Understanding Milk’s Bioactive Components: A Goal for the Genomics Toolbox¹

Robert. E. Ward and J. Bruce German*²

Department of Food Science & Technology, University of California-Davis, Davis, CA 95616; and *Nestlé Research Center, Vers-chez-les-Blanc, CH-1000 Lausanne 26, Switzerland

ABSTRACT The challenges to food research are to propel foods beyond the successes of safety, convenience, and inclusion of all the essential nutrients, and to build the knowledge of genetics, metabolism, and biomolecules necessary for developing foods that deliver optimal health to each individual. How then can scientific and biological principles be developed to assemble this knowledge? The evolutionary success of milk has afforded compelling examples of a food material designed by selective pressure to provide optimal health to healthy mammalian offspring. Milk contains components that are more than assembled essential amino acids and that provide biological activities that improve the competitive success of offspring who consume them. Many of these molecules are proteins that protect individuals from exogenous stresses, toxins, and pathogens; encourage optimal growth, development, and adaptation to a chosen environment; and promote metabolic regulation for physical and intellectual performance. These structures and their actions are the basis of nutritional benefits that were not recognized when freedom from amino acid deficiency was the sole criterion of protein quality. The rapidly expanding tools of biotechnology are enabling a new perception of ingested proteins, how they are regulated, and how they achieve their specific functions. Genomes and their analyses are revealing the molecular details of their remarkable structural complexity and design. Milk proteins, either exclusively synthesized in the mammary gland during lactation or transported from plasma and concentrated in the mammary gland, have been largely co-opted from other functions. Establishing the evolutionary lineage of orthologous milk proteins, including the physiological process from which they were recruited, will lead to identification of their bioactivity. While most emphasis has been placed on the genes per se, our approaches implicate the regulatory regions of the genome as additional targets of milk’s biological information content. Understanding the structures is guiding scientists to new food ingredients. Understanding structures and regulation will guide scientists to new benefits and ultimately to the knowledge to build a new generation of delicious foods that genuinely deliver on the promise of safety and maintenance of optimal health. J. Nutr. 134: 962S–967S, 2004.

KEY WORDS: • milk • genome • bioinformatics • evolution

The myriad physiological benefits of milk consumption argue for a better understanding of the mechanisms underlying its action. Whereas a small number of the milk proteins have been characterized, the many dynamic nutritional functions of this complete food suggest there are components that provide benefits in ways we do not understand. It is becoming clear that in addition to the required nutrients, milk contains complex macromolecular structures, which are either synthesized or assembled in the mammary epithelia, and these are in part responsible for its beneficial action. Now that the sequencing of the human genome is complete, we have the potential to obtain the nucleotide sequences of all the human proteins, and can begin to identify those that are responsible for or are in milk. There are many nutritional functions that milk provides that do not have a simple explanation and the possible components in milk that are responsible are also not known. Beyond this, milk has hundreds of components and structures whose presence has not been explained. Since most of these components and structures are not absolutely essential nutrients, the simple strategy of removing them from milk and examining the consequences to offspring will not necessarily provide answers. With such a range of questions, the arrival of genomics and the global, nonhypothesis driven technologies this science has engendered will be useful as systematic discovery tools. While the genes and proteins responsible for the synthesis and delivery of

¹ Published in a supplement to The Journal of Nutrition. Presented as part of the 94th American Oil Chemists’ Annual Meeting & Expo held in Kansas City, MO, May 4–7, 2003. This symposium was sponsored by the National Dairy Council, Kraft Foods Inc., The Whey Protein Institute, and the U.S. Dairy Export Council. Guest editors for the supplement publication were Peter J. Huth, National Dairy Council, Rosemont, IL; Donald K. Layman, University of Illinois, Urbana, IL; and Peter H. Brown, Kraft Foods Research and Development, Kraft Foods Inc., Glenview, IL.
² To whom correspondence should be addressed. Email: jgerman@ucdavis.edu.

