# To study the evolutionary origin of specificity in Triose-phosphate isomerase (TIM)

# And

# To extract protein-ligand interaction features using Canonical Correspondence Analysis (CCA)

**Ketika Garg**

*A dissertation submitted for the partial fulfilment of BS-MS dual degree in Science*

**Indian Institute of Science Education and Research, Mohali**
**April, 2017**

# Certificate of Examination

This is to certify that the dissertation titled “To study the evolutionary origin of specificity in Triose-Phosphate isomerase And To extract protein-ligand interaction features using Canonical Correspondence Analysis”submitted by Ms. Ketika Garg (Reg. No. MS12062) for the partial fulfilment of BS-MS dual degree programme of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr. Kuljeet Sandhu     Dr. Rajesh Ramachandran     Dr. Shashi Bhushan Pandit
(Supervisor)

Dated: 21 April 2017

# Declaration

The work presented in this dissertation has been carried out by me under the guidance of Dr. Shashi Bhushan Pandit at the Indian Institute of Science Education and Research Mohali.

This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.

Ketika Garg
(Candidate)

Dated: 21 April 2017

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Shashi Bhushan Pandit
(Supervisor)

# Acknowledgements

I would like to thank my supervisor, Dr. Shashi Bhushan Pandit  for suggesting interesting problems to work upon, helping me to understand them and develop effective approaches to solve them.

I would also like to thank my lab members, Santosh Behera, Preeti Choudhary, Rivi Verma, and Swapnil Tichkule for supporting my work and readily helping me to solve problems pertinent to the project. A special thanks to Ken, who allowed me to run my relentless analyses on his system.

# List of Figures

# List of Tables

# Contents

PART 1

PART 2

# Abstract

Recent studies have shown that enzymes can catalyze alternate reaction or substrate/s apart from their physiologically relevant activity. This ability of enzyme is referred to as enzyme promiscuity. Usually, promiscuous activities have low catalytic efficiency and specificity. However, these can become important under any genotype/environment perturbations. Jensen has hypothesised that in ancestral enzymes showed broad specificity (generalist) and these become specialized during evolution. Based on this, we studied whether ancestral enzymes exhibit low catalytic efficiency or have weak substrate affinity. For this, we used core glycolytic enzyme Trios-phosphate isomerase (TIM). We generated phylogenetic tree of TIM enzymes and overlay with known experimental kinetic parameters. We observed that catalytic efficiencies are similar in enzymes from both ancestral and recently evolved enzymes. However, binding affinity of ancestral enzymes is weaker in comparison to modern enzymes.

In the second project, we have used statistical methods to identify recurring patterns in protein sequences and compounds that can assist in understanding ligand-protein interactions. We have used canonical correspondence analysis (CCA) method with proteins represented as 6-mers string kernels and ligands expressed as atomic signatures. Based on preliminary analysis of 92 ligands, it can be suggested that CCA could be helpful in identifying important features of protein-ligand interactions. Further, this could be used in prediction of ligand binding sites.

# PART I

*To study the evolutionary origin of specificity in Triose-phosphate isomerase (TIM)*

# 1 Introduction

Enzymes are efficient catalysts that accelerate biochemical reactions up to rates at which biological processes are sustainable in an organism. Most of early studies on characterization of enzymes considered their function in isolation or in a limited set of interconnected metabolic reactions. However, the cataloguing of enzymes and everything known about them led to the representation of metabolism into metabolic pathways (*Reitz et al., 2004*). These metabolic pathways have been used extensively in understanding metabolic capability of an organism and have also provided possible alternate pathways when certain reactions are blocked due to mutation or deletion of enzyme/s (*McCloskey et al., 2013; Stobbe et al., 2014; Simeonidis et al., 2015*). The metabolic pathways are documented in databases such as KEGG, which are manually curated and periodically updated to keep up with discovery new metabolic reactions or enzyme functions (*Kanehisa et al., 2012*).

Traditionally, enzymes are described as efficient (acceleration of reaction) and specific (to its substrate) biocatalyst under notion of 'one enzyme-one substrate-one reaction'. However, recent studies have shown that enzymes can also catalyze alternate reaction or substrate/s apart from their physiologically relevant activity. This ability of enzyme to catalyze more than one biochemical reaction is referred to as enzyme promiscuity. This discovery has radically changed our understanding of enzymes and has very board implications from the evolution of enzymes (*Copley, 2003; Khersonsky & Tawfik, 2010; O'Brien & Herschlag, 1999*) to biotechnological applications (*Nobeli et al., 2009*).

## 1.1 Enzyme promiscuity and specificity

As discussed above, promiscuity refers to the ability of enzyme to catalyze alternate chemical transformations, usually, exhibited through same active sites. These adventitious secondary or promiscuous reactions are generally orders of magnitude less efficient than their evolved activities. For example, Malonate semialdehyde decarboxylase catalyzes decarboxylation (malonate semialdehyde) 3.5 fold more efficient than hydration reaction (2-oxo-3-pentynoate) (*Khersonsky & Tawfik, 2010*). These low levels of promiscuous activity, usually undetectable, can become important if substrate/enzyme concentration changes due to genotype/environmental perturbations. It has been suggested such low levels of reactions are part of underground metabolism in cell and confers robustness to metabolic networks (*D' Ari et al., 2008*).

Initial studies on promiscuous enzymes were mostly from serendipitous discovery of alternate function. With the availability of genome sequences, studies involving gene deletions in an organism, genome-scale metabolic reconstruction, high-throughput screening of enzyme activities and experimental elucidation of enzyme structures has increased identification of new promiscuous enzymes (*Nobeli et al., 2009; Khersonsky and Tawfik, 2010*). A recent study has shown that enzyme promiscuity is wide spread, for example *Escherichia coli* have 37% of its enzymes identified as promiscuous, which catalyze 65% of reactions (*Nam et al., 2012*).

It has been suggested that promiscuous activities can serve as a reservoir of novel catalytic activities and could be an important driving force in evolutionary divergence of enzymes families (*Nobeli et al., 2009; Khersonsky and Tawfik, 2010; Janssen et al., 2005*). One recent study has suggested possible role of promiscuity in the evolution of herbicide degradation pathway. Atrazine as an herbicide was effective in first years of its use. In subsequent years, it was discovered that atrazine is undergoing bacterial degradation, which led to discovery of atrazine degradation pathway in *Pseudomonas sp.* The first enzyme in degradation pathway is atrazine chlorohydrolase. Interestingly, this enzyme is 98% identical to melamine deaminase enzyme involved in melamine degrdation pathway (*Wackett, 2009*).

### 1.1.1 Mechanistic aspect of promiscuity

Many recent experimental studies suggested following insights into understanding of mechanistic and structural aspects of promiscuity: a. Conformational diversity of active sites can accommodate alternate substrates; b. existence of several subsites within active sites; c. differences in protonation states of catalytic residues between native and promiscuous function; d. metal substitutions following cofactor ambiguity and, e. formation of accidental hydrogen bonds, which can buffer opposing charges between the substrate and active site residues, or take on a/an acidic/basic/ nucleophilic role and catalyze promiscuous activities (**Figure 1**).

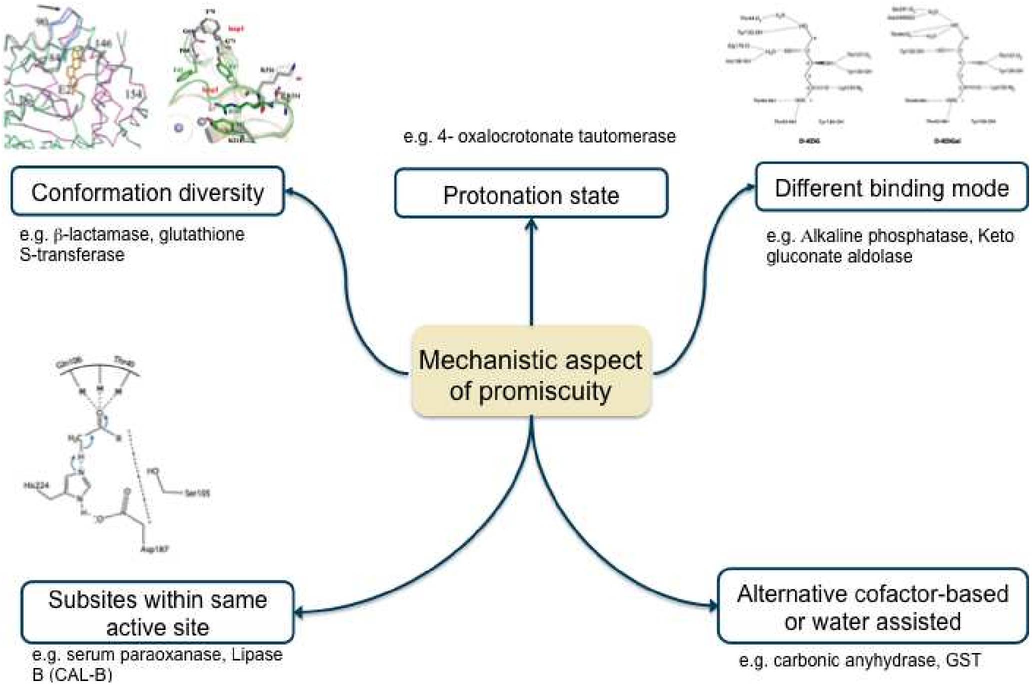

Conformation diversity
e.g. β-lactamase, glutathione S-transferase
e.g. 4- oxalocrotonate tautomerase
Protonation state
Different binding mode
e.g. Alkaline phosphatase, Keto gluconate aldolase
Mechanistic aspect of promiscuity
Subsites within same active site
e.g. serum paraoxanase, Lipase B (CAL-B)
Alternative cofactor-based or water assisted
e.g. carbonic anyhydrase, GST

*Figure 1. Summary of putative mechanism of enzyme promiscuity*

Of these, conformational flexibility can possibly be an important factor, as enzymes can exist in a continuum of conformational states. The primary state is the one, which engages with the native ligand ($P_N$) and other different states are generated through several structural variations, for example, in side chain rotamers, active-site loop rearrangements and fold transitions (Figure 2). Suppose one of the conformations ($P_4$) could accommodate a promiscuous ligand, and through selection the equilibrium could alter to increase the amount of this previously-minor conformation, which would lead up to the development of promiscuous activity, without substantially affecting the native function (*Khersonsky and Tawfik, 2010*; *Tokuriki and Tawfik, 2009*).

