Proteinas Canijas

Ale Zavala[a], Sergio Hernández[a], Pedro Miramontes[a], León Martínez[b]

[a]*Facultad de Ciencias, National Autonomous University of Mexico, Mexico City 04510, Mexico*
[b]*Metropolitan Autonomous University*

Abstract

¿Qué proponemos como abstract?

Keywords: Intrinsically disordered protein regions, ordered protein regions, Pfam family, protein domain, The discrete generalized beta distribution, megasecuence, familiome, self-organizing map, complexity

1. Background

Historically, the science of protein structure has privileged the study of ordered protein regions (OPRs) with regular and well-defined secondary structure, such as α-helix, β-conformations and turns. In fact, proteins that contain mainly ordered regions are the classic text-book example of what a typical protein should look like. This is probably associated with the fact that the most commonly used technique for determining protein structure, i. e. x-ray crystallography, is biased towards resolving protein regions with regular and well-defined secondary structures. However, it has recently become increasingly clear that intrinsically disordered protein regions play an important role in the structure of proteins [**?**].

Intrinsically disordered protein regions (IDPRs), are segments of polypeptides, that do not natively acquire well defined, regular, or repetitive, secondary structures and instead adopt many different structural ensembles with no single, preferred lowest energy conformation and have biological activities that are dependent on this disordered state. There are some computational tools that can predict the order/disorder state of a protein region using only the complete polypeptide sequence as input [**?**], [**?**]. IDPRs have been found to be prevalent in eukaryotes and are common in bacteria and archaea too (XXX? ¿PEDRO: que significa esta nota?) [**?**] [**?**] [**?**] [**?**]. A general function that has been proposed for many IDPRs is the binding to other molecules, including proteins, nucleic acids and other ligands [**?**] [**?**] [**?**]. However, it is possible that the importance of IDPRs goes beyond this currently accepted function, and that they may even play an important role in the folding and functioning of proteins. We would like to emphasize that in this study we have tried to jointly analyze both the OPRs and IDPRs. This approach derives from our belief that at least some of the properties of proteins are the product of the interactions between OPRs and IDPRs. Furthermore, we hypothesize that the distribution pattern of OPRs and IDPRs within a given protein domain is specific to that domain (similar to a fingerprint) and also specific to that domain's family. We will expand on the definition of protein families and other related concepts in the following sections.

Here, we will only mention that studying patterns of protein order/disorder at the level protein domains (instead of single proteins) allowed us to capture amino acid sequence variability that was, in principle, compatible with the tertiary structure conservation that characterizes a domain. Additionally, each representative of a domain acted as a new quasi-replicate thereof, allowing us to (naturally) increase the sample size of the data series to be analyzed with our proposed statistical tools (see below). Furthermore, we augmented our protein sequence data with gene ontology (GO) annotations in order to explore whether functional information of protein domains correlates with their order/disorder regularities.

Measures of complexity of patterns of OPRs and IDPRs have the potential to contain information on structural and functional aspects of proteins in the sense that members of a given protein family that are close to each other in the parameter space of some measures of complexity could also be close in terms of functional properties. If IDPRs are an integral part of proteins the patterns of alternating OPRs and IDPRs of varying lengths should be relatively conserved among members of a protein family. Potentially, measures of the complexity of such patterns could capture enough information thereof, such that they could serve as a proxy for characterizing specific proteins.

In this study we will use several measures of complexity of a polypeptide pattern. At this time "complexity" can be understood as how close or how far a sequences is from two extremes: randomness or periodicity. On one hand, he well known Kolmogorov's complexity index (K) provides a measure of how far a sequence is from being random and, on the other, Shannon's entropy (H) tells us how far is a sequence from the equiprobable distribution

(REFSXXX). Additionally, the recently proposed Discrete Generalized Beta function (DGBF) of rank-ordered distributions has also emerged as a measure of complexity. Interestingly, the DGBF can separate intermittent regimes from chaotic dynamics and may serve as an indicator of transitions between these two regimes in a wide array of phenomena [?].

Here, we propose the use of the above mentioned descriptors to place any given protein into a space of parameters and to discuss they way different families of proteins group in such a space. Our working model will be the Saccharomyces cerevisae proteome. We discuss the possibility that the DGBF of patterns of protein order/disorder may be able to detect evolutionary transitions in which ordered proteins acquire or expand their disordered regions or in which disordered protein start to limit their conformational repertoires.

