Analysis of sequence, structure and evolutionary information in protein superfamilies. Study in Enolase superfamily.

Martín Banchero[1]
1 Bioinformatics Unit, Fundación Instituo Leloir, Buenos Aires, Argentina

Background

Mutations of essential residues in a protein sequence may occur, only if a compensatory mutation takes place elsewhere within the protein to, preserve or restore activity [1]. These co-evolving mutations are of key interest since they may identify residues that interact within the protein to carry out particular functions such as catalytic reactions, structure stabilization, or allosteric regulation.

One of the most important and widely studied problems in protein sequence analysis is identifying which residues in a protein are responsible for protein function.

The enolase superfamily presents a TIM-Barrel superfold, a ubiquitous fold in all kingdoms of life and one of the most frequently observed. Although families within the enolase superfamily differ in their function, specificity, and even present different reaction mechanisms, these enzymes share a common catalytic step of abstraction of the alpha-proton of a carboxylate anion [2].

The objective of this work was to study the extent of the relationship of coevolving positions in the Enolase superfamily by using mutual information estimation (MI), and to analyse different information signals derived from the sequence, structure and evolution in this superfamily. Furthermore we hypothesize that related families contain a distinctive signature of MI, that differs from the MI signal of the superfamily to which they belong, which may involve residues necessary to maintain common features, such as fold or general protein function.

The results indicate that there is a sign of coevolution common to all the four families of the Enolasa superfamily and one sign specific to each family in particular. Furthermore, these results suggest that the superfamily signal obtained is related to the function, as it differs from the expected by chance and differs from the control, which was obtained with another superfamily with the same folding but different function.

We hope that this analysis will serve as a precedent for the development of methods of classification in families and superfamilies.

Materials and Methods

The selected dataset comprises 7251 enolase superfamily protein sequences retrieved from the expert-curated Structure-Function Linkage database [3]. At the same time, these sequences belong to the four families of interest. The families used, the number of sequences, and the PDB code of the reference structures are found in Table 1. The reference structures were chosen to take into account an X-ray resolution smaller than 2.5Å.

As control was used a superfamily with 12453 sequences and the same folding but with different function [4]. In this case, the superfamily is represented by clan CL0160, and all the sequences were obtained from the database Pfam. The number of sequences, Pfam family, and PDB code of the reference structures is shown in Table 2.

Multiple sequence alignments were performed for the superfamily and each family using MUSCLE [5].

Table 1: Dataset used for coevolution calculations.

Family name	Number of sequences	PDBcode(Reference structure)
Mandelate racemase	2327	3QPE
Mannonate dehydratase	256	2QJJ
Muconate cycloisomerase	2439	3CAW
Enolase	2229	1E9I

Table 2: Dataset used as a control. All the families belongs to the clan CL0160 of Pfam. In the last column, the range of residues of the structure belonging to the Indicated family. Sequence similarity networks were performed in order to verify the classification of the different sequences of families belonging to the Enolasa superfamily in the SFLD database. The Sequence similarity networks were calculated with the methodology described by Babbitt and collaborators 2009 [6]. In addition, this analysis allows a quick visualization of the relationships of similarity between each member of the Enolasa superfamily and the degree of similarity between the families.

Family name	Number of sequences	PDBcode(Reference structure)
PF08267	2913	2NQ5(3-309)
PF01717	4790	2NQ5(415-738)
PF01208	4750	1R3S

To know the degree of structural similarity between the members of the Enolase superfamily, all structures of this superfamily obtained from SFLD (table 3) were superimposed using the software Mammoth[7] and a clustering was made based on structural similarity. The values of RMSD100[8] which is a normalized value of RMSD are used as a measure of similarity. Using RMSD100 we become independent of the dimensions of the different proteins to compare.

Table 3: Dataset of structures by famlies.

