Biomedical Data Mining

Extended Abstract

The conserved structure of the DUF 2419 family

Kevin Zarghan

Abstract

The DUF 2419 superfamily has emerged as a remarkably versatile protein scaffold for the evolution of diverse catalytic activities. The DUF 2419 protein family was recently discovered and thus it is not well understood by the scientific community but some evidence from its orthologs indicates that it is likely involved in tRNA processing. This project is a preliminary effort to structure and function discovery. Here we overexpress the DUF 2419 using both human and bacterial cell lines to produce crystals that can be further analyzed using X-ray crystallography. C9orf64 is a gene located on chromosome 9, that in humans encodes the protein queuosine (Q) salvage protein. Queuosine is a micronutrient modification found on the wobble position of tRNAs. Recent publications indicate that DUF 2419 is involved in the methylation of cell lines in ovarian cancer, breast cancer, colon cancer, and acute myeloid leukemia.

There is still much to learn regarding its function in the cells of various cancers. In humans, the expression of the gene of interest is highest in the duodenum and small intestine but is also expressed in 24 other tissues. The protein was then exposed to JCSG and Salt Rx conditions for a total of 192 separate conditions in both 4mg/ml and 8mg/ml which are used to generate high-quality crystals that can then be further analyzed to confirm.

To achieve the ultimate goal of systems biology to model both living cells and organisms, we must know the functions of all their constituent parts. Even for the most intensively experimentally studied organisms there are many proteins for which we have no clue as to their function. For example, in the yeast *Saccharomyces cerevisiae* approximately 1000 proteins (17% of the genome) are still uncharacterized.

The Pfam database is a collection of protein families and domains that has been widely used for annotating sequenced genomes. Grouping each protein encoded by a genome into a family of homologous proteins can help to annotate its function. For example, if one or more members of a Pfam family have an experimentally determined function then this function can be tentatively assigned to the other proteins in that family. Using this approach, the majority of proteins encoded by a genome can be annotated despite the fact that not a single protein in that particular genome has ever been experimentally investigated. Even in the absence of functional information, grouping proteins into families can indicate those amino acids within the proteins that are conserved and hence are potentially functionally important. Approximately three-quarters of all known proteins now match one or another of the 10 000 protein families in Pfam.

Domains of unknown function, or DUFs, are a large set of families within the Pfam database that do not include any protein of known function. Although called DUFs, for many of these families it is not known whether they actually represent one protein domain or many. The DUF naming scheme was introduced by Chris Ponting through the addition of DUF1 and DUF2 to the SMART database. These two domains were found to be widely distributed in bacterial signalling proteins. Subsequently, the functions of these domains were identified and they have since been renamed as the GGDEF (PF00990,

Kevin Zarghan
San Diego State University, USA, E-mail: kzarghan@gmail.com

SMART accession SM00267) and EAL (PF00563, SM00052) domains, respectively. Both of these domains are involved in processing cyclic diguanylate, a universal bacterial second-messenger molecule. Although no further DUFs appeared in SMART, DUF1 and DUF2 were added to Pfam in 1997 and little did Chris Ponting realise that he was starting a trend that would see thousands of uncharacterized and largely anonymous families being added to the protein-family databases.

DUFs are created with the same care and attention as all other Pfam families. The only difference is that the curators are unable to identify any functional information from the scientific literature at the time that they are carrying out their analysis.

It is sometimes surprisingly difficult to determine the specific function of a protein. In some cases, identifying a nucleotide-binding P-loop motif might be considered to be sufficient to define a function for that protein. However, knowing that a protein binds a nucleotide does not tell us what biological process the protein is participating in or what action or role it might be carrying out.

Proteins of known function can also contain DUFs. For example, the very well characterized Dicer endonuclease contains a domain first named DUF283 (PF03368). The strong sequence conservation of this domain within Dicer proteins indicated that it was likely to convey an important function, yet at the time of curation this region was uncharacterized. Subsequently, it has been found that DUF283 shows sequence similarity to double-stranded RNA-binding domains, which indeed represents a highly likely function for a domain within the Dicer dsRNA endonuclease.

Identifying functions for DUFs is extremely important for characterizing lists of biological parts. Essentially, there are three ways to determine the function of an uncharacterized domain: the first involves identifying similarity to a domain of known function, either by sequence comparison or by structural analysis of a newly solved structure of one of the member proteins, the second involves using contextual information such as genomic context to computationally identify function, as employed by databases such as STRING and PROLINKS, and the third is through good old-fashioned molecular biology or biochemistry. Notably, Sir Rich Roberts put forward a proposal to stimulate experimentation on such uncharacterized proteins and there have been commendable attempts to functionally characterize proteins on a large scale. Martzen and coworkers identified that DUF27 (PF01661) may possess an adenosine phosphate-ribose 1′-phosphate processing activity. This activity was subsequently experimentally confirmed and this domain is now called the MACRO domain. One issue with identifying the functions of proteins classified as DUFs is that they are usually non-essential. A systematic knockout screen of *B. subtilis* has indicated that only 4% of essential genes have unknown function. These results imply that the knockout strategies that are routinely employed to identify a phenotype to help understand function are much less likely to be fruitful for identification of the function of DUFs.

Slowly, momentum is being gained and more functions for DUFs are being identified. Since we began adding DUFs to Pfam nearly ten years ago more than 270 of them have been renamed or reclassified, usually when a function has been identified. Pfam curators have not yet had time to systematically recheck all of the existing 2000+ DUFs to see whether new functional information for either the family or the individual protein has been identified. However, over the coming year we hope to revisit all of them and rename and re-annotate those where function is now known. This exercise should potentially identify 100+ families that have now been characterized. We ask users that if they know of any recently identified functions for these families they please contact the authors of the Pfam database.

Many of the DUF families had a rather limited membership when added to Pfam. As additional sequences are incorporated into the sequence database and added to the relevant families, we

sometimes determine that these families are actually subfamilies of much larger families. In such cases, the DUF subfamily is merged into the larger parent family. Just under 200 such merges have occurred after successive sequence inclusions.

Various tools are now available that can help to identify relationships between DUFs and other functionally characterized families. Profile-HMM comparison tools, such as *HHsearch*, *PRC*, *SIMPRO* and *SCOOP*, have proved to be very useful in this regard. In many cases, these programs can identify distant yet functionally relevant similarities that standard sequence and profile methods may miss. When these similarities are identified, it is possible to merge the two families into one large but more divergent single family. More often than not in such cases a single profile HMM is not sensitive enough to detect all the members of two or more distantly related families. When we are confident that two or more families are derived from a common evolutionary ancestor, we group them together in Pfam clans. Pfam clans are collections of families that are thought to have originated from a common evolutionary ancestor. As of Pfam release 23.0, 199 DUFs belong to clans in which there are one or more related families with known function. These distant relationships of DUFs with non-DUFs within a clan can also provide clues to the likely function of DUFs, but one must be especially cautious when transferring function.

This work is partly presented at 24th World Chemistry & Systems Biology Conference on October 03-04, 2018 in Los Angeles, USA

Kevin Zarghan
San Diego State University, USA, E-mail: kzarghan@gmail.com