2. Materials and Methods

2.1. Construction of the database of complete sequences represented in the yeast familiome

In this study, our aim is to characterize protein families in terms of the entropy, algorithmic complexity and characteristic beta function of their order-disorder pattern. A family of proteins is a group of proteins or protein domains that share patterns of significant sequence conservation, due to common ancestry, that frequently entail functional similarity [?]. A protein domain is a substructure produced by any contiguous part of a polypeptide chain whose structural features are independent of the rest of the protein. A domain usually contains between 40 to 350 amino acids, and it is the modular unit from which many larger proteins are constructed [?]. It is important to note that any given protein does not necessarily belong to a single family, as a given protein can contain several domains with different evolutionary histories, and indeed, many proteins belong to several families [?]. Although most protein domains that are identified using sequence-based approaches are have well-defined and relatively stable spatial structures, some can be fully or largely disordered or can contain conserved disordered regions [?], these are known as intrinsically disordered domains (IDDs; [?]). The protein families information is provided by the Pfam database ?] [?] [?] [?]. Pfam is a collection of protein domains and protein families in which each family is represented by two multiple sequence aligments and two profile-Hidden Markov Models, one of the two alignments is a high quality seed alignment [?] [?].

To build our yeast familiome database, we explored protein information in several biological databases. To every translated gene from Saccharomyces Genome Database (v2015; https://www.yeastgenome.org/; [?]) we associated an UniProt identifier and its complete polypeptide sequence ("uniprot" full file, v2014; [?]). Using the UniProt information, we linked to every yeast protein its corresponding Pfam families (v28, 2015; https://pfam.xfam.org/; [?] [?]). Subsequently, for each Pfam family from S. cerevisiae familiome we downloaded its Pfam-A seed alignment file (v28, 2015; [?] [?] (XXXXXXXX ?) , which contains only the aligned segments that belong to a protein family of a variety of species. Finally, we used the UniProt identifier provided by the Pfam-A seed alignment file and the UniProt full file to obtain the complete polypeptide sequence of each protein in the Pfam-A seed alignments of the yeast familiome.

2.2. Predicting intrinsically disordered residues in each polypeptide sequence of yeast familiome

In order to assign each residue from our complete polypeptide sequence yeast familiome database to either the "ordered" or "intrinsically disordered" categories we used the open-source DisEMBL prediction software. DisEMBL is based on artificial neural networks trained to predict three different definitions of disorder and displays the disordered segments of arbitrary length within a protein sequence [?]. We used the three differents algorithms of DisEMBL: loops/coils, hot loops and remark465 and the final assignment of each residue as either ordered or disordered residue was based on a majority rule decision between the three predictions. The order/disorder information was coded into the sequences as UPPERCASE/lowercase one-letter amino acid symbols, respectively. This procedure was performed for each complete polypeptide sequence of each Pfam-A seed alignment family in the yeast familiome. All the members in one family were concatenated together into one big megasequence.

2.3. Transferring the ordered/disordered information to the Pfam-A families seed alignment

In our study, we needed to associate sequences of the Pfam-A families seed alignment of the yeast familiome with the DisEMBL majority rule decision results. In order to do this we used options of the MAFFT program to maintain the Pfam-A families seed alignment unchanged, to maintain the gaps, the UPPERCASE/lowercase in the alignment and to preserve intact the order of the residues [?] [?] in the Pfam-A families seed alignment of the yeast familiome.

2.4. Gene Ontology annotations in the yeast familiome

To enrich our yeast familiome database, for those Pfam families where this information was available, we associated the molecular function annotation from Gene Ontology (v2018; http://geneontology.org/; [?] [?]). We needed a general molecular function annotation so we manually curated the specific GO molecular function annotations of the yeast familiome.

2.5. Megasequence construction

In order to have a sequence big enough to statistically represent the whole family in a robust way, we constructed what we called a megasequence which consisted in all the domain instances within a family glued together, so the statistical regularities will be magnified and easily observable in the so called megasequence.

We took all the aligned sequences for a given domain and spliced them one after another to obtain a family megasequence.

2.6. The discrete generalized beta distribution (DGBD)

All these megasequences of the yeast familiome were compared using a discrete generalized beta distribution (DGBD) [?] [?] which is a rank ordering distribution that takes the form:

$$f(r) = \frac{A(N+1-r)b}{r^a}$$

Where a and b are parameters to be found, N is the number of ranks and A is a normalization constant. This rank ordering distribution has been successfully used across a wide range of different phenomena regardless of the truncated scaling behavior shown typically in most of the rank-order distributions. The approach used in our analysis is as follows. We took all the aligned sequences and merge them together one after another to make a family megasequence, then we counted the frequencies of the different words of length 2 and ordered these distributions of sizes in decreasing order. Then, through a nonlinear fit of (1) we obtained the (a,b) pair which was used to represent the distribution.