In some cases, one active site can mediate both the native and promiscuous binding, with one key feature shared by all the activities. For example, (**Figure 1**) a mammalian lactonase, serum paraoxonase, has additional promiscuous esterase and phosphotriesterase activities. While all these three activities (the native, lactone hydrolysis and, the promiscuous esterases and phosphotriesterase) share one key feature, the coordination of the phosphoryl/carbonyl oxygen to the calcium present in the active site, the hydrolysis of lactones and esterase (His115-His134) and, phosphotriesterase are mediated by different sets of residues (*Khersonsky and Tawfik, 2010; Tawfik and Khersonsky, 2006*).

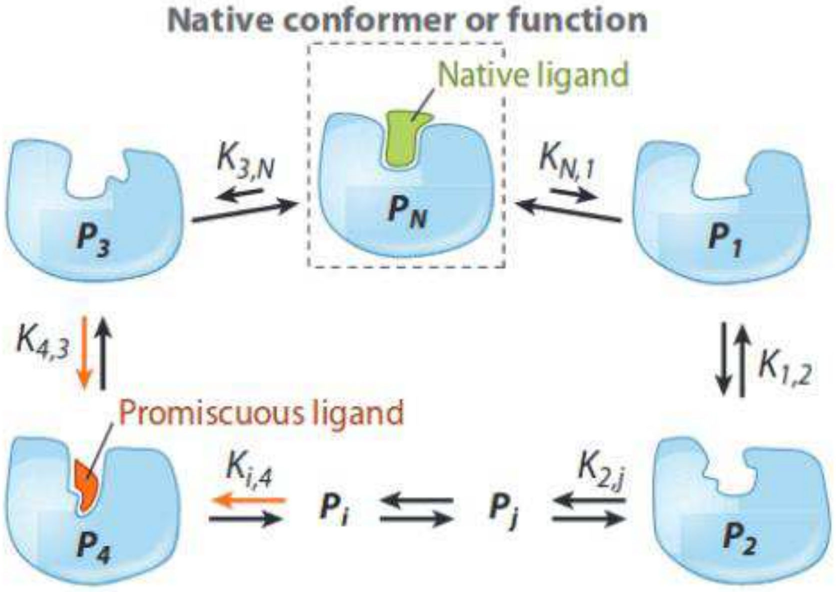


*Figure 2: Role of conformational flexibility in promiscuity (Adopted from* Tokuriki and Tawfik, 2009*)*

### 1.1.2 Quantifying degree of promiscuity

The experimental studies on promiscuous nature of enzymes, mostly, involve identifying alternate substrate/s or reactions. However, understanding the level or degree of promiscuous nature can assist in better characterization and classification of enzymes. Moreover, it can provide insights into evolutionary divergence of enzymes. Such studies can also help in understanding the evolution of specificity in enzymes. It has suggested that one can use diversity of substrates or chemical bonds broken or made during different reactions as a basis to quantify degree of promiscuity (*Khersonsky and Tawfik, 2010*).

An alternate approach could utilize experimentally determined rate constant associated with enzymatic reactions to characterize degree of promiscuity. Since, promiscuous reactions of an enzyme have been shown to have different kinetic rates than the native reactions, with the catalytic efficiencies of the former being lower.

#### 1.1.2.1 Michaelis menten equation

Michaelis Menten model (**Figure 3**) assumes that an enzyme (*E*) directly interacts with the substrate (S), resulting in an intermediate complex (ES). This leads to thermodynamic equilibrium, where the Michaelis constant ($K_m$) is a measure of the substrate concentration required for an effective catalysis (in the reactions following Michaelis Menten kinetics (Eq. 2)), which can often be translated into a measure of affinity of the enzyme for a particular substrate (Johnson and Goody, 2011). The $K_{cat}$ is the measure of the catalytic production rate of the product, considering the enzyme becomes saturated (Eq. 3).

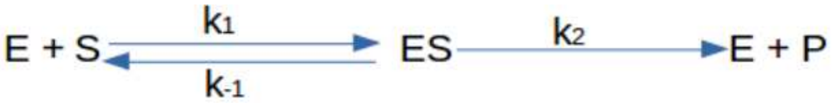


*Figure 3: Michaelis menten model of enzyme kinetics*

$$K_m = k_2 + k_{-1} / k_1 \qquad \textbf{Eq. 1}$$

$$V=V_{max}[S]/(K_m + [S]) \qquad \textbf{Eq. 2}$$

$$V_{max} = K_{cat}[E]_t$$

$$V=K_{cat}[E]_t\ [S]/(K_m + [S]) \qquad \textbf{Eq. 3}$$

The quantity $K_{cat}/Km$, specificity constant, is usually employed as a metric to objectively understand catalytic efficiency. Especially, it states the rate of catalysis as a function of substrate affinity. Hence, a high $K_{cat}/K_m$ should be a characteristic of a perfect enzyme. These measures can substantiate the differences between native and promiscuous functions. Intuitively, it could be considered that $K_m$ will be higher for promiscuous substrates and will exhibit lower $K_{cat}$ values. These will suggest importance of spatial orientation of substrate with respect to active site. Moreover, interactions forces which drive native substrate binding differ from that of promiscuous substrate binding; the former is mediated through enthalpy, for example hydrogen bonds, while the latter is determined by hydrophobic interactions and entropy (*Khersonsky and Tawfik, 2010*).

### 1.1.3 Jensen’s hypothesis

The possible role of promiscuity in enzyme evolution was hypothesized in seminal work by Jensen, which stated that ancestral enzymes probably possessed broad specificities, albeit with low catalytic efficiency and in the process of evolution through duplication, mutation and horizontal gene transfer led to diversification of gene families and apparently refined ancestral enzymes to become specific and catalytically efficient (*Jensen, 1976*). Thus, most modern enzymes are assumed to be “specialist” having evolved to specialize in one reaction on a unique primary substrate in an organism to increase their metabolic efficiency. On the contrary ancestral enzymes are referred as “generalist” having promiscuous characteristics. The central hypothesis of Jensen is summarized in **Figure 4**. Some modern enzymes exhibit promiscuity because these could confer fitness benefit to the organism under new selective pressures with promiscuous enzymes serving as starting point in the emergence of new enzyme functions and/or divergence of enzyme families. Moreover, it has been suggested that this 'floppiness' in enzymatic function has played an important role in the evolution of superfamilies of enzymes, transcriptional regulators and receptors (*Copley, 2015*).

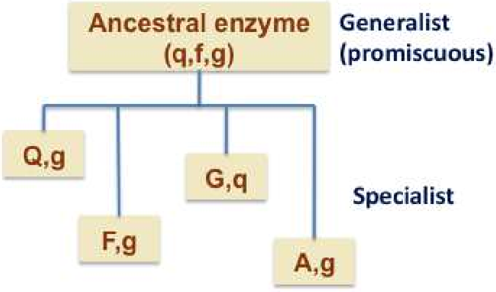


*Figure 4: Summary of Jensen’s hypothesis of enzyme evolution*

Based on Jensen’s hypothesis, one can expect that enzymes from phylogenetically ancestral organisms will harbor enzymes, which tend to be promiscuous compare to recently evolved enzymes. To explore this possibility as a result of Jensen’s hypothesis, we studied one of the core glycolytic enzymes, Triose phosphate isomerase (TIM) using catalytic efficiency ($K_{cat}/K_m$) as a measure of degree of promiscuity.

## 1.2 Triose phosphate isomerase (TIM)

TIM is one the most extensively studied enzymes, as it plays a central role in glycolysis (*Wierenga, et al., 2010*). It catalyzes the reversible inter-conversion of d-glyceraldehyde 3-phosphate (d-GAP) and dihydroxyacetone phosphate (DHAP) (see **Figure 5**). The reaction is characterized by an ene-diol intermediate, and two sequential proton exchanges. Two residues Glutamate (Glu) and Histidine (His) play an important role in catalysis, with Glu acting as an electrophile and His as a nucleophile during the isomerase reaction. It has been argued that TIM is catalytically perfect enzyme, where the rate of catalysis is of the same order as rates of diffusion (*Sharma and Guptasarma, 2015*). It has been suggested that TIM reaction center is so perfectly designed that a conservative 'Glu-to-Asp' mutation, which conserves the catalytically important carboxyl group and doesn't make any significant changes in the structure of the enzyme, except shifting the carboxyl group by about 0.1 nm away from the substrate. This leads to a 1000-fold reduction in the enzymatic activity (*Davenport et al., 1993*). The four catalytic sites *viz.* Asn11, Lys13, His95, and Glu167 along with loop 6 and 7 (numbers are based on *E. coli* sequence) are highly conserved across various organisms (*Wierenga, et al., 2010*).

