**Title: ?**

**Keywords:**

**Elementary functions, enzymatic function, evolution of function, enolase as a case study**

**INTRODUCTON**

Enzymes are essential macromolecules participating in many different processes in cells [REF]. To fully understand these processes one should know the precise role of the enzymes and how it is carried out. In order to affect these processes, a good place to start is to look at how nature achieved it by studying how the enzymatic function emerged and how it evolves.

The biochemical function of enzymes is to catalyze the transformation of substrate into product. The E.C. classifies the enzymes by the type of transformation and the substrate [EC]. There are 4585 biochemical functions in E.C. However, there can be several ways to perform the transformation, which is not captured by the E.C classification. The mechanism represents the description of the transformation as a sequence of elementary chemical reactions [IUPAC definition reference]. MACiE, the database of biochemical mechanisms, distinguishes between 335 different mechanisms. Amino acid residues involved in the mechanisms, in turn, have a number of roles - elementary functions - limited to few dozens. These numbers suggest that the whole diversity of enzymatic functions could be a result of (re)combination of elementary chemical reactions in the evolution. How are then these mechanisms formed and how are they supported by structure?

Earlier studies show that protein function can be represented as a combination of its basic units, Elementary Functional Loops (EFLs), segments of peptide chains of 25-30 amino acid residues possessing one or few residues that provide an elementary function [Closed loops, REFs\_Bioinfo1]. Previously, EFLs have been used to study emergence of new enzymatic functions by (re)combining already existing EFLs and reutilizing folds [BMC Evol Biol] and to explore the evolutionary relations between enzyme superfamilies [bioinfo 1, 2].

In this work we zoom in and focus on the evolution of function within a superfamily. Enzymes within a superfamily share a common fold and presumably have a common origin, but they can have different mechanisms and functions [REF].

Mechanistically diverse enzymatic superfamilies are of particular interest, because all members retain some conserved aspect of the function, such as, for example, common transition state, reaction step or intermediate stabilization [REF\_Babbit Mechanistically Diverse]. Enolase, one of the well studied mechanistically diverse enzymatic superfamilies in which enzymes share a common reaction step, is used as the subject of this study. In SFLD [REF SFLD] Enolase includes more than 10000 proteins in 20 families, corresponding to different functions. These proteins share a common fold consisting of two domains: α+β-capping domain in N terminal (determinant for substrate specificity) and a β/α-barrel in C terminal representing catalytic machinery. The common reaction step is abstraction of α-proton of a carboxylate substrate mediated by conserved residues that tend to be located in C terminal domain at the ends of the β-strands that constitute an active site. The families are arranged in seven subgroups by positions of the residues, involved in the conserved reaction step.

Our goals here are to identify EFLs corresponding to the elementary functions in the catalytic domain in Enolase, to find out if there exist common EFLs for the conserved reaction step, and to explore how the mechanisms are built from the EFLs.

In this work we derive a set of sequence profiles to represent EFLs. We use these profiles to show how proteins are related to each other by possessing different/similar EFLs. We also show that profiles can be used to find related proteins that should belong to Enolase superfamily and to predict their mechanisms.

**RESULTS**

From 80 derived sequence profiles that originate from different parts in structure (SM1 – all profiles) we picked those from C terminal domain that constitute the active site (Table 1). We annotate each profile using the information about elementary function or binding carried out by its key residues (many of them are identified in literature) together with the position of the elementary functional loop (EFL) in the β/α-barrel (the number of the β-strand). Figure 1 illustrates the structural location of the most common profiles in the C terminal. EFLs corresponding to these profiles are highlighted with different colors. The profile 101 has the α-proton abstraction elementary function in Glucarate dehydratase, Mandelate racemase, and Muconate cycloisomerase subgroups, where the catalytic Lysine is located at the end of the first β-strand (red EFL in Figure 1). In addition to the catalytic residues, the profiles of EFLs also describe the context by including the neighboring residues. The signatures (logo) in Figure 1 show several conserved positions, suggesting that they may be indirectly involved in the function. Profile 102 with the metal binding function on the second β-strand (yellow EFL in Figure 1) finds matches in four subgroups: Glucarate dehydratase, Mannonate dehydratase, Mandelate racemase, and Muconate cycloisomerase. Although the metal-binding Aspartate is conserved throughout the Enolase superfamily, in other three subgroups the context surrounding the Asp residue differs. The profiles 8 and 36 shown in Table 1 describe the specific context of the metal-binding EFL in Enolase and Methylaspartate ammonia-lyase, respectively. Another profile with the metal-binding function (profile 103, green EFL) is located at the third β-strand and has the common signature –E[DEQGS]P- in all the subgroups. The signature at the end of the fifth β-strand (profile 1054, blue EFL) is also similar in most of the subgroups -[KD]xxxxGG- except Mannonate Dehydratase. Depending on whether Lysine or Aspartate faces the active site, the elementary function is α-proton abstraction (in Enolase, Methylaspartate ammonia-lyase, and Muconate cycloisomerase) or pKa control (in Glucarate dehydratase and Mandelate racemase). In this case the context (two Glycines) is preserved even when the catalytic residue and the elementary function change. The fifth beta strand may have different profiles (Table 1) and different elementary functions depending on the family.

**Mechanism relations in families**

Enolase superfamily consists of 20 families with different biochemical functions. We use the profiles listed in Table 1 to show how the families are connected by the elementary functions. Figure 2 shows a graph of relations between families (ovals) and profiles (diamonds), where the edges denote matches between them. The profiles and their adjacent edges are colored by elementary function. For instance, there are 4 profiles finding matches in Enolase family: three having the metal binding elementary function (red) on 2β (the profile 8), 3β (103), and 4β (10) and a profile with pKa control elementary function (green) on 5β (1052). One can also study which families possess particular EFLs. For instance, profile 78 on 4β possessing metal binding residues have matches with 6 families in MR subgroup and MD family, meaning that these families all possess such EFL.

