Visualizing the dynamics of enzyme annotations in UniProt/SwissProt

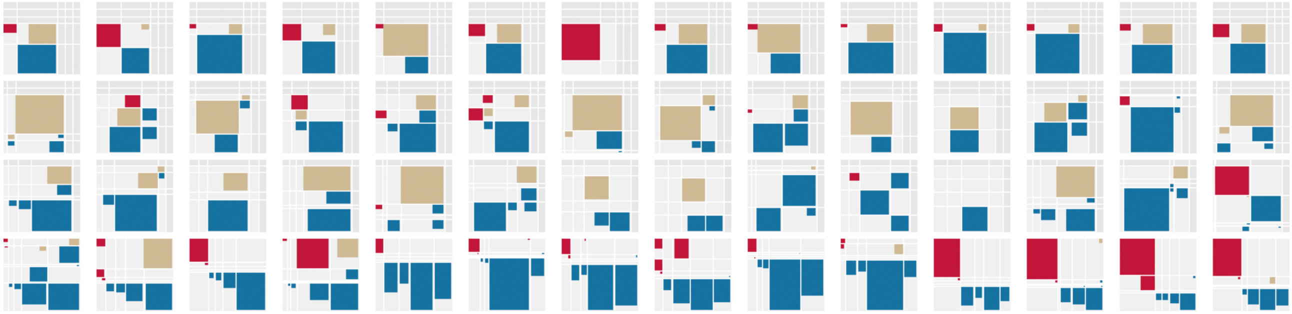
Sabrina A. Silveira[*]
Universidade Federal de Minas Gerais

Artur O. Rodrigues[†]
Universidade Federal de Minas Gerais

Raquel C. de Melo-Minardi[‡]
Universidade Federal de Minas Gerais

Carlos Henrique da Silveira[§]
Universidade Federal de Itajubá

Wagner Meira Jr.[¶]
Universidade Federal de Minas Gerais



Abstract

In this paper, we tackle the problem of visualizing evolution in enzyme annotations across several releases of UniProt/SwissProt data base. More specifically, we visualize the dynamics of the EC numbers which are a numerical and hierarchical classification scheme for enzymes, based on the chemical reactions they catalyze. An EC number consists of four numbers separated by periods and they represent a progressively finer classification of the catalized reaction. The proposed interactive visualization gives a macro view of the changes and presents further details on demand as, for instance, frequencies of types of changes segmented by levels of generalizations and specializations as well as by the enzyme families. Users can also explore entries meta data. With this visualizations we were able to evidence trends of specialization and growth of database as well as detect several exceptions where EC numbers were deleted, divided, created or annotation errors were detected.

Index Terms: Information visualization, Bioinformatics, Database dynamics, Enzymes, EC number, UniProt, SwissProt, Annotation, Processing.

1 Introduction

In recent decades there was a significant growth of biological data generated by experimental techniques such as the new generation DNA sequencing, protein sequencing and protein structure determination. Much of these data are organized and made publicly available to the scientific community in biological databases over the Internet. According to [13] these repositories not only store biological raw data but also relevant information related to them such as literature data, protein function, relationship between a protein and its encoding gene, among other meta data.

Given that these biological databases are growing at very high rates, most of these meta data are automatically assigned. In the majority of the cases, with no laboratory experiments at all, the roles of most genes in several organisms have been reported by homology propagation [4]. To ensure that these annotations remain reliable, studies about the confiability of the entries as well as measures of confidence should be developed. Many studies have called the attention to errors rates in the biological databases annotations [6, 9, 8, 12, 15, 11]

In fact, the automatic identification of theses errors is still an open problem and several challenges have to be faced. Without laboratory experiments to verify automatically assigned annotations, it is impossible to know for certain. However, most of the studies present comparisons of diverse functional annotation methods and show they are widely incompatible, which places a rough upper bound on their accuracy.

A major step toward automatic error detection is a description of how and to what extent biological databases entries annotations evolve. In other words, we have to be capable to understand why some entries seem to be more stable and and others more volatile and what are the factors that determines this different behavior.

The research and development of models and algorithms as well as visualization resources are very promising toward understanding how biological databases evolve. Interactive visualizations can be specially powerful to represent in a macro/micro perspective this voluminous, high-dimensional and complex datasets and to help users to unveil trends and exceptions in those data sets.

1.1 Enzyme annotations

By the late 1950's it had become evident that the nomenclature of enzymology, in a period when the number of known enzymes was increasing rapidly, was getting out of hand. In many cases the same enzymes became known by several different names, while conversely the same name was sometimes given to different enzymes. Many of the names conveyed little or no idea of the nature of the reactions catalysed, and similar names were sometimes given to enzymes of quite different types. To meet this situation, the General Assembly of the International Union of Biochemistry (IUB)

[*] e-mail: sabrina@dcc.ufmg.br
[†] e-mail: artur@dcc.ufmg.br
[‡] e-mail: raquelcm@dcc.ufmg.br
[§] e-mail: carlos.silveira@unifei.edu.br
[¶] e-mail: meira@dcc.ufmg.br

decided, in consultation with the International Union of Pure and Applied Chemistry (IUPAC), to set up an International Commission on Enzymes. Its objective was to consider the classification and nomenclature of enzymes and co-enzymes, their units of activity and standard methods of assay, together with the symbols used in the description of enzyme kinetics. The Commission prepared a report in 1961, it was adopted and has been widely used in scientific journals, textbooks, etc. since then. The size of the Enzyme Commission number (EC number) list has increased steadily since the publication of the first report and also many corrections were done.

