

**GENEFUN**  
**LSHG-CT-2004-503567**  
**D2.1**

*Reliability score for function transfer based on pairwise similarities*

**Angela del Pozo and Alfonso Valencia**  
**CNB –CSIC**

**[valencia@cnb.uam.es](mailto:valencia@cnb.uam.es)**

**[apozo@cnb.uam.es](mailto:apozo@cnb.uam.es)**

## **Evaluation of the potential error in the transference of function between related proteins.**

As part of our initial work in this project we decided to start by revisiting the previously published work by the groups of J. Thornton [3], B. Rost [4] and our work [1] on the relation between sequence similarity and functional annotations.

These groups selected three very different non-redundant sets of protein pairs and assessed the similarity of their database functional annotation in particular the three studies focused in the enzymatic function annotations.

For the development of the function transfer reliability index proposed in the objectives of the workpackage it is essential to obtain a reliable set of proteins to perform the calibration of the sequence to annotation relation.

In the following we describe the process implemented for the creation of that dataset, the comparison with previous results, and the final proposal of the sequence-to-annotation relation.

### **Background**

A number of studies have addressed the systematic comparison of the annotations deposited in databases for pairs of proteins at different levels of similarity.

Initially Devos and Valencia [1] analyzed the annotations of a large collection of protein pairs at different levels of sequence similarity, based on the corresponding structural alignments, conclude that general functional characteristics, i.e. the first EC digits and general definition of function, were conserved even for divergent proteins (less than 15% of similarity), while the detailed functional characteristics, i.e. the fourth E.C. digit which describes substrate specificity, were less conserved below 50% of identity.

Once this information is extrapolated to the level of genomes [2], assuming that they were annotated by transferring database information between similar proteins detected with simple pair-wise based methods, the expected errors could be as much as 40% of for the detailed characteristics (fourth EC digit) and as little as 5% for the general ones (i.e. general function, first EC digit). Interestingly this type of simple process of annotation transference is what has been used for the annotation of most genomes.

Todd et al. [3] performed a similar study focused more on the relation of function and the structural context. They used PDB with the same homologous superfamily extracted from the CATH database. In this case their assessment revealed an accuracy of 95% in the conservation of the EC fourth digit even in relatively low sequence identity of 30%.

After these two first systematic studies Rost [4] analyzed a new data set build to avoid possible redundancies in the selection of the protein pairs and the classes of functions analyzed. In this case the relation was weaker, and proteins with identities as high as 70% have different annotations at the level of the fourth EC digit.

### **Derivation of the non-redundant set of pair-wise alignments.**

We have focused our attention in a new dataset extracted from the Combinatorial Extension (CE) database as a source of well-organized pairwise alignments, defined by the structural similarity of the corresponding proteins.

From the CE database we selected very carefully a set of alignments to reduce the possible problem of data redundancy. We have considered pairs with good structural alignment (Z-score above 3.8). From the database we selected pairs complaining with the three criteria provided by CE database. These criteria are:

- The RMSD (Root Mean Square Deviation) between two chains is less than 2Å.
- The length difference between two chains is less than 10%
- The number of gap positions in alignment between two chains is less than 20% of aligned residue positions.
- At least 2/3 of residue positions in the represented chain are aligned.

The CE database is arranged in set of independent alignments based on a selected representative 3D structure (PDB chains in general corresponding to the first proteins solved). The corresponding alignments contain other proteins whose structure has been structurally aligned with the representative ones. The entire database is composed of 436508 alignments.

The way the information is organized in the set determines the way we check the pairs of proteins with the criteria mentioned above:

First, we applied the rules above to pairs formed among the representatives to detect related proteins. For related representatives we inspect the proteins that are aligned with them. If the two are aligned with the same proteins, i.e. given a representative protein A that is very closed to another representative protein B, and both are aligned with protein C, then we choose the best alignment (for example, protein A with protein C), excluding the other one from our dataset. The other cases are represented by proteins aligned with a representative sequence, that are also similar to a protein aligned with a different representative sequence, and the two representatives are related. That is the case if protein D is aligned with A, and E is aligned with B, and D and E are related. As in the first case the best of those alignments are select and the other pair discarded.

