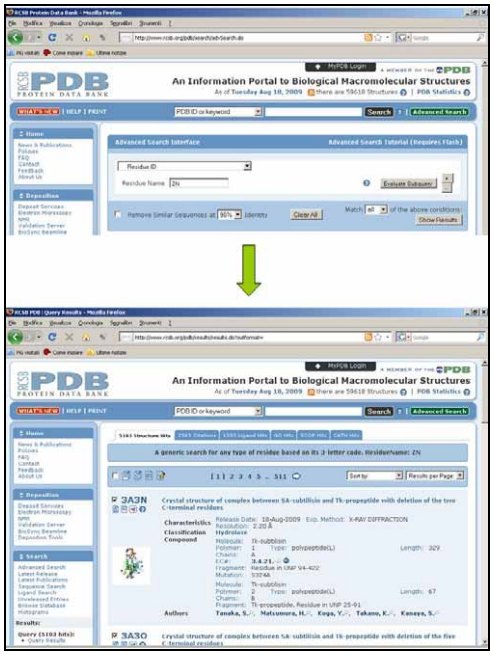
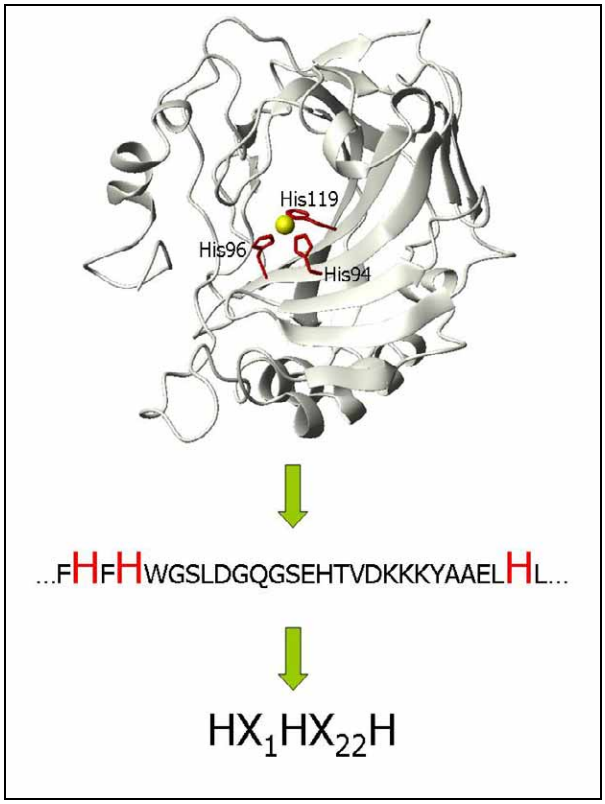
**ESI.** Example protocol for the identification of zinc proteins based on our method.

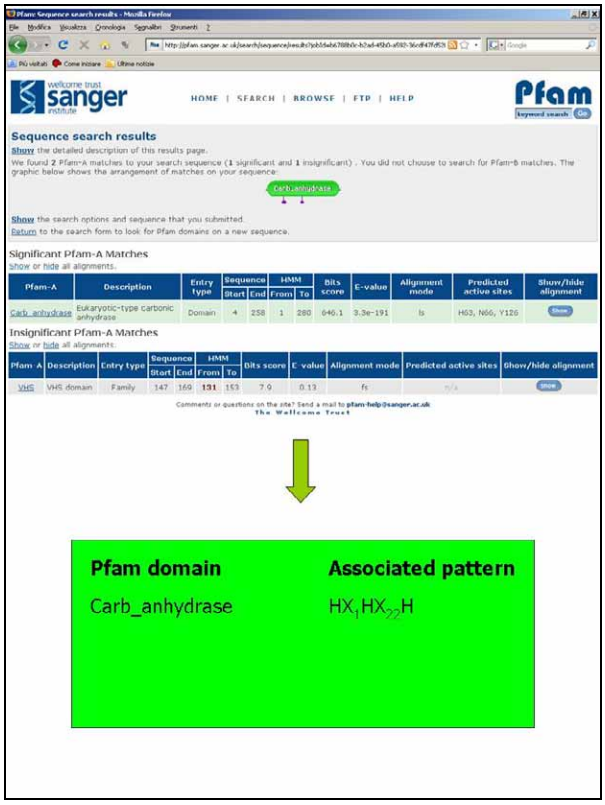
**Step 1.** All protein structures that bind zinc are retrieved from the PDB. A comprehensive retrieval strategy involves identifying all HET groups defined in PDB that contain zinc (see http://deposit.pdb.org/het_dictionary.txt) and then querying the PDB for structures that contain any of those HET groups. The Figure shows the result of a query for structures containing HET groups named "ZN".