Family	Number of strutures
Mandelate racemase	57
Mannonate dehydratase	15
Muconate cycloisomerase	49
Enolase	17
Total	**138**

The MI scores were calculated as described in [9]. To compare the MI values of the different families of the superfamily Enolase, one structural alignment was performed with POSA [10], through a rigid overlap of the four reference structures(Table 1), being one from each family. From this overlap, a sequence alignment was derived to which were removed all the columns of the alignment that contained at least one gap (not overlapping in space). This is important because is compared to the positions of the four proteins which are structurally equivalent (overlapping of equivalent positions of the four proteins in the space).

Then, using the edited sequence alignment, the MI values associated with these positions were compared between different families.

For each family, MI scores were represented in binary matrices of mutual information.

To obtain the binary matrices were used the values of MI belonging to the pairs of residues remaining in the edited alignment. The binary matrix was filled with zeros and ones according to whether the MI value of the represented pair equals or exceeds the cutoff value of 6.5. In case of exceeding or equalizing the cut-off is assigned a value one otherwise it is assigned zero.

To generate the matrices were used the positions of the edited alignment, consequently, all the resulting matrices have the same size. As a result, the matrices are symmetric and with a size of 215x215. For each matrix, was used only the upper diagonal matrix without including the diagonal. This due to the diagonal is the MI of a position with itself.

Using these matrices is more simple to analyze the distribution of the pairs that exceed the cut-off value of MI ≥ 6.5 along with the entire matrix.

The scripts to performed the matrices were performed using the programming language “R” [http://www.r-project.org/].

Due to the dots obtained for the different pairs of residues in the superfamily matrix of MI could have been obtained by chance, was performed a statistical analysis *of Boostrap* to validate the results. For this, random matrices were generated with the same number of dots as the original matrix.

One possibility would be to randomize all points inside the matrix, then perform the superposition of these matrices to obtain the number of pairs of residues with high MI that are superimposed and also expected by chance. However, this randomization does not consider that there are regions of the matrix with a higher density of points than others.

In order to respect the distribution of the data, it was decided to carry out randomization by regions. The test consists of generating submatrices of 5x5 within the matrix of each family and randomized the dots belonging to these 5x5 submatrices.

Results and Discussion

The sequence similarity networks were calculated as done by Babbitt and coworkers [6] as explained in the Methods section. This analysis allows us to understand the relationships of similarity within each family and between families; As well as verify the classification assigned in SFLD.

With a cut-off value (Ev) equal to 1e-2 and applying a clustering algorithm, sequential similarity classification alone would be sufficient to classify families correctly. However, this is not information available a priori.

Figure 1 shows the network that was obtained with an E-value of 1e-2. It is observed that families have subdivisions within each cluster, this could be due to the presence of different subfamilies within each family. The similarity network also shows that the Enolasa family is the one that presents less degree of similarity with the rest of the families, only has a connection with the Muconate family. On the other hand, Muconate presents greater similarity with Mandelate and less similarity degree with Mannonate. Moreover, the absence of lines with the control group confirms that the value of Ev used implies a relation between the sequences of the superfamily. Furthermore, there is no connection with the control group but with different families within the superfamily Enolase.

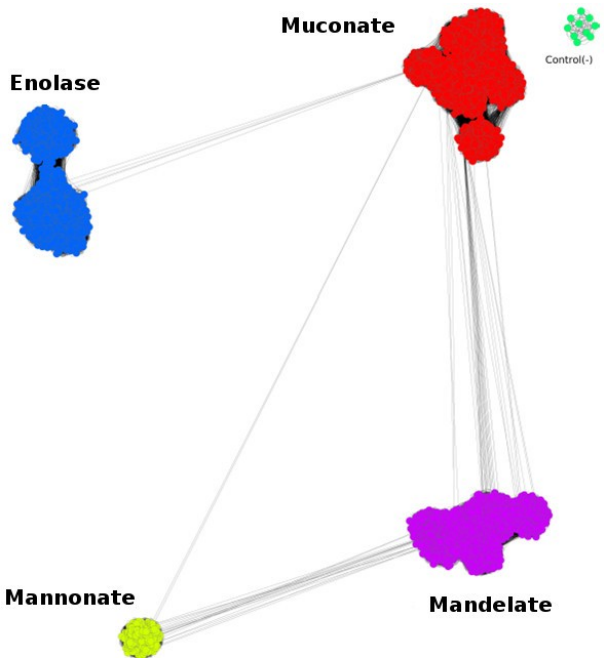

Figure 1: Sequence similarity network for the four families of the Enolasa superfamily. Each sequence is represented by a node and a line is drawn between the nodes only if the value of Ev between them is smaller than $1e^{-2}$. The nodes were colored according to the family to which they belong. It can be seen that the sequences are separated into four groups one for each family for this Ev value.

