UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Positional Amino Acid Frequency Patterns for Automatic Protein Annotation

Mestrado em Bioinformática e Biologia Computacional
Bioinformática

Andreia C. P. Silva

Dissertação orientada por:
Professor Doutor André Osório Falcão

2015

Resumo

Actualmente, com o avanço de tecnologias *high throughput,* como *next generation sequencing,* informação é gerada diariamente. Organizá-la e analisá-la de forma a poder recolher informação útil para aplicações médicas e farmacêuticas continua a ser um desafio. Para este fim é vital compreender tanto para genes como para proteínas características como: *i)* as suas funções bioquímicas, *ii)* a localização celular, *iii)* a extensão de participação nos processos celulares, *iv)* as interacções com outros genes e proteínas e também *v) a* estrutura, uma vez que esta se encontra intimamente relacionada com a função. A abordagem bioinformática continua a ser a única viável para esta situação, uma vez que experimentalmente, é impossível validar toda esta informação de forma expedita.

Entre as mais variadas áreas de trabalho da bioinformática, a anotação de função de proteínas contínua a ser essencial para compreender e definir o papel das proteínas. Neste trabalho apresenta-se uma nova abordagem à anotação automática de proteinas através da análise da sequência. Para tal propõe-se um modelo de *feature learning*.

Numa primeira fase, é executado um Position Specific Iterated BLAST - PSI-BLAST. Este algoritmo permite construir um perfil de probabilidades de cada aminoácido se encontrar numa dada posição da proteína, designado por *position specefic scoring matrix* (PSSM). É construído através de uma compilação iterativa de proteínas com baixa identidade de sequência, isto é homólogos distantes, permitindo assim perceber de uma forma mais clara a relevância efectiva de cada aminoácido para a função da proteína. Neste processo, o PSSM gerado em cada iteração é utilizado para apurar a busca na base de dados para a próxima iteração. O processo contínua até que mais nenhum homólogo distante seja acrescentado ao perfil.

Seguidamente, estes PSSMs, ou seja, as probabilidades de cada aminoácido numa dada proteína, são analisados com um algoritmo de *clustering*, k-means, que tem por objectivo particionar n observações por k grupos predefinidos, onde cada observação é associada ao grupo mais próximo pela distância euclidiana; encontrando, desta forma, padrões de probabilidades idênticos.

Posteriormente, é utilizado um terceiro algoritmo de exploração de regras de associação (*association rule mining*), com o objectivo de encontrar associações entre os *clusters* que representam os padrões probabilísticos de aminoácidos de cada posição e os termos do *Gene Ontology Consortium* (GOC). O último trata-se de um dicionário de expressões para descrever função de proteína, organizado em tês categorias: função molecular, processo biológico e componente celular a que pertence a proteína.

Desenvolveu-se uma fase de prova de conceito com um menor número de proteínas, não só para verificar se o método proposto era viável para um modelo de aprendizagem não supervisionada, mas também para resolver problemas e limitações que o método podesse apresentar, bem como para estabelecer os parâmetros a serem utilizados em cada um dos algoritmos na fase de treino do modelo. Nesta fase, verificou-se que alguns dos padrões de frequência dos aminoácidos em determinadas posições (PAFPs) eram irrelevantes para o modelo, uma vez que analisando os resultado dos PSI-BLASTs, se verificava que a probabilidade para qualquer um dos 20 aminoácidos era igual ou muito idêntica naquela posição em especifíco. Assim, estabeleceram-se vários limiares (entre 500 e 100) para o somatório dos valores de cada PAFP, a que este tinha que ser superior para ser considerado.

Uma vez completa esta fase e verificando-se que o método tinha potencial para ser usado como modelo de aprendizagem, sequências de proteínas com termos GO experimentalmente anotados foram obtidas da Swiss-Prot e o método acima descrito foi aplicado sobre cada uma. O k-means, com o algoritmo de inicialização de Forgy e k igual a 65, foi iterado sobre os 5 limiares assim como o arules, com variados parâmetros de suporte e confiança. Verificou-se que o limiar de 500 seria demasiado exigente, eliminando demasiados padrões que poderiam ser de interesse, diminuindo o campo de aprendizagem do modelo, ao passo que tanto o limiar de 100 como o de 200 eram demasiado laxos, incluindo demasiados padrões irrelevantes e toldando o modelo, pois desviaria o valor dos centróides determinado pelo k-means.

O modelo gerou para o limiar de 300, o maior número de regras com uma confiança de 40% e com um suporte equivalente a cerca de 30 proteínas, tendo identificado 280 termos GO para essas regras. O nível destes termos GO varia entre 1 e 10, sendo por isso termos de alto nível e com baixo conteúdo de informção; cerca de 516591 proteínas, da versão da Swiss-Prot de Julho de 2015, continham estes termos na sua anotação.

Para validar o modelo, 2591 sequências de proteínas com anotações experimetais foram obtidas, também da versão de Julho de 2015 da Swiss-Prot, versão seguinte à de onde foram retiradas as sequências de proteínas para treinar o modelo. Destas, tendo em conta os 280 termos GO acima referidos, o modelo foi capaz de inferir termos GO a todas. Verifica-se com frequência que o modelo atribuí mais proteínas a um termo GO do que as que estão originalmente anotadas com este termo. Contudo, não há informação disponível para perceber se de facto as proteínas não desempenham aquela função ou se simplesmente desempenham mas ainda não foi experimentalmente validado. Por outro lado, o modelo falha em atribuir algumas proteínas que contêm o GO na anotação. Novamente é dificil perceber se estes resultados são uma falha do modelo ou se o termo se encontra incorrectamente atribuído a uma proteína. Uma vez que o modelo foi baseado numa base de dados que, apesar de ser considerada a referência contém algumas inconsitências, é compreensível que o modelo seja fragilizado pelas últimas. Também deve ser notado que para o parâmentro de suporte utilizado para o algoritmo de *association rule learning* e para o limiar seleccionado, existem 7271 termos GO suportados por pelo menos 30 proteínas, assim seria expectável que o número de termos GO identificado fosse mais perto deste valor. Contudo, alguns dos termos identificados conferem, de facto, poder de inferência ao modelo. Estes, curiosamente, são termos GO com baixa representatividade na Swiss-Prot, usualmente os mais difíceis de identificar por modelos de inferência.

De qualquer forma, várias melhorias ao modelo são viáveis. Actualmente, as proteínas seleccionadas para o modelo necessitariam apenas que um termo GO da sua anotação fosse experimental para que a proteína fosse incluída no treino do modelo, contudo em muitos casos nem todos os termos GO tinham sido anotados com evidência experimental pelo que seria interessante considerar este pormenor, talvez com um métrica que penalizasse estes termos face aos que foram anotados com evidência experiemental.

Também, neste trabalho não se separou os termos GO de acordo com as categorias definidas pelo GOC, seria interessante verificar se ao realizar esta separação e correndo o k-means tendo em conta estes grupo,s se seria possível inferências categorizadas. Outra melhoria possível, e necessária, é inclusão um maior número de proteínas para treino do modelo.

Palavras-chave: anotação automática de proteínas; PSI-BLAST; k-means clustering; association rule learning; Gene Ontology

Abstract

Today most proteins contained in protein data bases have been annotated through electronic inference. Due to the amount of data being generated by high throughput methods, electronic inference remains the only viable path to understand proteins' biochemical function(s), cellular location(s), participation in cellular processes, as well as, its structure and interactions.

The feature learning model here proposed aims to introduce a new perspective on protein function annotation problem at a positional amino acid level. Initially, the probabilistic scores for each amino acid at each protein position is acquired, via a traditional PSI-BLAST search; this generates a PSSM with said information. Each protein's positional amino acid frequency pattern (PAFP) is sieved through a threshold to decrease the number of PAFPs irrelevant to the protein's function. Afterwards, these are clustered to their Euclidean closer relatives, via k-means algorithm; identifying, in this manner, s sort of fingerprint of amino acid score patterns. These are then associated to Gene Ontology terms retrieved for the training proteins, using arules package from R, i. e., establish association rules between the resulting k-means clusters of PAFPs and the GO terms.

The 300 threshold for the sum of PAFPs generated 280 GO terms, with a support of 0.0005, about 30 proteins, and a confidence of 40%. These terms were used to describe 516591 proteins out of 549008 in Swiss-Prot the release of July 2015. Most GO terms were, not leaf level, but higher. The model infers far more proteins to each GO term than the ones annotated to it, however it also fails to allocate proteins annotated with the GO term, resulting in high recall levels, but not equivalently high precision. However, note that these results do not mean the inference is incorrect but in fact that there is no evidence to support it one way or the other. Also, in the training set there are 7271 GO terms with a support of at least 30 proteins, it would be expectable for the model to return a similar number of identified GO terms. Despite, falling short of what was expected, the results strongly suggest that the existence of certain PAFPs within proteins may be important for their function. It is also interesting that the strongest signal was found on terms for which

the positive ratio is very low, which are typically very difficult classification problems. Results strongly suggest that it may be possible to find annotation clues by looking on amino acids substitution patterns alone. The results however were not perfect and more work will certainly be required to further validate the initial findings.

Keywords: Automatic protein annotation; PSI-BLAST; k-means clustering; association rule learning; Gene Ontology.

Index

List of Figures

List of Tables

1 Introduction

Cell and Molecular Biology aim to understand and define cellular roles for all proteins encoded in the Genome. This means that for each protein one must describe, among other characteristics, its biochemical function(s), cellular location(s), participation in cellular processes, structure and interactions. Previously, these characteristics were experimentally determined.

However, with the development of high-throughput technologies, able to sequence a whole genome or analyze thousands of genes/proteins simultaneously, the amount of information being produced is overwhelming. Public databases are constantly being updated and currently include over 7000 completely sequenced genomes of cellular organisms [1] contributing to more than fifty million unique protein sequences [2]. While this data holds enormous potential for biological and medical discovery, its amount, breath and complexity makes it extremely challenging to organize, store and analyze. Moreover, due to time and financial constraints it is only possible to experimentally validate/characterize a small fraction, rendering the computational approach vital. A clear representation of these situation is the fact that, UniProt/Swiss-Prot, a gold standard database for protein annotation, only has approximately 10% of its proteins annotated backed by experimental information; the remaining 90% have been annotated based on electronical methods.

Several organizational strategies, analysis and interpretation algorithms, in the fields of pattern recognition, machine learning and visualization, were developed by the bioinformatics community, so that the said data can be accurately characterized. Today, sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein function and structure prediction, prediction of gene expression, protein–protein interactions, genome-wide association studies and evolution modeling are some of the most rapidly progressing bioinformatics research areas.

In an attempt to store all the data being generated, a huge number of databases were created, covering almost everything from DNA and protein sequences, molecular structures to phenotypes and biodiversity. Each database used its own vocabulary to store information. Therefore, when the annotation of four small genomes was examined [3,4] in an attempt to estimate genome annotation error, the outcome was that at least 8% of the molecular function annotations were incorrect. Others, went further and suggest that depending on the definition of function used, misannotation level could be as high as 37%.

To circumvent this problem, several classification systems were proposed to standardize annotation and to facilitate computation. Seldom have these classification systems taken the structure of hierarchical ontologies. Enzyme Commission (EC) numbers [5] and Munich Information Center for Protein Sequences (MIPS) functional catalog [6] are two well accepted schemes; however the most commonly used functional classification is the Gene Ontology (GO).

The Gene Ontology Consortium (GOC) is a collaborative effort to address the need for consistent descriptions of gene products across data bases in a species-independent manner [7]. With time GOC has grown to include several databases, providing an extensive classification of functions, based on a dictionary of well-defined terms divided into three main categories: *i)* molecular function (MF), *ii)* biological process (BP) and *iii)* cellular component (CC) [8]. This way, biological databases are unified by sharing the same vocabulary. Allowing researchers to query any database with a gene/protein name or accession number and retrieve associated Gene Ontology (GO) terms or annotations based on computational or experimental evidence.

On one hand, it is undeniable that these efforts to unify the vocabulary were vital to standardize computation. On the other, they do not suffice. More recent studies that modeled annotation error based on Gene Ontology database, estimated that up to 49% of the computationally annotated sequences could be misannotated. Furthermore, several models of error propagation have shown that with sufficient initial error in the databases, error propagation can significantly degrade the quality of the annotations. This misannotation is not inherent to the vocabulary used; instead it is mainly because the

computational methods are based on previous annotations. And, while annotations remain an issue, the models for automatic annotation will remain, as well.

1.1 Rationale

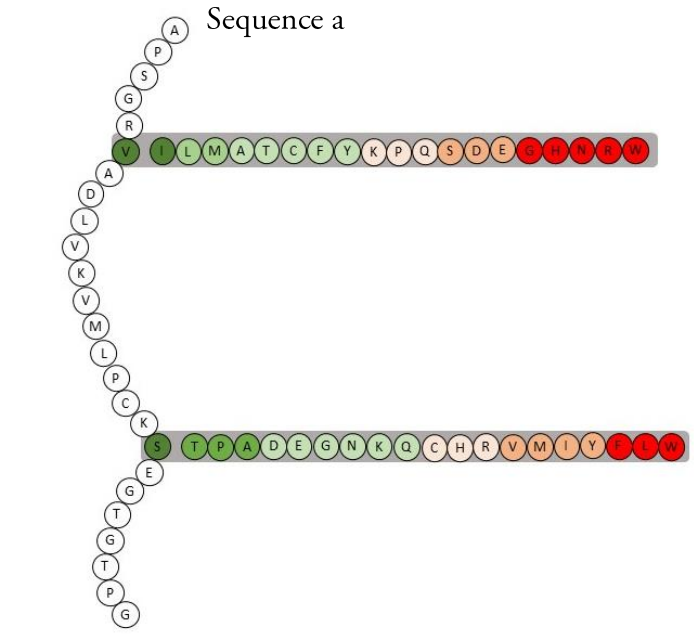
The rationale for this project is to understand whether a given positional amino acid frequency pattern, perpetuated throughout various distantly related proteins at the same or equivalent position, may enable protein function prediction.

Today it is known that an amino acid at a given protein position, in itself, is not enough to learn its role in the said protein's function. This is the traditional idea behind a multiple alignment (MA), where several sequences are aligned in order to identify conserved amino acids or motifs. When using MA tools, typically the result is a conserved region description for which the likelihood of finding each amino acid is present. Yet one fundamental property of MA is that each conserved protein region is found and defined as block. The actual contribution of each amino acid is lost in the context. The specific role of one amino acid is most of the times irrelevant if the amino acid is considered by itself. However, if several similar proteins share a similar substitution pattern for that specific position, this may imply that this conserved position may have in itself a specific biological meaning, without which certain biological functions or behaviors may not occur. This type of observation may be difficult to observe in MA, again due to the fact that only contiguous regions of amino acids are detected. Furthermore, the discovery of individual substitution patterns may prove to be ubiquitous even if the compared proteins are totally dissimilar allowing the identification of functional relationships between them.

Also, with PSI-BLAST, the MA issue in aligning distantly related proteins, becomes obsolete, because the latter beyond identifying distant relatives, also produces a quantitative profile of the probability of each amino acids at each position of the protein is compiled from distant relatives. These profiles will henceforth, for the sake of simplicity, be addressed a positional amino acid frequency patterns - PAFPs. In this manner, the amino acid patterns at certain positions, truly relevant to said proteins' function, are laid bare.

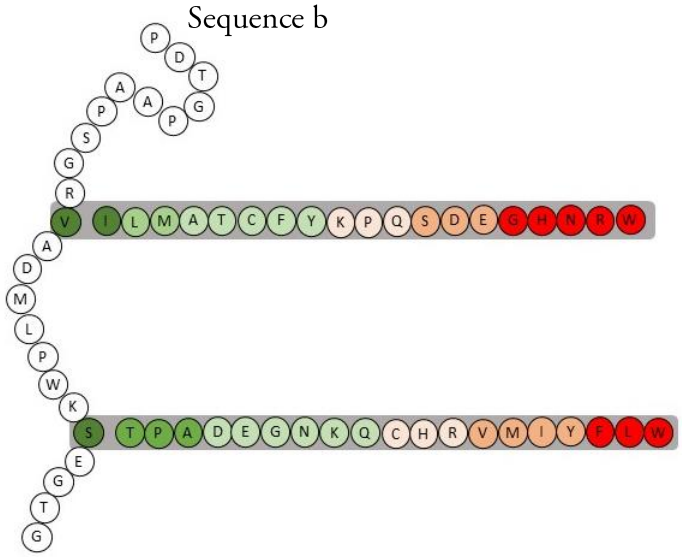
What is here proposed, is a feature learning method to identify recurrent PAFPs by analyzing the PSSMs of various experimentally annotated proteins, generating a sort of probabilistic score fingerprint for protein function, by clustering PAFPs and by using a machine learning/statistical method (association rule learning) to associate these fingerprints to protein function.

This idea is innovative and no significant studies have been found in the literature. Nonetheless, the use of PSI-BLAST and PSSMs have been used in automatic protein functional and structural annotation. Previously, it has been demonstrated the ability to use PSSMs to predict protein molecular function [9]. It is noteworthy, in this case, that the PSSMs were generated based on structural alignments and obtained for each potential molecular function GO term present among the initial protein structures. Here, all three categories established by GOC are considered and no structure is taken into account. Afterwards, a few improvements and, consequently, a widening of the predicting scope is obtained by using GO to direct the function prediction process, by splitting sets of sequences identified by PSI-BLAST into sub-alignments according to GO annotations [10]. Each GO term sub-alignment is then used to identify conserved residues within group, for which a PSSM profile is generated. This combination of steps enables the identification of conserved residues potentially associated with a particular function and produces a set of feature derived profiles from which protein function is predicted. Despite, the different step, the model here proposed aims for a similar outcome.



Function 1 - cerebrospinal fluid carrier of the thyroid hormone thyroxine

Funtion 2 - retinol-binding



Would it be possible for sequence b and c to have the same function since they have the same PAFPs?

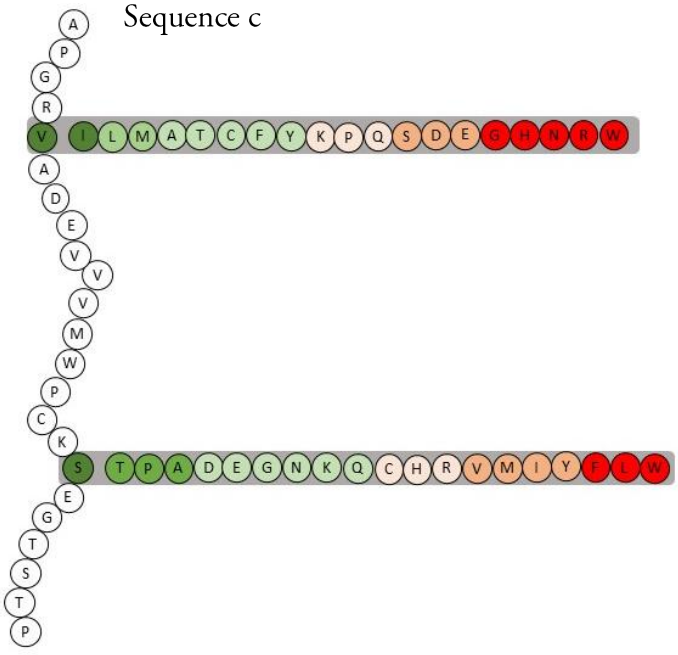


Figure 1 - Representation of the same PAFPs (grey box) in three distinct proteins, according to two randomly selected PAFPs from Table 2. The colour of the amino acids in the PAFPs, represents the likelihood of the amino acid in said position; ranging between dark green – very likely to be found at that position, bright red very unlikely to be found at that position. The project here proposed, aims to group these PAFPs and macth them to a putative protein function.

2 Methods for in silico automatic protein annotation

Accurately understand and annotate "anything that happens to or through a protein" [11] is key to understand life at molecular level. Maybe because this is quite a complex concept, presently there is no unified computational answer that is able to accurately characterize any given protein.

Protein function annotation can be predicted by several methods. Most traditional approach consist on identifying whole sequences or small pieces of these among proteins with experimentally determined function, inferring annotation from closest relatives. However, other strategies have also been developed.

2.1 Sequence alignment based methods

Proteins with similar sequence are often homologous [12] and, therefore, are estimated to have similar function. Consequently, proteins from a newly sequenced genome are commonly annotated by transference, i.e., based on similar protein sequences from previously annotated genomes, independent if annotation is experimentally or electronically inferred.

Several studies indicate that at least 60% sequence identity, and more likely closer to 80%, is required for accurate transfer of the third level of EC classification [11–16]. However, many cases of closely related proteins that do not share the same function have been reported [17]. Currently, there is no sequence-similarity threshold that guaranties function similarity. Furthermore, it is estimated that below 30% protein sequence identity, commonly called twilight zone, detection of a homologous relationship is not guaranteed by sequence alone [18].

Overall, the type of alignment applied and its outcome strongly depend on the characteristics of the query protein and on the characteristics of the ones in the query database. Notwithstanding, the use of BLAST (Basic Local Alignment Search Tool) [19]

and PSI-BLAST (position-specific iterated BLAST) [20] remains a common first step for inferring protein function through alignment of sequences. Thus, for a newly identified uncharacterized protein, without any further knowledge of similarity percentage to proteins in the database, this methods may not suffice.

However, because understanding protein alignment is vital to understanding the rationale for this project these methods are here described in further detail.

2.1.1 Global Alignment by Needleman-Wunsch

This alignment is better suited for closely related sequences which are of same length. Here, the alignment is carried out from beginning till end of the sequence to find out the best possible alignment by using optimal alignments of smaller subsequences.
This algorithm is divided into three separate steps:

i) Initialization of the matrix with the scores possible;

ii) Matrix filling with maximum scores;

iii) Trace back the residues for appropriate alignment.

The initialization step consist on building a matrix with the two sequences being aligned, where each of the amino acids of one sequence corresponds to the columns and the amino acids of the remaining one, correspond to the rows. The first column and row are filled with the scores for each of the amino acids.

Take the alignment of the two small sequences SEND and AND. The cells of the score matrix are labelled C(i,j), where i and j integer number between 1 and the length of each sequence.

		S	E	N	D
	C(1,1)	C(1,2)	C(1,3)	C(1,4)	C(1,5)
A	C(2,1)	C(2,2)	C(2,3)	C(2,4)	C(2,5)
N	C(3,1)	C(3,2)	C(3,3)	C(3,4)	C(3,5)
D	C(4,1)	C(4,2)	C(4,3)	C(4,4)	C(4,5)

		S	E	N	D
	0	-10	-20	-30	-40
A	-10				
N	-20				
D	-30				

		S	E	N	D
	done	left	left	left	left
A	up				
N	up				
D	up				

Afterwards, this matrix is filled by row starting at cell C(2,2). For any cell C(i,j) the maximum of the following is selected:

i) qdiag = C(i-1, j-1) + S(i, j) *ii)* qup = C(i-1, j) + g *iii)* qleft = C(i, j-1) + g

where S(i, j) is the substitution score for those letters and g is the gap penalty. Once the matrix is complete, the trace back process, which in this case starts at C(4,5) is carried out according to the value and origin of that value.

		S	E	N	D
	0	-10	-20	-30	-40
A	-10	1	-9	-19	-29
N	-20	-9	-1	-3	-13
D	-30	-19	-11	2	3

		S	E	N	D
	done	left	left	left	left
A	up	diag	left	left	left
N	up	diag	diag	diag	left
D	up	up	diag	diag	diag

		S	E	N	D
	done	left	left	left	left
A	up	diag	left	left	left
N	up	diag	diag	diag	left
D	up	up	diag	diag	diag

Sequences are aligned backwards, and according to the values in the backwards path established: diag, represents the letters from the two sequences are aligned; left, means that

a gap must be introduced to the left in the row sequence; up, represents that a gap is introduced in the column sequence. Hence, the result for this small example would be:

SEND

A-ND

The Needleman–Wunsch algorithm is still widely used for optimal global alignment, particularly when the quality of the global alignment is of the utmost importance. However, the algorithm is expensive with respect to time and space, proportional to the product of the length of two sequences making it unsuitable for long sequences. Recent development has focused on improving the time and space cost of the algorithm while maintaining quality. Fast Optimal Global Sequence Alignment Algorithm (FOGSAA), achieves a time gain of 70–90% for highly similar nucleotide sequences (with > 80% similarity), and 54–70% for sequences having 30–80% similarity [21].

2.1.2 Local Alignment by Smith-Waterman

This alignment is better suited for suspected similar sequences or even dissimilar sequences. It finds local regions with high level of similarity through an adaption of the previously described global alignment.