2.7. Shannon Entropy (H(X))

Another attribute added to the whole yeast familiome was the calculation of Shannon's Entropy for each of the sequences of the larger. Shannon's Entropy is defined as follows:

$$H(X) = \sum \frac{p\_i}{log(p\_i)}(2)$$

Where p_i is the probability of one of the N amino acids in the megasequence X. Applying H(X) to every family megasequence we have will reveal which sequences are the furthest from the normal distribution and so, which megasequence has the most structure in it [**?**].

2.8. Kolmogorov complexity

Kolmogorov complexity index when applied to a string of characters, in our case is the megasequence X, can be interpreted as the complexity of a computer program required to reproduce megasequence X. The calculation of Kolmogorov's complexity index can be approximated as follows:

$$k(seq) = \frac{length(compressed(seq))}{length(seq)}(3)$$

Where seq is the original megasequence of some family of proteins. The actual implementation of this function was done in Python computer language where zlib libraries were used to compress every sequence of the familiome [**?**]. In our experiment we used K as another attribute together with the already described, in order to understand the algorithmic complexity assumed to exist in every family. That is, if a set of instructions is behind the description of every protein in each family, we would expect that K captures this particular complexity.

2.9. Self-organizing map

The self-organizing map (SOM) is an unsupervised neural network used for data analysis and dimensionality reduction [**?**]. It has been long being applied into data analysis and biological sciences to detect similar profiles of analyzed data [**? ?**]. The basic algorithm is divided in two steps. First, an initial map is formed with N neurons arranged in a lattice which will represent M d-dimensional vectors. Each of the N neurons contains a single d-dimensional prototype that will be modified during the training of the map. Then, for each vector sample a winning prototype must be found. The second step consists in modifying the prototypes of all vectors within a neighborhood of this winning neuron, the magnitude of the modification is in proportion to the distance of the winning neuron. This process ends up unfolding a map where each prototype in neurons represents local averages of data, hence nearby neurons have similar prototypes. Once the SOM is formed, locally grouped families must be assigned to a group. This step is done by a hierarchical clustering using euclidean distance. The number of groups was determined by the Davies-Bouldin index.

3. Results

3.1. Saccharomyces cerevisiae familiome database

Our yeast familiome database contains 538 Pfam families (protein domains) and a total of NNNN instances of domains. The size range of the families goes from a family containing XXXX instances of a domain (Family number PF????) to a family containing YYYY instances (Family PF¿¿¿¿) All instances of a given domain were concatenated to obtain a megasequence, thus yielding a total of 538 megasequences. Each megasequence contains information of ordered/disordered status for all its residues, as well as the values of the parameters α and β from the expression of DGBD, Shannon's entropy, and Kolmogorov complexity. Additionally, for 260 families we have the specific GO molecular function corresponding to 18 different GO categories namely "hydrolase", "electron transfer activity", "protein dimerization activity", "isomerase", "motor activity", "transferase", "transmembrane transporter", "translation initiation factor activity", "binding", "translocases", "antioxidant activity", "structural constituent", "oxidoreductase", "ligase", "structural molecule activity", "catalytic activity", "lyase", and "copper chaperone activity". This information is shown in supplementary material S1-Yeast Familiome.

3.2. Discrete Generalized Beta Distribution to characterize S. cerevisiae familiome

The Discrete Generalized Beta Distribution used in this work is a novel probability function that it has been shown to be an alternative to fit data that does not fit Zipf's law perfectly nevertheless an underlying process alike seems to be taking place [?] In some instances the α exponent can be related to behaviors generating power laws, as is the case of scale invariance in turbulence in the so called inertial range where energy is transferred between different scales at the same rate, while β seems to be associated with chaotic, disordered fluctuations, for example the dissipative range for turbulence. In contrast with classical powerlaw like functions, the DGBD is able to encompass both the scale invariance and chaotic regime in depicting the whole process and its conflicting dynamics in the same graph. This gives a general representation of the phenomena under study [?].

There is a wide variety of a distinct phenomena studied under the DGBD as shown in [?] and specifically there is some research in the field of genomics as shown in It is worth noting that the role of exponents α and β as universality classifying parameters, as for example in critical phenomena, remains be investigated in further detail.

We constructed DGBD plots for the 580 families and we are showing the best DGBD plots according to the following criteria: first, a Pfam family needed to have the highest square of correlation coefficient; second, an sample size N of at least 30 different domain instances; and finally the DGBD alpha and beta values had to be ≥ 0. Every panel in each figure indicates the Pfam family, the square of correlation coefficient, the alpha and beta values, and its N value.