The families of sequences aligned with the representative sequences were also

inspected to eliminate redundant pairs. The Non-redundant set was constructed by searching for proteins that are aligned with the same representative and are very similar between them.

### **Mapping sequences to their functions.**

Once we build the non-redundant set of protein pairs, the following task was to map the protein structures from the CE database to the corresponding sequences in the Swissprot database, which provide the corresponding functional annotations. A number of conditions were imposed to guaranty the transference to the right domains.

To make the conversion from PDB chain to Swiss Prot accession number (AC), based on a correspondence table between PDB entries and Swiss-Prot databases build with Blast searches and pdbtosp.txt file (release 45.5).

This table was additionally used to guaranty that proteins in each pair do not correspond to the same structure or swissprot entry. At this point it was necessary to eliminate many duplicate pairs either because the same proteins was aligned several times, or because equivalent pairs appear after the conversion of PDBs in swissprot identifiers. During this process selected the best alignment in the cases in which more than one protein pair was present. After this process the database of alignments was composed of 240782 alignments.

Finally we have imposed some restrictions to guarantee the reliability and the quality of the alignments. We have rejected those alignments whose length covers less than 70% percent of the number of the residues in each structure, and has less than 50 amino acids included in the alignment. We have also excluded alignments with a sequence identity above to 95% to avoid possible self-comparisons.

### **Analysis of the EC numbers**

The first functional property studied has been the enzymatic function as described by the Enzyme Commission numbers (EC). That is the characteristic addressed by previous studies.

As a result of the selection process the non-redundant dataset contains 3630 structural alignments. 1228 (34.3%) of these proteins are enzymes according to the ENZYME database.

The number of proteins involved in the filtered alignments is 2758, and the total number of proteins with an assigned EC number is 1236 (44.81%), which is very similar to the rate enzyme/non-enzyme of 1/4 mentioned in previous publications [5]. Figure 3 shows the result of comparing proteins with and without EC codes assigned at various identity classes.

We define a parameter (ECC) for comparing EC numbers. ECC is defined as the agreement in the enzymatic function between the proteins aligned, and is defined as the number of E.C. code levels that two proteins share. So if a protein A has assigned the EC number 3.2.1.1, and other protein B has the EC number 3.2.1.17, their ECC will be of 3. The ECC ranks between 0 and 4.

### **Further evaluation of the composition of the dataset.**

We have further analyzed the dataset to reduce other possible bias. First we checked if some proteins were present in an excessive number of cases. The mean usage of proteins is approximately 2.6 times, what is a number small enough and a good indicator of data balance. The histogram in figure 4A, shows that around 90% of the proteins are used less than 7 times and the ones with larger frequencies (over 10 times) correspond to representatives. These proteins are used in total 44 times, representing only the 1.21% of the pairs.

We have made the same analysis with regard to the EC number (Fig. 5B). The situation here is more complicated since the EC codes are populated with protein sequences in a very inhomogeneous way, with 457 of the total 2021 codes representing only the 22.61% of the complete set of EC numbers. The ratio between numbers of EC per protein is 0.37 and the mean usage of the EC number is about 5.37 times, counting the many cases in which the two proteins of a pair have the same EC number. To study if any of the EC numbers are over-represented we calculated the mean and the standard deviation of the differences between the experimental distribution showed in figure 4B and the obtained across the ENZYME database. The values are -0.001 for the mean and 0.003 for the deviation, from where we concluded that the data is not biased in the sense there are no EC number over represented and basically their distribution is the expected from the composition of the ENZYME database.

Regarding the contribution of each EC code to the different values of ECC, 80% of the times that a EC is used they contribute to pairs that share three or four EC levels, and only 8% of the EC codes contribute to pairs that have conserved one single level.

