ADVISE: VISUALIZING THE DYNAMICS OF ENZYME ANNOTATIONS IN UNIPROT/SWISS-PROT

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 propose an interactive visualization called ADVISe (Annotation Dynamics Visualization), which tackles the problem of visualizing evolutions in enzyme annotations across several releases of the UniProt/SwissProt database. More specifically, we visualize the dynamics of Enzyme Commission numbers (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 represents a progressively finer classification of the catalyzed reaction. The proposed interactive visualization gives a macro view of the changes and presents further details on demand, such as frequencies of change types segmented by levels of generalization and specialization as well as by enzyme families. Users can also explore entry metadata. With this tool, we were able to identify trends of specialization, database growth and exceptions in which EC numbers were deleted, divided or created and revisions of past annotation errors.

Availability: A video introducing ADVISe is available at http://vimeo.com/arturhoo/advise and the source code can be downloaded from https://github.com/arturhoo/ADVISe.

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

INTRODUCTION

In recent decades, there has been a significant increase in the biological data generated by experimental techniques such as the new generation of DNA sequencing technologies, protein sequencing and protein structure determination. Much of these data are organized and publicly available to the scientific community in biological databases via the Internet. According to [14], these repositories store not only raw biological data but also relevant information such as literature data, protein function and the relationship between a protein and its encoding gene, among other metadata.

Because biological databases are growing at very high rates, most of these metadata are automatically assigned. In many cases, the roles of most genes in various organisms have been reported by homology propagation, without performing any laboratory experiments [4]. To ensure the reliability of these annotations, studies of the reliability of the entries and measures of confidence should be developed. Many studies have drawn attention to error rates in biological database annotations [6, 9, 8, 12, 16, 11].

In fact, the automatic identification of these errors remains an open problem, and several challenges must be overcome. In the absence of laboratory experiments to verify automatically assigned annotations, it will remain impossible to establish a definite conclusion. Many studies have presented comparisons of a diversity of methods of functional annotation, demonstrating that they are widely incompatible and constraining their accuracy.

A major step toward automatic error detection is the description of how and to what extent biological database entry annotations evolve. In other words, we must fully understand why some entries appear to be more stable while others remain more volatile as well as the factors that determine these contrasting behaviors.

The research and development of models and algorithms, coupled with constantly improving visualization resources, represent a promising approach toward understanding how biological databases evolve. Interactive visualizations can be particularly powerful for depicting voluminous, high-dimensional and complex datasets from a macro/micro perspective and to help users unveil trends and exceptions in those datasets.

∗e-mail: sabrinas@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

IEEE Symposium on Biological Data Visualization 2012
October 14 - 15, Seattle, Washington, USA
978-1-4673-4730-3/12/$31.00 ©2012 IEEE
1.1 Enzyme annotations

By the late 1950s, during a period in which the number of known enzymes was increasing rapidly, it had become evident that the nomenclature of enzymology was becoming unmanageable. In many cases, the same enzymes became known by several different names, while conversely, the same name was occasionally given to different enzymes [21]. Many of the names conveyed little or no idea of the nature of the reactions catalyzed, and similar names were sometimes given to enzymes of quite different types. To address this situation, the General Assembly of the International Union of Biochemistry (IUB) 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 and the symbols used in the description of enzyme kinetics. The Commission prepared a report in 1961 that was promptly adopted and has since been widely used in scientific journals, textbooks, and so on. The size of the Enzyme Commission number (EC number) list has increased steadily since the publication of the first report, and many corrections have been made.

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

3: hydrolase, which means the enzyme breaks a chemical bond with a water molecule.
3.4: peptidase, which means the broken bond is a peptide bond, i.e., a bond between amino acid residues in a protein chain.
3.4.21: endopeptidase, which breaks an intra-chain peptide bond in which a serine residue participates in the mechanism of catalysis.
3.4.21.4: trypsin, which indicates an enzyme that cleaves mainly at the carboxyl side of the amino acid residues lysine or arginine.

When a new enzyme is annotated, one can add from one to four levels to the EC number, depending on the level of detail of the existing knowledge. In the best scenario, everything is known about the catalyzed reaction as well as the specific substrates and products involved. However, in many cases, when not all of the details about the catalytic activity are known, partial EC numbers, in which the unknown levels are indicated with hyphens, are used to annotate enzymes. The EC number “3.4.21.-”, for example, indicates that the specific enzyme substrates are not known, although information about the reaction catalyzed is available.