The EC number is a numerical classification scheme for enzymes, based on the chemical reactions they catalyze. Every enzyme code consists of four numbers separated by periods. Those numbers represent a hierarchical progressively finer classification of the catalized reaction. For example, the code: 3.4.21.4 is a:

3: hydrolase, which means the enzyme breaks a chemical bond using a water molecule.

3.4: peptidase, which means the broken bond is a peptide bond, i.e., a bond between amino acids in a protein chain.

3.4.21: endopeptidase, because it breaks an intra-chain peptide bond.

3.4.21.4: trypsin, because enzyme has the specificity of cutting close the residues arginine and lisine.

When a new enzyme is annotated, one can add from one to four levels of the EC number, depending on the detail of existing knowledge. In the better case, we know all about the catalyzed reaction as well as the specific substrates and products involved. However, in many cases, when all the details about the catalytic activity are not known, partial EC numbers, in which hyphens are written in the unknown levels, are used to annotate enzymes. An EC number ”3.4.21.-”, for instance, means we don’t know enzyme substrates specifically although we have information about the reaction catalyzed.

In this paper, we model the problem of analyzing enzyme annotation dynamics and propose a technique to visualize the evolution of these annotations across several releases of UniProt/SwissProt data base. This paper is organized as follows: in section 2, we describe how we modeled the problem. Section 3 details the dataset presented in the visualization. In section 4, we talk about previous related researches and in section 5, we describe in detail the basis for the technique proposed as well as its capabilities. Finally, we discuss several insights we obtained in section 6 and conclude the work.

2 Problem modeling

Based on numerical and hierarchical nature of Enzyme Classification number, we proposed a model to characterize the EC changes observed over several versions of UniProt/SwissProt. First of all, our focus was on visualizing what types of changes happens and with what frequency they occur. Moreover, it is important to know the hierarchical level in which a change occurs, since a move in higher levels (leftmost) are more severe than in lower ones, thus we decided to segment changes by common prefix length, number of generalizations and number of specializations a specific EC number has suffered.

An example of EC number change characterized by our model is provided below.

$$3.1.3.2 \rightarrow 3.1.3.5$$

It happened in 77 Hydrolases of releases 5 to 6. Observe that the common prefix length is 3 (the first three levels from left to right remains the same), there was 1 generalization (number 2 was deleted) and 1 specialization (number 5 was written). This change means that an acid Phosphatase is now classified as a 5’-Nucleotidase.

More examples of EC moves characterized by our prefix / generalization / specialization model are provided in Table 1.

3 Data set

In this work we use the biological database UniProt [5], which aims to provide a centralized repository of protein sequences with comprehensive coverage and a systematic approach to protein annotation, incorporating, interpreting, integrating and standardizing data from a large number of disparate sources. It is the most comprehensive catalog of protein sequence and functional annotation. As stated by [5] the UniProt Knowledge base (UniProtKB) is an expertly curated database, a central access point for integrated protein information with cross-references to multiple sources.

In accordance with [1] UniProtKB consists of two sections, UniProtKB/SwissProt and UniProtKB/TrEMBL. SwissProt contains manually annotated records with information extracted from literature and curator-evaluated computational analysis. Annotation is done by biologists with specific expertise to achieve accuracy. TrEMBL contains computationally analyzed records enriched with automatic annotation and classification. As the SwissProt is considered the gold standard for protein annotation, in this work we use its data to observe and analyze the changes in EC annotation.

The major releases available in the ftp of UniProt database when this study was started (March 2009) were downloaded. We analyzed releases 1 (when SwissProt was integrated to UniProt) to 15 (the current release when this study was started).

In order to check if an EC move happened we need to look at a database entry EC annotation in two consecutive releases, therefore the mentioned releases were studied in pairs and the intersection of identifiers across two consecutive releases was taken.

The total number of entries as well as the number of entries annotated with EC number and its percentage for the fifteen used releases are provided in Table 2. Table 3 shows the number of entries in the set intersection of each release pair.

Table 3: Release pairs and number of entries in the intersection.

Release pair	Number of entries in $\cap$
1-2	141,249
2-3	151,318
3-4	162,812
4-5	166,933
5-6	181,005
6-7	193,382
7-8	207,069
8-9	222,181
9-10	241,189
10-11	260,065
11-12	269,152
12-13	276,011
13-14	356,036
14-15	392,597

4 Related work

We highlight to different context where information visualization techniques have been succesfuly used in visual analytics processes. Wikipedia articles evolution has motivated interesting visualization developments. In [**?**], authors investigate the dynamics of Wikipedia through an exploratory data analysis tool which was effective in revealing patterns within the article text changes. [**?**] proposed a color scheme approach to present edit histories of Wikipedia administrators. Furthermore, many authors [10, **?**, 14, 17] have studied visualizations to easy control and understanding software source code evolution or in mapping collaborative efforts of various developers.

In this work, we are interested in existence and quantification of specific events of change in enzyme hierarchical annotations. As far as we are concerned there are not other works that propose a visualization for this type of data.

5 Technique

The main objectives of the proposed visualization were:

1. to give a panoramic macro view of the evolution of EC number annotations

2. to allow users to explore the complete set of changes, including entry meta data, formulating and answering general questions about EC number changes

Concerning the first objective, we wanted to present in a single perspective the EC changes segmented by all the possible combinations of events, considering the three parameters of the model (common prefix length, number of generalizations and specializations) across all the database releases.