This distribution makes sense with a general trend in which related proteins tend to share the general enzymatic function.

### **Comparison of EC numbers for pairs of aligned proteins.**

The relation between EC number conservation (ECC) and the level of sequence identity of the corresponding protein pairs is represented in Figure 1. The EC numbers are conserved (more than 70% conservation) at sequence identity levels as high as 45%..

Below 25% of identity the fluctuations become important, therefore a correct assignment of the EC code based on a pairwise alignment is not guaranteed.

### **Comparison with previous studies.**

We have compared the current results with previous studies about the conservation of the enzymatic function carried out by Devos et al [1], Todd et al [3] and Rost [4]. The data have been extrapolated directly from the published figures therefore some deviations might be introduced.

The Figures 5A and 5B represent the percentages of pairs that conserve the four EC digits. Note that the three curves shows the same growing trend but they achieve optimal values (more than 80% of pairs with full enzymatic numbers conservation) at different levels of pairwise sequence identity in the rank of 40-70 percent.

Todd's curve shows always higher conservation of the sequence to function relation reaching more than 90% of conservation at lower levels of similarity (30%). The difference of own results with those of Todd could be related with a higher number of proteins without known enzymatic function and for a higher sequence redundancy of their dataset.

Figure 5B shows the results of Devos et al and Rost works. Apparently the difference with the work of Rost is related with the introduction in his dataset of a considerable number of pairs of alignments of small size (Devos, personal communication), and therefore an imperfect assignment of EC numbers to protein fragments, which makes their results difficult to interpret in a context of annotation of biological function. On the other hand it is worth stressing the similarity between the Devos's curves and the present work, specifically from 70% of sequence identity. This result stands out if we consider the datasets used, that are very different (FSSP in the case of Devos's work) and the process of selection of the data.

Note that Figure 5A includes non-enzymes as a class and Figure 5B does not what explains their apparent differences.

Figure 6A and 6B reinforce the ideas mentioned before representing accumulated values. We have calculated the percentage of pairs that share at least three enzymatic numbers to overcome possible fluctuations of the data.

### **Summary**

We have collected an unbiased data set composed by 3630 protein pairs, for which we have carefully assigned EC numbers. We have carefully assessed the dataset to reduce any possible bias, in proteins or EC codes.

The analysis of these new dataset has turned out to be very similar to our previous study that used a completely data set, and both of them are different to the other previous studies.

The results obtained in Figure 1, represents the bases for the calibration of the sequence-to-annotation relation, and the derivation of the function transfer reliability index.

### **Future work.**

Our next step will be to complete this analysis with other characteristics of protein function, including GO codes, swissprot keywords, binding sites, cellular localization and if possible post-translational modification.

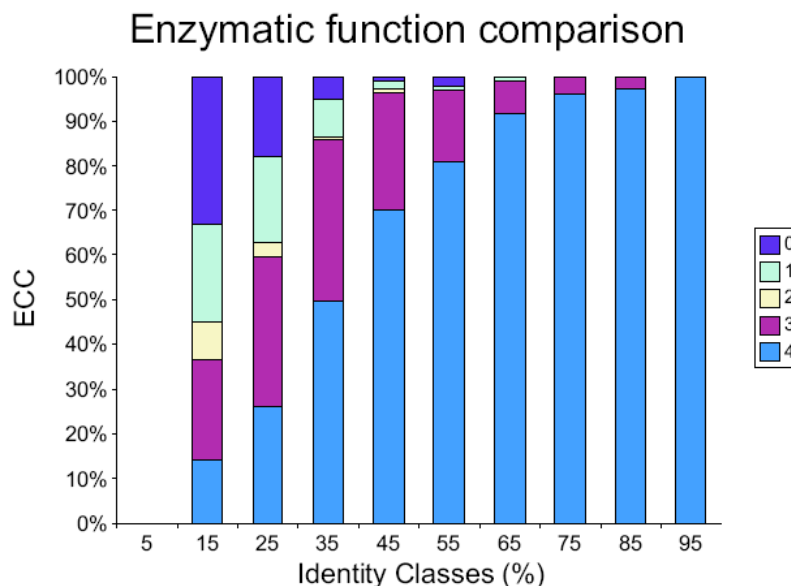
The corresponding results will be used to calibrate the sequence-similarity to function relationship and to deduce a formula applicable to standard sequence database searching programs.