One of the alterations to the Needleman-Wunsch algorithm, occurs at the first stage, where the negative scoring matrix cells are set to zero, which renders the (thus, positively scoring) local alignments visible. The process is the carried out as previously described, except for the tracing back step. This, instead of starting at the last filled position starts at the highest scoring matrix cell and proceeds until a cell with score zero is encountered, yielding the highest scoring local alignment.

It is noteworthy, that the algorithm used for BLAST is an optimization suggestion for a less time-consuming form of the algorithm used for Smith-Waterman. BLAST employs an alignment which finds "local alignments between sequences by finding short matches and from these initial matches (local) alignments are created", as well.

2.1.3 Basic Local Alignment Search Tool

Also using a heuristic method, BLAST finds similar sequences, by locating short matches – words, between the two sequences, instead of comparing sequences in full.

To identify relevant words a window, with user defined size, is slid across the sequence generating words that are then compared with database sequences. These comparisons are scored according to a scoring matrix - a commonly used one, for protein alignment, is BLOSUM62 - BLOcksSUBstitutionMatrix. Only words above the matrix determined threshold are kept. This process is called seeding.

Afterwards building an alignment is then possible. To do so, neighbouring words are also assembled, i. e., the alignment is extended in both directions of the original word in an attempt to extend it. However, each extension impacts the score of the alignment. Should this score be higher than a previously determined threshold, the alignment will be included in the results; should it be lower, the alignment will cease to extend, preventing areas of poor alignment from being included in the BLAST results.

Those alignments whose score is above the empirically determined cut-off S score are called High Scoring Segment Pair (HSP). The S score is determined by examining the distribution of the alignment scores modelled in comparison with the distribution or random alignment. The HSPs' scores of the extended regions are then created by using a *i)* substitution matrix, as before and *ii)* a gap penalty system.

Seldom more than one HSP is found in the same sequence, this may be due to loss of some sequence regions that would have been joined before determined evolutionary process, it is then imperative to consider them all in the alignment. The gap penalty system enables this situation by scoring the insertion and/or removal of gaps.

Once the alignment process is completed for a query and each subject sequence in the database, a report is generated, providing a list of those alignments with a value greater than the cut-off score S.

Despite being used to identify homologous sequences by searching and comparing a query sequence with those in the databases, Smith-Waterman and BLAST are quite different. BLAST is based on a heuristic algorithm, its results, in terms of hits found, may not be the best possible ones, as it will not identify remote homologs. The Smith-Waterman algorithm is a better alternative to identify these homologs. However, this accuracy comes at the expense of time and computer power.

2.1.4 Position Specific Iterated – BLAST (PSI-BLAST)

Database searches using position specific scoring matrices are often much better at detecting weak relationships between proteins than database searches that use a simple sequence as query, therefore PSI-BLAST is substantially more sensitive than the corresponding BLAST program.

PSI-BLAST's first step is to create a list of all closely related proteins using a standard BLAST. These proteins are, then, combined into a general "profile" sequence – a position specific score matrix (PSSM), which summarises significant features present in these sequences. This matrix is the length of the query protein * 20 matrix.

Analogously to BLAST, a query against the protein database is then run using this profile, instead of the 20*20 substitution matrix. A larger group of proteins is found and aligned to the query sequence.

In order to transform this alignment into another profile, several data manipulation stages take place: *i)* Only one row (alignment pair query-database sequence) above 98% identity is kept; *ii)* gaps are dismissed, meaning all sequences are the same length and *iii)* each alignment is attributed a weight, i. e, because a large set of closely related sequences carries about as much information as a single sequence, but due to its size it may easily "outvote" a small number of more divergent sequences, different weights are then assigned to the various sequences, with those having many close relatives receiving smaller weight [20]. This weighting process is based on a modified version of the Henikoff-Henikoff method [22]. Initially, for each column, a total weight of 1 is divided evenly among the

letter types that occur at that position, then the weight assigned to each letter type is divided evenly among the sequences that have that letter, afterwards for each sequence the weights from all sequences a summed.

Table 1 – Layman's summary of the Henikoff-Henikoff weighting method. Adapted from[22].

Sequence	calculation		results
G C **G** T **T** A **G** C	**1/4** + 1/3 + **1/3** + 1/4 + **1/4** + 1/3 + **1/4** + 1/2	2 1/2	0.31250
G A **G** T **T** G **G** A	**1/4** + 1/3 + **1/3** + 1/4 + **1/4** + 1/3 + **1/4** + 1/4	2 1/4	0.28125
C G **G** A **C** T **AA**	**1/2** + 1/3 + **1/3** + 1/2 + **1/2** + 1/3 + **1/2** + 1/4	3 1/4	0.40625

Posteriorly, to estimate the probability of a residue to be found at that column a data dependent pseudocount method, introduced by Tatusov [23], is applied. This method uses prior knowledge of the aa relationships embodied in the substitution matrix to generate residue pseudocount frequencies which are averaged with the observed frequencies.

This process is repeated until no new sequences are added to the alignment, hence no new information is added to the resulting PSSM.

Query seq	aa position	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
G	1	0	-3	-1	-2	-3	6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-3	-3
P	2	0	-2	-1	-1	-3	-2	-2	-3	-1	-3	-2	-1	6	-1	-2	2	1	-2	-3	-2
T	3	0	-2	-1	-1	-2	-2	6	-1	-1	-2	-1	0	-2	0	-1	0	2	0	-3	0
G	4	0	-3	-1	0	-3	5	-2	-4	-1	-4	-3	0	-2	-1	-2	0	-2	-3	-3	-3
T	5	2	-1	-1	1	-2	0	-1	-1	0	-2	-1	0	-1	0	-1	2	2	0	-2	-2
G	6	1	-2	1	2	-2	3	-1	-1	-1	-2	-1	-1	-1	0	-1	0	-1	-1	-2	-2
E	7	-1	-3	4	3	-3	-2	2	-3	0	-3	-2	1	-1	2	-1	0	-1	-3	-3	-2
S	8	1	-1	0	0	-3	0	-1	-2	0	-3	-2	0	1	0	-1	4	1	-2	-3	-2
K	9	-1	-3	-1	1	-3	-2	-1	-3	5	-3	-1	0	-1	1	2	0	-1	-2	-3	-2
C	10	-1	9	-4	-4	-3	-3	-3	-1	-3	-1	-2	-3	-3	-3	-4	-1	-1	-1	-2	-3
P	11	-1	-3	-2	-1	-4	-2	-2	-3	-1	-3	-3	-2	7	-1	-2	-1	-1	-2	-4	-3
L	12	-2	-1	-4	-3	0	-4	-3	1	-3	4	2	-3	-3	-2	-2	-3	-1	1	-2	-1
M	13	-1	-2	-3	-2	0	-3	-2	1	-1	2	6	-2	-3	-1	-2	-2	-1	1	-2	-1
V	14	0	-1	-3	-3	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1
K	15	-1	-3	-1	1	-3	-2	-1	-3	5	-3	-1	0	-1	1	2	0	-1	-2	-3	-2
V	16	0	-1	-3	-3	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1
L	17	-2	-1	-4	-3	0	-4	-3	1	-3	4	2	-3	-3	-2	-2	-3	-1	1	-2	-1
D	18	-2	-4	6	1	-4	-1	-1	-3	-1	-4	-3	1	-2	0	-2	0	-1	-3	-4	-3
A	19	4	-1	-2	-1	-2	0	-2	-1	-1	-2	-1	-1	-1	-1	-1	1	0	0	-3	-2
V	20	0	-1	-3	-3	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1
R	21	-1	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	2	6	-1	-1	-3	-3	-2
G	22	0	-3	-1	-2	-3	6	-2	-4	-2	-4	-3	-1	-2	-2	-2	0	-2	-3	-3	-3
S	23	0	-1	-1	0	-2	-1	-1	0	0	-2	-1	0	-1	0	2	3	1	-1	-3	-2
P	24	-1	-3	-2	-1	-4	-2	-2	-3	-1	-3	-3	-2	7	-1	-2	-1	-1	-2	-4	-3
A	25	4	0	-2	-1	-2	0	-2	-1	-1	-2	-1	-2	-1	-1	-2	1	0	0	-3	-2

Figure 2 – Example of an PSSM output for a peptide of homotetrameric plasma protein - transthyretin, whose structure was defined by x-ray crystallography [24].

2.2 Sequence motif-based methods

Evidence for function can also be inferred from known protein domains by matching a query sequence to a protein domain database, like Pfam - Protein Families Database. Other databases, such as dcGO, go further and include annotations referring to individual domains (evolutionary units that comprise proteins) and supra-domains (domain combinations that recur in different protein contexts with different partner domains) [25].

Aspects of a protein's function can be predicted without comparison to other full-length homologous protein sequences. Within protein domains there are shorter signatures, known as motifs, associated with particular functions [26]. These can be associated to function by querying motif databases such as PROSITE [27].

2.3 Structure-based methods

Structural similarity is thought to be a good indicator of function similarity since it is generally more well conserved than its sequence [17,26]. Many programs have been developed to screen an unknown protein structure against the Protein Data Bank [28] (PDB) and report similar structures. FATCAT (Flexible structure AlignmenT by Chaining AFPs (Aligned Fragment Pairs) with Twists) [29] CE (combinatorial extension) [30]) and DeepAlign (protein structure alignment beyond spatial proximity) [31] are some of the programs created towards the referred ending.

2.4 Homology modelling

However, due to the fact that many protein sequences have no x-ray crystallography or NMR solved structures, some function prediction servers, such as RaptorX, have also been developed to firstly predict the 3D model of a sequence and afterwards use a structure-based method to predict functions based on the predicted 3D model. In many cases instead of the whole protein structure, the 3D structure of a particular motif representing an active site or binding site can be targeted [26]. Databases such as Catalytic Site Atlas [32] have been developed that can be searched using novel protein sequences to predict specific functional sites.

In all instances, some prior knowledge of sequence or structural similarity is essential for any inference. Other methods escaping the prior paradigms have also been developed.

2.5 Genomic context-based methods

Recent methods for protein function prediction are not based on comparison of sequence or structure as the previous, instead they are based on some type of correlation between novel genes/proteins and those that already have annotations. This is known as phylogenetic profiling and is based on the observation that two or more proteins with the

same pattern of presence or absence in many different genomes most likely have a functional link [26,33]. Whereas homology-based methods are more often used to identify molecular functions of a protein, genomic context-based approaches are used to predict cellular function or the biological process in which a protein acts [33,34]. For example, proteins involved in the same signal transduction pathway are likely to share a genomic context across all species.

2.5.1 Gene fusion

Gene fusion occurs when two or more genes encode two or more proteins in one organism and have, through evolution, combined to become a single gene in another organism (or inverse for gene fission) [34,35]. This can occur as a result of: translocation, interstitial deletion, or chromosomal inversion.

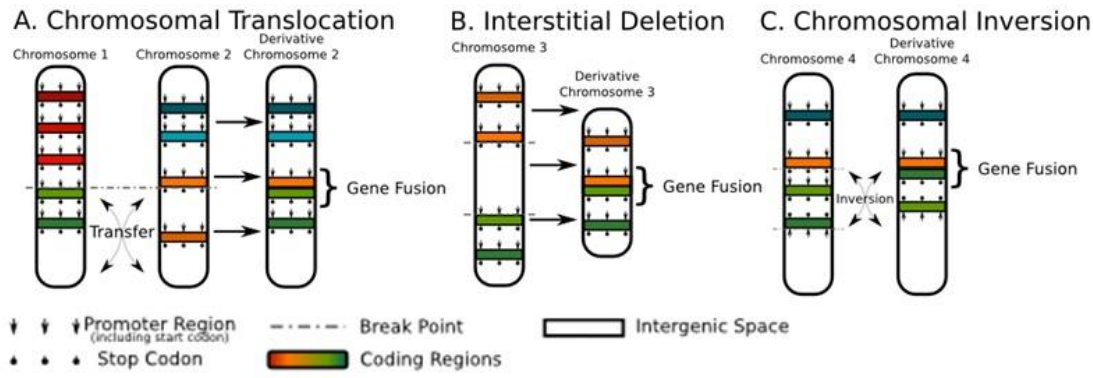


Figure 3 – Graphical representation of the various processes giving way to gene fusion. Adapted from [56].

Gene Fusion has been used to search all E. coli protein sequences for homology in other genomes. Over 6000 pairs of these protein sequences shared homology to proteins in other genomes, indicating the potential interaction between each of the pairs [35]. The latter would not have been predicted through homology-based methods, because the two sequences in each protein pair are non-homologous.

2.5.2 Co-location/Co-expression

In prokaryotes, clusters of genes that are physically close together in the genome are often conserved together through evolution, and tend to encode proteins that interact or are part of the same operon [34]. Thus, chromosomal proximity, also known as the gene neighbor method [36], can be used to predict functional similarity between proteins in prokaryotes.

Genes involved in similar functions are also often co-transcribed, so that an unannotated protein can often be predicted to have a related function to proteins with which it co-expresses [26]. The guilt by association algorithms developed based on this approach can be used to analyze large amounts of sequence data and identify genes with expression patterns similar to those of known genes [37,38]. In other words, a group of candidate genes, with an unknown function, are compared to a target group - genes known to be associated with a particular disease; the candidate genes are then ranked by their likelihood of belonging to the target group. Recently, however, some problems with this type of analysis have been reported. For instance: many proteins are multifunctional, thus the genes encoding them may belong to several target groups [39].

2.6 Network-based methods

Guilt by association type algorithms may be used to produce a functional association network for a given target group of genes or proteins. These networks serve as a representation of the evidence for shared/similar function within a group of genes, where nodes represent genes/proteins and are linked to each other by edges representing evidence of shared function [40].

2.6.1 Integrated networks

Several networks based on different data sources can be combined into a composite network, which can then be used by a prediction algorithm to annotate candidate genes or

proteins [41]. Many algorithms have been developed to predict function based on the integration of several data sources (e.g. genomic, proteomic, protein interaction, etc.). Testing on previously annotated genes indicates a high level of accuracy [40,42]. However, some function prediction algorithms are not directly interpretable and many require extremely high computational resources. Faster, more accurate algorithms such as GeneMANIA (Multiple Association Network Integration Algorithm) have been developed in recent years [41] in an attempt to surpass the aforementioned disadvantages.

Table 2 – Summary of the gains and limitations of the in silico methods described in this Chapter.

Method	Advantages	Disadvantages
Sequence aligment based methods	- commonly used - many fast tools available - most mature and reliable [43]	- no sequence-similarity threshold that guarantees function similarity - type of alignment to be applied depends on the characteristics of the query protein and database
Sequence motif-based methods	- Its sequence alignment based but considers several proteins, hence evolution – conserved blocks	- Its sequence alignment based - aligning distant relatives remains an issue - patterns from closely related sequences - amino acid contribution is lost in the context
Structure-based methods	- good indicator of function similarity (more conserved)	- Computationally demanding - Scarcity of crystallographic evidence, undermines confidence
Homology modelling	- considers several proteins, hence evolution – conserved blocks - only models the catalytic conserved block	- sequence or structural similarity is essential for any inference - again scarcity of crystallographic evidence, undermines confidence
Genomic context-based methods	- not based on sequence or structure similarity - used for cellular function or the biological process, instead of molecular function - prediction for non-homologous proteins	- used for cellular function or the biological process, instead of molecular function - genes encoding multifunctional proteins may belong to several target groups
Network-based methods	- high level of accuracy	- difficult to interpret - high computational resources

Overall, the above in silico methods, despite complementing each other in to automate protein annotation, do not seem to be able to do it in full. What is here proposed is an idea that offers a new approach to address this problem at a positional amino acid level, not yet considered by any of the aforementioned methods. Also, it is not expected for this model to fully automate protein annotation but it is hoped that in association with other methods it may in fact allow for it and/or reduce the range of existing possibilities.

3 Methods

Overview

As previously mentioned, the feature learning model here proposed to identify recurrent PAFPs to generate a sort of probabilistic score fingerprint for protein function, consists on:

i) Retrieving, from Swiss-Prot, for each protein annotated with an experimental evidence code, both the sequence and the GO annotation;

ii) with the retrieved sequences, acquiring quantitative information on the relevance of each amino acid at a given position - a typical PSI-BLAST search generates a PSSM, containing such information - PAFPs;

iii) clustering the closest PAFPs by their characteristics;

iv) using rule association to establish relations between the resulting clusters and protein function given by GO terms;

v) validating the resulting rules.

Therefore, it is necessary to understand if it has potential to become an annotation attribution model as well as to resolve issues that may arise throughout its implementation. Thus, a proof of concept, with a small scale data set will be carried out (Chapter 4).

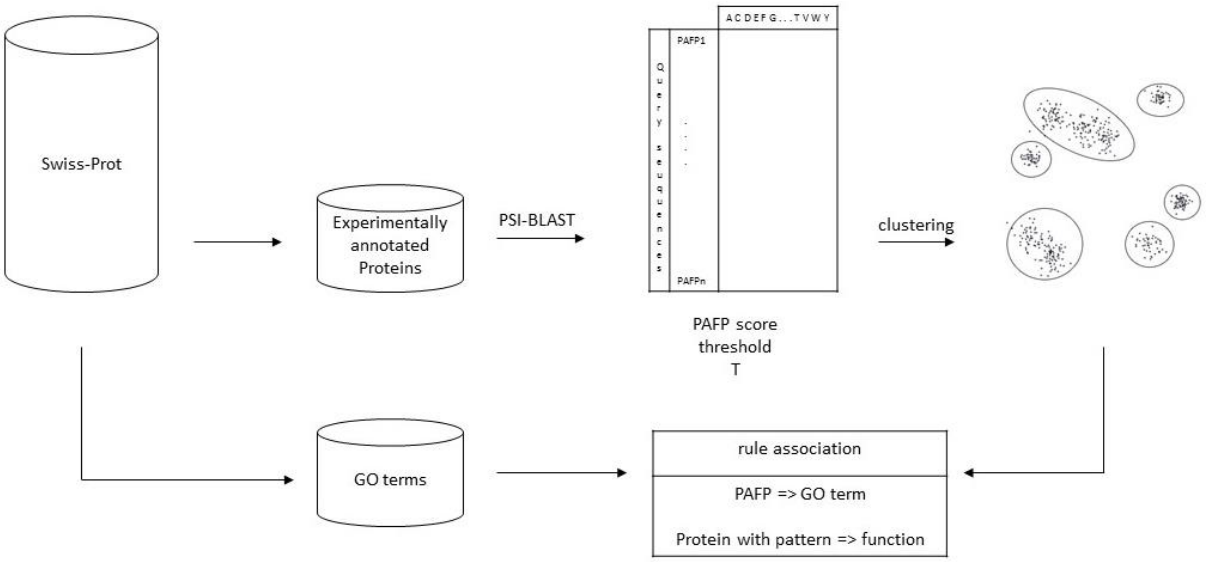


Figure 4 – Schema of the feature learning method here proposed. Initially, proteins with GO terms annotated with experimental evidence codes will be retrieved. PSI-BLAST will be run over sequences and only the PAFPs with sum of probabilities above a determined threshold are kept. These will then be clustered. Afterwards, proteins represented by GO terms and proteins represented by clusters will be, analysed in order to identify association rules among both data sets.

3.1 Data Clustering

In order, to identify PAFPs to train the model, it is necessary to group them by their similarities, since it is not expectable for them to be exactly equal throughout the data, clustering methods provide an approach to do so.

3.1.1 k – means

k-means [44] is one of the simplest flat unsupervised learning algorithms used to solve the clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed *a priori*.

This algorithm was elected mainly due to the nature of the project and also for its ease of implementation. Because, the data patterns being searched are not known and

cannot, at this point, be defined it is clear that an unsupervised learning method is necessary.

k-means, is, in fact, the simplest for grouping instances into clusters based on all variables without any target one. Most of the clever algorithms are much harder to implement efficiently and have a higher number parameters to set; on one other hand, they can be 100x faster. This is not relevant for the data being analysed, k-means is expected to perform reasonably fast. Also, because there is no evidence whether flat or hierarchical clustering methods are better at classifying, a simpler approach is favourable.

k-means' main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way, since different location causes different outcomes. Some implementations are done as to place them far apart as possible, such is the case of MacQueen and Hartigan-Wong [45]; others, like Lloyd-Forgy [46] algorithm, simply select for initial centroid values random points thorough out the data. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. This association is performed by calculating the Euclidean distance between the data point and the centroid. When all points have been associated, the first step, commonly denominated as seeding, is completed and an early groupage is done. At this point, k new centroids are re-calculated as barycentre of the clusters resulting from the previous step. Having these *k* new centroids, a new association between the same data set points and the nearest new centroid occurs. The *k* centroids change their location in each iteration until no more changes occur.

3.2 Association Rules

arules package for R [47] will be used to establish *a priori* association rules between clusters representing amino acid probability patterns and protein known annotations in the form of GO terms.

Mining association rules from transaction data, can be introduced [48], as follows:

i) Let I = {i1, i2. . . i*n*} be a set of n binary attributes called items.

ii) Let D = {t1, t2. . . t*m*} be a set of transactions called the database.

iii) Each transaction in D has an unique transaction ID and contains a subset of the items in I.

iv) A rule is defined as an implication of the form X ⇒ Y where X, Y ⊆ I and X ∩ Y = ∅.

The sets of items X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule.

Take into the account the following example, an over simplification of the data expected for this project.

Prot1	Cluster1, Cluster4, Cluster6, GO1, GO4, GO6
Prot2	Cluster1, GO3, GO5, GO6
Prot3	Cluster2, Cluster4, GO1, GO4, GO6
Prot4	Cluster1, Cluster5, Cluster6, GO3, GO5

The proteins can be considered the transaction identifiers and, both, the clusters and GO terms can be considered as the subset of items; clusters summarizing the amino acid probability patterns are introduced as LHS and the GO terms as RHS.

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on support and confidence. The latter is defined conf(X ⇒ Y) = sup (X ∪ Y)/ sup (X) and can be interpreted as an estimate of the probability P(Y|X), i.e., the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.

Returning to the example, the probability of finding cluster1 in a transaction that also contains the GO3 should 2/3, since cluster1 is present in 3 different transactions but only twice is it accompanied by GO3. This is performed for all clusters and all GO terms, rapidly increasing of generated rules. Therefore, association rules are required to satisfy both a minimum support and a minimum confidence constraint at the same time. The

generated rules should also have a maximum length, including LHS and RHS, depending on the information to be retrieved.

3.3 Model Validation

In order to validate the model, an updated version of Swiss-Prot, will be retrieved. Again, only proteins whose annotation contain any of the experimental evidence codes will be used. Subsequently, the new relevant entries will be run on PSI-BLAST. Afterwards, each protein position will be allocated, by Euclidean distance, to the previously calculated k-means clusters. Posteriorly, they will be matched to the GO terms previously identified in the rules resulting from the previous learned data.

For each of the newly selected proteins all GO terms, including ancestral GO terms, will be retrieved. The latter will be obtained, for each of the terms annotated for each protein, from Gene Ontology Database, using an altered version of the program [49] available at Annex 2. For each GO term it will be determined: *i)* the number of the selected proteins annotated with it - Positives; *ii)* the remaining number of proteins not annotated to it – Negatives; *iii)* the number of proteins inferred to it – Inferred Positives; *iv)* the number of proteins inferred to it, excluding the true positives - False Positives; *iv)* the number of false positives, excluding the inferred ones – True Positives; *v)* different between the positives and the true positives – False Negatives; and vi) the negatives minus de false positives. From these counts, recall (TP/(TP+FN)), precision (TP/IP) and F1 score - 2*(recall * precision/(recall + precision)) metrics are also calculated as well as Matthews correlation coefficient (MCC) for each GO term. The latter is given by

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation.

4 Proof of Concept

Initially, because this project was expected to be extremely strenuous in terms of computer processing, the previously described model was implemented over a small sample of proteins. This will also allow to *i)* understand if the model is indeed viable, *ii)* identify limitations of the method that are, otherwise, unpredictable and adapt to them, if possible at this stage and *iii)* define a set of parameters for the algorithms being used to train the model.

4.1 Data retrieval and processing

For this purpose, a set of 360 protein sequences in FASTA format were downloaded from Swiss-Prot/UniProt, 240 of which belong to the E.C. 1.1.1.1 family and 120 belonging to the E.C. 1.1.1.38 family. PSI-BLAST, from the ncbi-blast-2.2.30+ suit, was executed over each of the sequences using Swiss-Prot (uniprot_sprot.fasta) as Blast database; a cutoff E-value of 0.01 was used to select the PSI-BLAST's training sequences; the process ceased once no more new sequences were added to the training sequences or else until it reached 21 iterations.

