VOL.7 NO.1March 2018 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: http://scjournal.ius.edu.ba

IUS Soft Computing Research Group

A Computational Biology Approach in Function Annotation to Enzymes

M. Ljubijankić

Faculty of Engineering and Natural Sciences,
International University of Sarajevo International University of Sarajevo,
Hrasnicka cesta 15, Ilidža 71210 Sarajevo,
Bosnia and Herzegovina
maidaaa-k@hotmail.com

Article Info

Article history:

Article received on 13 January 2018
Received in revised form 25 February 2018

Keywords:

Enzyme Commission;function annotation; longest common subsequence

Abstract

Although homologous proteins do not necessarily exhibitidentical biochemical functions, local and global sequence similarity is widely used as an indication of functional identity. Enzyme Commission (EC) classified hundreds of thousands of enzymes into six essential classes. Then in each class, enzymes are given four digits numbers such that enzymes with identical functions carry the same EC number. EC numbers provide a well-defined, non-ambiguous method for annotation of enzyme function. In this article, in each of six enzyme class, enzymes are classified according to their EC numbers into enzyme subclasses. Among enzymes and enzyme subclasses a new similarity measure is defined, and it is seen that, similar enzymes according to this new similarity measure exhibit identical biochemical functions in 94% of the cases. This similarity measure is used for function annotation to enzymes, and an average accuracy rate of 94% is achieved. The technique is also used for function annotation to unknown enzymes.

1. INTRODUCTION

It is noticed that many reactions have EC numbers printed next to them. In the early days of enzyme science, many different enzymes were given the same name and, on the other hand, several different names were assigned to the same enzyme. To make a systematic enzyme classification, Dixon and Webb introduced a system in their 1958 book “Enzymes”, based on the reaction catalyzed by the enzyme. It provided the foundation for the current classification system. At about the same time, the International Union of Biochemistry has decided to form an official international commission on enzymes to develop a better classification and naming system. The first full report of the commission was published in 1965, using a six-category system that is still used today. In general, each enzyme receives a unique four-component identification number that also provides insight into the enzymatic activity, lists of names and synonyms,

references, and often commentary. Details about the classification system can be found at

http://enzyme-database.org/rules.php

and

https://en.wikipedia.org/wiki/Enzyme_Commission_number

2. ENZYME CLASSIFICATION TODAY

Besides different classifications based on the sequence and structure, proteins may be classified according to their function. Enzyme Commission initiative is one such classification, where enzymatic functions are classified in a hierarchical four-level numbering system (Alborzi, Devignes and Ritchie, 2017).

Classification and naming of enzymes have been a confusing and difficult task in the past with names usually representing little or ambiguous information about the enzyme itself. The rapid increase in the number of enzymes discovered made scientists to reconsider the classification and develop a method that will name them systematically. First such attempt was made in the 1950s when scientists started classifying enzymes in terms of their function, rather than by their structures. In the years after, classifications were based on the number of molecules involved in the reaction or according to the type of reaction catalyzed. All these attempts and classifications were the beginning of the current enzyme classification and nomenclature system (McDonald and Tipton, 2013).

They are now identified and named systematically with an EC number which actually represents a four-level description code that is used to classify enzymes depending on the overall chemical reactions (Martínez Cuesta et al., 2015). Four levels are described as follows:

1.First level represents the main class (division) to which enzyme belongs. There are 6 main classes and are classified according to the type of chemistry being carried out. The 6 main classes include:
EC 1 – Oxidoreductases catalyze oxidation/reduction reactions
EC 2 – Transferases transfer a chemical group
EC 3 – Hydrolases catalyze the hydrolysis of chemical bonds
EC 4 – Lyases cleave chemical bonds
EC 5 – Isomerases catalyze geometric and structural changes within a molecule
EC 6 - Ligases join two compounds coupled with the hydrolysis of a diphosphate bond in ATP or a similar triphosphate.

2.Second level indicates a subclass and describes chemical substrate type.

3.Third level shows sub-subclass and defines a more specific enzyme substrate class.