0022-3166/04 $8.00 © 2004 American Society for Nutritional Sciences.
milk are increasingly identified, functional genomic analysis, with the goal of annotating these components for nutritional function, may now begin.

A major initiative in food product development is in discovering functional ingredients, or components possessing nutritional benefits beyond the provision of essential nutrients and calories. Various dietary biomolecules have been associated anecdotally with functional nutritional benefits, and work is under way to characterize the mechanisms underlying these modes of action. However, beyond the causal benefits inferred from the consumption of certain foods, there is no obvious source pool to utilize in the discovery of new molecules. Milk, according to its role in biology, is composed of bioactive functional molecules, and thus represents an attractive source for the discovery of new food ingredients. In general, the use of model organisms in biological research is an efficient way of conducting investigations, because the results of many individual laboratories may be integrated to achieve a greater understanding of the system under study and possibly to other related systems. These integrative approaches have all been made possible by the dramatic increase in computing technologies over the past 3 decades. The power to build and interrogate very large databases is only possible now that this massive computing power is routinely available. In fact, bringing disparate biological data sets together with massive computing capabilities has become a scientific field in and of itself termed bioinformatics.

The field of bioinformatics is based in part on the principles of molecular evolution and the observation that conservation of biological sequence (nucleotide, protein, and carbohydrate) implies similarity of biological structure and function. Even beyond biological structures and functions, food functionalities are largely based on the principle of applying the basic biomaterial properties of specific biological structures to bulk polymer applications as food structures and functions (1). Bioinformatics is building the understanding of biological structure and function of life’s biopolymers and the same knowledge will be applicable to food structures. Foods do not need to rely on the common biopolymer properties of all proteins, polysaccharides, and lipids. Rather, the same structural features that are the basis of life can be investigated and understood in foods, and food processing can be designed to maintain this specificity and to extend the desired functionality. The physical properties of cheese are 1 such obvious example. Milk caseins were “designed” through evolution to spontaneously gel in the stomach of the neonate in response to the single peptide cleavage of k-casein catalyzed by chymosin. Cheese manufacturers discovered empirically that the same biological specificity could be co-opted into a bioprocessing system to produce a stable and nutritious food gel. This is a model for how food processing will become increasingly bio-guided. The principles of food structure formation, modification, and function will be increasingly guided by the knowledge of biological structures and their properties that are gained through genomics and bioinformatics. Thus, complete characterization of the profile of structures, functions, nutritional activities, and the constituent molecules in milk and those responsible for their synthesis will provide useful information for building new foods and for finding functional ingredients from others.

Coordinated research on the nutritional functions of milk is not only the starting point for the discovery of nutritionally active biomolecules, but also for the targets of their actions in the human intestine. Scientists are recognizing more and more the subtle benefits that milk provides to an infant throughout life. Milk assists the infant to adapt to its environment in many subtle ways. As breastfed children suffer fewer gastrointestinal disorders (2) and are statistically less likely to have metabolic complications later in life (3), the attention of scientists has increasingly focused on the functions provided for by milk (4). However, it is not sufficient to recognize that milk as an aggregate is beneficial. The challenge for the nutritional science researchers is to identify and characterize the specific molecules, structures, their targets, and their mechanisms of action in providing these benefits. Milk is an ideal model for establishing mechanisms of action for food and nutrition research in the 21st century. The composition is species, time, and physiologically specific, and the differences in composition address the different nutritional needs of the various offspring. Milk is a complete food, and therefore contains all the essential nutrients necessary for growth and development, again in a species-specific way. Furthermore, numerous structures have been characterized in milk which, although they are not nutrients per se, facilitate more efficient absorption and utilization of the nutrients present (3). More interesting for understanding the future of food research, the composition of milk is plastic with respect to numerous inputs, and varies between individuals, delivering varied composition and functions according to time, physiological state of the infant and the mother, and the varying nutritional needs of the infant (reviewed in (6)). How the milk of 1 species, or a component thereof, facilitates beneficial bioactivity in another is yet a further dimension to be explored. Nutrition is beginning to discover that food recommendations cannot be made without taking into account the variation in humans and their environment. A thorough investigation of the causes of the plasticity in milk will reveal pertinent information of how plasticity is developed in biological structures naturally designed through evolution to be foods. Perhaps the most compelling reason to approach milk as a model food fit for structure-function compositional deconstruction is that it is the product of mammalian genes for which we have now acquired the sequence. As such, the whole high-throughput bioinformatic toolbox is available for functional characterization. Research will not abandon its traditional models, strategies, and policies; bioinformatics will, however, dramatically augment them.