In the familiome database, the highest alpha value was 2.0027 and the lowest alpha value was 0.1003. Figure 1 shows selected cases for the scenario where $\alpha > \beta$.

In this database, we have the best 30 DGBD graphs with high alpha and low beta values and there are 5 different general GO molecular functions belonging to 11 different Pfam families. Seven of these eleven have a "binding" GO, 2 have "ligase" GO and "oxidoreductase", "isomerase" and "lyase" ontologies have one each (Figure 1). Althought there are seven general binding ontologies, we cannot group them because their specific ontologies are different. We have two "protein binding" , one "ATP binding", one "thiamine binding", one "DNA binding", one "metal ion binding", and one "phosphatidylinositol binding". For "ligase" ontology, we have two different Pfam families with specific ontologies "like aminoacyl-tRNA ligase activity" and "aminoacyl-tRNA editing activity". In the case of a family with "lyase" ontology, its specific ontology is "phosphatidylserine decarboxylase activity". For the families with "oxidoreductase" and "isomerase", there are not specific ontologies.

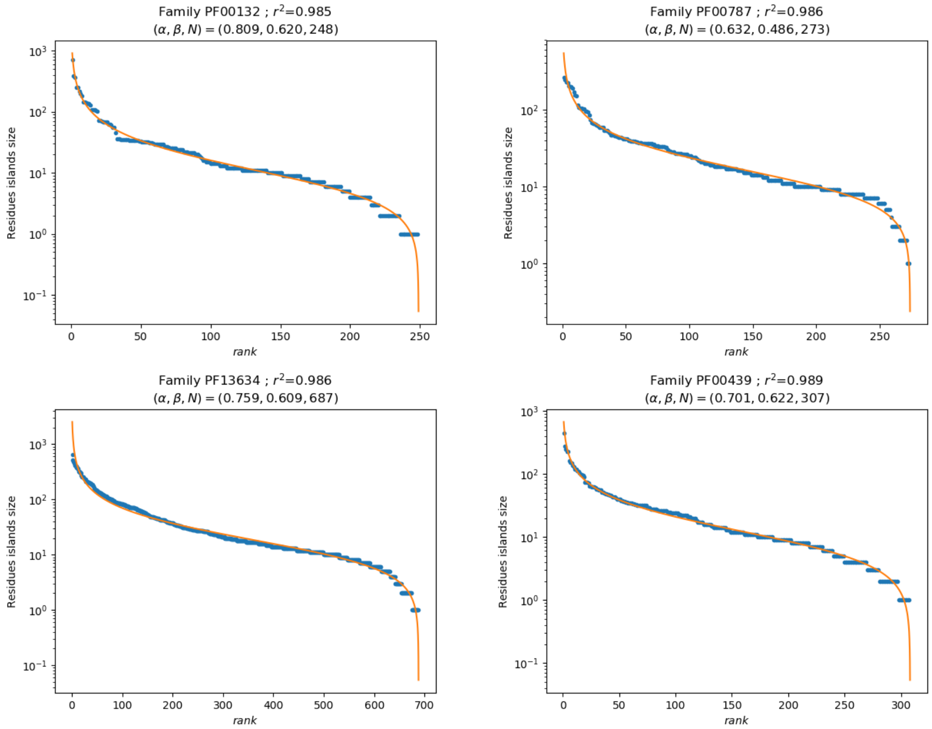

Figure 1: Semilog DGBD plots with high α and low β values. Orange lines represent the best fitting DGBD model for the corresponding experimental data points (shown in blue). Values for the α and β parameters of the DGBD as well as sample size are shown on top of each panel. PF00787 and PF00439 families have binding molecular function ontology, whereas PF00132 and PF13634 do not have an assigned GO. Notice that x- and y-axis scales are different among panels.

In the yeast familiome database, the highest β value was 1.6529 and the lowest β value was 0.0236. Figure 2 show the cases in which $\alpha < \beta$, namely 4 from among the best 30 DGBD graphs with high β and low α values. There can be found 5 different general GO molecular functions associated with different families. Seven families have“binding” ontology, whereas “transferase”, “translocase”, and “ligase” are found in only one family each (Figure 2). Although the binding function has a clearly higher prevalence, we have 9 different families with different molecular specific binding function. We have 3 families with “protein binding” and one family with both “protein binding” and “ATP binding”, two families are labelled “DNA binding” and one family is labelled “DNA binding” and “RNA polymerase activity”, one family has “RNA binding”, one family is labelled “metal ion binding”, and two families with other specific ontologies like “proton-transporting ATP synthase activity” and “aminoacyl-tRNA editing activity”. In this case, the distribution falls rapidly and with high beta values, the rank is minor in the distribution por lo qué...... [**?**].