The second and third levels are characterized by various criteria such as the chemical bond cleaved or formed, the reaction center, the transferred chemical group, and the cofactor used for catalysis.

4.Fourth level describes the substrate specificity.

General idea behind the EC classification is that enzymes are specific for a particular substrate. However, as many enzymes with the ability to catalyze more than one reactionwere discovered, EC classification was not unique and specific for one reaction. There was a need to add additional chemical reaction catalyzed by that same enzyme into the classification (Martínez Cuesta et al., 2015). Moreover, some enzymes were found to be quite good at performing this additional function. This ability to perform alternate chemistry is known as catalytic promiscuity. The most common type of promiscuity is the ability of enzyme to catalyze one chemical reaction with different substrates, referred to as “substrate ambiguity” (Pandya et al., 2014). Besides the phenomenon of substrate ambiguity, the promiscuity can include other, different phenomena, depending on the circumstances. Duarte, Amrein and Kamerlin (2013) reviewed several types of promiscuity: substrate promiscuity describes the catalysis of the same reaction with different substrates; catalytic promiscuity is the catalysis of chemically distinct reactions with different transition states; conditional promiscuity refers to the catalysis of different reactions under conditions different than the native one; product promiscuity is the ability to generate different products through the same reaction. Catalytic promiscuity is further divided into accidental and induced promiscuity where accidental describes the catalysis of non-native reactions which are catalyzed by the native, wild type enzymes, whilst induced represents a system with a completely new reaction as a result of mutations (overview in Figure 1).

The promiscuity is of great importance for enzyme evolution and is among the most common ways to evolve new enzyme functions. In terms of function, it can be advantageous to the cell through several mechanisms such as 1) proofreading, 2) scavenging of nutrients, 3) removal of antimetabolites, 3) balancing of metabolite pools, and 4) establishing system redundancy (Pandya et al., 2014).

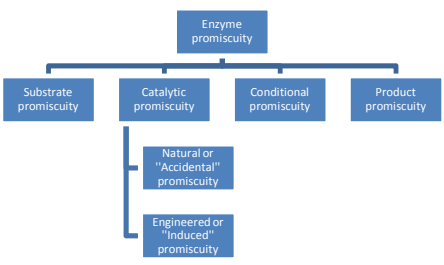

Figure 1. Schematic overview of enzyme promiscuity types as described in the main text.

When enzymes are classified according to six main classes, and EC numbers that are related to functions, their distribution in main classes and subclasses are as in Table 1, as of February 2018.

Table 1. Number of enzymes in six main classes, numbers of different functions in each class and function subclasses withmore than four enzymes.

Class	Enzymes	Functions	FSClass Enz> 4
1	39,869	1342	548
2	54,136	1043	478
3	67,213	1104	625
4	25,079	550	239
5	14,462	275	134
6	29,742	188	113
Total	230,501	4,472	2137

3. FROM SEQUENCE TO FUNCTION

In the past few decades, a lot of effort is made to predict structures and inter functions of proteins directly from the sequence.Although the possession of sequence similarity is usually indicative of underlying structural similarity, functional similarity prediction by sequence-based methods remains less reliable and annotating the function from sequence alone may be a challenging task(Lee, Redfern and Orengo, 2007; Petsko and Orengo, 2004). The most important question to be considered in this approach to function prediction is: ''what sequence similarity measures/thresholds should be used for the safely transferring function between related proteins?'' Besides this issue, nature provides us with examples where underlying sequence similarity doesn't imply functional similarity. Several studies in the past few decades have investigated this issue and tried to elucidate the sequence-function relationship. Most authors agree that sequence identity > 40% between two proteins can be enough to say that they share a common function (Table in the Appendix). Some authors also reported that approximately 90% of pairs of proteins with sequence identity > 40% conserve all four EC numbers. The authors also investigated the level of accuracy in annotation process and concluded that >60% pairwise sequence identity is required for a transfer with less than 30% errors and for errors below 10%, >75% sequence identity (Rost et al., 2003).

Important to emphasize is the fact that even identification of the same, common biochemical function does not mean the same or similar cellular and other higher-level functions. However, local alignments of motifs can usually identify at least one function of the protein. If long enough, motifs can be identified as a domain with a particular structure and function, what is of great importance for function prediction.