The new science of bioinformatics has as a stated goal the annotation of the human genome. This means that scientists are going through each of the genes, 1 by 1, determining what their biological functions to the organism are and then incorporating this knowledge into the intact genomic database. Imagine the exceptional scientific resource that the genome database is becoming with this goal in mind. However, many of the identified proteins in milk are being annotated for the biological activities that they or highly homologous genes are likely to perform in other cells and locations. Although the large scale effort is under way to characterize the genes in the human genome, even those with unknown functions, this research will not necessarily extend to their roles in milk. The goal for annotating milk genes is to assign their probable activities that they exhibit after being incorporated into milk and consumed. An overview of the gene ontology classifiers for some of the major proteins in milk shows there is an interesting collection of molecules, yet the nutritional functions are not obvious. For example, xanthine dehydrogenase/oxidase has been recognized for decades to be present in the fat globule membrane of milk. This enzyme catalyzes the conversion of hypoxanthine to xanthine, and from xanthine to uric acid, but given the substrate concentration and the large quantity of enzyme associated with the milk fat globular membrane, it may have another role in milk. Vorbach et al. (7) recently provided
evidence for the role xanthine dehydrogenase/oxidase plays in milk secretion. Although it is considered a “housekeeping” gene in many tissues, expression in the mammary gland increases during late pregnancy and is at the highest level following parturition. Due to its location, expression, and lack of enzymatic activity, Mather and Keenan (8) proposed that this enzyme might be involved in the reverse endocytotic process of milk fat globule secretion. Vorbach et al. (7) showed that while a homozygous knockout is lethal, a heterozygous mouse is fertile and capable of producing a litter. However, they are unable to maintain lactation and the pups die quickly of starvation. In fact, they were able to show that the mutation prevents the translocation of fat globules through the apical membrane of the epithelial cell into the alveolar lumen presumably because there was an insufficient amount of this protein. It is very difficult to imagine that the catalytic activity, converting hypoxanthine to xanthine, has anything to do with this function. Thus, its role is distinctly different in milk. There are other examples of proteins in milk serving different functions than their annotated cellular roles, despite the fact that this has never been the reason for characterization. Thus, once a protein has been identified as a participant in lactation, either as a constituent of milk or as a biosynthetic enzyme, further work will be needed to investigate whether it is acting in the same capacity as its annotated role.

A systematic investigation of the nutritional functions of milk constituents using modern tools of biology and bioinformatics is contingent upon identifying all the elements present and the incorporation of these constituents into electronically accessible databases. Despite an exhaustive search, the authors were unable to locate any comprehensive list of the proteins in the milk of any species. To identify the genes responsible for milk synthesis, and the proteins included in milk, a range of global non hypothesis-driven approaches are available, such as expression profiling, proteomics, and bioinformatics. However, this will require a concentrated effort to assemble the basic data of milk; i.e., its constituents and the genetic information necessary to produce them, i.e., the milk genome. Identifying the genes responsible for milk can use several approaches: metabolite composition analysis and reverse genetics; gene expression profiling of lactating mammary epithelial cells; and bioinformatics: genomic searching for lactation-specific sequences.

Milk component analysis and reverse gene discovery

The study of the components of milk has been a pursuit of scientists for the past several hundred years. In fact, much of biological science has used milk components as the targets of molecular based research during the past 100 y. Identifying the genes for milk proteins such as the enzymes, lysozyme, and beta lactoglobulin were hallmarks for biological science. Now with advances in proteomics, and metabolite analyses, the full spectrum of milk’s components is being identified and with the components has come a search for the genes that produce them.

