged according to overall chemistry - as measured by bond change similarity - and structures of the reactants - in line with the similarity of the structures of substrates and products. Half of the superfamilies shared average similarities of reactants higher than 50% (top two quadrants), whereas only about one-fourth exhibited average similarities of overall chemistry higher than 50% (right two quadrants). Particularly, there are only three instances where the overall chemistry is similar but the structures of the reactants significantly diverge (bottom right quadrant), highlighting

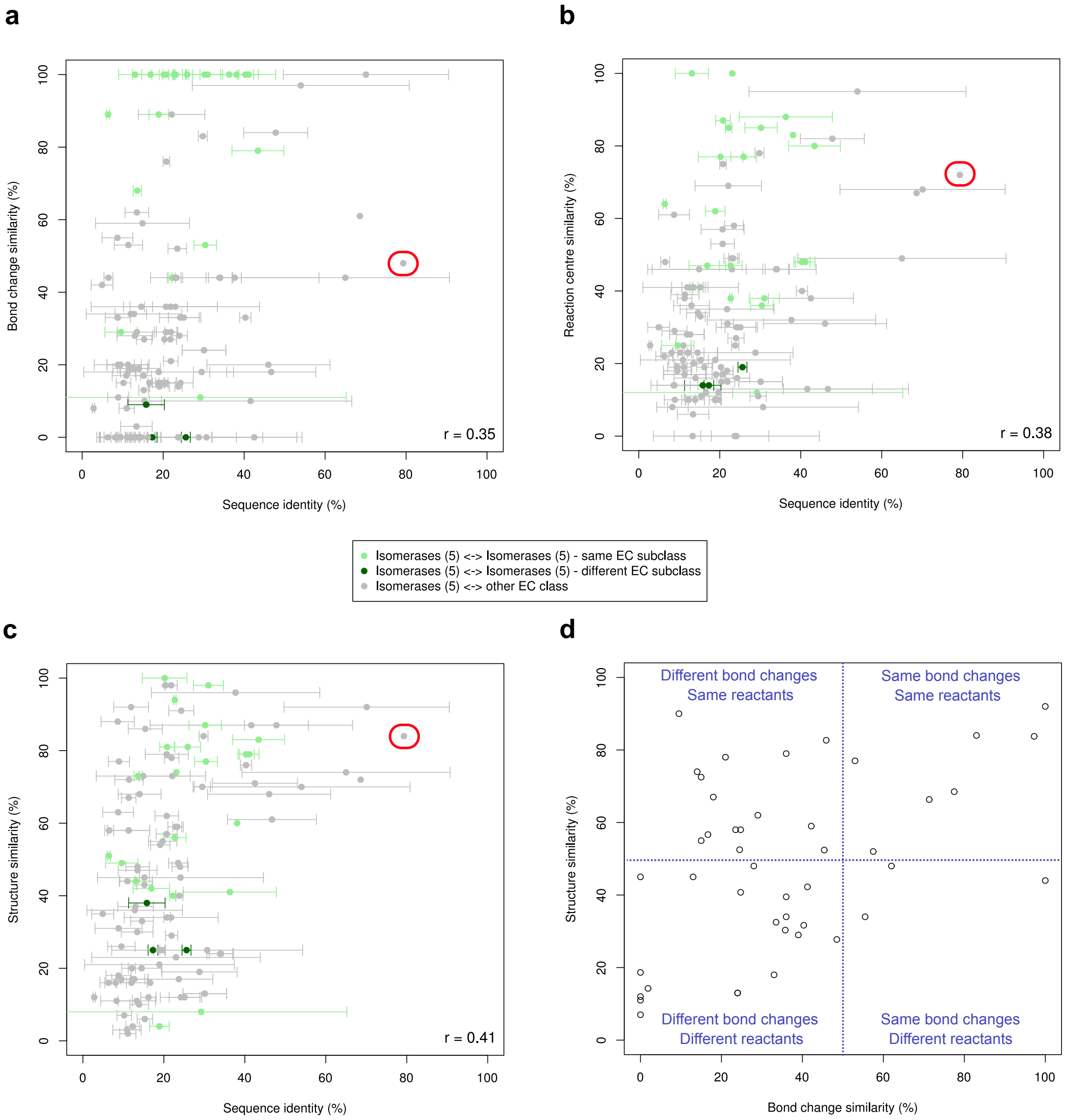


Figure 4.2: Sequence and functional similarity of the 145 changes of isomerase function. The three scatterplots represent global sequence identity against overall reaction similarity as calculated using three measures (a) bond change (b) reaction centre and (c) structure similarity of substrates and products. Each point represents one change of enzyme function involving two sets of enzymes catalysing two distinct functions each (Furnham *et al.*, 2012*b*). Average global sequence identities and standard deviations (error bars) are derived from all-against-all pairwise comparisons between sequences corresponding to one function and those corresponding to the second function. Encircled in red, the change EC 4.2.1.124 ->EC 5.4.99.31 (see main text). Pearson's correlation coefficients (r) range from 0.35 to 0.41 and indicate weak but significant linear relationships (p-value<0.001). (d) Distribution of bond change and structure similarities averaged by CATH superfamily.

that this is a rare event in the evolution of isomerase function.

4.3 An example - a family of SDRs acting on NDP-sugars from the UDP-galactose 4-epimerase superfamily

To explore one set of changes in more detail, eight changes of isomerase function involving a group of nine enzymes catalysing transformations between nucleoside diphosphate sugars (NDP-sugars) were studied. These metabolites are common in bacterial secondary metabolic pathways and they are necessary in molecular recognition and signalling processes (Singh *et al.*, 2012). Several studies have revealed the structural, functional and mechanistic determinants of this group of evolutionarily-related enzymes. They are epimerases (EC 5), dehydratases (EC 4), decarboxylases (EC 4) and oxidoreductases (EC 1) belonging to the subfamily of short-chain dehydrogenases/reductases (SDR) acting on NDP-sugars (Figure 4.3a) (Eixelsberger *et al.*, 2012; Frey & Hegeman, 2013; Hegeman *et al.*, 2002; Kowatz *et al.*, 2010). The changes in function involve two-domain enzymes comprising a catalytic NAD(P)-binding Rossmann-like domain (CATH 3.40.50.720) and a domain known as UDP-galactose 4-epimerase (CATH 3.90.25.10), which confers substrate specificity. The active site is located in the interdomain cavity where a conserved Tyr, Lys and Ser/Thr form a catalytic triad. Reactivity takes place on the C4, C5 and C6 atoms of the sugar substructure through a mechanism involving a transient oxidation intermediate mediated by NAD^+ (Tanner, 2008). The sequence data provide evidence that different catalytic amino acids are recruited to the active site in order to change the prevalent UDP-glucose 4-epimerase activity (EC 5.1.3.2) to other enzymatic activities. For instance, a base, Glu and an acid, Asp, are added to the catalytic triad in dTDP-glucose 4,6-dehydratase (EC 4.2.1.46) and GDP-mannose 4,6-dehydratase (EC 4.2.1.47) to perform the dehydration step which takes place in each of these overall reactions (Hegeman *et al.*, 2002). Since the reactivity takes place in the attached sugar moiety, the nucleoside diphosphate substructure (noted as X in Figure 4.3a) is not disrupted during catalysis and remains conserved in all enzymatic activities of this superfamily.

FunTree catalogues 8 changes of isomerase function within this family of enzymes (Figure 4.3a). They all share the same domain composition and therefore changes in function result directly from changes in sequence, rather than domain architecture. The analysis of sequence and functional similarities revealed that this family is divergent, with members sharing sequence identities in the 20% to 40% range. Bond change similarities revealed the already observed bimodal distribution due to the EC classification definitions (Figure 4.3b). Similarities by reaction centre remain low - not higher than 50% (Figure 4.3c)

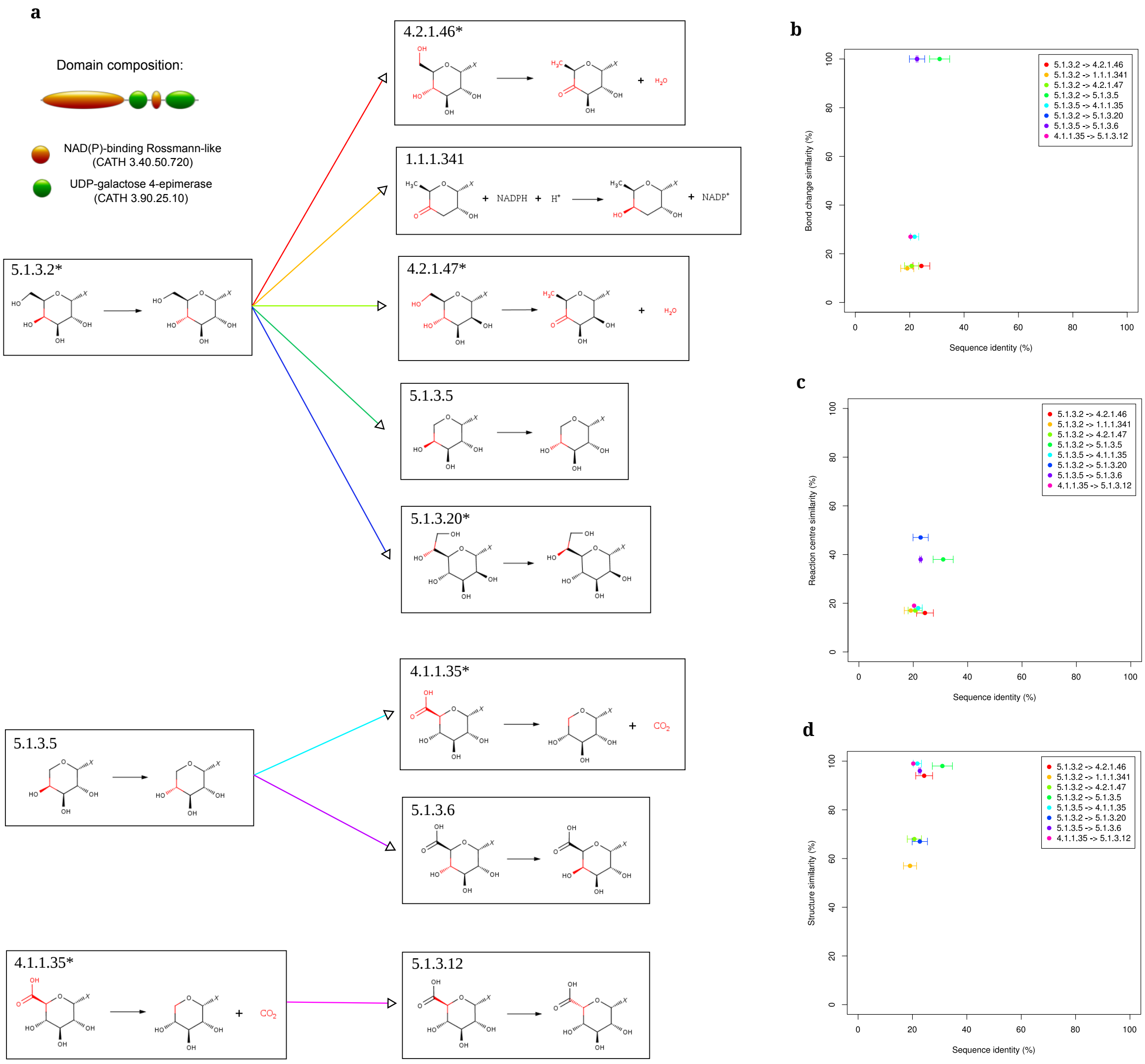


Figure 4.3: The evolution of SDRs acting on NDP-sugars. (a) Overview of the EC changes involving isomerases and domain composition of UDP-glucose 4-epimerases (EC 5.1.3.2). Biochemical reactions are represented in boxes. Black arrows inside boxes denote chemical transformations whereas coloured arrows linking boxes represent EC changes. EC numbers with an asterisk indicate reactions for which we found mechanistic evidence in MACiE (Holliday *et al.*, 2012) or in literature searches. Changing substructures are highlighted in red whereas X corresponds to nucleoside diphosphate moieties (ADP, TDP, GDP, CDP, UDP) in which the base may change, but the ribose diphosphate (or sometimes the 2'-deoxy derivatives) is broadly conserved. Three scatterplots illustrating sequence and functional similarity for this superfamily (b) bond change, (c) reaction centre and (d) structure similarity of substrates and products as in Figure 4.2.

whereas overall, this set of functional changes tend to conserve structural similarity, due to the common binding of a conserved nucleoside diphosphate (Figure 4.3d).

Taken together, this overview of sequence and functional relationships may help to identify possible sequences catalysing orphan isomerase reactions. For instance, comprehensive literature and database searches confirmed that the enzymatic activity UDP-glucosamine 4-epimerase (EC 5.1.3.16) is an orphan EC number. In 1959, it was first experimentally determined in rat liver by Maley (Maley & Maley, 1959). The high functional similarity to the activities UDP-glucose 4-epimerase (EC 5.1.3.2), UDP-arabinose 4-epimerase (EC 5.1.3.5) and UDP-glucuronate 4-epimerase (EC 5.1.3.6) suggests that the sequence catalysing EC 5.1.3.16 may belong to the UDP-galactose 4-epimerase superfamily. Ultimately, experimental analysis will reveal whether candidate sequences actually perform this reaction.

4.4 Conclusion

Using isomerases as an example, this chapter explored how enzyme chemistry may change over time, as enzymes evolve to perform different enzyme reactions.

The observation is that isomerases are more likely to evolve new functions in different EC primary classes, rather than evolve to perform different isomerase reactions. This is unlike the other EC classes where more than two-thirds of the exchanges happen within the same EC class. In addition we note that exchanges between isomerases and lyases (EC 4) are prevalent.

Isomerases change their overall chemistry and conserve the structure of their substrates more often than conserving the chemistry and changing substrates. This is also unlike other types of enzymes and reflects the mechanisms of isomerases, which can often incorporate mechanistic components from different classes to provide a different overall outcome while conserving the substrate binding abilities.

The chaotic nature of the sequence and function relationship in superfamilies including isomerases is evidenced by the lack of correlation between sequence and functional similarity. Variations in sequence are always very large revealing that changes happened long ago, emphasizing that evolutionary studies need to be undertaken on a superfamily basis. Here we gave an example of how combining knowledge from the chemistry and evolution of enzymes acting on nucleoside diphosphate sugars may help to characterise related orphan activities.

Chapter 5

Characterising Complex Biochemical Reaction Data

Most research studies on enzymes assume a one-to-one relationship between biochemical reaction and Enzyme Commission (EC) number, the widely accepted classification scheme used to characterise enzyme activity. However, this is an oversimplified description and almost one-third of all known EC numbers are linked to more than one biochemical reaction. Whereas enzyme databases often try to resolve this complexity by defining generic, alternative and partial reactions, EC numbers are often linked to different types of reactions. This complexity adds a new dimension to our understanding of enzyme function and is relevant for improving the classification of enzymes and to study the change of enzyme function during evolution.

5.1 Complexity in the description of enzyme function

Enzymes are life's catalysts that accelerate biochemical reactions up to the rates at which biological processes take place in living organisms. They play a central role in biology and have been thoroughly studied over the years. Since the 1960s, the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) has systematically encapsulated the functional information of enzymes into EC numbers. As discussed before, this is a hierarchical classification of enzymes based on multiple aspects of the overall chemistry of the biochemical reaction such as the chemical bonds that are broken or formed, cofactors being used and the nature of the substrates undergoing transformation (Tipton & Boyce, 2000). It is the global standard representation of molecular function for enzymes, and ultimately the genes, catalysing these transformations.

Whereas classical studies on enzyme catalysis highlighted the exquisite chemical specificity between enzyme and biochemical reaction (Fersht, 1999; Silverman, 2002), several studies have unearthed details of the additional ability of some enzymes to catalyse more than one biochemical reaction (Copley, 2003; O'Brien & Herschlag, 1999; Schulenburg & Miller, 2014). This phenomenon is widely known as enzyme promiscuity (Hult & Berglund, 2007); it has been described as being necessary for the evolution of enzyme function (Khersonsky & Tawfik, 2010); it has implications for biotechnology (Bornscheuer *et al.*, 2012; Nobeli *et al.*, 2009). Recently, various strategies to investigate the influence of chemistry in the evolution of promiscuous activities have been proposed (Khersonsky *et al.*, 2011) together with approaches exploring evolutionary trade-offs between native and promiscuous activities (Garcia-Seisdedos *et al.*, 2012). For such analyses, a transparent and accurate classification system to describe enzyme activity is particularly important.

The existing functional classification of enzymes has proved to be very powerful. It is manually curated and maintained by expert enzymologists, who use a controlled vocabulary and well-defined relationships in describing enzyme function (Friedberg, 2006) to convey the way biochemists think about reactions (McDonald & Tipton, 2014). It facilitates predefined comparisons between enzymes based on their chemistry and newly discovered enzymes are easily allocated in the different levels of its hierarchical classification. However, because of the diversity of chemical criteria used at different levels, the classification is not always coherent between EC classes (Kotera *et al.*, 2004; Latino & Aires-de Sousa, 2006; Sacher *et al.*, 2009). For instance, lyases (EC 4) are divided in subclasses depending on the type of chemical bond that is broken in the reaction whereas isomerases (EC 5) are divided based on the type of isomerisation. In addition, the EC classification is based on the overall catalysed reaction, which means that mechanistic steps and reaction intermediates are not considered. As a result, all enzymes carrying out the same overall reaction are generally assigned to the same EC number, even when they perform catalysis using different cofactors and mechanisms (O'Boyle *et al.*, 2007). For example, chloride peroxidase (EC 1.11.1.10) is used to describe three structurally distinct non-homologous enzymes, which change in their cofactor dependence in three different catalytic mechanisms and are deemed to have emerged from independent evolutionary origins (Holliday *et al.*, 2011; Omelchenko *et al.*, 2010). There are however exceptions to this rule, particularly in oxidoreductases (EC 1), where enzymes catalysing the same overall reaction using different cofactors are sometimes assigned different EC numbers. For instance, EC 1.1.1.32 and 1.1.1.33 represent two mevaldate reductases, both catalyse the conversion of (R)-mevalonate to mevaldate but respectively use NAD^+ and $NADP^+$ as a cofactor (Shearer *et al.*, 2014). On the other hand, homoserine dehydrogenase (EC 1.1.1.3) has broader cofactor specificity and interconverts L-homoserine and

L-aspartate 4-semialdehyde using both NAD$^+$ and NADP$^+$ but with a slight preference for the first (Jacques *et al.*, 2001).

Although reliable and rigorous, the manual process of naming and classifying each new enzyme is laborious and requires expert knowledge, therefore automatic approaches may help to accelerate this procedure. Similarly, the NC-IUBMB has also considered the current EC classification system to be a relic of the original attempts to develop a chemically sensible hierarchical classification. Ideas and methodologies envisioning a new system in which enzymes are assigned meaningless database identifiers have already been proposed (Tipton & Boyce, 2000) and automatic tools to search and compare enzyme reactions are useful to navigate through enzyme space and may help to improve future versions of the classification (Rahman *et al.*, 2014).