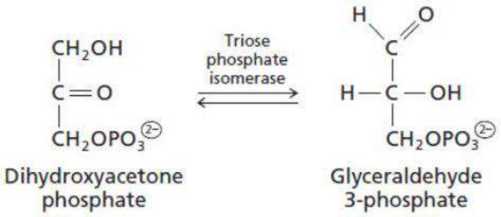

CH2OH
C=O
CH2OPO3 2−
Dihydroxyacetone phosphate
Triose phosphate isomerase
H
O
C
H—C—OH
CH2OPO3 2−
Glyceraldehyde 3-phosphate

*Figure 5: Reaction catalyzed by TIM (Adopted from* Davenport et al., 1993*)*

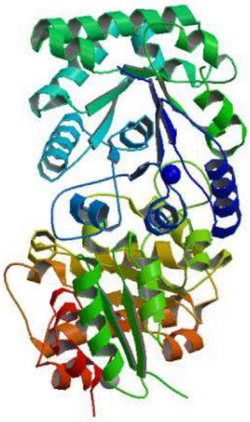


*Figure 6: Tertiary structure of human TIM (pdb id: 4zvj)*

TIM structure is characterized by classic alpha-beta barrel, which is an archetypal example of the 'TIM barrel fold', the most common fold amongst protein catalysts (*Voet et al., 2006*). TIM consists of alternating beta and alpha helices such that beta strands form core of the barrel (**Figure 6**). TIM was selected for this study, in the light of its prevalence among all the three domains of life, owing to its central role in glycolysis.

# 2 Methodology

## 2.1 Database searching for TIM orthologs

Since TIM is plays central role in glycolysis and is prevalent among all three domains of life, we have used this enzyme in our present study. All completely sequenced genomes were retrieved using FTP protocol from NCBI genome (ftp.ncbi.nih.gov) database (both RefSeq and GenBank) (*Pruitt et al., 2011; Benson et al., 2013*). We collated a list of 5077 both prokaryotic and eukaryotic genome sequences. We retrieved well-characterized TIM sequences from Uniprot database.

In order to identify TIM orthologs, we used well-known TIM sequences (Uniprot) as a query and searched them against all genome sequences as databases using BLAST+ (Camacho et al., 2008). The hits from the BLAST output were filtered on the basis of sequence identity and query coverage. For each genome, we select only one TIM ortholog, which has highest sequence identity and having ≥ 80% as query coverage. If we do not find any homolog, then we do not consider that genome sequence for further analysis. Here, the idea of finding homologs with such strict criteria is to reduce false positive orthologs, because TIM is most commonly occurring fold. This search resulted in 5074 hits.

Despite strict criterion, there is a possibility that we may get incorrect orthologs. To further refine list of orthologs, we used Hidden Markov Model (HMM) for TIM family (PF00121) from Pfam and scanned all BLAST parsed hit sequences using *hmmscan* program encoded in *Hmmer* package (*Bateman et al., 2016; Li W et al., 2015; Eddy, 2011*). From 'hmmscan' result, we selected sequences having significant match (e-value <0.01) to Pfam HMM profile. In addition to pruning orthologs, the *hmmscan* program also assigned domain boundaries for putative set of TIM orthologs. Of 5074 hits, hmmscan resulted in 5053 orthologs. Thus, obtained sequences were clustered at 90% sequence identity using CD-hit (*Li W, 2001; Li W, 2002*) that resulted in a total of 2272 representative sequences. These orthologous sequences were used for reconstruction of phylogeny. The sequences were mapped on Uniprot identifier to extract taxonomic lineages.

The experimentally determined kinetic parameters ($K_m$, $K_{cat}$ and $K_m/K_{cat}$) of TIM enzymes (Table 1) from some organisms were obtained from BRENDA (*Schomburg, 2002*) and SABIO-RK (*Wittig et al., 2011*).

*Table 1: Summary of kinetic parameters*

| **Values**<br>**Organsims** | **$K_m$ (mM)** | **$K_{cat}$($10^5$ min$^{-1}$)** | **$K_{cat}/K_m$**<br>**($10^5$min$^{-1}$mM$^{-1)}$)** |
|---|---|---|---|
| *Gallus gallus* | 0.47 | 2.6 | 5.5 |
| *Homo sapiens* | 0.39 | - | - |
| *Rattus norvegicus* | 0.87 | 1.5 | 1.8 |
| *Oryctolagus cuniculus* | 0.32 | 5.1 | 15.9 |
| *Saccharomyces cerevisiae* | 0.62 | 1.4 | 2.2 |
| *Trypanosoma brucei brucei* | 0.25 | 1.2 | 4.8 |
| *Entamoeba histolytica* | 0.83 | 0.0034 | 0.042 |
| *Plasmodium falciparum* | 0.35 | 2.7 | 7.7 |
| *Escherichia coli* | 1.03 | 5.4 | 5.2 |
| *Helicobacter pylori* | 3.46 | 0.88 | 0.25 |
| *Thermococcus onnurineus* | 1.7 | 14.3 | 8.6 |
| *Pyrococcus furiosus* | 1.2 | 15.2 | 12.8 |

## 2.2 Phylogeny reconstruction

We constructed phylogenies using Neighbor-Joining (NJ) and Maximum-likelihood (ML) methods. Since we have 2272 number of distinct organisms to be included in phylogenetic tree, its visualization and interpretation is a very difficult task. Hence, we first generate ML tree to identify taxonomic groups and selected one representative from such groups to make a set of 181 sequences, which are further reduced to 61 representative taxa. These 61 sequences are used for NJ based tree reconstruction. We explicitly included protein sequences with kinetic parameters in the representative set of 61 sequences.

Protein sequences are aligned using MUSCLE program (Edgar, 2004) and phylogenies are constructed using Molecular Evolutionary Genetics Analysis (MEGA) software (Tamura, Dudley, Nei, and Kumar 2007). MEGA can be utilized for deducing evolutionary relationships between different species based on a comparative analysis of an input of homologous protein sequences.

ML method is based on the general statistical principle of maximum-likelihood (*Felsenstein, 1981*). It initializes tree by constructing a sub-optimal tree. This tree with each iteration changes branch lengths, and calculates the likelihood of that tree topology, the process continues until likelihood of tree is maximized given the data. To compute distances between taxa, it uses similar models of amino acid substation as in NJ method.

NJ method utilizes distance between sequences to construct phylogeny. NJ starts with all the taxa rooted at a node, and then proceeds to build a pairwise distance matrix, according to which neighbors are grouped in a hierarchical fashion. The distance matrix can be generated using various models like, Dayhoff, Jones-Taylor-Thornton (JTT), Kimura's distance, each of which explain the differences in the amino acid sequences, in different ways. This method is employed to effectively sort out the large database of sequences, since it uses a polynomial time algorithm, which provides a nearly optimum tree at a fast computational speed (*Saitou and Nei, 1987*).

First, we generated ML tree for 2272 sequences and using visualization in MEGA grouped similar taxa together, which resulted in 181 representative taxa. We constructed phylogeny for this set of sequences. The ML tree was constructed using the G+I (Gamma + site invariant) model to generate site variation distribution and WAG model to construct the distance matrix. This tree was still quite difficult to interpret. Hence, we further reduced the number of representative sequences to 61, including sequence with kinetic parameters. For these 61 sequences, NJ tree was constructed using JTT model for amino acid substitution and sites are modeled using Gamma distribution ($\alpha$ =1.25). The bootstrap method using 500 replicates was used to find reliable branches.

# 3 Results and discussion

We have identified 5053 TIM orthologs in 5077 genomes. In general, both ML and NJ tree concur both the trees concur on the position of most of the taxonomic clades. The ML based phylogenetic tree for selected taxonomic clades *viz.* Eukaryota, Gamma Proteobacteria, Beta Proteobacteria, Alpha Proteobacteria, Bacteroidetes, Terrabacteria Group, Aquificae, Spirochaetes, FCB group, Synergistes, Chlymydia, Mycoplasma and Archaea shows (**Figure 7**) a clear segregation between Prokaryotes and Eukaryotes lineages. Interestingly, within prokaryotic lineage archaea and eubacteria forms 2 distinct groups. In fact, this is also evident from the sequence alignment between archaeal and eubacterial TIM orthologs. The archaea has been suggested as an ancestral clade in previous studies as well (*Keeling and Doolittle, 1997*). Importantly, in most cases TIM follows vertical pattern of inheritance. This makes it suitable to assess evolution of TIM specificity under Jensen's hypothesis.

Previous studies have argued that eukaryotic TIM was most likely derived from an early symbiotic bacterium of Alpha-Proteobacteria lineage, in fact *Rhizobium etli* sequences was suggested to be ancestral to eukaryotic lineage (*Keeling and Doolittle, 1997*). However, in both NJ and ML tree we observed that ancestral sequence could be of beta- or gamma-proteobacterial origin. Even though the ancestral node before bifurcating into eukaryotic and bacterial group does not have very high bootstrap values, we have seen this kind of lineage separation in tree constructed using other methods as well. This could be investigated further to understand evolutionary origin of eukaryotic TIM sequences.