In this paper, we tackle the problem of analyzing enzyme annotation dynamics and propose a technique to visualize the evolution of these annotations across several releases of the UniProt/SwissProt database. 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 discuss previous related studies, and in section 5, we describe in detail the basis of the technique proposed as well as its capabilities. Finally, we discuss several insights that we obtained in section 6 and conclude the work and present perspectives in section 7.

2 Problem modeling

Based on the numerical and hierarchical natures of the Enzyme Commission number, we proposed a model to characterize the EC changes observed over several versions of UniProt/SwissProt. Our initial focus was on the visualization of the types of changes that occur and the frequency with which they occur. Furthermore, it is important to know the hierarchical level in which a change occurs because an alteration at a higher level (leftmost) is more severe than at a lower level. Thus, we decided to segment changes by their common prefix length together with the number of generalizations and specializations associated with a specific EC number.

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

3.1.3.2 \rightarrow 3.1.3.5

This change occurred in 77 hydrolases of release 5 to 6. The common prefix length is 3 (the first three levels from left to right remained the same), there was 1 generalization (number 2 was deleted) and 1 specialization (number 5 was inserted). 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 Dataset

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 as well as the incorporation, interpretation, integration and standardization of data from a large number of disparate sources. The UniProt Knowledgebase (UniProtKB) is the most comprehensive catalog of protein sequence and functional annotation. As stated by [5], the UniProtKB is an expertly curated database and 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 the literature and curator-evaluated computational analysis. Annotation is performed by biologists with specific expertise to achieve accuracy. TrEMBL contains computationally analyzed records enriched with automatic annotation and classification. Because 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 repositories of the UniProt database at the beginning of this study (March 2009) were downloaded. We analyzed releases 1 (when SwissProt was integrated to UniProt) through 15 (the current release when this study was initiated).

To determine if an EC number change occurred, we examined 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 an EC number, and their percentage in the 15 releases are provided in Table 2. Table 3 shows the number of entries in the set intersection of each release pair.

4 Related work

We will review different contexts where information visualization techniques have been successfully used in visual analytic processes. In [18], the authors investigated the dynamics of Wikipedia articles through an exploratory data analysis tool that was effective in revealing patterns within a given set of changes in article texts. In [20], a color scheme approach was proposed to present edit histories of Wikipedia administrators. Furthermore, many authors [10, 13, 15, 19] have studied visualizations to facilitate control and understand software source code evolution or to map collaborative efforts of various developers.
Table 1: Example of EC numbers across consecutive database releases and our prefix/generalization/specialization model.

<table>
<thead>
<tr>
<th>Previous EC number</th>
<th>Actual EC number</th>
<th>UniProt id</th>
<th>Releases</th>
<th>Common prefix length</th>
<th>Degrees of generalizations</th>
<th>Degrees of specializations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q9K5T1</td>
<td>1.7.-</td>
<td>P41407</td>
<td>1 to 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.1.1.-</td>
<td>5.3.1.27</td>
<td>P42404</td>
<td>14 to 15</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>2.5.1.64</td>
<td>4.1.1.22</td>
<td>P95477</td>
<td>1 to 2</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Release</th>
<th>Release date (MM/DD/YYYY)</th>
<th>% of entries with EC</th>
<th>Number of entries with EC</th>
<th>Total number of entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12/15/2003</td>
<td>37</td>
<td>52,434</td>
<td>141,681</td>
</tr>
<tr>
<td>2</td>
<td>07/05/2004</td>
<td>38</td>
<td>57,931</td>
<td>153,871</td>
</tr>
<tr>
<td>3</td>
<td>10/25/2004</td>
<td>38</td>
<td>61,229</td>
<td>163,235</td>
</tr>
<tr>
<td>4</td>
<td>02/01/2005</td>
<td>38</td>
<td>63,221</td>
<td>168,297</td>
</tr>
<tr>
<td>5</td>
<td>05/10/2005</td>
<td>38</td>
<td>69,164</td>
<td>181,571</td>
</tr>
<tr>
<td>6</td>
<td>09/13/2005</td>
<td>38</td>
<td>74,468</td>
<td>194,317</td>
</tr>
<tr>
<td>7</td>
<td>02/07/2006</td>
<td>39</td>
<td>80,874</td>
<td>207,132</td>
</tr>
<tr>
<td>8</td>
<td>05/30/2006</td>
<td>40</td>
<td>89,245</td>
<td>222,289</td>
</tr>
<tr>
<td>9</td>
<td>10/31/2006</td>
<td>40</td>
<td>97,508</td>
<td>241,242</td>
</tr>
<tr>
<td>10</td>
<td>03/06/2007</td>
<td>40</td>
<td>105,225</td>
<td>260,175</td>
</tr>
<tr>
<td>11</td>
<td>05/29/2007</td>
<td>40</td>
<td>108,876</td>
<td>269,293</td>
</tr>
<tr>
<td>12</td>
<td>07/24/2007</td>
<td>40</td>
<td>111,230</td>
<td>276,256</td>
</tr>
<tr>
<td>13</td>
<td>02/26/2008</td>
<td>43</td>
<td>151,694</td>
<td>356,193</td>
</tr>
<tr>
<td>14</td>
<td>07/22/2008</td>
<td>43</td>
<td>168,849</td>
<td>392,667</td>
</tr>
<tr>
<td>15</td>
<td>03/24/2009</td>
<td>44</td>
<td>189,234</td>
<td>428,650</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Release pair</th>
<th>Number of entries in ∩</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>141,249</td>
</tr>
<tr>
<td>2-3</td>
<td>151,318</td>
</tr>
<tr>
<td>3-4</td>
<td>162,812</td>
</tr>
<tr>
<td>4-5</td>
<td>166,933</td>
</tr>
<tr>
<td>5-6</td>
<td>181,005</td>
</tr>
<tr>
<td>6-7</td>
<td>193,382</td>
</tr>
<tr>
<td>7-8</td>
<td>207,069</td>
</tr>
<tr>
<td>8-9</td>
<td>222,181</td>
</tr>
<tr>
<td>9-10</td>
<td>241,189</td>
</tr>
<tr>
<td>10-11</td>
<td>260,065</td>
</tr>
<tr>
<td>11-12</td>
<td>269,152</td>
</tr>
<tr>
<td>12-13</td>
<td>276,011</td>
</tr>
<tr>
<td>13-14</td>
<td>356,036</td>
</tr>
<tr>
<td>14-15</td>
<td>392,597</td>
</tr>
</tbody>
</table>

