The human “magnesome”: how to detect human proteins that can bind Mg ions

Damiano Piovesan and Rita Casadio

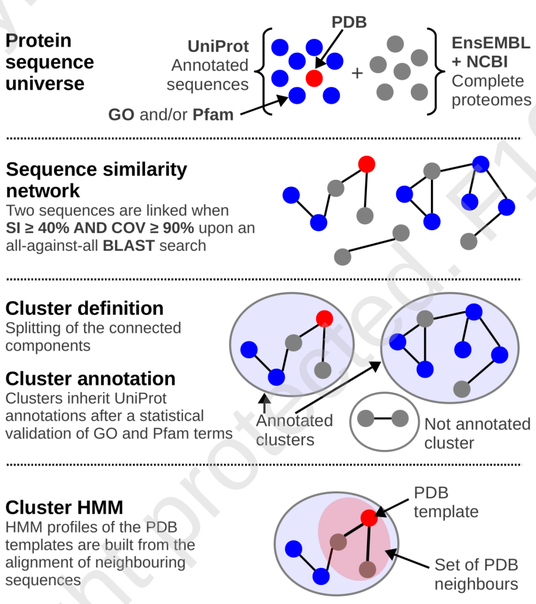
Biocomputing Group, CIRI - Health Science and Technologies / Department of Biology, University of Bologna, Italy

Abstract

How many human proteins bind Mg ions? We addressed this question computationally with our BAR-PLUS (BAR+), a non hierarchical clustering method that relies on the pair wise sequence comparison of about 14 millions proteins from over 300.000 species, of whom 998 are complete proteomes and include *Homo sapiens*. All the sequences cluster in over 900.000 clusters with the constraints that their sequences identity is ≥40% on ≥90% of the alignment length. From this they can inherit from the cluster in a validated manner functional and structural annotation (PDB +/ SCOP +/ Pfam +/ Goterms +/ Ligands). With our procedure we find that 2402 human proteins bind Mg ions in 88 clusters and another 1140 in 161 clusters bind cofactors through Mg ions. Some 30% of all the sequences are annotated for the first time as endowed with putative Mg binding sites. A cell localisation of the 3542 Mg binding proteins (the human Magnesome) is developed considering that the most populated Cell Components are: Intracellular part, Endomembrane system, Cell periphery, Protein complex. In turn the most populated Biological Processes are: Cellular metabolic process, Primary metabolic process, Multicellular organismal development, Macromolecule metabolic process, Nitrogen compound metabolic process, Small molecule metabolic process, Anatomical structure development, Cell cycle, Cell death. The most populated Molecular Functions are: Hydrolase activity, Tranferase activity, Nucleotide binding, Ion Binding, Protein binding, Oxideraductase activity, Signal transducer activity, Isomerase activity. We also characterise typical Mg binding signatures useful in annotating Mg binding sites from protein sequences.