5.1 Multivariate display

We have a multivariate problem where the fundamental task is to compare multiple instances of several variables at once and to allow users to identify similarities and differences among them. Small Multiples of Tufte [16] or Trellis Displays of Cleveland [2, 3] are a straightforward approach to present our data. They consist of splitting the data into multiple graphs that are presented close to each other in the screen allowing easier examination of the data in a given graph, and comparison of values and patterns among graphs to be relatively simple.

According to Few [7], individual graphs display a subset of a data set originally divided according to a categorical variable and the several graphs differ only in terms of the data being displayed. Every graph ideally shares the same type, shape, and size and, consequently the same categorical and quantitative scales. Scales in each graph must start and end with the same values (otherwise the accurate comparison is more difficult). Graphs can be arranged horizontally or vertically or as a matrix in a meaningful order.

5.1.1 Basic frame

With that in mind, we proceed our explanation of the proposed visual representation. The basic graph of the proposed Small Multiple representation, which we call from now on *frame*, is presented in Figure 1. It is a 2D plot where we present in the x-axis the number of specializations and in the y-axis, the number of generalizations. Both x and y-axes vary in the interval [0,4].

Notice some remarkable positions in the frame:

Position (0,0): entries with no changes in the corresponding pair of versions.

Diagonal: entries which suffered the same level of generalizations and specializations, potentially error corrections. They are presented in beige in Quadmap.

Lower right matrix: entries that suffered more levels of specializations than generalizations, in other words, knowledge about the catalyzed reaction has increased. They are presented in blue in Quadmap.

Upper left matrix: entries that suffered more levels of generalizations than specializations, in other words, knowledge about catalyzed reaction has decreased. They are presented in red in Quadmap.

Prohibited positions: if a change keeps a common prefix of size 3, it is impossible to have 2 degrees of generalization. Events like this are presented in a darker shade of gray.

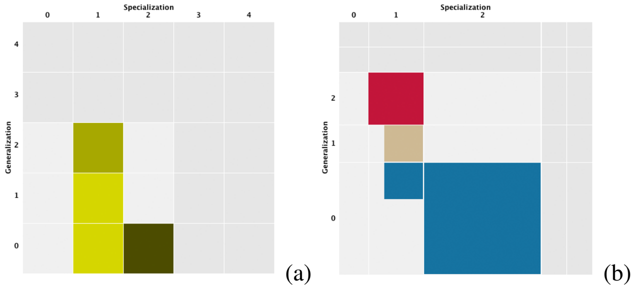


Figure 1: Basic frames for the proposed small multiple visualization. In (a), we present the Heatmap version and in (b), the Quadmap. In (a), the more dark the green, the higher the value represented. Likewise, in (b), the bigger the rectangle area, the higher the value. Red represents entries which presented more levels of generalizations than specializations, in other words, points above the diagonal. In blue, we present the points with more levels of specializations than generalizations, that means, points bellow the diagonal. Finally, in beige, we represent diagonal points which present the same levels of generalizations and specializations.

Several frames like this are then arranged in a small multiple fashion as in Figure 6. In x-axis, we represent the consecutive pairs of released versions. The y-axis presents the possible common prefixes in [0,4].

5.1.2 Heatmap

In a first version of the graph, we use a Heatmap representation where color is a pre-attentive attribute that encodes the frequency of that configuration of change.

The aim of this representation was to give an overview of the complete data evidencing trends and exceptions across the 15 releases. An interesting feature of this representation is that values in the lower right triangular matrix represents specializations and in the upper left triangular matrix, generalizations. Consequently, it is easy to recognize global trends towards generalization or specialization in enzyme reaction annotations.

5.1.3 Quadmap

Heatmaps present relevant trends in terms of generalization and specialization occurrences but we see two possible drawbacks in that approach.

Firstly, color is not a pre-attentive attribute able to precisely encode quantitative data. Most certainly, one can perceive that an intense color represents a higher value than a less intense one. However, it is very difficult to precisely estimate the values from color intensities.

The second drawback is that our heatmap presents too much blank space. According to Tufte [16], the data density of a graph is the proportion of the total size of the graph that is dedicated to displaying data. Tufte prefers high data density graphs as the human perceptual system is capable of detecting subtle patterns, trends and exceptions. On account of that, we decided to propose a second complementary view, hoping to reduce blank (non-data) space and also improve quantity estimation.

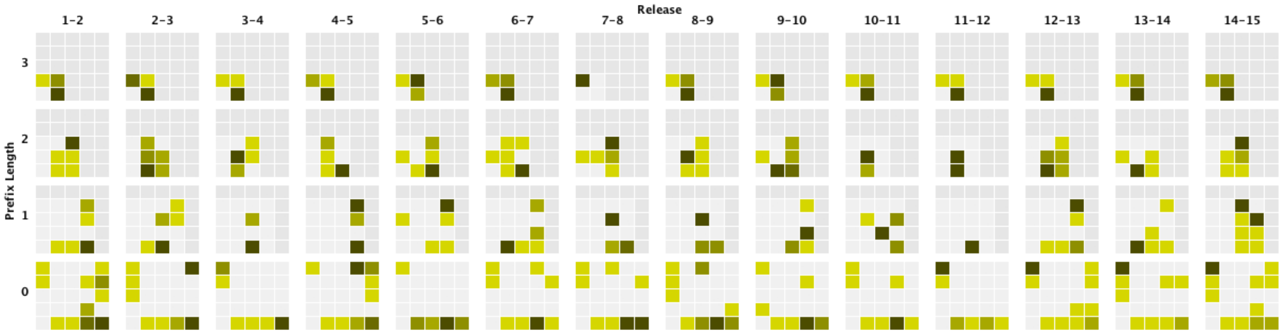


Figure 2: Multivariate view with Heatmaps.