Proteomics

A new set of technologies is being assembled to examine the hundreds of proteins present in biological samples simultaneously in a single analytical platform. This new field has been made possible primarily by the rapid advances in the mass spectrometry of peptides. The increasingly accurate mass determinations obtained with modern mass spectrometers make it possible to infer sequence information which, when combined with genetic information, leads to rapid assignment of protein sequences from mass spectral data. These proteomic techniques are amenable to and have already been shown as proof of principle capable of discovery of milk proteomes during lactation and compartments of milk. Studies to date have separated proteins from milk or mammary subsets such as cytoplasmic lipid droplets (9) and the Golgi complex (10). Comparison of gels or chromatographs from different samples, such as lactating vs. nonlactating epithelial organelles, detects the proteins and/or protein modifications that differentiate the samples. Those interesting candidates can then be partially sequenced and identified. An additional advantage promised from this technology is that it can potentially provide phenotypic information, as the analytes have been subjected to post-translational modifications. However, despite the scientific promise of its great analytical utility, the technical obstacles presented by proteomics remain formidable. For example, in milk and in the cells that produce it, the dynamic range of protein concentration is very large. Thus, it is difficult to achieve both separation and detection on a single gel. In addition, it is difficult to achieve consistency in the gels, which makes cross-laboratory comparison difficult. Additionally, 2D gel separation tends to enrich the soluble proteins at the expense of those that are membrane bound, although new technologies and strategies are being developed to address this problem (11). Finally, and perhaps most importantly, proteomics separates and analyzes proteins based solely on their molecular weight and charge and provides no information on their 3-dimensional structure in solution/membranes. Because many of the subtypes of protein functions are dictated by their structures in solution that are in turn modified extensively and dynamically by post-translational modifications, proteomics still remains more of a metabolite analytical platform than a true protein activity tool.

Metabolite profiling

The lipids and carbohydrates in milk are biosynthetic products of proteins expressed in the mammary gland acting on dietary substrates. Although neither class of molecule is necessarily produced in the mammary epithelia, it is the enzymes and transporters in these cells that are responsible for building complex structures from the building blocks supplied to their basolateral membrane. For example, using fluorophore-assisted carbohydrate electrophoresis (FACE), Kelder et al. (12) showed there were substantially more free oligosaccharides in the milk of humans than in a range of ruminants and rodents. The compositional tools of modern analytical chemistry that are capable of solving higher order structures are being applied to the study of milk’s complex lipids, carbohydrates, and small metabolites. These molecules in milk are not the direct products of single genes, but instead are the products of biosynthetic pathways whose catalysts and regulators are the products of “milk genes.” Establishing which enzymes are responsible for embellishing milk with these complex structures, how the enzymes are regulated, their evolutionary origins, and investigating the nutritional roles of their biosynthetic products are other areas of research being addressed by genomics and bioinformatics (13). Exciting dividends of this work are new functional ingredients discovered for foods, the strategies to regulate their abundance in lactating mammals, and the means to produce them in foreign expression systems.
Global gene expression analyses, from arrays to colored beads

Expression profiling is a powerful global analysis tool used to discover all of the genes being transcribed in a particular tissue sample, and to discover the differences in the expression of genes from different samples. This tool set and its application was the first approach taken to facilitate the extraction of information from the lactating mammary gland. The approach has been successful in demonstrating the remarkable number of genes that are differentially expressed in the murine (14) and bovine (15) lactating mammary epithelial cells. Despite the theoretical utility of such an investigation, there are drawbacks to this technology with respect to identifying genes involved in lactation. Perhaps the biggest barrier to such an investigation, as far as human lactation is concerned, is the availability of the lactating tissue required. Although these experiments will be valuable in studying lactation in ruminants and in model organisms such as the mouse and rat, parallel investigations utilizing other platforms will be necessary to fully elucidate the structural and nutritional complexity of the milk of humans. A second drawback in using cDNA expression analysis is the lack of correlation with protein concentration (16), and a lack of information as to the extent and specificity of post-translational modifications. As there are still points of regulation between the synthesis of RNA and the final protein concentration in a cell, gene expression differences alone do not provide clear phenotypic information. Finally, lactation is supported by tissues other than the mammary gland, and the mammary gland itself is composed of multiple cell types, and thus an array would be needed from each tissue/cell type at each time point to get a complete picture of the lactation transcriptome.