There are also other limitations regarding the ability of the EC classification to accurately represent enzyme function (Babbitt, 2003). In fact, it had to be adapted after discovering that homologous enzymes annotated with the same EC number can manifest different levels of substrate specificity (Cornish-Bowden, 2014) (also known as substrate promiscuity or ambiguity). For instance, UDP-glucose 4-epimerases (EC 5.1.3.2) display different substrate specificities depending on the taxonomic lineage. Due to differences in the amino acid composition of the enzyme active sites, the bacterial epimerases only act upon UDP-glucose whereas the eukaryotic relatives additionally catalyse the transformation of UDP-N-acetylglucosamine (Daenzer *et al.*, 2012). Similar trends of variation in substrate specificity are common for other isomerases (EC 5) (Gall *et al.*, 2014) and enzymes from other EC classes (McDonald & Tipton, 2014). Even though this limitation has partially been addressed by introducing specificity information in the “Comments” section of several EC entries (Kotera *et al.*, 2008) together with reaction data and structured tables in resources such as KEGG (Kanehisa *et al.*, 2012) and BRENDA (Schomburg *et al.*, 2013*a*), there is still a need to represent this phenomenon in a more computer-friendly format in order to obtain accurate comparisons between EC numbers.

Although enzyme function depends on the sequence and structure of the enzymes performing catalysis, the EC classification does not consider this information during the assignment process (Cornish-Bowden, 2014). However, scientific interest in groups of enzyme functions that are poorly classified by the EC (e.g. activities on polymeric biomolecules like sugars, proteins or DNA) led to the development of alternative schemes that consider homology as organisational principle. For instance, proteolytic enzymes are hydrolases (EC 3) and lyases (EC 4) exhibiting broad substrate specificity, classified using sequence and structure analyses in the MEROPS database (Rawlings *et al.*, 2014). Likewise carbohydrate-active enzymes are classified in the CAZy resource using similar principles

(Lombard *et al.*, 2014).

Currently, the EC classification is effectively used as the link between enzyme information (genes, sequence and structure) and chemistry data (biochemical reactions) in common resources like UniprotKB (The Uniprot Consortium, 2013). However there is already evidence suggesting that the correspondence between enzymes, EC numbers and biochemical reactions is not as simple as previously thought (Babbitt, 2003; Egelhofer *et al.*, 2010). The relationship between enzyme and EC number is complex and rarely one-to-one (Holliday *et al.*, 2011). Some enzymes are annotated with multiple EC numbers (multifunctional) (McDonald & Tipton, 2014) whereas some EC numbers are associated with many unrelated enzymes (Omelchenko *et al.*, 2010). Early studies defined "average" sequence similarity thresholds for the accurate transfer of EC numbers to enzymes using homology (Rost, 2002; Tian & Skolnick, 2003; Todd *et al.*, 2001) and subsequent efforts integrated other sources of information such as structural data (Laskowski *et al.*, 2005*b*) and catalytic residue conservation (George *et al.*, 2005) to help understanding this relationship. Although there are several studies that have deliberatively excluded multifunctional enzymes in order to avoid these complexities (des Jardins *et al.*, 1997; Todd *et al.*, 2001), some approaches aiming to predict biochemical reactions in metabolites have successfully handled reactions associated with more than one EC number (Mu *et al.*, 2011). To some extent, KEGG circumvents the need for using EC numbers to link enzymes and biochemical reactions by directly connecting reactions to groups of orthologous enzymatic genes (Kotera *et al.*, 2014). This association might considerably simplify the process of linking chemical and genomic information.

Driven by the observation of striking differences in the way biochemical reactions are represented using the EC classification in several databases (Altman *et al.*, 2013), together with evidence from various studies excluding biochemical reactions associated with more than one EC number (Latino & Aires-de Sousa, 2009; Latino *et al.*, 2008), here the relationship between EC number and biochemical reaction is explored. Although some reviews commented on certain aspects of EC number diversity (Bernard *et al.*, 2014; Sorokina *et al.*, 2014), to the author's best knowledge, studies addressing this connection in a systematic manner are lacking. This relationship was first explored for a chemically diverse class of enzymes catalysing geometrical and structural rearrangements between isomers, the isomerases, and then this knowledge was used to develop an automatic approach to gain an overview of reaction diversity across the EC classification.

5.2 Characterising complexity

5.2.1 Overview

There are 5385 four-digit EC numbers in the 9th April 2014 release of the NC-IUBMB list, 4237 of them (79%) are associated with 6494 unique biochemical reactions bearing structural information in the 70.0+ release of KEGG database (Kanehisa *et al.*, 2012), accessed using the KEGG website and Advanced Programming Interface (API) (Kawashima *et al.*, 2003). The remaining 21% lack structural data. Although most EC numbers are linked to one biochemical reaction, almost a third are associated with more than one (Figure 5.1a). Comparatively, oxidoreductases (EC 1) exhibit the highest fraction of multiple reactions whereas isomerases (EC 5) the lowest (Figure 5.1b). Similarly, some unusual cases were identified where individual EC numbers are linked to over 20 biochemical reactions, with one extreme outlier, classified as an unspecific monooxygenase (EC 1.14.14.1) with up to 66 reactions (Figure 5.1c). In isomerases, the total number of EC numbers in the database is 245, for which 222 are associated with 298 biochemical reactions and 23 are not linked to any reaction. Among the EC numbers linked to isomerase reactions, 42 are associated with more than one reaction.

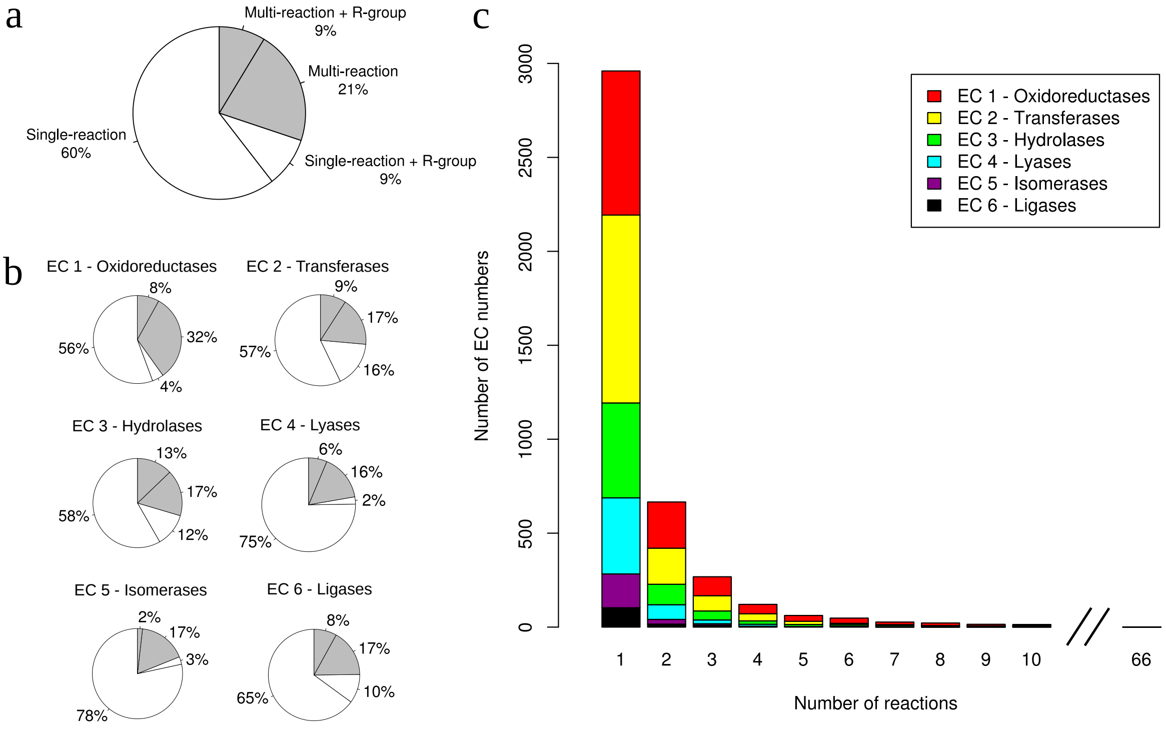


Figure 5.1: Survey of EC numbers associated with more than one biochemical reaction. (a) Overall distribution. White and gray slices indicate single and multi-reaction EC numbers, respectively. "R-group" represents EC numbers containing a Markush label in at least one reaction (see *generic* reactions in main text) (b) Distribution by EC class (c) Distribution of EC numbers according to the number of reactions.

5.2.2 Relationship between EC number and biochemical reaction in isomerases

In general, the intrinsic diversity in isomerase multi-reaction EC numbers was interpreted in terms of the chemical variability between the reactions linked to the same EC number. In the context of catalytic promiscuity, previous studies defined reactions to be different if they differ in the types of bond changes (formed and cleaved), the reaction mechanism or both (Kaltenbach & Tokuriki, 2014; Kazlauskas, 2005). The reactions associated with the 42 multi-reaction isomerase EC numbers were manually analysed on the basis of bond and stereochemistry changes and EC numbers were divided into three groups according to *same*, *partial* and *different overall* chemistry of the reaction (Figure 5.2). According to our observations, the first group was then further divided into two subgroups: *different* reactants and *generic* reaction. Since the EC number only describes the *overall* reaction, we do not include mechanisms in this analysis. Below is an explanation of each subgroup.

In the *different* reactants subgroup, reaction diversity arises due to the presence of different chemical substituents on a common structural scaffold. For example, the so-called “arginine racemase” (EC 5.1.1.9) describes the racemisation of arginine, lysine and ornithine. The three reactions involve a chiral inversion of the common Cα in the amino acid (Figure 5.2a).

Generic reactions are used to represent multiple reactions by means of the chemical composition of their reactants. They are represented using Markush labels (e.g. R-groups) (Brecher, 2008), which serve as chemical wildcards for other reactions. Almost one in five EC numbers are associated to at least one *generic* reaction, half of them refer to multi-reaction EC numbers and the other half represent single-reaction EC numbers (Figure 5.1a). Although the association between Markush labels from the *generic* reaction and the corresponding chemical substructures in exemplar reactions is direct for multi-reaction EC numbers, this correspondence in single-reaction EC numbers is challenging where comparisons with all the other EC numbers are required.

Multi-reaction EC numbers where at least one reaction is *generic* are the subject of this study. *Generic* relationships according to chemical composition are of two types. First, some cases resemble the characteristics of the *different* reactants subgroup but the various chemical substituents are collectively displayed in an additional *generic* reaction, which represents the rest of reactions. For instance, amino acid racemase (EC 5.1.1.10) is linked to five reactions. Four of them describe racemisations of glutamine, serine, ornithine and cysteine and the extra one represents all of them by encapsulating the diversity of the amino acid side chain into a R-group (Figure 5.2b). In some cases however, the *generic*

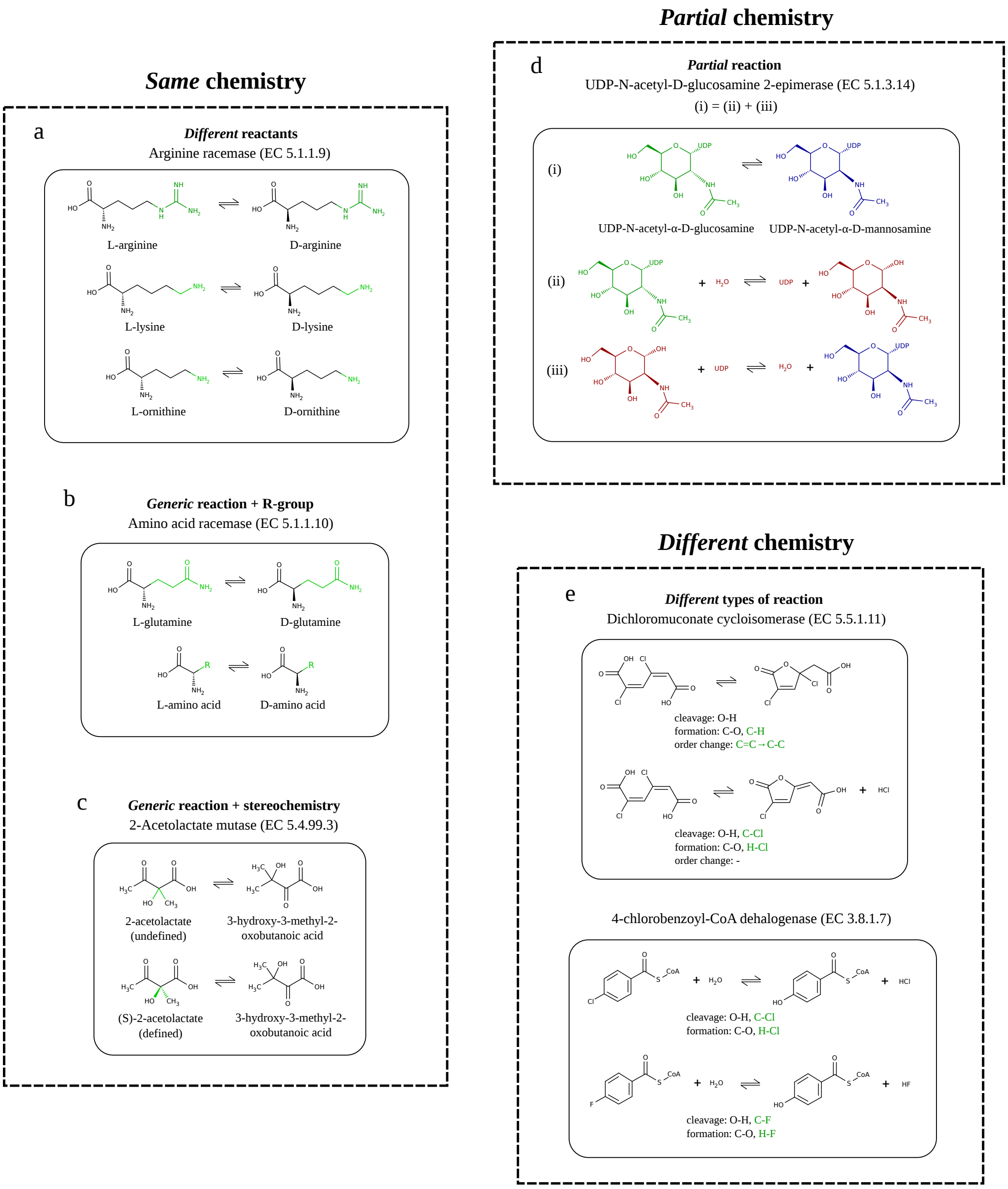

Same chemistry
a
Different reactants
Arginine racemase (EC 5.1.1.9)
L-arginine
D-arginine
L-lysine
D-lysine
L-ornithine
D-ornithine
b
Generic reaction + R-group
Amino acid racemase (EC 5.1.1.10)
L-glutamine
D-glutamine
L-amino acid
D-amino acid
c
Generic reaction + stereochemistry
2-Acetolactate mutase (EC 5.4.99.3)
2-acetolactate (undefined)
3-hydroxy-3-methyl-2-oxobutanoic acid
(S)-2-acetolactate (defined)
3-hydroxy-3-methyl-2-oxobutanoic acid
Partial chemistry
d
Partial reaction
UDP-N-acetyl-D-glucosamine 2-epimerase (EC 5.1.3.14)
(i) = (ii) + (iii)
(i)
UDP-N-acetyl-α-D-glucosamine
UDP-N-acetyl-α-D-mannosamine
(ii)
(iii)
Different chemistry
e
Different types of reaction
Dichloromuconate cycloisomerase (EC 5.5.1.11)
cleavage: O-H
formation: C-O, C-H
order change: C=C→C-C
cleavage: O-H, C-Cl
formation: C-O, H-Cl
order change: -
4-chlorobenzoyl-CoA dehalogenase (EC 3.8.1.7)
cleavage: O-H, C-Cl
formation: C-O, H-Cl
cleavage: O-H, C-F
formation: C-O, H-F

Figure 5.2: Examples of isomerase EC numbers associated with more than one biochemical reaction. (a) Arginine racemase (EC 5.1.1.9) is an isomerase acting on *different* reactants. The variability in chemical substituents is highlighted in green and the common scaffold in black. (b) Amino acid racemase (EC 5.1.1.10) is an example of *generic* reaction on the basis of R-group. Same colouring as in (a). (c) 2-acetolactate mutase (EC 5.4.99.3) is an example of *generic* reaction based on stereochemistry. The stereochemistry of C2 in acetolactate is represented as straight (undefined), up and down (defined) bonds and highlighted in green. (d) UDP-N-acetyl-D-glucosamine 2-epimerase (EC 5.1.3.14) belongs to *partial* reaction, (i) *overall* reaction - epimerisation of UDP-N-acetyl-α-D-glucosamine (green) and UDP-N-acetyl-α-D-mannosamine (blue), (ii) first *partial* reaction - hydrolysis and epimerisation of UDP-N-acetyl-α-D-glucosamine and (iii) second *partial* reaction - addition of UDP to N-acetyl-α-D-mannosamine. Intermediate compounds are highlighted in red. (e) Dichloromuconate cycloisomerase (EC 5.5.1.11) and 4-chlorobenzoyl-CoA dehalogenase (EC 3.8.1.7) catalyse different types of reactions. Shared bond changes are coloured in black, whereas different bond changes in green.

reaction is the common structural scaffold shared among all reactions. As a result, there is no R-group involved, and the reactants of the *generic* reaction are substructures of the reactants of the rest of reactions. For example, in Figure 5.2a the reactants in the epimerisation of L-ornithine are substructures of the reactants in the epimerisation of L-arginine, hence the former could also be a *generic* reaction of the latter. Although the latter *generic* relationship is evident in our manual analysis, in the process of developing an automatic method to assign EC numbers to reaction diversity groups (see *5.2.3 Automatic analysis*) this was considered as an example of *different* reactants. Other isomerase EC numbers fall into this category such as chalcone isomerase (EC 5.5.1.6), which catalyses reversible cyclisation of chalcone into flavanone as common structural scaffold. In addition, it also performs the same reaction in hydroxy-substituted derivatives of chalcone and flavanone (Kimura *et al.*, 2001).

The second case of representation by *generic* reaction arises due to differences in the definition of stereochemistry between the *generic* reaction and rest of the reactions. Here, undefined stereochemistry (in the form of wiggly or non-stereo bond) characterises one of the chiral carbons in the *generic* reaction, whereas stereochemistry is defined for that atom in the rest of the reactions. Although a previous study reported data challenges due to the lack of stereochemical completeness in KEGG metabolites and reactions (Ott & Vriend, 2006), to some extent recent versions of the database have incorporated these recommendations to improve the handling of stereochemistry and related data inconsistencies. Taken together, the common existence of cases of defined and undefined stereochemistry in several EC numbers supported the formulation of this diversity group. For example, acetolactate mutase (EC 5.4.99.3) is associated with two reactions: the isomerisations of 2-acetolactate (generic reaction, undefined stereochemistry) and (S)-2-acetolactate (specific reaction, defined stereochemistry) (Figure 5.2c). As in *generic* reactions on the basis of R-group, cases of undefined stereochemistry in the form of wiggly bonds were detected in our automatic method, however the cases of non-stereo bonds were regarded as examples of *different* reactants.