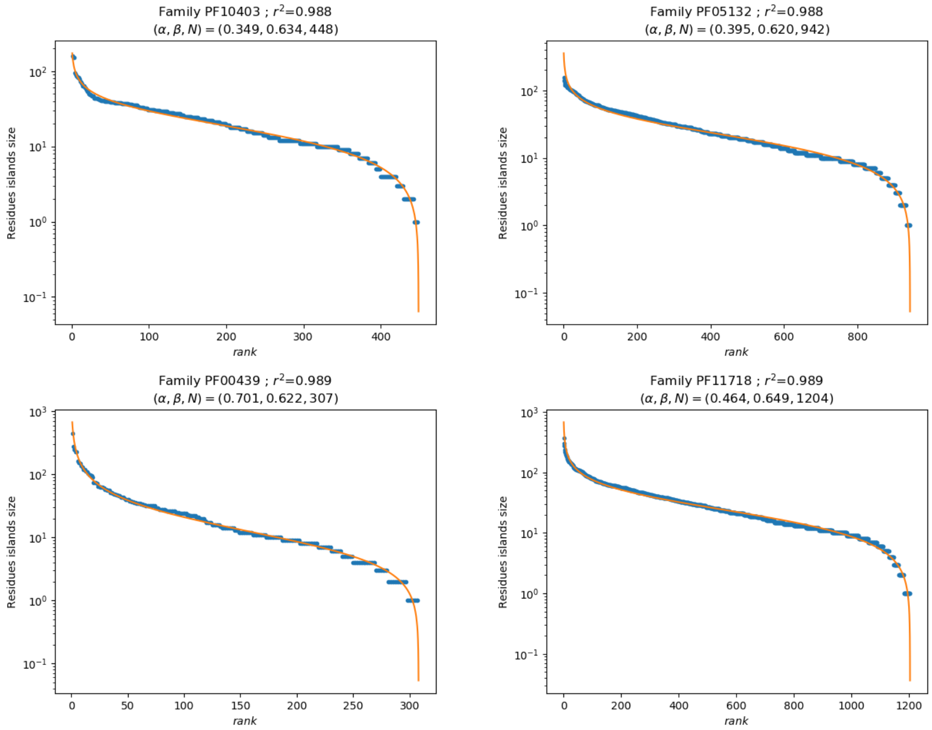

Figure 2: Semilog DGBD plots with high β and low α values. PF05132 family has transferase and binding ontology. Orange lines represent the best fitting DGBD model for the corresponding experimental data points (shown in blue). Values for the α and β parameters of the DGBD as well as sample size are shown on top of each panel. PF00439 and PF10403 share binding ontology. Notice that x- and y-axis scales are different among panels.

In the yeast familiome database, 28 different families had α and β with values nearest to 1 ±0.10. There are 3 different generals GO molecular functions belonging to different families. Six Pfam families have a "binding ontology", two have "ligase ontologies", and "hydrolase", "initiation factor" and "cooper chaperone ontologies" are found in only one family each (Figure 3). We have 6 families with different "binding" ontologies and other ontologies. One family have "zinc ion binding" and "nucleic acid binding", one family have metal "ion binding" and "translation initiation factor activity"; another one family have "ATP binding", "nucleotide binding" and "aminoacyl-tRNA ligase activity"; another one family have "copper ion binding" and "copper chaperone activity"; another one family have "RNA binding", and the last family have "ATP-dependent 3'-5' DNA helicase activity" and "aminoacyl-tRNA editing activity".