Methods

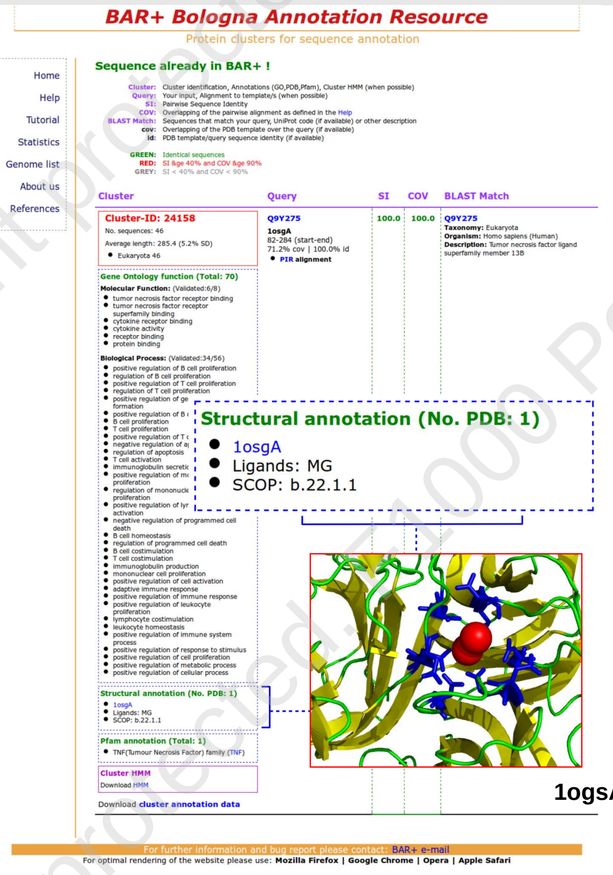
BAR+ for protein annotation



Piovesan et al.
NAR – Web server issue 2011
http://bar.biocomp.unibo.it/bar2.0

Our method collects sequences from the protein universe (UniProtKB) including also some 988 genomes. By this all the features [PDB (+/- SCOP classification) (red circles), GO terms (including Molecular Function, Biological Process and Cellular Localization) and Pfam models (blue circles) are also included. An extensive BLAST alignment is performed of all the 13,495,736 sequences in a GRID environment. The sequence similarity network is built by connecting two sequences only if their sequence identity is ≥40% with an overlapping (Coverage, COV) ≥90%. 913,762 clusters are obtained by splitting of the connected components. By this any cluster may contain from two up to 87,893 sequences (one cluster containing ABC transporters from Prokaryotes, Eukaryotes and Archaea). Stand alone sequences are called Singletons (30,4% of the total protein universe). Sequences inherit the annotations within a cluster. When clusters are endowed with PDB template/s, a Cluster HMM is generated by considering all the sequences that have an identity ≥40% and a coverage≥90% with the structure/s (pink subset). The Cluster-HMM can be used to align all the other sequences in the cluster to template/s.

Magnesium binding sites with BAR+



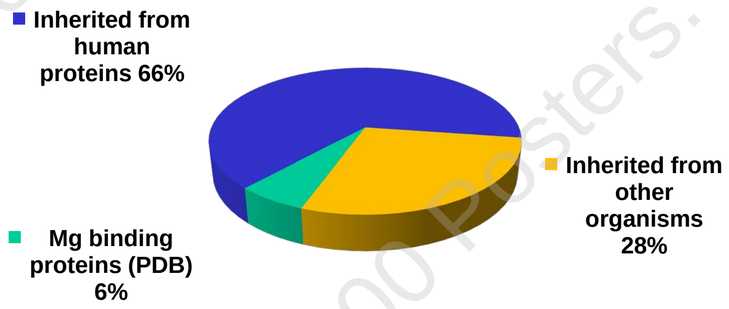
The BAR+ web server at work. The result page identifies a cluster that comprises 46 sequences from Eukaryotes with an average length of 285 residues and only one PDB structure (Chain A of the Tumor necrosis factor ligand superfamily member 13B [1osgA]. This protein chain binds two Mg ions and its GO terms (function) are available in UniProtKB. Within the cluster all the other partially annotated or non annotated sequences inherited all the features in a validated manner. The page also lists validated GO terms, the PDB template, the SCOP classification, and the Pfam domain (term “Tumor necrosis factor”) (PF00229). BAR+ gives the target/template alignment for computational modelling of the 3D structure of all the other sequences in the cluster. Among these two are other two from *Homo sapiens* that inherit all the cluster specific annotation, including the Mg binding sites.

1osgA:

SITE	1 AC1	5 GLN A 234	MG A 702	GLN B 234	ASN B 243	
SITE	2 AC1	5 GLN C 234				
SITE	1 AC2	3 ASN A 235	MG A 701	ASN B 243		
SITE	1 AC3	3 ASN D 235	MG D 802	ASN E 243		
SITE	1 AC4	4 GLN D 234	GLN E 234	MG E 801	GLN F 234	

Results

The human sequences whose structure in PDB has at least one Mg ion as “ligand” are only 337 (“Mg binding protein (PDB)”). With BAR+ we increase this number up to 5073. This occurs either by direct transfer of annotation from human Mg binding proteins (3542, “Inherited from human proteins”) or when human sequences fall into clusters where there are Mg binding proteins from other organism (1531, “Inherited from other organisms”).