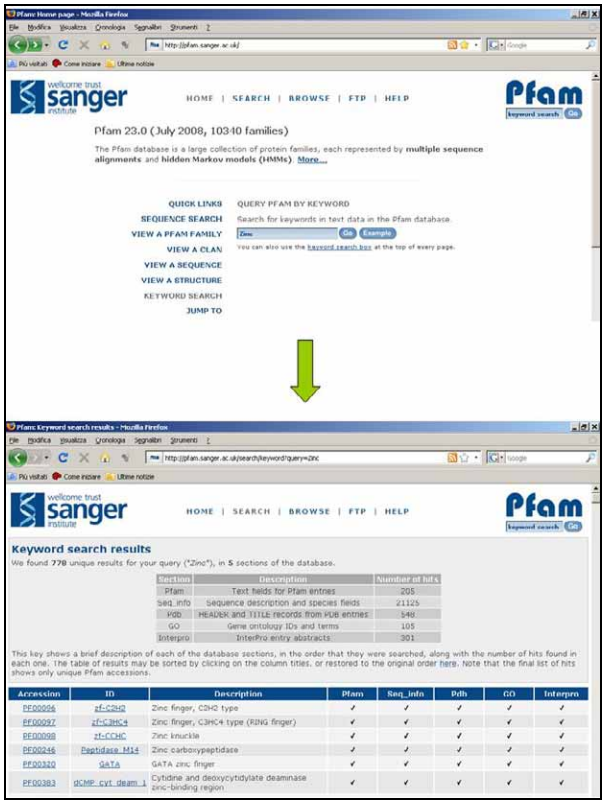
**Step 2.** After removing proteins where zinc is not present under physiological conditions, zinc-binding PDB structures are analysed to identify the amino acid residues that coordinate the metal. The zinc ligands are then mapped onto the protein sequence to define a metal-binding pattern. The Figure shows the zinc ligands and the corresponding pattern for human carbonic anhydrase II (PDB code 2ILI). Note that the removal of non-physiological zinc proteins relies on the analysis of literature, which is much facilitated by grouping PDB structures according to CATH and/or SCOP databases.



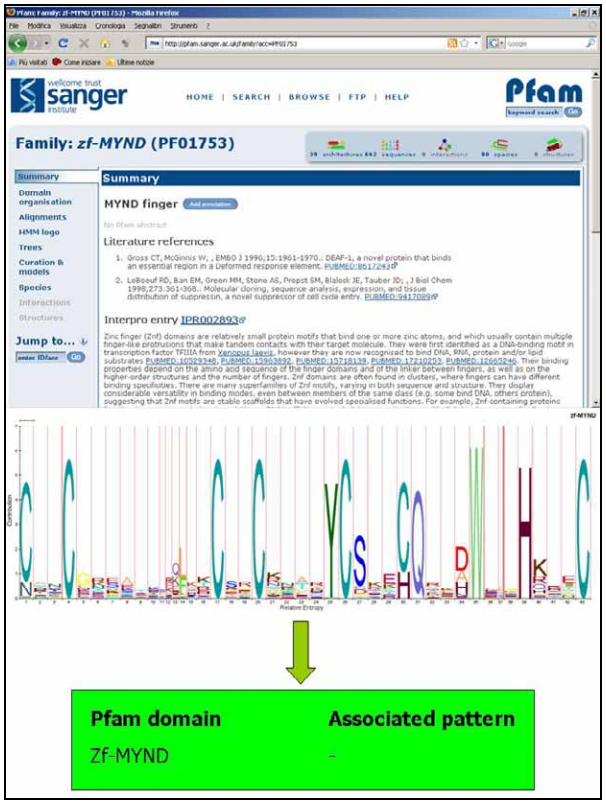
**Step 3.** Zinc-binding PDB structures are analysed to identify the protein domains that contain the zinc-binding sites. This is done by comparing their amino acid sequences against the Pfam database. The Figure shows that human carbonic anhydrase II (PDB code 2ILI) contains the Pfam domain named "Carb_anhydrase", which contains the zinc-binding site. The "Carb_anyhdrase" domain is thus taken as zinc-binding and associated with the $HX_1HX_{22}H$ pattern (see Step 2). In this way, a library of zinc-binding Pfam domains associated with zinc-binding patterns is built.