In the past few decades a great effort has been made in studying enzymes and enzyme classes since EC classification brought some questions important to answer, such as: How diverse is the function in one EC class? What sequence similarity can be observed between enzymes in the same EC class? Can sequence and structural information be used to group an uncharacterized protein in a particular EC class?

Accordingly, several ways have been developed to associate proteins from structural databases and EC numbers. SIFTS is one such approach and it represents a collaboration between the Protein Data Bank in Europe and UniProt where PDB chain entries are linked to external biological resources such as Pfam, and IntEnz. However, only few approaches have been developed for automatically assigning EC numbers to structural domains. The dcGO ontology database provides information about Pfam domains associated with a particular GO term using the idea that association of GO term to UniProtKB protein which contains a particular domain is automatically association of that GO term to that domain.

The principle and idea are quite straightforward but are only available for GO terms and not for EC numbers (Alborzi, Devignes and Ritchie, 2017). Mentioned approaches try to correlate structural information to enzyme function. There are, however, approaches which rely on sequence information in annotating enzyme function and try to classify particular enzyme to a particular EC class. Additionally, such approaches investigate functional diversity in enzyme classes and between homologous proteins.

Several studies have taken an advantage of supervised machine learning systems to try to decipher divergent functions in homologous proteins. Shah and Hunter (1998) in their work focused on the mapping the sequence information to functional enzyme class to try to identify divergent functions in homologous proteins. They investigated if conserved domains in proteins can be used to identify alternative functions by using machine learning systems and tried to assign the correct EC class to similar sequences, based on the modular structure of the proteins.

Their results suggested that such approach may be useful in discrimination among functionally distinct homologs. There were, however, several factors impeding the discrimination, such as the fact that sequence change due to a mutation may affect the catalytic activity and domains may not be expressive enough to capture these changes. Additionally, enzyme promiscuity, as already discussed, may influence a one-to-one mapping between proteins and functions.

However, a recent work done by Baier, Copp&Tokuriki (2016) reviewed recent studies which systematically characterize the enzyme promiscuity to provide insights into the functional repertoire and evolutionary potential between subgroups within a superfamily. Such studies, as

stated: “performed large-scale function profiling, in which a diverse set of enzymes belonging to the same enzyme superfamily is assayed against a set of substrates in an “all versus all” manner”. Such approach provided information about the extent to which distinct functions are connected via promiscuous activities.

In addition, initiatives and databases are making the investigation at any level easier and efficient. Some of them, such as The Structure Function Linkage Database (SFLD) and Enzyme Function Initiative (EFI) incorporate and integrate information about protein sequence, structure and function at one place and provide a way to annotate new sequences more attentively.

4. A NEW SIMILARITY AMONG ENZYMES AND ENZYME CLASSES

As seen from table in the Appendix, sequence identity thresholds for functional similarity is too high. This sequence identity is computed through pairwise alignments. In this research, we define a new similarity measure among enzymes and enzyme classes. It will be shown that this similarity measure can be used for function annotation.

4.1 A New Similarity Measure Between the Two Enzymes

The new similarity measure among enzymes is defined as the longest common subsequence of the two enzymes as in Figure 2. Although the motif QLAE is also common between the two enzymes, it is not counted since it is not the longest. Longest common subsequence may be the conserved domain which is responsible for the common functions of the two enzymes.

LongestCommonSubsequence[

VSFDQLAEIGVIYYNAKMQQEELDALATEREYK,

RDVVTLNQLAEAFNNDIDAYNAKMQQFYKEHY] =

YNAKMQQ

Figure 2. The number of amino acids (the length) of the longest common The string length of the subsequence YNAKMQQ = 7, is defined as the measure of closeness between the two sequences.

Closeness of an enzyme to a function subclass is the maximum of the closenesses of the enzyme to all enzymes in the subclass. If the enzyme is a member of the subclass, its closeness to itself is excluded from the list.