To verify that knowing the structures is not enough to divide the superfamily into families (even in the hypothetical situation of having the structures of all sequences), the result was evaluated making a clusterization by structural similarity and representing the result as a dendogram(Fig.2).

To performed the dendogram were obtained the values of RMSD100 as described in Materials and Methods. This dendogram was cut to generated 4 clusters. In the next step the clusters were analyzed to evaluate if the structures were grouped in a similar way to the similarity network (correctly according to SFLD).

The cut-off value of the dendogram was RMSD100 equal to 2.69, while the average of values of RMSD100 of each cluster are shown in Table 4.

Table 4: The table shows mean values of RMSD 100 for each of the four clusters obtained.

Cluster ID	$RMSD_{100}$
1	2,62
2	1,46
3	2,56
4	2,69

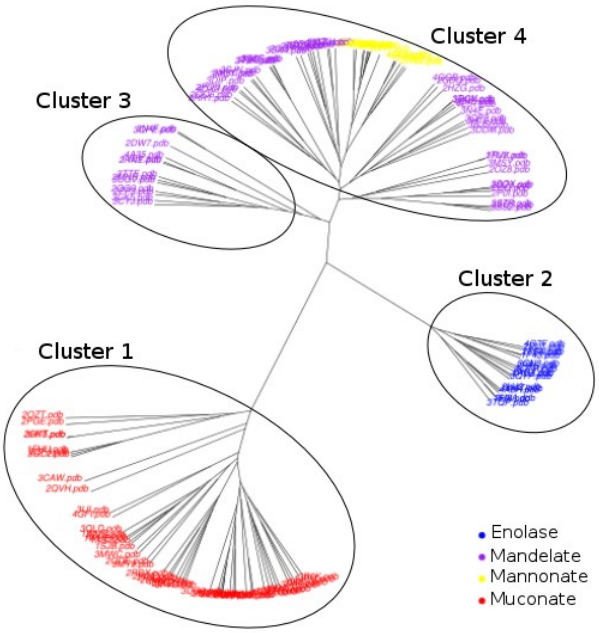

Figure 2: Tree without root as a result of structural comparison. The distribution of the members of the four families in four groups generated according to the structure. It can be observed that the sequences of the Mannonate family is integrate to cluster 4 together with Mandelate. This last family is divided in two groups which are cluster 3 and cluster 4.

\

Although two families are well separated (cluster 1 and 2) there are two other families that did not (cluster 3 and 4). This shows that the structural similarity in this case does not allow families to be grouped as the SFLD database (used for sequential information classification).

In this case the structural similarity contributes with information of the structures which is complementary to the information obtained only from sequences.The MI was calculated for each family in order to find the positions that could have coevolved. The positions with high MI of the four families were compared, in order to verify if there is an MI signal shared by the four families. Furthermore, the binary matrices of MI with the exclusive points of each family of Enolase were represented in Figure 3.a and in Figure 3.b for the clan CL0160. This was done to have a clearer view of the distribution of dots along the matrices.

The binary matrices of MI obtained for each superfamily were represented in Figure 4.a for Enolase and in Figure 4.b for CL0160.

