Understanding enzyme function evolution from a computational perspective
Jonathan D Tyzack¹, Nicholas Furnham², Ian Sillitoe³, Christine M Orengo³ and Janet M Thornton¹

In this review, we will explore recent computational approaches to understand enzyme evolution from the perspective of protein structure, dynamics and promiscuity. We will present quantitative methods to measure the size of evolutionary steps within a structural domain, allowing the correlation between change in substrate and domain structure to be assessed, and giving insights into the evolvability of different domains in terms of the number, types and sizes of evolutionary steps observed. These approaches will help to understand the evolution of new catalytic and non-catalytic functionality in response to environmental demands, showing potential to guide de novo enzyme design and directed evolution experiments.

Addresses
¹ EMBL-EBI, Wellcome Genome Campus, CB10 1SD, United Kingdom
² London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, United Kingdom
³ Institute of Structural and Molecular Biology, University College London, Gower Street, London, WC1E 6BT, United Kingdom

Current Opinion in Structural Biology 2017, 47:131–139
This review comes from a themed issue on Catalysis and regulation
Edited by Christine Orengo and Janet Thornton

http://dx.doi.org/10.1016/j.sbi.2017.08.003
0959-4409/© 2017 Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Introduction
Enzymes are the product of billions of years of evolution giving us molecular machines that are critical for life. Across the vast space of protein structure, evolution has settled upon a limited number of structural folds that support an incredibly diverse chemistry acting on a multitude of substrates. Enzymes have evolved substrate and function specificity to improve the organism’s overall fitness in response to environmental demands, but enzymes can have multiple substrates and functions, contradicting the traditional view of one enzyme one reaction and the high specificity implied in the lock and key and induced-fit paradigms. Furthermore, it is advantageous for organisms to be able to adapt quickly to a changing environment which manifests itself at the molecular level in the inherent promiscuity present in many enzymes.

Exploiting enzyme promiscuity to develop novel functionality has been the focus of much recent research effort, where directed evolution techniques can improve the catalytic performance of even very weakly active starting points to become commercially relevant enzymes. In this review, we discuss the biochemical role of enzymes in terms of specificity and functionality, and then focus on recent developments in the application of structural bioinformatics methods to understand the evolution of specificity and guide de novo enzyme design.

Functional changes in enzyme evolution
Evolution is a random generator of possible improvements in the face of the environmental challenges an organism experiences, where survival of the fittest ensures the retention of successful solutions into future generations. Evolution is powerful and has produced enzymes with varying degrees of substrate and function specificity where beneficial to the organism. Figure 1 shows the various types of functional changes observed in enzyme evolution, where more common changes in chemistry and substrates are supplemented with rarer gain or loss of function in the form of moonlighting and pseudo-enzymes respectively. These rarer functional changes will be considered briefly first, before the discussion moves to the more common evolutionary pathways.

Gain and loss of enzyme function: moonlighting and pseudo-enzymes
The acquisition of new functionality within an enzyme family as a result of gene duplication and mutation is commonly observed, see [1*] for some selected examples, but some enzymes can exhibit secondary, markedly different non-enzymatic functionality, when an enzyme with exactly the same chemical structure can moonlight to perform different roles in different cell compartments or environments [2,3]. These moonlighting enzymes, although rare, are being increasingly documented, where secondary function may be controlled by the varying ligand concentrations, by local phosphorylation levels and their different homo/hetero oligomeric states.

The additional functionality from moonlighting enzymes usually arises from a different site in the same structure and is distinct from gene fusions, multiple RNA splice variants or pleiotropic effects (where one gene influences two or more seemingly unrelated phenotypic traits), which can all affect enzyme catalysis. The evolution of secondary functional sites on an enzyme and the
regulation of expression are active research questions [4] and demonstrate that multi-functional proteins are a design possibility, opening up opportunities for multi-functional polypeptide drugs and synthetic pathways. Researchers are becoming increasingly aware of the contribution of moonlighting enzymes but it is challenging to identify moonlighting functionality with bioinformatics methods [5] where small changes in structure or context can have dramatic effects on functionality [6]. Most discoveries of moonlighting occur from observation and serendipity and a database of moonlighting enzymes with approaching 500 entries has recently been curated from the literature [7], where one example is alpha-crystallin, the structural protein in the lens of the eye, that also has lactate dehydrogenase and argininosuccinate lyase activity.