The system will be implemented in an open web server and the results extrapolated to the available genome annotations. In the final phase we will develop a new method for the extrapolation of the pair-wise results to the more complex intra family relationships.

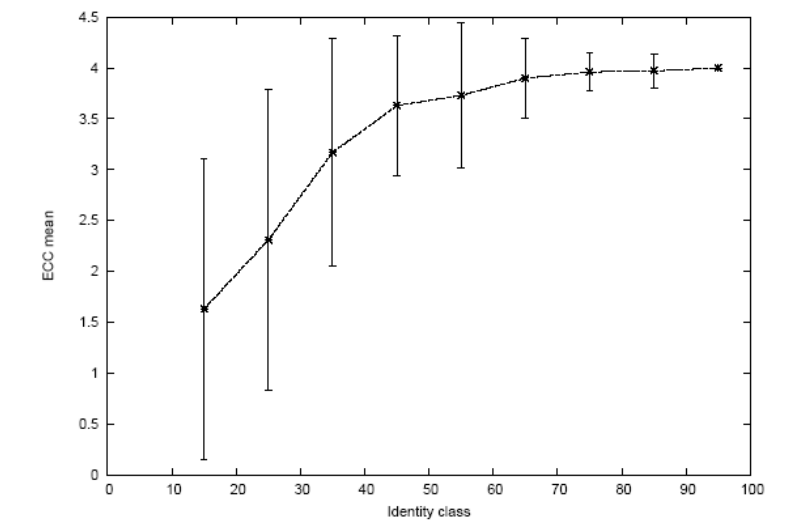
## References

- [1] Devos D. and Valencia A. Practical limits of function prediction. *Proteins*. 2000 Oct 1;41(1):98-107.
- [2] Devos D. and Valencia A. Intrinsic errors in genome annotation. *Trends Genet*. 2001 Aug;17(8):429-31.
- [3] Todd A., Orengo C. and Thornton J. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol*. 2001 Apr 6;307(4):1113-43.
- [4] Rost B. Enzyme function less conserved than anticipated. *J Mol Biol*. 2002 Apr 26;318(2):595-608.
- [5] Hegyi H. and Gerstein The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol*. 1999 Apr 23;288(1):147-64.

