UCD Conway Institute of Biomolecular and Biomedical Research

&

UCD Conway Institute of Biomolecular and Biomedical Research

2019 - 2020

Investigation of Short Linear Motifs avoided in Major Milk Proteins and their relation to Milk Digestion by Proteolysis

---

Atharva Bhadke

Student number: 19205288

M.Sc. Biological & Biomolecular Science (Negotiated Learning)

Supervisor: Prof. Denis C. Shields

Date of submission: 15 January 2021

CONTENTS

ACKNOWLEDGEMENTS

Firstly, I would like to thank Prof. Denis Shields for kindly accepting me as a project student in his team at Shields Lab, UCD Conway Institute of Biomolecular and Biomedical Research. The missions given by him to carry out the project have helped me massively improve my computational as well as biological skills. I am ever-grateful to Prof. Denis for this.

Moreover, I would like to thank Dr. Jean Manguy, PhD. for assisting me throughout the project, as well as introducing me to the world of R programming. His programming sessions and expertise helped me accomplish the project with a greater computational refinement.

I would also like to acknowledge Professor Richard Edwards from the University of New South Wales (Sydney, Australia) whose laboratory developed and is still maintaining SLiMSuite software, which gives immense information of SHort Linear Motifs.

Lastly, I am thankful to Dr. Joanna Kacprzyk and Dr. Gavin Stewart for helping me sustain throughout the project as my course coordinators.

ABSTRACT

Milk protein digestion is of major importance in mammals, and is a process where the young one of the neonates derives all its nutrition from during the first few months after birth. Protein digestion takes place through two important proteolytic machinery namely proteasome and lysosome. While proteasome breaks down intracellular proteins in a highly specific process via the ubiquitin-proteasome system, lysozyme degrades extracellular proteins. This project analyses the various peptides and enzymes found in the milk protein sequences of humans (*Homo sapiens*) and cows (*Bos taurus*) to check for the presence of certain short, linear motifs or SLiMs. Since the beginning of time, milk has been an essential food for mammals, adn this project aimed to identify the slimes present in milk proteins to check for their distribution and specificity amongst humans and cattle. The common major milk proteins found in both these organisms were picked out, and they were screened against two datasets of motifs, one of which contained 2-lettered dipeptides while the other was a collection of proteolytic enzymes found in the body. The results showed that certain amino acids like alanine and valine were highly avoided in the same, while enzymes like chymotrypsin and pepsin were underrepresented as well. They also showed that certain properties of amino acids such as hydrophobicity, non-aromaticity and non-polarity play a crucial role in milk digestion by proteolysis.

LIST OF ABBREVIATIONS

DD: Dipeptides dataset

DE: Digestive enzyme dataset

SLiMs : Short Linear Motifs

Ub: Ubiquitin

UPC: Unrelated Protein Clusters

UPS: Ubiquitin-Proteasome System

INDEX OF TABLES

INDEX OF FIGURES

INTRODUCTION

Milk contains a wide variety of proteins that adds to its distinctive properties, and after many of them are digested, they provide a balanced source of amino acids for the purpose of the infant's growth. While some proteins such as lipase and amylase help in digestion and utilisation of micronutrients and macronutrients from milk, other proteins such as caseins and lactalbumin resists themselves against proteolysis in the gastrointestinal tract of humans (Lönnerdal, 2003). In humans (*Homo sapiens*) as well as cows (*Bos taurus*), milk proteins help in the growth of 'good' bacteria such as lactobacilli and limit the growth of harmful pathogens by decreasing the pH in the intestinal area (Holton et al., 2014; Ziegler et al., 1990). All in all, the mother's milk provides the essential nutrition to the breat-fed infant which defends the infant against potential infection, while faciaition the development of the necessary physiological attributes of the newborn. Although milk proteins provide the above said advantages, many of them contribute to digestion and uptake of other nutrients from breast milk. Furthermore, along with delivering proteins, sugars, lipids and minerals to the developing neonate, human milk also contains a plethora of indigenous enzymes such as pepsin, which is known to be present in the stomach of infants by 16 weeks of gestation (Holton et al., 2014). As the neonatal GI tract varies in terms of period of maturation, it is suggested that these milk enzymes might fill the voids in the development of enzymatic digestion of milk. Holton et al. (2014) took this point in study and compared the protein cleavage patterns at multiple time points to check for retention of the activities of the digestive enzymes. They analysed the change in the proteins when it reaches the infant's stomach, further cleavage possibility and novel

cleavage patterns form gastric enzymes. It was seen that the enzymes continue their activity in the infant stomach, and displayed increased activity in several cases (Holton et al., 2014). Moreover, it has also been identified that certain peptides play a significant role with the amino acid sequences of native milk proteins, classified as 'bioactive' peptides (Clare and Swaisgood, 2000). They are a result of hydrolysis reactions, for example, those catalysed by digestive enzymes. Bioactive peptides are directly involved in a number of physiological processes such as responses to gastrointestinal, hormonal or immunological signals. These peptides further can be isolated and put to commercial use, for example, phosphopeptides from casein fractions are used in pharmaceutical and dietary supplements. They provide immense therapeutic value for either prevention or treatment of diseases (Clare and Swaisgood, 2000). As is most of the case with bioactive peptides and milk digestion, they involve a lot of protein-protein interactions, which are of fundamental importance and help to carry out a large chunk of the human body's physiological functions.