The Quadmap representation was inspired in 2D scatter plots where the points are rectangles whose area represent frequency. Even though area is not the most precise visual attribute to encode quantity, it is more precise than colors. Notice, in Figure 1, that it is easier to estimate quantities in the Quadmap (b) than in the Heatmap (a).

It is important to highlight that, in Quadmaps axis are different from one frame to the other what goes against the rule of axis and scale preserving in Small Multiples. This happens because rectangle sizes distort tics in axes. Despite, we believe this option helps to emphasize trends and exceptions by using colored pixels to represent quantities more precisely than in Heatmaps.

5.2 Analytical interaction and navigation

5.2.1 Filtering, scales and normalization options

The efficaciousness of the information visualization techniques hinge on their ability to clearly and accurately represent information and on the capacity to fathom underlying information through interaction. Indeed, no matter how rich the display is, questions will arise, making interaction a necessary instrument in the pursuit of answers. Furthermore, contrasting different perspectives can lead to different insights. The proposed visualization provides predefined filters and different scaling and normalization options:

1. **Logarithmic scale on the frequencies:** rectangle area or heatmap colors are computed according to a logarithmic scale of frequencies

2. **Normalization of frequencies globally or by frame:** global normalization leads to a more realistic view of frequencies while local (or frame) normalization, despite contradicting Small Multiple rules, emphasizes a part-to-whole relationship into a give frame.

3. **Filter by only changes or presentation of the complete data set**: data is very unbalanced as we have much more stable entries than changes. In conclusion, when we exhibit the complete dataset, the changes are de-emphasized.

5.2.2 Hierarchical navigation

A particularly interesting way to create dense graphics is through what Tufte calls micro/macro readings [16]. These graphics convey one layer of information on a micro scale and another layer on a zoomed out, macro scale. One nice consequence of this technique is that information is consumed hierarchically. The viewer may glance from the distance to observe a global trend and, later, peer closely to examine individual pieces of that trend. Our multivariate view is a macro view of the whole set of changes in the dataset. Users can click on each frame and see it zoomed in a micro view. In other words, as we can see in Figure 3, users can select a specific release and common prefix length and view a detailed description of the respective frame.

Furthermore, users can click on the points in the micro view and see interactive histograms of each type of change. Through these histograms users can see the enzyme families which have suffered that change. These histograms are composed by small rectangles representing each change and by clicking on individual points users can see details about that specific entry.

6 Discussions

In this section, we describe the insights we obtained through the proposed interactive visualization.

6.1 Trends

6.1.1 Stable enzyme annotations

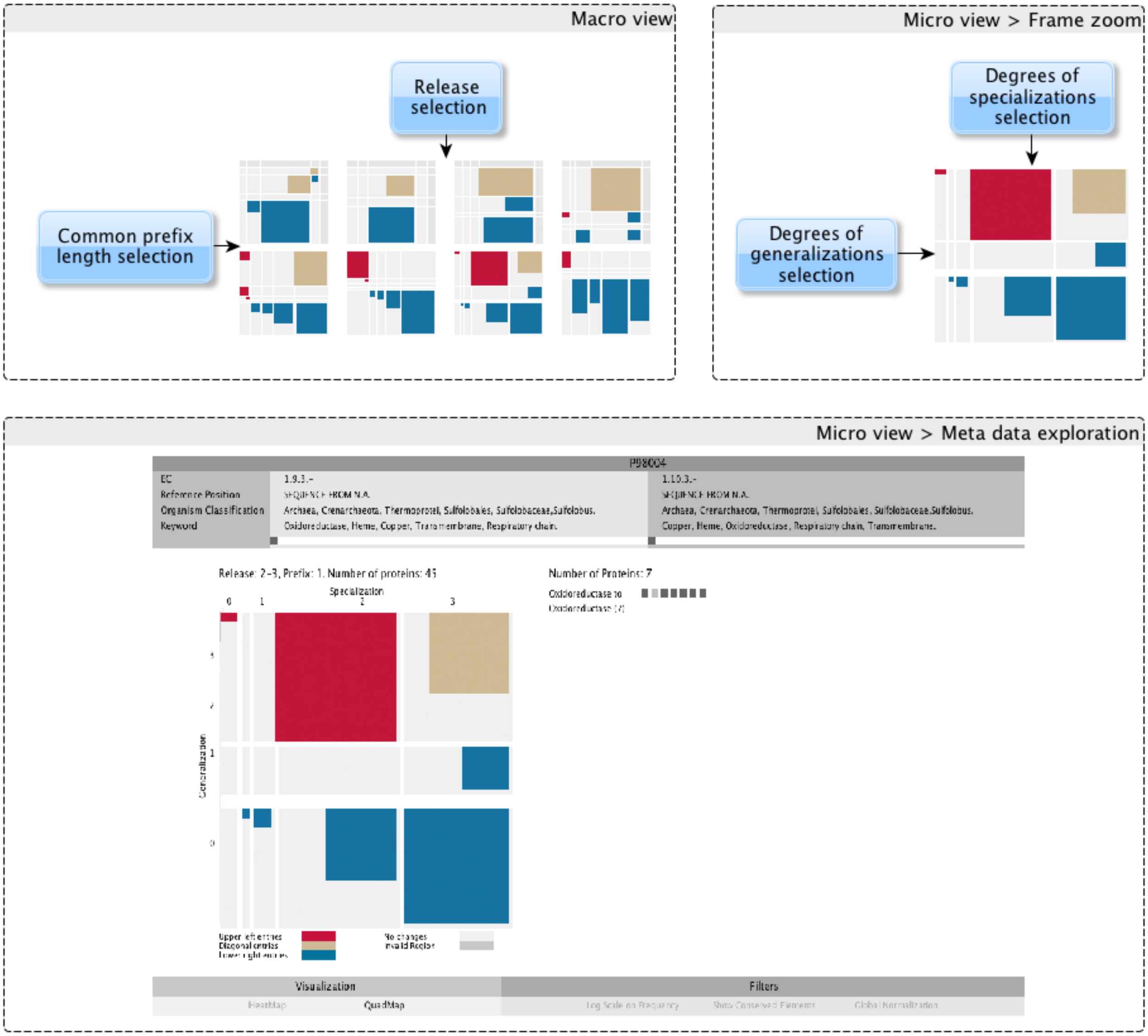
The most common event over the entire data set is located at the bottom left corner of each frame and it represents pairs of observed EC numbers that remained constant in a certain pair of versions. It means that the two EC numbers involved were equal (i.e. 3.1.3.2 to 3.1.3.2) or that there was no EC number (i.e. -.-.-.- to -.-.-.-).