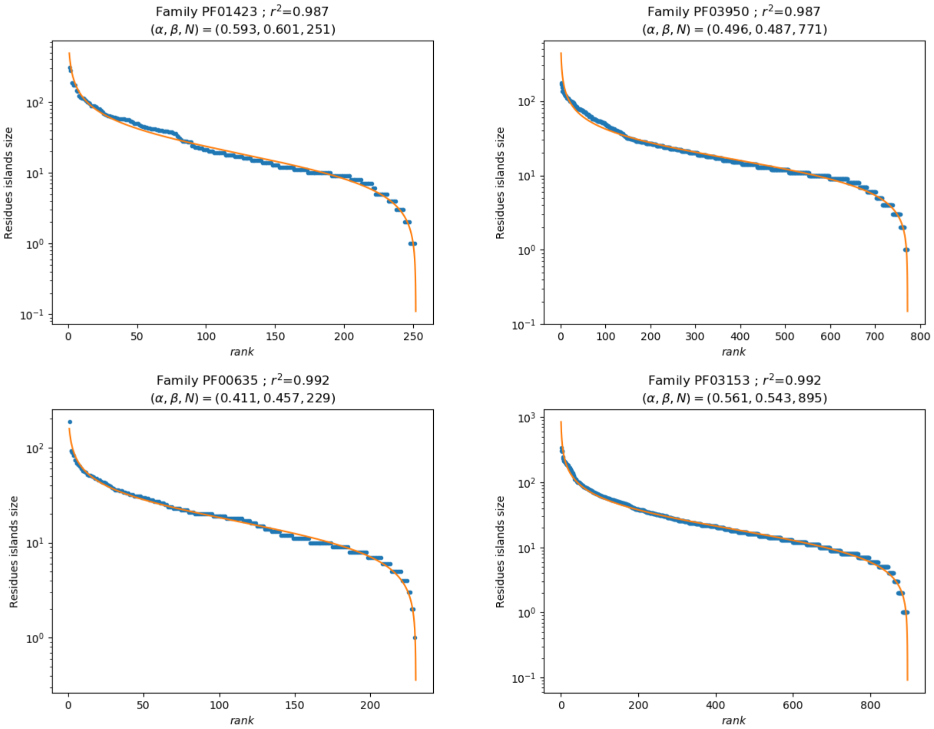

Figure 3: Semilog DGBD plots when α equals β. PF03950 has binding and initiating factor ontologies. Orange lines represent the best fitting DGBD model for the corresponding experimental data points (shown in blue). Notice that x- and y-axis scales are different among panels.

In general, the specific ontologies are different between the parameters $\alpha > \beta$, $\alpha < \beta$, and $\alpha = \beta$. There is no consensus in these ontologies, although the three cases are represented by binding and ligase general ontologies.

3.3. ***Self-organizing map.***

In this work we use a self-organizing map with input data from in a 7-dimensional space. This can be seen in figure 5 and 6 where two maps were constructed using the before mentioned groups of features respectively. As shown in both of these figures, the lack of a more complex structure like the one shown in figure 4 gives an idea of how the conjunction of both sets of features is required to accomplish such rich structure. Three variables correspond to gene ontology features such as "binding", "transferase" and "hydrolase", which are coded in a binary variable whether the attribute is in the family or not. The other four features are: parameters α and β from expression (1), Shannon's entropy and Kolmogorv complexity.

The map was trained with a lattice of 14x14 units. In figure 4 we can see the final map where we can detect 6 groups of clustered families located in the darker blue areas and these patterns and groups cannot be formed neither using gene ontologies nor the complexity features alone, and in figure 5 we can see in a heatmap each one of different entries of the prototypes, this visualization allows us to see the distribution of the different attributes. In this map all of the 7 attributes were used. From this final figure we can infer that the gene ontology attributes are completely independent from each other.

In figure 4 there are 6 groups of clustered families with different general molecular function annotation between every group and delimited by darker blue areas that are a group of neurons that does not have any family representation. The cluster with blue circles have the general molecular function annotation binding. It is the biggest cluster with 93 different families. The molecular function annotation more specific for these families are: "protein binding", "DNA binding", "metal ion binding", "ATP binding", "RNA binding", "nucleic acid binding", "thiamine pyrophosphate binding", "GTP binding", "nucleotide binding", "NAD binding", "calcium ion binding", "heme binding", "FMN binding", "coenzime binding", "flavin adenine dinucleotide binding", "phosphatidylinositol binding", "pyridoxal phosphate binding", "chromatin binding", "GTPase binding", "ubiquitin binding", "iron-sulfur cluster binding", "translation binding" and "histone binding". There are 16 different families from this group that had others molecular function like ”oxidoreductase”, ”catalytic activity”, ”copper chaperone activity”, ”translation initiation factor activity” and ”protein dimerization activity”. The cluster with purple ”X” have the general molecular function annotation of ”transferase” and have 26 different families. The molecular function annotation more specific for these families are: ”phosphotransferase activity”, ”prenyltransferase activity”, ”methyltransferase activity”, and others. The red pentagon cluster have the general molecular function annotation of ”hidrolase” and have 22 different families. The molecular function annotation more specific for these families are: ”endonuclease activity”, ”thiol-dependent ubiquitinyl hydrolase activity”, ”deubiquitinase activity”, and others.