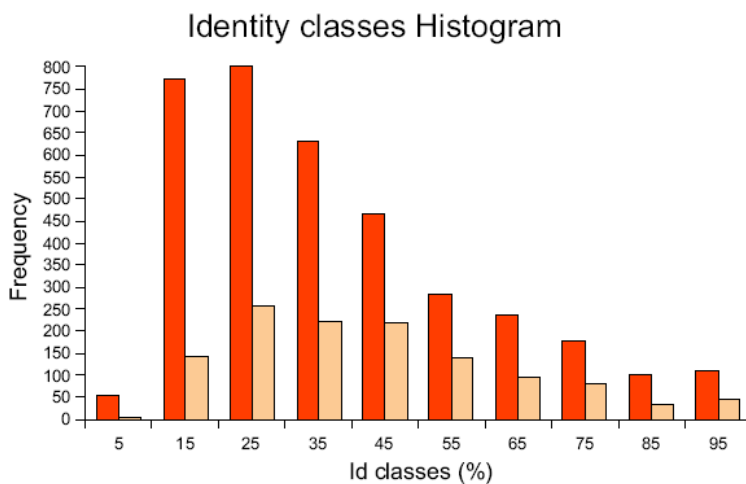
## Figures



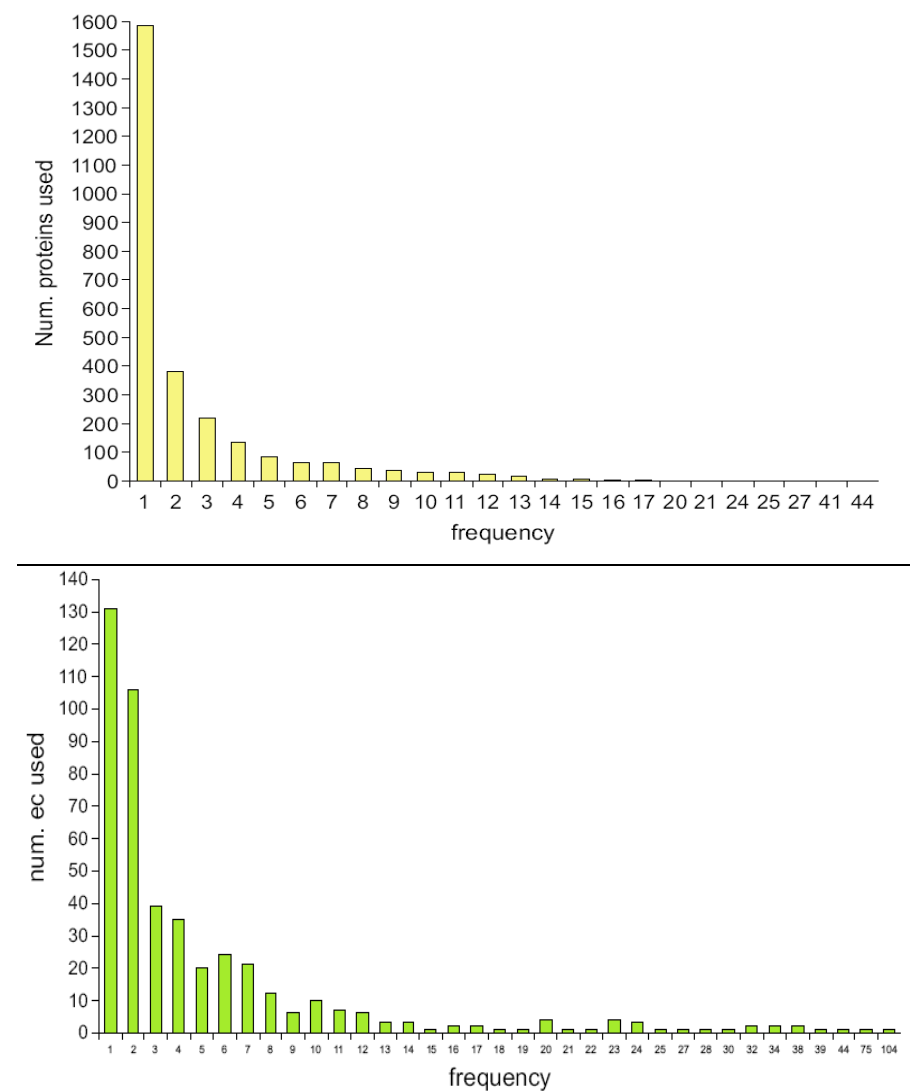
**Figure 1: Distribution representing the level of agreement of the four-digit classification (ECC) across the structural alignments taken from the database Combinatorial Extension (CE).** The ECC is calculated counting the number of code levels shared by the proteins aligned, so, ECC equals to 0 means no coincidence was found between the EC numbers of the proteins while 4 means a complete matching. The data source was filtered to reach a non-redundant (without double pairs), representative (without different chains describing the same structure) and reliable (imposing some restrictions to choose only worthy pairs) set.