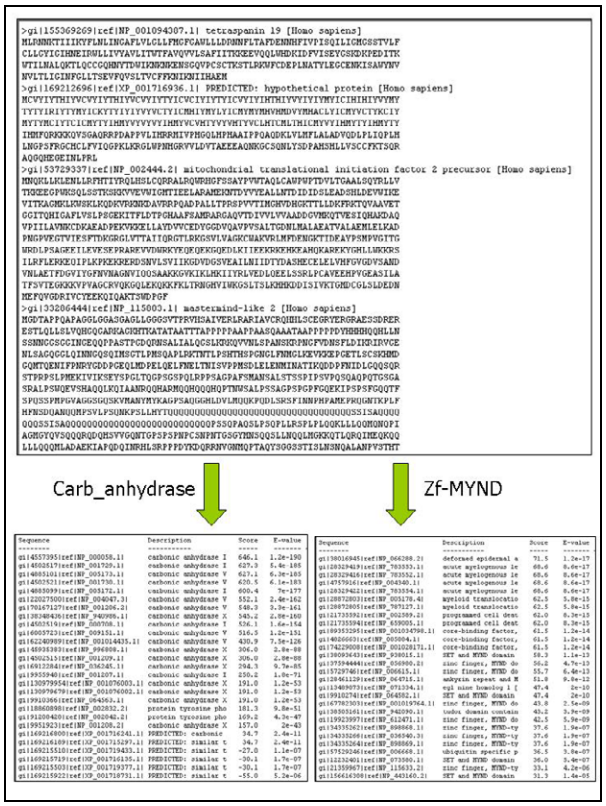

**Step 4.** The Pfam database is queried for all domains whose annotation contains the word "zinc" (the Figure shows the result of such a query) or the word "Zn". By these queries, one retrieves most (but not all) domains identified in Step 3, other domains which have been characterized as zinc-binding though a structure is not available, and other domains which are not in fact zinc-binding though the word "zinc" or "Zn" is mentioned in their annotation.



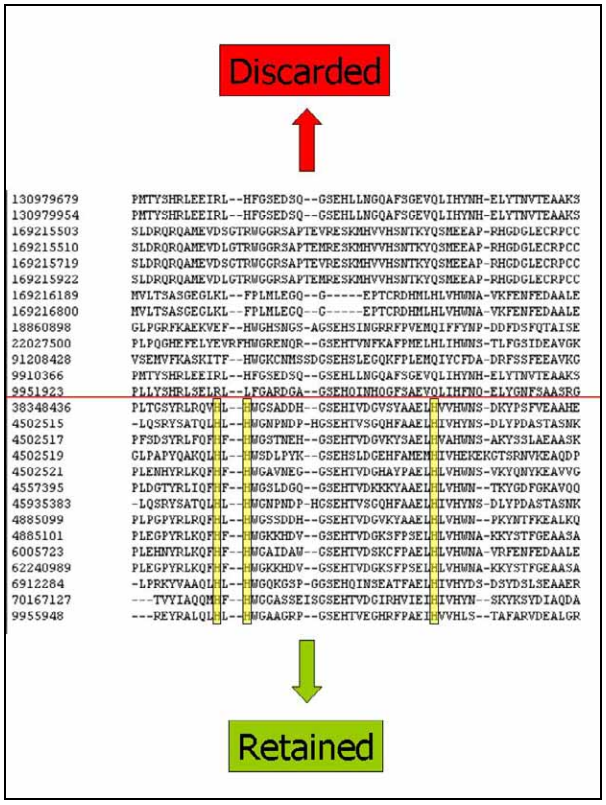
**Step 5.** Pfam domains retrieved in Step 4 and having no associated zinc-binding pattern are manually classified as zinc-binding (and then added to the library) or not, on the basis of literature analysis and the occurrence of conserved putative ligands in the sequences. The Figure shows the annotation and the sequence profile (revealing the presence of conserved Cys and His residues) of the Pfam domain named "zf-MYND", representing a zinc finger domain with unknown structure.