We overlay $K_m/K_{cat}$ and $K_m$ values on phylogenetic tree generated using NJ method with 500 bootstraps (**Figure 8**). Considering catalytic efficiency ($K_m/K_{cat}$) TIM ancestral TIM enzymes are not less efficient, in fact, ancestral archaeal enzymes are highly efficient. Hence, viewing TIM evolution with respect to catalytic efficiency Jensen's hypothesis is not evident. This could be because of central role of this enzyme in glycolysis. Hence, possibly organism cannot afford to have a less efficient enzyme. Archaea used in present study are hyper-thermophiles, which have atypical TIM and sequences are quite different from eubacterial or eukaryotic lineages. It has been suggested that this could be due to their adaptation toward thermostability (Maes, 1999; Verhees, 2004).

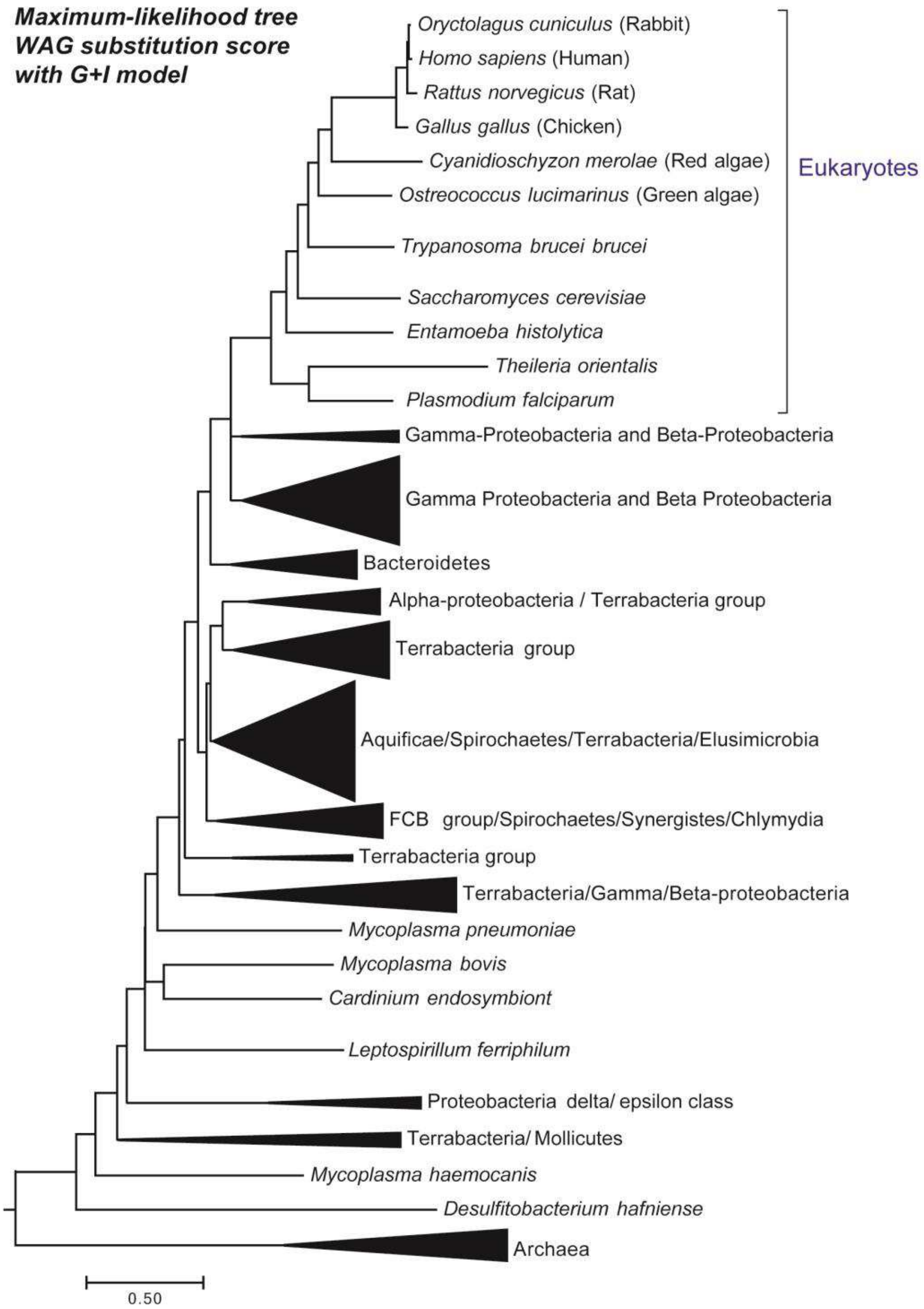

Maximum-likelihood tree
WAG substitution score
with G+I model
Oryctolagus cuniculus (Rabbit)
Homo sapiens (Human)
Rattus norvegicus (Rat)
Gallus gallus (Chicken)
Cyanidioschyzon merolae (Red algae)
Ostreococcus lucimarinus (Green algae)
Trypanosoma brucei brucei
Saccharomyces cerevisiae
Entamoeba histolytica
Theileria orientalis
Plasmodium falciparum
Eukaryotes
Gamma-Proteobacteria and Beta-Proteobacteria
Gamma Proteobacteria and Beta Proteobacteria
Bacteroidetes
Alpha-proteobacteria / Terrabacteria group
Terrabacteria group
Aquificae/Spirochaetes/Terrabacteria/Elusimicrobia
FCB group/Spirochaetes/Synergistes/Chlymydia
Terrabacteria group
Terrabacteria/Gamma/Beta-proteobacteria
Mycoplasma pneumoniae
Mycoplasma bovis
Cardinium endosymbiont
Leptospirillum ferriphilum
Proteobacteria delta/ epsilon class
Terrabacteria/ Mollicutes
Mycoplasma haemocanis
Desulfitobacterium hafniense
Archaea
0.50

*Figure 7: Maximum likelihood tree for 181 TIM sequences*

Taking $K_m$ values of TIM enzymes in various evolutionary lineages (**Figure 8**), ancestral enzymes have higher $K_m$ , less affinity, compared to more recently evolved enzymes, which have lower $K_m$. This suggests that in course of evolution modern enzymes have increased specificity towards its substrates. Hence, this indicates that Jensen's idea of ancestral enzyme having broader specificity may be possible correct in this context.

However, one thing to note is that we are looking at specificity to only one substrate; whether archaeal enzyme have broad specificity is yet to studied experimentally.