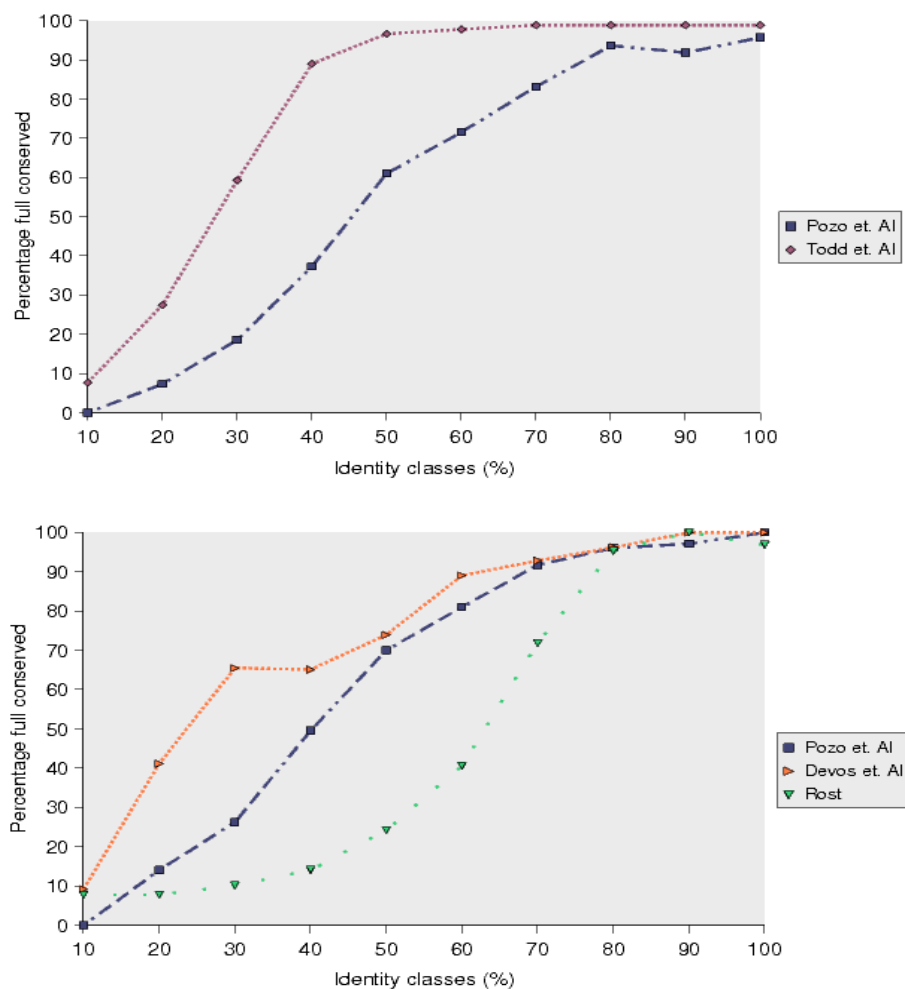
**Figure 2: Mean value of the ECC variable across the sequence similarity levels.** Note the curve grows as expected and the fluctuations become less relevant as sequence identity levels higher.