It is a well known fact that there are enzymes releasing intermediate products of an *overall* reaction from the active site (McDonald & Tipton, 2014). Reactions leading to these intermediates are known as *partial* reactions. Similarly, an enzyme may subsequently catalyse two or more *partial* reactions with or without releasing any intermediates, these are considered as *consecutive* reactions. For example, in Figure 5.2d UDP-N-acetyl-D-glucosamine 2-epimerase (EC 5.1.3.14) catalyses the epimerisation of UDP-N-acetyl-α-D-glucosamine and UDP-N-acetyl-α-D-mannosamine (*overall* reaction). This transformation comprises two successive *partial* reactions in the mechanism - hence, they are *consecutive*. First, the UDP moiety is hydrolytically eliminated from the anomeric carbon and epimerisation takes place at C2 (first *partial* reaction). Second, the UDP moiety is added to the anomeric carbon (second *partial* reaction). Combining these two *consecutive* reactions leads to the *overall* reaction. Whereas this example summarises this group in its simplest form, we also found three other alternatives of *partial* reactions linked to the same EC number:

- Two *consecutive* reactions and the *overall* reaction are all linked to the same EC number.
- Two *consecutive* reactions are linked to the same EC number, the *overall* reaction is linked to a different EC number.
- One *consecutive* reaction and the *overall* reaction are linked to the same EC number, the other *consecutive* reaction might be assigned to a different EC number or correspond to an uncatalysed reaction.

Previous studies have alternatively used the concept of "multi-step reaction" to refer to our definition of *overall* reaction composed of more than one *partial* reactions that occur consecutively (Kotera *et al.*, 2004). However, the term step in a reaction usually implies one mechanistic step of the *overall* reaction. As mechanisms are not included in the EC classification, we preferred using the term *partial* reaction in order to avoid confusion.

Finally, EC numbers might also be linked to at least two *different* types of reactions. Dichloromuconate cycloisomerase (EC 5.5.1.11) catalyses two types: first, the isomerisation of 2,4-dichloro-cis,cis-muconate and 2,4-dichloro-2,5-dihydro-5-oxofuran-2-acetate and also, the conversion of 2,4-dichloro-cis,cis-muconate into trans-2-chlorodienelactone and chloride (Figure 5.2e) (Kuhm *et al.*, 1990; Pieper & Stadler-Fritzsche, 1993). Although the two reactions share the cleavage of O–H and formation of C–O bonds, they differ in other bond changes, so they are considered to be *different*. However the product of the first isomerisation might eliminate chloride to yield trans-2-chlorodienelactone in an uncatalysed manner and therefore the second reaction would be the result of an isomerisation and successive elimination, which can also be interpreted as an example of *partial* reaction as described before. Other examples of EC numbers that can also be categorised under both *different* types of reaction and *partial* reaction involve sugar isomerisations such as those catalysed by D-arabinose isomerase (EC 5.3.1.3) and ribose-5-phosphate isomerase (EC 5.3.1.6) where the ring opening and closure might be uncatalysed. Perhaps a more definite example of *different* reaction types is 4-chlorobenzoyl-CoA dehalogenase (EC 3.8.1.7). This EC number involves the dehalogenation of 4-chlorobenzoyl-CoA into 4-hydroxybenzoyl-CoA and also the hydrolysis of the fluoro, bromo and iodo derivatives (Figure 5.2e).

Following our manual classification, 30 of the 42 multi-reaction isomerase EC numbers were solely assigned to one of the groups, whereas the diversity of the remaining 12 EC numbers was explained by more than one group. Overall, 57 group assignments were manually designated: 24 *different* reactants, 17 *generic* reactions (R-group and stereochemistry), 5 *partial* reactions and 11 *different* types of reactions. Among the EC numbers assigned to more than one group, we found 2-acetolactate mutase (EC 5.4.99.3) (Figure 5.2c). In addition to the transfer of a methyl group from C2 to C3 in (S)-2-acetolactate, this isomerase also catalyses the transfer of an ethyl group from C2 to C3 in (S)-2-aceto-2-hydroxybutanoate. This EC number could be assigned to both groups: *generic* reaction on the basis of stereochemistry and *different* reactants (Figure 5.3). Similarly, although dichloromuconate cycloisomerase (EC 5.5.1.11) is an example of *different* types of reactions (Figure 5.2e), a potentially uncatalysed elimination of chloride may also link these two reactions in a *partial* relationship.

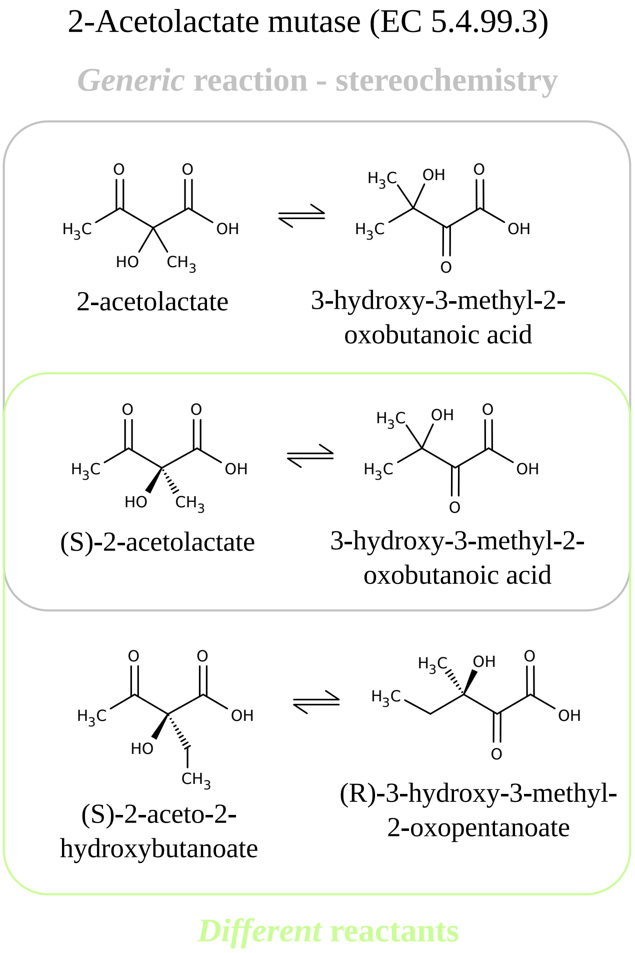

2-Acetolactate mutase (EC 5.4.99.3)
Generic reaction - stereochemistry
2-acetolactate
3-hydroxy-3-methyl-2-oxobutanoic acid
(S)-2-acetolactate
3-hydroxy-3-methyl-2-oxobutanoic acid
(S)-2-aceto-2-hydroxybutanoate
(R)-3-hydroxy-3-methyl-2-oxopentanoate
Different reactants

Figure 5.3: 2-Acetolactate mutase (EC 5.4.99.3) is an example of EC number assigned to two groups of reaction diversity: *different* types of reaction and *partial* reactions.

5.2.3 Automatic analysis - extending diversity groups found in isomerases to the EC classification

The automatic extraction of chemical attributes from biochemical reactions such as bond changes and reaction centres is necessary to compare enzymes based on the chemistry of their catalysed reactions. In order to calculate chemical attributes EC-BLAST was used, a recently-developed algorithm to obtain accurate atom-atom mapping, extract bond changes and reaction centres and perform similarity searches between enzyme reactions (Rahman *et al.*, 2014).

The strategy[1] comprised a set of conditional statements combining bond change results from EC-BLAST, which allowed the detection of *different* types of reaction; comparisons of substrate and product structures and identification of R-groups and stereochemistry using Open Babel (O'Boyle *et al.*, 2011) and in-house scripts, which helped to find *generic* and *partial* reactions (Figure 5.4). Cases of different reactants comprised the remaining multi-reaction EC numbers and were not detected by the conditions addressing the other diversity groups.

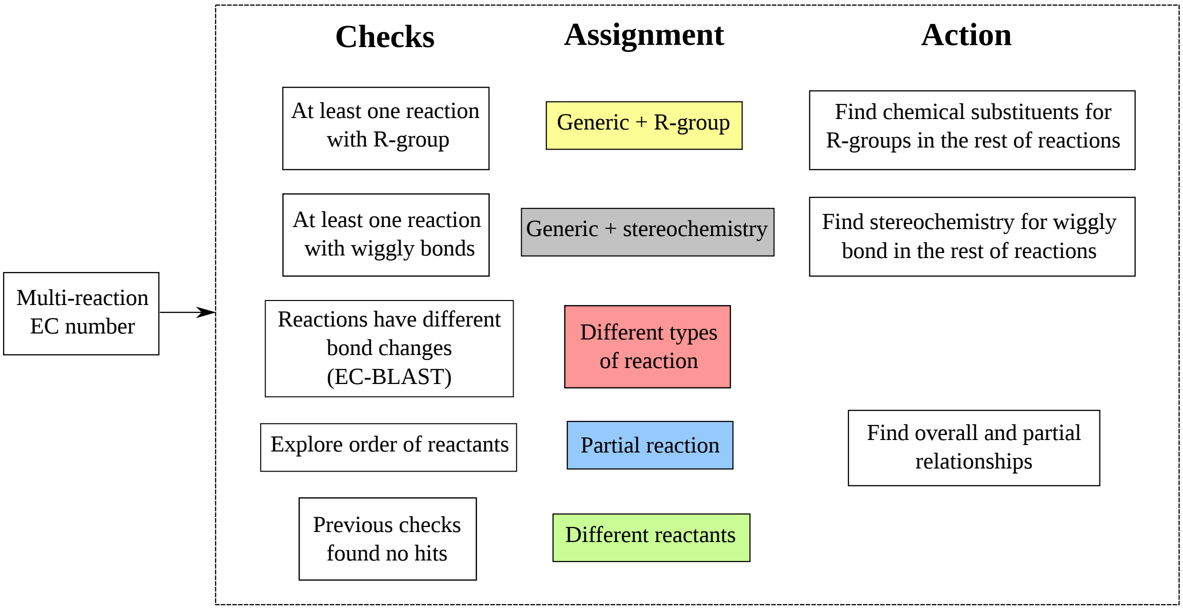

Checks
Assignment
Action
Multi-reaction EC number
At least one reaction with R-group
Generic + R-group
Find chemical substituents for R-groups in the rest of reactions
At least one reaction with wiggly bonds
Generic + stereochemistry
Find stereochemistry for wiggly bond in the rest of reactions
Reactions have different bond changes (EC-BLAST)
Different types of reaction
Explore order of reactants
Partial reaction
Find overall and partial relationships
Previous checks found no hits
Different reactants

Figure 5.4: Workflow illustrating the automatic analysis of multi-reaction EC numbers.

[1] The study of reaction diversity across the EC classification was done in collaboration with Handan Melike Dönertaş, a visiting student from the Department of Biological Sciences, Middle East Technical University, Ankara (Turkey). During her three months stay, Melike developed a method based on the 42 multi-reaction isomerase EC numbers to automatically label the type of diversity in any multi-reaction EC number (*different* reactants, *generic* reaction on the basis of R-group and stereochemistry, *partial* reaction and *different* types of reactions). My contribution to her project involved supervision.

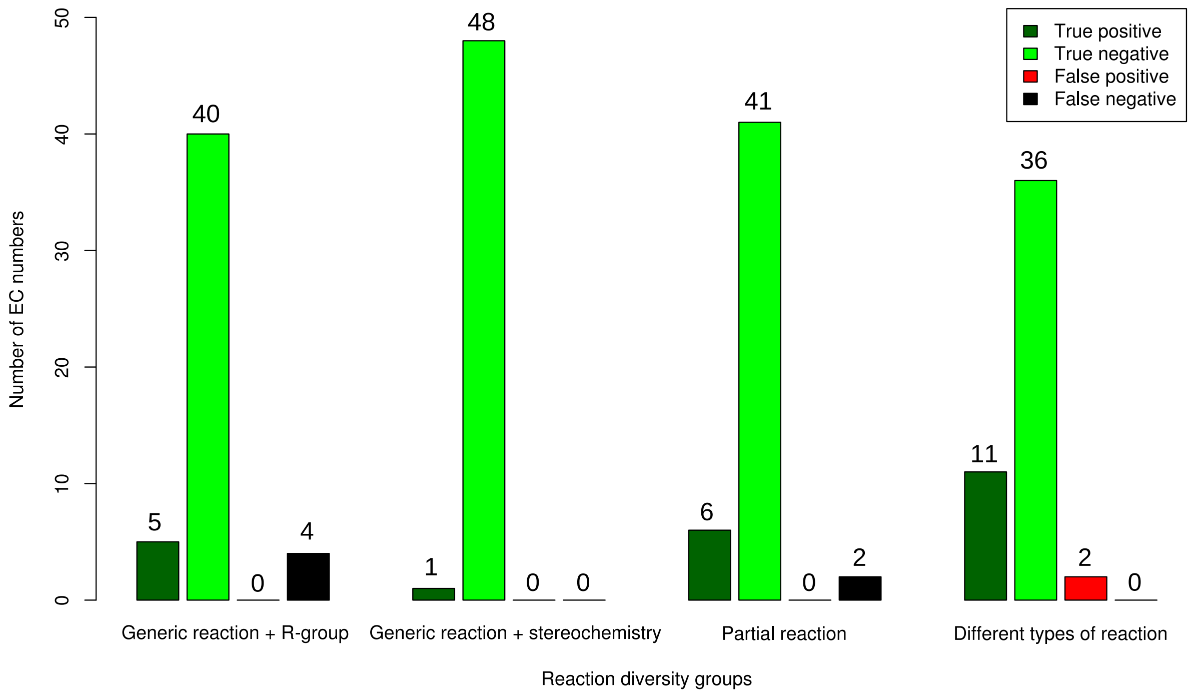


Figure 5.5: Results of the test to evaluate the automatic method labelling multi-reaction EC numbers according to the reaction diversity group.

The performance of the method was tested by assessing its ability to correctly identify the type of diversity in fifty randomly-selected multi-reaction EC numbers from the whole of the EC classification. The test dataset comprised 22 oxidoreductases (EC 1), 19 transferases (EC 2), 5 hydrolases (EC 3), 2 lyases (EC 4) and 2 ligases (EC 6), which were manually assigned to a reaction diversity group allowing performance to be evaluated. The selection of test multi-reaction EC numbers was carried out randomly, but it was assured that it covers the whole diversity space of the EC classification. True positives, true negatives, false positives and false negatives were calculated for each EC number in the test set and group of reaction diversity using manually-defined annotations as reference (Figure 5.5). Overall, the method successfully assigned the correct diversity group in 41 of the total of 50 test EC numbers. Almost all the assignments were true positives or true negatives, however nine remaining cases could not be correctly assigned due to data errors, detection problems and atom-atom mapping accuracy.

- EC 1.1.1.222 was assigned to *different* reaction types due to an atom-atom mapping error in R03337 and R03339. This issue was addressed in the present development version of EC-BLAST.

- EC 3.1.3.5 was assigned to *different* reaction types due to a data error, namely the different protonation state of the phosphate moiety in R02323 compared to the rest of reactions, leads EC-BLAST to detect a nonexistent formation of a O–H bond.

- EC 1.2.1.4 was not assigned to *generic* reaction + R-group due to a data error concerning protonation states. Whereas R00634 presents a carboxylate group, R00711 and R05099 consider the corresponding chemical species as carboxylic acid.

- EC 2.4.1.295 was not assigned to *generic* reaction + R-group due to difficulties in identifying R-groups. This is handled in development versions of the automatic method.

- EC 3.5.1.2 was not assigned to *generic* reaction + R-group due to a data error concerning protonation states. Whereas R06134 presents a carboxylate group, R01579 and R00256 consider the corresponding chemical species as carboxylic acid.

- EC 2.4.1.178 was not assigned to *generic* reaction + R-group due to similar difficulties in identifying R-groups as EC 2.4.1.295.

- EC 1.1.1.42 was not assigned to *partial* reaction due to difficulties in detecting *partial* reactions. Here, R00268 and R01899 are *partial* reactions of the overall R00267 such that R00267 = R00268 + R01899.

- EC 6.3.4.10 was not assigned to *partial* reaction due to similar difficulties in detecting *partial* reactions as in EC 1.1.1.42. Here, R01074 and R05145 are *partial* reactions of the *overall* R04582 such that R04582 = R01074 + R05145, considering C04763 = C06249 and C04727 = C06250.

- EC 1.1.1.100 was not considered in the test due to a data error. R04534 is unbalanced because a proton is missing on the right-hand side of the reaction.

5.2.4 Relationship between EC number and biochemical reaction in the EC classification

A schematic diagram illustrating the various groups of reaction diversity is shown in Figure 5.6a. There are 1277 multi-reaction EC numbers in the entire EC classification, 90% of them (1153) could be analysed using our method. The most common group was *different* reactants including almost half of the examples. *Different* reaction types followed with 29% and ultimately *partial* and *generic* reactions made up the rest (Figure 5.6b). The overall distribution was similar in oxidoreductases (EC 1), transferases (EC 2) and hydrolases (EC 3), which were correspondingly the EC classes involving the highest number of multi-reaction EC numbers (Figure 5.6c) and not surprisingly, also the EC classes with the largest number of EC numbers in the EC classification (McDonald *et al.*, 2009). Exceptionally, the most common diversity group in ligases (EC 6) is *different* reaction types, instead of *different* reactants. Also, the method did not identify any example of EC numbers involving *generic* reactions in lyases (EC 4) and ligases (EC 6).

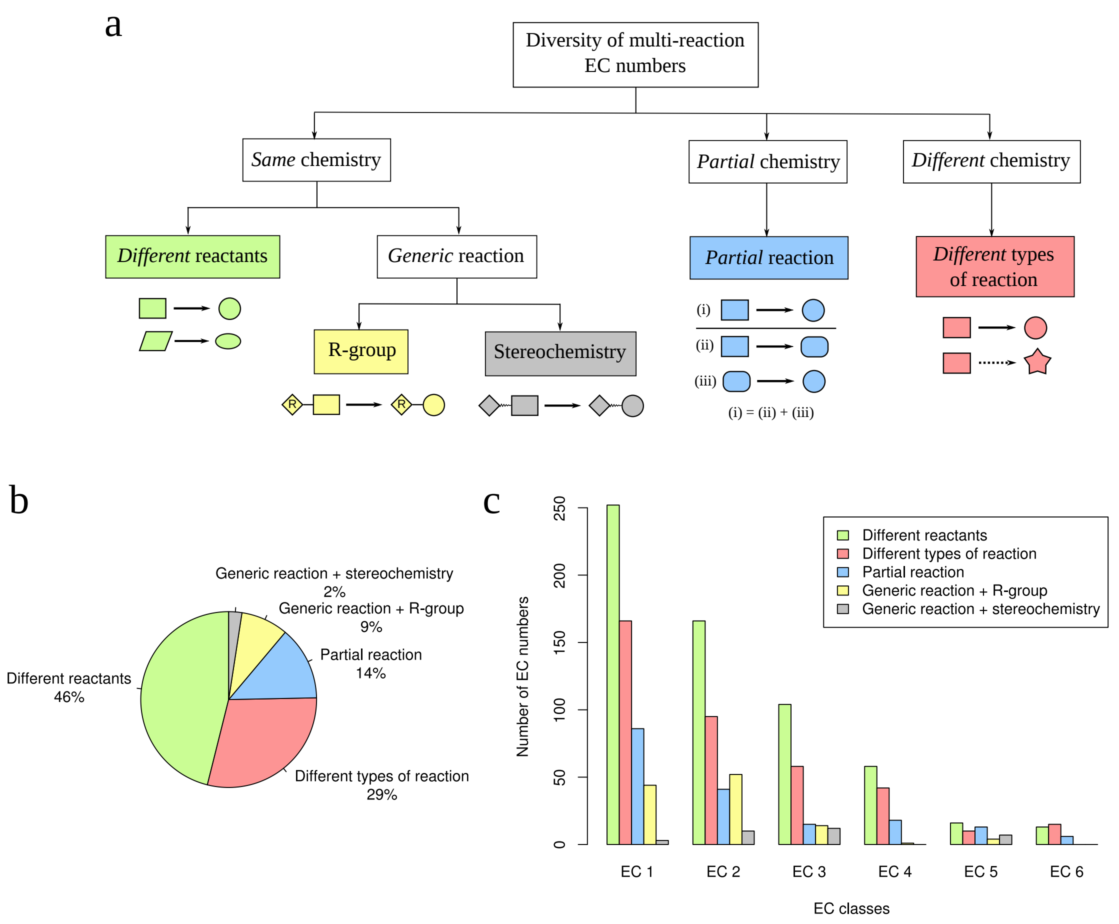


Figure 5.6: An overview of reaction diversity in the EC classification. (a) A schematic diagram summarising the groups of reaction diversity. (b) Frequency of reaction diversity group assignments. (c) Total number of multi-reaction EC numbers by EC class for each group of reaction diversity.