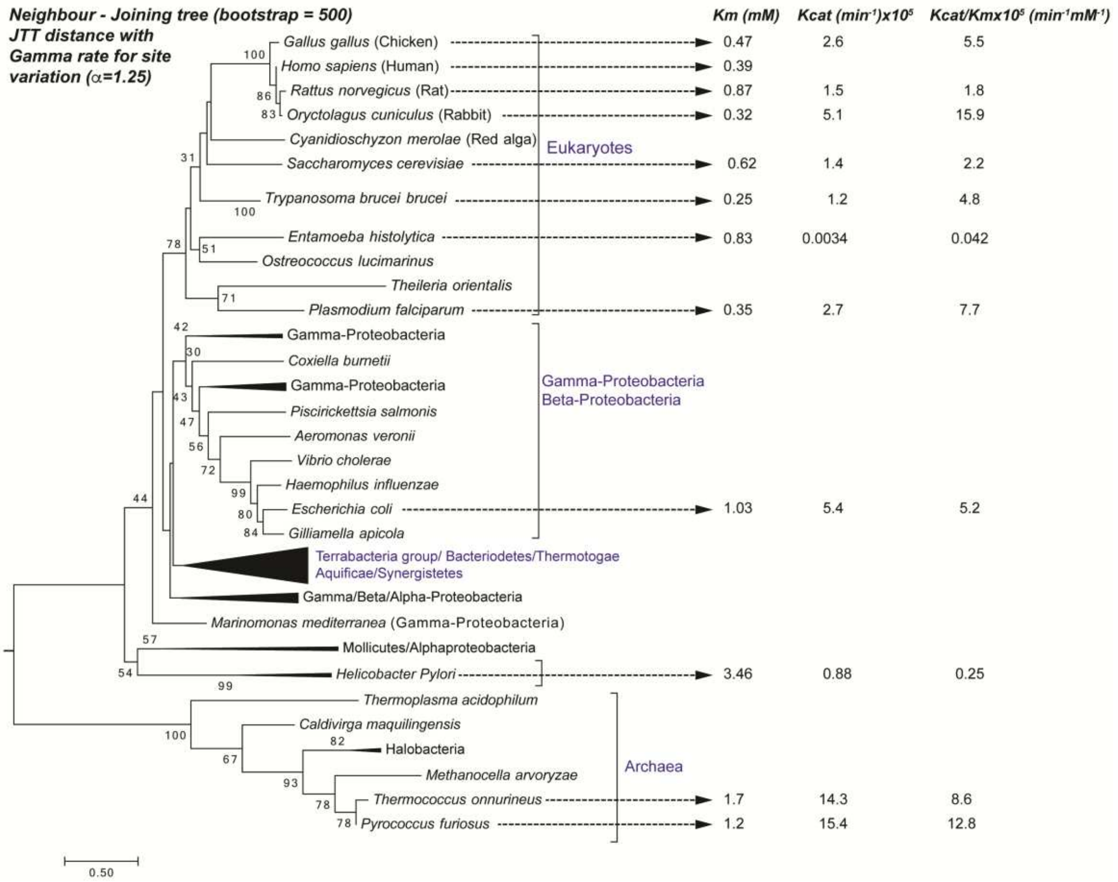

Neighbour - Joining tree (bootstrap = 500)
JTT distance with Gamma rate for site variation (α=1.25)
Km (mM)
Kcat (min⁻¹)x10⁵
Kcat/Kmx10⁵ (min⁻¹mM⁻¹)
Gallus gallus (Chicken) 0.47 2.6 5.5
Homo sapiens (Human) 0.39
Rattus norvegicus (Rat) 0.87 1.5 1.8
Oryctolagus cuniculus (Rabbit) 0.32 5.1 15.9
Cyanidioschyzon merolae (Red alga)
Saccharomyces cerevisiae 0.62 1.4 2.2
Eukaryotes
Trypanosoma brucei brucei 0.25 1.2 4.8
Entamoeba histolytica 0.83 0.0034 0.042
Ostreococcus lucimarinus
Theileria orientalis
Plasmodium falciparum 0.35 2.7 7.7
Gamma-Proteobacteria
Coxiella burnetii
Gamma-Proteobacteria
Piscirickettsia salmonis
Aeromonas veronii
Vibrio cholerae
Haemophilus influenzae
Escherichia coli 1.03 5.4 5.2
Gilliamella apicola
Gamma-Proteobacteria
Beta-Proteobacteria
Terrabacteria group/ Bacteriodetes/Thermotogae
Aquificae/Synergistetes
Gamma/Beta/Alpha-Proteobacteria
Marinomonas mediterranea (Gamma-Proteobacteria)
Mollicutes/Alphaproteobacteria
Helicobacter Pylori 3.46 0.88 0.25
Thermoplasma acidophilum
Caldivirga maquilingensis
Halobacteria
Methanocella arvoryzae
Archaea
Thermococcus onnurineus 1.7 14.3 8.6
Pyrococcus furiosus 1.2 15.4 12.8
100
86
83
31
100
78
51
71
42
30
43
47
56
72
99
80
84
44
57
54
99
100
82
67
93
78
78
0.50

*Figure 8: Neighbor joining tree for 61 sequences with overlay of kinetic data*

The ability of enzymes to bind with different substrates, even if in a non-optimized manner at first, is responsible for the existing repertoire of enzymes and their activities. Through the changes in environmental requirements, selection pressure shifts from one enzymatic activity to another, and the new activity selected is further optimized, further leading to the development of a new, more specific enzyme, due to gene duplication and divergence. Investigating enzyme evolution in the context of their kinetic parameters can open up the possibility of classifying enzymes and their promiscuity by the kinetic parameters. Moreover, such studies can provide insights that can assist in rational design of enzymes exploiting their promiscuity (*Copley, 2015*).

# PART II

*To extract protein-ligand interaction features using Canonical Correspondence Analysis (CCA)*

# 1 Introduction

## 1.1 Protein-ligand interactions

The knowledge of protein interaction partners is often starting point to describe its molecular function. In prediction of macromolecular interactions, finding interacting ligands could greatly assist in function annotation. Moreover, understanding of protein-ligand interaction can assist in rational design of drugs or substrates for new enzymatic activities.

Many sequence and/or structure-based approached have been developed to predict and model protein-ligand interactions. However, detailed insights into their mode of interaction are usually limited. With availability of protein tertiary structures molecular docking provided useful hints of protein-ligand interaction. However, such approaches are limited by lack of tertiary structures (*Yamanishi et al., 2011; Lacapère J-J et al., 2007*).

Recently, statistical methods have been developed as an alternative approach to investigate protein-ligand interactions (Keiser et al., 2009; *Yamanishi et al., 2011).* This is also feasible because of many protein structures are now available in PDB database. Many of these methods use both chemical and genomic information in their prediction approaches and usually referred as chemogenomics. For example, new drug targets have been predicted using statistical mining a comprehensive network of drug-target associations (Keiser et al., 2009). Usually, chemogenomics involves mining a given chemical space, for example drugs, in the context of its relation to the biological space, for example, drug targets. The fundamental assumption behind this approach is that similar molecule can bind similar targets (*Jacob and Vert, 2008*). Subsequent to prediction, one can analyze most important features and gain insights into protein-ligand interactions.

The ligand-protein interactions are often due to common chemical structures that are usually shared by the ligands and binding site residues (*Yamanishi et al., 2011*). It has been shown that binding site residues have relatively high sequence and structure conservation in comparison to other residues. Based on this assumption, in the present work we sought to derive the conservation patterns between chemical substructures and binding site residues using statistical method canonical correspondence analysis between chemical substructures derived from small molecules and features extracted from protein sequences. The main idea in this analysis is it will extract chemical substructures and sequence features, which jointly appear in the interaction pairs and disappear in the other pairs. The broad goal of such a study would be to identify and define rules for molecular recognition between chemical substructures and protein functional binding sites.

In the present approach, we represent proteins as *k-mers* of their respective amino acid sequences, and their associated ligands as stereo-chemical signatures of their respective chemical structures. The relation between these two defined sets of proteins and ligands, and their respective feature vectors, was given by creating a dataset of protein and ligand interactions.

## 1.2 Stereo-chemical signature

Chemical structures can be numerically characterized by using Graph Theory (*Schultz, 1989*). A molecule can be represented as a signature, containing a vector of its associated atomic signatures and their occurrence. An atomic signature is a canonical representation of the subgraph surrounding a particular atom. This subgraph includes all atoms and bonds up to a predefined distance from the given atom. This distance is called the signature height *h*. Moreover, this subgraph has all the vertices labeled in a canonical order, which at a molecular level help in giving a global description of the molecules (*Faulon et al., 2004; Carbonell and Faulon, 2010*). This method is optimized to reduce the size of the signature information, by avoiding duplicates in the storage. This is done by checking for isomorphic graph; when the bijection between two sets of vertices of two graphs is actually an isomorphic map of a graph's map onto the other one despite different forms, hence, only one of the graph is taken into account. A prerequisite for this is graph canonicalization, which assigns a unique label each atom, and is invariant to the atom order (*Carbonell et al., 2013*).

Mathematically, molecular signature descriptors can be defined as a 2D representation of the molecular graphs as undirected graphs: *G(V, E, C)*, where *V* is atoms present in the molecule, *E* is the bonds connecting the atoms, and *C* is the atom type (*Carbonell et al., 2013*). Then the molecular signature of G is given by:

$$^{h}\sigma(G)=\sum_{x_i\in V} {}^{h}\sigma(x_i)$$

Where, ${}^{h}\sigma(x_i)$ is atomic signature of G rooted at atom $x_i$ of height $h$.

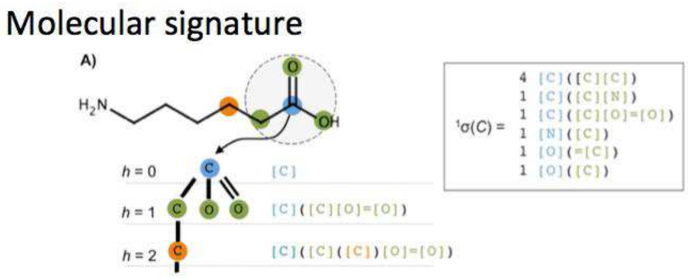


*Figure 9***:** *An example of molecular signature*

The signature descriptor of height *h* (**Figure 9**) of an atom *x* belonging to the molecular graph *G* is a canonical representation of the subgraph of *G*, which contains all the atoms present at a distance of *h* from the atom, *x.* In other words, if the height, *h* = 1, the signature will comprise of all the immediate neighbors of the atom, *x;* at *h* = 2, these neighbors will become the vertices, whose immediate neighbors will be considered, in a progressive fashion (*Carbonell et al., 2013*). These descriptors are then expressed as a string of characters corresponding to the canonized subgraph. Branch levels are indicated by a set of parentheses following the parent vertex (*http://www.issb.genopole.fr/~faulon/signatures.php*). But, the generic molecular descriptors do not take in account of a compound's chirality, that is their property of asymmetry, which, in turn, affects their chemical and biological properties. Therefore, stereochemistry information needs to be included in the molecular descriptors, to discern the chirality (*Carbonell et al., 2013*).

The approach used in this project, SSCAN (*Faulon, 2012*) takes in account the chirality of compounds, and represents ligand as a wholesome stereo-chemical signature. It does so by simultaneously describing chirality and signature descriptor, through an iterative algorithm, which follows a predefined set of conventions called Cahn-Inglod-Prelog priority rules (*Carbonell et al., 2013; Provisional Nomenclature Recommendations, 2005*).

## 1.3 The statistical approach: Canonical Correspondence Analysis (CCA)

CCA falls under the category of ordination techniques, which, in statistics, are exploratory data analytics. Ordination methods help in representing a multivariate data matrix with reduced dimensions. The term 'canonical', in mathematics, refers to reducing functions or relations to their simplest yet most comprehensive form, without a loss of generality (*Blasius and Greenacre, 2014*). Constrained ordination analyzes two data sets, response matrix and explanatory matrix, and allows for measuring the relationship between the two. It computes axes, which are the linear combination of the explanatory variables and best explain the variance within the response variables. Correspondence analysis differs from Principal Component Analysis in the type of data employed in analysis. CCA uses categorical data, rather than continuous data. CCA is basically CA with the canonical axes being the linear combinations of the explanatory variables, by using weighted multiple regression (*Borcard et al. 2011*).