In Figure 4, we present the more realistic view of the data set aggregating stable (entries in point (0, 0) at each frame) and the changes in other points in a global normalization and normal scale. We can see there is a global predominance of points (0, 0) in the last row. These points represent mostly entries with unknown EC numbers (-.-.-.-) and that remained unknown. Notice that the areas are clearly growing what reflects the grow in UniProt/SwissProt database over the fifteen analyzed releases.

In Figure 5, we show the same data normalized by frame, what reveals stable entries are predominant in almost every frame. Exceptions exist and they are going to be discussed in section 6.2.

6.1.2 Generalization versus Specialization

Consider, for each frame, a diagonal that extends from the bottom left corner to the top right corner. The matrix of points below this diagonal, called lower right triangular matrix, represents changes in which there are more specializations than generalizations. In a similar manner, the matrix of points above this diagonal, called upper left triangular matrix, represents changes in which there are more generalizations than specializations. In the Heatmap of Figure 6, for instance, the lower triangular matrices have more points than the superior ones, and therefore in the entire data set there are more specializations than generalizations. In Figure 6, where we present only changes in normal scale and local normalization, we can see a predominance of blue rectangles representing this trend. Once again, exceptions are evidenced and some of them are going to be discussed further in 6.2.



P98004		
EC	1.9.3.-	1.10.3.-
Reference Position	SEQUENCE FROM N.A.	SEQUENCE FROM N.A.
Organism Classification	Archaea, Crenarchaeota, Thermoprotei, Sulfolobales, Sulfolobaceae,Sulfolobus.	Archaea, Crenarchaeota, Thermoprotei, Sulfolobales, Sulfolobaceae,Sulfolobus.
Keyword	Oxidoreductase, Heme, Copper, Transmembrane, Respiratory chain.	Copper, Heme, Oxidoreductase, Respiratory chain, Transmembrane.

Figure 3: Navigation scheme.

Figure 6 also emphasizes that the line representing 0 generalizations in the last row of frames (common prefix length = 0) in multivariate matrix is a frequent type of change. It reveals an interesting trend of specialization for entries without annotation (-.-.-.-) as they tend to receive EC levels along all the releases.

6.2 Exceptions

6.2.1 Annotation deletion

The four points, in the red rectangles of the last line in Figure 7, whose parameters are common prefix length = 0, generalization = 4 and specialization = 0, in releases, 12-13, 13-14 and 14-15, represent a drastic change in which the four levels of involved EC numbers were deleted. The Table 4 shows the frequencies related to each point.

EC must be assigned to protein catalytic subunits. This implies that in large protein complexes only one or a few of the subunits will be annotated with an EC number. Indeed, proteins can have non catalytical functions like transport of substances, immunological or structural, for instance. In some cases, automatic annotation can assign EC numbers to a whole complex with non-catalytic subunits. These points in the graph represent corrections where the curators completely removed the EC numbers since the concerned subunits are not enzymes. We present three examples of UniProt/SwissProt entries that experienced four-level EC number deletion from version 12 to 13.