The resulting outputs where then processed in order to obtain a PSSM, describing the probability of each position of each protein having a determined amino acid in it. For this purpose the program at Annex 1 was used. Resulting matrix contains log-odds, these were then transformed into exponential probabilities, to better distinguish between them, and summed by position in the protein, i. e., each element in the probability vector for each amino acid was summed.

4.2 k-means

4.2.1 k selection

In order to select the number of centroids with which to start the k-means algorithm several k-means, with Hartigan-Wong implementation (by default), were run, in R, varying the number of centroids between 1 and 200, with a step of 5, until the within groups sum of squares did not vary.

Several thresholds were established, ranging from 100 to 5000 in order to understand which would be an optimal threshold to select relevant PAFPs to be further analyzed. The k-means algorithm was then run with 65 centroids and the Hartigan-Wong implementation. This was all carried out in R.

4.2.2 Verifying cluster coherence

k-means starts, despite the seeding algorithm, by selecting random PAFPSs from the data set as initial centroids. As previously mentioned, depending on the seeding algorithm this might be completely random or pseudo-random, in the sense that the data can be previously compartmentalized and the initial centroids selected from each of these compartments, in a way that guaranties maximum distance among them. Nevertheless, because this process is essentially random, every time k-means is run the initial centroid are different as well the numbering attributed to the clusters, because the latter is done according to the initial centroid selection. To understand if throughout the iterations the PAFPs are assigned coherently to the initial centroids through various k-means runs, despite the randomly selected initial centroids and their numbering, a 100 different k-means were run over the data, with the Hartigan-Wong algorithm. In order to have an intersection control value, the following procedure was also applied to 10 k-means, generated from random values. For each cluster in each pair of k-means an intersection

value was obtained, in an attempt to identify the most similar k-means, generating a matrix of 65 by 65 clusters; in total 4950 of these matrixes. For each of this matrixes an altered Hungarian algorithm was executed in order to obtain the highest intersection value between clusters, i. e., the highest the intersection value the more similar two k-means are, despite the numbering of the clusters.

The Hungarian algorithm, also known as the Kuhn-Munkres algorithm, is a combinatorial optimization algorithm that solves the assignment problem in polynomial time. This algorithm is based on the following theorem: If a number is added to or subtracted from all of the entries of any one row or column of a cost matrix, then an optimal assignment for the resulting cost matrix is also an optimal assignment for the original cost matrix [50,51].

As previously mentioned, for the purpose of this project, it was necessary to obtain maximum intersection values for each input matrix, in analogy maximum cost, instead of minimum cost. Therefore, the numbers in the input matrix were reversed and the same algorithm was run over the data.

4.3 Results

Initially, 360 proteins, 240 from the EC 1.1.1.1 and 120 from EC 1.1.1.38 families, were randomly selected. Moreover, they were also selected at random from a wide range of species within the mentioned families. Hence, providing proteins with very similar sequences within each family but diverse sequences between the two families. This along with inclusion of several species is expected to widen the scope of the learning algorithm.

As expected, running PSI-BLAST on an Intel Core duo CPU P8600 @ 2.40GHz with 3 GB ram, was a rather time consuming, yet not computational heavy. Each protein took on average 5 minutes to process. The selected proteins have a total of 150070 amino acids.

At this point, several k-means were run with a varying k values in order to determine optimal number of clusters. However the results were rather incoherent and the algorithm did not converge. In order to surpass this problem a rationale was established: not every

position in a protein is highly relevant to its structure and/or function. By observing the matrixes originated from the PSI-BLASTs this is quite conspicuous; on one hand, certain protein positions are highly conserved, therefore only specific amino acids are permitted at that position, having very high probabilistic values when compared to the remaining amino acids. On the other hand, positions that are not relevant have quite indistinguishable low probabilistic values for whichever amino acid, meaning that in that position any amino acid is acceptable, for it will have little impact on the proteins' structure and/or function. For this reason these PAFPs were excluded from further analysis.

As to understand which would be the adequate probabilistic threshold, that would distinguish between relevant and none relevant positions, several were established and the number positions and information in those positions was evaluated.

Table 3 – Evaluated thresholds and corresponding number of selected PAFPs.

Threshold values	# Selected PAFPs
5000	457
500	2404
400	3486
300	3928
250	3972
200	4310
100	5707

The 5000 threshold was too strict: only PAFPs that included scores higher than 8 were selected with this value, hence skewing the data and dismissing several potential PAFPs, whose relevance to the study, at this point, was not understood.

4.3.1 k-means

Having figured out which range of thresholds would be adequate to carry out the study, it then became possible to determine the k value. This, was selected once the within groups sum of squares did not vary, this was verified for k = 65 for the every threshold.

Despite the stabilization of the within groups sum of squares, the k-means results could not be checked across the various thresholds, since the algorithm starts with random k vectors every run, which means that a cluster named 32 in one run can actually be named 1 in the next. To address this issue, a 100 k-means were run over the whole data (no threshold was applied) with k = 65 and another 10 were run over similarly structured but random data. Both these data were treated as having the same common origin. Each cluster of each k-means was then compared with remaining clusters of each remaining k-means, generating for each pair of k-means a matrix of intersections among clusters. Over the latter, an altered Hungarian algorithm was run in search of maximum overlap/intersection between clusters, i. e. matching the clusters across the k-means, so that that it would be possible to verify the closest matches between the generated clusters. Because this process was executed over the total data and the number of PAFPs amounted to 27624 altogether. It was expected that among the 100 k-means of real data the Hungarian output would sum up to values close to the number of PAFPs when in comparison to the random data, that would generate much lower values. In fact, it was verified that for the real data the values were between 19000 and 22000. In contrast, the values for the random data never exceeded the value 800. Meaning that despite the first step of k-means being random and potentially generating different results for each run, it is safe to say that it is fairly coherent. Furthermore, one must keep in mind that this experiment was carried out with all the data, none of the least conserved positions were discarded at this point, contributing with a lot of irrelevant PAFPs and even so the results were reasonably coherent. This essay was not repeated for the thresholds previously mentioned due to its high time consumption. However, this is not detrimental to the study since even with all the data, coherence was verified.

4.3.2 Association rule learning

At this point, because this was merely a proof of concept and in order to understand if the rationale was indeed correct and the project should continue to be carried out as previously envisioned, only the 500 threshold was run for the posterior phase. The resulting PAFPs were allocated to each of the 65 centroids, making it possible to create a sparse matrix that contained each protein (line) and the clusters (column) that represent it, i. e., if a determined protein is represented by a cluster that intersection has a non-zero value. This data was then introduced into the *a priori* rule association algorithm – R's arules package, having the protein's name and family (either EC 1.1.1.1 or EC 1.1.1.38) as right hand side of the rules and the latter matrix as left hand side rules; keeping the length of the rules at a maximum of 3; a maximum support of 30% and a minimum confidence of 80%.

The resulting rules were in fact able to distinguish between the two families, however due to the small number of proteins used for training, the results had very little relevance. On the other hand, the model used did produce expected discerning of protein characteristics, i. e. certain clusters number were only present in one of the families and absent in the other. Other not so flagrant presences and absence of clusters are also verifiable.

5 Results and Discussion

5.1 Data Retrieval and Processing

Once the concept was corroborated by the data, a larger training set was also retrieved from Swiss-Prot. From the 547357 protein sequences in Swiss-Prot, release 2015_07 on February 2015, the ones annotated with any of the experimental evidence codes: Inferred from Experiment – EXP, Inferred from Direct Assay - IDA, Inferred from Physical Interaction – IPI, Inferred from Mutant Phenotype – IMP, Inferred from Genetic Interaction – IGI and Inferred from Expression Pattern – IEP, were kept has training set. Approximately a tenth of the proteins contained in Swiss-Prot - 57047 proteins were selected to be used as training set.

The decision to use proteins annotated with experimental evidence codes is due to fact that for over two decades, now, the majority of sequences found in public data bases have been annotated using computational prediction methods alone, which raises awareness for annotation accuracy and database quality.

PSI-BLAST were then run for the aforementioned protein sequences on a server with CentOs, version 5.11, 8 Gb of RAM and a Intel ®Xeon E5630 2.53Ghz processor. And as described previously, the resulting PSSM outputs where then processed (Annex 1) in order to obtain a matrix of probabilities. Again, several thresholds were established, this time ranging between 100 and 500 with a step of a 100, to eliminate negligible PAFPs.

As previously mentioned this work aims to establish a relation between located amino acid probability patterns within proteins and their functions. Therefore, in order to generate a training model to conceptually prove the hypothesis, GO terms for the relevant proteins were also retrieved from Swiss-Prot. Analogously to the proof of concept, ancestral terms for each of the terms annotated for each protein were also obtained from Gene Ontology Database, using an altered version of the program [49] available at Annex 2.

5.2 Protein data and annotation

The retrieved proteins were annotated with a total of 22793 distinct GO terms. Afterwards, ancestral terms were also obtained and the GO terms' count increased to 26762.

The GO terms obtained directly from the annotations (for the sake of simplicity, henceforth designated originals) were not leaf terms, some were constant in many proteins' annotations. The most frequent terms refer to cellular location; being nucleus, cytoplasm and integral membrane component the highest rating ones; over 10000 proteins contain these annotations. However, it is important to keep in mind that no information over the evidence code of that GO term is taken into consideration in this model.

On the other end of the spectrum (not shown), there were 10372 GO terms that were represented by less than five proteins and 3840 represented by only one protein. These terms are in general in closer to leaf level, hence richer in detail content.

For the ancestral terms, retrieved in association with the previous 22793 terms, the most frequent terms are biological_process, cellular_component, representing above 50000 proteins; molecular_function only appears in forth position, below cell part. According to a critical assessment of protein function annotation [52], for the existing available tools, there is a substantial difference in the ability to predict the two GO categories: molecular function versus biological process. This is can be partly explained by the fact that biological process has a larger number of terms, branching factor, maximum depth and number of leaf terms, than the molecular function category. Therefore, the former is more represented than the latter in data sets used to train the models, it would not surprising then that the model here proposed suffers from the same bias.

On the other side of the spectrum, 9460 GO ancestral terms represent less than five proteins and 3427 GO terms represent one single protein. In comparison, with the homologous results for the original GO terms only, it is visible a decrease in the number of proteins represented. Moreover, from 3840 proteins with original terms and the 3427 ancestral terms, only 2886 are the same.

5.3 PSI-BLAST

When PSI-BLAST was run over the sequences, surprisingly 169 of these did not generate any output, essentially because no distant homolog relatives exist in the Swiss-Prot database. Thus, the rest of the learning process was carried out with 56878 sequences.

Despite the PSI-BLAST process being run on much powerful machines than the one used for the proof of concept, it was still a very time consuming process; the all process took up approximately 3 months, even though the sequences were processed in parallel. Had they been run without parallelism on a single machine, the process would have taken approximately 10 months.

5.4 k-means and arules

Due to the amount of data being processed it was not feasible to use R, due to its memory limitations. The k-means algorithm was then implemented, using a Forgy initialization method, in Python 2.7, where the first 65 k vectors were randomly selected throughout the data. This implementation of k-means was run for all 5 thresholds, maintaining 65 centroids, until it converged.

Contrary to R, instead of loading all the information to the machine's memory, in Python the calculations were done in segments, defined by the proteins' length; such is the case of the located PAFPs selection for each of the previously established thresholds.

As stated, the seeding process was done according to the Forgy method, i. e., the initial vectors were chosen at random, and the new centroids were calculated through the Euclidean distances to the previous centroids. This was preferred over the Hartigan-Wong algorithm for its simplicity and swiftness of implementation. Since it was previously verified that the data was quite coherent, it is inferable that the difference in seeding process should have little impact on the outcome.

From this point onwards all thresholds were run both in k-means and arules. The *a priori* algorithm from arules package was run over the data, establishing association rules between the clusters summarizing the located patterns (LHS) and GO terms (RHS). A combination of various parameters were tested: *i)* maximum length of the rules generated was kept at 4; *ii)* the minimum support, varying between 0.05 and 0.0005 and *iii)* an interval from 90 to 70% of minimum confidence; for the thresholds of 500 and 300, the length of the rules generated was widened to a maximum length of 5 and the confidence was lowered even further to 40%, with a step of 10. All possible combinations of the three referred parameters were executed.

Because ancestral GO terms were also included in this analysis, many of the resulting association rules where referent to them, hence limiting the scope of the project. Therefore, in order to obtain rules regarding the lower level GO terms, the eighteen highly recurring ancestral GO terms, were removed from the analysis.

The most stringent threshold (T500) selected approximately 4 million PAFPs, as for the least restricting threshold (T100) selected about 15.3 million PAFPs. The T400, T300 and T200 generated between 7 million and 9 million PAFPs. The abrupt increase in PAFPs between the T200 and T100 is rather conspicuous, implying that last threshold is too lax and is including too many PAFPs of little relevance, i. e., that it is unimportant for function whichever amino acid is found at that position. Nevertheless, all combinations of the previously mentioned arules/*a priori* algorithm parameters were tested.

Table 4 – Summary of the number of selected proteins and PAFPs for each of the thresholds.

	T500	**T400**	**T300**	**T200**	**T100**
#selectedProts	56856	56877	56877	56877	56878
#selectedPAFPs	4 006 821	7 044 490	7 382 181	8 929 717	15 278 788

Keeping the support at a minimum of 5.0% and varying the confidence between 90% and 80% generated no rules for most of the thresholds. For confidence levels of 75% to

70% a reasonable amount of rules, approximately between 300 and 800 from T500 to the T100, were generated but only 15 GO terms were actually referenced (Tab. 4), even posterior to the removal of the 18 most incident and ancestral ones. As expected, these were the same across the 5 thresholds. However, since this GO terms were closer to leaf level they were not removed, as they could be significant to understand protein function.

Table 5 – Resulting 15 GO terms from the arules with a minimum support of 0.05 at 70% confidence.

Id	Name	supp	conf	Id	name	supp	conf
GO:0043168	anion binding	0.073	0.70	GO:0001883	purine nucleoside binding	0.058	0.72
GO:0044238	primary metabolic process	0.092	0.72	GO:0032549	ribonucleoside binding	0.059	0.74
GO:0035639	purine ribonucleoside triphosphate binding	0.059	0.72	GO:0032555	purine ribonucleotide binding	0.057	0.72
GO:0097367	carbohydrate derivative binding	0.060	0.72	GO:0017076	purine nucleotide binding	0.059	0.73
GO:0000166	nucleotide binding	0.070	0.70	GO:0032550	purine ribonucleoside binding	0.059	0.73
GO:0036094	small molecule binding	0.071	0.70	GO:1901265	nucleoside phosphate binding	0.069	0.70
GO:0032553	ribonucleotide binding	0.059	0.74	GO:0001882	nucleoside binding	0.059	0.74
GO:0065007	biological regulation	0.052	0.70				

Note: by order of occurrence in *a priori* algorithm output, no sort was applied.

Lowering the support, to 0.005, and testing for the same interval of confidence did increase the number of rules generated but did not significantly improve the amount GO terms identified. For the three highest thresholds, within the confidence interval of 90-80%, the number of rules was significantly lower in comparison when the confidence is dropped to 75-70%. And again, the identified GOs coincide with the above specified. At the 70-75% several new GOs are identified, nevertheless the number of GOs is still quite low, narrowing the scope of the learning model. For the two lowest thresholds, the 90-80% confidence interval generates no rules at all, this is very likely due to the amount of uninteresting PAFPs still included in the data, lowering significantly the support and confidence for the relevant PAFPs.

In an attempt to circumvent the aforementioned issues, the minimum support level was again lowered to 0.05% with the same range of confidence values. At this support level, and for the T500, T400 and T300, with 90-80% confidence there was slight increase in the rules generated and identified GOs as the confidence is decreased. However, at lower confidence levels, 75-70%, a significant increase in the rules generated was verified, this is true for the identified GOs as well. Identifying approximately double the GO terms identified at the support of 0.5% with the 70% confidence level. This was not the case for the lowest thresholds T200 and T100. Indeed it is verifiable an increase both in rules and GO terms identified, however this is not comparable to the results obtained for the highest thresholds. The same pattern is verifiable: at 90-80% confidence the rules generated and GO terms identified are in number equal to the obtained with the 0.005 support (corresponding to approximately 30 proteins) and 75-70% confidence. Again at lower support levels the number of rules increase and the GO terms duplicate.

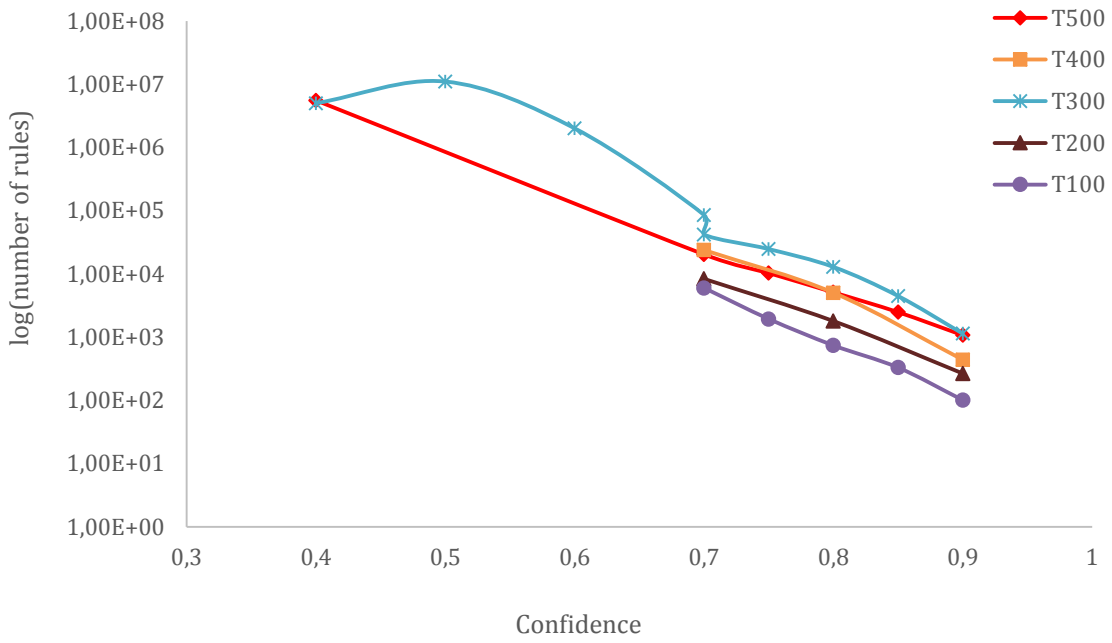


Figure 5 – Logarithmic representation of the number of resulting rules for thresholds 500, 400, 300, 200 and a 100, at a support of 0.0005 (≈30 proteins) and confidence ranging between 90 and 40% and a maximum rule length of 4; except for the T300, where maximum length of 5 is also represented.

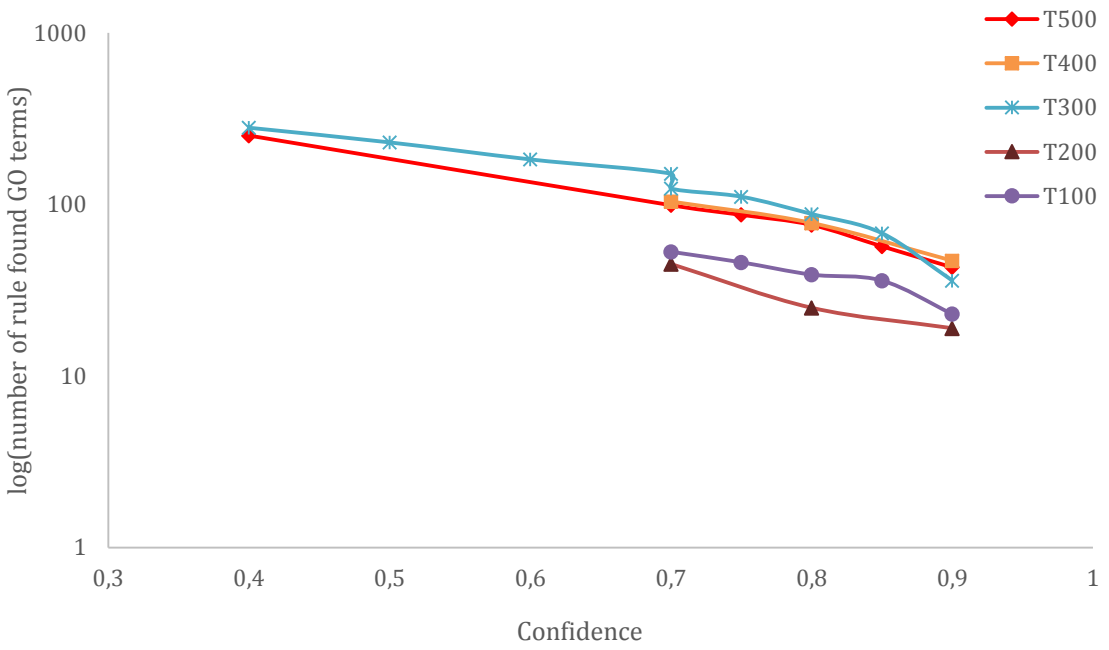


Figure 6 - Logarithmic representation of the number of rule-found GO terms for thresholds 500, 400, 300, 200 and a 100, at a support of 0.0005 (≈30 proteins) and confidence ranging between 90 and 40% and a maximum rule length of 4; except for the T300, where maximum length of 5 is also represented.

Notwithstanding, the number GO terms identified by this combination of parameters are in the tens, in contrast to the numbers of GO terms identified with the same combination of parameters for the higher thresholds, which in the hundreds. As previously mentioned, this is very likely due to the PAFPs that are not well conserved and are included in the data at this thresholds.

Overall, T300 generated the largest number of relevant GO terms independently of support and confidence. Nevertheless, 124 identified GOs, remains a very low number when compared with the number GO terms with a support of 0,05%, i. e., when compared with the number GO terms that are contained in the annotation of at least 30 proteins, considering the 57047 proteins in the training set.

Table 6 – Summary table of conditions tested in arules and outputs.

		T500		T400		T300		T200		T100	
Support	Confidence	Number of Rules	Number of GosRules	Number of Rules	Number of GosRules	Number of Rules	Number of GosRules	Number of Rules	Number of GosRules	Number of Rules	Number of GosRules
0.05	0.9	0	NA	0	NA	0	NA	0	NA	0	NA
	0.85	0	NA			0	NA			0	NA
	0.8	0	NA	0	NA	2	2	0	NA	0	NA
	0.75	13	5			50	14			24	5
	0.7	339	15	448	15	639	15	243	7	814	15
0.005	0.9	81	8	67	1	230	3	0	NA	0	NA
	0.85	201	12			1299	13			0	NA
	0.8	563	25	1015	23	3792	39	4	4	6	4
	0.75	1494	30			8063	46			229	19
	0.7	4560	42	6374	46	14585	48	1327	25	2340	24
0.0005	0.9	1073	43	439	47	1137	36	266	19	101	23
	0.85	2480	57			4470	68			330	36
	0.8	5107	76	5031	78	12869	88	1794	25	740	39
	0.75	10295	87			24687	111			1938	46
	0.7	20297	99	23805	104	42007	124	8337	45	5993	53
	0.7					85300	151				
	0.6					2,00E+6	183				
	0.5					1,10E+7	230				
	0.4	5532000	252			5,00E+6	280				

Note: grey shaded cells refer to rules with a maximum length of 4 and the blue shaded cell refer to rules with a maximum length of 5.

In order to understand if this results could be surpassed a compromise in terms of confidence was made. The latter was lowered to 40% in steps of 10%. Predictably, the number of rules generated increases inversely to the confidence level and so does the number of identified GOs. However, the number of rules generated is within the tens of thousands and the number of identified GOs reaches a maximum at 280 GOs.

By analysing the results, the T300 is least strict threshold that manages to exclude the highest number irrelevant PAFPs. Also, T300 includes PAFPs from all the proteins in the training set except for one (Tab. 5), this avoids restraining the model at this step, allowing a wider basis for the learning process. Therefore, this was the threshold selected for model validation.

Crosschecking the generated rules selected to train the model, with the clusters, it is verifiable that not all clusters are significant. The 280 rules generated had several combinations of the clusters, ranging from one GO term identified by one cluster to one GO term identified by 4 clusters (Annex 4). However, and despite being the vast majority, only some clusters were used to generate rules, clusters 7, 12, 20, 28, 30, 33, 47, 50 are missing, this represents they were not used in any rules.

Table 7 – Number of proteins contained in the clusters not used for rule association in comparison with some clusters, selected at random, from the ones used for rule association.