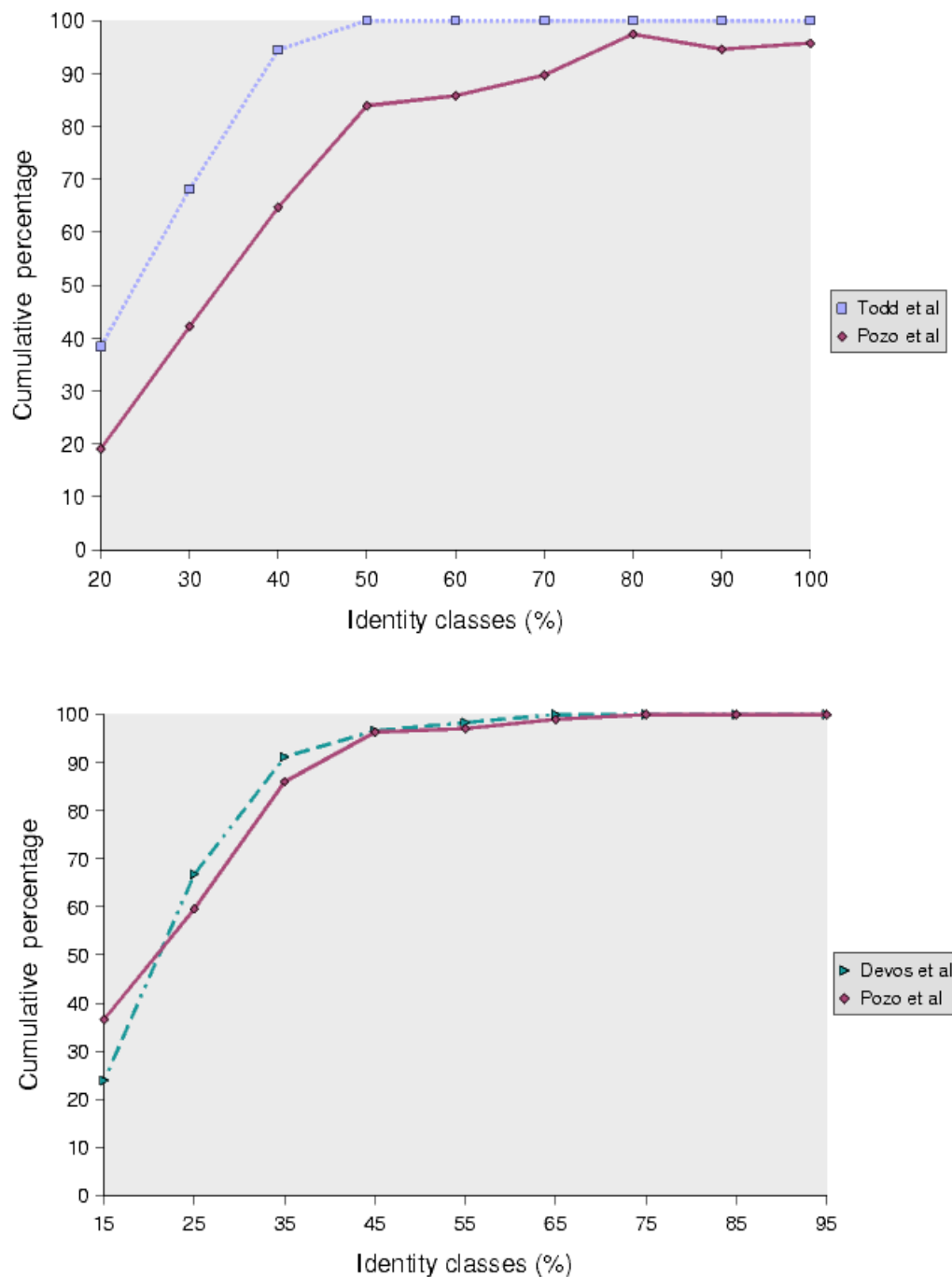
**Figure 3: Distributions of sequence identity values in the dataset.** Frequency of the total number of pairs (dark orange) against the frequency of pairs enzyme/non-enzyme (light orange). In each identity class the rate between both distributions is around 0.4.



**Figure 4: Statistical study of the data set.** Frequency usage of each protein (A) and each EC number (B). This study correspond to a statistical inspect of the data to give away any possible bias introduced in the pairs as result of the filtering process.



**Figure 5: Comparisons of our results with previous publications.** The curves represent the percentage of pairs with all the enzymatic numbers conserved (in the present work we mean ECC equals to 4 ). All the curves represent the same trend of increasing similar function with the increasing in sequence similarity. Todd's curve reaches higher conservation values faster than the others. Our current results show values very closed to our previous work (Devos and Valencia) with a different dataset.



**Figure 6: Comparisons of our results with previous publications.** The curves represent the percentage of pairs with at least 3 numbers conserved. This result overcome fluctuations in the assignment of the enzymatic numbers. The similarity of the the curves of Devos and this work becomes more evident.