Table 4: Frequency of four-level EC number deletion from releases 11 to 15

Pair of releases	Frequencies
11-12	146
12-13	1,357
13-14	1,006
14-15	1,976

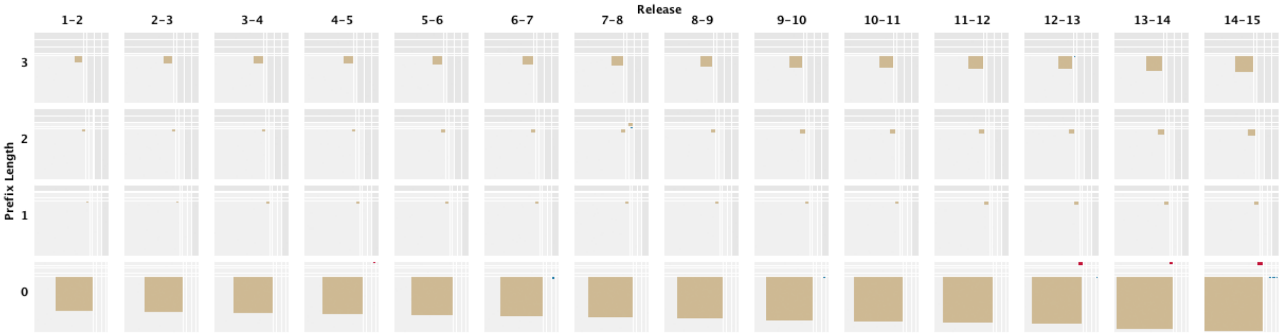


Figure 4: Normal scale, Stable entries + changes, Global normalization

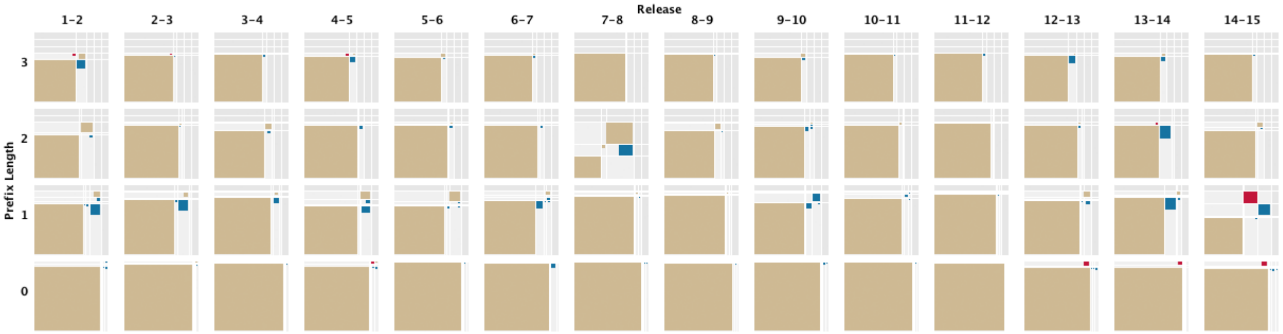


Figure 5: Normal scale, Stable entries + changes, Local normalization

- Identifier Q6FSJ2, which was annotated as 1.10.2.2 in version 12, is subunit 7 of Cytochrome b-c1, but not the subunit with Reductase Activity

- Identifier Q8LX28, whose annotation was 3.6.3.14 in version 12, is subunit 8 of ATP Synthase, which is part of the membrane proton channel

- Identifier Q6AY96, which was annotated as 2.7.11.1 in version 12, is a subunit of a transcription factor, but not the subunit with Serine/Threonine Kinase activity.

6.2.2 Deleted EC numbers

In Figure 7, the point with parameters common prefix length = 2, generalization = 2 and specialization = 2 in releases 7-8, a total of 1,900 EC number changes are represented. The three most numerous changes depicted in this point are, respectively, 2.7.1.37 to 2.7.11.1 (frequency 918), 2.7.1.112 to 2.7.10.1 (frequency 215) and 2.7.1.112 to 2.7.10.2 (frequency 165). As stated by IUBMB, the EC number 2.7.1.37 was deleted and divided in 2005 into 2.7.11.1, 2.7.11.8, 2.7.11.9, 2.7.11.10, 2.7.11.11, 2.7.11.12, 2.7.11.13, 2.7.11.21, 2.7.11.22, 2.7.11.24, 2.7.11.25, 2.7.11.30 and 2.7.12.1. The same happened to the EC number 2.7.1.112, that was deleted and divided into 2.7.10.1 and 2.7.10.2. In such cases, Transferase annotations, more specifically 2.7.*.* (transferring phosphorus-containing groups), underwent a revision caused by a change in the Enzyme Classification system, not by a change in enzyme function annotation.

Something similar happened in the point with parameters common prefix length = 1, generalization = 2 and specialization = 3 in releases 14-15 (frequency 212). This point can be better visualized in the Quadmap of figure 7 and it represents the EC move 2.5.1.- (transferring alkyl or aryl groups, other than methyl groups) to 2.2.1.9 (2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylic-acid synthase). The EC 2.5.1.64 was created in 2003 and deleted in 2008 when it was divided into 2.2.1.9 and 4.2.99.20. In this case, the annotation changes are due to the creation of a new EC (2.2.1.9), in other words, there was a change in the Enzyme Classification system.

6.2.3 Created EC numbers

In some cases, enzymes were integrated to UniProt/SwissProt when their catalytic activity was already known but there were no appropriate EC numbers defined by IUBMB to describe such catalytic activity . For instance, in Figure 7, the point with parameters common prefix length= 3, generalization = 0 and specialization = 1 in releases 12-13, represents a total of 637 EC number changes. One of the frequent EC moves represented by this point is 2.8.1.- (sulfurtransferases) to 2.8.1.8 (EC created in 2006 to represent lipoyl synthase), with frequency 117. The UniProt entry Q7UH37 experienced this change. It was integrated to UniProt in 10 MAY 2004 and its associated function was lipoyl synthase. However, there was not an EC number related to lipoyl synthase at that moment and this entry remained with the same incomplete EC 2.8.1.- until release 13 (26 FEB 2008), when it was annotated with EC 2.8.1.8.