4.2 An Enzyme is the Closest to its Own Function Subclass

In each of the six classes, we consider function subclasses with more than four enzymes. When one enzyme is chosen randomly from each function subclass, and the closenesses

4.2 An Enzyme is the Closest to its Own Function Subclass

In each of the six classes, we consider function subclasses with more than four enzymes. When one enzyme is chosen randomly from each function subclass, and the closenesses

of this enzyme to all function subclasses are computed, this enzyme is found to be the closest to its own subclass around 94% of the times (Table 2).

Table 2. An enzyme is found to be the closest around 84.5% of the times to its own function subclass.

Class	Funct. SC	FSClass Enz> 4	%
1	1342	548	97
2	1043	478	97
3	1104	625	93
4	550	239	93
5	275	134	88
6	188	113	94
Total	4,472	2137	Av: 94

On the other side, if one enzyme israndomly chosenfrom 100 function classes fromeach of six classes the average closeness of enzymes from one class is the highest with its own class among others (Table 3).

Table 3. The average closeness of enzymes from one class is the highest with its own class among others.

Class	1	2	3	4	5	6
1	**132.5**	15.3	9.6	11.2	10.6	7.7
2	9.4	**114.9**	8.2	8.0	7.2	7.6
3	7.5	8.3	**103.3**	8.3	7.1	7.9
4	8.5	14.9	12.0	**116.7**	7.5	7.7
5	10.2	21.9	11.8	12.7	**87.4**	7.6
6	7.7	17.6	22.8	14.2	7.7	**131.8**

These two observations suggest a usage of the closeness measure in function annotation. When an enzyme is given with unknown function, the function of the subclass that is closest to the unknown enzyme may be annotated as the function of this unknown enzyme.

4.3 Function Annotation to Unknown Enzymes in PDB

To obtain dataset of proteins of unknown functions, we used PDB website’s advanced search option. The fields used for the advanced search are: (i) Text search containing the phrase “unknown function”; (ii) Experimental method is “X-ray”; and (iii) Macromolecule type is protein. A search on 29 April 2018 using the Advanced Search web interface of the PDB yielded 3482 X-ray crystallographic protein structures that are annotated as proteins of unknown function. Since our proteins of interest are enzymes, we checked refinements for obtained results. In the refinements, we used proteins present in

Enzyme Classification section which are grouped here mostly according to data from an external resource, UniProtKB database. After filtering the results, a total of 477 PDB entries were described as enzymes and were used for our analysis. Among these, 163 are hydrolases, 111 transferases, 92 oxidoreductases, 62 lyases, 34 isomerases and 15 ligases.

Since among these 477 enzymes of unknown function some had EC numbers attached, we checked protein annotation rules in PDB. Generally, annotations and data about the protein function are integrated into PDB from different external resources (description of all resources used by PDB is available at:

http://www.rcsb.org/pages/external-resources).

UniProtKB (http://www.uniprot.org)provides the most information related to the function, but also data about catalytic activity and subunit structure. Regarding EC information, PDB uses data from ENZYME database available at ExPASy (http://expasy.org) and UniProtKB database.

The main source for the data in the ENZYME database comes from recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) with minor part of the data being extracted from the literature.

EC numbers in UniProtKB database are assigned according to data from publications or according to UniProt rules, the so-called UniRule. In turn, the annotations and conditions for a particular UniRule are derived from homologous proteins whose annotations are experimentally determined and for which publication is available. Thus, if there are enough information about homologous proteins for which information and publications are available, even newly characterized proteins can have assigned EC number in UniProtKB. On the other hand, ENZYME does not assign EC numbers for newly characterized enzymes.

Accordingly, although the protein function in PDB is characterized as unknown, if EC number and function is annotated according to at least the UniRule, that particular EC number will appear in PDB as well.

4.4 Function Annotation by the Closest Function Subclass

To use closeness measure for function annotation, 119 non redundant enzymes with unknown function is chosen from PDB database, some of which are annotated functions using several ad hoc techniques.

Table 4. 119 unknown nonredundant enzymes are taken from each of six enzyme groups some of whichhave annotated functions using several ad hoc techniques.