Pseudo-enzymes are proteins that closely resemble an active enzyme, but at some point have lost their catalytic functionality and are retained in a genome for the beneficial new functionality that they have acquired [8], such as roles in regulatory and signaling pathways. Pseudo-enzymes are considered in more detail elsewhere in this edition, and the discussion here will move on to the evolution of different functionality from the same binding pocket giving changes in substrate specificity and chemistry.

**Different types of enzyme substrate specificity**

Enzymes are able to give many orders of magnitude speed-up in essential reactions such as respiration, digestion and photosynthesis by stabilising transition states, thereby reducing activation energies and enabling reactions to proceed on timescales that can support life [9]. At a fundamental level, catalytic residues must be positioned around substrates in the correct orientations to stabilise transition states [10]. However, the specificity of the binding event between enzyme and substrate can vary depending on the extent to which the pocket has been optimised over evolutionary time [11]. An enzyme only needs to offer selectivity over detrimental side-reactions on potential substrates it is likely to encounter in its expressed location, where it becomes beneficial for evolution to deliver greater specificity. Indeed, some enzymes have evolved group or bond specificity such as those acting on some proteins and carbohydrates (see Figure 2 for trypsin and amylase examples), a more efficient solution then having to evolve a set of highly specific enzymes for every occurrence of each bond or group.

Enzyme specificity, defined according to the range of substrates and their similarity either for an individual enzyme or family of enzymes, exists on a continuum between highly specific and highly unspecific (promiscuous), demonstrating the concept that enzymes only evolve specificity when it is advantageous for the organism. It is challenging to define mutually exclusive categories to characterise the varied specificity observed, but despite this four categories of enzyme specificity have emerged [12]:

a) high specificity (an enzyme catalyzes one reaction at one site in one substrate to produce one product)
b) group specificity (an enzyme acts on a specific group (i.e. a given bond (cleaving or ligating) in a defined and restricted molecular environment))
c) bond specificity (an enzyme acts on a specific bond regardless of molecular environment)
d) low specificity (an enzyme can act at multiple sites in multiple substrates where site of reaction is influenced but not dictated by reactivity and accessibility considerations).
A classic example of low specificity is isoform 3A4 from the Cytochrome P450 family of enzymes that is involved in the oxidation of a broad range of xenobiotics at multiple sites in the substrate via several pathways, facilitating their subsequent conjugation and elimination from the organism [13]. Across the P450 family, the different isoforms have varying expression levels, tissue distributions, binding pocket sizes and characteristics giving rise to markedly different substrate profiles, causing them to occupy different positions on the specificity continuum.

There is a further important consideration given the polymeric nature of the molecules of life. Many enzymes which act on proteins, DNA, RNA, sugar chains and lipids, show bond and group specificity, but polymer promiscuity, that is, although they have some side-chain specificity, they act on multiple different polymers. It would be extremely inefficient to have separate enzymes operating on all possible monomer combinations. For example, the alpha amylase cleavage of glycosidic links in starch and glycogen is bond specific [14], but interestingly also stereo specific and unable to cleave beta links in cellulose, disqualifying cellulose as a source of glucose for animals. Examples of group specific enzymes include pepsin, trypsin and chymotrypsin cleaving the amino groups of aromatic, amino groups of basic and carboxyl groups of aromatic amino acids respectively [15]. The phosphorylation of hexoses by hexokinase is also group specific, in contrast to the highly specific glucokinase acting only on glucose [16].

**Modes of evolution: creeping and leaping**

Enzyme evolution is complex [17**] where evolutionary events such as single-point mutations, indels and domain fusions can cause substrate specificity creep due to changes in the binding cavity or more remarkable leaps in chemistry often due to changes to catalytic residues. Most evolutionary changes cause a relatively minor change in function described as ‘creeping evolution’ but occasionally there can be a radical shift or ‘leap’ in function, for example, when a change allows the binding of a completely different substrate or when a substrate binds in ‘reverse mode’ or when the change provides an alteration in enzyme mechanism (Figure 3).