Major Milk Proteins in Humans (Homo sapiens)

In humans, almost all of the milk proteins are synthesised by the mammary gland, with significant exceptions such as serum albumin, which appears from maternal circulation (Lönnerdal, 2003). Milk proteins are classified in three groups namely; caseins, mucins and whey proteins (Patton and Huston, 1986). In human milk, there is a wide range of bioactive agents, and various targets in the GI tract for those agents have been identified (Goldman, 2000; Goldman et al., 1997; "Koldovsky: Growth factors and cytokines in milk - Google Scholar," n.d.)). Although there is proof of undergone modification in the GI tract in breastfeeding infants, components of breastmilk, along

with interacting with the GI tract, also appear to show significant effects on various other organ systems in humanas (Goldman, 2000). In infants, many of the milk proteins are digested for the same, while some proteins like lipase and amylase help in the digestion of micro- and macronutrients from milk (Hendricks and Guo, 2014). Although some components of human milk are resistant to proteolysis, which is essential for digestion of the proteins, they might promote the growth of 'good' bacteria in the infant's GI tract (Guo and Hendricks, 2008; Lönnerdal, 2003). Nutrient absorption is assisted by proteins thereby allowing breastfed infants to take up the nutrients successfully. Simultaneously, protease inhibitors help sustain the physiological function and stability of the binding proteins, thus allowing efficient nutrient uptake. This is done by limiting the activity of enzymes that break up these proteins (Hendricks and Guo, 2014). Thus, the need for classification of the components of milk proteins is handy, and classifying them helps understand their innate nature and method of action. The types of proteins in human milk are:

Whey Proteins:

1. α-lactalbumin
2. Lactoferrin
3. secretory IgA (SIgA).

Other significant proteins include:

1. lysozyme
2. folate-binding protein (FBP)
3. bifidus factor

4. caseins
5. lipase and amylase
6. α_1-antitrypsin and antichymotrypsin
7. haptocorrin.

The above list of human milk proteins is taken from Hendricks and Guo(2014) and their characteristics are explained below.

1. α-lactalbumin

It comprises of more than 25% of the whey protein content of human milk and is essential in biosynthesising lactose, by binding calcium and zinc ions. It may be involved in facilitating the absorption of divalent cations by generating peptides, and hence increasing the absorption of mineral (Hendricks and Guo, 2014)

2. Lactoferrin

Lactoferrin tightly binds with iron, thus preventing the spread of pathogenic bacteria and thus facilitates the uptake of iron, thereby rendering the iron unavailable to the microflora. By limiting the growth of bacteria by disrupting their breakdown of carbohydrates, lactoferrin slows down the reproduction of harmful organisms such as fungus and spores to prevent illness in infants (Hendricks and Guo, 2014).

3. Secretory IgA (SIgA)

Human milk contains five basic types of antibodies, namely SIgA, IgA, IgM, IgD and IgE. IgA is the most abundant of them all, and is usually found in the form of SIgA. SIgA is composed of two IgA molecules combined with a secretory component. The latter

works as a defense mechanism for antibody molecules, thus safeguarding them from the digestion enzymes and gastric acid. AS the infant's digestive system is extremely sensitive, an excess of the chemicals which ward off diseases may potentially harm it. SIgA helps protect mucosal surfaces even those other than the gut, here most of the milk digestion takes place (Hendricks and Guo, 2014)

4. Lysozyme

An anti-infective agent found in human milk, lysozyme is a glycoprotein which hydrolyses N-acetylglucosamine and N-acetylmuramic acid in bacterial walls (Newburg, 2001). Alike lactoferrin, lysozyme is present in other exocrine secretions as well. Breaking down mostly gram-positive bacteria align with a few gram-negative bacteria, its concentration in milk increases along with prolonged lactation. It is found in higher concentration in human milk as compared to bovine milk, and the value is around 1000-fold in the same (Hamosh, n.d.; Hendricks and Guo, 2014)

5. Folate-binding protein (FBP)

Folate-binding protein, or FBP, is found in the soluble form as well as the particulate form in human milk. The soluble FBP is helpful for surviving proteolytic digestion, and it shows tolerance to low gastric pH. This property has been found common in humans as well as goats in milk digestion. Pickering et al (2013) put forth the theory that FBO could slow down the release and uptake of folate in the small intestine, as the gradual release of folate increases tissue utilisation (Hendricks and Guo, 2014; Pickering et al., 2013).

6. Bifidus factor

Also known as B_{12}-binding protein, is one of the oldest known factors in human milk which provides resistance to diseases. Promoting the growth of the beneficial organism *bifidobacteria*, it is present in a variety of food products such as yogurt and probiotic supplements. As the name suggests, the protein binds with vitamin B_{12} in the intestinal tract and prevents its uptake by harmful organisms (Hendricks and Guo, 2014)

7. Caseins

Caseins are highly digestible milk proteins which form stable aggregates in alignment with calcium and phosphorus. Due to this, they are found in higher concentrations in human milk as compared to other minerals. Caseins form micelles in colloidal dispersions, and their size ranges from 20-55nm as compared to casein micelles in bovine milk. β-casein is the main casein found in human milk, and its nature is highly phosphorylated. During digestion, β-casein forms phosphopeptides which increase calcium absorption by increasing its solubility. This makes the breast milk rich in bioavailability of calcium, a necessary ingredient required for growth of infants, be it human or bovine. Casein phopshoptpeitdes also contribute to the absorption of divalent vacations such as zinc. *K*-casein, another variant of casein, is a highly glycosylated human milk protein which helps in defending the body against infection. It encourages the growth of *Bifidobacterium bifidum*, an acid-producing anerobe which reduces the growth of pathogenic microorganisms in infants. This is due to the C-terminus preotelysis product of *κ*-casein (Hendricks and Guo, 2014).