Group	Unknown Enzymes
1	21
2	27
3	42
4	11
5	7
6	11
Total	119

For each of these 119 enzymes, the closenesses of all 4,472 function subclasses are computed, the one which is the closest to this enzyme is found and the function of this subclass is annotated as the function of this enzyme.The annotated functions to theseenzymes are as follows:

Table 5. Class 1: 21 unknown enzymes

PDB	C Prediction	UniRule
1A4U	1.2.1.71	1.1.1.1
1BIQ	1.-.-.-	1.17.4.1
1CPO	1.3.7.7	1.11.1.10
1DRA	1.2.1.71	1.5.1.3
1A27	1.2.1.38	1.1.1.62
1FRV	2.1.2.13	1.12.2.1
1AP5	1.3.98.1	1.15.1.1
1NPD	1.1.1.25	1.1.1.25
1TEH	1.2.1.71	1.2.1.1
2HI1	1.2.1.38	1.1.1.262
2JIM	1.2.1.71	1.7.99.4
2YYY	1.1.1.25	1.2.1.59
3E9N	1.20.4.4	1.1.1.36
3EGO	1.1.1.25	1.1.1.169
3KH5	1.20.4.4	1.1.1.34
3KZV	1.20.4.4	1.-.-.-
3OAJ	1.1.1.25	1.13.11.-
3S4M	1.2.1.71	1.16.3.1
4F3Y	1.14.11.16	1.3.1.26
5H2F	1.1.1.300	1.10.3.9
5M8L	1.3.7.7	1.14.18.-

Table 6. Class 2: 27 unknown enzymes

PDB	C Prediction	UniRule
1A7J	2.1.1.43	2.7.1.19
1ACM	3.4.19.12	2.1.3.2
1AXW	2.1.1.	2.1.1.45
1BQD	2.1.2.3	2.6.1.1
1HQM	2.1.1.43	2.7.7.6
1J7D	2.1.1.43	
1J7D	2.1.2.13	6.3.2.19
1JQE	2.1.2.3	2.1.1.8
1NXZ	2.1.1.	
1ON0	2.1.1.43	
1OXQ	2.1.1.319	
1OXQ	2.1.1.43	
1ECJ	2.1.3.2	2.4.2.14
1PT8	2.1.1.43	
1SG9	2.1.3.2	
1VH0	3.4.21.91	
1HQM	2.1.1.43	2.7.7.6
1XMT	2.1.1.43	
2GA8	2.1.1.43	
2OVF	2.1.1.43	
2QGQ	2.1.1.43	
1HQM	2.1.1.43	2.7.7.6
3B76	2.1.1.319	6.3.2.-
1PT8	2.1.1.43	
3DO8	2.1.1.43	2.7.7.3
3EWB	2.1.1.	2.3.3.13
3FLO	2.1.1.43	

Table 7. Class 3: 42 unknown enzymes

PDB	C Prediction	UniRule
1AI8	3.1.3.16	3.4.21.5
1AKO	3.1..	3.1.11.2
1AUK	2.7.7.49	3.1.6.8
1CEF	3.1.13.	3.4.16.4
1ELF	3.1.1.29	3.4.21.36
1FKW	3.1.1.3	3.5.4.4
1FWI	3.1.2.2	3.5.1.5
1GPL	3.1.3.48	3.1.1.3
1GPQ	3.1.1.	3.2.1.17
1T9H	3.1.3.48	3.6.1.-
1ANI	3.1.26.11	3.1.3.1
1W8I	3.1.6.	3.1.-.-
1ZZM	3.1.6.	3.1.21.-
2CMU	3.1.6.	3.5.3.6
2D69	3.4.19.12	3.1.3.5
2FKB	3.1.4.	3.6.-.-
2R11	3.1..	3.1.1.1
2RLA	3.1.13.	3.5.3.1
2W87	3.1.4.4	3.2.1.-
1W8I	3.1.6.	3.1.-.-
3CS0	3.1.26.11	3.4.21.-
3ENH	3.1.3.48	3.4.24.57
3FEF	3.1.26.11	3.2.1.67
3K35	3.1.6.1	3.5.1.-
1T9H	3.1.3.48	3.6.1.-
3CS0	3.1.26.11	3.4.21.-
3PBG	3.1.21.	3.2.1.85
3PGA	3.1.1.29	3.5.1.1
3QUQ	3.1.6.	3.6.1.1
2RLA	3.1.13.	3.5.3.1
2Q47	3.1.1.3	3.1.3.48
3SF8	3.1.6.	3.4.-.-
3V2I	3.4.19.12	3.1.1.29
4DW8	3.1.6.	3.8.1.-
4DZ4	3.1.26.11	3.5.3.11
4N0N	3.4.19.12	3.4.19.12
4PII	3.1.1.3	3.2.2.-
3RSD	3.1..	3.1.27.5
5B7I	3.1.6.	3.1.-.-
5C3I	3.1.6.	3.6.4.12
5PTP	3.1.26.11	3.4.21.4
8PCH	3.1.1.29	3.4.22.16