The core structure of the protein is rarely changed by such evolutionary events despite low sequence identity between relatives, providing a robust scaffold that enables changes in different relatives to give varying degrees of diverse chemistry [18**]. The focus of the remainder of this review is on the application of bioinformatics methods to understand enzyme promiscuity and the evolution of substrate specificity for a given catalytic site, usually confined to a single domain of the protein. However, nature likes to recycle and the emergence of novel functionality from the recombination of different...
The creeping and leaping modes of evolution. Creeping involves incremental changes to the binding cavity leading to minor changes in functionality (example based on binding pocket mutations to give the substrate creep from EC:1.4.3.4 (monoamine oxidase) to EC:1.4.3.2 (l-amino-acid oxidase). Leaping involves a radical shift in function, dramatically altering substrate binding or chemistry (example based on binding pocket mutations to give the leap between EC:1.2.1.10 (acetaldehyde dehydrogenase) and EC:4.1.3.39 (4-hydroxy-2-oxopentanoate pyruvate lyase)). Inset molecular structures generated with MarvinSketch [35].

<table>
<thead>
<tr>
<th>Mode of evolution</th>
<th>EC:1.4.3.4</th>
<th>EC:1.4.3.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Creeping</td>
<td>Monoamine oxidase</td>
<td>L-amino-acid oxidase</td>
</tr>
<tr>
<td>(b) Leaping</td>
<td>acetaldehyde dehydrogenase</td>
<td>4-hydroxy-2-oxopentanoate pyruvate lyase</td>
</tr>
</tbody>
</table>

domains [19,20] giving rise to new protein–protein interactions and quaternary structure [21,22] is also commonly observed, but not discussed further here.

The large scale study of enzyme evolution is becoming possible from the increasing amount of sequence and structural data available [1*] where alignment methods can be applied to identify evolutionary relationships between enzymes [23*]. It has been observed that changes in substrate specificity, probably arising from incremental binding site mutations, are far more likely than changes in chemistry, which probably require many complementary mutations to key catalytic residues without disrupting enzyme activity. However, the impact of longer range mutations further from the catalytic centre can sometimes have dramatic effects on the binding site meaning that the subtle but cumulative effects of second and third shell mutations cannot be ignored [24,25]. There is a growing body of evidence that non-additive interactions amongst mutations (epistasis) are common in adaptive evolution [26–28], amplifying the effect of later mutations and making the fitness landscape more extreme, enabling more diverse evolutionary trajectories to be accessing [29].

However, whilst substrate changes are more common, the capability of evolution to deliver dramatic leaps in chemistry is demonstrated by the remarkable observation that all changes between EC (Enzyme Commission) classes are observed. Most EC classes retain the same primary class, but isomerases EC5 are exceptional in that they are more likely to evolve a different function than remain an isomerase [30], with conversions to lysases EC4 occurring more frequently than expected [31].

The size of the evolutionary steps required to deliver these evolutionary creeps and leaps can be measured from the extent of the change in sequence and structure [30]. In order to explore the relationship between sequence changes and functional changes, it is necessary to develop quantitative measurements of changes in function and changes in specificity. The change in chemistry can be measured using recently developed computational methods to compare the similarity of enzyme reactions by
decomposing them into feature vectors derived from the bond changes they catalyze, and the reaction centres and substrates they operate on \([32^*,33^*]\). This allows the correlation between change in substrate structure and change in enzyme structure/sequence to be investigated.

The ability to quantitatively measure enzyme reaction similarity also reveals inconsistencies and irregularities inherent in the current hierarchical EC numbering system. For instance, almost a third of all known EC numbers are associated with more than one enzyme reaction in the KEGG database \([34]\).

**Measuring changes in specificity and function**
Recent work in our group has used FunTree \([23^*]\) and CATH \([36]\) to identify evolutionary relationships between enzymes within a homologous family and investigate the correlation between change in substrate structure (calculated using ECBLAST \([32^*]\)) and change in enzyme structure (calculated using SSAP \([37^*]\)) generating plots at the CATH domain level. An example domain has been chosen from each of the CATH structural classes (a) All Alpha, (b) All Beta and (c) Mixed Alpha/Beta, with Figure 4 showing the analysis for CATH domains (a) 1.10.600.10 (e.g. farnesyl diphosphate synthase).