Protein Mg binding sites are well characterized regarding their structural geometry. However the pattern of residues that are involved in Mg binding is still poorly characterised. From our analysis with BAR+ four different groups can be defined. Our focus is the human genome and in the following we consider human proteins. The groups comprises clusters where human proteins bind Mg as the only ion, Mg and other ions, Mg and other ligands (routinely not ions), and Mg with other ions and other ligands.

	Cluster	Sequence	PDB	RMSD (Å)	SD	Human Inherited	Seq with conserved sites	Human inherited (Id<30%)	Seq with conserved sites (Id<30%)
Mg	14	14	14	#	#	73	66	1	1
Mg + Ions	11	11	14	0.47	0.00	59	59	5	5
Mg + Ligands	45	47	123	0.90	0.89	332	318	39	39
Mg + Ions + Ligands	68	130	436	1.38	1.76	2492	2489	37	37

When more PDB structure fall into the same cluster their RMSDs (Root Mean Square Deviations) is very low (<2 Å) for all groups. This indicates that BAR+ clusters preserve the structural specificity. About the 99% of the inherited sequences (“Seq with conserved sites” column) when aligned conserve identical Mg binding residues (data not shown). This is so also for very distantly related sequences (sequence identity <30%, last column) that are in the same cluster. Therefore when a target sequence falls into a cluster characterised by Mg ligand the binding site annotation can be safely inherited.

Below we list the cellular localization (**Cellular Component of Gene Ontology**) of the human sequences of the four groups. (the “**human magnesome”**) For each GO term the number of human sequences (direct plus inherited) falling in a cluster with the corresponding validated annotation is reported. The selected terms are those that are the most distant from the ontology root. This is the most specific annotation of the cellular localization obtained with BAR+. Similarly GO terms of biological process and molecular function can be obtained.

Cellular Component (localization)

Number of sequences	Go terms
Mg	
23	endoplasmic reticulum lumen
21	cell body
8	cell part
2	membrane-bounded organelle
2	intracellular membrane-bounded organelle
Mg + Ions	
33	site of polarized growth
13	intracellular membrane-bounded organelle
9	intracellular
4	cellular_component
3	type III intermediate filament
2	membrane-bounded organelle
Mg + Ligands	
118	azurophil granule
37	cytoplasmic mRNA processing body
19	cytoplasmic membrane-bounded vesicle
16	intracellular
15	intracellular membrane-bounded organelle
14	mitochondrion
11	neuron projection
11	cell part
8	mitochondrial intermembrane space
8	external side of plasma membrane
8	endoplasmic reticulum part
7	cellular_component
6	mitochondrial matrix
5	nuclear part
3	positive transcription elongation factor complex b
3	nucleus
2	soluble fraction
2	extracellular space
2	gamma-tubulin small complex
2	chromatin
2	mitochondrial matrix
2	histone pre-mRNA 3'end processing complex
Mg + Ions + Ligands	
1817	cell surface
117	endoplasmic reticulum part
92	dendrite cytoplasm
56	mitochondrial matrix
48	cell division site
47	ruffle
44	cell septum
44	membrane raft
37	endoplasmic reticulum
24	cell leading edge
23	plasma membrane enriched fraction
22	internal side of plasma membrane
15	cell cortex
15	intracellular membrane-bounded organelle
12	protein complex
11	mitochondrial tricarboxylic acid cycle enzyme complex
9	mitochondrion
9	cell part
8	macromolecular complex
8	Golgi cisterna membrane
7	extracellular region
6	nuclear chromosome
6	nucleus
5	soluble fraction
4	mitochondrial part
4	plant-type cell wall
3	cytoplasmic membrane-bounded vesicle
3	protein kinase CK2 complex
2	intracellular
2	microtubule organizing center
2	cytosolic small ribosomal subunit