Table 8. Class 4: 11 unknown enzymes

PDB	C Prediction	UniRule
1BKS	4.3.2.1	4.2.1.20
1SQC	4.2.3.4	5.4.99.-
1TO3	4.2.3.4	4.1.-.-
1XV2	4.2.1.1	4.1.1.5
1YAS	4.2.3.5	4.1.2.39
2DGD	4.2.1.1	4.1.1.76
2HNE	4.2.3.4	4.2.1.68
3KKL	4.3.2.1	3.2.-.-
3RPH	4.3.2.1	4.2.1.93
3T6C	4.2.1.10	4.2.1.-
4PII	4.2.1.10	4.2.99.18

Table 9. Class 5: 7 unknown enzymes

PDB	C Prediction	UniRule
1A0C	3.1.1.64	5.3.1.5
1DTN	5.3.1.6	5.1.2.2
1I60	5.3.1.6	
1PIN	5.3.3.19	5.2.1.8
1SQC	5.3.1.6	5.4.99.-
2A6P	5.3.1.6	5.4.2.1
4G9B	5.3.1.6	5.4.2.6

Table 10. Class 6: 11 unknown enzymes

PDB	C Prediction	UniRule
1NZJ	6.3.1.1	
1R8G	6.3.1.1	6.3.-.-
1Z2U	6.3.5.4	6.3.2.19
2FO3	6.3.5.4	6.3.2.19
2H2Y	6.3.5.4	6.3.2.19
2R84	6.3.1.1	
3D54	6.3.1.1	6.3.5.3
3D54	6.3.5.4	6.3.5.3
4Q5E	6.3.5.4	2.7.-.-
4Q5E	6.3.5.4	6.3.2.19
4XOM	6.3.1.1	

The discrepancy between closeness prediction and UniRule prediction is not essential. Such a situation is observed in the example of the survival protein E (SurE) from *Thermotogamaritima* (PDB ID 1ilv). Zhang and colleagues (2001) determined the crystal structure of this protein and in their work, described it as having the function of acid phosphatase (EC 3.1.3.2). Additionally, manual assertion inferred from sequence similarity was done for the same protein in UniProtKB database (UniProt accession code: P96112), describing the function of acid phosphatase as well. However, the new annotation was later added by UniProtKB's type of evidence that is used in manual assertions, the so-called curator inference evidence. For this type of evidence, the information has been inferred by a curator based on his/her scientific knowledge or on the scientific content of an article (Zhang, et.al. 2001), and new EC number is assigned to the protein (EC 3.1.3.5), describing a new function – nucleotidase.

For the case of survival protein E (SurE) from *T. maritima*with sequence,

```
>SurE_Paerophilum

MKILVTNDDGVHSPGLRLLYQFALSLGDVDVVAPESPKS
ATGLGITLHKPLRMYEVDLCGFRAIATSGTPSDTVYLAT
FGLGRKYDIVLSGINLGDNTSLQVILSSGTLGAAFQAAL
LGIPALAYSAYLENWNELLNNKEAVEIMGAVVSSTASYV
LKNGMPQGVDVISVNFPRRLGRGVRAKLVKAAKLRYAQQ
VVERVDPRGVRYYWLYGRDLAPEPETDVYVVLKEGGIAI
TPLTLNLNAVDAHREVDMDSLNRMVEYINASL
```

is today in EC database, and annotated to the function EC 3.1.3.5. If it is disregarded, the next closest enzyme has a longest common subsequence of length 35 and is still in function subclass of EC 3.1.3.5. Therefore, our technique would annotate the same function, even if it was not in EC database.