8. Lipase and amylase

Lipase and amaysiase are digestive enzymes present in milk which also help in absorption of certain micronutrients (Lönnerdal, 2003). When newborns suffer from low lipase activity or poor lipid usage, bile salt-stimulated lipase helps counteract this by hydrolising cholesterol esters, diacylphosphatidylglycerols, di- and triacylglycerols, and micellar and water-soluble substrates. This helps in smooth digestion of the lipids. Amylase, in human milk, is present in a significantly good quantity. It is hypothesised that its high presence could be to make up for the lower amylase activity in the pancreas in neonates. It also helps in digestion of complex carbohydrates when the infant is fed the initial diet after the breastfeeding session (Hendricks and Guo, 2014)

9. α_1-antitrypsin and antichymotrypsin

The above are protease inhibitors in human milk ,and work in tandem to keep away the pancreatic enzymes. In vitro studies have shown that α-1-antitrypsin is capable of preventing lactoferrin proteolytic degradation. Studies however indicate that the influence of α-1-antitrypsin and antichymotrypsin may only delay protein breakdown rather than prevent it, as the overall nitrogen balance of breastfed infants is not significantly affected (Hendricks and Guo, 2014; Lönnerdal, 2003).

10. Haptocorrin

Haptocorrin (once referred to as vitamin B_{12} binding protein) is thought to be the primary means of facilitating the absorption of vitamin B_{12} in early infancy. It binds with vitamin B_{12} to form the haptocorrin complex. This complex can bind to the membranes of the human intestine, where the intestinal cells absorb the haptocorrin-associated vitamin

B_{12}. The synthesis of vitamin B_{12} is promoted later in life by an intrinsic component secreted by the gastric mucosa. Infants are heavily dependent on haptocorrin for the absorption of vitamin B_{12} as they do not have the sufficient amount of intrinsic factor to do the same (Adkins and Lönnerdal, 2001; Hendricks and Guo, 2014).

Table 1. *Biological functions of proteins in human milk (Hendricks, G. M., and M. Guo. "3 - Bioactive Components in Human Milk." ScienceDirect, Woodhead Publishing, 1 Jan. 2014)*

Protein compound	**Biological function**
κ-Casein	Ion carrier, inhibits microbial adhesion to mucosal membranes
α-Lactalbumin	Ion carrier (Ca^{2+}), part of lactose synthase
Lactoferrin	Anti-infective, iron carrier
Lysozyme	Anti-infective
Bile salt-dependent lipase	Production of FFA with antiprotozoal and antibacterial activity
Glutathione peroxidase	Anti-inflammatory (prevents lipid oxidation)
Platelet-activating factor (PAF): acetylhydrolase	Protects against necrotizing enterocolitis (hydrolysis of PAF)
Cytokines	Modulate functions and maturation of the immune system
SIgA	Immune protection
IgM	Immune protection
IgG	Immune protection
IgD	Immune protection
IgE	Immune protection

Major Milk Proteins in cows (Bos taurus)

In this study, a comparison of proteins found in human milk and bovine milk is to be done, to compare the presence of the common SLiMs (described in detail further in this report). Due to its similarity with human milk and abundant nutrients, bovine milk has been researched since decades. Over the course of time, the constituents of bovine milk have evolved mainly due to feeding and breeding, the latter being a crucial determinant in the protein profile of bovine milk. Although the composition of milk varies through different cow breeds, it always contains caseins and whey proteins along with fats and carbohydrate in addition to the necessary vitamins and minerals (Smithers and Copeland, 1998). Overall, bovine milk proteins and related peptides are classified into four different groups as follows (Le et al., 2017):

1. Caseins

a. α_{S1}-casein
b. α_{S2}-casein
c. β-casein
d. κ-casein

2. Serum Proteins

a. α-lactalbumin (α-La)
b. β-lactoglobulin (β-Lg)
c. bovine serum albumin (BSA)

3. Immunoglobulins

4. Minor Whey Proteins

5. Proteose Peptones

These are low molecular weight peptides derived from caseins and well as proteose peptone component 3.

6. Milk Fat Globule Membrane (MFGM) Proteins

Table 2 describes the protein contents of bovine milk as adapted from (Tremblay et al., 2003).

Table 2. *Protein contents of bovine milk (Dupont, D., et al. "Quantitation of Proteins in Milk and Milk Products." Advanced Dairy Chemistry, 27 Oct. 2012.)*

Protein	**Concentration ($g\ L^{-1}$)**
α_{S1}-CN	10
α_{S2}-CN	2.6
β-CN	9.3
κ-CN	3.3
ɣ-CN	0.8
β-Lg	3.2
α-La	1.2
BSA	0.4
Ig	0.8
PP, 8F, 8S	0.5
PP3	0.3
Lactoferrin	0.1
Transferrin	0.1
MFGM	0.4
Total	33

Protein Breakdown

Proteins play a huge role in maintaining the functional repertoire of the cells in the body, and are involved in virtually all critical physiological processes (Neduva and Russell, 2006). Protein degradation causes them to carry out their defined functions, and there are two major defined ways in which proteins degrade, with protein degradation being inherently irreversible. Intracellular proteins are degraded by the proteasome via the ubiquitin-proteasome system via the process of ubiquitination.