5.3 Discussion

5.3.1 Overall

Although there is literature reported by the NC-IUBMB discussing specific cases of reaction diversity across the EC classification (McDonald & Tipton, 2014), the aim of this study was to systematically explore aspects of the chemical diversity in the description of enzyme function in a specific EC primary class - manually and automatically in the entire EC classification. In order to extract bond changes from biochemical reactions the EC-BLAST algorithm was used, which is based on chemical concepts, such as the principle of minimum chemical distance and chemical bond energies, in order to guide the atom-atom mapping and chemical matrices for similarity searches (Rahman *et al.*, 2014). As suggested in a recent review by Chen *et al.* (2013), the incorporation of chemical knowledge adds accuracy to existing strategies to perform reaction comparison.

To what extent do the findings of this study overlap with those discovered in previous accounts on enzyme promiscuity? To some degree, the working definitions of substrate and product promiscuity (Hult & Berglund, 2007) somewhat resemble our diversity groups of *different* reactants and *generic* reactions. Likewise, catalytic promiscuity partly corresponds to *different* reaction types. However, whereas promiscuity definitions are genuinely attributed to enzymes in order to describe their ability to catalyse more than one biochemical reaction, our characterisation of reaction diversity applies to multi-reaction EC numbers, which adds an extra level of chemical variability to the existing definitions of enzyme function.

The surprising observation of this study is that almost one-third of the EC numbers involving more than one biochemical reaction have *different* reaction types, bearing key differences in catalysed bond changes. Whereas some of them also correspond to *partial* reactions, many are cases of catalytic promiscuity within the same EC number where the annotated enzyme catalyses two or more distinct reactions. Manual analysis revealed that most cases are similar to 4-chlorobenzoyl-CoA dehalogenase (EC 3.8.1.7) (Figure 5.2e) indicating that whereas some bond changes are shared, the rest individually characterise each of the different reactions.

The rationale behind why the NC-IUBMB and reaction databases have assigned multiple biochemical reactions to the same EC number is to some extent comprehensible. For instance, the product of some catalysed reactions sometimes undergoes a fast and uncatalysed reaction while still in the active site. These EC numbers comprise two reactions: one comprising only the catalysed reaction and another consisting of the catalysed+uncatalysed *consecutive* reactions. Whereas some enzymologists might preferably associate the EC number only with the catalysed reaction, the fact that the uncatalysed reaction takes place in the enzyme's confinement supports the catalysed+uncatalysed interpretation.

However the complexity in the relationship between biochemical reaction and EC number goes beyond this study and cases of *generic* relationships are also common in single-reaction EC numbers (Figure 5.1a) and across different EC numbers. For example, as highlighted before, EC 5.1.1.10 was defined by the NC-IUBMB after the discovery of an enzyme that broadly catalyses racemisations of several amino acids (Lim *et al.*, 1993). The biochemical reaction contains an R-group and it effectively represents reactions catalysed by specific amino acid racemases, which are also assigned different EC numbers, e.g. alanine (EC 5.1.1.1) and serine (EC 5.1.1.18). Although this and other examples (Kotera *et al.*, 2014) were attempts to incorporate an enzyme property such as substrate specificity to guide the EC classification, this might lead in some cases to EC numbers being no longer chemically independent from each other, which adds further complications to a classification based solely on the chemistry of the overall reaction.

5.3.2 Improving the EC classification

The ability of the NC-IUBMB to manually update the EC classification in the form of transferred and deleted entries when new enzyme data becomes available is necessary. For example, during the fifty years succeeding the creation of the EC entry for phosphoglycerate mutase in 1961 (EC 5.4.2.1), evidence supporting two distinct mechanisms concerning different usage of the cofactor 2,3-diphosphoglycerate by this enzyme accumulated in the literature (Foster *et al.*, 2010). In 2013, the original EC number was transferred to EC 5.4.2.11 (cofactor-dependent) and EC 5.4.2.12 (cofactor-independent). In addition, several expert recommendations concerning definition and handling of EC numbers in biological databases have already been suggested in different contexts. For example, Green and Karp advised about the problems associated with the assignment of partial EC numbers (those containing a dash, e.g. EC 5.1.1.-) to genes and proposed changes to the specification of these ambiguous identifiers (Green & Karp, 2005). Similarly, we suggest approaches to clarify multi-reaction EC numbers, which will hopefully help to improve the EC classification (McDonald & Tipton, 2014) and serve to guide standards for the reporting of enzyme data (Apweiler *et al.*, 2010; Gardossi *et al.*, 2010; Tipton *et al.*, 2014) and existing initiatives for the assignment of enzyme function (Anton *et al.*, 2013; Bastard *et al.*, 2014; Gerlt *et al.*, 2011).

A multi-reaction EC number belonging to the groups' *different* reactants or *generic* reactions could either be combined into a single-reaction EC number (**collective** approach) or split into as many distinct EC numbers (**specific** approach). In the first place, diversity could be represented by R-group definitions, which would encapsulate chemical substituents at different positions in the reactants. When necessary, stereochemically-undefined bonds could also be employed to indicate the non-stereoselectivity of some biochemical reactions (Figure 5.7a). Secondly, the **specific** strategy arises when there are significant changes of substrate specificity between enzymes annotated with the same multi-reaction EC number. Instead of defining a *generic* reaction, it might be more sensible to re-define several EC numbers according to the distinct patterns of substrate specificity (Schomburg *et al.*, 2014). However, although EC-BLAST provides a robust method to measure chemical differences between overall reactions in a continuous manner, defining the cut-offs required to designate separate EC numbers (for example, between different substrates) is a priori arbitrary and would need to be addressed explicitly.

A proposed *modus operandi* when dealing with *different* reaction types involves using the

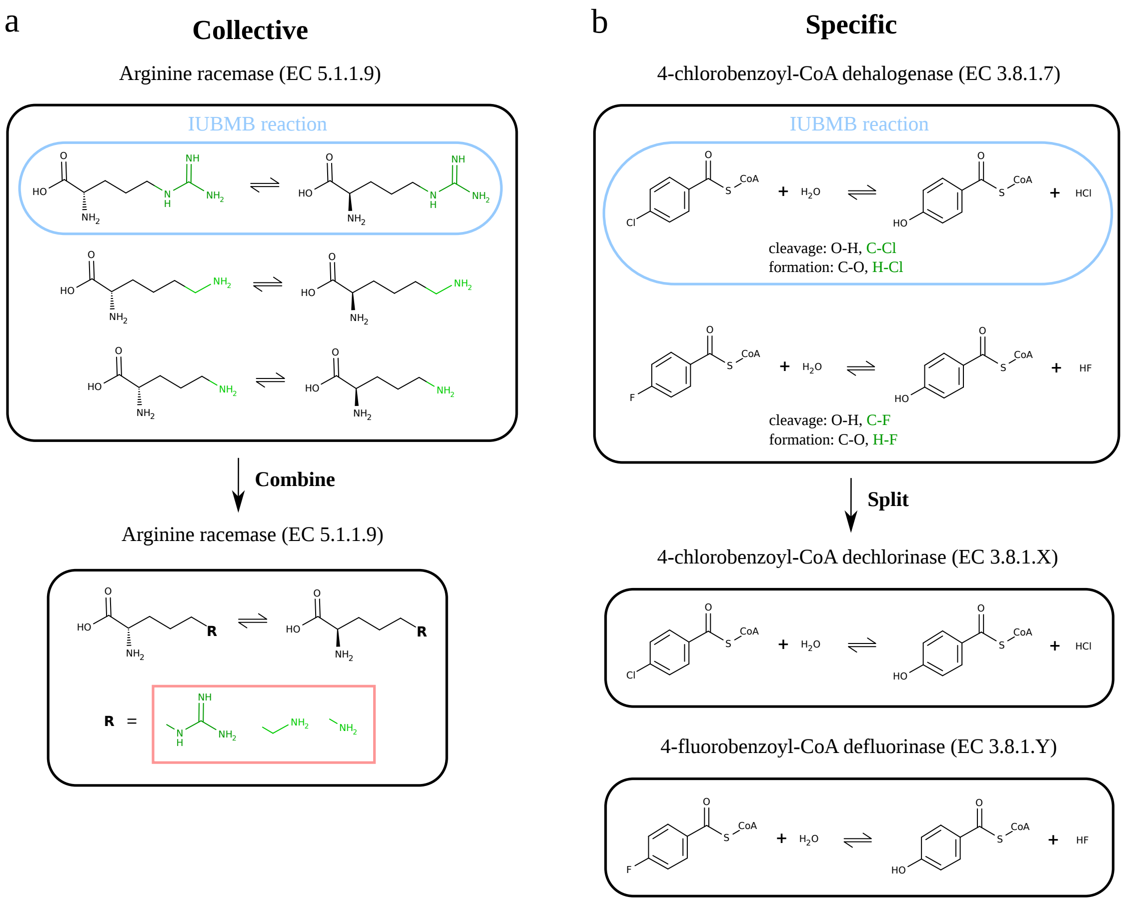

a
Collective
Arginine racemase (EC 5.1.1.9)
IUBMB reaction
Combine
Arginine racemase (EC 5.1.1.9)
R =
b
Specific
4-chlorobenzoyl-CoA dehalogenase (EC 3.8.1.7)
IUBMB reaction
cleavage: O-H, C-Cl
formation: C-O, H-Cl
cleavage: O-H, C-F
formation: C-O, H-F
Split
4-chlorobenzoyl-CoA dechlorinase (EC 3.8.1.X)
4-fluorobenzoyl-CoA defluorinase (EC 3.8.1.Y)

Figure 5.7: Examples of the **collective** and **specific** approaches. (a) The *different* reactants of arginine racemase (EC 5.1.1.9) are combined into a single-reaction EC number using R-group. (b) The two *different* types of reaction catalysed by 4-chlorobenzoyl-CoA dehalogenase (EC 3.8.1.7) are split and re-defined into two single-reaction EC numbers.

specific approach to divide the multi-reaction EC number into multiple EC numbers, one for each *different* reaction (Egelhofer *et al.*, 2010) (Figure 5.7b). Regarding *partial* reactions, we recommend to collectively reduce the multi-reaction EC number by combining all *partial* reactions with required enzyme catalysis into a single-reaction EC number, while setting uncatalysed reactions aside.

Both **collective** and **specific** approaches have several benefits. For instance, three main advantages characterise the **collective** approach. First, it is a compact way to arrange reaction information in a clear and structured manner. Second, it conveys how chemists and biochemists represent reactions in the literature, databases and patents (Geyer, 2013; Warr, 2011; Zass, 1990). Third, diversity can be captured using Markush labels such as R-groups (Brecher, 2008; Simmons, 1991), which would be subsequently described in associated files, tables or chemical libraries (Warr, 2014). Alternatively, diversity in the reactants could be encoded using recent developments in the description of chemical patterns (Schomburg *et al.*, 2013*b*). Also, the **collective** approach brings together reactions that are often evolutionarily-related. The precise definition of R-groups will also help previous studies that were limited in their ability to handle *generic* structures. Although some strategies did not explicitly define R-groups in their representation of biochemical reactions (Triviño & Pazos, 2010), several studies preprocessed oxidoreductase (EC 1) and hydrolase (EC 3) reactions by replacing every R-group by a hydrogen atom (Hu *et al.*, 2010; Sacher *et al.*, 2009) or methyl group (Mu *et al.*, 2006) in order to calculate physicochemical and topological properties in atoms and bonds involved in reaction centres. Using more specific substitutions, R-groups were manually replaced by methyl, adenine, cytosine or other chemical moieties depending on the type of biochemical reaction (Latino & Aires-de Sousa, 2009; Latino *et al.*, 2008). These studies suggest that having EC number-specific definitions of R-groups based on experimental evidence is a necessary step in order to implement the **collective** approach across the classification.

Whereas the **collective** approach relies on presenting a common structural scaffold and diversity encoded as chemical placeholders, the **specific** approach is divisive and explicitly distinguishes between reactions that are considered as chemically distinct. A clear advantage of the latter is when subtle differences between biochemical reactions are captured using different EC numbers, for instance, distinct bond changes. The description of enzyme function will then be more detailed and it will help to dissect some of the complexities in the relationship between enzyme sequence, structure and function (Holliday *et al.*, 2011). The terms of application of the **collective** and **specific** approaches to combine or split multi-reaction EC numbers are summarised in the following recommendations:

- Reactions sharing the *same overall* chemistry (identical bond changes) should be combined into a single-reaction EC number (corresponding to groups: *different* reactants and *generic* reaction). The chemical diversity observed as different embodiments of a *generic* structure would be encapsulated using R-group definitions and stereochemically-undefined bonds in associated libraries and chemical patterns.

- If reactions have *different overall* chemistry (distinct bond changes), the EC number should be split in multiple single-reaction EC numbers (group: *different* types of reaction). Similarly, reactions catalysed by enzymes annotated with the same EC number that display distinct substrate specificities or cofactor dependencies should also be split in as many single-reaction EC numbers as patterns of specificity exist (groups: *different* reactants and *generic* reaction).

- Reactions sharing *partial overall* chemistry (several *partial* reactions integrate into an *overall* reaction) should be treated carefully. The *partial* reactions that take place in the active site of the enzyme should be combined into a single-reaction EC number (group: *partial* reaction) with chemical diversity encapsulated in libraries as described before. Uncatalysed *partial* reactions should be considered separately.

This systematic analysis is relevant for the functional annotation of sequenced genomes and by extension, it has implications for our ability to build and compare genome-scale metabolic reconstructions (Monk *et al.*, 2014; Oberhardt *et al.*, 2011; Saha *et al.*, 2014). There is a direct correspondence between EC numbers and terms representing the molecular function of protein-coding genes in the Gene Ontology (GO) (Ashburner *et al.*, 2000), which implicitly adopted EC numbers as part of their classification. This ontology is currently the widely-used standard for the automatic assignment of function to proteins and genes (Radivojac *et al.*, 2013). We observed that multi-reaction EC numbers/GO terms are commonly transferred between similar enzymes during this process. Such a predicted assignment of function does not consider that enzymes annotated with the same multi-reaction EC number might have different reaction specificities in different species, which may lead to a general overestimation of the catalytic capabilities of organisms as predicted from their genomes.

5.4 Conclusion

This study adds an additional level of chemical complexity to our current description of enzyme function using EC numbers. Remarkably, almost a third of all known EC numbers are associated with more than one biochemical reaction. Existing approaches to characterise this diversity are ineffective, therefore this complexity was decomposed into four categories: *different* reactants, *generic*, *partial* and *different* types of reaction with the aid of computational methods to automatically compare biochemical reactions. Several recommendations to improve the characterisation were proposed, which will hopefully help to improve our understanding and description of biochemical reactions.

Chapter 6

Conclusions

The scope of this dissertation was to understand more about the chemistry and evolution of isomerase function (EC 5). Isomerases are life's catalytic toolkit to interconvert isomers, molecules with the same molecular formula but different chemical structures. As one of the six classes in the EC system, isomerases populate the genome and contribute to the metabolism of all living organisms. They also serve in various applications for biotechnology and chemical synthesis. Like most enzymes, the overall chemistry of isomerases is captured in the EC system. Unlike for other EC primary classes (Holliday *et al.*, 2014; Hu *et al.*, 2010; Latino *et al.*, 2008), an evaluation of the isomerase classification using a computational approach revealed the need for revision and improvement.

First, the EC system of isomerases needs simplification. The total of six subclasses can be summarised into two if isomerism is considered and the discovery of groups of similar overall reactions in complex sub-subclasses suggests that dissecting them into new sub-subclasses would help to resolve part of their complexity. Second, the prevalence of complex biochemical reaction data calls for more accurate description, especially multi-reaction EC numbers. A method to discover and classify complex biochemical reactions is currently under development. Reactions should be represented by the same EC number if the bond changes and the substrate's reacting site are identical and split into different EC numbers if the bond changes differ, the reacting site is compromised or the substrate experiences an uncatalysed reaction occurring outside the enzyme's active site. The best way to communicate these main findings and recommendations to the enzyme community, EC and reaction databases is currently under investigation.

The comprehensive study of the chemistry of isomerases helped to develop EC-BLAST, a tool to search and compare biochemical reactions (Rahman *et al.*, 2014). A pipeline was produced to test the ability of the algorithm to detect changes of stereochemistry in isomerase reactions. It is based on trial and error and subsequently developed into a method which is routinely used today to check the chemical accuracy of the tool. EC-

BLAST provides a robust alternative to existing approaches to measure the functional similarity between proteins based on ontologies (Gabaldón & Koonin, 2013), although its range of application is limited to enzymes only.

During evolution, isomerase function often exchanges with other EC classes, especially lyases (EC 4). This is also unlike other EC classes where most exchanges occur within the same EC class (Furnham *et al.*, 2012*a*). However, further analyis on whether this trend is a direct consequence of the EC classification of isomerases (Nath & Mitchell, 2012) or the result of chemical similarity between isomerases and other EC classes is needed. In addition isomerases conserve substrates and products more often than conserving bond changes. These observations suggest that isomerases might have originated from enzymes from other EC primary classes, which performed different overall chemistry in similar substrates and products. More analyses and experiments are needed to explore this hypothesis and to help to determine prevalent exchanges for other EC classes.

The EC classification of isomerase reactions only depends on aspects of the overall chemistry such as bond changes, functional groups and substrate specificity. In contrast to other EC primary classes such as oxidoreductases (EC 1), no mechanistic components e.g. cofactors guide the isomerase class. Therefore, the description of isomerase chemistry based on bond changes, reaction centres and substrates and products presented in this dissertation is enough to investigate coherence in the EC system.