Two clustered families have different molecular functions but join two different groups of clustered families. The yellow plus sign (+) has 8 different families and the molecular function of ”binding” and ”hydrolase” and the dark yellow downward pointing triangles cluster have 7 different families and the molecular function of ”binding” and ”transferase”. These clusters have only these two functions delimited by less dark blue areas and link two big families groups, ”binding” and ”transferase”, and the other families groups are ”binding” and ”hydrolase”.

Finally, the green square cluster has 49 different families with the rest of the general ontologies that do no has any particular grouping. The ontologies are ”oxidoreductase”, ”isomerase”, ”ligase”, ”lyase”, ”translocase”, ”transmembrane transporter”, ”structural constituent”, ”catalytic activity”, ”antioxidant activity”, ”structural molecule activity”, ”translation initiation factor activity”, ”motor activity”, and ”electron transfer activity”.

The information in each cluster is different from each one, however, the information in every neuron is very important to clustering one or more families and it has to be direct with the general ontologies and complexity features like DGBD, the Shannon entropy, and the Kolmogorov complexity. In the future, we hope we can predict the family ontology with this kind of result, however, we know that we have to obtain more clear results with this methodology.

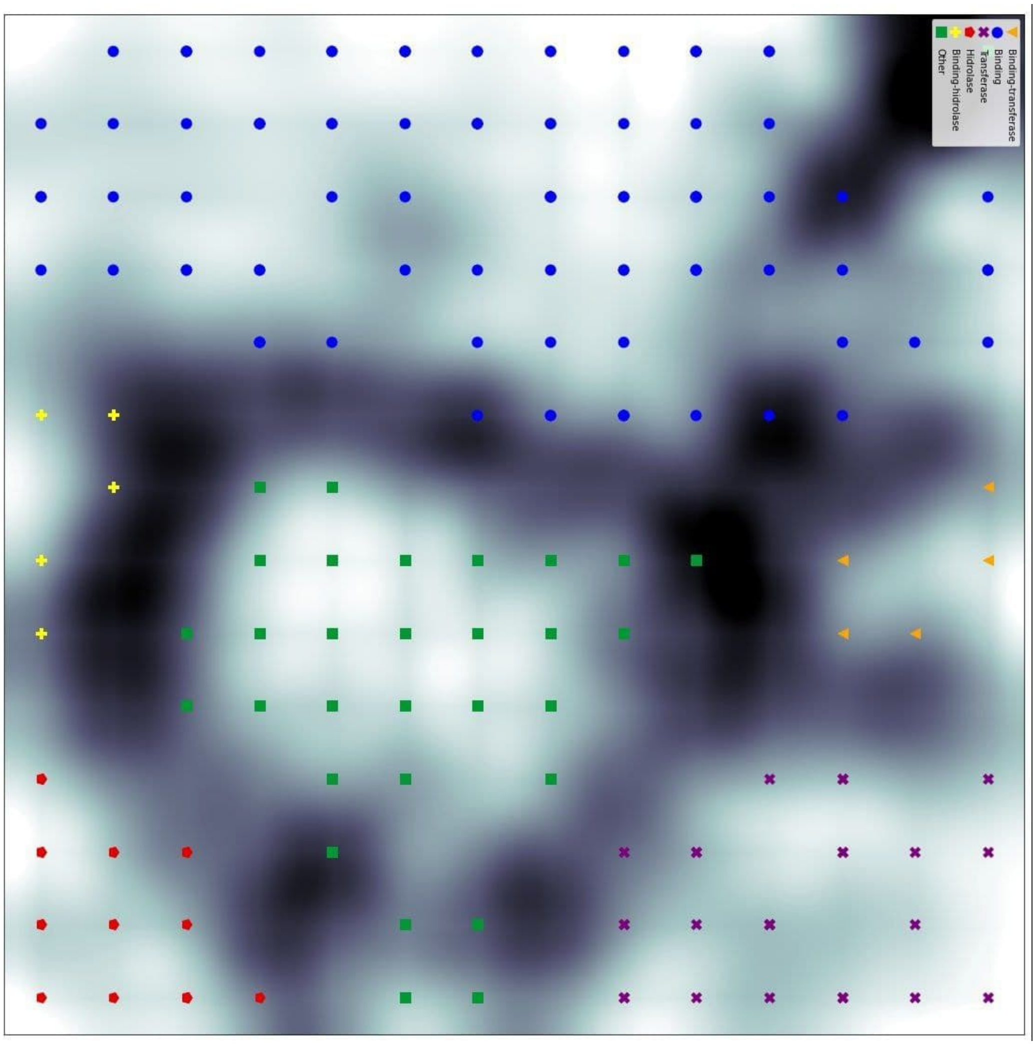

Figure 4: Self organized map of the whole familiome with all of the selected features. This map shows the clustered families projected into a 2D space preserving the topology of the data in the original dimension.