Unused clusters		some used Clusters	
cluster	#proteins	cluster	#proteins
7	31566	1	30869
12	19035	22	10460
20	12399	24	16420
28	52254	35	15587
30	42627	45	42039
33	40175	64	11944
47	40614		
50	26104		

Out of the 65 predefined k clusters, only 57 of them generate rules for the set of parameters tested. It is interesting, to verify that despite not contributing to generate rules, these centroids coordinates were calculated from several PAFPs, hence proteins (Tab. 7). One could argue that they were not used to generate rules either because many proteins did not contain those clusters or because many proteins did contain those clusters, however by

comparing with the second half of table 7, one can verify that is not the case, the unused clusters have a similar occurrence in proteins to the used ones.

Due to time constraints the k-means with this parameter alteration were not run again. Notwithstanding, it would be interesting to verify if the resulting rules were identical, or if the lines spread through the remaining 8 clusters, that would then have to be spread across the 57, would change the average of the cluster vectors, resulting in different rules.

5.5 Validation

A superficial analysis, indicates that the 280 GO terms selected by the T300 with a support of 0.05%, and a confidence of 40% are used in the annotation of 516,591 proteins out of 549,008, the current number of proteins contained in Swiss-Prot's release 2015_08. It is noteworthy that 28,530 of these proteins have no attributed GO terms in their annotation. Translating into a total of 3887 proteins that are not annotated with any of these GO terms. This indicates that the 280 identified GO terms are not leaf level on the GO tree, but higher level GO terms. There are 13 GO terms with a level of 1, with information content circa 8.5, and only 6 GO terms shared among level 9 and 10 with an information content circa of 11 (Annex 7), incidentally the lower terms are the least represented in Swiss-Prot, with only 0.3 to 1% of the proteins containing this annotation.

In order to further validate the model, proteins' sequence whose annotation contained any of the aforementioned experimental evidence codes were retrieved, as well as, their GO terms and ancestral, as previously described. The previous 57047 relevant proteins were subtracted from the 59145 newly retrieved ones.

Removing the proteins previously used as training set 57047 (56878+169), from the 59145 proteins, resulted in 2591 proteins to constitute the validation set. However, this result is rather odd, since difference between the two sets should result in 2098 proteins, not 2591; there is an excess of 493 proteins. In order to understand these results the proteins in the training set were crosschecked with the newly retrieved ones and, where in fact, there should be 57047 proteins common to both sets, there is only 56554. By

subtracting the latter from the training set, it is then possible to identify these 493 odd proteins and justify the discrepancy.

The selection of the training and validation set was done using the following regular expression was used (code snippet at Annex 5):

('.*EXP:.*|.*IDA:.*|.*IPI:.*|.*IMP:.*|.*IGI:.*|.*IEP:.*')

each of the terms within parenthesis corresponding to experimental evidence codes. Hence, as previously described, it would be enough for a protein to have a single one of these codes to be selected. Analysing the Swiss-Prot release 2015_07, 57047 proteins had at least one of these terms in its annotation; analyzing the Swiss-Prot release 2015_08, 59145 proteins had at least one of these terms. However, 493 of the proteins selected in the training set had their annotations changed and proteins that once, in the 7th release, had at least GO term annotated with an experimental evidence code, no longer had, in the 8th release; this is true for both the GO term as well as for the experimental annotation (annex 9).

Nevertheless, the newly found 2591 proteins were treated as previously described: *i)* a PSSM was obtained for each; *ii)* the PAFPs were exponentiated to better distinguish them, *iii)* a threshold of 300 was applied to each of the proteins selecting the interesting PAFPs; *iv)* each protein position is then attributed to the previous calculated clusters (Annex 3), by Euclidean distance. Posteriorly, they were matched to the GO terms previously identified and to the rules resulting from the previous learned data.

Similarly to the training set, not every protein generated a PSSM, but in this case only one protein did not generate a PSSM for lack of identifiable distant relatives. A total of 2590 PSSMs were obtained. The model managed to allocate the 2590 proteins to the 57 clusters based on the previously established rules.

So far, in this project, it has not been considered to differentiate between molecular functions, biological process or cellular component (the GOC function defining categories), this is, mainly, because the method aimed to include as much information about the proteins' function as possible, and, has stated previously, function is described by these categories. Nevertheless, it seems this approach might have been too eager; by not developing a GOC-category-specific GO term algorithm, conservation patterns that might

more directly match one of these GOC-category-specific GO term, might be lost in the process, i. e. a cluster that might accurately describes a GO term from one of the categories, to instantiate: a cluster that accurately describes the biological process category GO term, might in the method here used be clustered to a wider group, because all categories are clustered together, preventing the identification of that GOC-category-specific GO term.

Also, according to the algorithm here used, it would be enough for one protein to be selected if only one evidence code was established experimentally, independently of the GO term that had been annotated with that evidence term. This represents that some proteins were annotated with GO terms other than the ones experimentally established. In other words, even if a protein only has one single experimentally annotated GO term and has several others with evidence codes of electronic inference, the protein would still be selected for the training set. In this way, the model inference capability is further degraded, because some of the annotations might indeed be erroneous since they are not experimentally annotated in full. Also, because no difference between these annotations is made, i. e., no weight is set for experimental and non-experimental annotated GO terms; and because arules is parameterized with confidence and support levels, the rules are generated with both in an undifferentiated manner. These electronical annotated GO terms skews the confidence and support values for the rules, requiring these parameters to be lowered.

Considering the number of ancestral GO terms included in the training set, it is expectable for these to be more frequent the more proteins they represent, they will therefore contribute with very little relevant information but because they are so frequent most of the rules with higher support will be referent to these. To diminish this effect, the higher recurrent terms were removed, however, the prevalence of rules with ancestral GO terms, is not expected to have been completely eliminated. In other words, it is expectable that the rules with higher confidence and support levels be referent to ancestral GO terms higher in the GO tree. This is the reason why lowering the parameters enables finding rules with GO terms not included in the more rules generated with higher percentages of support.

After obtaining for each GO term the relevant classification statistics (Positives, Negatives, Inferred Positives, False Positives, True Positives, False Negatives, True

Negatives, recall, precision and F1 score as well as MCC - tab. 8, 9 and annex 8), it is evident that the model hardly ever infers exactly the same number of proteins to each GO term as the ones truly annotated to it.

In some situations, the model infers more proteins to a GO term than the ones annotated to it, resulting in high counts of false positives (tab. 9); in others, it infers less proteins than the ones annotated to a GO term, resulting in high counts of false negatives. Some of the worst situations: term GO:0016866 [intramolecular transferase activity] (annex 8) where all the proteins inferred with the GO term by the model were not annotated to it; another example is GO:0048869 [cellular developmental process], where only 3 proteins, out of hundreds effectively annotated to the term, were inferred. In such cases, it is not possible to obtain neither recall nor precision or F1 values, since there are no true positives and these metrics are calculated upon this value; expectedly, the MCC has values slightly lower than zero, implying there is not a positive correlation between inference and truth values, in other words, the model results are indistinguishable from a random estimation. There are other cases where the model had no discriminatory capability, for instance the term GO:0031559, where all proteins were inferred to the term. In this situations (tab. 8, fig. 7), recall values are generally high, while precision values are lower, this is due to the fact that all existing proteins are inferred to the GO term, contributing to the high recall values, but since half of them are erroneously inferred, the precision values are lower. It is also, noticeable that F1 values vary proportionally to the number of inferred proteins. Furthermore, it is noteworthy that this GO term, and the remaining ones where this oocurs, are indeed very common in the Swiss-Prot annotations of the 2590 validation proteins, suggesting these are ancestral terms, which contribute with very little relevant information. Crosschecking with the information on annex 7, it is visible that these GO terms have very high level and low information content.

Table 8 – GO terms with the lowest MCC.

GO term	P	N	IP	FP	TP	FN	TN	Recall	Prec	F1	MCC
GO:0006807	680	1910	1667	1207	460	220	703	0,676	0,276	0,392	0,0401
GO:0016020	798	1792	1874	1275	599	199	517	0,751	0,32	0,449	0,040
GO:0044422	837	1753	1558	1033	525	312	720	0,627	0,337	0,438	0,036
GO:0043170	800	1790	1876	1297	579	221	493	0,724	0,309	0,433	-0,001
GO:0044260	740	1850	1623	1160	463	277	690	0,626	0,285	0,392	-0,001
GO:0050789	1165	1425	2583	1422	1161	4	3	0,997	0,449	0,619	-0,013
GO:0050794	1075	1515	2542	1494	1048	27	21	0,975	0,412	0,579	-0,041

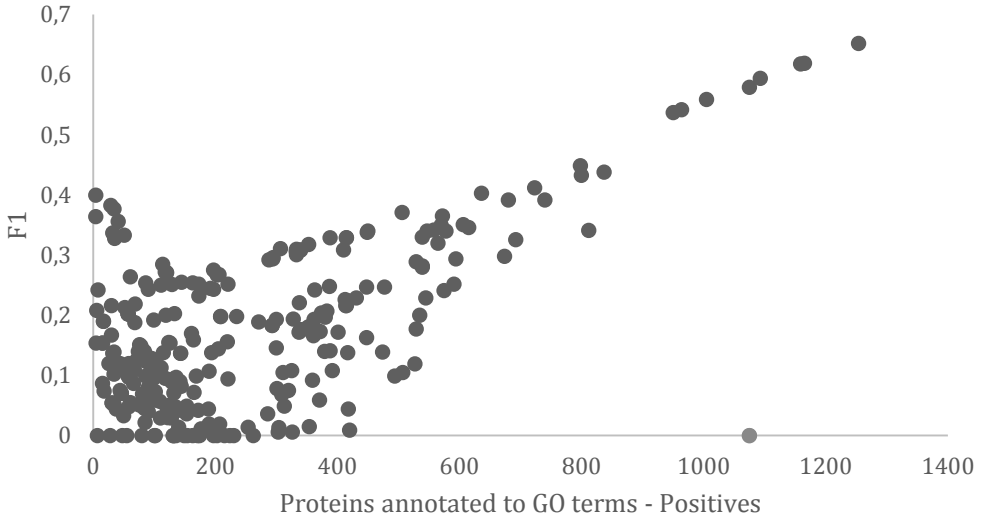


Figure 7 – Model evaluation in terms of F1 score and positives.

Notwithstanding, by selecting the highest F1 along with the highest MCC values it is possible to identify the GO terms that indeed grant inference capability to the model (tab. 9); these are more specific GO terms that contribute with more meaningful information for function annotation, it is verifiable that the recall values are rather high as well as the MCC values. On the other hand, the precision values are somewhat low. To instatiate: GO:0005231 [excitatory extracellular ligand-gated ion channel activity] and GO:0009069 [serine family amino acid metabolic process] (tab. 9) correspond to the highest information content values described at annex 7 and are represented by a small amount of Swiss-Prot

proteins; nevertheless, the model managed to infer proteins correctly to these go terms with high recall values and moderate precision, resulting in high F1 and MCC values.

Table 9 – GO terms with the highest F1 scores and Matthews' correlation coefficient, in decreasing order of MCC.

GO term	P	N	IP	FP	TP	FN	TN	Recall	Prec	F1	MCC
GO:0005506	29	2561	86	64	22	7	2497	0,759	0,256	0,383	0,431
GO:0020037	34	2556	88	65	23	11	2491	0,676	0,261	0,377	0,409
GO:0005231	4	2586	6	4	2	2	2582	0,5	0,333	0,4	0,407
GO:0009069	4	2586	7	5	2	2	2581	0,5	0,286	0,364	0,377
GO:0004930	51	2539	171	134	37	14	2405	0,725	0,216	0,333	0,376
GO:0046906	41	2549	88	65	23	18	2484	0,561	0,261	0,356	0,369
GO:0004713	6	2584	42	37	5	1	2547	0,833	0,119	0,208	0,312
GO:0004888	86	2504	388	328	60	26	2176	0,698	0,155	0,254	0,285
GO:0005230	8	2582	25	21	4	4	2561	0,5	0,16	0,242	0,279

These results strongly suggest that the proposed approach was able to find relevant terms, even when the number of positives is very low, which makes the problem very difficult. Low precision values imply that several false positives are being found, however due to the incompleteness of most annotations, it is difficult to assess the importance of this statistic.

Assessing protein function annotation is complicated by the complex nature of protein function, it is likely, that functional annotations do not fully describe a protein's function [10]. Annotations can be too general, in some cases proteins are annotated with a general GO term when a more specific GO term better describes its function. Some proteins, especially the multi-domain ones, may also have more functions than those they are annotated with. It is, therefore, plausible to consider that predictions are more specific than the existing annotations and even those which are apparently completely different from existing annotations might actually be correct. Furthermore, as previously mentioned, the model is trained with annotations that include some GO terms with non-experimental evidence codes, if any of these is incorrectly annotated, which is likely [53] , then the model

might be skewed. For example, the controversial protein E1V4Y0 that was, until 2010, annotated, by electronic inference, with galactonate dehydratase, when in fact its sequence similarity score is well below the cutoff trusted for proteins of that family, in fact by sequence similarity alone the protein should have been annotated with gluconate dehydratase (GO:0047929); only, after experimental validation was the protein annotated as such. It is noteworthy, that despite being annotated experimentally as gluconate dehydratase, its record name, in Swiss-Prot, remains associated to the galactonate dehydratase (GO:0008869) family. Moreover, in the same Swiss-Prot annotation where the above information can be found, in the subsection Function it is stated that this protein has low dehydratase activity both with D-mannonate (GO:0008927) and D-gluconate (GO:0047929), it has no significant role in the in vivo degradation of these compounds and has no detectable activity with a panel of 70 other acid sugars (in vitro) [54]. Evidencing that experimental validation previously obtained is not a 100 % reliable. Other examples are evidenced by [53].

Furthermore, the fact that the Swiss-Prot, the gold standard for protein annotation, is not static in relation to the annotation content, it makes it quite difficult to truly understand the extent of the error of the model. For example, the set of 493 proteins of the 56878, that were used as training set of this model, retrieved from an older release of Swiss-Prot, which upon validation of the model, with a more recent release of Swiss-Prot, ceased having at least one experimentally annotated GO term and being annotated with several different electronically inferred GO terms, some of this proteins and their GO terms can be found at annex 9, they were not all included in this document because it would be an overwhelming amount of information, instead a few representative examples are shown.

6 Concluding Remarks

The model here proposed for automatic protein annotation using positional amino acid frequency patterns, does manage to identify GO terms for protein annotation with high recall values, but not equivalent precision. As with most protein annotation methods, these values are not perfect. Also, from the 7271 GO terms annotated to 30 or less proteins, the model only managed to identify 280. Despite, falling short of what was expected, the results strongly suggest that the existence of certain PAFPs within proteins may be important for their function. It is also interesting that the strongest signal was found on terms for which the positive ratio is very low, which are typically very difficult classification problems. Results strongly suggest that it may be possible to find annotation clues by looking on amino acids substitution patterns alone. The results however were not perfect and more work will certainly be required to further validate the initial findings.

To circumvent the multilabel classification limitation, some groups [9,10] have described methods where the models are trained separately for each of the GO terms specific families, i. e., after selecting the proteins experimentally annotated, these are separated into groups according to their GO term annotations. Afterwards, they are aligned and the resulting functional subalignment is used to construct PSSMs. It would be interesting, in the future, to take an identical approach. Especially, because it may allow to predict functions that would otherwise be lost, for instance, by identifying a molecular function GO term, one might be able to infer, in an unrelated manner, the cellular component where the said protein could be found. This sort of information is valuable when attempting to design a wet lab experiment, diminishing the number of possible paths to procure, by indicating a more likely route; if not invalidating some, at least enabling a sort of hierarchy when establishing what to experiment first.

Also, it would be interesting to take into account, that some of the GO terms included were not all experimentally annotated and understand to what extent they influence the model. These could be considered in the model by incorporating a weight system, where the experimentally annotated would have higher weight than the non-experimental. Also,

quality analysis could be carried out with two different sets of data: *i)* only the terms experimentally annotated, *ii)* with the electronic annotated terms. This analysis would also allow to understand if the data set could be enriched by including the latter, widening the support range of the model. Previously, it has been demonstrated that using more extensive electronic annotations results in improved precision compared to a set of non-electronic annotations [10].

Bibliography

1. Online DG. https://gold.jgi.doe.gov/index. [accessed 30 Oct 2015].

2. UniProt. https://www.ebi.ac.uk/uniprot/TrEMBLstats. [accessed 30 Sep 2015].

3. Brenner SE. Errors in genome annotation. Trends in Genetics. doi:10.1016/S0168-9525(99)01706-0. 1999. pp. 132–133.

4. Devos D, Valencia A. Intrinsic errors in genome annotation. Trends in Genetics. doi:10.1016/S0168-9525(01)02348-4. 2001. pp. 429–431.

5. NC-IUBMB. Enzyme nomenclature. http://www.chem.qmul.ac.uk/iubmb/enzyme/. 1992 [accessed 18 Sep 2015].

6. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Güldener U, Mannhaupt G, Münsterkötter M, Mewes HW. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res. doi:10.1093/nar/gkh894. 2004;32: 5539–5545.

7. Gene Ontology Consortium. Gene Ontology Documentation. http://geneontology.org/page/documentation. [accessed 18 Sep 2015].

8. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. doi:10.1038/75556. 2000;25: 25–29.

9. Pazos F, Sternberg MJE. Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci U S A. doi:10.1073/pnas.0404569101. 2004;101: 14754–14759.

10. Wass MN, Sternberg MJE. ConFunc - Functional annotation in the twilight zone. Bioinformatics. doi:10.1093/bioinformatics/btn037. 2008;24: 798–806.

11. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Automatic prediction of protein function. Cell Mol Life Sci. doi:10.1007/s00018-003-3114-8. 2003;60: 2637–2650.

12. Reeck GR, de Haën C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, Chambon P, McLachlan AD, Margoliash E, Jukes TH. “Homology” in proteins and nucleic acids: a terminology muddle and a way out of it. Cell. doi:10.1016/0092-8674(87)90322-9. 1987. p. 667.

13. Devos D, Valencia A. Practical limits of function prediction. Proteins Struct Funct Genet. doi:10.1002/1097-0134(20001001)41:1<98::AID-PROT120>3.0.CO;2-S. 2000;41: 98–107.

14. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol. doi:10.1006/jmbi.2000.3550\rS0022-2836(00)93550-2 [pii]. 2000;297: 233–249.

15. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol. doi:10.1016/j.jmb.2003.08.057. 2003;333: 863–882.

16. Addou S, Rentzsch R, Lee D, Orengo CA. Domain-Based and Family-Specific Sequence Identity Thresholds Increase the Levels of Reliable Protein Function Transfer. J Mol Biol. doi:10.1016/j.jmb.2008.12.045. 2009;387: 416–430.

17. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. Q Rev Biophys. doi:10.1017/S0033583503003901. 2003;36: 307–340.

18. Rost B. Twilight zone of protein sequence alignments. Protein Eng. doi:10.1093/protein/12.2.85. 1999;12: 85–94.

19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. J Mol Biol. doi:10.1016/S0022-2836(05)80360-2. 1990; 403–410.

20. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research. doi:10.1093/nar/25.17.3389. 1997. pp. 3389–3402.

21. Chakraborty A, Bandyopadhyay S. FOGSAA: Fast Optimal Global Sequence Alignment Algorithm. Suplementary Material. Sci Rep. doi:10.1038/srep01746. 2013;3: 1746.

22. Henikoff S, Henikoff JG. Position-based sequence weights. J Mol Biol. doi:10.1016/0022-2836(94)90032-9. 1994;243: 574–578.

23. Tatusov RL, Altschul SF, Koonin E V. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc Natl Acad Sci U S A. doi:10.1073/pnas.91.25.12091. 1994;91: 12091–12095.

24. Neto-Silva RM, Macedo-Ribeiro S, Pereira PJB, Coll M, Saraiva MJ, Damas AM. X-ray crystallographic studies of two transthyretin variants: Further insights into amyloidogenesis. Acta Crystallogr Sect D Biol Crystallogr. doi:10.1107/S0907444904034316. 2005;61: 333–339.

25. Vogel C, Berzuini C, Bashton M, Gough J, Teichmann SA. Supra-domains: Evolutionary Units Larger than Single Protein Domains. Journal of Molecular Biology. doi:10.1016/j.jmb.2003.12.026. 2004. pp. 809–823.

26. Sleator RD, Walsh P. An overview of in silico protein function prediction. Archives of Microbiology. doi:10.1007/s00203-010-0549-9. 2010. pp. 151–155.

27. Sigrist CJA, Cerutti L, De Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. doi:10.1093/nar/gkp885. 2009;38.

28. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. doi:10.1093/nar/28.1.235. 2000;28: 235–242.

29. Ye Y, Godzik A. FATCAT: a web server for flexible structure comparison and structure similarity searching. Nucleic Acids Res. doi:10.1093/nar/gkh430\n32/suppl_2/W582 [pii]. 2004;32: W582–5.

30. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. doi:10.1093/protein/11.9.739. 1998;11: 739–747.

31. Wang S, Ma J, Peng J, Xu J. Protein structure alignment beyond spatial proximity. Sci Rep. doi:10.1038/srep01448. 2013;3: 1448.

32. Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res. doi:10.1093/nar/gkh028. 2004;32: D129–D133.

33. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. Protein function in the post-genomic era. Nature. doi:10.1038/35015694. 2000;405: 823–826.

34. Gabaldón T, Huynen MA. Prediction of protein function and pathways in the genome era. Cellular and Molecular Life Sciences. doi:10.1007/s00018-003-3387-y. 2004. pp. 930–944.

35. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. Science. doi:10.1126/science.285.5428.751. 1999;285: 751–753.

36. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A. doi:10.1073/pnas.96.6.2896. 1999;96: 2896–2901.

37. Walker MG, Volkmuth W, Sprinzak E, Hodgson D, Klingler T. Prediction of gene function by genome-scale expression analysis: Prostate cancer-associated genes. Genome Res. doi:10.1101/gr.9.12.1198. 1999;9: 1198–1203.

38. Klomp JA, Furge KA. Genome-wide matching of genes to cellular roles using guilt-by-association models derived from single sample analysis. BMC Research Notes. doi:10.1186/1756-0500-5-370. 2012. p. 370.

39. Pavlidis P, Gillis J. Progress and challenges in the computational prediction of gene function using networks [v1 ; ref status : approved 1 , http://f1000r.es/SqmJUM]. F1000 Res. doi:10.3410/f1000research.1-14.v1. 2012;1: 1–6.

40. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol. doi:10.1038/msb4100129. 2007;3 : 88.

41. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol. doi:10.1186/gb-2008-9-s1-s4. 2008;9 Suppl 1: S4.

42. Peña-Castillo L, Tasan M, Myers CL, Lee H, Joshi T, Zhang C, Guan Y, Leone M, Pagnani A, Kim WK, Krumpelman C, Tian W, Obozinski G, Qi Y, Mostafavi S, Lin GN, Berriz GF, Gibbons FD, Lanckriet G, Qiu J, Grant C, Barutcuoglu Z, Hill DP, Warde-Farley D, Grouios C, Ray D, Blake JA, Deng M, Jordan MI, Noble WS, Morris Q, Klein-Seetharaman J, Bar-Joseph Z, Chen T, Sun F, Troyanskaya OG, Marcotte EM, Xu D, Hughes TR, Roth FP. A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome Biol. doi:10.1186/gb-2008-9-s1-s2. 2008;9 Suppl 1: S2.

43. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol. doi:10.1038/nrm2281. 2007;8: 995–1005.

44. MacQueen JB. Kmeans Some Methods for classification and Analysis of Multivariate Observations. 5th Berkeley Symp Math Stat Probab 1967. doi:citeulike-article-id:6083430. 1967;1: 281–297.

45. Hartigan J a., Wong M a. A K-Means Clustering Algorithm. J R Stat Soc. 1979;28: 100–108.

46. Forgy E. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics. http://ci.nii.ac.jp/naid/10009668881/. 1965;21: 768–769.

47. Hahsler M, Hornik K, Buchta C. Introduction to arules – A computational environment for mining association rules and frequent item sets. J Stat Softw. 2005;14: 1–25.

48. Agrawal R, Imielinski T, Swami A. Database mining: A performance perspective. IEEE Trans Knowl Data Eng. doi:10.1109/69.250074. 1993;5: 914–925.

49. Telliott99. http://telliott99.blogspot.pt/2010/12/go-gene-ontology.html. 2010.

50. Kuhn HW. The Hungarian method for the assignment problem. Nav Res Logist Q. 1955; 83–97.

51. Munkres J. Algorithms for the Assignment and Transportation Problems. Journal of the Society for Industrial and Applied Mathematics. doi:10.1137/0105003. 1957. pp. 32–38.

52. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML, Piovesan D, Casadio R, Wang Z, Cheng J, Fang H, Gough J, Koskinen P, Törönen P, Nokso-Koivisto J, Holm L, Cozzetto D, Buchan DWA, Bryson K, Jones DT, Limaye B, Inamdar H, Datta A, Manjari SK, Joshi R, Chitale M, Kihara D, Lisewski AM, Erdin S, Venner E, Lichtarge O, Rentzsch R, Yang H, Romero AE, Bhat P, Paccanaro A, Hamp T, Kaßner R, Seemayer S, Vicedo E, Schaefer C, Achten D, Auer F, Boehm A, Braun T, Hecht M, Heron M, Hönigschmid P, Hopf TA, Kaufmann S, Kiening M, Krompass D, Landerer C, Mahlich Y, Roos M, Björne J, Salakoski T, Wong A, Shatkay H, Gatzmann F, Sommer I, Wass MN, Sternberg MJE, Škunca N, Supek F, Bošnjak M, Panov P, Džeroski S, Šmuc T, Kourmpetis YAI, van Dijk ADJ, ter Braak CJF, Zhou Y, Gong Q, Dong X, Tian W, Falda M, Fontana P, Lavezzo E, Di Camillo B, Toppo S, Lan L, Djuric N, Guo Y, Vucetic S, Bairoch A, Linial M, Babbitt PC, Brenner SE, Orengo C, Rost B, Mooney SD, Friedberg I. A large-scale evaluation of computational protein function prediction. Nat Methods. doi:10.1038/nmeth.2340. 2013;10: 221–7.

53. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. doi:10.1371/journal.pcbi.1000605. 2009;5.

54. Wichelecki DJ, Balthazor BM, Chau AC, Vetting MW, Fedorov AA, Fedorov E V, Lukk T, Patskovsky Y V, Stead MB, Hillerich BS, Seidel RD, Almo SC, Gerlt JA. Discovery of function in the enolase superfamily: D-mannonate and d-gluconate dehydratases in the D-mannonate dehydratase subgroup. Biochemistry. doi:10.1021/bi500264p. 2014;53: 2722–2731.

55. University of Cape Town Computational Biology Group. http://www.cbio.uct.ac.za/ITGOM/tools/itgom.php. [accessed 22 Oct 2015].

56. Leonard G. https://commons.wikimedia.org/wiki/File:Gene_Fusion_Types.png. 2012 [accessed 16 Oct 2015].

Annex 1 - Pssmreader.py

```
import os

def read_asn(fname):
##    0|-|Gap
##    1|A|Alanine
##    2|B|Asp or Asn
##    3|C|Cysteine
##    4|D|Aspartic Acid
##    5|E|Glutamic Acid
##    6|F|Phenylalanine
##    7|G|Glycine
##    8|H|Histidine
##    9|I|Isoleucine
##    10|K|Lysine
##    11|L|Leucine
##    12|M|Methionine
##    13|N|Asparagine
##    14|P|Proline
##    15|Q|Glutamine
##    16|R|Arginine
##    17|S|Serine
##    18|T|Threoine
##    19|V|Valine
##    20|W|Tryptophan
##    21|X|Undetermined or atypical
##    22|Y|Tyrosine
##    23|Z|Glu or Gln
##    24|U|Selenocysteine
##    25|*|Termination
##    26|O|Pyrrolysine
##    27|J|Leu or Ile

    fil = file(fname, "rt")
    lins = fil.readlines()
    fil.close()
    inscores = False
```

```
    matrix = []
    numCols = 0
    numRows = 0
    col = 0
    row = 0
    matrix.append([])
    #will only use the cols for the 20 amino acids
    good_cols=[1,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,22]
    for lin in lins:
        slin=lin.strip()
        if inscores==True:
            if col in good_cols: matrix[-1].append(int(slin[:-1]))
            if col<numRows-1: col+=1
            else:
                col=0
                row+=1
                if row==numCols:
                    #print "finished!"
                    break
                matrix.append([])
        else:
            if slin=="scores {": inscores =True
            elif slin[:7]=="numRows":
                txt, res=slin.split(" ")
                numRows=int(res[:-1])
                #print "N rows", numRows
            elif slin[:7]=="numColu":
                txt, res=slin.split(" ")
                numCols=int(res[:-1])
                #print "N columns", numCols
    return matrix

inpath = "ckpsSelectedPsiBlast_56878\\ScorematFiles”
if not os.path.exists(path):
    os.makedirs(path)

for fname in os.listdir(path):
    if fname.endswith(".ckp"):
        print fname
        mat = read_asn(path + "\\" + fname)
```

```
            scorematf = " ckpsSelectedPsiBlast_56878\\ScorematFiles\\"
+ fname[0:6] + ".txt"
            fil = open(scorematf , "wt")
            fil.write("A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y\n")
            i=1
            for m in mat:
                s="%d %3d" % (i,m[0])
                for c in m[1:]:
    s+=" %3d" % c
                fil.write(s+"\n")
    i+=1
            fil.close()
```

Annex 2 - GOUtils.py

based on #http://telliott99.blogspot.pt/2010/12/go-gene-ontology.html

```
import os
import pickle

def load_data(fn):
    FH = open(fn,'r')
    data = FH.read().strip()
    FH.close()
    return data

def loadGODB(fn=None):
    if not fn:
        fn = 'db/gene_ontology_ext.obo'
    #fn = 'db/short.txt'
    FH = open(fn,'r')
    data = FH.read()
    FH.close()
    L = data.strip().split('\n\n')
    D = dict()
    for e in L:
        if not '[Term]' == e[:6]:  continue
        lines = e.split('\n')
        goD = dict()
        for line in lines[1:]:
            k,v = line.split(':',1)
            v = v.strip()
            # easier if they're all lists
            if k == 'id':
                k = 'go_id'
            if k in goD:
                goD[k].append(v)
            else:
                goD[k] = [v]
        D[goD['go_id'][0]] = goD
```

```
    return D

def alt_id_match(D, go_id):
    for k, v in D.items():
        if 'alt_id' in v:
            if go_id in v['alt_id']:
                match = k
    return match

def descend(D, go_id, seen, pairs):
    seen.append(go_id)
    if go_id in D.keys():
        goD = D[go_id]
    else:
        goD = alt_id_match(D, go_id)
    if not 'is_a' in goD:
        #pairs.append((go_id,'None'))
        return pairs
    L = goD['is_a']
    L = [item.split()[0] for item in L]
    for item in L:
        pairs.append((go_id, item))
        if not item in seen:
            descend(D, item, seen, pairs)
    return pairs

def show_item(D,target):
    goD = D[target]
    print target
    for k in sorted(goD.keys()):
        print k
        for item in goD[k]:
            print '  ' + item[:50],
            if len(item) > 50:
                print '..'
            else:
                print

def handle_request(D,target,debug=False):
    if debug:
```

```
        show_item(D,target)
    pairs = list()
    seen = list()
    pairs = descend(D,target,seen,pairs)
##    if debug:
##        print len(pairs)
##        print len(list(set(pairs)))
##        for pair in pairs:
##            for item in pair:
##                print D[item]['go_id'][0], D[item]['name'][0]
##                print '-'*10
##        print '-'*50
    return pairs

def dictGOToProt(dicti, key, value):
    if not dictGOProt.has_key(key):
        dictGOProt[key] = [value]
    else:
        if value not in dictGOProt[key]:
            dictGOProt[key].append(value)
    return dicti

D = loadGODB()
queryGOFolder = "GOsCorrigido_57047_Orig+Ancs\\"
outf = "dictGOProt_Orig+Ancs.txt"
output = open(outf, "w")
dictGOProt = {}

for c,fname in enumerate(os.listdir(queryGOFolder)):
    #print str(c+1) + " --> " + str(queryGOFolder + fname)
    with open(queryGOFolder+fname, "a+") as f:
        content = f.readlines()
        f.write("\n" + "ANCESTRAIS por GOTerm:" + "\n")
        for line in content:
            target = line.split("; ")[0]
            Prot = fname.split(".")[0]
            if "GO:" in target:
                dictGOToProt(dictGOProt, target, Prot)
                #show_item(D, target)
                pairs = handle_request(D,target,debug=False)
```

```
                f.write(str(target) + "  ",)
                for t in pairs:
                    #print t
                    dictGOToProt(dictGOProt, t[0], Prot)
                    dictGOToProt(dictGOProt, t[1], Prot)
                    f.write(str(t))
                f.write("\n")

ks = dictGOProt.keys()
ks.sort()
for i in ks:
    output.write(str(i) + " --> " + str(dictGOProt[i]) + "\n")
output.close()

dictfile = "Object_dictGOProt"
with open (dictfile, "wb") as df:
    pickle.dump(dictGOProt, df)
```

Annex 3 - k-means centroid means for the T300, from left to right cluster 0 to 64

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	# PAFPs trainning	# PAFPs validation
0	0	-1	-3	-3	-3	6	0	-4	-3	-4	-1	-2	4	0	-2	0	-2	-3	-1	-1	23118	1704
1	-1	0	1	0	0	0	0	0	0	-1	0	6	0	1	0	1	0	-2	0	0	394990	15968
2	-2	-2	-1	-1	-1	-1	0	0	-1	-2	0	6	0	4	-2	1	0	-3	-1	0	54429	2486
3	-5	11	-5	-6	-2	-5	-1	-4	-6	-5	-2	-4	-4	-4	-4	-3	-4	-4	-2	0	3034	236
4	-4	-4	-2	-2	-2	-2	0	0	-2	-3	-1	6	0	5	-3	0	-1	-4	-3	-1	10323	419
5	0	3	0	0	4	0	2	0	0	0	2	0	1	0	0	0	0	0	10	4	34974	1695
6	-2	-1	1	7	-2	-2	0	-3	0	-3	-1	-1	-1	1	-1	-1	-2	-2	0	-1	75870	3465
7	-1	1	-1	-1	7	0	1	0	-1	1	1	0	0	-1	-1	-1	-1	0	3	4	149670	6867
8	-1	1	-1	-1	4	0	3	0	-1	0	1	0	0	0	0	-1	0	0	4	8	59219	2638
9	-3	-3	-2	-2	-2	-2	0	-1	-2	-2	0	6	0	6	-3	0	-1	-3	-2	-1	8792	418
10	0	0	-1	-2	-2	7	-1	-3	-2	-4	-2	0	-1	-2	-1	-1	-2	-3	0	-1	225273	10027
11	-1	0	6	2	0	0	1	-1	0	-2	0	2	0	0	0	0	0	-1	0	0	222658	10014
12	-3	0	8	0	-2	-2	0	-4	-2	-4	-1	0	-2	-1	-2	-1	-2	-4	0	-2	65980	2838
13	-5	0	-1	-3	-3	-3	0	-1	-2	-5	-2	9	-4	-1	-4	0	-2	-5	-2	-1	82116	3055
14	-3	-3	-3	0	-4	-4	1	-3	-1	-3	0	-1	0	9	-2	-2	-3	-4	-1	-2	10776	437
15	-2	11	-1	-2	-1	0	0	-2	-3	-2	-1	0	0	-2	-2	-1	-1	-1	2	0	67636	2843
16	6	2	-1	-1	-1	1	0	0	-1	-1	0	-1	0	-1	-1	1	0	0	0	-1	81244	3197
17	0	0	-1	-1	-1	-1	0	-2	-1	-2	-1	-1	8	0	-1	0	-1	-2	1	-1	178028	8315
18	-4	11	-4	-5	-3	-4	-3	-4	-5	-5	-4	-4	-3	-4	-4	-3	-4	-4	-1	-3	40649	2310
19	1	6	1	0	1	2	1	1	0	0	1	1	2	0	0	1	1	1	2	2	106084	4224
20	0	0	-1	-2	-2	0	0	-4	-2	-4	-2	0	-1	-1	-2	7	1	-3	-1	-1	23530	1073
21	-1	1	-3	-2	0	-2	0	0	-2	0	10	-1	-1	0	-2	-3	-1	0	0	0	13733	616
22	7	0	-4	-3	-3	-1	-2	-2	-3	-3	-1	-3	-2	-3	-4	0	-2	-1	-1	-3	25030	1061
23	-2	-1	0	0	-1	-2	2	-2	0	-1	1	1	1	8	0	-1	-2	-2	0	0	51304	2167
24	-1	3	-1	0	3	0	2	0	-1	0	1	0	0	0	0	-1	0	0	11	3	36990	1733
25	-4	-1	-2	-3	-2	-3	11	-4	-3	-4	-2	-1	-2	-1	-2	-3	-3	-4	-1	0	47247	2911
26	0	-1	0	1	0	0	1	0	0	0	0	2	2	6	0	0	0	-1	0	0	244239	10943
27	-2	0	-2	-2	-2	-2	1	-2	1	-2	-1	-1	-2	0	8	-2	-2	-3	0	-1	48294	2130
28	1	2	1	1	4	3	1	2	1	2	2	2	3	2	1	3	3	2	2	2	593507	26957
29	0	10	0	-1	0	1	1	0	-1	-1	0	0	0	0	-1	0	0	0	4	0	61087	2499
30	0	0	0	0	1	0	0	5	0	3	5	0	0	0	0	0	1	3	1	0	239284	10610
31	-2	0	0	-1	0	0	0	0	0	-1	0	6	0	1	-1	1	0	-2	0	0	287874	10932
32	-2	0	0	-1	1	-2	6	-2	-1	-2	0	6	-2	0	-1	1	-1	-2	0	2	871	33
33	0	0	0	0	0	6	0	-2	0	-2	0	0	0	0	0	0	-1	-1	0	0	315291	16341
34	0	0	0	0	-1	0	0	-1	6	-1	0	0	0	1	3	0	0	-1	0	0	81265	3828
35	-1	-1	0	1	-1	-1	2	-1	1	0	1	2	1	6	0	0	-1	-2	0	0	135207	6352
36	-1	-2	0	5	-3	-3	0	-3	2	-3	0	0	-2	6	0	-1	-2	-4	0	-2	4028	172
37	1	5	0	-2	-1	6	0	-2	-2	-2	-2	0	0	-1	-1	0	-1	-1	0	-1	4876	198
38	0	0	2	6	-1	0	0	-1	1	-1	0	0	0	2	0	0	0	-1	0	0	109499	5227
39	-1	0	-1	-2	-1	-2	0	-1	-1	-2	0	0	-1	-1	-1	1	7	-1	0	-1	61212	2909
40	0	9	0	0	1	1	2	0	-1	0	0	0	1	0	0	0	0	0	3	1	73019	2908
41	-3	-2	-3	-1	-3	-4	-2	-4	8	-5	-2	-2	-3	-1	2	-3	-3	-4	-1	-3	33824	1876
42	0	1	0	0	1	0	1	2	0	3	8	0	0	1	0	0	0	1	2	1	43637	2039
43	0	6	2	-1	0	0	0	-1	-1	-1	0	6	0	0	0	0	0	-1	0	-1	368	10
44	0	0	0	0	0	0	1	0	3	-1	0	0	0	1	6	0	0	-1	0	0	182876	8113
45	0	0	0	0	0	1	0	-1	0	-1	0	0	6	1	0	0	0	0	0	0	820284	38491
46	1	6	1	0	1	2	1	1	0	1	1	1	2	0	0	1	1	1	2	2	136665	5370
47	0	0	0	0	0	0	0	-1	0	-1	0	0	6	0	0	0	0	-1	0	0	269848	13069
48	1	8	1	0	1	2	1	0	0	0	1	1	2	0	0	1	1	1	2	2	75104	2954
49	0	0	1	1	2	0	6	0	1	0	0	2	1	1	2	1	0	0	1	3	79699	3637
50	-2	0	-4	-3	1	-4	-1	1	-3	6	2	-3	-2	-2	-3	-3	-2	0	0	-1	138796	7443
51	0	2	0	0	4	1	2	1	0	1	2	0	2	0	1	0	0	0	8	4	101339	4578
52	0	1	-2	-2	0	-2	-1	3	-2	0	0	-2	-1	-2	-2	-2	0	6	0	0	113649	4676
53	0	1	0	0	3	1	1	1	0	1	1	0	1	0	1	0	0	1	6	3	181024	7827
54	-2	1	-2	-2	3	-2	2	-1	-2	-1	0	-2	-2	-2	-2	-2	-2	-2	3	9	58401	2764
55	-4	0	0	-2	-2	-2	1	0	-1	-3	-1	8	-2	0	-2	0	-1	-3	-1	0	285478	10134
56	-1	0	-1	0	-1	-2	5	-2	0	-1	0	0	1	6	1	0	-1	-2	0	1	1533	79
57	0	1	2	1	2	1	6	0	0	0	1	2	1	1	2	1	0	0	2	3	51125	2406
58	-2	0	-3	-3	0	-3	-1	7	-3	1	2	-2	-1	-1	-3	-3	-1	2	1	0	49638	1968
59	0	1	0	0	4	0	2	0	0	0	1	0	0	0	0	0	0	0	3	6	201552	8964
60	0	0	1	0	2	0	8	0	0	0	1	2	1	1	2	0	0	0	2	3	75887	3550
61	-1	-2	-3	-1	-2	-1	1	-2	-2	-1	0	0	4	6	-2	0	-2	-3	-2	-1	13809	652
62	-4	0	-4	-4	9	-4	0	-2	-4	-1	0	-3	-3	-4	-3	-3	-3	-2	2	2	41745	2198
63	-3	0	-3	-2	2	-2	0	-2	-3	-1	0	-2	-2	-1	-1	-3	-2	-2	12	2	47122	2220
64	-2	2	0	0	0	-1	10	-1	-1	-2	1	1	0	1	0	-1	0	-2	2	3	26528	1211

Annex 4 - Resulting rules for T300, minimum of confidence of 40% and minimum support of 0,0005

clusters	conf	GO term	description	#rules w/ GO	hits_spdb
22,3,56	1.000	GO:0004252	serine-type endopeptidase activity	8074	3184
22,3,56	1.000	GO:0070011	peptidase activity, acting on L-amino acid peptides	8689	28392
22,3,57	1.000	GO:0008236	serine-type peptidase activity	8179	4246
27,41,9	1.000	GO:0000166	nucleotide binding	335780	227918
13,18,25,37	1.000	GO:0016740	transferase activity	100615	119322
22,3,56	1.000	GO:0017171	serine hydrolase activity	8179	4256
27,41,9	1.000	GO:0097367	carbohydrate derivative binding	262375	210010
27,41,9	1.000	GO:0032549	ribonucleoside binding	244932	103294
27,41,9	1.000	GO:0017076	purine nucleotide binding	240097	206028
27,41,9	1.000	GO:0005524	ATP binding	171435	86706
22,3,56	1.000	GO:0016787	hydrolase activity	37554	117856
27,41,9	1.000	GO:0032559	adenyl ribonucleotide binding	172212	87776
27,41,9	1.000	GO:0032555	purine ribonucleotide binding	237491	102681
27,41,9	1.000	GO:0032553	ribonucleotide binding	244265	104903
27,41,9	1.000	GO:0035639	purine ribonucleoside triphosphate binding	235635	101195
22,3,56	1.000	GO:0008233	peptidase activity	8690	18134
39,41,9	1.000	GO:0043168	anion binding	396053	123883
14,22,23,3	1.000	GO:0043169	cation binding	102974	123562
27,41,9	1.000	GO:0030554	adenyl nucleotide binding	174173	88416
22,3,56	1.000	GO:0004175	endopeptidase activity	8334	9808
27,41,9	1.000	GO:0036094	small molecule binding	380270	234291
27,41,9	1.000	GO:0001883	purine nucleoside binding	236592	102352
27,41,9	1.000	GO:0001882	nucleoside binding	245604	205732

27,41,9	1.000	GO:1901265	nucleoside phosphate binding	335797	226051
27,41,9	1.000	GO:0032550	purine ribonucleoside binding	236347	102321
22,29,3,41	0.982	GO:0003676	nucleic acid binding	28309	136633
14,22,29,3	0.979	GO:0043226	organelle	554651	274732
14,22,23,3	0.974	GO:0080090	regulation of primary metabolic process	30639	63857
14,22,23,3	0.974	GO:0019222	regulation of metabolic process	34153	344577
14,22,23,3	0.974	GO:0060255	regulation of macromolecule metabolic process	29583	186498
14,22,23,3	0.974	GO:0043229	intracellular organelle	514435	126687
14,22,23,3	0.974	GO:0003677	DNA binding	23361	51904
14,22,23,3	0.974	GO:0050789	regulation of biological process	256418	474851
14,22,23,3	0.974	GO:0065007	biological regulation	336799	243778
14,22,23,3	0.974	GO:0050794	regulation of cellular process	203467	283128
14,22,23,3	0.974	GO:0031323	regulation of cellular metabolic process	31102	165995
14,4,46,8	0.970	GO:0043170	macromolecule metabolic process	260976	63937
22,24,25,61	0.969	GO:0044710	single-organism metabolic process	415531	504592
32,53,56,62	0.967	GO:0016491	oxidoreductase activity	2806	61780
16,3,42,64	0.959	GO:0043231	intracellular membrane-bounded organelle	423654	70265
16,3,42,64	0.959	GO:0043227	membrane-bounded organelle	522330	89232
14,22,23,3	0.949	GO:2001141	regulation of RNA biosynthetic process	24517	52267
14,22,23,3	0.949	GO:0009889	regulation of biosynthetic process	25935	113533
14,22,23,3	0.949	GO:0006355	regulation of transcription, DNA-templated	24482	43643
14,22,23,3	0.949	GO:0010556	regulation of macromolecule biosynthetic process	25241	98664
14,22,23,3	0.949	GO:0051171	regulation of nitrogen compound metabolic process	25524	69314
14,22,23,3	0.949	GO:0051252	regulation of RNA metabolic process	24550	53761
14,22,23,3	0.949	GO:0031326	regulation of cellular biosynthetic process	25749	101647
14,22,23,3	0.949	GO:2000112	regulation of cellular macromolecule biosynthetic process	25059	52255
14,22,23,3	0.949	GO:0010468	regulation of gene expression	25711	68926

14,22,23,3	0.949	GO:0019219	regulation of nucleobase-containing compound metabolic process	25404	60292
14,22,23,3	0.949	GO:1903506	regulation of nucleic acid-templated transcription	24483	50995
14,21,61,62	0.947	GO:0006807	nitrogen compound metabolic process	203695	498947
14,34,4,8	0.944	GO:0044260	cellular macromolecule metabolic process	175711	345895
14,25,5,61	0.939	GO:0034641	cellular nitrogen compound metabolic process	91770	503729
22,24,25,61	0.938	GO:0044281	small molecule metabolic process	75661	263832
18,2,5,63	0.937	GO:0044425	membrane part	57118	139234
14,18,37,41	0.935	GO:0008104	protein localization	1844	21940
14,18,37,41	0.935	GO:0033036	macromolecule localization	1866	24034
14,18,37,41	0.935	GO:0006810	transport	3414	109392
14,18,37,41	0.935	GO:0051234	establishment of localization	3507	105802
14,18,37,41	0.935	GO:0071702	organic substance transport	1999	32732
14,18,37,41	0.935	GO:0051179	localization	3863	148560
16,3,42,64	0.932	GO:0005634	nucleus	27105	32457
18,37,41,63	0.932	GO:0016020	membrane	146716	117656
18,5,58,61	0.921	GO:0043436	oxoacid metabolic process	29483	99722
18,5,58,61	0.921	GO:0006082	organic acid metabolic process	29774	160458
18,5,58,61	0.921	GO:0019752	carboxylic acid metabolic process	28619	168107
14,34,4,8	0.917	GO:1901360	organic cyclic compound metabolic process	100719	418474
14,34,4,8	0.917	GO:0046483	heterocycle metabolic process	75464	401433
14,34,4,8	0.917	GO:0006139	nucleobase-containing compound metabolic process	50339	273483
14,34,4,8	0.917	GO:0006725	cellular aromatic compound metabolic process	77983	370913
14,34,4,8	0.917	GO:0090304	nucleic acid metabolic process	26791	129108
18,2,23,63	0.915	GO:0022857	transmembrane transporter activity	2301	72887
18,2,23,63	0.915	GO:0005230	extracellular ligand-gated ion channel activity	1034	538
18,2,23,63	0.915	GO:0022892	substrate-specific transporter activity	2447	35964
18,2,23,63	0.915	GO:0015276	ligand-gated ion channel activity	1034	1089