Despite the drawbacks, there is a great deal of information that will be learned about lactation using microarray experiments. By measuring the relative transcription of all the genes in the mammary gland at a range of time points throughout pregnancy and lactation, it is possible to identify a large subset of the genes responsible for the synthesis and secretion of milk. Clustering these genes into similar expression profiles will allow the identification of groups that act coordinately to fulfill a specific component of milk synthesis, and will provide candidates for further laboratory validation. In addition, the 5′ (and sometimes 3′) regions of the genes that are coordinately regulated can be investigated to identify which DNA binding motifs constitute the regulatory elements and confer this expression profile. One further usage of microarrays, which will facilitate characterizing the molecular details of lactation, and the range of variation accessible to genetic breeding of agriculturally important ruminants for milk traits, will be in genotyping. Hybridization of mRNA to oligonucleotide arrays provides rapid SNP (single nucleotide polymorphism) data. As scientists move rapidly to understand how different variations in gene sequence relate to variations in milk properties and compositions, the technologies of rapid genotyping will make it possible to dramatically accelerate breeding programs and can be used, for example, to probe the differences in genotype for 2 individuals with different milk phenotypes.

The bioinformatic approach

The most comprehensive strategy to approach identification of the genes that contribute to lactation is through the emerging field of bioinformatics and computational biology. The complete sequence of the human genome is now available, and the mouse and rat are nearing completion. In addition, sequencing of the bovine is underway as is the chimpanzee, and other mammals are soon to follow. With these databases in place, bioinformatics can draw on the large biological knowledge base of lactation and milk to address the task of discovering which of the thousands of genes in any particular genome are responsible for producing milk. The challenge is to find some nucleotide sequence pattern that would suggest a gene is activated during pregnancy or lactation. One attractive aspect of such an investigation is that the discovery would not be limited to a particular tissue. The question then becomes, by integrating the information we have already gathered on milk protein expression and on the hormonal control of lactation, can we identify de novo genes involved in lactation based only on an analysis of their sequence?

To support the idea that this type query is feasible, witness the precise control of lactation-specific gene expression induced by the lactogenic hormones, prolactin, insulin and the glucocorticoids. Although many of these protein activities in milk have been co-opted from other physiologic duties, lysozyme and albumin being 2 examples, others such as α-lactalbumin and the caseins are only expressed in the lactating mammary gland. Each and every human diploid cell contains 2 copies of these genes, yet they are only expressed by females at the proper time and location in response to increasingly well defined hormonal inputs.

Despite the seeming simplicity of the principle, detection of genes with tissue-specific expression is moving forward slowly for the following reasons. Whereas microbial genes are organized into operons, which are coordinately regulated by a small number of transcription factors, eukaryotic gene regulation is far more complex. In order to initiate transcription, a large agglomeration of proteins is needed to complex with the polII RNA polymerase, and these are known as general transcription factors as they are involved in the synthesis of most genes. In addition, specific transcription factors, unique to each gene, bind to the regulatory sequence, or cis region, to properly recruit and position the transcription complex. It is in identifying the binding sites for these specific factors that a sequence-based gene finding program concentrates. However, unlike the coding regions of proteins, these elements are highly degenerate, and often only a portion of the whole motif is conserved. Additionally, whereas prokaryotic genomes are small and the regulatory regions are found in close proximity to the genes, eukaryotic genomes are much larger, and regulatory regions have been located as far as 85 kb downstream from the transcriptional start site.