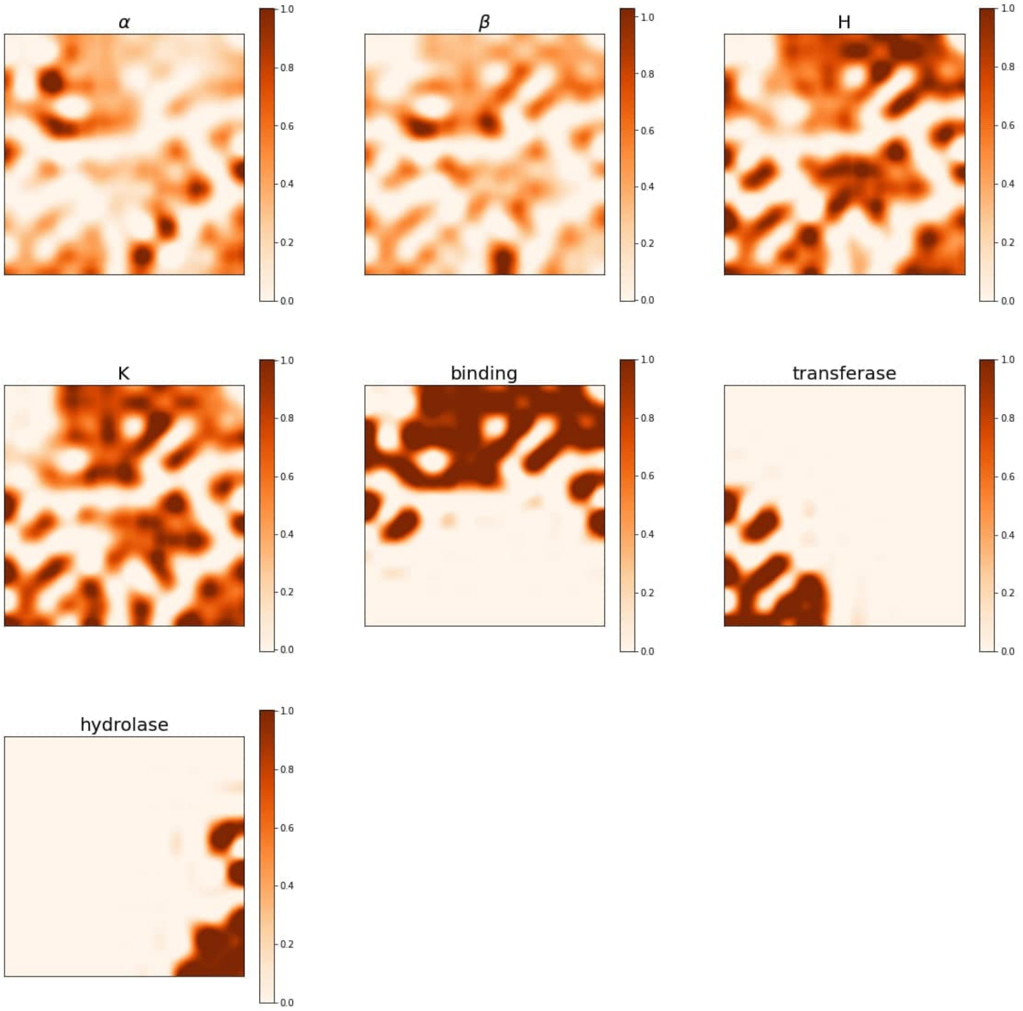

Figure 5: Components of SOM in . This figure shows the distribution of the features used to create the SOM which are α is alpha, β is beta, H is Shannon entropy and K is Kolmogorov complexity, the last three are obtained from the gene ontology functions associated with each family.

Figure 5 shows the distribution of stimuli intensities for each stimulus type remaining after the SOM algorithm reduced dimensions. (¿Qué piensa Checo de esta frase?)The separate stimuli are respectively: DGBD parameter α, DGBD parameter β, Shannon entropy (H), Kolmogorov complexity (K), "binding" ontology, "transferase" ontology and "hydrolase" ontology. Notice how the maps obtained from only one of α, β, H, or K share all their blobs and only differ in their intensity, whereas maps obtained from either "binding", or "transferase", or "hydrolase" ontology information, each contain only a different subset of the blobs. We interpret this pattern to indicate that α, β, H and K all contain a superset of the information contained in the GO General Molecular Functions.

The other four features are: parameters α and β from expression (1), Shannon's entropy and Kolmogorv complexity.

4. Conclusions.

5. References.