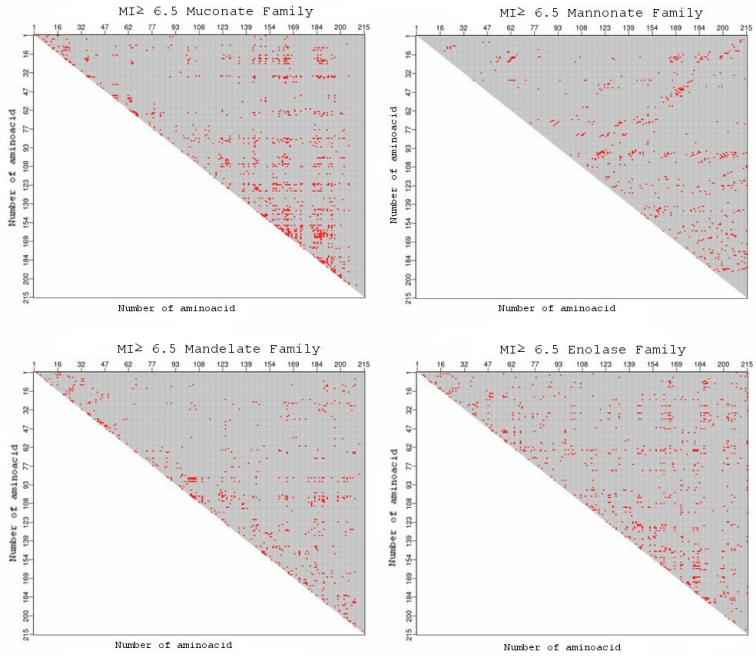

Figure 3.a: MI matrices for each family of the Enolasa superfamily. The positions of the sequence are represented in each matrix. The MI score greater or equal than 6.5 is represented as a red dot, otherwise it is gray.

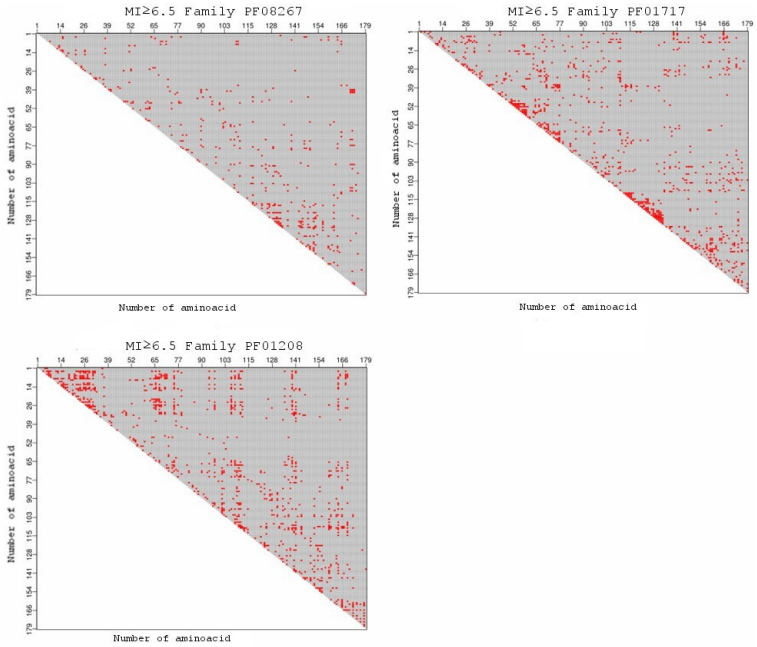

Figure 3.b: MI matrices for each family of the clan CL0160. The positions of the sequence are represented in each sequence. The MI score greater or equal than 6.5 is represented as a red dot, otherwise it is gray.

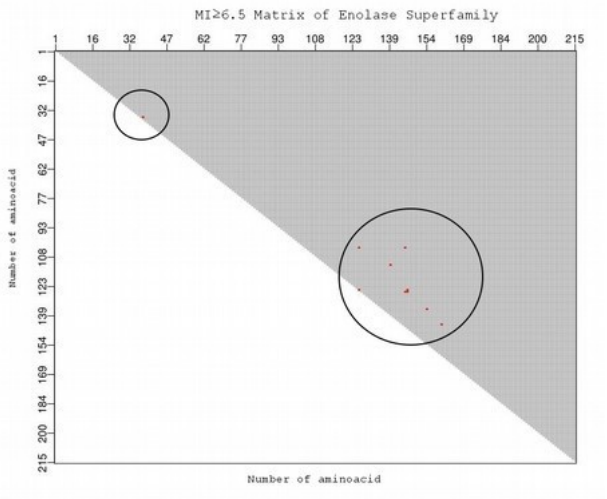

Figure 4.a: Superfamily binary matrix. Ten pairs of residues with high MI were observed that have in common the four families of the superfamily Enolasa. These are shown in red and indicated by the circles.