![Figure 4](image)

*Substrate similarity of the main reactant (calculated using ECBLAST) against domain structure similarity (calculated using SSAP) for each evolutionary relationship identified with FunTree for CATH domains: (a) 1.10.600.10, (b) 2.40.110.10, (c) 3.90.1150.10. One member enzyme of each domain family is shown in the top left.*
synthase), (b) 2.40.110.10 (e.g. butyryl-CoA dehydrogenase) and (c) 3.90.1150.10 (e.g. aspartate aminotransferase), where each data point is color coded depending on the EC level at which the change occurs. For each example domain we have fitted a line to the points to gain an approximate correlation.

Inspecting the plots for all domains it is clear that a trend exists whereby bigger changes in substrate structure generally require bigger changes in protein sequence and structure, although there are some exceptions where the linear trend is broken and the data points form a cluster. Also, some domains apparently show an anti-correlation, where the most similar substrates are catalyzed by the most dissimilar enzymes. Each family is complex and requires detailed analysis. Examining the evolution within each CATH domain allows comparisons to be made in terms of the number of evolutionary changes observed, the diversity of the chemistry performed, the number of changes in function (as defined by a change in EC class) and the amount of structural change required to accommodate new substrates and mechanisms. This will help to give insights into the evolvability of different domains and may help in understanding which domains have the potential to be multi-specific in terms of finding suitable starting points for directed evolution.

For instance, in the examples shown, CATH domain 1.10.600.10 only shows changes in substrate, whereas domain 2.40.110.10 shows changes at all EC levels, including changes in the bonds broken/formed, the basic reaction and the substrate specificity and domain 3.90.1150.10 shows changes of substrate (4th level EC) and changes at the primary EC class level. Domain 3.90.1150.10 contains 29 data points, almost three times as many as the other domains, which may indicate a more versatile and evolvable domain.

What determines specificity?
Understanding enzyme evolution and the factors that facilitate specificity would be helpful in de novo enzyme design and it has been proposed that a key enabler of promiscuity is the flexibility of the enzyme [38]. Evolution gives cycles of destabilisation and restabilisation allowing conformational space to be sampled [39] whilst maintaining the positions of important catalytic residues. The trade-off between stability and activity is demonstrated in the adaptive evolution of RubisCO [40] and highlights the strong biophysical constraints that influence the evolution of enzymes.

It has also been suggested that the location of active site residues can influence the evolvability. The presence of active site residues in flexible, loosely packed loops distinct from the core scaffold is thought to facilitate high evolvability [41**] where mutations to key residues are less likely to have a destabilising effect on the protein.

For example, in TIM barrels, the location of the active site at one end of the barrel involving many loops allows them to incorporate more easily a wide range of cofactors and contributes to their prevalence in modern proteomes where they support very diverse substrates and chemistry [42].

A further important feature to facilitate promiscuous function may be the presence of water networks to stabilise binding of non-native substrates, which, if providing advantageous functionality, would be refined over evolutionary time by the replacement of entropically unfavourable water-mediated interactions with polar protein–ligand interactions [43].

Resurrecting ancestral enzymes
Understanding natural evolution gives opportunities to resurrect ancestral enzymes as suitable starting points for directed evolution [44–46]. It has been proposed that early enzymes were generalists with broad functionality and substrate specificity [47**], giving rise to more specialist enzymes over evolutionary time. However, it is also possible that enzymes could evolve to be more promiscuous given the right circumstances. The stability of early enzymes in the face of harsher prevailing conditions [48] might also make them suitable candidates for repurposing and more tolerant to the high mutational load of directed evolution.

The evolution of specificity from a common ancestor is demonstrated by the flavin-dependent monooxygenases where a promiscuous FAD binding domain gave rise to more specific functionality from various domain fusion events [49]. Domain fusion events are also implicated in the structurally and functionally diverse HADSF enzyme superfamily where enzymes with an inserted CAP domain show wider substrate promiscuity [50]. The Cytochrome P450s also have a broad substrate range making them repurposing candidates for drugs and pharmaceuticals [51].