CCA helps in dimension reduction for two co-dependent databases. The two databases are: a response matrix (say, *X {n x p}*), and an explanatory matrix (say, *Y {n x q}*). Ordination of *Y* is constrained such that the resulting ordination vectors are a linear combination of variables in *X*. It computes a general singular value decomposition of a matrix *Q {p x q},* which constitutes weighted averages for the columns of *Y* using the row totals of *X.*

The history of CCA stems from ecological analyses; where the multivariate method is used to elucidate the relationships between species abundance and the environmental variables present at the site (*Braak and Verdonschot, 1995*). In ecology, the matrix *X* is sites *vs* species and the matrix *Y* is sites *vs* environmental variables. Using the matrix X, site and species weights are computed, from which a correlation matrix of Y is computed. This correlation matrix and site weights are used to create a weighted matrix, *A,* in which each entry is an average environmental variable for a given species. Further the matrix *A* is decomposed into eigenvalues through Singular Value Decomposition (SVD). The aim is to reduce the dimensions of the matrix A, by transforming the data, such that the first two canonical dimensions (or axes) can explain the majority of other dimensions. SVD is a way of linear transformation, which operates upon a matrix and reduces redundancy within it. Suppose there is a matrix *W* {*r x t*}, using SVD, it can be decomposed as:

In the present approach, we applied CCA method to extract conserved patterns between chemical substructures and binding site residues such that both concurring features of protein and ligand appear and disappear together. Suppose *n* proteins have *p* k-mers and *m* ligands have *q* molecular signatures, such that they can be represented as linear combinations of their features, *u* and *v* respectively, with weights *α* and *β* respectively, and *I* is the indicator function. Then correlation maximized is (adopted from *Yamanishi et al., 2010*):

$$corr(u, v) = \frac{\sum_{i,j} I(\mathbf{x}_i, \mathbf{y}_j)\alpha^T\mathbf{x}_i \cdot \beta^T\mathbf{y}_j}{\sqrt{\sum_i d_{x_i}(\alpha^T\mathbf{x}_i)^2}\sqrt{\sum_j d_{y_j}(\beta^T\mathbf{y}_j)^2}}$$

where, the indicator function = 1 when there is an interaction between a protein and a ligand, otherwise it is 0 (*Yamanishi et al., 2011*). This data exploratory analysis can provide new insights into protein ligand interactions.

# 2 Methodology

Three kinds of matrices were generated, described as follows:

## 2.1 Protein *k-mer* Matrix

All PDB sequences were downloaded from RCSB site and atom record sequence was extract for each pdb chain using an in-house code. The proteins having length < 40 amino acids were removed from the analysis. We used CD-hit to remove redundant sequences with sequence identity ≥ 70% from the set of sequences. Using an in-house code overlapping stretches of 6 residues long was extracted, which we referred to as 6-mers. For example, if a sequence reads like, FPDGEDTPE, then the 6-mers would be: FPDGED, PDGEDT, DGEDTP, GEDTPE. The count of unique 6-mers is the signature for that sequence. Apart from the obvious advantage of computational efficiency, k-mers can prove to be good signatures to represent proteins and connect proteins otherwise unrelated in their secondary structures and functions (*Havukkala, 2010*). Subsequently, a matrix (**Figure 10 E**) was generated between the proteins and the 6-mers, such that each entry represents the occurrence of the 6mer {column value}, in the chain of a protein {row value}.

## 2.2 Ligand-Protein Matrix

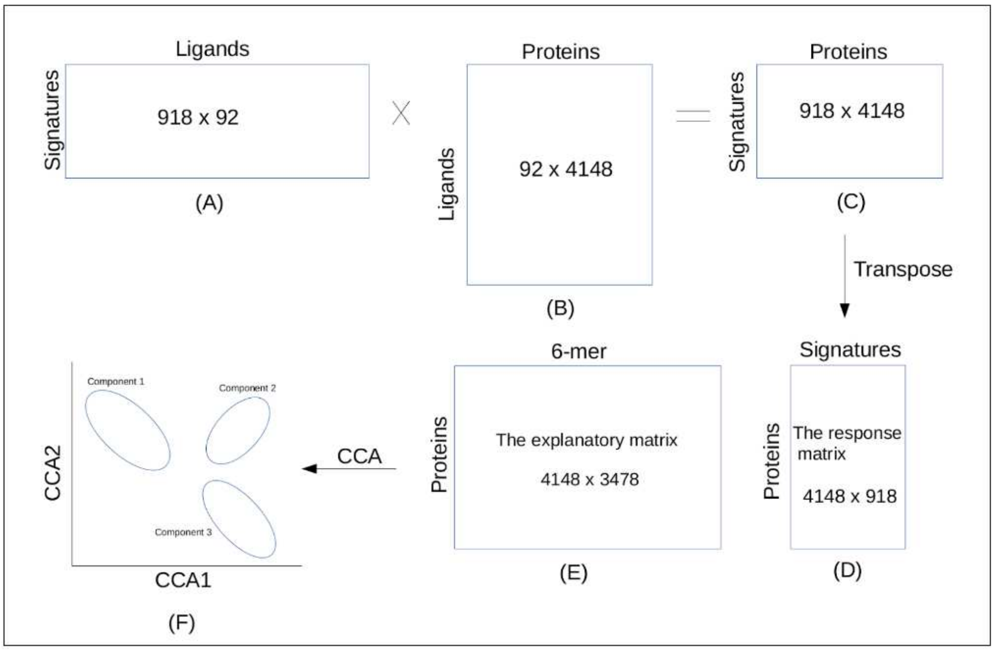
The protein structures were parsed through a program called LPC (Ligand - Protein Contacts) (*Sobolev et al., 1999*), which allows discerning the ligands, which interact with a given protein. Any ligand, which is interacting with 6 or more residues of the protein were considered for ligand protein matrix. A binary indicator matrix (**Figure 10 B**) was constructed with this information between ligands and proteins, with the entry of 1 denoting an interaction, and 0 denoting otherwise.

## 2.3 Signature-Ligand Matrix

The ligands interacting with at least one protein from the Ligand-Protein Matrix were downloaded as .sdf {structure data files} file, from RCSB, excluding the hydrogen atoms. The .sdf files were then converted into .mol format using OpenBabel (*O'boyle et al., 2011*) and further stereochemical signatures were extracted using SSCAN (*Faulon, 2012*) algorithm, at the height of 6. A frequency matrix (**Figure 10 A**) was computed between ligands and signatures. The top 92 ligands interacting with the most number of proteins were selected and consequently, a subset covering the selected ligands and their associated proteins were extracted, from the matrices.

## 2.4 Matrix manipulation and CCA

The matrices Signatures-Ligands and Ligands-Proteins were multiplied and transposed to generate the response matrix, Proteins-Signatures. While, the matrix, Proteins-6mers was used as the explanatory matrix. CCA was conducted on these two matrices using the R package, ade4 (*Dray et al., 2007*), treating 6-mers as environmental variables and signatures as species (**Table 2**)



.

*Figure 10: Summary of matrices used for CCA*

*Table 2: Parallel between variables used in ecological and our present study*

| **Ecological Variables** | **Variables used in this study** |
|---|---|
| Species | Chemical signatures of ligands |
| Sites | Proteins |
| Environmental Variables | 6-mers present in the proteins |

# 3 Results and discussion

From the summary of the analysis, it is observed that 6-mers and their patterns of occurrence can explain that 58.3% of the total constrained inertia or variance in the distribution of the ligand signatures. This constrained inertia is project on different axes, in order to explain most of the variance on minimum number of axes. The first two axes have the highest eigenvalues of 0.90 and 0.81, respectively. This explains most of the variance (**Figure 11**). Therefore, the first two axes would be ideal to plot the distribution scores on.

