

MAO: Multiple Alignment Ontology

Julie D. Thompson¹, Patrice Koehl², Stephen R. Holbrook³, Kazutaka Katoh⁴,
Eric Westhof⁵, Dino Moras¹, Olivier Poch¹

Keywords: ontology, multiple alignment, protein sequence, RNA sequence, 3D structure

1 Introduction.

Ontologies have become important in bioinformatics as they provide a structured representation of the knowledge available in a particular domain [1]. Ontologies contain concepts, which are explicitly defined and include relationships between concepts. One of the most widely used bio-ontologies is the Gene Ontology (GO), which formalises knowledge about biological processes, molecular functions and cell components. Other ontologies also exist, such as protein-protein interactions (PSI-MI) or genome sequence annotation (SO). Many of these have been collected together at the Open Biomedical Ontologies (OBO) web site (<http://obo.sourceforge.net/>). OBO is an umbrella site for well-structured controlled vocabularies intended for shared use across different biological domains.

2 Multiple alignment ontology.

We present here MAO, a Multiple Alignment Ontology registered at the OBO web site. MAO is a task-oriented ontology for data retrieval and exchange in the fields of DNA/RNA alignment, protein sequence and protein structure alignment. Multiple alignments provide an ideal workbench for the integration, cross-validation and analysis of all the information pertaining to a particular sequence [2]. By placing the sequence in the context of the overall family, multiple alignments permit not only a horizontal analysis of the sequence along its length, but also a hierarchical, vertical view of its evolution. One of the earliest applications of multiple alignments was in phylogenetic analyses, ranging from the study of the evolution of a particular sequence to the classification of organisms and the construction of the “tree of life” initially based on alignments of ribosomal RNA sequences. Another important application is the structural and functional characterisation of protein families, using either homology-based methods to propagate information from a known to an unknown protein or mean *ab initio* predictions for a family of sequences. Here, multiple alignments are exploited at a number of levels, from the determination of molecular function to the characterisation of structural domains and the identification of functionally active sites or point mutations associated with somatic and inherited diseases. More recently, multiple alignments have played a fundamental role in most of the computational methods used in genomics and proteomics, from gene identification and sequence validation to the characterisation of molecular and cellular networks. The knowledge gained by these methods, combined with new high-throughput experimental techniques such as differential RNA-expression analysis, can help to identify potential therapeutic agents, providing a framework for rational and reliable drug discovery.

¹ IGBMC, 1 rue L. Fries, Illkirch, France, E-mail: (julie,moras,poch)@igbmc.u-strasbg.fr

² UC Davis, Davis, CA, USA. E-mail: koehl@cs.ucdavis.edu

³ LBNL, 1 Cyclotron Road, Berkeley CA, USA. E-mail: SRHolbrook@lbl.gov

⁴ Bioinformatics Center, ICR, Kyoto, Japan. E-mail: kkatoh@kuicr.kyoto-u.ac.jp

⁵ IBMC, 15 rue René Descartes, Strasbourg, France. E-mail: E.Westhof@ibmc.u-strasbg.fr

Multiple alignment techniques are now evolving in response to these new requirements and enormous progress has been achieved recently. Current developments are moving away from a single all-encompassing algorithm towards co-operative, knowledge based systems which exploit the new structural and functional data available. The success of these new techniques relies on efficient data mining and sharing of information between the different algorithms. Clearly, standards are now needed, not only for data integration, but also for knowledge extraction and presentation to the biologist in a user-friendly format. The purpose of MAO is to standardise descriptions of alignments and the associated structural and functional information in order to allow the different alignment techniques to communicate with each other. Most of the features associated with multiple alignments are defined as MAO concepts, ranging from a single residue to sub-families of sequences and/or 3D structures. The ontology is represented as a directed acyclic graph (DAG). The top-level concept is the multiple alignment, which could represent a single protein domain or may include complete, full-length sequences. The concepts are organised in a hierarchical structure as shown in Figure 1. Attributes are assigned to the concepts where appropriate, in order to permit the integration of more complex information such as residue function or activity, sequence feature conservation or 3D structural location. An important criterium in the design of MAO was the possibility to link with other biological ontologies, in particular those included in OBO. MAO has therefore been implemented in the common shared syntax defined by OBO, using the DAG-Edit software. Cross-links are provided to related ontologies such as GO, SO, PSI, Interpro, and the NCBI organismal classification, but the list of inter-relations will obviously grow as new domain ontologies are developed. Work is now in progress to construct an instance knowledge base of multiple alignments annotated using the MAO terms. This knowledge base will incorporate high quality, global multiple alignments that cover most of the known protein fold space. Information as diverse as gene structure, protein 3D structure/function or specific residue interactions will be combined together with taxonomic and evolutionary information to produce a detailed description of a protein family. An important part of this development will be the analysis and cross-validation of this mass of heterogeneous information and the presentation of the pertinent information in a user-friendly, graphical interface. The potential applications for such a knowledge base are numerous, but will include such fields as the definition of characteristic motifs for specific protein folds, or the automatic annotation of the ever-increasing number of hypothetical proteins being produced by the high-throughput genome sequencing projects.

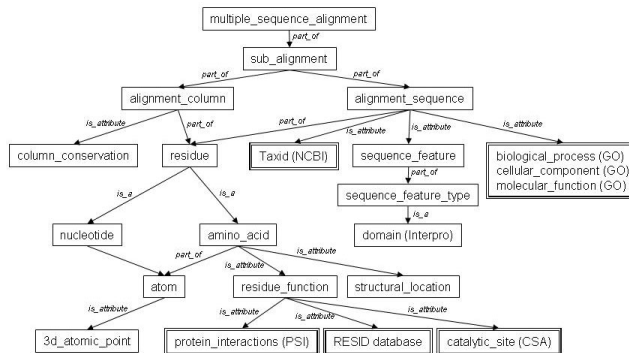


Figure 1: Part of the MAO ontology, showing the scope and some major interactions.

References

- [1] Bard, J.B. and Rhee, S.Y. 2004. Ontologies in biology: design, applications and future challenges. *Nat Rev Genet.* 3:213-222.
- [2] Lecompte, O., Thompson J.D., Plewniak, F., Thierry, J. and Poch O. 2001 Multiple alignment of complete sequences (MACS) in the post-genomic era. *Gene* 270:17-30.