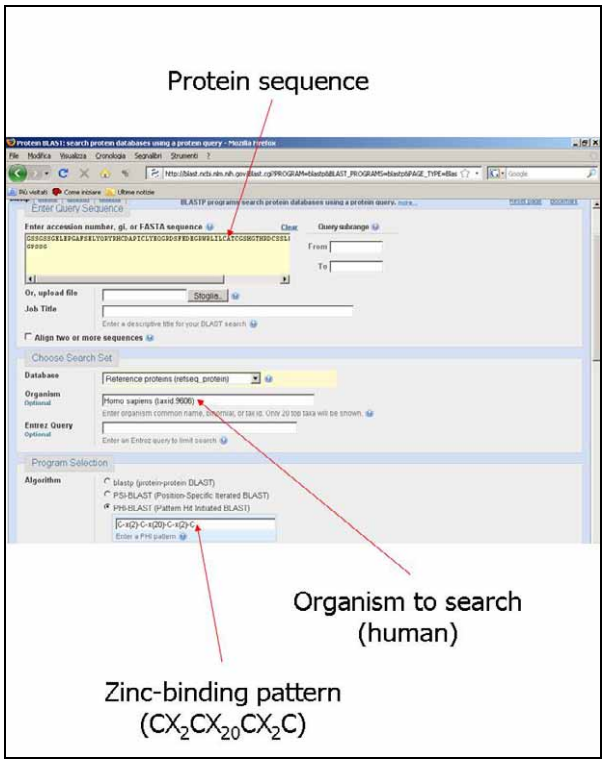

**Step 6.** The complete predicted proteome of an organism is scanned against the library of zinc-binding Pfam domains for proteins containing at least one such domain, using the program HMMER. This produces a set of proteins containing zinc-binding domains with an associated pattern, and a set of proteins containing zinc-binding domains with no associated pattern. The Figure shows the result of the search for human proteins containing the Carb_anyhdrase (27 hits) and the zf-MYND domain (28 hits).



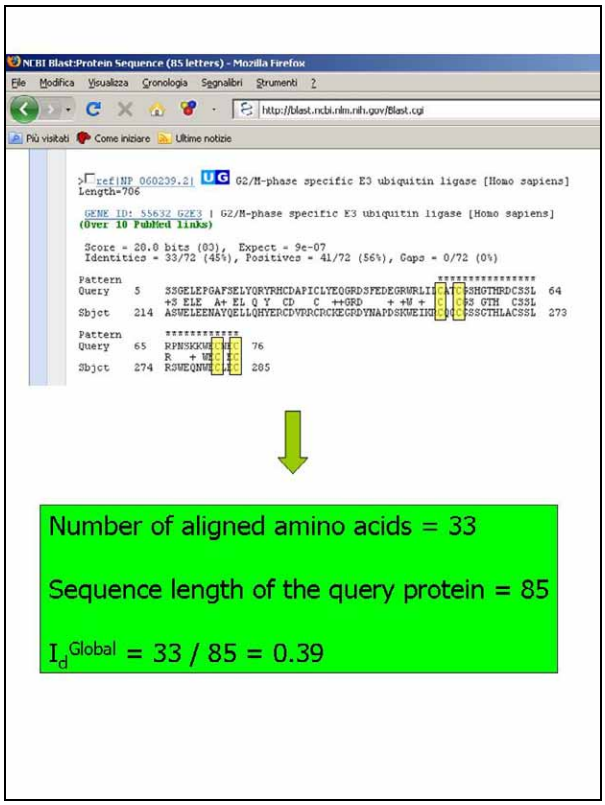
**Step 7.** For proteins containing a zinc-binding domain associated with a zinc-binding pattern, sequences are checked for the occurrence of the pattern, and those that lack the pattern are filtered out. The Figure shows that, out of the 27 human proteins containing a Carb_anhydrase domain (see Step 6), 13 do not have the zinc-binding pattern, and are thus filtered out. These proteins correspond to the so-called carbonic anhydrase-related proteins which, despite their sequence homology to the catalytic isozymes, cannot bind zinc and are devoid of $CO_2$ hydration activity.




**Step 8.** Zinc proteins which do not contain any known Pfam domain, but have a zinc-binding site similar to that of a protein with known structure are searched using PHI-BLAST. PHI-BLAST input consists of the amino acid sequence of the zinc protein with known structure plus the zinc-binding pattern extracted from that structure. The Figure shows PHI-BLAST settings to search zinc proteins with sites similar to that of mouse PHD finger protein 7 (PDB code 1weq).



**Step 9.** PHI-BLAST hits retrieved in Step 8 are evaluated based on the $I_d^{Global}$ parameter, defined as the ratio between the number of amino acids aligned by PHI-BLAST and the length of the sequence of the query protein. Proteins with $I_d^{Global} > 0.2$ are taken as zinc proteins. The Figure shows the PHI-BLAST alignment between mouse PHD finger protein 7 (PDB code 1weq) and human protein NP_060239.2, which is taken as a zinc protein ($I_d^{Global} = 0.39$) despite no zinc-binding Pfam domains could be detected.




**Step 10.** The predicted zinc proteome is obtained by collecting all the zinc proteins identified by (i) a zinc-binding domain with an associated zinc-binding pattern (e.g., the proteins with the Carb_anhydrase domain and the $HX_1HX_{22}H$ pattern retrieved in Step 7), (ii) a zinc-binding domain only (e.g., the proteins with the zf-MYND domain retrieved in Step 6), and (iii) a zinc-binding pattern only (e.g., the NP_060239.2 protein retrieved in Step 9). In human, predicted zinc proteins represent about 10% of the entire proteome.