However, the study of the evolution of enzyme function needs also to be informed by mechanistic data (Furnham *et al.*, 2012*a*; Nath *et al.*, 2014), which is not considered in this dissertation. For instance, although the analysis of bond changes and reaction centres demonstrated that racemase and epimerase reactions (EC 5.1) are similar in overall chemistry, these enzymes use different mechanisms such as proton or hydride transfers to produce enolates as catalytic intermediates. Twelve EC 5.1 mechanisms are available in MACiE and the four sub-subclasses are represented (Holliday *et al.*, 2012). The means of producing and collapsing the intermediate are different between mechanisms. Some involve covalent catalysis using an organic cofactor such as pyridoxal 5'-phosphate (PLP) or NAD^+ (Frey & Hegeman, 2013; Okazaki *et al.*, 2009; Palani *et al.*, 2013; Tanner, 2008), some stabilise the intermediate using a divalent metal cofactor like Mg^{2+}, Zn^{2+}, Ni^{2+} or Co^{2+} (Desguin *et al.*, 2014; Petsko *et al.*, 1993) and some do not employ any cofactor but only a pair of catalytic amino acids acting as conjugate acid and base, respectively. For instance, the bacterial glutamate racemase (EC 5.1.1.3) first uses a cysteine to deprotonate the substrate L-glutamate, which leads to a planar enolate intermediate. Then, the intermediate collapses by deprotonating a second cysteine and the product D-glutamate is generated. Taken together, mechanisms add an extra layer of functional information to evolutionary analyses and increasing the coverage of resources like MACiE using the enzymology literature seems necessary. Alternatively, in order to inform an overall chemistry representation with mechanistic components, Gasteiger and coworkers proposed a physicochemical description of the reaction centre (Hu *et al.*, 2010; Rose & Gasteiger, 1994). However, this approach does not include fundamental aspects of enzyme mechanisms such as catalytic amino acids and cofactors, which are considered in other methods that compare enzyme mechanisms directly (Almonacid *et al.*, 2010).

From the chemistry point of view, this study depends on the quality of reaction data available in databases (Ott & Vriend, 2006). To some extent the manual curation efforts tried to resolve some discrepancies, however many reactions were not balanced therefore atom-atom mapping becomes impossible. Similarly, stereochemistry assignments are sometimes not valid. Whereas multiple strategies to correct unbalanced reactions (Chen *et al.*, 2013; Kraut *et al.*, 2013; Shaw *et al.*, 2012) and to reconcile biochemical reactions across databases (Bernard *et al.*, 2014) have been recently presented, novel improvements of the algorithms and further data curation and integration are essential (Kumar *et al.*, 2012). In addition, the quality of the manual curation performed in this dissertation depends on the author's ability to interpret reactions, as well as the experimental information available in the primary literature. Overall, the analysis relied only upon the reaction equation and the performance of EC-BLAST to compute accurate atom-atom mappings.

Our study of the evolution of isomerase function relies on the quality of functional annotations in the reviewed section of UniprotKB and PDB (Berman *et al.*, 2013; The Uniprot Consortium, 2013). Recent studies exploring biases in functional annotations discovered that the reviewed section of UniprotKB presents the lowest levels of misannotation compared to other resources (Furnham *et al.*, 2009; Schnoes *et al.*, 2009, 2013). As a result, several strategies were developed to automatically detect misannotations based on sequence and genomic context correlations (Hsiao *et al.*, 2010) and history of changes in functional annotations (Silveira *et al.*, 2014).

How useful is it to have a functional classification of isomerases based on chemical attributes? In general, classification brings order to chaos and helps to extract knowledge from structured data. In the context of this study, classification using EC-BLAST is useful to navigate and compare overall reactions. Although the method is automatic, quantitative and consistent, the best way to classify isomerase function depends on the purpose of the classification and requires the development of criteria. Although the EC classification which is based on overall chemistry is the current standard for reporting isomerase function, several studies have recently shown that other aspects such as substrate substructures not directly involved in the reacting moiety also contribute to the catalytic efficiency of enzymes (Khersonsky *et al.*, 2011) and affect how enzyme function evolves (Furnham *et al.*, 2012*a*; Martinez Cuesta *et al.*, 2014). A recent study explored how changing substrates by removing functional groups that are distantly located from the reacting chemical bonds can considerably reduce the catalytic activity of enzymes (Barelier *et al.*, 2014). Therefore combining bond changes and reaction centres with structural information about the substrates and products and mechanisms is needed to capture the essence of enzyme chemistry in a functional classification. This might help to connect overall reactions to mechanisms and to study the evolution and design of enzyme function.

In the 1980s, the divergence of the amino acid sequence and three-dimensional structure of related proteins was first shown to be monotonic (Chothia & Lesk, 1986), which led to the foundations of homology-based protein structure prediction. Soon after, researchers turned into sequence-based database searching to discover that sequence similarity is also correlated to functional similarity so functions of unknown proteins can be transferred from those of their homologues up to certain extent (Friedberg, 2006). Exceptionally, a single mutation can drastically change function (Schmidt, 2003; Terao *et al.*, 2006) or unrelated proteins might also perform the same function (Omelchenko *et al.*, 2010). In fact, current algorithms for protein function prediction from sequence take this relationship for granted using machine learning principles (Radivojac *et al.*, 2013). However the shape of the association is still unknown. In this dissertation, the relationship between sequence and functional similarity in superfamilies containing isomerases was shown to be remarkably nonlinear, yet monotonic. Other approaches to obtain similarity between enzymes based on the three-dimensional structures or catalytic sites need to be explored in future studies.

The analysis of the evolution of isomerase function is based on exploring the evolution of separate domains. However many enzymes are multidomain and change their domain composition and function during evolution (Bashton & Chothia, 2007). In addition, enzyme function is assigned on a whole-sequence basis without associating specific functions to the composite domains (Lopez & Pazos, 2009). Therefore cataloguing the functional evolution of each individual domain is a complex process, which can lead to multiple different evolutionary pathways. Further research on domain-based assignment of enzyme function and more experimental studies exploring how enzyme function changes with multidomain architecture would complement and broaden this dissertation.

In order to understand more about the relationship between the chemistry and evolution of enzyme function, further research is crucial. This dissertation is isomerase-centric therefore other EC classes need to be investigated. Although the methods are useful to explore and improve the EC classification, understanding the relationships between sequence, structure, function and mechanism requires more work. It now seems even more obvious that future efforts should continue revolving around the question of how enzymes evolve new function. In order to address this fundamental topic, more experimental data is required. Choosing a family or a metabolic pathway to study in detail is not easy and depends on what is already known and the feasibility of conducting experiments. Each family or pathway is complex and different, hence in order to address the problems of finding candidate sequences for orphan enzymes or searching for potential enzyme activities of proteins with unknown function, each family or pathway will have to be considered separately. To be able to continue exploring the origins of biochemistry and further developing enzyme design, these issues need to be resolved in the near future. The field of high-throughput determination of protein function could not be better positioned to help to address the challenges of enzyme research today.

Bibliography

Abe, I. 2014 The Oxidosqualene Cyclases: One Substrate, Diverse Products. In *Nat. prod. discourse, divers. des.* (ed. G. T. Osbourn, Anne; Goss, Rebecca; Carter), chap. 16, pp. 297–317. Wiley, 1st edn. (doi: 10.1002/9781118794623.ch16)

Aguirre, Y., Cabrera, N., Aguirre, B., Pérez-Montfort, R., Hernandez-Santoyo, A., Reyes-Vivas, H., Enríquez-Flores, S., de Gómez-Puyou, M. T., Gómez-Puyou, A. *et al.* 2014 Different contribution of conserved amino acids to the global properties of triosephosphate isomerases. *Proteins*, **82**(2), 323–35. (doi: 10.1002/prot.24398)

Aharoni, A., Gaidukov, L., Khersonsky, O., McQ Gould, S., Roodveldt, C. & Tawfik, D. S. 2005 The evolvability of promiscuous protein functions. *Nat. Genet.*, **37**(1), 73–6. (doi: 10.1038/ng1482)

Alcántara, R., Axelsen, K. B., Morgat, A., Belda, E., Coudert, E., Bridge, A., Cao, H., de Matos, P., Ennis, M. *et al.* 2012 Rhea–a manually curated resource of biochemical reactions. *Nucleic Acids Res.*, **40**(Database issue), D754–60. (doi: 10.1093/nar/gkr1126)

Almonacid, D. E. & Babbitt, P. C. 2011 Toward mechanistic classification of enzyme functions. *Curr. Opin. Chem. Biol.*, **15**(3), 435–42. (doi: 10.1016/j.cbpa.2011.03.008)

Almonacid, D. E., Yera, E. R., Mitchell, J. B. O. & Babbitt, P. C. 2010 Quantitative comparison of catalytic mechanisms and overall reactions in convergently evolved enzymes: implications for classification of enzyme function. *PLoS Comput. Biol.*, **6**(3), e1000 700. (doi: 10.1371/journal.pcbi.1000700)

Altman, T., Travers, M., Kothari, A., Caspi, R. & Karp, P. D. 2013 A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, **14**, 112. (doi: 10.1186/1471-2105-14-112)

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990 Basic local alignment search tool. *J. Mol. Biol.*, **215**(3), 403–10. (doi: 10.1016/S0022-2836(05)80360-2)

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**(17), 3389–402. (doi: 10.1093/nar/25.17.3389)

Amyes, T. L. & Richard, J. P. 2013 Specificity in transition state binding: the Pauling model revisited. *Biochemistry*, **52**(12), 2021–35. (doi: 10.1021/bi301491r)

Anandarajah, K., Kiefer, P. M., Donohoe, B. S. & Copley, S. D. 2000 Recruitment of a double bond isomerase to serve as a reductive dehalogenase during biodegradation of pentachlorophenol. *Biochemistry*, **39**(18), 5303–11. (doi: 10.1021/bi9923813)

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E. & Murzin, A. G. 2014 SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.*, **42**(Database issue), D310–4. (doi: 10.1093/nar/gkt1242)

Anton, B. P., Chang, Y.-C., Brown, P., Choi, H.-P., Faller, L. L., Guleria, J., Hu, Z., Klitgord, N., Levy-Moonshine, A. *et al.* 2013 The COMBREX Project: Design, Methodology, and Initial Results. *PLoS Biol.*, **11**(8), e1001 638. (doi: 10.1371/journal.pbio.1001638)

Apostolakis, J., Sacher, O., Körner, R. & Gasteiger, J. 2008 Automatic determination of reaction mappings and reaction center information. 2. Validation on a biochemical reaction database. *J. Chem. Inf. Model.*, **48**(6), 1190–8. (doi: 10.1021/ci700433d)

Apweiler, R., Armstrong, R., Bairoch, A., Cornish-Bowden, A., Halling, P. J., Hofmeyr, J.-H. S., Kettner, C., Leyh, T. S., Rohwer, J. *et al.* 2010 A large-scale protein-function database. *Nat. Chem. Biol.*, **6**(11), 785. (doi: 10.1038/nchembio.460)

Asano, Y. & Hölsch, K. 2012 Isomerizations. In *Enzym. catal. org. synth.* (eds K. Drauz, H. Gröger & O. May), pp. 1607–1684. Wiley-VCH Verlag, 3rd edn. (doi: 10.1002/9783527639861.ch39)

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S. *et al.* 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**(1), 25–9. (doi: 10.1038/75556)

Babbitt, P. C. 2003 Definitions of enzyme function for the structural genomics era. *Curr. Opin. Chem. Biol.*, **7**(2), 230–7. (doi: 10.1016/S1367-5931(03)00028-0)

Baier, F. & Tokuriki, N. 2014 Connectivity between catalytic landscapes of the metallo-β-lactamase superfamily. *J. Mol. Biol.*, **426**(13), 2442–56. (doi: 10.1016/j.jmb.2014.04.013)

Bairoch, A. 2000 The ENZYME database in 2000. *Nucleic Acids Res.*, **28**(1), 304–5. (doi: 10.1093/nar/28.1.304)

Banerjee, S., Anderson, F. & Farber, G. K. 1995 The evolution of sugar isomerases. *Protein Eng. Des. Sel.*, **8**(12), 1189–1195. (doi: 10.1093/protein/8.12.1189)

Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., Pogson, C. I., Wilson, I. A., Corran, P. H., Furth, A. J., Milman, J. D. *et al.* 1975 Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 angstrom resolution using amino acid sequence data. *Nature*, **255**(5510), 609–14.

Bar-Even, A., Flamholz, A., Noor, E. & Milo, R. 2012 Rethinking glycolysis: on the biochemical logic of metabolic pathways. *Nat. Chem. Biol.*, **8**(6), 509–17. (doi: 10.1038/nchembio.971)

Barelier, S., Cummings, J. A., Rauwerdink, A. M., Hitchcock, D. S., Farelli, J. D., Almo, S. C., Raushel, F. M., Allen, K. N. & Shoichet, B. K. 2014 Substrate deconstruction and the nonadditivity of enzyme recognition. *J. Am. Chem. Soc.*, **136**(20), 7374–82. (doi: 10.1021/ja501354q)

Barker, J. A. & Thornton, J. M. 2003 An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics*, **19**(13), 1644–1649. (doi: 10.1093/bioinformatics/btg226)

Bartlett, G. J., Borkakoti, N. & Thornton, J. M. 2003 Catalysing new reactions during evolution: economy of residues and mechanism. *J. Mol. Biol.*, **331**(4), 829–60. (doi: 10.1016/S0022-2836(03)00734-4)

Bartlett, G. J., Porter, C. T., Borkakoti, N. & Thornton, J. M. 2002 Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **324**(1), 105–21. (doi: 10.1016/S0022-2836(02)01036-7)

Bashton, M. & Chothia, C. 2007 The generation of new protein functions by the combination of domains. *Structure*, **15**(1), 85–99. (doi: 10.1016/j.str.2006.11.009)

Bastard, K., Smith, A. A. T., Vergne-Vaxelaire, C., Perret, A., Zaparucha, A., De Melo-Minardi, R., Mariage, A., Boutard, M., Debard, A. *et al.* 2014 Revealing the hidden functional diversity of an enzyme family. *Nat. Chem. Biol.*, **10**(1), 42–9. (doi: 10.1038/nchembio.1387)

Bawden, D. 1991 Classification of chemical reactions: potential, possibilities and continuing relevance. *J. Chem. Inf. Model.*, **31**(2), 212–216. (doi: 10.1021/ci00002a006)

Berman, H. M., Kleywegt, G. J., Nakamura, H. & Markley, J. L. 2013 The future of the protein data bank. *Biopolymers*, **99**(3), 218–22. (doi: 10.1002/bip.22132)

Bernard, T., Bridge, A., Morgat, A., Moretti, S., Xenarios, I. & Pagni, M. 2014 Reconciliation of metabolites and biochemical reactions for metabolic networks. *Brief. Bioinform.*, **15**(1), 123–35. (doi: 10.1093/bib/bbs058)

Blake, C. C., Koenig, D. F., Mair, G. A., North, A. C., Phillips, D. C. & Sarma, V. R. 1965 Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution. *Nature*, **206**(4986), 757–61.

Bornscheuer, U. T., Huisman, G. W., Kazlauskas, R. J., Lutz, S., Moore, J. C. & Robins, K. 2012 Engineering the third wave of biocatalysis. *Nature*, **485**(7397), 185–94. (doi: 10.1038/nature11117)

Bouvier, J. T., Groninger-Poe, F. P., Vetting, M., Almo, S. C. & Gerlt, J. A. 2014 Galactaro δ-lactone isomerase: lactone isomerization by a member of the amidohydrolase superfamily. *Biochemistry*, **53**(4), 614–6. (doi: 10.1021/bi5000492)

Bray, T., Doig, A. J. & Warwicker, J. 2009 Sequence and structural features of enzymes and their active sites by EC class. *J. Mol. Biol.*, **386**(5), 1423–36. (doi: 10.1016/j.jmb.2008.11.057)

Brecher, J. 2008 Graphical representation standards for chemical structure diagrams (IUPAC Recommendations 2008). *Pure Appl. Chem.*, **80**(2), 277–410. (doi: 10.1351/pac200880020277)

Brown, S. D. & Babbitt, P. C. 2012 Inference of functional properties from large-scale analysis of enzyme superfamilies. *J. Biol. Chem.*, **287**(1), 35–42. (doi: 10.1074/jbc.R111.283408)

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A. *et al.* 2014 The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **42**(1), D459–71. (doi: 10.1093/nar/gkt1103)

Chen, L. & Gasteiger, J. 1997 Knowledge Discovery in Reaction Databases: Landscaping Organic Reactions by a Self-Organizing Neural Network. *J. Am. Chem. Soc.*, **119**(17), 4033–4042. (doi: 10.1021/ja960027b)

Chen, W. L., Chen, D. Z. & Taylor, K. T. 2013 Automatic reaction mapping and reaction center detection. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **3**(6), 560–593. (doi: 10.1002/wcms.1140)

Chiang, R. A., Sali, A. & Babbitt, P. C. 2008 Evolutionarily conserved substrate substructures for automated annotation of enzyme superfamilies. *PLoS Comput. Biol.*, **4**(8), e1000 142. (doi: 10.1371/journal.pcbi.1000142)

Chothia, C. & Lesk, A. M. 1986 The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**(4), 823–6.

Claudel-Renard, C., Chevalet, C., Faraut, T. & Kahn, D. 2003 Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.*, **31**(22), 6633–9. (doi: 10.1093/nar/gkg847)

Conti, P., Tamborini, L., Pinto, A., Blondel, A., Minoprio, P., Mozzarelli, A. & De Micheli, C. 2011 Drug discovery targeting amino acid racemases. *Chem. Rev.*, **111**(11), 6919–46. (doi: 10.1021/cr2000702)

Copley, S. 2003 Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr. Opin. Chem. Biol.*, **7**(2), 265–272. (doi: 10.1016/S1367-5931(03)00032-2)

Copley, S. D. 2012 Toward a systems biology perspective on enzyme evolution. *J. Biol. Chem.*, **287**(1), 3–10. (doi: 10.1074/jbc.R111.254714)

Cornish-Bowden, A. 2014 Current IUBMB recommendations on enzyme nomenclature and kinetics. *Perspect. Sci.*, **1**(1-6), 74–87. (doi: 10.1016/j.pisc.2014.02.006)

Crawley, M. J. 2007 *The R Book.* Chichester, UK: John Wiley & Sons, Ltd. (doi: 10.1002/9780470515075)

Daenzer, J. M. I., Sanders, R. D., Hang, D. & Fridovich-Keil, J. L. 2012 UDP-galactose 4'-epimerase activities toward UDP-Gal and UDP-GalNAc play different roles in the development of Drosophila melanogaster. *PLoS Genet.*, **8**(5), e1002 721. (doi: 10.1371/journal.pgen.1002721)

Daugherty, A. B., Govindarajan, S. & Lutz, S. 2013 Improved biocatalysts from a synthetic circular permutation library of the flavin-dependent oxidoreductase old yellow enzyme. *J. Am. Chem. Soc.*, **135**(38), 14 425–32. (doi: 10.1021/ja4074886)

de Beer, T. A. P., Berka, K., Thornton, J. M. & Laskowski, R. A. 2014 PDBsum additions. *Nucleic Acids Res.*, **42**(Database issue), D292–6. (doi: 10.1093/nar/gkt940)

De Ferrari, L., Aitken, S., van Hemert, J. & Goryanin, I. 2012 EnzML: multi-label prediction of enzyme classes using InterPro signatures. *BMC Bioinformatics*, **13**(1), 61. (doi: 10.1186/1471-2105-13-61)

De Ferrari, L. & Mitchell, J. B. O. 2014 From sequence to enzyme mechanism using multi-label machine learning. *BMC Bioinformatics*, **15**(1), 150. (doi: 10.1186/1471-2105-15-150)

de Luca, A., Horvath, D., Marcou, G., Solov'ev, V. & Varnek, A. 2012 Mining chemical reactions using neighborhood behavior and condensed graphs of reactions approaches. *J. Chem. Inf. Model.*, **52**(9), 2325–38. (doi: 10.1021/ci300149n)

Dellus-Gur, E., Toth-Petroczy, A., Elias, M. & Tawfik, D. S. 2013 What makes a protein fold amenable to functional innovation? Fold polarity and stability trade-offs. *J. Mol. Biol.*, **425**(14), 2609–21. (doi: 10.1016/j.jmb.2013.03.033)

des Jardins, M., Karp, P. D., Krummenacker, M., Lee, T. J. & Ouzounis, C. A. 1997 Prediction of enzyme classification from protein sequence without the use of sequence similarity. *Proc Int Conf Intell Syst Mol Biol*, **5**, 92–9.