6.2.4 Annotation errors

Another example of exception we detected is presented in Figure 7 by the red point with parameters common prefix length= 1, generalization = 3 and specialization = 2 in releases 14-15. This point represents a single kind of change which happened 261 times. The EC move was 2.1.1.61, which was created in 1982 and associated with tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase function, to 2.8.1.-, which is associated with sulfurtransferase function. The EC number 2.1.1.61 was not deleted,

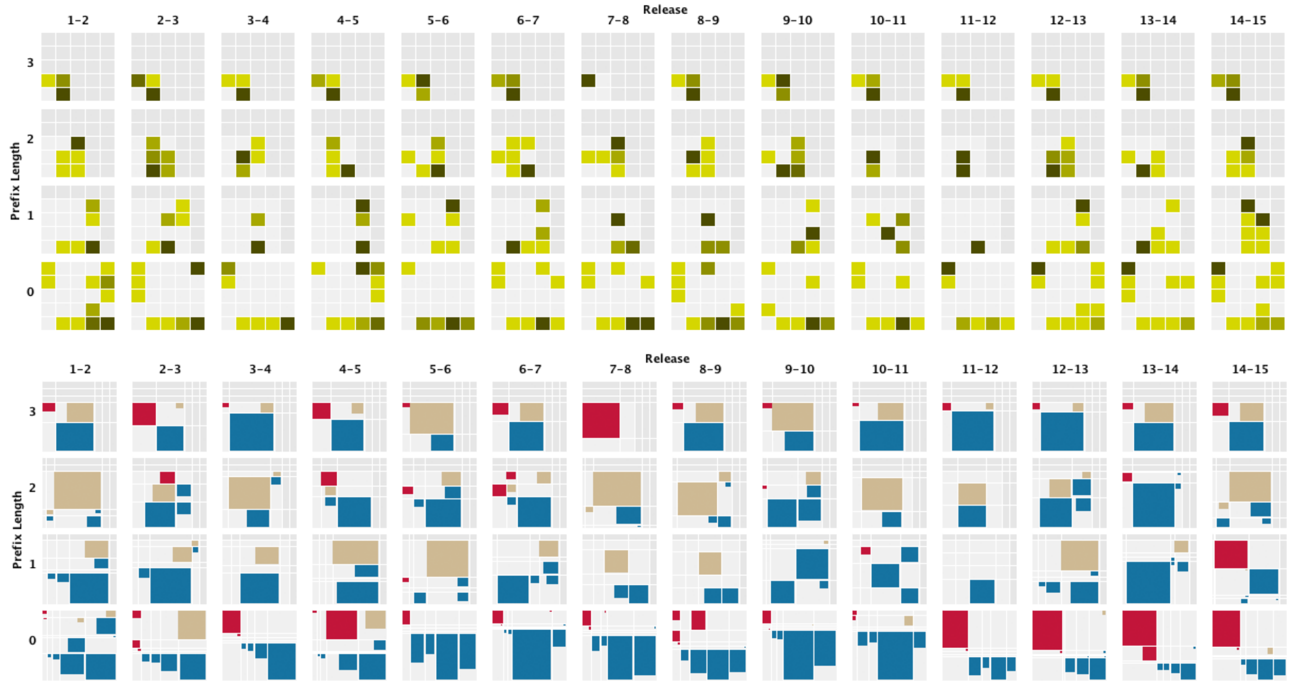


Figure 6: Normal scale, Only changes, Local normalization

so the move was a correction in order to annotate the involved entries with a more approprite catalytic function.

7 Conclusion

In this paper, we proposed a technique to visualize evolution in enzyme annotations, in special EC numbers, across several releases of UniProt/SwissProt data base. We modeled the changes of consecutive releases using parameters as the common prefix length and levels of generalization and specialization. The proposed interactive visualization gives a macro view of the changes and presents further details on demand as, for instance, frequencies of types of changes segmented by levels of generalizations and specializations as well as by the enzyme families. Users can further explore entries meta data. By visual means, we were able to evidence trends of specialization and growth of the database as well as detect several exceptions where EC numbers were deleted, divided, created or annotation errors were detected.

Acknowledgements

This work was supported by the Brazilian agencies Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científco e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Financiadora de Estudos e Projetos (FINEP) and Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais.

References

[1] R. Apweiler, M. Martin, C. O'Donovan, M. Magrane, Y. Alam-Faruque, R. Antunes, D. Barrell, B. Bely, M. Bingley, D. Binns, et al. The universal protein resource (uniprot) in 2010. *Nucleic Acids Res*, 38:D142–D148, 2010.

[2] R. Becker, W. Cleveland, M. Shyu, and S. Kaluzny. Trellis display: a framework for visualizing 2d and 3d data. Technical report, 1994.

[3] R. Becker, W. Cleveland, M. Shyu, and S. Kaluzny. Trellis display: User's guide. Technical report, 1994.

[4] S. Brenner et al. Errors in genome annotation. *Trends in Genetics*, 15(4):132–133, 1999.

[5] U. Consortium et al. Reorganizing the protein space at the universal protein resource (uniprot). *Nucleic Acids Res*, 40:D71–D75, 2012.

[6] D. Devos and A. Valencia. Intrinsic errors in genome annotation. *Trends in Genetics*, 17(8):429–431, 2001.

[7] S. Few. *Now you see it*. 2009.

[8] W. Gilks, B. Audit, D. de Angelis, S. Tsoka, and C. Ouzounis. Percolation of annotation errors through hierarchically structured protein sequence databases. *Mathematical biosciences*, 193(2):223–234, 2005.

[9] M. Green and P. Karp. Genome annotation errors in pathway databases due to semantic ambiguity in partial ec numbers. *Nucleic acids research*, 33(13):4035–4039, 2005.

[10] R. Holt. Gase: visualizing software evolution-in-the-large. In *Proceedings of the Third Working Conference on Reverse Engineering*, pages 163–167, 1996.