In this work, we are interested in the existence and quantification of specific events of change in enzyme hierarchical annotations. To the best of our knowledge, there are no other works that propose a visualization of this type of data.

5 ADVISY

The main objectives of the proposed visualization were the following:

1. to provide a panoramic macro view of the evolution of EC number annotations;
2. to permit users to explore the complete set of changes, including entry metadata, and the formulation and resolution of general questions about EC number changes.

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

5.1 Multivariate display

We have a multivariate problem in which the fundamental task is to simultaneously compare multiple instances of several variables and to permit users to identify similarities and differences among them. Small Multiples of Tufte [17] or Trellis Displays of Cleveland [2, 3] are a straightforward approach to present our data. These approaches consist of splitting the data into multiple graphs that are presented close to each other in the screen, permitting easier examination of the data in a given graph and relatively simple comparison of values and patterns among graphs.

According to Few [7], individual graphs within multiple graphs display a subset of a dataset originally divided according to a categorical variable, and the several graphs differ only in terms of the data displayed. Every graph ideally shares the same type, shape and size and, consequently, the same categorical and quantitative scales. The 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 the above in mind, we proceed with our explanation of the proposed visual representation. The basic graph of the proposed Small Multiple representation, which we will refer to as frame, is presented in Figure 1. It is a two-dimensional plot in which we present the number of specializations in the x-axis and the number of generalizations in the y-axis. Both x and y-axes vary in the interval [0,4].
Note some remarkable positions in the frame:

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

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

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

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

**Invalid positions:** if a change retains a common prefix of size 3, it is impossible to have 2 degrees of generalization. These types of events are presented in a dark shade of gray.

First, color is not a pre-attentive attribute that is able to precisely encode quantitative data. One can perceive that an intense color represents a higher value than a less intense color. 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 [17], 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 because the human perceptual system is capable of detecting subtle patterns, trends and exceptions. Therefore, we decided to propose a second, complementary view, with the aim of reducing blank (non-data) space as well as improving quantity estimation.

The Quadmap representation was inspired in two-dimensional scatter plots where the points, which we will refer to as positions, are rectangles in which area represents frequency. Although area is not the most precise visual attribute to encode quantity, it is more precise than color. Note, 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 the axes in Quadmaps are different from one frame to the other, going against the rule of preservation of axis and scale in Small Multiples. This occurs because rectangle sizes distort ticks in axes so, to identify the diagonal, lower right and upper left matrix we coloured these elements in beige, blue and red respectively. Nevertheless, 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 or linear scale on the frequencies:** rectangle areas in Quadmap or Heatmap colors are computed according to a logarithmic scale or absolute value 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 given frame.