Desai, D. K., Nandi, S., Srivastava, P. K. & Lynn, A. M. 2011 ModEnzA: Accurate Identification of Metabolic Enzymes Using Function Specific Profile HMMs with Optimised Discrimination Threshold and Modified Emission Probabilities. *Adv. Bioinformatics*, **2011**, 743 782. (doi: 10.1155/2011/743782)

Desguin, B., Goffin, P., Viaene, E., Kleerebezem, M., Martin-Diaconescu, V., Maroney, M. J., Declercq, J.-P., Soumillion, P. & Hols, P. 2014 Lactate racemase is a nickel-dependent enzyme activated by a widespread maturation system. *Nat. Commun.*, **5**, 3615. (doi: 10.1038/ncomms4615)

Dessailly, B. H., Dawson, N. L., Mizuguchi, K. & Orengo, C. A. 2013 Functional site plasticity in domain superfamilies. *Biochim. Biophys. Acta*, **1834**(5), 874–89. (doi: 10.1016/j.bbapap.2013.02.042)

Dobson, P. D. & Doig, A. J. 2005 Predicting enzyme class from protein structure without alignments. *J. Mol. Biol.*, **345**(1), 187–99. (doi: 10.1016/j.jmb.2004.10.024)

Eddy, S. R. 2011 Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**(10), e1002 195. (doi: 10.1371/journal.pcbi.1002195)

Egelhofer, V., Schomburg, I. & Schomburg, D. 2010 Automatic Assignment of EC Numbers. *PLoS Comput Biol*, **6**(1), e1000 661. (doi: 10.1371/journal.pcbi.1000661)

Eixelsberger, T., Sykora, S., Egger, S., Brunsteiner, M., Kavanagh, K. L., Oppermann, U., Brecker, L. & Nidetzky, B. 2012 Structure and mechanism of human UDP-xylose synthase: evidence for a promoting role of sugar ring distortion in a three-step catalytic conversion of UDP-glucuronic acid. *J. Biol. Chem.*, **287**(37), 31 349–58. (doi: 10.1074/jbc.M112.386706)

El Yacoubi, B. & de Crécy-Lagard, V. 2014 Integrative data-mining tools to link gene and function. *Methods Mol. Biol.*, **1101**, 43–66. (doi: 10.1007/978-1-62703-721-1_4)

Elias, M. & Tawfik, D. S. 2012 Divergence and convergence in enzyme evolution: parallel evolution of paraoxonases from quorum-quenching lactonases. *J. Biol. Chem.*, **287**(1), 11–20. (doi: 10.1074/jbc.R111.257329)

Everitt, B. & Hothorn, T. 2011 *An Introduction to Applied Multivariate Analysis with R*. Use R! New York, NY: Springer New York. (doi: 10.1007/978-1-4419-9650-3)

Faulon, J.-L., Misra, M., Martin, S., Sale, K. & Sapra, R. 2008 Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. *Bioinformatics*, **24**(2), 225–33. (doi: 10.1093/bioinformatics/btm580)

Fersht, A. 1999 *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*. New York: W. H. Freeman.

Fischer, J. D., Holliday, G. L., Rahman, S. A. & Thornton, J. M. 2010 The structures and physicochemical properties of organic cofactors in biocatalysis. *J. Mol. Biol.*, **403**(5), 803–24. (doi: 10.1016/j.jmb.2010.09.018)

Foster, J. M., Davis, P. J., Raverdy, S., Sibley, M. H., Raleigh, E. A., Kumar, S. & Carlow, C. K. S. 2010 Evolution of bacterial phosphoglycerate mutases: non-homologous isofunctional enzymes undergoing gene losses, gains and lateral transfers. *PLoS One*, **5**(10), e13 576. (doi: 10.1371/journal.pone.0013576)

Fraley, C., Raftery, A. E., Murphy, T. B. & Scrucca, L. 2012 mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical Report No. 597. Tech. rep., Department of Statistics, University of Washington.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P. *et al.* 2013 STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**(Database issue), D808–15. (doi: 10.1093/nar/gks1094)

Frey, P. A. & Hegeman, A. D. 2013 Chemical and stereochemical actions of UDP-galactose 4-epimerase. *Acc. Chem. Res.*, **46**(7), 1417–26. (doi: 10.1021/ar300246k)

Frey, P. A., Hegeman, A. D. & Ruzicka, F. J. 2008 The Radical SAM Superfamily. *Crit. Rev. Biochem. Mol. Biol.*, **43**(1), 63–88. (doi: 10.1080/10409230701829169)

Friedberg, I. 2006 Automated protein function prediction–the genomic challenge. *Brief. Bioinform.*, **7**(3), 225–42. (doi: 10.1093/bib/bbl004)

Furnham, N., Garavelli, J. S., Apweiler, R. & Thornton, J. M. 2009 Missing in action: enzyme functional annotations in biological databases. *Nat. Chem. Biol.*, **5**(8), 521–5. (doi: 10.1038/nchembio0809-521)

Furnham, N., Holliday, G. L., de Beer, T. A. P., Jacobsen, J. O. B., Pearson, W. R. & Thornton, J. M. 2014 The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.*, **42**(1), D485–9. (doi: 10.1093/nar/gkt1243)

Furnham, N., Sillitoe, I., Holliday, G. L., Cuff, A. L., Laskowski, R. A., Orengo, C. A. & Thornton, J. M. 2012*a* Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies. *PLoS Comput. Biol.*, **8**(3), e1002 403. (doi: 10.1371/journal.pcbi.1002403)

Furnham, N., Sillitoe, I., Holliday, G. L., Cuff, A. L., Rahman, S. A., Laskowski, R. A., Orengo, C. A. & Thornton, J. M. 2012*b* FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucleic Acids Res.*, **40**(Database issue), D776–82. (doi: 10.1093/nar/gkr852)

Gabaldón, T. & Koonin, E. V. 2013 Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.*, **14**(5), 360–6. (doi: 10.1038/nrg3456)

Gall, M., Thomsen, M., Peters, C., Pavlidis, I. V., Jonczyk, P., Grünert, P. P., Beutel, S., Scheper, T., Gross, E. *et al.* 2014 Enzymatic Conversion of Flavonoids using Bacterial Chalcone Isomerase and Enoate Reductase. *Angew. Chem. Int. Ed. Engl.*, **53**(5), 1439–42. (doi: 10.1002/anie.201306952)

Galperin, M. Y. & Koonin, E. V. 2012 Divergence and convergence in enzyme evolution. *J. Biol. Chem.*, **287**(1), 21–8. (doi: 10.1074/jbc.R111.241976)

Garcia-Seisdedos, H., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. 2012 Probing the mutational interplay between primary and promiscuous protein functions: a computational-experimental approach. *PLoS Comput. Biol.*, **8**(6), e1002 558. (doi: 10.1371/journal.pcbi.1002558)

Gardossi, L., Poulsen, P. B., Ballesteros, A., Hult, K., Svedas, V. K., Vasić-Racki, D., Carrea, G., Magnusson, A., Schmid, A. *et al.* 2010 Guidelines for reporting of biocatalytic reactions. *Trends Biotechnol.*, **28**(4), 171–80. (doi: 10.1016/j.tibtech.2010.01.001)

Gariev, I. A. & Varfolomeev, S. D. 2006 Hierarchical classification of hydrolases catalytic sites. *Bioinformatics*, **22**(20), 2574–6. (doi: 10.1093/bioinformatics/btl413)

Gaston, M. A., Zhang, L., Green-Church, K. B. & Krzycki, J. A. 2011 The complete biosynthesis of the genetically encoded amino acid pyrrolysine from lysine. *Nature*, **471**(7340), 647–50. (doi: 10.1038/nature09918)

George, R. A., Spriggs, R. V., Bartlett, G. J., Gutteridge, A., MacArthur, M. W., Porter, C. T., Al-Lazikani, B., Thornton, J. M. & Swindells, M. B. 2005 Effective function annotation through catalytic residue conservation. *Proc. Natl. Acad. Sci. U. S. A.*, **102**(35), 12 299–304. (doi: 10.1073/pnas.0504833102)

Gerlt, J. A., Allen, K. N., Almo, S. C., Armstrong, R. N., Babbitt, P. C., Cronan, J. E., Dunaway-Mariano, D., Imker, H. J., Jacobson, M. P. *et al.* 2011 The enzyme function initiative. *Biochemistry*, **50**(46), 9950–62. (doi: 10.1021/bi201312u)

Gerlt, J. A. & Babbitt, P. C. 2001 Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu. Rev. Biochem.*, **70**, 209–46. (doi: 10.1146/annurev.biochem.70.1.209)

Gerlt, J. A., Babbitt, P. C., Jacobson, M. P. & Almo, S. C. 2012 Divergent evolution in enolase superfamily: strategies for assigning functions. *J. Biol. Chem.*, **287**(1), 29–34. (doi: 10.1074/jbc.R111.240945)

Geyer, P. 2013 Markush structure searching by information professionals in the chemical industry - Our views and expectations. *World Pat. Inf.*, **35**(3), 178–182. (doi: 10.1016/j.wpi.2013.05.002)

Gherardini, P. F., Wass, M. N., Helmer-Citterich, M. & Sternberg, M. J. E. 2007 Convergent evolution of enzyme active sites is not a rare phenomenon. *J. Mol. Biol.*, **372**(3), 817–45. (doi: 10.1016/j.jmb.2007.06.017)

Glasner, M. E., Fayazmanesh, N., Chiang, R. A., Sakai, A., Jacobson, M. P., Gerlt, J. A. & Babbitt, P. C. 2006*a* Evolution of structure and function in the o-succinylbenzoate synthase/N-acylamino acid racemase family of the enolase superfamily. *J. Mol. Biol.*, **360**(1), 228–50. (doi: 10.1016/j.jmb.2006.04.055)

Glasner, M. E., Gerlt, J. A. & Babbitt, P. C. 2006*b* Evolution of enzyme superfamilies. *Curr. Opin. Chem. Biol.*, **10**(5), 492–7. (doi: 10.1016/j.cbpa.2006.08.012)

Green, M. L. & Karp, P. D. 2005 Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucleic Acids Res.*, **33**(13), 4035–9. (doi: 10.1093/nar/gki711)

Green, M. L. & Karp, P. D. 2007 Using genome-context data to identify specific types of functional associations in pathway/genome databases. *Bioinformatics*, **23**(13), i205–i211. (doi: 10.1093/bioinformatics/btm213)

Greenacre, M. 2007 *Correspondence Analysis in Practice.* Taylor and Francis, 2nd edn.

Grethe, G., Goodman, J. M. & Allen, C. H. 2013 International chemical identifier for reactions (RInChI). *J. Cheminform.*, **5**(1), 45. (doi: 10.1186/1758-2946-5-45)

Gutteridge, A. & Thornton, J. M. 2005 Understanding nature's catalytic toolkit. *Trends Biochem. Sci.*, **30**(11), 622–9. (doi: 10.1016/j.tibs.2005.09.006)

Hamma, T. & Ferré-D'Amaré, A. R. 2006 Pseudouridine synthases. *Chem. Biol.*, **13**(11), 1125–35. (doi: 10.1016/j.chembiol.2006.09.009)

Hammes-Schiffer, S. 2013 Catalytic efficiency of enzymes: a theoretical analysis. *Biochemistry*, **52**(12), 2012–20. (doi: 10.1021/bi301515j)

Hanson, A. D., Pribat, A., Waller, J. C. & de Crécy-Lagard, V. 2010 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list–and how to find it. *Biochem. J.*, **425**(1), 1–11. (doi: 10.1042/BJ20091328)

Hartmans, S., Smits, J. P., van der Werf, M. J., Volkering, F. & de Bont, J. A. 1989 Metabolism of Styrene Oxide and 2-Phenylethanol in the Styrene-Degrading Xanthobacter Strain 124X. *Appl. Environ. Microbiol.*, **55**(11), 2850–5.

Hatzimanikatis, V., Li, C., Ionita, J. A., Henry, C. S., Jankowski, M. D. & Broadbelt, L. J. 2005 Exploring the diversity of complex metabolic networks. *Bioinformatics*, **21**(8), 1603–9. (doi: 10.1093/bioinformatics/bti213)

Hegeman, A., Gross, J. & Frey, P. 2002 Concerted and stepwise dehydration mechanisms observed in wild-type and mutated Escherichia coli dTDP-glucose 4, 6-dehydratase. *Biochemistry*, **41**(8), 2797–2804. (doi: 10.1021/bi011748c)

Heng, L. 2006 Constructing the TreeFam database. Ph.D. thesis, Chinese Academic of Science.

Herschlag, D. & Natarajan, A. 2013 Fundamental challenges in mechanistic enzymology: progress toward understanding the rate enhancements of enzymes. *Biochemistry*, **52**(12), 2050–67. (doi: 10.1021/bi4000113)

Hilterhaus, L. & Liese, A. 2012 Industrial Application and Processes Using Isomerases. In *Enzym. catal. org. synth.* (eds K. Drauz, H. Gröger & O. May), Ec 4, pp. 1685–1691. Wiley-VCH Verlag, third edit edn. (doi: 10.1002/9783527639861.ch39)

Höcker, B., Jürgens, C., Wilmanns, M. & Sterner, R. 2001 Stability, catalytic versatility and evolution of the $(\beta\alpha)(8)$-barrel fold. *Curr. Opin. Biotechnol.*, **12**(4), 376–81. (doi: 10.1021/cr030191z)

Holliday, G. L., Almonacid, D. E., Mitchell, J. B. O. & Thornton, J. M. 2007*a* The chemistry of protein catalysis. *J. Mol. Biol.*, **372**(5), 1261–77. (doi: 10.1016/j.jmb.2007.07.034)

Holliday, G. L., Andreini, C., Fischer, J. D., Rahman, S. A., Almonacid, D. E., Williams, S. T. & Pearson, W. R. 2012 MACiE: exploring the diversity of biochemical reactions. *Nucleic Acids Res.*, **40**(Database issue), D783–9. (doi: 10.1093/nar/gkr799)

Holliday, G. L., Fischer, J. D., Mitchell, J. B. O. & Thornton, J. M. 2011 Characterizing the complexity of enzymes on the basis of their mechanisms and structures with a bio-computational analysis. *FEBS J.*, **278**(20), 3835–45. (doi: 10.1111/j.1742-4658.2011.08190.x)

Holliday, G. L., Mitchell, J. B. O. & Thornton, J. M. 2009 Understanding the functional roles of amino acid residues in enzyme catalysis. *J. Mol. Biol.*, **390**(3), 560–77. (doi: 10.1016/j.jmb.2009.05.015)

Holliday, G. L., Rahman, S. A., Furnham, N. & Thornton, J. M. 2014 Exploring the biological and chemical complexity of the ligases. *J. Mol. Biol.*, **426**(10), 2098–111. (doi: 10.1016/j.jmb.2014.03.008)

Holliday, G. L., Thornton, J. M., Marquet, A., Smith, A. G., Rébeillé, F., Mendel, R., Schubert, H. L., Lawrence, A. D. & Warren, M. J. 2007*b* Evolution of enzymes and pathways for the biosynthesis of cofactors. *Nat. Prod. Rep.*, **24**(5), 972–87. (doi: 10.1039/b703107f)

Hsiao, T.-L., Revelles, O., Chen, L., Sauer, U. & Vitkup, D. 2010 Automatic policing of biochemical annotations using genomic correlations. *Nat. Chem. Biol.*, **6**(1), 34–40. (doi: 10.1038/nchembio.266)

Hu, Q.-N., Zhu, H., Li, X., Zhang, M., Deng, Z., Yang, X. & Deng, Z. 2012 Assignment of EC numbers to enzymatic reactions with reaction difference fingerprints. *PLoS One*, **7**(12), e52 901. (doi: 10.1371/journal.pone.0052901)

Hu, X., Yan, A., Tan, T., Sacher, O. & Gasteiger, J. 2010 Similarity perception of reactions catalyzed by oxidoreductases and hydrolases using different classification methods. *J. Chem. Inf. Model.*, **50**(6), 1089–100. (doi: 10.1021/ci9004833)

Huang, R., Hippauf, F., Rohrbeck, D., Haustein, M., Wenke, K., Feike, J., Sorrelle, N., Piechulla, B. & Barkman, T. J. 2012 Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates. *Proc. Natl. Acad. Sci. U. S. A.*, **109**(8), 2966–71. (doi: 10.1073/pnas.1019605109)

Hult, K. & Berglund, P. 2007 Enzyme promiscuity: mechanism and applications. *Trends Biotechnol.*, **25**(5), 231–8. (doi: 10.1016/j.tibtech.2007.03.002)

Husson, F., Lê, S. & Pagès, J. 2011 *Exploratory Multivariate Analysis by Example Using R.* Chapman & Hall/CRC, 1st edn.

IUPAC 2014 *Compendium of Chemical Terminology. Gold Book.* 2nd edn. (doi: 10.1351/goldbook)

Izrailev, S. & Farnum, M. A. 2004 Enzyme classification by ligand binding. *Proteins*, **57**(4), 711–24. (doi: 10.1002/prot.20277)

Jacques, S. L., Nieman, C., Bareich, D., Broadhead, G., Kinach, R., Honek, J. F. & Wright, G. D. 2001 Characterization of yeast homoserine dehydrogenase, an antifungal target: the invariant histidine 309 is important for enzyme integrity. *Biochim. Biophys. Acta*, **1544**(1-2), 28–41. (doi: 10.1016/S0167-4838(00)00203-X)

Jochum, C., Gasteiger, J. & Ugi, I. 1980 The Principle of Minimum Chemical Distance (PMCD). *Angew. Chem. Int. Ed. Engl.*, **19**(7), 495–505.

Kaltenbach, M. & Tokuriki, N. 2014 Dynamics and constraints of enzyme evolution. *J. Exp. Zool. B. Mol. Dev. Evol.*, (August 2013), 1–20. (doi: 10.1002/jez.b.22562)

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. 2012 KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**(Database issue), D109–14. (doi: 10.1093/nar/gkr988)

Karp, P. 2004 Call for an enzyme genomics initiative. *Genome Biol.*, **5**(8), 401. (doi: 10.1186/gb-2004-5-8-401)

Kawashima, S., Katayama, T., Sato, Y. & Kanehisa, M. 2003 KEGG API: A web service using SOAP/WSDL to access the KEGG system. *Genome Informatics*, **14**, 673–674.