18,2,23,63	0.915	GO:0022891	substrate-specific transmembrane transporter activity	2259	39901
18,2,23,63	0.915	GO:0005215	transporter activity	2479	78367
18,2,23,63	0.915	GO:0005216	ion channel activity	1099	5430
18,2,23,63	0.915	GO:0022836	gated channel activity	1036	2310
18,2,23,63	0.915	GO:0015075	ion transmembrane transporter activity	2247	43102
18,2,23,63	0.915	GO:0022838	substrate-specific channel activity	1099	3964
18,2,23,63	0.915	GO:0022834	ligand-gated channel activity	1034	961
18,2,23,63	0.915	GO:0015267	channel activity	1099	7007
18,2,23,63	0.915	GO:0022803	passive transmembrane transporter activity	1099	4756
14,22,3,8	0.913	GO:0016070	RNA metabolic process	22634	129759
22,29,3,41	0.911	GO:1901362	organic cyclic compound biosynthetic process	23042	154687
22,29,3,41	0.911	GO:0044249	cellular biosynthetic process	78905	613402
22,29,3,41	0.911	GO:0019438	aromatic compound biosynthetic process	22210	125786
22,29,3,41	0.911	GO:0018130	heterocycle biosynthetic process	22563	147976
22,29,3,41	0.911	GO:0097659	nucleic acid-templated transcription	21369	29239
22,29,3,41	0.911	GO:1901576	organic substance biosynthetic process	94600	587283
22,29,3,41	0.911	GO:0034654	nucleobase-containing compound biosynthetic process	21569	100448
22,29,3,41	0.911	GO:0044271	cellular nitrogen compound biosynthetic process	23277	196364
22,29,3,41	0.911	GO:0006351	transcription,DNA-templated	21369	29372
22,29,3,41	0.911	GO:0032774	RNA biosynthetic process	21426	33819
22,29,3,41	0.911	GO:0034645	cellular macromolecule biosynthetic process	21500	102282
22,29,3,41	0.911	GO:0009058	biosynthetic process	108948	649027
22,29,3,41	0.911	GO:0009059	macromolecule biosynthetic process	21605	147235
13,14,35,37	0.887	GO:0005525	GTP binding	1560	14478
13,14,35,37	0.887	GO:0032561	guanyl ribonucleotide binding	1560	14895
13,14,35,37	0.887	GO:0019001	guanyl nucleotide binding	1560	14909
18,2,23,63	0.881	GO:0044459	plasma membrane part	1355	20187

14,18,24,56	0.879	GO:0016021	integral component of membrane	18712	74452
14,18,24,56	0.879	GO:0031224	intrinsic component of membrane	20140	84772
14,24,32,62	0.878	GO:1901564	organonitrogen compound metabolic process	15545	452635
13,14,35,37	0.876	GO:0007165	signal transduction	13704	33581
13,14,35,37	0.876	GO:0035556	intracellular signal transduction	1852	9115
13,14,35,37	0.876	GO:0007264	small GTPase mediated signal transduction	1541	2353
13,32,41,63	0.868	GO:0016772	transferase activity, transferring phosphorus-containing groups	28123	69940
14,15,36,63	0.865	GO:0006520	cellular amino acid metabolic process	6638	125141
18,2,23,63	0.864	GO:0097060	synaptic membrane	905	1119
18,2,23,63	0.864	GO:0030054	cell junction	1008	7169
18,2,23,63	0.864	GO:0044456	synapse part	913	3213
18,2,23,63	0.864	GO:0045211	postsynaptic membrane	904	855
22,42,43,5	0.861	GO:0044444	cytoplasmic part	501601	170466
2,35,36,51	0.857	GO:0044422	organelle part	19157	140661
2,35,36,51	0.857	GO:0044446	intracellular organelle part	15080	115940
13,14,35,37	0.835	GO:0045184	establishment of protein localization	1663	21083
13,14,35,37	0.835	GO:0015031	protein transport	1620	19388
14,24,32,41	0.829	GO:0016874	ligase activity	1359	34021
13,32,41,63	0.827	GO:0016301	kinase activity	20584	25294
18,5,56,62	0.824	GO:0046914	transition metal ion binding	2191	46452
13,32,41,63	0.823	GO:0004672	protein kinase activity	17241	10241
13,32,41,63	0.823	GO:0016773	phosphotransferase activity, alcohol group as acceptor	18848	17316
18,5,56,62	0.813	GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	794	3515
18,56,62,63	0.800	GO:0020037	heme binding	857	5455
18,56,62,63	0.800	GO:0046906	tetrapyrrole binding	869	6995

18,56,62,63	0.800	GO:0005506	iron ion binding	851	12817
18,29,43,8	0.789	GO:0032502	developmental process	21965	103077
18,5,56,62	0.780	GO:0004497	monooxygenase activity	717	3595
18,2,23,31	0.778	GO:0043234	protein complex	4591	83868
18,2,23,31	0.778	GO:0032991	macromolecular complex	6564	127477
13,36,41,63	0.764	GO:0004674	protein serine/threonine kinase activity	7305	5300
13,18,40,63	0.758	GO:0060089	molecular transducer activity	3018	20818
13,18,40,63	0.751	GO:0004871	signal transducer activity	2658	11912
13,18,40,63	0.740	GO:0004872	receptor activity	2263	9869
13,18,40,63	0.733	GO:0038023	signaling receptor activity	1978	10151
13,48,61,8	0.732	GO:0044711	single-organism biosynthetic process	12695	181296
18,2,23,63	0.729	GO:1902495	transmembrane transporter complex	468	6003
18,2,23,63	0.729	GO:0034702	ion channel complex	465	1196
18,2,23,63	0.729	GO:1990351	transporter complex	468	6069
18,2,23,63	0.729	GO:0098796	membrane protein complex	488	32432
13,18,40,63	0.725	GO:0004888	transmembrane signaling receptor activity	1826	6982
14,2,41,46	0.723	GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	1826	40750
14,2,41,46	0.723	GO:0016817	hydrolase activity, acting on acid anhydrides	1846	54734
14,2,41,46	0.723	GO:0017111	nucleoside-triphosphatase activity	1716	37367
14,2,41,46	0.723	GO:0016462	pyrophosphatase activity	1824	39976
21,4,41,6	0.710	GO:0016829	lyase activity	1552	26792
18,37,41,63	0.705	GO:0005886	plasma membrane	6125	54764
13,15,32,41	0.697	GO:0006796	phosphate-containing compound metabolic process	10832	133560
13,15,32,41	0.697	GO:0006793	phosphorus metabolic process	12005	158827
10,11,3,37	0.694	GO:0044767	single-organism developmental process	17492	61910
13,24,58,61	0.667	GO:0050896	response to stimulus	20093	173170
21,4,41,6	0.661	GO:1901605	alpha-amino acid metabolic process	1152	86605

18,2,23,63	0.661	GO:0044765	single-organism transport	730	86855
18,2,23,63	0.661	GO:1902578	single-organism localization	830	54277
13,32,41,48	0.660	GO:0019538	protein metabolic process	10963	133686
23,43,58,63	0.659	GO:1901575	organic substance catabolic process	396	74400
23,43,58,63	0.659	GO:0009056	catabolic process	487	103425
13,18,32,63	0.657	GO:0004930	G-protein coupled receptor activity	382	3746
13,41,56,64	0.652	GO:0044283	small molecule biosynthetic process	1902	63376
13,41,56,64	0.652	GO:1901566	organonitrogen compound biosynthetic process	629	197213
1,23,26,43	0.649	GO:0016043	cellular component organization	1522	92937
1,23,26,43	0.649	GO:0071840	cellular component organization or biogenesis	1770	70229
18,2,23,63	0.644	GO:0006811	ion transport	297	38971
15,18,32,63	0.640	GO:0007166	cell surface receptor signalling pathway	1225	9831
3,37,51	0.638	GO:0044421	extracellular region part	1474	16396
13,15,32,41	0.634	GO:0044267	cellular protein metabolic process	5754	90302
13,41,56,64	0.630	GO:0016053	organic acid biosynthetic process	637	49266
13,41,56,64	0.630	GO:0046394	carboxylic acid biosynthetic process	637	51592
17,18,29,43	0.627	GO:0048856	anatomical structure development	793	26989
18,2,23,63	0.627	GO:0055085	transmembrane transport	261	21935
18,2,23,63	0.627	GO:0034220	ion transmembrane transport	259	27627
24,3,31,57	0.625	GO:0048518	positive regulation of biological process	21368	121108
14,3,42,46	0.620	GO:0001071	nucleic acid binding transcription factor activity	13912	12267
13,15,32,41	0.614	GO:0043412	macromolecule modification	4139	44035
13,15,32,41	0.614	GO:0036211	protein modification process	3374	25841
13,15,32,41	0.614	GO:0006464	cellular protein modification process	3374	33228
13,15,32,41	0.607	GO:0016310	phosphorylation	1916	11853
13,15,32,41	0.607	GO:0006468	protein phosphorylation	1717	5318
14,3,42,46	0.600	GO:0003700	sequence-specific DNA binding transcription factor activity	13624	12557

22,43,57,63	0.596	GO:0016788	hydrolase activity, acting on ester bonds	391	33881
19,3,31,53	0.590	GO:0048519	negative regulation of biological process	18300	94328
24,3,41,5	0.583	GO:0048522	positive regulation of cellular process	12784	55121
23,24,40,61	0.574	GO:0016866	intramolecular transferase activity	114	5392
23,24,40,61	0.574	GO:0016853	isomerase activity	114	16867
11,3,37,53	0.566	GO:0005509	calcium ion binding	40	4360
18,32,35,5	0.558	GO:0006629	lipid metabolic process	1502	66824
22,43,57,63	0.558	GO:0052689	carboxylic ester hydrolase activity	28	5834
19,41,5,61	0.558	GO:0048037	cofactor binding	221	30293
14,19,2,41	0.558	GO:0016887	ATPase activity	76	20999
16,23,26,43	0.558	GO:0005975	carbohydrate metabolic process	154	63869
19,3,31,53	0.557	GO:0048523	negative regulation of cellular process	13555	44395
1,18,29,43	0.552	GO:0044707	single-multicellular organism process	475	23343
1,18,29,43	0.552	GO:0032501	multicellular organismal process	490	22304
21,4,41,6	0.548	GO:1901607	alpha-amino acid biosynthetic process	214	31198
21,4,41,6	0.548	GO:0008652	cellular amino acid biosynthetic process	218	49918
13,14,35,37	0.546	GO:0051649	establishment of localization in cell	327	18663
24,3,31,5	0.540	GO:0010604	positive regulation of macromolecule metabolic process	761	39472
24,3,31,5	0.540	GO:0009893	positive regulation of metabolic process	1145	71866
14,24,31,32	0.534	GO:0019637	organophosphate metabolic process	35	176192
23,48,52,61	0.532	GO:0016627	oxidoreductase activity, acting on the CH-CH group of donors	5	3704
[16,18,25,32	0.527	GO:0005737	cytoplasm	207	147574
3,31,55,64	0.525	GO:0010605	negative regulation of macromolecule metabolic process	2640	37082
3,31,55,64	0.525	GO:0010629	negative regulation of gene expression	1347	13736
3,31,55,64	0.525	GO:0009892	negative regulation of metabolic process	3048	60060
18,37,40,54	0.524	GO:0005515	protein binding	175	33855

23,24,48,61	0.523	GO:0044255	cellular lipid metabolic process	518	41954
14,23,46,61	0.522	GO:0008270	zinc ion binding	60	23177
1,13,14,37	0.520	GO:0003924	GTPase activity	135	8977
23,48,52,61	0.519	GO:0050662	coenzyme binding	12	18720
22,3,49,56	0.517	GO:0005576	extracellular region	482	24021
3,31,55,64	0.515	GO:0051253	negative regulation of RNA metabolic process	1079	9253
3,31,55,64	0.515	GO:0009890	negative regulation of biosynthetic process	1149	21204
3,31,55,64	0.515	GO:0010558	negative regulation of macromolecule biosynthetic process	1149	18772
3,31,55,64	0.515	GO:1903507	negative regulation of nucleic acid-templated transcription	1079	8729
3,31,55,64	0.515	GO:0031327	negative regulation of cellular biosynthetic process	1149	19231
3,31,55,64	0.515	GO:0031324	negative regulation of cellular metabolic process	2056	30229
3,31,55,64	0.515	GO:2000113	negative regulation of cellular macromolecule biosynthetic process	913	9765
3,31,55,64	0.515	GO:0045934	negative regulation of nucleobase-containing compound metabolic process	1123	10339
3,31,55,64	0.515	GO:1902679	negative regulation of RNA biosynthetic process	1079	8800
3,31,55,64	0.515	GO:0051172	negative regulation of nitrogen compound metabolic process	1123	10984
3,31,55,64	0.515	GO:0045892	negative regulation of transcription, DNA-templated	913	8813
23,24,40,61	0.508	GO:0031559	oxidosqualene cyclase activity	29	44
23,48,52,61	0.506	GO:0050660	flavin adenine dinucleotide binding	4	4117
24,3,41,46	0.500	GO:0006357	regulation of transcription from RNA polymerase II promoter	6674	9976
13,14,21,37	0.500	GO:0046907	intracellular transport	173	26533
24,3,41,5	0.500	GO:0010628	positive regulation of gene expression	260	10847
24,48,61,63	0.500	GO:0008610	lipid biosynthetic process	45	31347

13,18,32,63	0.495	GO:0007186	G-protein coupled receptor signalling pathway	54	5285
23,48,52,61	0.494	GO:0044248	cellular catabolic process	16	66633
18,24,29,32	0.492	GO:0048583	regulation of response to stimulus	18	67838
18,2,23,63	0.492	GO:0044700	single organism signaling	65	2394
18,2,23,63	0.492	GO:0005231	excitatory extracellular ligand-gated ion channel activity	65	266
18,2,23,63	0.492	GO:0023052	signaling	65	2412
18,2,23,63	0.492	GO:0007268	synaptic transmission	65	1254
18,2,23,63	0.492	GO:0007154	cell communication	68	9092
18,2,23,63	0.492	GO:0007267	cell-cell signaling	65	2202
24,3,31,57	0.486	GO:0031325	positive regulation of cellular metabolic process	414	39512
13,18,29,35	0.484	GO:0004713	protein tyrosine kinase activity	5	1012
21,4,41,6	0.484	GO:0009069	serine family amino acid metabolic process	40	10408
23,48,52,61	0.481	GO:0003995	acyl-CoA dehydrogenase activity	4	225
23,48,52,61	0.468	GO:0044282	small molecule catabolic process	3	13608
23,48,52,61	0.468	GO:0044712	single-organism catabolic process	5	31131
18,5,56,62	0.462	GO:0031090	organelle membrane	15	37243
0,15,41,63	0.460	GO:0006950	response to stress	23	57048
24,3,31,57	0.458	GO:0009891	positive regulation of biosynthetic process	74	22551
24,3,31,57	0.458	GO:0010557	positive regulation of macromolecule biosynthetic process	74	10529
24,3,31,57	0.458	GO:0031328	positive regulation of cellular biosynthetic process	74	11418
23,48,52,61	0.456	GO:0046395	carboxylic acid catabolic process	2	11044
23,48,52,61	0.456	GO:0016054	organic acid catabolic process	2	10843
24,3,31,41	0.447	GO:0045893	positive regulation of transcription, DNA-templated	52	8417
24,3,31,41	0.447	GO:0051254	positive regulation of RNA metabolic process	52	8550
24,3,31,41	0.447	GO:1902680	positive regulation of RNA biosynthetic process	52	8112
24,3,31,41	0.447	GO:1903508	positive regulation of nucleic acid-templated	52	8026

			transcription		
24,3,31,42	0.447	GO:0045935	positive regulation of nucleobase-containing compound metabolic process	54	9903
24,3,31,42	0.447	GO:0051173	positive regulation of nitrogen compound metabolic process	54	11460
10,29,43,60	0.446	GO:0051246	regulation of protein metabolic process	12	39461
3,37,45,53	0.446	GO:0050793	regulation of developmental process	94	41651
13,18,32,63	0.444	GO:0065008	regulation of biological quality	34	36463
13,14,35,37	0.443	GO:0016192	vesicle-mediated transport	26	7563
3,37,45,53	0.435	GO:0051239	regulation of multicellular organismal process	95	34531
13,21,23,36	0.429	GO:1901135	carbohydrate derivative metabolic process	2	172418
0,14,22,63	0.426	GO:0044428	nuclear part	5	30031
3,31,55,64	0.426	GO:0000122	negative regulation of transcription from RNA polymerase II promoter	13	2585
10,3,37,45	0.425	GO:0009653	anatomical structure morphogenesis	28	15056
15,29,32,41	0.425	GO:0005829	cytosol	12	14482
13,14,35,37	0.423	GO:0031982	vesicle	6	23624
29,43,59,60	0.420	GO:0032268	regulation of cellular protein metabolic process	5	30538
18,23,40,56	0.420	GO:0006508	proteolysis	11	8560
18,29,31,37	0.419	GO:0048869	cellular developmental process	2	19490
13,14,21,37	0.417	GO:0098588	bounding membrane of organelle	6	17778
13,14,35,37	0.412	GO:0031988	membrane-bounded vesicle	2	11370
14,15,36,60	0.408	GO:0006259	DNA metabolic process	3	40685
18,29,37,38	0.407	GO:0022610	biological adhesion	1	6260
18,29,37,38	0.407	GO:0007155	cell adhesion	1	6508
18,23,24,56	0.406	GO:0005615	extracellular space	1	5564
14,15,36,48	0.403	GO:0006996	organelle organization	1	37867

Annex 5 - Snippet of code used to run over Swiss-Prot (uniprot_sprot.dat) one protein annotation at a time, to identify experimentally annotated ones

```
    anot = ""
    for textlin in open(SwissProtTxt):
        fimbloco = re.findall("^//",textlin)
        if len(fimbloco) > 0:
            anot += textlin
            m =
re.findall('.*EXP:.*|.*IDA:.*|.*IPI:.*|.*IMP:.*|.*IGI:.*|.*IEP:.*',
anot)
            add = False
            if len(m) != 0:
                for GO in m:
                    if not add and GO[0:8] == 'DR   GO;':
                        add = True
            if add:
                selectedSeqs(anot[60:75].split(';')[0])
            anot = ""
            #print anot
        else:
            anot += textlin
```

Annex 6 - Most frequent, in decreasing order, ancestral GO terms in the 57047 training set; these were obtained via the code snippet at Annex 2 – GOUtils.py. Only GO terms representing over 10000 proteins are included.

GO term	count	Name (or association)
GO:0008150	52940	biological_process
GO:0005575	52495	cellular_component
GO:0044464	48439	cell part
GO:0003674	47136	molecular_function
GO:0009987	43250	cellular process
GO:0044424	41792	intracellular part
GO:0044699	37666	single-organism process
GO:0005488	34622	binding
GO:0008152	32251	metabolic process
GO:0043226	31215	organelle
GO:0044763	31049	single-organism cellular process
GO:0071704	29874	organic substance metabolic process
GO:0044237	29145	cellular metabolic process
GO:0043229	28855	intracellular organelle
GO:0043227	28322	membrane-bounded organelle
GO:0044238	27633	primary metabolic process
GO:0065007	27059	biological regulation
GO:0043231	25941	intracellular membrane-bounded organelle
GO:0044444	25434	cytoplasmic part
GO:0050789	25110	regulation of biological process
GO:0003824	23953	catalytic activity
GO:0050794	23309	regulation of cellular process
GO:0044422	21372	organelle part
GO:0043170	20871	macromolecule metabolic process
GO:0044446	20798	intracellular organelle part
GO:0043167	20521	ion binding
GO:0097159	20296	organic cyclic compound binding
GO:1901363	20124	heterocyclic compound binding
GO:0044260	19319	cellular macromolecule metabolic process

GO:0016020	18882	membrane
GO:0006807	17061	nitrogen compound metabolic process
GO:0044425	16223	membrane part
GO:0044710	16119	single-organism metabolic process
GO:0050896	15657	response to stimulus
GO:1901360	15444	organic cyclic compound metabolic process
GO:0005634	15369	nucleus
GO:0032991	15232	macromolecular complex
GO:0034641	15112	cellular nitrogen compound metabolic process
GO:0019222	15086	regulation of metabolic process
GO:0009058	15025	biosynthetic process
GO:0006725	14587	cellular aromatic compound metabolic process
GO:1901576	14528	organic substance biosynthetic process
GO:0046483	14401	heterocycle metabolic process
GO:0071840	14293	cellular component organization or biogenesis
GO:0044249	14155	cellular biosynthetic process
GO:0032502	13957	developmental process
GO:0016043	13903	cellular component organization
GO:0031323	13562	regulation of cellular metabolic process
GO:0060255	13267	regulation of macromolecule metabolic process
GO:0080090	13255	regulation of primary metabolic process
GO:0006139	13186	nucleobase-containing compound metabolic process
GO:0044767	13045	single-organism developmental process
GO:0005737	12809	cytoplasm
GO:0043169	12578	cation binding
GO:0046872	12348	metal ion binding
GO:0031224	12337	intrinsic component of membrane
GO:0043234	12275	protein complex
GO:0051179	11931	localization
GO:0048518	11881	positive regulation of biological process
GO:0016021	11829	integral component of membrane
GO:0090304	11088	nucleic acid metabolic process
GO:0003676	10997	nucleic acid binding
GO:0036094	10982	small molecule binding
GO:0005515	10868	protein binding
GO:0043168	10844	anion binding

GO:0051234	10783	establishment of localization
GO:0048519	10728	negative regulation of biological process
GO:0006810	10439	transport
GO:0019538	10295	protein metabolic process
GO:0009889	10277	regulation of biosynthetic process
GO:0048522	10274	positive regulation of cellular process
GO:1901265	10163	nucleoside phosphate binding
GO:0000166	10162	nucleotide binding
GO:0031326	10153	regulation of cellular biosynthetic process
GO:0010468	10064	regulation of gene expression
GO:0006950	10039	response to stress

Annex 7 - Description, level, information content (obtained from [55]) and representation in Swiss-Prot of the 280 GO terms selected by the T300 with a support of 0.0005 and a confidence of 40%

GO	Name	level	IC	# SP prot w/ GO	% SP prot w/ GO
GO:0004252	serine-type endopeptidase activity	6	7,06	3184	0,58
GO:0004497	monooxygenase activity	3	6,79	3595	0,65
GO:0008104	protein localization	3	7,67	21940	4
GO:0080090	regulation of primary metabolic process	4	6,57	63857	11,63
GO:0019222	regulation of metabolic process	3	4,15	344577	62,76
GO:0044444	cytoplasmic part	6	3,68	170466	31,05
GO:2000113	negative regulation of cellular macromolecule biosynthetic process	7	10	9765	1,78
GO:0060089	molecular transducer activity	1	4,98	20818	3,79
GO:0007165	signal transduction	4	7,25	33581	6,12
GO:0007166	cell surface receptor signalling pathway	5	9,54	9831	1,79
GO:0044283	small molecule biosynthetic process	4	6,89	63376	11,54
GO:0016866	intramolecular transferase activity	3	6,57	5392	0,98
GO:1901362	organic cyclic compound biosynthetic process	4	5,65	154687	28,18
GO:0044707	single-multicellular organism process	2	7,46	23343	4,25
GO:0030054	cell junction	1	8,39	7169	1,31
GO:0097060	synaptic membrane	6	10,49	1119	0,2
GO:0044710	single-organism metabolic process	2	3,05	504592	91,91
GO:0044711	single-organism biosynthetic process	3	5,67	181296	33,02
GO:0010605	negative regulation of macromolecule metabolic process	5	8,38	37082	6,75
GO:0031982	vesicle	2	6,73	23624	4,3
GO:0048869	cellular developmental process	3	8,41	19490	3,55
GO:0016021	integral component of membrane	4	3,92	74452	13,56
GO:0016829	lyase activity	2	4,55	26792	4,88
GO:0031988	membrane-bounded vesicle	3	7,47	11370	2,07
GO:0016740	transferase activity	2	2,55	119322	21,73
GO:0004713	protein tyrosine kinase activity	6	8,86	1012	0,18
GO:1901575	organic substance catabolic process	3	4,53	74400	13,55
GO:0048518	positive regulation of biological process	3	7,55	121108	22,06
GO:0048519	negative regulation of biological process	3	6,71	94328	17,18
GO:0033036	macromolecule localization	2	7,33	24034	4,38
GO:1902495	transmembrane transporter complex	5	8,07	6003	1,09
GO:0017171	serine hydrolase activity	3	6,36	4256	0,78
GO:0060255	regulation of macromolecule metabolic process	4	5,43	186498	33,97
GO:0098588	bounding membrane of organelle	4	7,72	17778	3,24
GO:0004672	protein kinase activity	5	5,55	10241	1,87
GO:0003924	GTPase activity	7	6,76	8977	1,64