Moreover, Proudfoot et al. (2004) clearly demonstrated that the annotation of SurE proteins as acid phosphatases is not accurate. They explained that in contrast to nonspecific acid phosphatases, SurE proteins from *E. coli*, *T. maritima*, and *P. aerophilum* show strict specificity to nucleoside 5'(3')-monophosphates and, accordingly, should be annotated as 5'(3')-nucleotidases.

The survival protein E (SurE) from *E. coli*, with sequence,

```
>SURE_Ecoli

MRILLSNDDGVHAPGIQTLAKALREFADVQVVAPDRNRS
GASNSLTLESSLRTFTFENGDIAVQMGTPTDCVYLGVNA
LMRPRPDIVVSGINAGPNLGDDVIYSGTVAAAMEGRHLG
FPALAVSLDGHKHYDTAAAVTCSILRALCKEPLRTGRIL
NINVPDLPLDQIKGIRVTRCGTRHPADQVIPQQDPRGNT
LYWIGPPGGKCDAGPGTDFAAVDEGYVSITPLHVDLTAH
SAQDVVSDWLNSVGVGTQW
```

is also in EC databasetoday and annotated to the function EC 3.1.3.5. If it is disregarded, the next closest enzyme has a longest common subsequence of length 237 and still in function subclass of EC 3.1.3.5. Therefore, our technique would annotate the same function, even if it was not in EC database.

When SurE protein details are accessed in PDB, one will notice that EC number describing the protein is EC 3.1.3.5, suggesting that PDB updated the data as the external resources data are updated. Interesting to note, however, is the annotated function in the header of this PDB entry, which is still ''structural genomics unknown function''.

5. DISCUSSION

Although homologous proteins do not necessarily exhibitidentical biochemicalfunctions, local and global sequence similarity is widely used as anindication of functional identity. In this article between enzymes and enzyme classes a new similarity measure is defined by the longest common subsequence, and it is seen that, similar enzymes according to this new similarity measure exhibit almost identical biochemicalfunctions. This similarity measure is used to annotation of enzyme functions, and an average accuracy rate of 94% is achieved. The same measure is used to annotate functions to 119 unknown enzymes, and its success is shown in two enzymes which previously had assigned the wrong EC number in UniProtKB, but were correctly assigned immediately with this similarity measure.

REFERENCES

Addou, S., Rentzsch, R., Lee, D., &Orengo, C. (2009). Domain-Based and Family-Specific Sequence Identity Thresholds Increase the Levels of Reliable Protein Function Transfer. Journal Of Molecular Biology, 387(2), 416-430. http://dx.doi.org/10.1016/j.jmb.2008.12.045

Alborzi, S., Devignes, M., & Ritchie, D. (2017). ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains. BMC Bioinformatics, 18(1).

Baier, F., Copp, J., &Tokuriki, N. (2016). Evolution of Enzyme Superfamilies: Comprehensive Exploration of Sequence–Function Relationships. Biochemistry, 55(46), 6375-6388.

Devos, D., & Valencia, A. (2000). Practical limits of function prediction. Proteins: Structure, Function, And Genetics, 41(1), 98-107.

Duarte, F., Amrein, B., & Kamerlin, S. (2013). Modeling catalytic promiscuity in the alkaline phosphatase superfamily. Physical Chemistry Chemical Physics, 15(27), 11160.

Lee, D., Redfern, O., &Orengo, C. (2007). Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*, *8*(12), 995-1005.

Martínez Cuesta, S., Rahman, S., Furnham, N., & Thornton, J. (2015). The Classification and Evolution of Enzyme Function. Biophysical Journal, 109(6), 1082-1086.