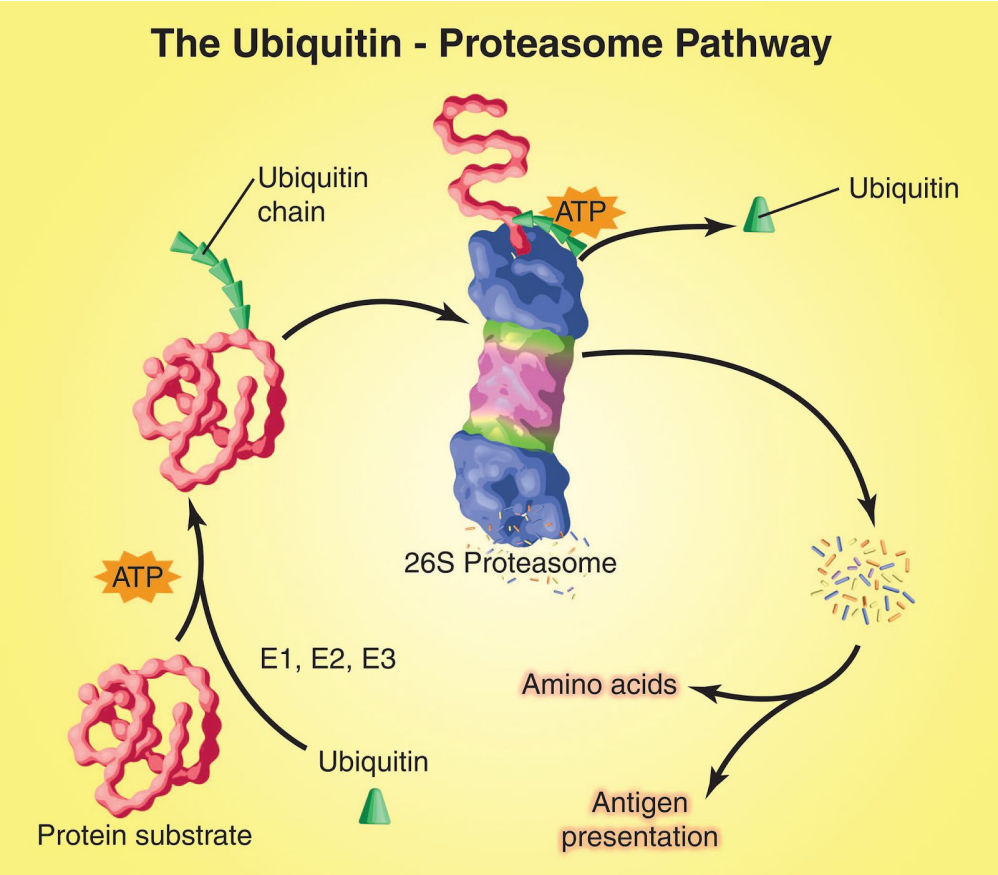

Figure 1. *The ubiquitin (Ub)-proteasome pathway (UPP) of protein degradation (Lecker, Stewart H., et al. "Protein Degradation by the Ubiquitin–Proteasome Pathway in Normal and Disease States." Journal of the American Society of Nephrology, vol. 17, no. 7, 31 May 2006)*

Ub is connected to proteins that are on the way to degrade by an ATP-dependent process that involves three enzymes. A chain of five Ub molecules links to the substrate of the protein and is sufficient for the complex to be identified by the 26S proteasome. In addition to ATP-dependent reactions, Ub is removed and the protein is compacted and injected into the proteasome, where it is digested to peptides. The peptides are degraded to amino acids by peptidases in the cytoplasm or used in antigen presentation. *(*Source: *Lecker, Stewart H., et al. "Protein Degradation by the Ubiquitin–Proteasome Pathway in Normal and Disease States." Journal of the American Society of Nephrology, vol. 17, no. 7, 31 May 2006*)

Figure 1 shows the overall process of the ubiquitin-proteasome pathway. In the process, a substrate protein molecule is tagged with a chain of ubiquitin protein molecules, and this is carried out by multiple enzymes in an orderly process. These are E1 (Ubiquitin activating enzyme),E2 (Ubiquitin conjugating enzyme) and E3 (Ubiquitin ligase). Finally, the adjoining Ubiquitin is removed from the subrstre via deubiquitination and is performed by DUBs, or deubiquitinating enzymes (Lecker et al., 2006; Wang and Robbins, 2014).

On the other hand, extracellular or secreted proteins are degraded by lysosomes. Lysosomes also degrade other cytoplasmic components and protein aggregates outside the cells (Wang and Robbins, 2014)

.

Short Linear Motifs (SLiMs)

Even though many of such interactions are interposed by large domain-domain interfaces, Neduva and Russell (2006) state that around 15% to 40% of the interactions may be mediated by a short, linear motif in one of the binding partners. And because of their degenerate and tiny nature, they are often difficult to characterise and identify (Neduva and Russell, 2006). Short Linear Motifs, or SLiMs, are short (usually less than 10AA long with five or less defined residues), linear (comprised of adjacent amino acids in a protein's primary sequence) and have a motif (a defined sequence pattern, which is necessary for function and recurs in the relevant proteins) (Edwards et al., 2007). They consist of a stretch of amino acids anywhere between 3 to 10 in length, and as less as 2 sites could be a determining factor in the defining activity of SLiMs. Hence, the process of identifying and characterising SLiMs can be extremely complicated (Edwards et al., 2007). According to Davey (2010), SLiMs can p;ay the role of being target sites for proteolytic cleavage, or as sites of post-translational modifications (PTMs). They can also perform the roles of being crucial determinants of subcellular localisation and mediators of protein-protein interactions. For interacting with other proteins in their vicinity, SLiMs also have specific sites for phosphorylation and other modifications within them (Neduva and Russell, 2006). Although they have the ability to encode a functional and a specific interaction with a limited number of residues, they typically mediate interactions between a short linear region in one protein and a globular domain in another (E. Davey et al., 2012; Edwards and Palopoli, 2015). Regarding SLiMs, they are not position-specific intrinsically, that is, a functional residue does not require a specific amino acid at that position for functionality (Davey, 2010). As for their

