

MilkMine - milking the dairy literature

S.G. Edwards ¹, B. Webber ¹, C. Holt ², L. Sawyer ¹

¹University of Edinburgh, UK; ²Hannah Research Institute, UK
stephen@milker.org.uk

The *MilkMine* project aims to bring together data from empirical research with emerging text-mining techniques. The vast amount of literature on milk proteins and their genes, as well as bioactive milk-protein derived peptides means that researchers often struggle to keep up with the information expansion and may not recognise vital biological links, particularly those that extend outwith their area of expertise. *MilkMine* is an attempt to provide a single informatics resource to help researchers mine this information mountain and to help focus development of research in the food, health and medical industries.

MilkMine is based on a generic data warehousing system, InterMine, enabling the integration of many traditional data types, such as UniProt protein data, genomes and comparative genomic data, as well as protein interaction data. During the data integration step, links are made between the data types, for example, finding orthologues for a given milk protein. Already a useful tool, this structure has been extended here to make use of text-mining techniques.

Milk and lactation specific terminology was identified to complement the underlying UMLS metathesaurus. This valuable, milk-related ontology can be used to identify biological concepts in free text and has been applied to milk literature contained in the PubMed database. The resultant semantically enriched literature allows the derivation of relationships between milk proteins, genes and peptides, as well as other biological concepts, such as diseases or biological processes. In this way we can create new hypotheses using the basic principle that if “A is linked to B”, and if “B is linked to C” then we can infer an association between A and C. Filtering and downstream processing of the many generated relationships help to reduce the number of non-significant interactions and improve the scoring of novel ones, for example, by removal of known associations where an A to C link is widely known.

The *MilkMine* resource is accessible online through an excellent query interface allowing users to generate and perform complex queries across the data model. In comparison with labour-based research, conceptual research is more cost-effective and it is hoped that this system will lead to the discovery of novel functional relationships among milk proteins under physiological and processing conditions, leading to manufacturing and health benefits for mother, child and consumer. These hypotheses can then be verified in the laboratory.