[11] S. Hung, J. Wasmuth, C. Sanford, and J. Parkinson. Detect - a density estimation tool for enzyme classification and its application to plasmodium falciparum. *Bioinformatics*, 26(14):1690–1698, 2010.

[12] C. Jones, A. Brown, and U. Baumann. Estimating the annotation error rate of curated go database sequence annotations. *BMC bioinformatics*, 8(1):170, 2007.

[13] A. Lesk and J. Wiley. *Database annotation in molecular biology*. Wiley Online Library, 2005.

[14] F. V. Rysselberghe. Studying software evolution information by visualizing the change history. In *Proceedings of the 20th IEEE International Conference on Software Maintenance*, pages 328–337, 2004.

[15] A. Schnoes, S. Brown, I. Dodevski, and P. Babbitt. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology*, 5(12):e1000605, 2009.

[16] E. Tufte. *Envisioning information*. 1990.

[17] L. Voinea, A. Telea, and J. van Wijk. Cvsscan: visualization of code evolution.

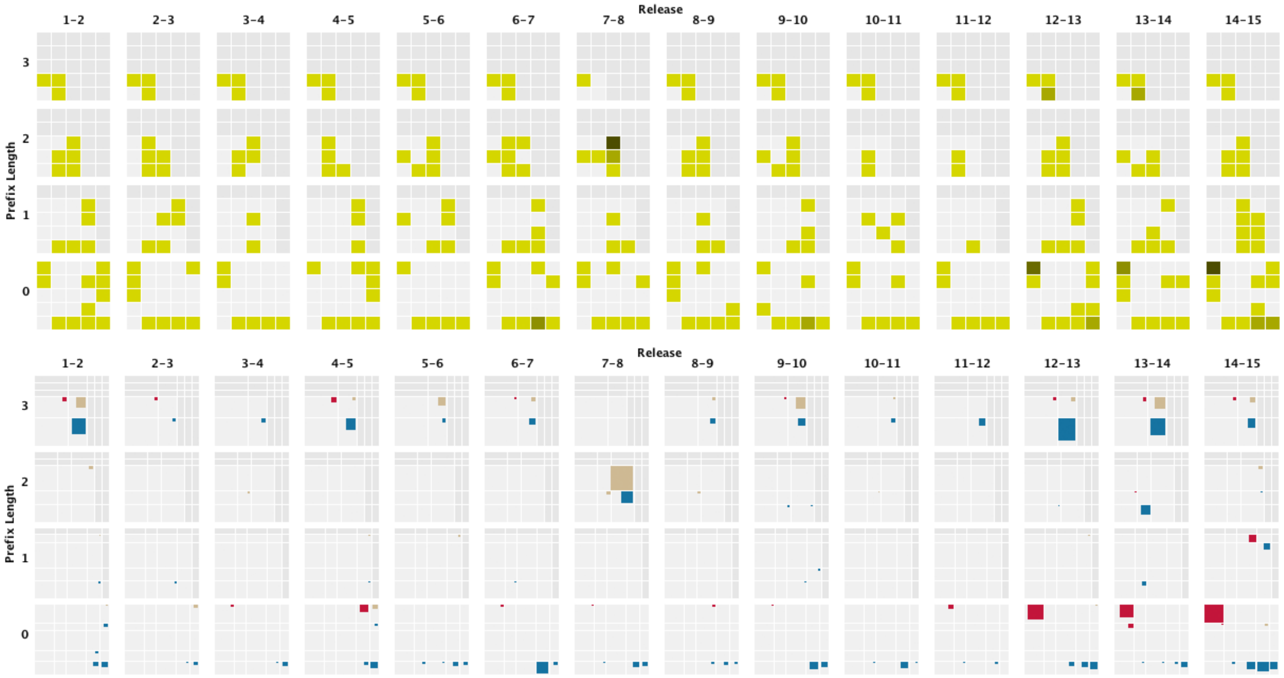


Figure 7: Normal scale, Only changes, Global normalization

Table 1: Example of EC numbers across consecutive database releases and our prefix / generalization / specialization model

Previous EC number	Actual EC number	UniProt id	releases	Common prefix length	Number of Generalizations	Number of Specializations
-.-.-.-	-.-.-.-	Q9K5T1	1 to 2	0	0	0
3.1.4.14	1.7.-.-	P41407	7 to 8	0	4	2
1.1.1.-	1.-.-.-	P52895	5 to 6	1	2	0
5.3.-.-	5.3.1.27	P42404	14 to 15	2	0	2
2.5.1.64	2.5.1.-	P17109	13 to 14	3	1	0
4.1.1.22	4.1.1.22	P95477	1 to 2	4	0	0

Table 2: Releases 1 to 15 of UniProt/SwissProt.

Release	Release date (MM/DD/YYYY)	% of entries with EC	Number of entries with EC	Total of entries
1	12/15/2003	0.37	52,434	141,681
2	07/05/2004	0.38	57,931	153,871
3	10/25/2004	0.38	61,229	163,235
4	02/01/2005	0.38	63,221	168,297
5	05/10/2005	0.38	69,164	181,571
6	09/13/2005	0.38	74,468	194,317
7	02/07/2006	0.39	80,874	207,132
8	05/30/2006	0.40	89,245	222,289
9	10/31/2006	0.40	97,508	241,242
10	03/06/2007	0.40	105,225	260,175
11	05/29/2007	0.40	108,876	269,293
12	07/24/2007	0.40	111,230	276,256
13	02/26/2008	0.43	151,694	356,194
14	07/22/2008	0.43	168,849	392,667
15	03/24/2009	0.44	189,234	428,650