Kazlauskas, R. J. 2005 Enhancing catalytic promiscuity for biocatalysis. *Curr. Opin. Chem. Biol.*, **9**(2), 195–201. (doi: 10.1016/j.cbpa.2005.02.008)

Kelley, L. A., Gardner, S. P. & Sutcliffe, M. J. 1996 An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Eng.*, **9**(11), 1063–5. (doi: 10.1093/protein/9.11.1063)

Kersey, P. J., Allen, J. E., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., Hughes, D. S. T., Humphrey, J., Kerhornou, A. *et al.* 2014 Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.*, **42**(Database issue), D546–52. (doi: 10.1093/nar/gkt979)

Kharchenko, P., Chen, L., Freund, Y., Vitkup, D. & Church, G. 2006 Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics*, **7**(1), 177. (doi: 10.1186/1471-2105-7-177)

Khersonsky, O., Malitsky, S., Rogachev, I. & Tawfik, D. 2011 Role of chemistry versus substrate binding in recruiting promiscuous enzyme functions. *Biochemistry*, **50**(13), 2683–2690. (doi: 10.1021/bi101763c)

Khersonsky, O. & Tawfik, D. S. 2010 Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.*, **79**, 471–505. (doi: 10.1146/annurev-biochem-030409-143718)

Kim, I.-K., Roldão, A., Siewers, V. & Nielsen, J. 2012 A systems-level approach for metabolic engineering of yeast cell factories. *FEMS Yeast Res.*, **12**(2), 228–48. (doi: 10.1111/j.1567-1364.2011.00779.x)

Kimura, Y., Aoki, T. & Ayabe, S. 2001 Chalcone isomerase isozymes with different substrate specificities towards 6'-hydroxy- and 6'-deoxychalcones in cultured cells of Glycyrrhiza echinata, a leguminous plant producing 5-deoxyflavonoids. *Plant Cell Physiol.*, **42**(10), 1169–73. (doi: 10.1093/pcp/pce130)

Kiss, G., Çelebi Ölçüm, N., Moretti, R., Baker, D. & Houk, K. N. 2013 Computational enzyme design. *Angew. Chem. Int. Ed. Engl.*, **52**(22), 5700–25. (doi: 10.1002/anie.201204077)

Körner, R. & Apostolakis, J. 2008 Automatic determination of reaction mappings and reaction center information. 1. The imaginary transition state energy approach. *J. Chem. Inf. Model.*, **48**(6), 1181–9. (doi: 10.1021/ci7004324)

Kotera, M., Goto, S. & Kanehisa, M. 2014 Predictive genomic and metabolomic analysis for the standardization of enzyme data. *Perspect. Sci.*, **1**(1-6), 24–32. (doi: 10.1016/j.pisc.2014.02.003)

Kotera, M., McDonald, A. G., Boyce, S. & Tipton, K. F. 2008 Functional group and substructure searching as a tool in metabolomics. *PLoS One*, **3**(2), e1537. (doi: 10.1371/journal.pone.0001537)

Kotera, M., Okuno, Y., Hattori, M., Goto, S. & Kanehisa, M. 2004 Computational Assignment of the EC Numbers for Genomic-Scale Analysis of Enzymatic Reactions. *J. Am. Chem. Soc.*, **126**(50), 16 487–16 498. (doi: 10.1021/ja0466457)

Kowatz, T., Morrison, J. P., Tanner, M. E. & Naismith, J. H. 2010 The crystal structure of the Y140F mutant of ADP-L-glycero-D-manno-heptose 6-epimerase bound to ADP-beta-D-mannose suggests a one base mechanism. *Protein Sci.*, **19**(7), 1337–43. (doi: 10.1002/pro.410)

Kraut, H., Eiblmaier, J., Grethe, G., Löw, P., Matuszczyk, H. & Saller, H. 2013 Algorithm for reaction classification. *J. Chem. Inf. Model.*, **53**(11), 2884–95. (doi: 10.1021/ci400442f)

Krebs, H. A. 1940 The citric acid cycle and the Szent-Györgyi cycle in pigeon breast muscle. *Biochem. J.*, **34**(5), 775–9.

Kresge, N., Simoni, R. D. & Hill, R. L. 2005 Otto Fritz Meyerhof and the elucidation of the glycolytic pathway. *J. Biol. Chem.*, **280**(4), e3.

Kuhm, A. E., Schlömann, M., Knackmuss, H. J. & Pieper, D. H. 1990 Purification and characterization of dichloromuconate cycloisomerase from Alcaligenes eutrophus JMP 134. *Biochem. J.*, **266**(3), 877–83.

Kumar, A., Suthers, P. F. & Maranas, C. D. 2012 MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinformatics*, **13**(1), 6. (doi: 10.1186/1471-2105-13-6)

Kumar, C. & Choudhary, A. 2012 A top-down approach to classify enzyme functional classes and sub-classes using random forest. *EURASIP J. Bioinform. Syst. Biol.*, **2012**(1), 1. (doi: 10.1186/1687-4153-2012-1)

Kumar, N. & Skolnick, J. 2012 EFICAz2.5: application of a high-precision enzyme function predictor to 396 proteomes. *Bioinformatics*, **28**(20), 2687–8. (doi: 10.1093/bioinformatics/bts510)

Künzler, D. E., Sasso, S., Gamper, M., Hilvert, D. & Kast, P. 2005 Mechanistic insights into the isochorismate pyruvate lyase activity of the catalytically promiscuous PchB from combinatorial mutagenesis and selection. *J. Biol. Chem.*, **280**(38), 32 827–34. (doi: 10.1074/jbc.M506883200)

Landrum, G. 2013 RDKit: Open-source cheminformatics.

Lang, M., Stelzer, M. & Schomburg, D. 2011 BKM-react, an integrated biochemical reaction database. *BMC Biochem.*, **12**(1), 42. (doi: 10.1186/1471-2091-12-42)

Laskowski, R. A., Watson, J. D. & Thornton, J. M. 2005*a* ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, **33**(Web Server issue), W89–93. (doi: 10.1093/nar/gki414)

Laskowski, R. A., Watson, J. D. & Thornton, J. M. 2005*b* Protein function prediction using local 3D templates. *J. Mol. Biol.*, **351**(3), 614–26. (doi: 10.1016/j.jmb.2005.05.067)

Latino, D. A. R. S. & Aires-de Sousa, J. 2006 Genome-scale classification of metabolic reactions: a chemoinformatics approach. *Angew. Chem. Int. Ed. Engl.*, **45**(13), 2066–9. (doi: 10.1002/anie.200503833)

Latino, D. A. R. S. & Aires-de Sousa, J. 2007 Linking Databases of Chemical Reactions to NMR Data: an Exploration of 1 H NMR-Based Reaction Classification. *Anal. Chem.*, **79**(3), 854–862. (doi: 10.1021/ac060979s)

Latino, D. A. R. S. & Aires-de Sousa, J. 2009 Assignment of EC numbers to enzymatic reactions with MOLMAP reaction descriptors and random forests. *J. Chem. Inf. Model.*, **49**(7), 1839–46. (doi: 10.1021/ci900104b)

Latino, D. A. R. S. & Aires-de Sousa, J. 2011 *Classification of Chemical Reactions and Chemoinformatic Processing of Enzymatic Transformations*, vol. 672 of *Methods in Molecular Biology*. Totowa, NJ: Humana Press. (doi: 10.1007/978-1-60761-839-3)

Latino, D. A. R. S., Zhang, Q.-Y. & Aires-de Sousa, J. 2008 Genome-scale classification of metabolic reactions and assignment of EC numbers with self-organizing maps. *Bioinformatics*, **24**(19), 2236–44. (doi: 10.1093/bioinformatics/btn405)

Leber, M., Egelhofer, V., Schomburg, I. & Schomburg, D. 2009 Automatic assignment of reaction operators to enzymatic reactions. *Bioinformatics*, **25**(23), 3135–42. (doi: 10.1093/bioinformatics/btp549)

Lee, S. M., Jellison, T. & Alper, H. S. 2012 Directed evolution of xylose isomerase for improved xylose catabolism and fermentation in the yeast Saccharomyces cerevisiae. *Appl. Environ. Microbiol.*, **78**(16), 5708–16. (doi: 10.1128/AEM.01419-12)

Lees, J., Yeats, C., Redfern, O., Clegg, A. & Orengo, C. 2010 Gene3D: merging structure and function for a Thousand genomes. *Nucleic Acids Res.*, **38**(Database issue), D296–300. (doi: 10.1093/nar/gkp987)

Legendre, P. & Legendre, L. F. J. 2012 *Numerical Ecology.* Elsevier Science, 3rd edn.

Leonard, E., Ajikumar, P. K., Thayer, K., Xiao, W.-H., Mo, J. D., Tidor, B., Stephanopoulos, G. & Prather, K. L. J. 2010 Combining metabolic and protein engineering of a terpenoid biosynthetic pathway for overproduction and selectivity control. *Proc. Natl. Acad. Sci. U. S. A.*, **107**(31), 13 654–9. (doi: 10.1073/pnas.1006138107)

Lespinet, O. & Labedan, B. 2005 Orphan enzymes? *Science*, **307**(5706), 42. (doi: 10.1126/science.307.5706.42a)

Light, S. & Elofsson, A. 2013 The impact of splicing on protein domain architecture. *Curr. Opin. Struct. Biol.*, **23**(3), 451–8. (doi: 10.1016/j.sbi.2013.02.013)

Lim, Y. H., Yokoigawa, K., Esaki, N. & Soda, K. 1993 A new amino acid racemase with threonine alpha-epimerase activity from Pseudomonas putida: purification and characterization. *J. Bacteriol.*, **175**(13), 4213–7.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. 2014 The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.*, **42**(Database issue), D490–5. (doi: 10.1093/nar/gkt1178)

Lopez, D. & Pazos, F. 2009 Gene ontology functional annotations at the structural domain level. *Proteins*, **76**(3), 598–607. (doi: 10.1002/prot.22373)

Lukk, T., Sakai, A., Kalyanaraman, C., Brown, S. D., Imker, H. J., Song, L., Fedorov, A. A., Fedorov, E. V., Toro, R. *et al.* 2012 Homology models guide discovery of diverse enzyme specificities among dipeptide epimerases in the enolase superfamily. *Proc. Natl. Acad. Sci. U. S. A.*, **109**(11), 4122–7. (doi: 10.1073/pnas.1112081109)

Lundqvist, T., Fisher, S. L., Kern, G., Folmer, R. H. A., Xue, Y., Newton, D. T., Keating, T. A., Alm, R. A. & de Jonge, B. L. M. 2007 Exploitation of structural and regulatory diversity in glutamate racemases. *Nature*, **447**(7146), 817–22. (doi: 10.1038/nature05689)

Maley, F. & Maley, G. F. 1959 The enzymic conversion of glucosamine to galactosamine. *Biochim. Biophys. Acta*, **31**(2), 577–8.

Martinez Cuesta, S., Furnham, N., Rahman, S. A., Sillitoe, I. & Thornton, J. M. 2014 The evolution of enzyme function in the isomerases. *Curr. Opin. Struct. Biol.*, **26C**, 121–130. (doi: 10.1016/j.sbi.2014.06.002)

Mavridis, L., Nath, N. & Mitchell, J. B. O. 2013 PFClust: a novel parameter free clustering algorithm. *BMC Bioinformatics*, **14**(1), 213. (doi: 10.1186/1471-2105-14-213)

May, J. W., James, A. G. & Steinbeck, C. 2013 Metingear: a development environment for annotating genome-scale metabolic models. *Bioinformatics*, **29**(17), 2213–5. (doi: 10.1093/bioinformatics/btt342)

McDonald, A. & Tipton, K. 2014 Fifty–five years of enzyme classification: advances and difficulties. *FEBS J.*, **281**(2), 583–592. (doi: 10.1111/febs.12530)

McDonald, A. G., Boyce, S. & Tipton, K. F. 2009 ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.*, **37**(Database issue), D593–7. (doi: 10.1093/nar/gkn582)

Meng, E. C. & Babbitt, P. C. 2011 Topological variation in the evolution of new reactions in functionally diverse enzyme superfamilies. *Curr. Opin. Struct. Biol.*, **21**(3), 391–7. (doi: 10.1016/j.sbi.2011.03.007)

Monk, J., Nogales, J. & Palsson, B. O. 2014 Optimizing genome-scale network reconstructions. *Nat. Biotechnol.*, **32**(5), 447–52. (doi: 10.1038/nbt.2870)

Moriya, Y., Shigemizu, D., Hattori, M., Tokimatsu, T., Kotera, M., Goto, S. & Kanehisa, M. 2010 PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res.*, **38**(Web Server issue), W138–W143. (doi: 10.1093/nar/gkq318)

Mu, F., Unkefer, C. J., Unkefer, P. J. & Hlavacek, W. S. 2011 Prediction of metabolic reactions based on atomic and molecular properties of small-molecule compounds. *Bioinformatics*, **27**(11), 1537–1545. (doi: 10.1093/bioinformatics/btr177)

Mu, F., Unkefer, P. J., Unkefer, C. J. & Hlavacek, W. S. 2006 Prediction of oxidoreductase-catalyzed reactions based on atomic properties of metabolites. *Bioinformatics*, **22**(24), 3082–8. (doi: 10.1093/bioinformatics/btl535)

Muller, C., Marcou, G., Horvath, D., Aires-de Sousa, J. & Varnek, A. 2012 Models for Identification of Erroneous Atom-to-Atom Mapping of Reactions Performed by Automated Algorithms. *J. Chem. Inf. Model.*, **52**(12), 3116–22. (doi: 10.1021/ci300418q)

Nagao, C., Nagano, N. & Mizuguchi, K. 2014 Prediction of detailed enzyme functions and identification of specificity determining residues by random forests. *PLoS One*, **9**(1), e84 623. (doi: 10.1371/journal.pone.0084623)

Nath, N. & Mitchell, J. B. O. 2012 Is EC class predictable from reaction mechanism? *BMC Bioinformatics*, **13**(1), 60. (doi: 10.1186/1471-2105-13-60)

Nath, N., Mitchell, J. B. O. & Caetano-Anollés, G. 2014 The natural history of biocatalytic mechanisms. *PLoS Comput. Biol.*, **10**(5), e1003 642. (doi: 10.1371/journal.pcbi.1003642)

Nguyen, T. T., Brown, S., Fedorov, A. A., Fedorov, E. V., Babbitt, P. C., Almo, S. C. & Raushel, F. M. 2008 At the periphery of the amidohydrolase superfamily: Bh0493 from Bacillus halodurans catalyzes the isomerization of D-galacturonate to D-tagaturonate. *Biochemistry*, **47**(4), 1194–206. (doi: 10.1021/bi7017738)

Nguyen, T. T., Fedorov, A. A., Williams, L., Fedorov, E. V., Li, Y., Xu, C., Almo, S. C. & Raushel, F. M. 2009 The mechanism of the reaction catalyzed by uronate isomerase illustrates how an isomerase may have evolved from a hydrolase within the amidohydrolase superfamily. *Biochemistry*, **48**(37), 8879–90. (doi: 10.1021/bi901046x)

Nobeli, I., Favia, A. D. & Thornton, J. M. 2009 Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.*, **27**(2), 157–167. (doi: 10.1038/nbt1519)

Nobeli, I., Ponstingl, H., Krissinel, E. B. & Thornton, J. M. 2003 A structure-based anatomy of the E.coli metabolome. *J. Mol. Biol.*, **334**(4), 697–719. (doi: 10.1016/j.jmb.2003.10.008)

Nobeli, I., Spriggs, R. V., George, R. A. & Thornton, J. M. 2005 A ligand-centric analysis of the diversity and evolution of protein-ligand relationships in E.coli. *J. Mol. Biol.*, **347**(2), 415–36. (doi: 10.1016/j.jmb.2005.01.061)

Nosrati, G. R. & Houk, K. N. 2012 SABER: a computational method for identifying active sites for new reactions. *Protein Sci.*, **21**(5), 697–706. (doi: 10.1002/pro.2055)

Oberhardt, M. A., Puchaka, J., Martins dos Santos, V. A. P. & Papin, J. A. 2011 Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis. *PLoS Comput. Biol.*, **7**(3), e1001 116. (doi: 10.1371/journal.pcbi.1001116)

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. & Hutchison, G. R. 2011 Open Babel: An open chemical toolbox. *J. Cheminform.*, **3**(1), 33. (doi: 10.1186/1758-2946-3-33)

O'Boyle, N. M., Holliday, G. L., Almonacid, D. E. & Mitchell, J. B. O. 2007 Using reaction mechanism to measure enzyme similarity. *J. Mol. Biol.*, **368**(5), 1484–99. (doi: 10.1016/j.jmb.2007.02.065)

O'Brien, P. J. 2006 Catalytic promiscuity and the divergent evolution of DNA repair enzymes. *Chem. Rev.*, **106**(2), 720–52. (doi: 10.1021/cr040481v)

O'Brien, P. J. & Herschlag, D. 1999 Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.*, **6**(4), R91–R105. (doi: 10.1016/S1074-5521(99)80033-7)

Oh, M., Yamada, T., Hattori, M., Goto, S. & Kanehisa, M. 2007 Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. *J. Chem. Inf. Model.*, **47**(4), 1702–12. (doi: 10.1021/ci700006f)

Okazaki, S., Suzuki, A., Mizushima, T., Kawano, T., Komeda, H., Asano, Y. & Yamane, T. 2009 The novel structure of a pyridoxal 5'-phosphate-dependent fold-type I racemase, alpha-amino-epsilon-caprolactam racemase from Achromobacter obae. *Biochemistry*, **48**(5), 941–50. (doi: 10.1021/bi801574p)

Oksanen, J. 2011 Multivariate Analysis of Ecological Communities in R : vegan tutorial.

Omelchenko, M. V., Galperin, M. Y., Wolf, Y. I. & Koonin, E. V. 2010 Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biol. Direct*, **5**, 31. (doi: 10.1186/1745-6150-5-31)

Ott, M. a. & Vriend, G. 2006 Correcting ligands, metabolites, and pathways. *BMC Bioinformatics*, **7**, 517. (doi: 10.1186/1471-2105-7-517)

Palani, K., Burley, S. K. & Swaminathan, S. 2013 Structure of alanine racemase from Oenococcus oeni with bound pyridoxal 5'-phosphate. *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun.*, **69**(Pt 1), 15–9. (doi: 10.1107/S1744309112047276)

Pandya, C., Brown, S., Pieper, U., Sali, A., Dunaway-Mariano, D., Babbitt, P. C., Xia, Y. & Allen, K. N. 2013 Consequences of domain insertion on sequence-structure divergence in a superfold. *Proc. Natl. Acad. Sci. U. S. A.*, **110**(36), E3381–7. (doi: 10.1073/pnas.1305519110)

Paradis, E. 2012 *Analysis of Phylogenetics and Evolution with R.* Springer.

Paradis, E., Claude, J. & Strimmer, K. 2004 APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**(2), 289–90. (doi: 10.1093/bioinformatics/btg412)

Patthy, L. 1999 Protein-coding genes. In *Protein evol.*, pp. 1–11. London: Wiley-Blackwell, 2nd edn.