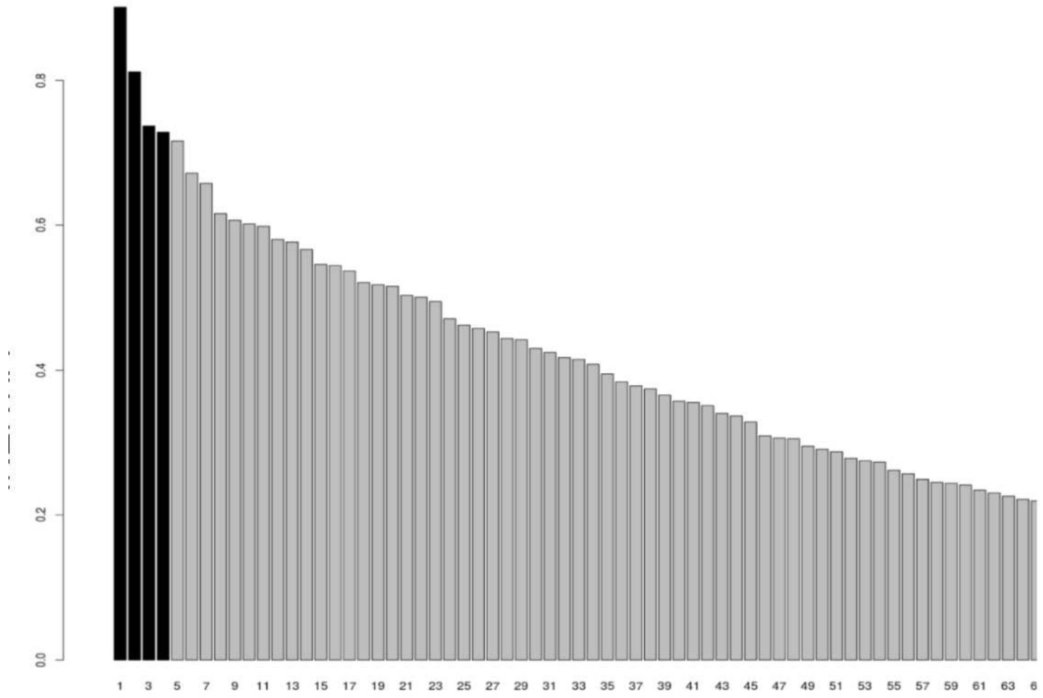


*Figure 11*: *Barplot (scree plot) showing variance*

The 6-mers distribution (**Figure 12**) and molecular signature (species) distribution (**Figure 13**) and the superimposition of the two (**Figure 14)** suggest that there are clusters in the distribution of the signatures and, the groups of 6-mers point towards one cluster. The length and the direction of the arrow depict the strength of correspondence between the associated 6-mer and the signature. In general, correlations amongst the 6-mers and signature can be ascribed to the ligands bound to proteins.

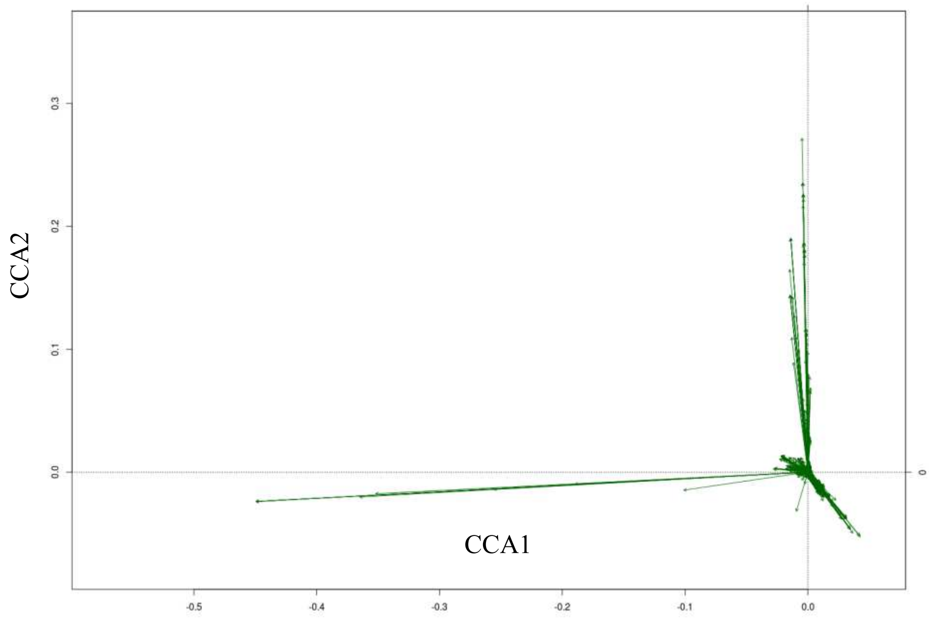


*Figure 12: Visualization of 6-mers*

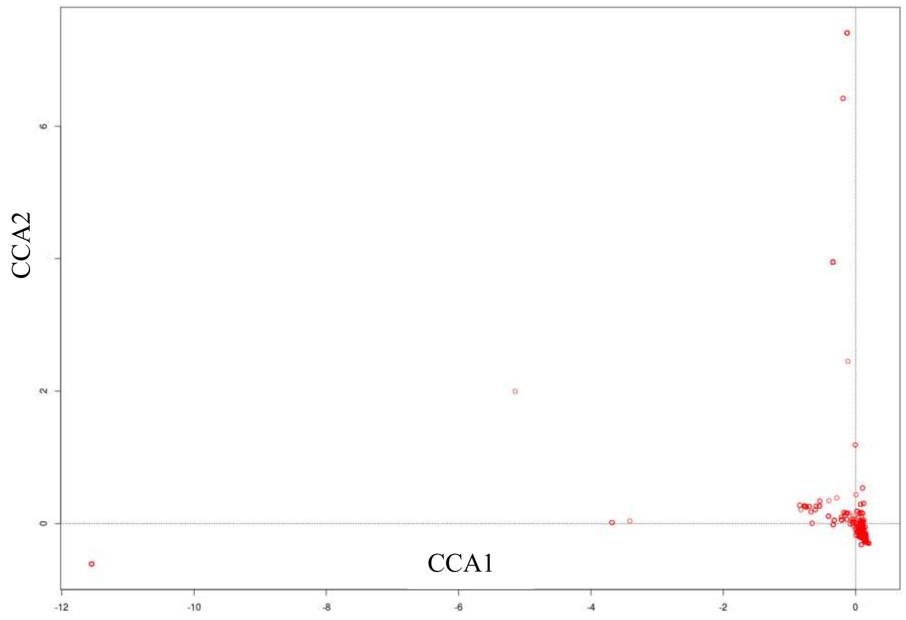


*Figure 13: Signature distribution*

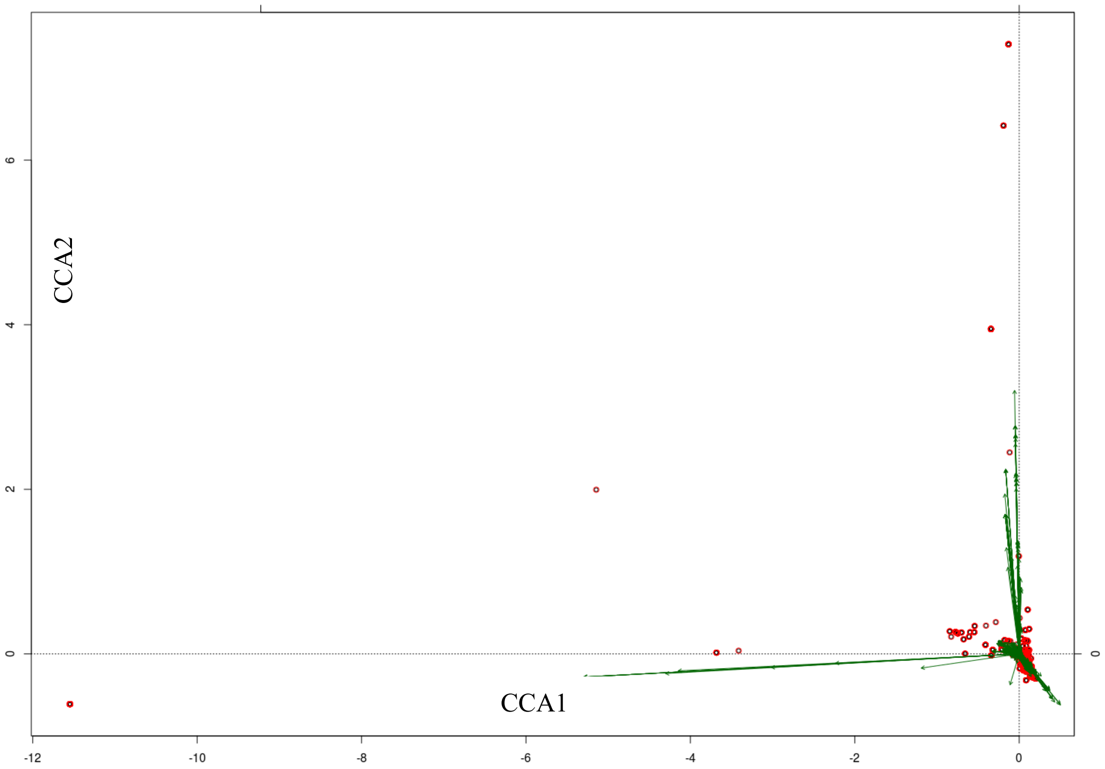


*Figure 14: Canonical correspondence analysis map of both signature and 6-mers.*

A few 6-mers clustered together were extracted to validate the data (**Table 3**). The 6-mers were then traced back to the proteins they belong to, which in turn were traced to the ligands with which they interact. It can be noted that there's correspondence between these 6-mers and the ligands BCL, LDA, SCN, MPD, etc. A cluster of signatures towards which this cluster of 6-mers was pointing was extracted too, and it could be seen that the majority of those signatures are present in the ligand BCL, and a few others like LDA, and MPD.

In conclusion, this result can be attributed to the existence of inherent patterns in protein-ligand interactions. In this preliminary analysis, we only considered 92 ligands. The analysis on whole dataset may lead to better understanding of ligand-protein interactions. Moreover, 6-mers thought to be in a correspondence with a few signatures could be looked at in the spatial context of the protein they belong to, in other words, the said 6-mers could be placed in the 3D structures of the proteins they hail from, in order to investigate if they're present near or at the ligand binding sites of the protein and are directly involved in the binding with ligand signatures. This initial analysis provides usefulness of CCA in elucidating patterns from protein and ligands that could be exploited to predict their interactions.