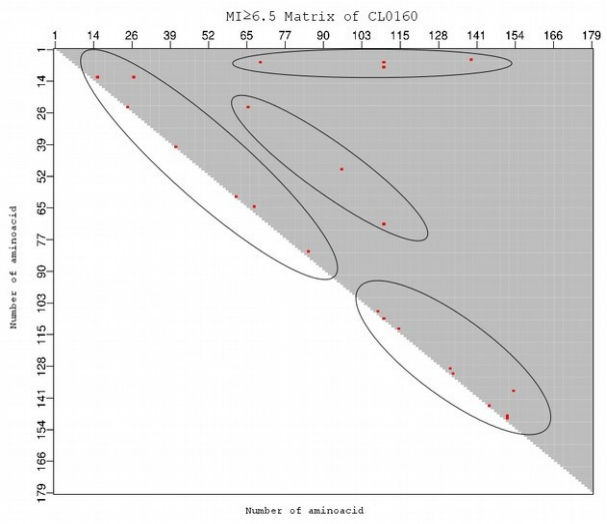

Figure 4.b: Binary matrix for Clan CL0160. The twenty three pairs of residues with high MI that the three families have in common are shown in red.

As mentioned in Materials and Methods, a statistical analysis *Boostrap* was performed to validate the results, due to the points obtained for differents pairs of residues in the MI matrix of superfamily could have been obtained by chance.

The result of the statistical evaluation of the superfamily MI matrix is a histogram of frequencies that is observed in Figure 5. In this, it is observed the frequencies of shared positions with significant values of MI for different random superfamily matrices.

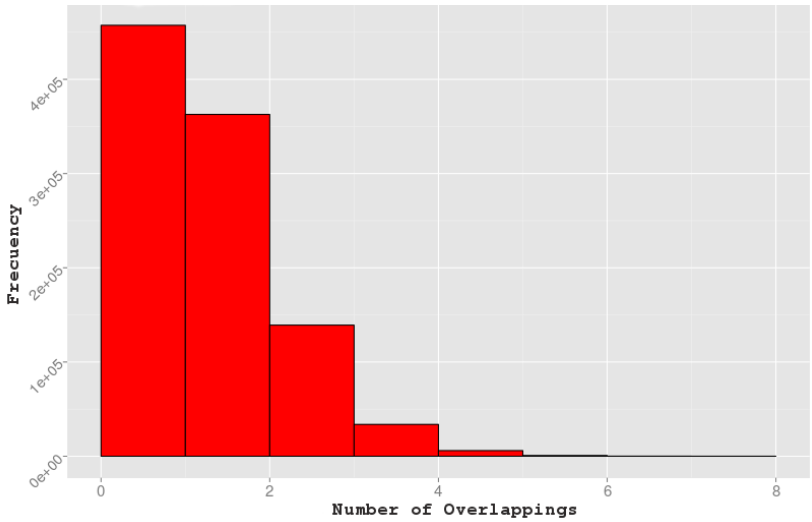

Figure 5: Histogram of frequencies. Represents the number of positions at which a positive value is superimposed for the four random matrices generated with the same number of dots as the original matrix of Enolase superfamily.

In this analysis were generated 1x106 random matrices, in only one overlapping of the 4 matrices we get eight matches, this is expected due to the number of dots obtained decreases exponentially, this in principle leads us to think that the dots obtained for the superfamily Enolasa are not due to chance.

If we think of the probability of obtaining 10 points by overlapping four random matrices, just like those obtained for the superfamily matrix of Enolase, it would be less than 1e10-6. This result is significant considering the procedure of creating the random matrices. They have not been filled randomly with the same amount of points, but it has been taken into account that there could be an error given the obligatory sequence of amino acids.