GCRD is an example of how the biochemical function and the mechanism is constructed as combination of elementary functions. There are profiles that match each of 7 β strands in this family.

Bigger subgroups are not that homogeneous in terms of profiles and elementary functions, it could be hard to identify a common mechanism in the subgroup. As in case of MC, some families have matches with only common profiles (such as MC (anti)), indicating that function mechanisms in this family are diverse compared to other families. In MR subgroup proteins of DGR family have matches with profile 78, but not profile 101, while DAD family proteins have matches with profile 101 but not 78. This indicates that certain EFLs of families from different subgroups can be more similar than of families inside one subgroup.

**Superfamily prediction**

Figure 2 shows that some profiles find matches with bigger number of families/subgroups than others. Therefore, we intended to find the minimal set of profiles that would manage to identify most members of Enolase superfamily and to find other proteins possessing these EFLs that are likely to be enolase-related. In order to show that, we introduce Enolase superfamily predictor. Predictor is represented by a set of profiles, such that protein is classified as superfamily-related based on a match with at least one of the predictor’s profiles. Predictor's performance is measured by sensitivity (or superfamily coverage).

We suggest a procedure, when at each step we choose a profile, which together with profiles from previous steps, achieves best superfamily coverage. We show result of the procedure for Enolase in Table/Figure X (file EP\_cumulative\_coverage\_greedy). Depending on the purpose of predictor and defined coverage threshold, sets of different sizes can be used. If the threshold is above 85%, then predictor consisting of profiles 1054, 102, and 101 would be good enough. For coverage threshold above 90% profiles 46 and 78 should be added to predictor as well. Union of all the profiles used in this work achieves sensitivity of 94,6537%.

**Prediction of functional mechanisms**

As we showed in Figure 2, the combinations of profiles (and hence combinations of elementary functions) define the biochemical functions and their mechanisms. Now we show that profiles can also be useful in prediction of mechanisms and functions. Since we are interested in identifying the individual groups of proteins with particular mechanism/function, we exclude profiles that find matches in several subgroups (3 or more). We perform a search through UniRef90 (non-redundant UniProt with 90 percent or less sequence identity) with the remaining profiles. For each protein sequence, we store the information about their matches with the profiles. Then we cluster together all proteins that have matches with exactly the same set of profiles (see M&M). Figure 3 shows the resulting graph with profiles and proteins as nodes and edges between them representing the sequence matches. The graph consists of 5 main connected components. The nodes representing proteins are colored by the subgroup to which they belong according to the SFLD database. The rest of the protein nodes are colored in black: these are proteins that are either not in SFLD or they are in SFLD but which don’t belong to any subgroup. It appears that clusters mainly contain nodes of the same color, pointing to that the proteins in the same cluster are likely to have the same mechanism. The more profiles match the cluster (and less colors are represented in a given cluster), the more confident one can be about the mechanism in the cluster. This allows one to predict the mechanism/function of the proteins not annotated in SFLD (black nodes), because they are likely to have similar mechanisms as annotated proteins in the same cluster.

Table 2(SM?) shows the detailed information about each cluster. The biggest cluster in Component 1 have matches with profiles 8 and 10 corresponds to EFLs having the metal binding elementary function on 2β and 3β, respectively (Table 1). This cluster contains 1089 red nodes - proteins that belong to Enolase subgroup in SFLD, and 299 black nodes - 6 of them are proteins in Enolase superfamily in SFLD but that are not assigned any subgroup, and 293 that are not in the superfamily. We conclude that these 299 proteins have Enolase subgroup-specific metal binding EFLs and, therefore, are likely to have similar function/mechanisms.

Component 3 includes clusters of both green (MR) and orange (MD) colors. These clusters, corresponding to different subgroups, are mainly connected by profile 78. This, again, indicates that all these proteins possess similar EFLs.

Since the approach is based on profile sequence comparison, we believe that it can be used in other, less investigated, superfamilies to give a clue about similarities of mechanisms in different groups of proteins, indicate their differences, and find new unannotated proteins.

**CONCLUSIONS**

Originally, EFLs were devised as units of biochemical function that allow one to survey emergence and evolution of enzymatic functions, and to find evolutionary connection between contemporary functional superfamilies and folds even though they are apparently unrelated. Here, we apply EFL-based approach to the analysis of evolutionary relations inside the Enolase superfamily. We show that the profiles of EFLs derived on proteins from different functional families provide sufficient resolution for establishing the minimal set of profiles that can serve as a predictor for the whole Enolase superfamily, as well as for predicting functional mechanisms.

**Most important result:** EFL-based approach allowed to show basis for grouping Enolase into one superfamily, which was not possible earlier based on the comparison of complete

proteins.

**Second:** EFLs allowed to show that evolution of mechanisms inside subgroups does not follow simple tree-like scenario, the same profiles can be found in different subgroups (as in the case of …) revealing peculiarities in emergence and evolution of different families.

**Third:**

Separation of the "Enolase-superfamily" function into EFLs allows to predict Enolase function in unannotated protein sequences. It provides 20 percent in addition to annotated enolases.

**Future directions:**

Enolase superfamily proved to be a perfect case study and a benchmark for the EFL-based approach to the evolutionary analysis of the protein function.

**Materials and Methods**

**Clustering of protein matches**

By removing profiles the number of components will increase together with increase of information about mechanisms and loss of some proteins, such that the more profiles we have, the more certain are proteins mechanisms.