*Table 3: Summary of relation between 6-mer and ligands*

| 6-mers extracted | PDB id with 6-mers | Ligands the proteins bind to |
|---|---|---|
| DFWVGP | 3WMM_L<br>4CAS_B<br>4IN7_L | BCL<br>BCL<br>BCL |
| FDFWVG | 3WMM_L<br>4CAS_B<br>4IN7_L | BCL<br>BCL<br>BCL |
| LFDFWV | 3WMM_L<br>4CAS_B<br>4IN7_L | BCL<br>BCL<br>BCL |
| VGFFGV | 3WMM_L<br>4CAS_B<br>4IN7_L | BCL<br>BCL<br>BCL |
| LGKIGD | 3KHS_A<br>3WMM_M<br>4CAS_C | TRS<br>BCL<br>BCL |
| REGYPL | 1EYS_H<br>1VRN_H<br>4IN7_H | BCL<br>LDA<br>BCL, LDA |
| EEEAIA | 1LGH_B<br>1XQ9_B<br>3KXL_A | BCL<br>SCN<br>SCN |
| TLLGVL | 1SDI_A<br>2PNO_A<br>3WMM_L | ACY, MPD<br>LMT, GSH<br>BCL |

# Bibliography

- Bateman A (2002) The Pfam Protein Families Database. Nucleic Acids Research 30:276–280.
- Blasius J, Greenacre M (2014) Visualization and Verbalization of Data. Visualization and Verbalization of Data: Chapman and Hall/CRC.
- Borcard D, Legendre P, Gillet François (2011) Numerical Ecology with R. New York: Springer.
- Braak CJFT, Verdonschot PFM (1995) Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquatic Sciences 57:255–289.
- Benson DA *et al.* (2013) GenBank. Nucleic acids research. 1;41(D1):D36-42.
- Berman et al. (2000) The Protein Data Bank *Nucleic Acids Research*, 28: 235-242.
- Camacho C. *et al.* (2008) BLAST+: architecture and applications. BMC Bioinformatics 10:421
- Carbonell P, Faulon J-L (2010) Molecular signatures-based prediction of enzyme promiscuity.Bioinformatics 26:2012–2019.
- Carbonell P, Carlsson L, Faulon J-L (2013) Stereo Signature Molecular Descriptor. Journal of Chemical Information and Modeling 53:887–897.
- Computational Protein Design. SlideShare Available at:
- https://www.slideshare.net/pablocarb/lecture02-6180470 [Accessed April 2017].
- Copley S (2003) Enzymes with extra talents: moonlighting functions and catalytic promiscuity. Current Opinion in Chemical Biology 7:265–272.
- Copley SD (2015) An evolutionary biochemist's perspective on promiscuity. Trends in Biochemical Sciences 40:72–78.
- D'ari R, Casadesús J (1998) Underground metabolism. BioEssays 20:181–186.
- Davenport R *et al.* (1993) Structure Of The Triosephosphate Isomerase Phosphoglycolohydroxamate Complex: An Analogue Of The Intermediate On The Reaction Pathway.
- Dray, S. and Dufour, A.B. and Chessel, D. (2007): The ade4 package-II: Two-table and K-table methods. R News. 7(2): 47-52
- Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Computational Biology 7.
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792–1797.

- Faulon JL (2012). Signature molecular descriptor. http://www.issb.genopole.fr/faulon/sscan.php [Accessed October 2016]
- Faulon J-L, Collins MJ, Carr RD (2004) The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequences. Journal of Chemical Information and Computer Sciences 44:427–436.
- Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17:368–376.
- Havukkala I (2010) Biodata mining and visualization: novel approaches. New Jersey: World Scientific.
- Jacob L, Vert J-P (2008) Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics 24:2149–2156.
- Janssen DB, Dinkla IJT, Poelarends GJ, Terpstra P (2005) Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. Environmental Microbiology 7:1868–1882.
- Jensen RA (1976) Enzyme Recruitment in Evolution of New Function. Annual Review of Microbiology 30:409–425.
- Johnson KA, Goody RS (2011) The Original Michaelis Constant: Translation of the 1913 Michaelis–Menten Paper. Biochemistry 50:8264–8269.
- Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2011) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research 40.
- Keeling PJ, Doolittle WF (1997) Evidence that eukaryotic triosephosphate isomerase is of alpha-proteobacterial origin. Proceedings of the National Academy of Sciences 94:1270–1275.
- Keiser MJ *et al.* (2009) Predicting new molecular targets for known drugs. Nature 462:175–181.
- Khersonsky O, Tawfik DS (2006) The Histidine 115-Histidine 134 Dyad Mediates the Lactonase Activity of Mammalian Serum Paraoxonases. Journal of Biological Chemistry 281:7649–7656.
- Lacapère J-J, Pebay-Peyroula E, Neumann J-M, Etchebest C (2007) Determining membrane protein structures: still a challenge! Trends in Biochemical Sciences 32:259–270.
- Li W, Jaroszewski L, Godzik A (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17:282–283.

- Li W, Jaroszewski L, Godzik A (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. Bioinformatics 18:77–82.
- Li W *et al.* (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Research 43.
- Maes D *et al.* (1999) The crystal structure of triosephosphate isomerase (TIM) from Thermotoga maritima: A comparative thermostability structural analysis of ten different TIM structures. Proteins: Structure, Function, and Genetics 37:441–453.
- Mccloskey D, Palsson BO, Feist AM (2014) Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Molecular Systems Biology 9:661–661.
- Molecular Signatures. Molecular signatures Available at: http://www.issb.genopole.fr/~faulon/signatures.php [Accessed January 2017].
- National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988]. Available from: https://www.ncbi.nlm.nih.gov/. [Accessed August 2016]
- Nam H, Lewis NE, Lerman JA, Lee D-H, Chang RL, Kim D, Palsson BO (2012) Network Context and Selection in the Evolution to Enzyme Specificity. Science 337:1101–1104.
- Nobeli I, Favia AD, Thornton JM (2009) Protein promiscuity and its implications for biotechnology. Nature Biotechnology 27:157–167.
- O'brien PJ, Herschlag D (1999) Catalytic promiscuity and the evolution of new enzymatic activities. Chemistry & Biology 6.
- O'boyle NM *et al.* (2011) Open Babel: An open chemical toolbox. Journal of Cheminformatics 3:33.
- Olga Khersonsky And Dan S. Tawfik (2010) Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. Annual Review of Biochemistry 79:471–505.
- Open Babel, version 2.3.1, http://openbabel.org - (accessed Oct 2016)
- Provisional Nomenclature Recommendations (2005). Pure and Applied Chemistry 77.
- Pruitt KD, Tatusova T, Brown GR, Maglott DR (2011) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Research 40.
- Reitz M, Sacher O, Tarkhov A, Trümbach D, Gasteiger J (2004) Enabling the exploration of biochemical pathways. Org Biomol Chem 2:3226–3237.

- Roland BP *et al.*(2015) Triosephosphate isomerase I170V alters catalytic site, enhances stability and induces pathology in a Drosophila model of TPI deficiency. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease 1852:61–69.
- Saitou N, Nei M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution
- Schomburg I (2002) BRENDA, enzyme data and metabolic information. Nucleic Acids Research 30:47–49.
- Schultz HP (1989) Topological organic chemistry. 1. Graph theory and topological indices of alkanes. Journal of Chemical Information and Modeling 29:227–228.
- Sharma P, Guptasarma P (2015) 'Super-perfect' enzymes: Structural stabilities and activities of recombinant triosephosphate isomerases from Pyrococcus furiosus and Thermococcus onnurineus produced in Escherichia coli. Biochemical and Biophysical Research Communications 460:753–758.
- Sobolev V., Sorokine A., Prilusky J., Abola E.E., Edelman M.(1999) Automated analysis of interatomic contacts in proteins. Bioinformatics, 15, 327-332
- Stobbe MD, Jansen GA, Moerland PD, Kampen AHCV (2012) Knowledge representation in metabolic pathway databases. Briefings in Bioinformatics 15:455–470.
- Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. Molecular Biology and Evolution 24:1596–1599.
- Tokuriki N, Tawfik DS (2009) Protein Dynamism and Evolvability. Science 324:203–207.
- Verhees CH *et al.* (2004) The unique features of glycolytic pathways in Archaea. Biochemical Journal 377:819–822.
- Voet D, Voet JG, Pratt CW (2006) Fundamentals of biochemistry: life at the molecular level. New York: John Wiley & Sons.
- Wackett, L. P. (2009) Questioning our perceptions about evolution of biodegradative enzymes, Current opinion in microbiology 12, 244-251.
- Weisstein, Eric W. "Isomorphic Graphs." From *MathWorld*--A Wolfram Web Resource. http://mathworld.wolfram.com/IsomorphicGraphs.html [Accessed March 2017]
- Wierenga RK, Kapetaniou EG, Venkatesan R (2010) Triosephosphate isomerase: a highly evolved biocatalyst. Cellular and Molecular Life Sciences 67:3961–3982.

- Wittig U *et al.* (2011) SABIO-RK--database for biochemical reaction kinetics. Nucleic Acids Research 40.
- Yamanishi Y, Pauwels E, Saigo H, Stoven V (2011) Extracting Sets of Chemical Substructures and Protein Domains Governing Drug-Target Interactions. Journal of Chemical Information and Modeling 51:1183–1194.