McDonald, A., & Tipton, K. (2013). Fifty-five years of enzyme classification: advances and difficulties. FEBS Journal, 281(2), 583-592.

Pandya, C., Farelli, J., Dunaway-Mariano, D., & Allen, K. (2014). Enzyme Promiscuity: Engine of Evolutionary Innovation. Journal Of Biological Chemistry, 289(44), 30229-30236.

Petsko, G., and Ringe, D. (2004). *Protein structure and function*. Oxford: Oxford University Press.

Proudfoot, M., Kuznetsova, E., Brown, G., Rao, N., Kitagawa, M., & Mori, H. et al. (2004). General Enzymatic Screens Identify Three New Nucleotidases in Escherichia coli. Journal Of Biological Chemistry, 279(52), 54687-54694.

Rost, B. (2002). Enzyme Function Less Conserved than Anticipated. Journal Of Molecular Biology, 318(2), 595-608. http://dx.doi.org/10.1016/s0022-2836(02)00016-5

Rost, B., Wrzeszczynski, K., Ofran, Y., Nair, R., & Liu, J. (2003). Automatic prediction of protein function. Cellular And Molecular Life Sciences (CMLS), 60(12), 2637-2650.

Sangar, V., Blankenberg, D., Altman, N., &Lesk, A. (2007). Quantitative sequence-function relationships in proteins based on gene ontology. BMC Bioinformatics, 8(1), 294. http://dx.doi.org/10.1186/1471-2105-8-294

Shah, I., & Hunter, L. (1998). Identification of Divergent Functions in Homologous Proteins by Induction over Conserved Modules. In Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB-98) (pp. 157-164). Montreal: American Association for Artificial Intelligence (AAAI Press).

Tian, W., & Skolnick, J. (2003). How Well is Enzyme Function Conserved as a Function of Pairwise Sequence Identity?. Journal Of Molecular Biology, 333(4), 863-882.

Todd, A., Orengo, C., & Thornton, J. (2001). Evolution of function in protein superfamilies, from a structuralperspective 1 1Edited by A. R. Fersht. Journal Of Molecular Biology, 307(4), 1113-1143.

Wilson, C., Kreychman, J., & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. Journal of Molecular Biology, 297(1), 233-249.

Zhang, R., Skarina, T., Katz, J., Beasley, S., Khachatryan, A., Vyas, S., Arrowsmith, C. H., Clarke, S. Edwards, A., Joachimiak, A., and Savchenko, A. (2001). Structure of Thermotogamaritima Stationary Phase Survival Protein SurE. Structure, 9(11), 1095-1106

APPENDIX: Summary of studies exploring sequence identity threshold

	Dataset composition	Conclusions
2338 p	Homologous pairs of PDB structural domains or sequences	50% identity is required for conservation of all four EC digits. (Devosand Valencia, 2000)
n/a	29,454 representative pairs of structural domains from SCOP comparisons at various levels ofthe SCOP hierarchy (family, superfamily, fold)	40% identity is required for conservation of all four EC digits. (Wilson et al., 2000).
65303 p	Homologous pairs of structural domains from CATH and their sequence relatives from SwissProt and GenBank.	40% and 60% identity is required for conservation of all four EC digits in single- and multidomain proteins. (Todd et al., 2001).
26243 p	Whole protein sequences with an EC annotation.	Below 70% identity, both the first and the fourth EC digits start to diverge. (Rost, 2002).
22645 p	Whole protein sequence homologues including enzymes and non-enzymes (derived by PSI-BLAST).	40% (60%) identity is required for conservation of the first three (all four) EC digits. (Tian and Skolnick, 2003).
7868 f	Protein families with different numbers of members and with seed and full alignments of proteins in each family.	Between 40% and 60% sequence identity shows the highest change in the identical functions.The threshold at about 40% sequence identity, at which the observed behavior changes. (Sangar et al., 2007).
721 suf 3210 sf	3210 enzyme functional subfamilies identified in 721 CATH-Gene3D enzyme superfamilies.	For more than 60% sequence identity, proteins share the same EC number in 90% of cases. (Addou et al., 2008).