Pauling, L. 1946 Molecular Architecture and Biological Reactions. *Chem. Eng. News*, **24**(10), 1375–1377. (doi: 10.1021/cen-v024n010.p1375)

Payen, A. & Persoz, J. 1833 Mémoire sur la Diastase, les principaux Produits de ses Réactions, et leurs applications aux arts industriels. *Ann. chim. phys*, **53**, 73–92.

Penny, D. & Hendy, M. D. 1985 The Use of Tree Comparison Metrics. *Syst. Biol.*, **34**(1), 75–82. (doi: 10.1093/sysbio/34.1.75)

Peralta-Yahya, P. P., Zhang, F., del Cardayre, S. B. & Keasling, J. D. 2012 Microbial engineering for the production of advanced biofuels. *Nature*, **488**(7411), 320–8. (doi: 10.1038/nature11478)

Petsko, G. A., Kenyon, G. L., Gerlt, J. A., Ringe, D. & Kozarich, J. W. 1993 On the origin of enzymatic species. *Trends Biochem. Sci.*, **18**(10), 372–6. (doi: 10.1016/0968-0004(93)90091-Z)

Pieper, D. & Stadler-Fritzsche, K. 1993 Metabolism of 2-chloro-4-methylphenoxyacetate by Alcaligenes eutrophus JMP 134. *Arch. Microbiol.*, **160**(3), 169–178. (doi: 10.1007/BF00249121)

Plata, G., Fuhrer, T., Hsiao, T.-L., Sauer, U. & Vitkup, D. 2012 Global probabilistic annotation of metabolic networks enables enzyme discovery. *Nat. Chem. Biol.*, **8**(10), 848–54. (doi: 10.1038/nchembio.1063)

Pouliot, Y. & Karp, P. 2007 A survey of orphan enzyme activities. *BMC Bioinformatics*, **8**(1), 244. (doi: 10.1186/1471-2105-8-244)

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G. *et al.* 2012 The Pfam protein families database. *Nucleic Acids Res.*, **40**(Database issue), D290–301. (doi: 10.1093/nar/gkr1065)

Quester, S. & Schomburg, D. 2011 EnzymeDetector: an integrated enzyme function prediction tool and database. *BMC Bioinformatics*, **12**(1), 376. (doi: 10.1186/1471-2105-12-376)

R Core Team 2012 *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K. *et al.* 2013 A large-scale evaluation of computational protein function prediction. *Nat. Methods*, **10**(3), 221–7. (doi: 10.1038/nmeth.2340)

Rahman, S. A., Bashton, M., Holliday, G. L., Schrader, R. & Thornton, J. M. 2009 Small Molecule Subgraph Detector (SMSD) toolkit. *J. Cheminform.*, **1**(1), 12. (doi: 10.1186/1758-2946-1-12)

Rahman, S. A., Cuesta, S. M., Furnham, N., Holliday, G. L. & Thornton, J. M. 2014 EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nat. Methods*, **11**(2), 171–174. (doi: 10.1038/nmeth.2803)

Rawlings, N. D., Waller, M., Barrett, A. J. & Bateman, A. 2014 MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.*, **42**(Database issue), D503–9. (doi: 10.1093/nar/gkt953)

Reitz, M., Sacher, O., Tarkhov, A., Trumbach, D. & Gasteiger, J. 2004 Enabling the exploration of biochemical pathways. *Org. Biomol. Chem.*, **2**(22), 3226–37. (doi: 10.1039/B410949J)

Remmert, M., Biegert, A., Hauser, A. & Söding, J. 2012 HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**(2), 173–5. (doi: 10.1038/nmeth.1818)

Roberts, D. W. 2012 *labdsv: Ordination and Multivariate Analysis for Ecology.*

Rodionova, I. A., Scott, D. A., Grishin, N. V., Osterman, A. L. & Rodionov, D. A. 2012 Tagaturonate-fructuronate epimerase UxaE, a novel enzyme in the hexuronate catabolic network in Thermotoga maritima. *Environ. Microbiol.*, **14**(11), 2920–34. (doi: 10.1111/j.1462-2920.2012.02856.x)

Rose, J. R. & Gasteiger, J. 1994 HORACE: An automatic system for the hierarchical classification of chemical reactions. *J. Chem. Inf. Model.*, **34**(1), 74–90. (doi: 10.1021/ci00017a010)

Rost, B. 2002 Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**(2), 595–608. (doi: 10.1016/S0022-2836(02)00016-5)

Rother, K., Hoffmann, S., Bulik, S., Hoppe, A., Gasteiger, J. & Holzhütter, H.-G. 2010 IGERS: inferring Gibbs energy changes of biochemical reactions from reaction similarities. *Biophys. J.*, **98**(11), 2478–86. (doi: 10.1016/j.bpj.2010.02.052)

Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L. J. M., Guo, Y., Hériché, J.-K., Hu, Y., Kristiansen, K. *et al.* 2008 TreeFam: 2008 Update. *Nucleic Acids Res.*, **36**(Database issue), D735–40. (doi: 10.1093/nar/gkm1005)

Sacher, O., Reitz, M. & Gasteiger, J. 2009 Investigations of enzyme-catalyzed reactions based on physicochemical descriptors applied to hydrolases. *J. Chem. Inf. Model.*, **49**(6), 1525–34. (doi: 10.1021/ci800277f)

Saha, R., Chowdhury, A. & Maranas, C. D. 2014 Recent advances in the reconstruction of metabolic models and integration of omics data. *Curr. Opin. Biotechnol.*, **29C**, 39–45. (doi: 10.1016/j.copbio.2014.02.011)

Savignon, T., Costa, E., Tenorio, F., Manhães, A. C. & Barradas, P. C. 2012 Prenatal hypoxic-ischemic insult changes the distribution and number of NADPH-diaphorase cells in the cerebellum. *PLoS One*, **7**(4), e35 786. (doi: 10.1371/journal.pone.0035786)

Schmidt, S. 2003 Metabolites: a helping hand for pathway evolution? *Trends Biochem. Sci.*, **28**(6), 336–341. (doi: 10.1016/S0968-0004(03)00114-2)

Schnell, B., Faber, K. & Kroutil, W. 2003 Enzymatic Racemisation and its Application to Synthetic Biotransformations. *Adv. Synth. Catal.*, **345**(67), 653–666. (doi: 10.1002/adsc.200303009)

Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. 2009 Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLoS Comput Biol*, **5**(12), e1000 605. (doi: 10.1371/journal.pcbi.1000605)

Schnoes, A. M., Ream, D. C., Thorman, A. W., Babbitt, P. C. & Friedberg, I. 2013 Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS Comput. Biol.*, **9**(5), e1003 063. (doi: 10.1371/journal.pcbi.1003063)

Schomburg, I., Chang, A., Placzek, S., Söhngen, C., Rother, M., Lang, M., Munaretto, C., Ulas, S., Stelzer, M. *et al.* 2013*a* BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res.*, **41**(Database issue), D764–72. (doi: 10.1093/nar/gks1049)

Schomburg, I., Chang, A. & Schomburg, D. 2014 Standardization in enzymology – Data integration in the worlds enzyme information system BRENDA. *Perspect. Sci.*, **1**(1-6), 15–23. (doi: 10.1016/j.pisc.2014.02.002)

Schomburg, K. T., Wetzer, L. & Rarey, M. 2013*b* Interactive design of generic chemical patterns. *Drug Discov. Today*, **18**(13-14), 651–8. (doi: 10.1016/j.drudis.2013.02.001)

Schulenburg, C. & Miller, B. G. 2014 Enzyme recruitment and its role in metabolic expansion. *Biochemistry*, **53**(5), 836–45. (doi: 10.1021/bi401667f)

Scornavacca, C., Zickmann, F. & Huson, D. H. 2011 Tanglegrams for rooted phylogenetic trees and networks. *Bioinformatics*, **27**(13), i248–56. (doi: 10.1093/bioinformatics/btr210)

Seebeck, F. P. & Hilvert, D. 2003 Conversion of a PLP-dependent racemase into an aldolase by a single active site mutation. *J. Am. Chem. Soc.*, **125**(34), 10 158–9. (doi: 10.1021/ja036707d)

Shaw, R., Debsarma, S. & Kundu, S. 2012 An algorithm for removing stoichiometric discrepancies in biochemical reaction databases. *Curr. Sci.*, **103**(11), 1328–1334.

Shearer, A. G., Altman, T. & Rhee, C. D. 2014 Finding sequences for over 270 orphan enzymes. *PLoS One*, **9**(5), e97 250. (doi: 10.1371/journal.pone.0097250)

Shi, J., Blundell, T. L. & Mizuguchi, K. 2001 FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**(1), 243–57. (doi: 10.1006/jmbi.2001.4762)

Sillitoe, I., Cuff, A. L., Dessailly, B. H., Dawson, N. L., Furnham, N., Lee, D., Lees, J. G., Lewis, T. E., Studer, R. a. *et al.* 2013 New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.*, **41**(Database issue), D490–8. (doi: 10.1093/nar/gks1211)

Silveira, S. D. A., de Melo-Minardi, R. C., da Silveira, C. H., Santoro, M. M. & Meira, W. 2014 ENZYMAP: exploiting protein annotation for modeling and predicting EC number changes in UniProt/Swiss-Prot. *PLoS One*, **9**(2), e89 162. (doi: 10.1371/journal.pone.0089162)

Silverman, R. B. 2002 *The Organic Chemistry of Enzyme-catalyzed Reactions.* Academic Press.

Simmons, E. S. 1991 The grammar of Markush structure searching: vocabulary vs. syntax. *J. Chem. Inf. Model.*, **31**(1), 45–53. (doi: 10.1021/ci00001a007)

Singh, S., Phillips, G. N. & Thorson, J. S. 2012 The structural biology of enzymes involved in natural product glycosylation. *Nat. Prod. Rep.*, **29**(10), 1201–37. (doi: 10.1039/c2np20039b)

Smith, A. A. T., Belda, E., Viari, A., Medigue, C. & Vallenet, D. 2012 The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. *PLoS Comput. Biol.*, **8**(5), e1002 540. (doi: 10.1371/journal.pcbi.1002540)

Socha, R. D. & Tokuriki, N. 2013 Modulating protein stability - directed evolution strategies for improved protein function. *FEBS J.*, **280**(22), 5582–95. (doi: 10.1111/febs.12354)

Song, L., Kalyanaraman, C., Fedorov, A. A., Fedorov, E. V., Glasner, M. E., Brown, S., Imker, H. J., Babbitt, P. C., Almo, S. C. *et al.* 2007 Prediction and assignment of function for a divergent N-succinyl amino acid racemase. *Nat. Chem. Biol.*, **3**(8), 486–91. (doi: 10.1038/nchembio.2007.11)

Sorokina, M., Stam, M., Médigue, C., Lespinet, O. & Vallenet, D. 2014 Profiling the orphan enzymes. *Biol. Direct*, **9**(1), 10. (doi: 10.1186/1745-6150-9-10)

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. & Willighagen, E. 2003 The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.*, **43**(2), 493–500. (doi: 10.1021/ci025584y)

Stitt, M. & Gibon, Y. 2014 Why measure enzyme activities in the era of systems biology? *Trends Plant Sci.*, **19**(4), 256–265. (doi: 10.1016/j.tplants.2013.11.003)

Sumner, J. 1926 The isolation and crystallization of the enzyme urease. *J. Biol. Chem.*, **69**, 435–441.

Tamuri, A. U. & Laskowski, R. A. 2010 ArchSchema: a tool for interactive graphing of related Pfam domain architectures. *Bioinformatics*, **26**(9), 1260–1. (doi: 10.1093/bioinformatics/btq119)

Tanner, M. E. 2008 Transient oxidation as a mechanistic strategy in enzymatic catalysis. *Curr. Opin. Chem. Biol.*, **12**(5), 532–8. (doi: 10.1016/j.cbpa.2008.06.016)

Terao, Y., Miyamoto, K. & Ohta, H. 2006 Introduction of single mutation changes arylmalonate decarboxylase to racemase. *Chem. Commun. (Camb).*, (34), 3600–2. (doi: 10.1039/b607211a)

The Uniprot Consortium 2013 Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**(Database issue), D43–7. (doi: 10.1093/nar/gks1068)

Tian, W. & Skolnick, J. 2003 How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.*, **333**(4), 863–82. (doi: 10.1016/j.jmb.2003.08.057)

Tipton, K. & Boyce, S. 2000 History of the enzyme nomenclature system. *Bioinformatics*, **16**(1), 34–40. (doi: 10.1093/bioinformatics/16.1.34)

Tipton, K. F., Armstrong, R. N., Bakker, B. M., Bairoch, A., Cornish-Bowden, A., Halling, P. J., Hofmeyr, J.-H., Leyh, T. S., Kettner, C. *et al.* 2014 Standards for Reporting Enzyme Data: The STRENDA Consortium: What it aims to do and why it should be helpful. *Perspect. Sci.*, **1**(1-6), 131–137. (doi: 10.1016/j.pisc.2014.02.012)

Todd, A. E., Orengo, C. A. & Thornton, J. M. 2001 Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.*, **307**(4), 1113–43. (doi: 10.1006/jmbi.2001.4513)

Todd, A. E., Orengo, C. A. & Thornton, J. M. 2002 Plasticity of enzyme active sites. *Trends Biochem. Sci.*, **27**(8), 419–26. (doi: 10.1016/S0968-0004(02)02158-8)

Tohsato, Y., Matsuda, H. & Hashimoto, A. 2000 A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 376–83.

Tohsato, Y. & Nishimura, Y. 2009 Reaction Similarities Focusing Substructure Changes of Chemical Compounds and Metabolic Pathway Alignments. *Inf. Media Technol.*, **4**(2), 390 – 399. (doi: 10.11185/imt.4.390)

Triviño, J. C. & Pazos, F. 2010 Quantitative global studies of reactomes and metabolomes using a vectorial representation of reactions and chemical compounds. *BMC Syst. Biol.*, **4**, 46. (doi: 10.1186/1752-0509-4-46)

Uberto, R. & Moomaw, E. W. 2013 Protein similarity networks reveal relationships among sequence, structure, and function within the Cupin superfamily. *PLoS One*, **8**(9), e74 477. (doi: 10.1371/journal.pone.0074477)

Vacca, R. A., Giannattasio, S., Graber, R., Sandmeier, E., Marra, E. & Christen, P. 1997 Active-site Arg - Lys substitutions alter reaction and substrate specificity of aspartate aminotransferase. *J. Biol. Chem.*, **272**(35), 21 932–7. (doi: 10.1074/jbc.272.35.21932)

Vamvaca, K., Vögeli, B., Kast, P., Pervushin, K. & Hilvert, D. 2004 An enzymatic molten globule: efficient coupling of folding and catalysis. *Proc. Natl. Acad. Sci. U. S. A.*, **101**(35), 12 860–4. (doi: 10.1073/pnas.0404109101)

Vick, J. E. & Gerlt, J. A. 2007 Evolutionary potential of (β/α)8-barrels: stepwise evolution of a new reaction in the enolase superfamily. *Biochemistry*, **46**(50), 14 589–97. (doi: 10.1021/bi7019063)

Vongvilai, P., Linder, M., Sakulsombat, M., Svedendahl Humble, M., Berglund, P., Brinck, T. & Ramström, O. 2011 Racemase activity of B. cepacia lipase leads to dual-function asymmetric dynamic kinetic resolution of α-aminonitriles. *Angew. Chem. Int. Ed. Engl.*, **50**(29), 6592–5. (doi: 10.1002/anie.201007373)

Voordeckers, K., Brown, C. A., Vanneste, K., van der Zande, E., Voet, A., Maere, S. & Verstrepen, K. J. 2012 Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication. *PLoS Biol.*, **10**(12), e1001 446. (doi: 10.1371/journal.pbio.1001446)

Warr, W. A. 2011 Representation of chemical structures. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **1**(4), 557–579. (doi: 10.1002/wcms.36)

Warr, W. A. 2014 A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol. Inform.*, **33**, 469–476. (doi: 10.1002/minf.201400052)

Wehrens, R. & Buydens, L. M. C. 2007 Self- and Super-organizing Maps in R : The kohonen. *J. Stat. Softw.*, **21**(5), 1–19.

White, D. & Gramacy, R. B. 2012 maptree: Mapping, pruning, and graphing tree models.

Wittig, U., Rey, M., Kania, R., Bittkowski, M., Shi, L., Golebiewski, M., Weidemann, A., Müller, W. & Rojas, I. 2014 Challenges for an enzymatic reaction kinetics database. *FEBS J.*, **281**(2), 572–82. (doi: 10.1111/febs.12562)

Yamada, T., Letunic, I., Okuda, S., Kanehisa, M. & Bork, P. 2011 iPath2.0: interactive pathway explorer. *Nucleic Acids Res.*, **39 Suppl 2**(May), W412–5. (doi: 10.1093/nar/gkr313)

Yamada, T., Waller, A. S., Raes, J., Zelezniak, A., Perchat, N., Perret, A., Salanoubat, M., Patil, K. R., Weissenbach, J. *et al.* 2012 Prediction and identification of sequences coding for orphan enzymes using genomic and metagenomic neighbours. *Mol. Syst. Biol.*, **8**(581), 581. (doi: 10.1038/msb.2012.13)

Yamanishi, Y., Hattori, M., Kotera, M., Goto, S. & Kanehisa, M. 2009 E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. *Bioinformatics*, **25**, i179–i186. (doi: 10.1093/bioinformatics/btp223)

Yamanishi, Y., Mihara, H., Osaki, M., Muramatsu, H., Esaki, N., Sato, T., Hizukuri, Y., Goto, S. & Kanehisa, M. 2007 Prediction of missing enzyme genes in a bacterial metabolic network. Reconstruction of the lysine-degradation pathway of Pseudomonas aeruginosa. *FEBS J.*, **274**(9), 2262–73. (doi: 10.1111/j.1742-4658.2007.05763.x)

Yamanishi, Y., Vert, J.-P. & Kanehisa, M. 2005 Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinformatics*, **21 Suppl 1**(suppl 1), i468–77. (doi: 10.1093/bioinformatics/bti1012)

Young, E., Lee, S.-M. & Alper, H. 2010 Optimizing pentose utilization in yeast: the need for novel tools and approaches. *Biotechnol. Biofuels*, **3**(1), 24. (doi: 10.1186/1754-6834-3-24)

Yu, C., Zavaljevski, N., Desai, V. & Reifman, J. 2009 Genome-wide enzyme annotation with precision control: catalytic families (CatFam) databases. *Proteins*, **74**(2), 449–60. (doi: 10.1002/prot.22167)

Yu, X., Doroghazi, J. R., Janga, S. C., Zhang, J. K., Circello, B., Griffin, B. M., Labeda, D. P. & Metcalf, W. W. 2013 Diversity and abundance of phosphonate biosynthetic genes in nature. *Proc. Natl. Acad. Sci. U. S. A.*, **110**(51), 20 759–64. (doi: 10.1073/pnas.1315107110)

Zass, E. 1990 A user's view of chemical reaction information sources. *J. Chem. Inf. Model.*, **30**(4), 360–372. (doi: 10.1021/ci00068a004)

Zhao, S., Kumar, R., Sakai, A., Vetting, M. W., Wood, B. M., Brown, S., Bonanno, J. B., Hillerich, B. S., Seidel, R. D. *et al.* 2013 Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature*, **502**(7473), 698–702. (doi: 10.1038/nature12576)