involvement in physiological processes, SLiMs are involved in myriad of processes such as post translational modification, cell signalling, digestion, cell adhesion, gene expression, membrane binding, subcellular localisation and protein folding (Edwards and Palopoli, 2015). Regarding evolution of SLiMs in protein sequences, there are two key principles as follows: divergent evolution, where individual SLiM occurrences are conserved, and convergent evolution, wherein there is an independent evolution of SLiM occurrences in unrelated proteins (E. Davey et al., 2012). With respect to their sites, they are mostly found in intrinsically disordered regions of proteins, at least in their unbound state (E. Davey et al., 2012). SLiMs are assessed at the dataset level so that scientists can explore the function of the particular dataset through known motifs, or explore the possible function of the motifs through their distribution in the dataset . Majority of the time, a particular motif is analysed for overrepresentation in most experiments, but this project assesses the underrepresentation for a set of motifs involved in milk protein digestion (Edwards and Palopoli, 2015). Hence, common motifs which are involved in milk digestion had to be carefully picked out. According to Davey et al. (2012), finding out novel classes of shared SLiMs among proteins with a common function in complicated, as the signal is very weak due to the minute nature of SLiMs and can be easily overshadowed by the vast potential of false positives due to the same reason (E. Davey et al., 2012). Disordered regions of proteins tend to have a weak three-dimensional structure to be seen experimentally and show distinct bias towards particular amino acids. They seem to be enriched in amino acids Proline (P), Lysine (K), Serine (S), Glutamic acid (E), Glycine (G) and Glutamine (Q), and tend to be depleted in C (Cysteine), F (Phenylalanine), I (Isoleucine), L (Leucine), W (Tryptophan) and Y

(Tyrosine). A detailed explanation of the individual amino acids found in the resulting underrepresented motifs has been given in the discussion.

In this project, the avoidance, or under representation of SLiMs involved in milk protein digestion was analysed, by screening the proteins through two datasets of dipeptide motifs and motifs representing certain digestive enzymes important for digestion in the stomach and the surrounding area near the gut. This would show the motifs less involved in the process of breaking down of proteins and hence help understand milk digestion better.

A hypothesis regarding the presence of motifs was also assumed. Milk has been evolutionarily developed to provide optimum nutrition for the neonate which helps in its essential physiological growth. The hypothesis states that there are more motifs underrepresented in the milk proteins, which have been taken as cases as compared to secreted proteins in the stomach and surrounding area, the latter considered as controls. The thought behind this concept was that if there were a higher number of cleavage sites, in this case, SLiMs, in the milk proteins, they would cause quicker breakdown of milk proteins which is not feasible for the baby nutritionally as well as evolutionarily. In teh beginning of the child's lifetime, a lower rate of proteolysis would be favourable so that it could get the necessary immunity from milk's contents. A stronger rate of proteolysis would mean that no antibodies would be formed during milk digestion in the infant's body, and it could be hazardous in the long run for it. Moreover, the secreted proteins are not specifically as important as the milk proteins in the neonate, hence there would be a greater number of motifs which delay the process of proteolysis present in them as compared to milk proteins.

MATERIALS AND METHODS

Datasets

Protein sequences: For the case studies, major milk proteins from humans (*Homo sapiens*) and cows (*Bos taurus*) were identified from the UniProt database, a widely-used and known database where millions of protein sequences are present and updated regularly according to latest research (The UniProt Consortium, 2015). Further, the common proteins in this selection were taken to compare against a control dataset of secreted proteins found in the stomach, as milk digestion majorly occurs in the stomach (Holton et al., 2014). As for the controls, common secretory proteins found in the stomach of both humans and cows, so that the digestion patterns could be compared to. The main problem faced was the small size of the dataset. Although this was a drawback regarding getting statistical results closer to accuracy, the small size meant a lesser threshold of false positives and the quicker time it took to run computationally. As the focus was on milk digestion in the stomach, a similar sized control dataset was taken. The cases and the controls (common in both humans and cows) are tabulated below:

Table 3. *Dataset of cases and controls*

CASES			
Sr. no.	**Protein**	**Organism**	**Uniprot ID**
1	Alpha-lactalbumin	*Homo sapiens*	P00709
		Bos taurus	P00711
2	Alpha-S1-casein	*Homo sapiens*	P47710
		Bos taurus	P02662
3	Beta-casein	*Homo sapiens*	P05814
		Bos taurus	P02666
4	Kappa-casein	*Homo sapiens*	P07498
		Bos taurus	P02668
5	Albumin	*Homo sapiens*	P02768
		Bos taurus	P02769
6	Lactotransferrin	*Homo sapiens*	P02788
		Bos taurus	P24627
CONTROLS			
Sr. no.	**Protein**	**Organism**	**Uniprot ID**
1	Gastrin	*Homo sapiens*	P01350
		Bos taurus	P01352
2	Pro-glucagon	*Homo sapiens*	P01275

		Bos taurus	P01272
3	Secretin	*Homo sapiens*	P09683
		Bos taurus	P63296
4	C-X-C motif chemokine	*Homo sapiens*	Q6UXB2
		Bos taurus	A4IFR0
5	VIP peptides	*Homo sapiens*	P01282
		Bos taurus	P81401
6	Spexin	*Homo sapiens*	Q9BT56
		Bos taurus	Q0VC44

The function of the cases are discussed in the introduction, however, the controls were selected at a later stage after trying out other controls such as protein with amino acids with similar length as that of the cases, non-secretory proteins and proteins involved in milk allergens. However, as milk digestion was to be studied, the current control dataset of secreted proteins, most of them which are found in the stomach was considered. Theri brief function is as follows, as found on UniProt:

Table 4. *Functions of the control proteins*

CONTROL PROTEIN	FUNCTION	SOURCE
Gastrin	Stimulates the production and secretion of hydrochloric acid by the stomach mucosa and the pancreas to secrete its digestive enzymes	(Bundgaard et al., 1995)
Pro-glucagon	It plays a crucial function in the synthesis and homeostasis of glucose.Regulates blood glucose by increasing gluconeogenesis and decreasing glycolysis. Also reduces food intake. Inhibits gastric emptying in humans.	(Drucker, 2003; Kieffer and Habener, 1999)
Secretin	This hormone is involved in various processes such as regulation of the pH of the duodenal content, food intake and water homeostasis.Also plays a key role by acting as a gastrointestinal hormone by regulating the pH of the duodenal content	(Afroze et al., 2013)
C-X-C motif chemokine	It acts as a chemoattractant for monocytes, macrophages and dendritic cells . Plays the role of anti-inflammation in the stomach	(Lee et al., 2013; Pisabarro et al., 2006)
VIP peptides	They increase glycogenesis hence relaxing the smooth muscles of the stomach, the gallbladder and the trachea	(Ma et al., 2004)
Spexin	It plays a major role in energy metabolism and storage, by inhibiting proliferation of the adrenocortical cells with slight stimulation on corticosterone release. Intraperitoneal administration of the peptide induces a reduction in food consumption and body weight. Also stops the uptake of long chain fatty acid into adipocytes	(Kim et al., 2014; Mirabeau et al., 2007)

Motifs

Two datasets of dipeptide motifs were generated for comparison, namely DD and DE. They DE datasets, or the 'Digestive Enzyme' dataset, is a selection of digestive enzyme preferences which are involved in the breakdown of proteins for the digestion of nutrients in the stomach. It is a set of regular expressions which corresponds to a particular protease as tabulated below (Betts and Russell, 2003).

Table 5. *Explanation of the DE motif dataset*

MOTIF	DESCRIPTION
R	Avoided in Pepsin P1
P	Avoided in Trypsin and Chymotrypsin P1'
[FLYWM]	Chymotrypsin and Pepsin preferred P1 (Aromatic + LM)
[KR]	Trypsin preferred in P1
[STNQCH]	Polar
[ILVMAG]	Non-polar, non-aromatic
[DE]	Negative

Regarding the DD motif dataset, it consists of 400 possible combinations of all the 20 amino acids (20 x 20 =). According to Davey (2010), much of the defined position of SLiMs are degenerate. That is, a functional residual does not require a specific amino acid at a particular position for functionality. Hence, this dataset was considered. Examples of DD motifs are ‘AA’, ‘KL’, ‘NP’... etc.

Execution of Analyses

SLiMProb, a computational tool used for searching SLiMs was used to ‘run’ the motifs as regular expressions against the cases and the control sequences. It is built on Python, and allows the user to search for motifs in the particular dataset (Edwards et al., 2020). It is based on SLiMFinder as explained in the introduction (Edwards et al., 2007), and improves a step upon it by an updated algorithm for refined statistical values. It produces the underrepresented and overrepresented motifs in a particular dataset using regular expressions, adn uses UPC (Unrelated Protein Clusters: clusters of proteins derived from a whole genome) correction via anaysign evolutionary relationships of the input sequences. It uses BLAST for the same, and clusters proteins such that no protein in an UPC detects homology with a protein in another UPC (Altschul et al., 1990; Edwards and Palopoli, 2015). It calculates various parameters as shown:

4.3. SLiMProb summary table

In addition to the basic information about the dataset and motif (Table 4.2), additional statistical calculations are made based on the SLiMChance algorithm of SLiMFinder. Four statistics are calculated (*N*, *E*, *p* and *pUnd*) for each of three levels of motif occurrence (Occ, Seq and UPC):

- *N* = The observed number of occurrences of the motif in the dataset
- *E* = The expected number of occurrences of the motif in the dataset, given the motif and dataset composition. (If the background=FILE option is used, motif occurrence in the background dataset will be used to generate the *E* value instead.)
- *p* = The probability of seeing the observed number of occurrence (*N*) *or more*, given the expected number of occurrences, *E*. For UPC and Seq calculations (below) the Binomial distribution is used for this calculation, using the mean probability of a motif occurrence for each Sequence/UPC, as implemented in SLiMFinder. For the occurrence calculation, the Poisson distribution is used to calculate *p* given *E*.
- *pUnd* = The probability of seeing the observed number of occurrence (*N*) *or less*, given the expected number of occurrences, *E*. For UPC and Seq calculations (below) the Binomial distribution is used for this calculation, using the mean probability of a motif occurrence for each Sequence/UPC, as implemented in SLiMFinder. For the occurrence calculation, the Poisson distribution is used to calculate *p* given *E*.
- Occ = Calculations are made based on the total number of occurrences of the motif in the dataset, irrespective of what sequences the occurrences are in and any evolutionary relationships between them.
- Seq = Calculations are made on a sequence-by-sequence basis, such that the number of occurrences within each sequence does not matter, only whether a sequence contains *any* occurrences of the motif or not. Evolutionary relationships are not considered. (This is of most use for the *N_Seq* value.)
- UPC = Calculations are made according to the number of UPC a motif occurs in, adjusting for evolutionary relationships as performed by SLiMFinder. If all sequences are unrelated (or efilter=F), this should be the same as Seq.

Example. *E_Seq* is the expected number of sequences containing 1+ occurrences of the motif, ignoring evolutionary relationships.