Conclusions

There are a group of ten pairs of amino acids with high MI that are equivalent for the 4 families, this number of pairs is not expected by chance. It was ascertained by comparison with another superfamily with the same fold (CL0160), that these ten pairs of amino acids are not a product of the structure and are specific to the superfamily analyzed. With these results, We can conclude that exists a signature of MI for the superfamily, which is intimately related to the general function of that particular superfamily. Some groups coevolve (residues with high MI) specific for each family within the Enolasa superfamily, these will be important for the proper function of each family.

Due to the sequence analysis, it was observed that the Enolase family had the lowest degree of sequence similarity with the rest of the families. On the other hand, Muconate and Mandelate presented greater similarity. In addition, Mandelate does not present a relation with the Enolasa family as shown by our similarity network.

The results by comparison of the three-dimensional structure suggest that there is a partial coincidence with the results of the analysis of sequence similarities, due to the Mannonate family presents greater structural similarity with Mandelate (more than Mandelato within itself). From this analysis, we can conclude that the structure alone does not give enough information to classify the sequences in families, even in a hypothetical situation to have a structure for each sequence.

For the four families, a group of 10 pairs of residues was determined that coevolved in common.

Having found an MI signal shared for the four families (own superfamily) and MI for each family, the results show that the MI could provide information for the classification of sequences in families, or to assign families to superfamilies.

Due to the proteins of a superfamily have a common ancestor, we hypothesize that functional diversity within superfamilies is given by a series of concerted changes that have to leave a recognizable coevolutionary sign. Our objective will be to identify this signal to divide the superfamilies among families.

Bibliography

[1] L. C. Martin, G. B. Gloor, S. D. Dunn, and L. M. Wahl, “Using information theory to search for co-evolving residues in proteins.,” Bioinformatics, vol. 21, no. 22, pp. 4116–24, Nov. 2005.

[2] J. F. Rakus, A. A. Fedorov, E. V Fedorov, M. E. Glasner, J. E. Vick, P. C. Babbitt, S. C. Almo, and J. A. Gerlt, “Evolution of Enzymatic Activities in the Enolase {Superfamily:D-Mannonate} Dehydratase from Novosphingobium aromaticivoran,” Biochemistry, vol. 46, no. 45, pp. 12896–12908, Nov. 2007.

[3] S. D. Brown, J. a Gerlt, J. L. Seffernick, and P. C. Babbitt, “A gold standard set of mechanistically diverse enzyme superfamilies.,” Genome Biol., vol. 7, no. 1, p. R8, Jan. 2006.

[4] J.-L. Ferrer, S. Ravanel, M. Robert, and R. Dumas, “Crystal structures of cobalamin-independent methionine synthase complexed with zinc, homocysteine, and methyltetrahydrofolate.,” The Journal of biological chemistry, vol. 279, no. 43, pp. 44235–8, Oct. 2004.

[5] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput.,” Nucleic acids research, vol. 32, no. 5, pp. 1792–7, Jan. 2004.

[6] H. J. Atkinson, J. H. Morris, T. E. Ferrin, and P. C. Babbitt, “Using sequence similarity networks for visualization of relationships across diverse protein superfamilies.,” PloS one, vol. 4, no. 2, p. e4345, Jan. 2009.

[7] A. R. Ortiz, C. E. M. Strauss, and O. Olmea, “MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison.,” Protein science : a publication of the Protein Society, vol. 11, no. 11, pp. 2606–21, Nov. 2002.

[8] O. Carugo and S. Pongor, “A normalized root-mean-square distance for comparing protein three-dimensional structures.,” Protein science : a publication of the Protein Society, vol. 10, no. 7, pp. 1470–3, Jul. 2001.

[9] C. M. Buslje, J. Santos, J. M. Delfino, and M. Nielsen, “Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information,” Bioinformatics, vol. 25, no. 9, pp. 1125–1131, May 2009.

[10] Y. Ye and A. Godzik, “Multiple flexible structure alignment using partial order graphs.,” Bioinformatics (Oxford, England), vol. 21, no. 10, pp. 2362–9, May 2005.