3. **Filter by only changes or presentation of the complete data set:** only entries that suffered changes are showed or the whole dataset (including stable entries). The data are very unbalanced because we have many more stable entries than changes. In conclusion, when we visualize 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 refers to as micro/macro readings [17]. These graphics convey one layer of information on a micro scale and another layer on a zoomed-out, macro scale. A favorable consequence of this technique is that information is consumed hierarchically. The viewer may scan from a distance to observe a global trend and, later, scrutinize closely to examine individual components of that trend. Our multivariate view is a macro view of the entire set of changes in the dataset. Users can click on each frame and see it zoomed in...
Figure 2: (a) Multivariate view of Heatmap and Quadmap with linear scale, only changes presented and local normalization. (b) Multivariate view of Heatmap and Quadmap with linear scale, only changes presented and global normalization. In (a), local normalization highlights changes that are numerous inside the frames and in (b), global normalization highlights changes that are numerous across all of the releases. Basic frames and axis for Heatmap and Quadmap were detailed in Figure 1.
a micro view. In other words, as observed in Figure 3, users can select a specific release and common prefix length and view a detailed description of the respective frame.

Additionally, users can click on the positions in the micro view and see interactive histograms of each type of change. Through these histograms, users can identify the enzyme families that are subject to that change. These histograms are composed by small rectangles representing each change, and by clicking on individual rectangles, users can view details about that specific entry.

5.3 Implementation

ADVISe was implemented in Processing\(^1\), release 1.5.1. The dataset accessed by our visualization was downloaded from UniProt and filtered using Java Development Kit 6\(^2\) to get the data we were interested in: EC number annotation and line types Reference Position (RP), Organism Classification (OC) e Keyword (KW) from UniProt text files. These data were processed by some Python\(^3\) scripts (version 2.6.5) and stored in a MySQL\(^4\) database version 5.1.61.

6 DISCUSSION

In this section, we describe the insights we obtained from ADVISe.

6.1 Trends

6.1.1 Stable enzyme annotations

The most common event spread over the entire dataset is located at the bottom left corner of each frame, position (0,0), and represents pairs of observed EC numbers that remained unchanged in a pair of releases. In this case the two EC numbers involved were equal (i.e., 3.1.3.2 to 3.1.3.2) or there was no EC number (i.e., -.-.-.- to -.-.-.-).

In Figure 4 (a), we present a more realistic view of the dataset, aggregating stable entries (position (0, 0) at each frame) and changes in other positions with global normalization and a linear scale. We can observe a global predominance of entries with no generalization or specialization and prefix length 0. These entries usually have undefined EC numbers (-.-.-.-) that have remained so. Note that the area of this specific position is clearly growing across releases, reflecting the growth in the UniProt/SwissProt database over the fifteen analyzed releases.

In Figure 4 (b), we show the same data normalized by frame, revealing that stable entries are predominant in almost every frame. Exceptions do exist and will be discussed in section 6.2.

6.1.2 Generalization versus Specialization

In the Heatmap of Figure 2 (a), we can observe that the lower triangular matrices have more entries than the upper triangular matrices and thus, in the entire dataset, there were more generalizations than generalizations. In the Quadmap of Figure 2 (a), in which we present only changes in linear scale and local normalization, we can observe a predominance of blue rectangles representing this trend. Once again, exceptions are apparent, and some will be discussed further in section 6.2.

Figure 2 (a) also emphasizes that the line representing no generalizations in the bottom row of frames (common prefix length 0) in the multivariate matrix is a frequent type of change. It reveals an interesting trend of specialization for entries without annotation (-.-.-.-) because they tend to receive EC levels in each release.

6.2 Exceptions

6.2.1 Annotation deletion

The four positions indicated by red rectangles on the bottom row of Figure 2 (b), in which the parameters are common prefix length 0, 4 degrees of generalization and no specialization in releases 12-13, 13-14 and 14-15, represent a drastic change in which the four levels of EC numbers were deleted. Table 4 shows the frequencies associated with each position.