Figure 2. *SLiMProb Summary Table (Edwards et al., 2007)*

In addition to the above values, SLiMProb calculates the motif avoidance probability based on Poisson law =$P(X \leq N\_Occ)$, X being a Poisson random variable with $\lambda = E\_Occ$ parameter. To improve the calculation of this motif avoidance probability, a 'pUnd_Occ_UPC' statistical calculation, implemented by Dr. Denis Shields (UCD) was applied in the calculator of this project.

$$\text{pUnd\_Occ\_UPC} = P(X \leq N\_Occ\_UPC)$$

X being a Poisson random variable with λ=E_Occ_UPC parameter

$$\text{With } N\_\_ = N\_ \times \frac{N\_\_}{N\_Seq} \quad \text{and } E\_\_ = E\_ \times \frac{E\_\_}{E\_Seq}$$

pUnd_Occ_UPC refined the results of getting avoided motifs in sequences, and is useful while comparing multiple datasets, in this case, the cases and the controls. It gives the probability of a given motif to be underrepresented by chance in the dataset, and is considered as another version of the p-value, or the 'true' p-value. This is because the default 'pUnd_Occ' statistic is calculated based solely on the sequence similarity observed in the datasets, adn not on the number of occurrences of a certain motif in the particular data set. As the sequences were related in this case the observed number of occurrences of the motifs were considered as $N\_\_ = N\_ \times \frac{N\_\_}{N\_Seq}$ while the expected number of occurrences were considered as

$E\_\_ = E\_ \times \frac{E\_\_}{E\_Seq}$, according to the new pUnd_Occ_UPC statistics. These calculations were done via R programming and integrated with the datasets in the initial stage itself. Overall, the statistic indicates that higher the value of pUnd_Occ_UPC, the greater the probability that the motif is underrepresented (by chance) and vice versa. For in-depth information, the SLiMProb manual and its core files field can be referred via SLiMProb V2.5.1.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a multivariate technique that takes a data table which contains various observations described by interrelated quantitative variables and analyses the same. By extracting the information from the table, it gives a visual output in the form of a well-defined graph which projects a new set of orthogonal variables known as principal components. Mathematically, the calculations depend upon the decomposition of the eigenvalues of the positive semidefinite matrices and upon the singular value decomposition of regular matrices (Abdi and Williams, 2010). The main aim of PCA is to extract the most crucial information from the data table, compress this information, and project the compressed, yet simplified information to allow the user to analyse the structure and distribution of the variables.

In this project, PCA was carried out on the final SLiMProb results after getting the underrepresented motifs, when the R programming was done succeeding the SLiMProb runs of the protein datasets against the motifs. “R” software was used again for the same, with the emphasis being on the FactoMineR package (https://cran.r-project.org/web/packages/FactoMineR/index.html).

RESULTS

SLiMProb interpretation

The main motive of the project was to identify the underrepresented motifs which were involved in milk digestion in cattle and humans. In order to do this, initially, the datasets of the protein sequences (cases and controls) were run against the common dataset of the motif files (DD and DE) in SLiMProb (Edwards et al., 2007). As a result four tables, two of which consisted of the motifs present in the cases against the DD dataset, and two of the DE dataset were obtained. The motifs were then conjoined individually to create a complementary motif, for example, if the motif [FLYM][DE] was to be screened against the dataset, the motif [DE][FLYM] was also screened for for the detection of motifs in the dataset in another combination. This was done to improve the accuracy of prediction. Moreover, most of the defined positions do not specifically require an amino acid at that position for functionality, instead, requires one from a particular set of amino acids (Van Roey et al., 2014), adding to the explanation of the step. All of these files and the files mentioned below have been attached in the supplementary files.

Thereafter, these files were treated with R programming, wherein the N_Occ_UPC and E_Occ_UPC values were calculated. These values give the number of occurrences of them in the Unrelated Protein Clusters (UPCs) of the dataset and the expected occurrences in the dataset respectively. Thereafter, their ‘Ratio’ was calculated for each dataset. ‘Ratio’ here refers to Ratio = N_Occ_UPC/E_Occ_UPC. This calculation would give a comparable value totally to find out the difference between the actual number of

underrepresented motifs against the expected ones. Further, the Poisson distribution was calculated based on the 'ppois' formula of the calculator in R, which is a cumulative probability function. This gives the probability that the random variable, in this case, the motif detection, will be lower than or equal to a value. This was considered as we had to check for underrepresentation of motifs.

Table 6. *SLiMProb results and analysis of the human dataset*

Dataset	Motif	N_Occ_UPC	E_Occ_UPC	Ratio	pUnd_Occ_UPC
case_human	AN	1	6.583304	0.151899	0.01049
case_human	GV	1	5.33784	0.187342	0.030461
case_human	LE	6.4	13.17318	0.485836	0.023382
case_human	VC	1	5.390531	0.18551	0.029138
case_human	[FLYWM][DE]	25.6	42.6171	0.600698	0.002492
case_human	[ILVMAG][STNQCH]	121.6667	152.5833	0.797378	0.004716
control_human	[FLYWM][KR]	10.66667	17.42825	0.612033	0.040103
control_human	[KR][FLYWM]	10.66667	17.42825	0.612033	0.040103
control_human	[KR][ILVMAG]	17.33333	26.33367	0.658219	0.035975

Furthermore, Bonferroni correction was done, which states that for n independent tests in an test, the nominal significant p-value for each test should be reduced from 0.05 to

0.05/n toto make sure that the probability of one significant effect is 0.05 (Weisstein, n.d.). The p-value threshold was taken as 0.000125 for the DD dataset (0.05/400) and 0.00102040816 for the DE dataset (0.05/49). As the dataset was small, moreover, we were checking for underrepresentation of motifs, it meant that there was a very limited space to work with. Hence, Bonforreonicorrection menat absolutely no results if applied. Hence the standard threshold of the p-value < 0.05 was considered, as looking through the result still now, the p-vlaues ranged from 0.01XX to 0.94XX. If the p-values were close enough to either side of the 0.05 threshold, some other alternative would have been considered (Di Leo and Sardanelli, 2020). Note: the p-value is the value of pUnd_Occ_UPC or the true p-value, as explained in the Materials and Methods section. Thereafter, teh underrepresented motifs were filtered and plotted for better visualisation using the ‘ggplot2’ package in R. The results are shown in Table 5 and are plotted in the figure below.