GO:0045184	establishment of protein localization	4	8,07	21083	3,84
GO:0007155	cell adhesion	3	9,86	6508	1,19
GO:2001141	regulation of RNA biosynthetic process	7	7,06	52267	9,52
GO:0043436	oxoacid metabolic process	5	5,53	99722	18,16
GO:0043231	intracellular membrane-bounded organelle	6	4,15	70265	12,8
GO:0043234	protein complex	2	4,89	83868	15,28
GO:0044700	single organism signaling	2	7,24	2394	0,44
GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	4	3,84	40750	7,42
GO:0016192	vesicle-mediated transport	4	8,79	7563	1,38
GO:1901566	organonitrogen compound biosynthetic process	4	6,13	197213	35,92
GO:0031559	oxidosqualene cyclase activity	4	13,72	44	0,01
GO:0019538	protein metabolic process	4	6,31	133686	24,35
GO:0005829	cytosol	7	7,44	14482	2,64
GO:0016817	hydrolase activity, acting on acid anhydrides	3	3,65	54734	9,97
GO:0097367	carbohydrate derivative binding	2	2,77	210010	38,25
GO:0044428	nuclear part	8	6,42	30031	5,47
GO:0006259	DNA metabolic process	6	6,96	40685	7,41
GO:0003995	acyl-CoA dehydrogenase activity	4	8	225	0,04
GO:0019438	aromatic compound biosynthetic process	4	5,74	125786	22,91
GO:0018130	heterocycle biosynthetic process	4	5,67	147976	26,95
GO:0051253	negative regulation of RNA metabolic process	7	10,02	9253	1,69
GO:0044422	organelle part	2	3,89	140661	25,62
GO:0001071	nucleic acid binding transcription factor activity	1	5,03	12267	2,23
GO:0017076	purine nucleotide binding	5	3,18	206028	37,53
GO:0009892	negative regulation of metabolic process	4	7,4	60060	10,94
GO:0009893	positive regulation of metabolic process	4	8,25	71866	13,09
GO:0009890	negative regulation of biosynthetic process	5	8,68	21204	3,86
GO:0009891	positive regulation of biosynthetic process	5	9,8	22551	4,11
GO:0006950	response to stress	2	7,51	57048	10,39
GO:0051254	positive regulation of RNA metabolic process	7	10,75	8550	1,56
GO:0043229	intracellular organelle	5	3,27	126687	23,08
GO:0010628	positive regulation of gene expression	6	10,72	10847	1,98
GO:0034702	ion channel complex	6	9,6	1196	0,22
GO:0003676	nucleic acid binding	3	3,28	136633	24,89
GO:0003677	DNA binding	4	3,87	51904	9,45
GO:0035556	intracellular signal transduction	5	8,65	9115	1,66
GO:0016772	transferase activity, transferring phosphorus-containing groups	3	3,43	69940	12,74
GO:0043227	membrane-bounded organelle	2	3,71	89232	16,25
GO:0050789	regulation of biological process	2	3,66	474851	86,49
GO:0097659	nucleic acid-templated transcription			29239	
GO:1901576	organic substance biosynthetic process	3	4,37	587283	106,97
GO:0004930	G-protein coupled receptor activity	5	7,37	3746	0,68
GO:0046914	transition metal ion binding	5	4,59	46452	8,46
GO:0006357	regulation of transcription from RNA polymerase II promoter	9	10,62	9976	1,82
GO:0046483	heterocycle metabolic process	3	3,07	401433	73,12

GO:0016043	cellular component organization	2	6,73	92937	16,93
GO:0065007	biological regulation	1	3,61	243778	44,4
GO:0071840	cellular component organization or biogenesis	1	6,3	70229	12,79
GO:0005886	plasma membrane	4	4,56	54764	9,98
GO:0032549	ribonucleoside binding	4	3,88	103294	18,81
GO:0022803	passive transmembrane transporter activity	3	6,35	4756	0,87
GO:0065008	regulation of biological quality	2	6,88	36463	6,64
GO:0007186	G-protein coupled receptor signaling pathway	6	10,02	5285	0,96
GO:0005230	extracellular ligand-gated ion channel activity	8	9,29	538	0,1
GO:0005524	ATP binding	8	4,01	86706	15,79
GO:0005525	GTP binding	8	6,12	14478	2,64
GO:0016787	hydrolase activity	2	2,57	117856	21,47
GO:0006810	transport	3	5,12	109392	19,93
GO:0006629	lipid metabolic process	3	6,87	66824	12,17
GO:0020037	heme binding	4	5,97	5455	0,99
GO:0006139	nucleobase-containing compound metabolic process	4	3,43	273483	49,81
GO:0050793	regulation of developmental process	3	8,69	41651	7,59
GO:0009889	regulation of biosynthetic process	4	5,53	113533	20,68
GO:0006811	ion transport	5	6,57	38971	7,1
GO:0005231	excitatory extracellular ligand-gated ion channel activity	9	10,29	266	0,05
GO:1990351	transporter complex	3	8,07	6069	1,11
GO:0016788	hydrolase activity, acting on ester bonds	3	4,26	33881	6,17
GO:0050794	regulation of cellular process	3	5,2	283128	51,57
GO:0019637	organophosphate metabolic process	4	3,71	176192	32,09
GO:0043412	macromolecule modification	4	7,07	44035	8,02
GO:0036211	protein modification process	5	7,99	25841	4,71
GO:0051239	regulation of multicellular organismal process	3	9,45	34531	6,29
GO:0044425	membrane part	2	3,02	139234	25,36
GO:0034654	nucleobase-containing compound biosynthetic process	5	5,99	100448	18,3
GO:0044282	small molecule catabolic process	4	8,8	13608	2,48
GO:0051234	establishment of localization	2	5,02	105802	19,27
GO:0010604	positive regulation of macromolecule metabolic process	5	9,33	39472	7,19
GO:0022891	substrate-specific transmembrane transporter activity	3	3,65	39901	7,27
GO:0008652	cellular amino acid biosynthetic process	8	7,18	49918	9,09
GO:0022892	substrate-specific transporter activity	2	3,6	35964	6,55
GO:0044271	cellular nitrogen compound biosynthetic process	4	5,68	196364	35,77
GO:0046907	intracellular transport	4	8,35	26533	4,83
GO:0046906	tetrapyrrole binding	3	5,9	6995	1,27
GO:0016773	phosphotransferase activity, alcohol group as acceptor	4	4,96	17316	3,15
GO:0046395	carboxylic acid catabolic process	7	8,95	11044	2,01
GO:0050896	response to stimulus	1	5,4	173170	31,54
GO:0016301	kinase activity	4	4,43	25294	4,61
GO:0009058	biosynthetic process	2	3,51	649027	118,22
GO:0015075	ion transmembrane transporter activity	4	3,96	43102	7,85
GO:0019001	guanyl nucleotide binding	6	6,08	14909	2,72

GO:0010557	positive regulation of macromolecule biosynthetic process	6	10,53	10529	1,92
GO:0015267	channel activity	4	6,35	7007	1,28
GO:0032559	adenyl ribonucleotide binding	7	4	87776	15,99
GO:0006351	transcription, DNA-templated	8	7,38	29372	5,35
GO:0032555	purine ribonucleotide binding	6	3,89	102681	18,7
GO:0010558	negative regulation of macromolecule biosynthetic process	6	9,39	18772	3,42
GO:0032553	ribonucleotide binding	5	3,85	104903	19,11
GO:0016020	membrane	1	2,42	117656	21,43
GO:0032774	RNA biosynthetic process	7	7,21	33819	6,16
GO:0051246	regulation of protein metabolic process	5	8,5	39461	7,19
GO:0035639	purine ribonucleoside triphosphate binding	4	3,9	101195	18,43
GO:0016310	phosphorylation	5	7,21	11853	2,16
GO:0051649	establishment of localization in cell	3	8,12	18663	3,4
GO:0044248	cellular catabolic process	3	4,39	66633	12,14
GO:0044249	cellular biosynthetic process	3	4,3	613402	111,73
GO:0016627	oxidoreductase activity, acting on the CH-CH group of donors	3	6,28	3704	0,67
GO:0023052	signaling	1	7,24	2412	0,44
GO:0016054	organic acid catabolic process	5	8,95	10843	1,98
GO:0034645	cellular macromolecule biosynthetic process	5	6,43	102282	18,63
GO:0008233	peptidase activity	3	4,38	18134	3,3
GO:0000166	nucleotide binding	4	2,47	227918	41,51
GO:0031224	intrinsic component of membrane	3	3,87	84772	15,44
GO:0008236	serine-type peptidase activity	5	6,36	4246	0,77
GO:0005737	cytoplasm	5	3,27	147574	26,88
GO:0043169	cation binding	3	3,51	123562	22,51
GO:0052689	carboxylic ester hydrolase activity	4	7,31	5834	1,06
GO:0031090	organelle membrane	3	5,43	37243	6,78
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	10	12,38	2585	0,47
GO:0005634	nucleus	7	5,7	32457	5,91
GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	3	6,68	3515	0,64
GO:0008270	zinc ion binding	6	5,41	23177	4,22
GO:0015276	ligand-gated ion channel activity	7	9,01	1089	0,2
GO:0032561	guanyl ribonucleotide binding	7	6,09	14895	2,71
GO:0044456	synapse part	2	9,72	3213	0,59
GO:0006508	proteolysis	7	8,01	8560	1,56
GO:0005506	iron ion binding	6	5,63	12817	2,33
GO:0022610	biological adhesion	1	8,92	6260	1,14
GO:0044459	plasma membrane part	5	6,12	20187	3,68
GO:1901607	alpha-amino acid biosynthetic process	9	8	31198	5,68
GO:1903508	positive regulation of nucleic acid-templated transcription			8026	
GO:0016070	RNA metabolic process	6	6,02	129759	23,64
GO:0032501	multicellular organismal process	1	7,43	22304	4,06
GO:0044281	small molecule metabolic process	3	3,5	263832	48,06
GO:0010556	regulation of macromolecule biosynthetic process	5	6,29	98664	17,97

GO:0005215	transporter activity	1	2,52	78367	14,27
GO:0005216	ion channel activity	6	6,79	5430	0,99
GO:0006725	cellular aromatic compound metabolic process	3	3,1	370913	67,56
GO:1903506	regulation of nucleic acid-templated transcription			50995	
GO:1903507	negative regulation of nucleic acid-templated transcription			8729	
GO:0045211	postsynaptic membrane	7	10,62	855	0,16
GO:0045892	negative regulation of transcription, DNA-negative regulation of transcription, DNA- templated	9	10,2	8813	1,61
GO:0045893	positive regulation of transcription, DNA-templated	9	10,79	8417	1,53
GO:0044712	single-organism catabolic process	3	5,68	31131	5,67
GO:0005975	carbohydrate metabolic process	3	6,17	63869	11,63
GO:0044255	cellular lipid metabolic process	4	7,37	41954	7,64
GO:0055085	transmembrane transport	5	7	21935	4
GO:0032268	regulation of cellular protein metabolic process	6	9,01	30538	5,56
GO:0016853	isomerase activity	2	4,85	16867	3,07
GO:0006082	organic acid metabolic process	4	5,22	160458	29,23
GO:0030554	adenyl nucleotide binding	6	3,97	88416	16,1
GO:1901605	alpha-amino acid metabolic process	8	6,69	86605	15,77
GO:1901135	carbohydrate derivative metabolic process	3	4,6	172418	31,41
GO:0051252	regulation of RNA metabolic process	6	6,97	53761	9,79
GO:0010629	negative regulation of gene expression	6	9,54	13736	2,5
GO:0016053	organic acid biosynthetic process	5	7,03	49266	8,97
GO:1902680	positive regulation of RNA biosynthetic process	8	10,77	8112	1,48
GO:0046394	carboxylic acid biosynthetic process	7	7,03	51592	9,4
GO:0032550	purine ribonucleoside binding	5	3,89	102321	18,64
GO:0022836	gated channel activity	5	7,94	2310	0,42
GO:0005515	protein binding	2	5,33	33855	6,17
GO:0022834	ligand-gated channel activity	6	8,99	961	0,18
GO:0022838	substrate-specific channel activity	5	6,77	3964	0,72
GO:0032502	developmental process	1	6,16	103077	18,78
GO:0004175	endopeptidase activity	5	5,91	9808	1,79
GO:0038023	signaling receptor activity	3	5,9	10151	1,85
GO:0016887	ATPase activity	7	4,6	20999	3,82
GO:0031328	positive regulation of cellular biosynthetic process	6	10,48	11418	2,08
GO:0031327	negative regulation of cellular biosynthetic process	6	9,37	19231	3,5
GO:0031326	regulation of cellular biosynthetic process	5	6,28	101647	18,51
GO:0031325	positive regulation of cellular metabolic process	5	9,37	39512	7,2
GO:0031324	negative regulation of cellular metabolic process	5	8,77	30229	5,51
GO:0031323	regulation of cellular metabolic process	4	5,6	165995	30,24
GO:0019752	carboxylic acid metabolic process	6	5,44	418474	76,22
GO:0090304	nucleic acid metabolic process	5	5,6	129108	23,52
GO:0036094	small molecule binding	2	2,06	234291	42,68
GO:0004872	receptor activity	1	5,71	9869	1,8
GO:0004871	signal transducer activity	2	4,99	11912	2,17
GO:0009069	serine family amino acid metabolic process	9	8,93	10408	1,9
GO:0001883	purine nucleoside binding	4	3,89	102352	18,64

GO:0001882	nucleoside binding	3	3,19	205732	37,47
GO:0050662	coenzyme binding	3	5,09	18720	3,41
GO:0050660	flavin adenine dinucleotide binding	5	6,3	4117	0,75
GO:0005615	extracellular space	3	8,8	5564	1,01
GO:0006355	regulation of transcription, DNA-templated	8	7,15	43643	7,95
GO:1901360	organic cyclic compound metabolic process	3	3,07	418474	76,22
GO:0006520	cellular amino acid metabolic process	7	5,86	125141	22,79
GO:2000112	regulation of cellular macromolecule biosynthetic process	6	7,03	52255	9,52
GO:0004888	transmembrane signaling receptor activity	4	6,86	6982	1,27
GO:0043226	organelle	1	2,39	274732	50,04
GO:0006796	phosphate-containing compound metabolic process	4	3,85	133560	24,33
GO:0004674	protein serine/threonine kinase activity	6	7,02	5300	0,97
GO:0070011	peptidase activity, acting on L-amino acid peptides	4	4,66	28392	5,17
GO:1901265	nucleoside phosphate binding	3	2,26	226051	41,17
GO:0071702	organic substance transport	4	6,91	32732	5,96
GO:0010468	regulation of gene expression	5	6,86	68926	12,55
GO:0044446	intracellular organelle part	6	4,23	115940	21,12
GO:0006468	protein phosphorylation	7	9,16	5318	0,97
GO:0045935	positive regulation of nucleobase-containing compound metabolic process	6	10,57	9903	1,8
GO:0045934	negative regulation of nucleobase-containing compound metabolic process	6	9,85	10339	1,88
GO:0044267	cellular protein metabolic process	5	6,84	90302	16,45
GO:0019219	regulation of nucleobase-containing compound metabolic process	5	6,74	60292	10,98
GO:0017111	nucleoside-triphosphatase activity	6	3,88	37367	6,81
GO:0006464	cellular protein modification process	6	7,99	33228	6,05
GO:1902679	negative regulation of RNA biosynthetic process	8	10,19	8800	1,6
GO:0034220	ion transmembrane transport	6	7,93	27627	5,03
GO:0044765	single-organism transport	4	5,68	86855	15,82
GO:0009059	macromolecule biosynthetic process	4	5,99	147235	26,82
GO:0051171	regulation of nitrogen compound metabolic process	4	6,68	69314	12,63
GO:0051172	negative regulation of nitrogen compound metabolic process	5	9,85	10984	2
GO:0007268	synaptic transmission	5	12,33	1254	0,23
GO:0007267	cell-cell signaling	4	11,85	2202	0,4
GO:0007154	cell communication	3	7,18	9092	1,66
GO:0007264	small GTPase mediated signal transduction	6	11,13	2353	0,43
GO:0009056	catabolic process	2	3,62	103425	18,84
GO:0051179	localization	1	4,85	148560	27,06
GO:1902578	single-organism localization	2	10,47	54277	9,89
GO:0008610	lipid biosynthetic process	4	8,13	31347	5,71
GO:0006996	organelle organization	3	8,54	37867	6,9
GO:0032991	macromolecular complex	1	4,18	127477	23,22
GO:0044260	cellular macromolecule metabolic process	4	5,02	345895	63
GO:0051173	positive regulation of nitrogen compound metabolic process	5	10,56	11460	2,09
GO:0003700	sequence-specific DNA binding transcription factor activity	2	5,03	12557	2,29
GO:0048856	anatomical structure development	2	7,04	26989	4,92

GO:0043168	anion binding	3	3,61	123883	22,56
GO:0098796	membrane protein complex			32432	
GO:0022857	transmembrane transporter activity	2	3,15	72887	13,28
GO:0005576	extracellular region	1	2,9	24021	4,38
GO:0006793	phosphorus metabolic process	3	3,09	158827	28,93
GO:0015031	protein transport	5	8,21	19388	3,53
GO:0048037	cofactor binding	2	4,7	30293	5,52
GO:0016462	pyrophosphatase activity	5	3,85	39976	7,28
GO:0048523	negative regulation of cellular process	4	8,43	44395	8,09
GO:0048522	positive regulation of cellular process	4	8,91	55121	10,04
GO:0048583	regulation of response to stimulus	3	6.74	67838	12,36
GO:1901564	organonitrogen compound metabolic process	3	3.98	452635	82,45
GO:0016874	ligase activity	2	4.60	34021	6,2
GO:0044421	extracellular region part	2	2.92	16396	2,99
GO:0006807	nitrogen compound metabolic process	2	2.74	498947	90,88
GO:0009653	anatomical structure morphogenesis	3	8.85	15056	2,74
GO:0044767	single-organism developmental process	2	6.87	61910	11,28
GO:0034641	cellular nitrogen compound metabolic process	3	3.08	503729	91,75
GO:0016491	oxidoreductase activity	2	3.24	61780	11,25
GO:0005509	calcium ion binding	5	7.44	4360	0,79
GO:0043170	macromolecule metabolic process	3	4.01	63937	11,65

Annex 8 - Statistics per GO term identified in the association rule learning step

GO term	P	N	IP	FP	TP	FN	TN	recall	prec	F1	MCC
GO:0031559	0	2590	9	9			2581	0	0	0	NA
GO:0044444	950	1640	2590	1640	950	0	0	1	0,367	0,537	NA
GO:0043231	964	1626	2590	1626	964	0	0	1	0,372	0,542	NA
GO:0043227	1005	1585	2590	1585	1005	0	0	1	0,388	0,559	NA
GO:0043229	1093	1497	2590	1497	1093	0	0	1	0,422	0,594	NA
GO:0043226	1159	1431	2590	1431	1159	0	0	1	0,447	0,618	NA
GO:0065007	1254	1336	2590	1336	1254	0	0	1	0,484	0,652	NA
GO:0005506	29	2561	86	64	22	7	2497	0,759	0,256	0,383	0,431
GO:0020037	34	2556	88	65	23	11	2491	0,676	0,261	0,377	0,409
GO:0005231	4	2586	6	4	2	2	2582	0,5	0,333	0,4	0,407
GO:0009069	4	2586	7	5	2	2	2581	0,5	0,286	0,364	0,377
GO:0004930	51	2539	171	134	37	14	2405	0,725	0,216	0,333	0,376
GO:0046906	41	2549	88	65	23	18	2484	0,561	0,261	0,356	0,369
GO:0004497	32	2558	81	62	19	13	2496	0,594	0,235	0,337	0,361
GO:0016705	35	2555	81	62	19	16	2493	0,543	0,235	0,328	0,344
GO:0004713	6	2584	42	37	5	1	2547	0,833	0,119	0,208	0,312
GO:0004888	86	2504	388	328	60	26	2176	0,698	0,155	0,254	0,285
GO:0005230	8	2582	25	21	4	4	2561	0,5	0,16	0,242	0,279
GO:0003924	30	2560	7	3	4	26	2557	0,133	0,571	0,216	0,272
GO:0004674	68	2522	591	529	62	6	1993	0,912	0,105	0,188	0,267
GO:0017111	114	2476	259	206	53	61	2270	0,465	0,205	0,285	0,261
GO:0016310	90	2500	362	307	55	35	2193	0,611	0,152	0,243	0,258
GO:0006468	69	2521	351	305	46	23	2216	0,667	0,131	0,219	0,257
GO:0016462	118	2472	286	231	55	63	2241	0,466	0,192	0,272	0,248
GO:0038023	111	2479	393	330	63	48	2149	0,568	0,16	0,25	0,245
GO:0016887	61	2529	68	51	17	44	2478	0,279	0,25	0,264	0,245
GO:0016818	120	2470	286	231	55	65	2239	0,458	0,192	0,271	0,245
GO:0016817	120	2470	286	231	55	65	2239	0,458	0,192	0,271	0,245
GO:0004672	99	2491	780	696	84	15	1795	0,848	0,108	0,192	0,238
GO:0005524	288	2302	1496	1235	261	27	1067	0,906	0,174	0,292	0,235
GO:0030554	295	2295	1500	1234	266	29	1061	0,902	0,177	0,296	0,234
GO:0032559	294	2296	1496	1232	264	30	1064	0,898	0,176	0,294	0,232
GO:0004872	129	2461	412	344	68	61	2117	0,527	0,165	0,251	0,23
GO:0032553	353	2237	1689	1364	325	28	873	0,921	0,192	0,318	0,224
GO:0004871	145	2445	444	369	75	70	2076	0,517	0,169	0,255	0,223
GO:0001883	333	2257	1667	1361	306	27	896	0,919	0,184	0,307	0,221
GO:0032550	333	2257	1667	1361	306	27	896	0,919	0,184	0,307	0,221
GO:0007268	30	2560	6	3	3	27	2557	0,1	0,5	0,167	0,22
GO:0032549	333	2257	1672	1366	306	27	891	0,919	0,183	0,305	0,22
GO:0035639	333	2257	1663	1358	305	28	899	0,916	0,183	0,305	0,219

GO:0032555	339	2251	1671	1361	310	29	890	0,914	0,186	0,309	0,218
GO:0016773	119	2471	793	702	91	28	1769	0,765	0,115	0,2	0,218
GO:0017076	340	2250	1677	1366	311	29	884	0,915	0,185	0,308	0,217
GO:0034220	72	2518	7	2	5	67	2516	0,069	0,714	0,126	0,217
GO:0001882	333	2257	1704	1397	307	26	860	0,922	0,18	0,301	0,214
GO:0016772	173	2417	907	782	125	48	1635	0,723	0,138	0,232	0,209
GO:0060089	163	2427	451	373	78	85	2054	0,479	0,173	0,254	0,208
GO:0019752	197	2393	333	260	73	124	2133	0,371	0,219	0,275	0,207
GO:0097367	388	2202	1746	1395	351	37	807	0,905	0,201	0,329	0,206
GO:0016301	133	2457	820	723	97	36	1734	0,729	0,118	0,203	0,206
GO:0044281	307	2283	941	747	194	113	1536	0,632	0,206	0,311	0,205
GO:0016740	333	2257	1330	1072	258	75	1185	0,775	0,194	0,31	0,201
GO:0046914	173	2417	175	131	44	129	2286	0,254	0,251	0,252	0,199
GO:0043436	203	2387	350	276	74	129	2111	0,365	0,211	0,267	0,196
GO:0007186	52	2538	61	49	12	40	2489	0,231	0,197	0,213	0,196
GO:0006082	206	2384	354	279	75	131	2105	0,364	0,212	0,268	0,195
GO:0015276	17	2573	25	21	4	13	2552	0,235	0,16	0,19	0,188
GO:0022834	17	2573	25	21	4	13	2552	0,235	0,16	0,19	0,188
GO:0000166	415	2175	1933	1546	387	28	629	0,933	0,2	0,329	0,187
GO:1901265	415	2175	1933	1546	387	28	629	0,933	0,2	0,329	0,187
GO:0005525	57	2533	52	41	11	46	2492	0,193	0,212	0,202	0,185
GO:0032561	57	2533	52	41	11	46	2492	0,193	0,212	0,202	0,185
GO:0019001	57	2533	52	41	11	46	2492	0,193	0,212	0,202	0,185
GO:0055085	99	2491	7	2	5	94	2489	0,051	0,714	0,095	0,184
GO:0006796	192	2398	600	503	97	95	1895	0,505	0,162	0,245	0,183
GO:0006508	83	2507	12	6	6	77	2501	0,072	0,5	0,126	0,181
GO:0007267	44	2546	6	3	3	41	2543	0,068	0,5	0,12	0,18
GO:0006793	197	2393	620	521	99	98	1872	0,503	0,16	0,243	0,177
GO:0044700	46	2544	6	3	3	43	2541	0,065	0,5	0,115	0,176
GO:0023052	46	2544	6	3	3	43	2541	0,065	0,5	0,115	0,176
GO:0036094	450	2140	2028	1607	421	29	533	0,936	0,208	0,34	0,17
GO:1901564	221	2369	382	306	76	145	2063	0,344	0,199	0,252	0,169
GO:0043168	449	2141	2035	1615	420	29	526	0,935	0,206	0,338	0,167
GO:0031982	142	2448	12	5	7	135	2443	0,049	0,583	0,09	0,158
GO:0003995	5	2585	8	7	1	4	2578	0,2	0,125	0,154	0,156
GO:0043169	506	2084	1430	1071	359	147	1013	0,709	0,251	0,371	0,156
GO:1903507	76	2514	30	22	8	68	2492	0,105	0,267	0,151	0,152
GO:0045211	15	2575	24	21	3	12	2554	0,2	0,125	0,154	0,152
GO:1901607	15	2575	11	9	2	13	2566	0,133	0,182	0,154	0,152
GO:1902679	78	2512	30	22	8	70	2490	0,103	0,267	0,149	0,15
GO:0008652	32	2558	12	9	3	29	2549	0,094	0,25	0,137	0,147
GO:0045892	74	2516	26	19	7	67	2497	0,095	0,269	0,14	0,145
GO:0006629	164	2426	63	45	18	146	2381	0,11	0,286	0,159	0,144
GO:0016787	410	2180	716	542	174	236	1638	0,424	0,243	0,309	0,143
GO:0051253	84	2506	30	22	8	76	2484	0,095	0,267	0,14	0,143