Exploiting enzyme promiscuity
The innate substrate promiscuity of many enzymes, showing weak activity with off-target but similar molecules, can help to drive the acquisition of new functions. This inherent promiscuity exists since over evolutionary time, the pressure of natural selection ceases when further catalytic or specificity improvements do not improve fitness [52**]. Therefore, a perfectly specific active site is unnecessary and in a changing environment it is likely to be beneficial for enzymes to retain an inherent promiscuity and the corresponding ability to evolve new functions. This is demonstrated by the acquisition of resistance to beta-lactam antibiotics by populations of bacteria over many generations, one of the key challenges to modern medicine, where progress has been made in understanding the resistance profile of the different
beta-lactamases and pinpointing resistance properties to key sequence sites [53].

Promiscuity is thought to be one factor that facilitates natural evolution, and can be exploited by directed evolution [54,55] which has brought enzyme catalyzed synthesis to many industries including pharmaceuticals, textiles and food [56–58]. Promiscuity can vary greatly between orthologous enzymes [59] as a result of neutral sequence divergence so it is important to consider multiple orthologues as start points in directed evolution.

The importance and diversity of directed evolution is demonstrated by recent progress in developing enzymes to improve CO2 fixation [45,60]; detoxify organophosphates from contaminated soil and water [61]; catalyze the formation of organosilicon compounds [62]; and destroy latent HIV pro-virus in cells [63]. Directed evolution methods reinforce the fact that there is nothing magical about the honing of mechanisms by natural evolution, and sophisticated, highly active and novel mechanisms have been shown to emerge from ultra-high-throughput techniques, such as the emergence of a catalytic tetrad to rival the efficiencies of natural aldolases [64].

Opportunities for the future

Enzyme promiscuity is an important factor in enzyme evolution where new functions emerge at the edges of current functionality [45] from the refinement of weak non-native interactions as required by environmental demands. The development of de novo enzymes using directed evolution aims to exploit this inherent promiscuity and bioinformatics methods can be used to identify suitable start points and guide these approaches. Any insights that can be garnered from natural evolution to inform de novo enzyme design are invaluable, such as the identification of evolutionary labile structures and molecules that are likely to show some promiscuous activity with current enzymes.

Quantitative measures of substrate specificity and promiscuity would be useful additions to the meta data associated with PDB structures, helping to guide biologists in their selection of starting points for de novo enzyme design. Understanding the evolution of function and molecular mechanism may also help to generate more knowledge for enzyme design.

The study of the evolution of substrate specificity is an active avenue of research attempting to link the flexibility, modularity, and stability of enzyme structure to function and fitness. The potential of exploiting enzyme promiscuity to give commercial catalysts of outstanding efficiency and specificity is beginning to be fully realized.

Conflict of interest

Nothing declared.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest


One of the first enzyme evolution studies to apply bioinformatics methods to analyze change in function and catalytic machinery across structural domains on a large-scale basis.


A review of integrative approaches to explore enzyme evolution and functional diversity within enzyme superfamilies, highlighting recent large-scale experimental methods to determine activity profiles across enzyme superfamilies.


Highlights the power of using similarity networks to give insights about enzyme evolution from sequence and structural data.


Funtree is a unique resource bringing together protein sequence, structure and function information that allows the investigation of novel functionality within a structurally defined superfamily.


ECBLAST is a web tool to perform quantitative similarity comparisons between enzyme reactions at three levels: bond change, reaction center and compound structure.


Reaction Decoder Tool is a command-line tool to quantify the chemical similarity between reactions at three levels: bond change, reaction center and compound structure. It can be applied to quantify the size of evolutionary steps from a functional perspective.


This software performs and scores structural alignments. It can be applied to quantify the size of evolution steps from a structural perspective.


Proposes a rationalization for enzyme fold evolvability in terms of the connectivity of the core scaffold to the active site making some structures inherently more predisposed to the evolution of multiple functions.