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

<table>
<thead>
<tr>
<th>Pair of releases</th>
<th>Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>11-12</td>
<td>146</td>
</tr>
<tr>
<td>12-13</td>
<td>1,357</td>
</tr>
<tr>
<td>13-14</td>
<td>1,006</td>
</tr>
<tr>
<td>14-15</td>
<td>1,976</td>
</tr>
</tbody>
</table>

EC numbers 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-catalytic functions such as transport of substances or an immunological or structural role. In some cases, automatic annotation can assign EC numbers to a whole complex, including non-catalytic subunits. Positions that symbolize such cases in ADVISe represent corrections in which the curators completely removed the EC numbers because the related subunits are not enzymes. We present three examples of UniProt/SwissProt entries that experienced four-level EC number deletion from releases 12 to 13:

- Identifier Q6FSJ2, which was annotated as 1.10.2.2 in version 12, is subunit 7 of cytochrome b-c1 but is not the subunit with reductase activity.
- Identifier Q8LX28, which was annotated as 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 is not the subunit with serine/threonine kinase activity.

6.2.2 Deleted EC numbers

In Figure 2 (b), a total of 1,900 EC number changes are represented by the position with common prefix length 2, 2 degrees of generalization and 2 degrees of specialization in releases 7 to 8. The three most numerous changes depicted in this position are, respectively, 2.7.1.37 to 2.7.11.1 (918 entries), 2.7.1.112 to 2.7.10.1 (215 entries) and 2.7.1.112 to 2.7.10.2 (165 entries). As stated by IUBMB, 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. Similarly, EC number 2.7.1.112 was deleted and divided into 2.7.10.1 and 2.7.10.2. In such cases, transferase annotations, and more specifically, EC numbers beginning with 2.7 (transferring phosphorus-containing groups), underwent a revision caused by a change in the EC number classification system and not a change in enzyme function annotation.

A similar phenomenon occurred at the position with a common prefix length 1, 2 degrees of generalization and 3 degrees of specialization in releases 14 to 15 (212 entries). This position can be better visualized in the Quadmap of Figure 2 (b) and represents the EC number change 2.5.1.- (transferring alkyl or aryl groups other than

\(^1\)http://processing.org/
\(^2\)http://www.oracle.com/technetwork/java/index.
\(^3\)http://www.python.org/
\(^4\)http://www.mysql.com/
methyl groups) to 2.2.1.9 (2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylic-acid synthase). The EC number 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 EC number classification system.

6.2.3 Created EC numbers
In some cases, enzymes were integrated into the UniProt/SwissProt database when their catalytic activity was already known but there were no appropriate EC numbers defined by IUBMB to describe this specific catalytic activity. For example, in Figure 2 (b), the position with common prefix length 3, no generalizations and 1 degree of specialization in releases 12 to 13 represents a total of 637 EC number changes. A representative EC number change depicted by this position is 2.8.1.- (sulfurtransferases) to 2.8.1.8 (EC created in 2006 to represent lipoyl synthase), with 117 entries. The UniProt entry Q7UH37 exhibited this change. It was integrated to UniProt on 10 May 2004, and its associated function was lipoyl synthase. However, there was not an EC number related to lipoyl synthase at that time, and this entry remained with the same incomplete EC number, 2.8.1.-, until release 13 (26 Feb 2008), when it was annotated with EC number 2.8.1.8.

6.2.4 Annotation errors
Another exception we detected is presented in Figure 2 (b) by the red position with common prefix length 1, 3 degrees of generalization and 2 degrees of specialization in releases 14 to 15. This position represents a single type of change that occurred 261 times. The EC number change was 2.1.1.61, which was created in 1982 and is associated with tRNA (5-methylaminomethyl-2-thiouridylate-methyltransferase) activity, to 2.8.1.-, which is associated with sulfurtransferase activity. The EC number 2.1.1.61 was not deleted, and thus, the EC number change was a correction to annotate the associated entries with a more appropriate catalytic function.

7 Conclusions and future work
In this paper, we proposed ADVISSe, an interactive tool to visualize the dynamics of enzyme annotation evolution, and specifically, EC numbers, across several releases of the UniProt/SwissProt database. We modeled the changes of consecutive releases with the parameters of 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 such as frequencies of types of changes segmented by levels of generalization and specializations as well as by enzyme family. Users can further explore entry metadata. By visual inspection, we were able to identify trends of specialization and database growth as well as detect several exceptions in which EC numbers were deleted, divided or created or annotation errors were corrected.

In future work, we intend to implement a consensus view to summarize each line and generate a frame that is representative of the trends related to each common prefix length. As a consequence, we believe we will be able to spot relevant exceptions relative to the pattern. We will highlight these exceptions automatically to simplify the visual analytical process. Furthermore, we want to investigate methods to allow users to annotate insights from specific positions of the frames so that we can collect relevant data from expert users for further studies. Last, but not least, we are planning...
to systematically measure user insights and impressions about the proposed visualization.

ACKNOWLEDGEMENTS

This work was supported by the Brazilian agencies Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico 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