GO:0046907	102	2488	7	3	4	98	2485	0,039	0,571	0,073	0,142
GO:0044255	131	2459	17	10	7	124	2449	0,053	0,412	0,094	0,134
GO:0010558	95	2495	30	22	8	87	2473	0,084	0,267	0,128	0,132
GO:0045934	95	2495	30	22	8	87	2473	0,084	0,267	0,128	0,132
GO:0001071	143	2447	47	34	13	130	2413	0,091	0,277	0,137	0,132
GO:0003700	143	2447	47	34	13	130	2413	0,091	0,277	0,137	0,132
GO:2000113	87	2503	26	19	7	80	2484	0,08	0,269	0,123	0,132
GO:1901135	81	2509	6	3	3	78	2506	0,037	0,5	0,069	0,13
GO:0006357	115	2475	45	34	11	104	2441	0,096	0,244	0,138	0,129
GO:0016829	86	2504	15	10	5	81	2494	0,058	0,333	0,099	0,128
GO:0006811	132	2458	10	5	5	127	2453	0,038	0,5	0,071	0,127
GO:0007264	34	2556	38	33	5	29	2523	0,147	0,132	0,139	0,127
GO:0031327	102	2488	30	22	8	94	2466	0,078	0,267	0,121	0,126
GO:0009890	104	2486	30	22	8	96	2464	0,077	0,267	0,12	0,125
GO:0010605	190	2400	34	22	12	178	2378	0,063	0,353	0,107	0,124
GO:0007154	89	2501	6	3	3	86	2498	0,034	0,5	0,064	0,123
GO:0004252	34	2556	116	107	9	25	2449	0,265	0,078	0,121	0,123
GO:0035556	124	2466	95	78	17	107	2388	0,137	0,179	0,155	0,12
GO:0051172	111	2479	30	22	8	103	2457	0,072	0,267	0,113	0,12
GO:0022836	58	2532	25	20	5	53	2512	0,086	0,2	0,12	0,119
GO:0044421	220	2370	102	77	25	195	2293	0,114	0,245	0,156	0,116
GO:0015075	101	2489	28	21	7	94	2468	0,069	0,25	0,108	0,114
GO:0031224	572	2018	1157	841	316	256	1177	0,552	0,273	0,365	0,113
GO:0017171	37	2553	120	111	9	28	2442	0,243	0,075	0,115	0,113
GO:0008236	37	2553	120	111	9	28	2442	0,243	0,075	0,115	0,113
GO:0005216	63	2527	25	20	5	58	2507	0,079	0,2	0,113	0,113
GO:0022838	63	2527	25	20	5	58	2507	0,079	0,2	0,113	0,113
GO:0031324	169	2421	32	22	10	159	2399	0,059	0,313	0,099	0,112
GO:0097060	26	2564	24	21	3	23	2543	0,115	0,125	0,12	0,112
GO:0032774	300	2290	207	158	49	251	2132	0,163	0,237	0,193	0,111
GO:0009892	221	2369	34	22	12	209	2347	0,054	0,353	0,094	0,11
GO:0006520	79	2511	81	70	11	68	2441	0,139	0,136	0,137	0,11
GO:0031988	132	2458	5	2	3	129	2456	0,023	0,6	0,044	0,11
GO:0022803	67	2523	25	20	5	62	2503	0,075	0,2	0,109	0,108
GO:0015267	67	2523	25	20	5	62	2503	0,075	0,2	0,109	0,108
GO:0036211	209	2381	448	383	65	144	1998	0,311	0,145	0,198	0,108
GO:0006464	209	2381	448	383	65	144	1998	0,311	0,145	0,198	0,108
GO:0016491	161	2429	215	183	32	129	2246	0,199	0,149	0,17	0,108
GO:0006351	293	2297	188	144	44	249	2153	0,15	0,234	0,183	0,107
GO:0097659	293	2297	188	144	44	249	2153	0,15	0,234	0,183	0,107
GO:0044765	301	2289	31	18	13	288	2271	0,043	0,419	0,078	0,104
GO:0051649	153	2437	8	4	4	149	2433	0,026	0,5	0,049	0,104
GO:0010629	136	2454	30	22	8	128	2432	0,059	0,267	0,097	0,104
GO:0022891	118	2472	28	21	7	111	2451	0,059	0,25	0,095	0,102
GO:0007166	126	2464	357	320	37	89	2144	0,294	0,104	0,154	0,102

GO:0034654	328	2262	239	184	55	273	2078	0,168	0,23	0,194	0,099
GO:1902578	320	2270	31	18	13	307	2252	0,041	0,419	0,075	0,099
GO:0007155	75	2515	5	3	2	73	2512	0,027	0,4	0,051	0,097
GO:0022857	128	2462	28	21	7	121	2441	0,055	0,25	0,09	0,097
GO:0044456	57	2533	24	20	4	53	2513	0,07	0,167	0,099	0,095
GO:0004175	79	2511	127	114	13	66	2397	0,165	0,102	0,126	0,095
GO:0034702	44	2546	9	7	2	42	2539	0,045	0,222	0,075	0,094
GO:0016053	59	2531	24	20	4	55	2511	0,068	0,167	0,097	0,093
GO:0046394	59	2531	24	20	4	55	2511	0,068	0,167	0,097	0,093
GO:0043412	235	2355	494	422	72	163	1933	0,306	0,146	0,198	0,093
GO:1901605	34	2556	25	22	3	31	2534	0,088	0,12	0,102	0,093
GO:0016021	558	2032	1057	781	276	282	1251	0,495	0,261	0,342	0,092
GO:0022610	84	2506	5	3	2	82	2503	0,024	0,4	0,045	0,091
GO:0016070	414	2176	318	239	79	335	1937	0,191	0,248	0,216	0,09
GO:0022892	144	2446	28	21	7	137	2425	0,049	0,25	0,082	0,089
GO:0005886	387	2203	571	452	119	268	1751	0,307	0,208	0,248	0,088
GO:1901576	547	2043	1146	858	288	259	1185	0,527	0,251	0,34	0,088
GO:0050660	15	2575	8	7	1	14	2568	0,067	0,125	0,087	0,087
GO:0031090	313	2277	17	9	8	305	2268	0,026	0,471	0,049	0,087
GO:0005615	139	2451	7	4	3	136	2447	0,022	0,429	0,042	0,087
GO:0044249	539	2051	1046	785	261	278	1266	0,484	0,25	0,33	0,084
GO:0048037	66	2524	26	22	4	62	2502	0,061	0,154	0,087	0,082
GO:0052689	34	2556	4	3	1	33	2553	0,029	0,25	0,052	0,082
GO:1902495	46	2544	11	9	2	44	2535	0,043	0,182	0,07	0,081
GO:1990351	46	2544	11	9	2	44	2535	0,043	0,182	0,07	0,081
GO:0048519	448	2142	179	128	51	397	2014	0,114	0,285	0,163	0,081
GO:0044459	194	2396	141	118	23	171	2278	0,119	0,163	0,138	0,08
GO:0044248	123	2467	9	6	3	120	2461	0,024	0,333	0,045	0,079
GO:0009059	372	2218	217	166	51	321	2052	0,137	0,235	0,173	0,079
GO:0006725	578	2012	1063	784	279	299	1228	0,483	0,262	0,34	0,079
GO:0009058	570	2020	1282	958	324	246	1062	0,568	0,253	0,35	0,078
GO:0044710	636	1954	2400	1788	612	24	166	0,962	0,255	0,403	0,078
GO:0005215	165	2425	29	22	7	158	2403	0,042	0,241	0,072	0,077
GO:0005737	527	2063	109	71	38	489	1992	0,072	0,349	0,119	0,076
GO:0034645	361	2229	206	159	47	314	2070	0,13	0,228	0,166	0,075
GO:0016627	18	2572	9	8	1	17	2564	0,056	0,111	0,074	0,074
GO:0019438	362	2228	311	246	65	297	1982	0,18	0,209	0,193	0,074
GO:0070011	96	2494	135	122	13	83	2372	0,135	0,096	0,112	0,074
GO:0008233	97	2493	136	123	13	84	2370	0,134	0,096	0,112	0,072
GO:0006810	379	2211	149	112	37	342	2099	0,098	0,248	0,14	0,071
GO:0051234	388	2202	152	114	38	350	2088	0,098	0,25	0,141	0,07
GO:0044283	90	2500	35	30	5	85	2470	0,056	0,143	0,08	0,069
GO:0044711	205	2385	212	182	30	175	2203	0,146	0,142	0,144	0,069
GO:0046395	31	2559	6	5	1	30	2554	0,032	0,167	0,054	0,069
GO:0016054	31	2559	6	5	1	30	2554	0,032	0,167	0,054	0,069

GO:0007165	363	2227	813	671	142	221	1556	0,391	0,175	0,242	0,067
GO:1901362	380	2210	350	278	72	308	1932	0,189	0,206	0,197	0,066
GO:0008270	121	2469	12	9	3	118	2460	0,025	0,25	0,045	0,066
GO:0044712	90	2500	8	6	2	88	2494	0,022	0,25	0,04	0,065
GO:1901360	606	1984	1216	896	320	286	1088	0,528	0,263	0,351	0,065
GO:0005509	61	2529	12	10	2	59	2519	0,033	0,167	0,055	0,064
GO:0090304	477	2113	537	412	125	352	1701	0,262	0,233	0,247	0,064
GO:0010468	431	2159	492	386	106	325	1773	0,246	0,215	0,229	0,064
GO:0031326	413	2177	498	395	103	310	1782	0,249	0,207	0,226	0,063
GO:0048523	359	2231	76	56	20	339	2175	0,056	0,263	0,092	0,063
GO:0048518	591	1999	479	344	135	456	1655	0,228	0,282	0,252	0,061
GO:0019538	338	2252	711	595	116	222	1657	0,343	0,163	0,221	0,06
GO:0016788	122	2468	7	5	2	120	2463	0,016	0,286	0,03	0,059
GO:0034641	615	1975	1158	851	307	308	1124	0,499	0,265	0,346	0,058
GO:0018130	353	2237	322	261	61	292	1976	0,173	0,189	0,181	0,058
GO:0044267	271	2319	589	508	81	190	1811	0,299	0,138	0,189	0,058
GO:0044425	723	1867	1600	1121	479	244	746	0,663	0,299	0,412	0,057
GO:0060255	539	2051	730	551	179	360	1500	0,332	0,245	0,282	0,057
GO:0009889	417	2173	512	408	104	313	1765	0,249	0,203	0,224	0,057
GO:0051171	415	2175	464	369	95	320	1806	0,229	0,205	0,216	0,057
GO:0046483	565	2025	1053	794	259	306	1231	0,458	0,246	0,32	0,056
GO:0016043	494	2096	96	67	29	465	2029	0,059	0,302	0,099	0,056
GO:0044282	38	2552	7	6	1	37	2546	0,026	0,143	0,044	0,055
GO:2000112	374	2216	450	366	84	290	1850	0,225	0,187	0,204	0,055
GO:0010556	383	2207	465	377	88	295	1830	0,23	0,189	0,207	0,055
GO:0051179	417	2173	174	133	41	376	2040	0,098	0,236	0,138	0,054
GO:0006139	529	2061	846	647	199	330	1414	0,376	0,235	0,289	0,054
GO:0009056	172	2418	18	14	4	168	2404	0,023	0,222	0,042	0,052
GO:0080090	539	2051	741	562	179	360	1489	0,332	0,242	0,28	0,052
GO:0044707	311	2279	110	88	22	289	2191	0,071	0,2	0,105	0,052
GO:0019219	379	2211	459	374	85	294	1837	0,224	0,185	0,203	0,051
GO:0031323	594	1996	759	560	199	395	1436	0,335	0,262	0,294	0,05
GO:0048522	529	2061	271	200	71	458	1861	0,134	0,262	0,177	0,049
GO:0015031	103	2487	32	28	4	99	2459	0,039	0,125	0,059	0,049
GO:0003676	448	2142	693	552	141	307	1590	0,315	0,203	0,247	0,049
GO:0032501	325	2265	121	97	24	301	2168	0,074	0,198	0,108	0,049
GO:0045184	109	2481	32	28	4	105	2453	0,037	0,125	0,057	0,046
GO:1901575	153	2437	15	12	3	150	2425	0,02	0,2	0,036	0,046
GO:0050896	692	1898	848	597	251	441	1301	0,363	0,296	0,326	0,045
GO:0044446	812	1778	794	520	274	538	1258	0,337	0,345	0,341	0,045
GO:0098588	207	2383	6	4	2	205	2379	0,01	0,333	0,019	0,045
GO:2001141	340	2250	385	321	64	276	1929	0,188	0,166	0,176	0,043
GO:0051252	348	2242	385	320	65	283	1922	0,187	0,169	0,178	0,042
GO:0005975	140	2450	3	2	1	139	2448	0,007	0,333	0,014	0,042
GO:1903506	339	2251	384	321	63	276	1930	0,186	0,164	0,174	0,041

GO:0044271	401	2189	319	257	62	339	1932	0,155	0,194	0,172	0,041
GO:0006807	680	1910	1667	1207	460	220	703	0,676	0,276	0,392	0,041
GO:0016020	798	1792	1874	1275	599	199	517	0,751	0,32	0,449	0,04
GO:0016874	58	2532	24	22	2	56	2510	0,034	0,083	0,048	0,04
GO:0006355	337	2253	382	320	62	275	1933	0,184	0,162	0,172	0,04
GO:0008104	124	2466	34	30	4	120	2436	0,032	0,118	0,05	0,038
GO:0033036	126	2464	34	30	4	122	2434	0,032	0,118	0,05	0,037
GO:0098796	128	2462	12	10	2	126	2452	0,016	0,167	0,029	0,037
GO:0050662	50	2540	10	9	1	49	2531	0,02	0,1	0,033	0,037
GO:0003677	300	2290	289	246	43	257	2044	0,143	0,149	0,146	0,036
GO:0044422	837	1753	1558	1033	525	312	720	0,627	0,337	0,438	0,036
GO:0016192	85	2505	6	5	1	84	2500	0,012	0,167	0,022	0,036
GO:0005515	418	2172	38	28	10	408	2144	0,024	0,263	0,044	0,034
GO:1901566	135	2455	35	31	4	131	2424	0,03	0,114	0,048	0,033
GO:0071702	189	2401	36	31	5	184	2370	0,026	0,139	0,044	0,03
GO:0071840	507	2083	139	105	34	473	1978	0,067	0,245	0,105	0,029
GO:0048583	371	2219	71	58	13	358	2161	0,035	0,183	0,059	0,019
GO:0030054	110	2480	25	23	2	108	2457	0,018	0,08	0,029	0,018
GO:0032502	575	2015	605	463	142	433	1552	0,247	0,235	0,241	0,017
GO:0048856	309	2281	98	84	14	295	2197	0,045	0,143	0,068	0,014
GO:0043234	392	2198	183	152	31	361	2046	0,079	0,169	0,108	0,014
GO:0019222	674	1916	849	622	227	447	1294	0,337	0,267	0,298	0,011
GO:0032991	474	2116	260	209	51	423	1907	0,108	0,196	0,139	0,011
GO:0051239	286	2304	44	38	6	280	2266	0,021	0,136	0,036	0,011
GO:0044428	326	2264	6	5	1	325	2259	0,003	0,167	0,006	0,006
GO:0010628	177	2413	11	10	1	176	2403	0,006	0,091	0,011	0,006
GO:0044767	535	2055	502	398	104	431	1657	0,194	0,207	0,2	0,001
GO:0043170	800	1790	1876	1297	579	221	493	0,724	0,309	0,433	-0,001
GO:0044260	740	1850	1623	1160	463	277	690	0,626	0,285	0,392	-0,001
GO:0005576	190	2400	29	27	2	188	2373	0,011	0,069	0,019	-0,002
GO:0016866	7	2583	12	12	0	7	2571	0	0	0	-0,004
GO:0009653	100	2490	1	1	0	100	2489	0	0	0	-0,004
GO:0000122	28	2562	4	4	0	28	2558	0	0	0	-0,004
GO:0006259	80	2510	2	2	0	80	2508	0	0	0	-0,005
GO:0032268	197	2393	2	2	0	197	2391	0	0	0	-0,008
GO:0051246	203	2387	2	2	0	203	2385	0	0	0	-0,008
GO:0006996	230	2360	2	2	0	230	2358	0	0	0	-0,009
GO:0008610	55	2535	9	9	0	55	2526	0	0	0	-0,009
GO:0019637	102	2488	5	5	0	102	2483	0	0	0	-0,009
GO:0009893	354	2236	28	25	3	351	2211	0,008	0,107	0,015	-0,009
GO:0010604	254	2336	28	26	2	252	2310	0,008	0,071	0,014	-0,009
GO:0016853	48	2542	12	12	0	48	2530	0	0	0	-0,009
GO:0048869	224	2366	3	3	0	224	2363	0	0	0	-0,01
GO:0045893	131	2459	6	6	0	131	2453	0	0	0	-0,011
GO:1903508	131	2459	6	6	0	131	2453	0	0	0	-0,011

GO:1902680	132	2458	6	6	0	132	2452	0	0	0	-0,011
GO:0051254	136	2454	6	6	0	136	2448	0	0	0	-0,011
GO:0031325	304	2286	25	23	2	302	2263	0,007	0,08	0,013	-0,011
GO:0050789	1165	1425	2583	1422	1161	4	3	0,997	0,449	0,619	-0,013
GO:0010557	149	2441	7	7	0	149	2434	0	0	0	-0,013
GO:0045935	154	2436	7	7	0	154	2429	0	0	0	-0,013
GO:0051173	163	2427	7	7	0	163	2420	0	0	0	-0,013
GO:0031328	171	2419	7	7	0	171	2412	0	0	0	-0,014
GO:0009891	174	2416	7	7	0	174	2409	0	0	0	-0,014
GO:0065008	303	2287	18	17	1	302	2270	0,003	0,056	0,006	-0,016
GO:0050793	262	2328	6	6	0	262	2322	0	0	0	-0,016
GO:0005634	545	2045	737	590	147	398	1455	0,27	0,199	0,229	-0,017
GO:0005829	215	2375	9	9	0	215	2366	0	0	0	-0,018
GO:0006950	420	2170	23	21	2	418	2149	0,005	0,087	0,009	-0,019
GO:0050794	1075	1515	2542	1494	1048	27	21	0,975	0,412	0,579	-0,041

Annex 9 - Proteins whose annotation changed in two sequential Swiss-Prot releases

GO term	Q61618
Annotation in Swiss-Prot release 2015_07	GO:0005887; C:integral component of plasma membrane; IMP:MGI, GO:0042629; C:mast cell granule; IDA:GOC, GO:0005886; C:plasma membrane; IDA:MGI, GO:0001609; F:G-protein coupled adenosine receptor activity; IDA:MGI, GO:0001973; P:adenosine receptor signaling pathway; IMP:MGI, GO:0002553; P:histamine secretion by mast cell; IDA:MGI, GO:0050850; P:positive regulation of calcium-mediated signaling; IDA:MGI, GO:0050729; P:positive regulation of inflammatory response; IMP:MGI, GO:0002687; P:positive regulation of leukocyte migration; IMP:MGI, GO:0043306; P:positive regulation of mast cell degranulation; IDA:MGI, GO:0070257; P:positive regulation of mucus secretion; IMP:MGI, GO:0014068; P:positive regulation of phosphatidylinositol 3-kinase signaling; IDA:MGI
Annotation in Swiss-Prot release 2015_08	GO:0016021; C:integral component of membrane; IEA:UniProtKB-KW, GO:0005886; C:plasma membrane; IEA:UniProtKB-SubCell, GO:0001609; F:G-protein coupled adenosine receptor activity; IEA:InterPro

GO term	Q39147
Annotation in Swiss-Prot release 2015_07	GO:0008725; F:DNA-3-methyladenine glycosylase activity; IGI:TAIR
Annotation in Swiss-Prot release 2015_08	GO:0005634; C:nucleus; IEA:UniProtKB-SubCell, GO:0003677; F:DNA binding; IEA:InterPro, GO:0008725; F:DNA-3-methyladenine glycosylase activity; IEA:UniProtKB-EC, GO:0052822; F:DNA-3-methylguanine glycosylase activity; IEA:UniProtKB-EC, GO:0052821; F:DNA-7-methyladenine glycosylase activity; IEA:UniProtKB-EC, GO:0043916; F:DNA-7-methylguanine glycosylase activity; IEA:UniProtKB-EC, GO:0006284; P:base-excision repair; IEA:InterPro

GO term	Q76HN1
Annotation in Swiss-Prot release 2015_07	GO:0004415; F:hyaluronon glucosaminidase activity; IDA:RGD, GO:0008219; P:cell death; IDA:RGD, GO:0030214; P:hyaluronan catabolic process; IDA:RGD, GO:0008285; P:negative regulation of cell proliferation; IDA:RGD
Annotation in Swiss-Prot release 2015_08	GO:0005737; C:cytoplasm; ISS:UniProtKB, GO:0031410; C:cytoplasmic vesicle; ISS:UniProtKB, GO:0005615; C:extracellular space; ISS:UniProtKB, GO:0036117; C:hyaluranon cable; ISS:UniProtKB, GO:0005764; C:lysosome; ISS:UniProtKB, GO:0050501; F:hyaluronan synthase activity; ISS:UniProtKB, GO:0004415; F:hyaluronon glucosaminidase activity; IEA:UniProtKB-EC, GO:0008134; F:transcription factor binding; ISS:UniProtKB, GO:0005975; P:carbohydrate metabolic process; IEA:InterPro, GO:0051216; P:cartilage development; ISS:UniProtKB, GO:0071347; P:cellular response to interleukin-1; ISS:UniProtKB, GO:0071467; P:cellular response to pH; ISS:UniProtKB, GO:0036120; P:cellular response to platelet-derived growth factor stimulus; ISS:UniProtKB, GO:0071493; P:cellular response to UV-B; ISS:UniProtKB, GO:0030213; P:hyaluronan biosynthetic process; ISS:UniProtKB, GO:0030212; P:hyaluronan metabolic process; ISS:UniProtKB, GO:0006954; P:inflammatory response; ISS:UniProtKB, GO:0030308; P:negative regulation of cell growth; ISS:UniProtKB, GO:0045766; P:positive regulation of angiogenesis; ISS:UniProtKB, GO:0045785; P:positive regulation of cell adhesion; ISS:UniProtKB, GO:0030307; P:positive regulation of cell growth; ISS:UniProtKB, GO:0010634; P:positive regulation of epithelial cell migration; ISS:UniProtKB, GO:0050679; P:positive regulation of epithelial cell proliferation; ISS:UniProtKB, GO:0045927; P:positive regulation of growth; ISS:UniProtKB, GO:1900106; P:positive regulation of hyaluranon cable assembly; ISS:UniProtKB, GO:0046677; P:response to antibiotic; ISS:UniProtKB, GO:0000302; P:response to reactive oxygen species; ISS:UniProtKB, GO:0009615; P:response to virus; ISS:UniProtKB