There are 2 general informatic strategies being used to characterize the regulatory regions of eukaryotic genes given genomic sequence; 1 involves identification of conserved nucleotide binding sites across a range of genes with similar expression profiles, and the other enumerates the binding sites for transcription factors known to participate in the regulation of the genes [reviewed in (18)]. One factor, covered in this reference, that is proving to be particularly helpful in the identification of eukaryotic regulatory regions is phylogenetic footprinting. In brief, alignment of syntetic genomic sequences allows for the detection of nongenic conserved regions. In order for this procedure to be effective, a sufficient amount of evolutionary time must have passed for the 2 divergent species to have accumulated enough mutations in the regions that are not being selected for, so that the conserved regions are obvious. This procedure has been effectively used to detect the regulatory regions in 1 megabase of orthogonal human and mouse sequences (19). One drawback of this method, specifically applicable to the study of milk genes from
mice and humans, is that it can only detect the regulatory regions of genes for functions which the 2 species share, and therefore would miss those which have developed since divergence. Because the composition of milk differs greatly across mammalian species, and because this is a result of the cooption of different gene sets in the respective species, a significant fraction of regulatory regions will be missed by human/mouse comparison. In response to this problem, Bofelli et al. (20) created phylogenetic shadowing, a process of syntenic alignments of genomic regions from a range of primates to detect regulatory regions specific to this order. Although a pairwise alignment is not informative for any 2 given primates due to the high level of conservation, by using 7 or so species the aggregate accumulation of mutations allows for detection of conserved nongenic sequence motifs.

Identifying all the response elements in the cis region of a gene and the transcription factors which bind to them does not in itself provide an explanation for the complex ways in which regulatory regions of genes for functions which the 2 species share, and therefore would miss those which have developed since divergence. Because the composition of milk differs greatly across mammalian species, and because this is a result of the cooption of different gene sets in the respective species, a significant fraction of regulatory regions will be missed by human/mouse comparison. In response to this problem, Bofelli et al. (20) created phylogenetic shadowing, a process of syntenic alignments of genomic regions from a range of primates to detect regulatory regions specific to this order. Although a pairwise alignment is not informative for any 2 given primates due to the high level of conservation, by using 7 or so species the aggregate accumulation of mutations allows for detection of conserved nongenic sequence motifs.

Future directions

As milk contributes to the survival and fitness of mammalian young, it is not simply a participant in evolution, but rather a driver. Perhaps the mammalian strategy of altricial birth, milk provision for postnatal nutrition, and low birth number, combined with a catastrophic meteor impact, are all responsible for the extraordinary success of mammals in the last 220 million years. As the production of milk taxes the metabolism of the mother, evolutionary pressure would tend to select those individuals whose milk represents an optimization between efficiency and effectiveness. Therefore, it is hard to imagine there is much in milk that is not interesting from a nutritional perspective.

We are now in a position in which analytical techniques can realistically determine all of the constituents of milk and those proteins responsible for its production, and can therefore look for the mechanisms through which it exerts its benefit. New strategies will have to be invented to probe the bioactivities of milk's molecules. The payoff will be a previously unknown dexterity in food production and nutrient provision. Although determining the functions of some milk proteins, such as xanthine oxidoreductase, will challenge the deductive powers of scientists, the tools of structure function solution will make even this most curious example more understandable. In fact, as each identified human gene is annotated for function, clues as to why it participates in lactation should emerge. Two entry points into probing the function of milk constituents will be in determining the compositional diversity among mammals, and the plasticity in regards to lactation period. As for ways in which we can glean nutrition information divulged through our study of milk, perhaps the different composition between colostrum and mature milk can serve as a guide. The striking feature of colostrum is the redundancy of its composition with respect to antimicrobial proteins, lipids, and carbohydrates. Clearly, the composition is complementing the immunomaturity of the newborn by insuring the pathogen load is minimized, and this is a strategy that can currently be applied to immunocompromised adults. As biotechnology tools become more sophisticated, milk will continue to provide challenges and rewards in its evolutionary knowledge banks.

LITERATURE CITED