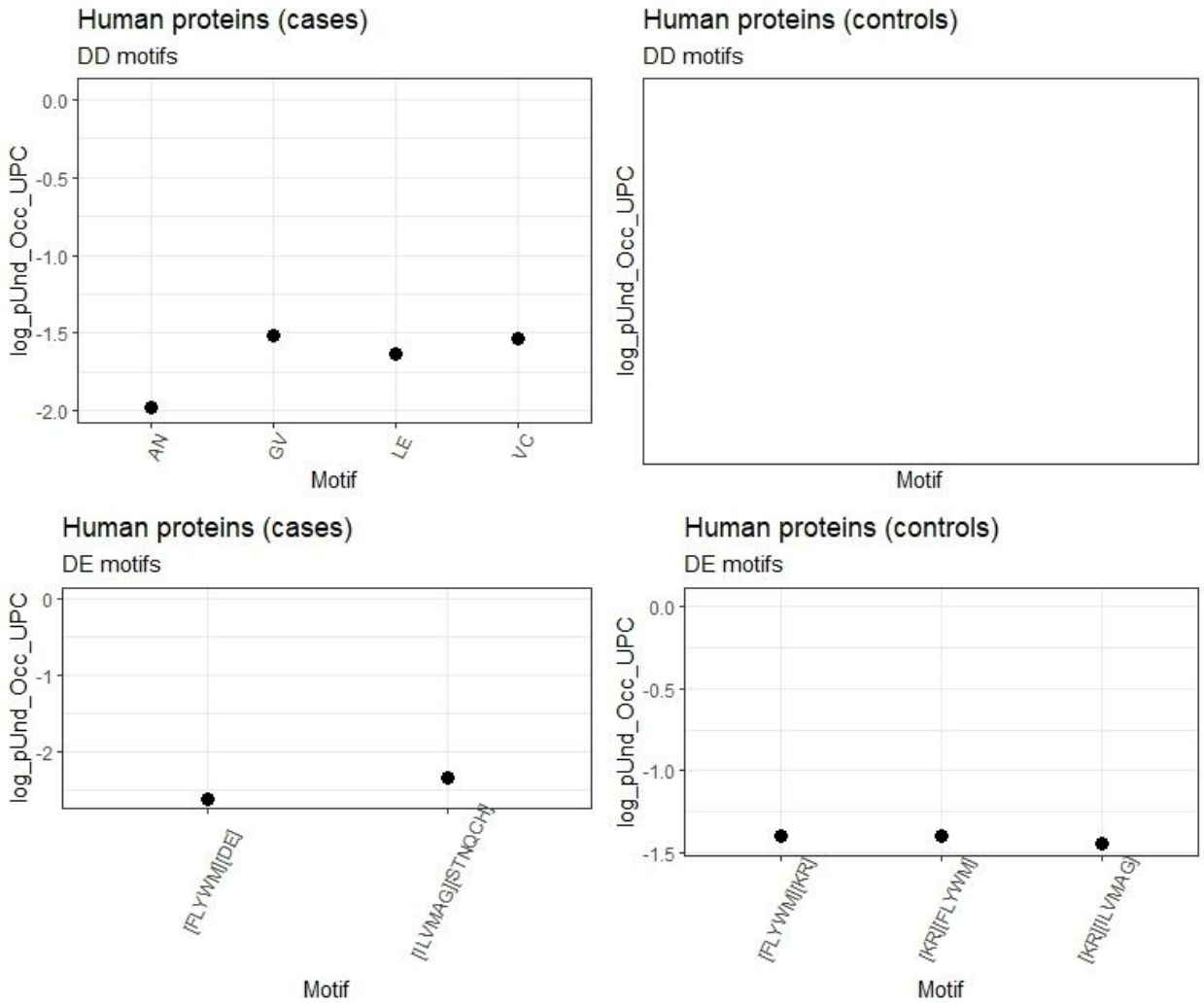

Figure 3. *SLiMProb runs and analysis of the human protein dataset*

Figure 3 shows that the dipeptide motifs AN, GV, LE and VC were heavily underrepresented in the cases in humans, while the control dataset did not show any underrepresented motifs. As for the DE motif dataset, Chymotrypsin and Pepsin, along with negatively charged proteins were underrepresented in the cases along with polar, non-polar and non-aromatic proteins. As for the controls, Chymotrypsin and Pepsin, along with Trypsin and its reverse peptide, and Trypsin and non-polar and non-aromatic proteins were underrepresented.

The same procedure was done with the bovine datasets, and the following results were obtained:

Table 7. *SLiMProb results and analysis of the bovine dataset*

Dataset	Motif	N_Occ_UPC	E_Occ_UPC	Ratio	pUnd_Occ_UPC
case_bovine	AA	5	12.51662	0.399469	0.014666
case_bovine	NA	1	5.347639	0.186998	0.030211
case_bovine	RV	1	4.997222	0.200111	0.040521
case_bovine	SV	3.75	7.881271	0.475812	0.045908
case_bovine	[DE][FLYWM]	32.8	45.65126	0.71849	0.021322
case_bovine	[FLYWM][DE]	30.83333	45.65126	0.67541	0.009106
case_bovine	[ILVMAG][STNQCH]	132.5	155.5833	0.851634	0.029665
case_bovine	[R][ILVMAG]	15.75	23.24884	0.677453	0.047012
control_bovine	[FLYWM][R]	4.2	11.06347	0.379628	0.014471

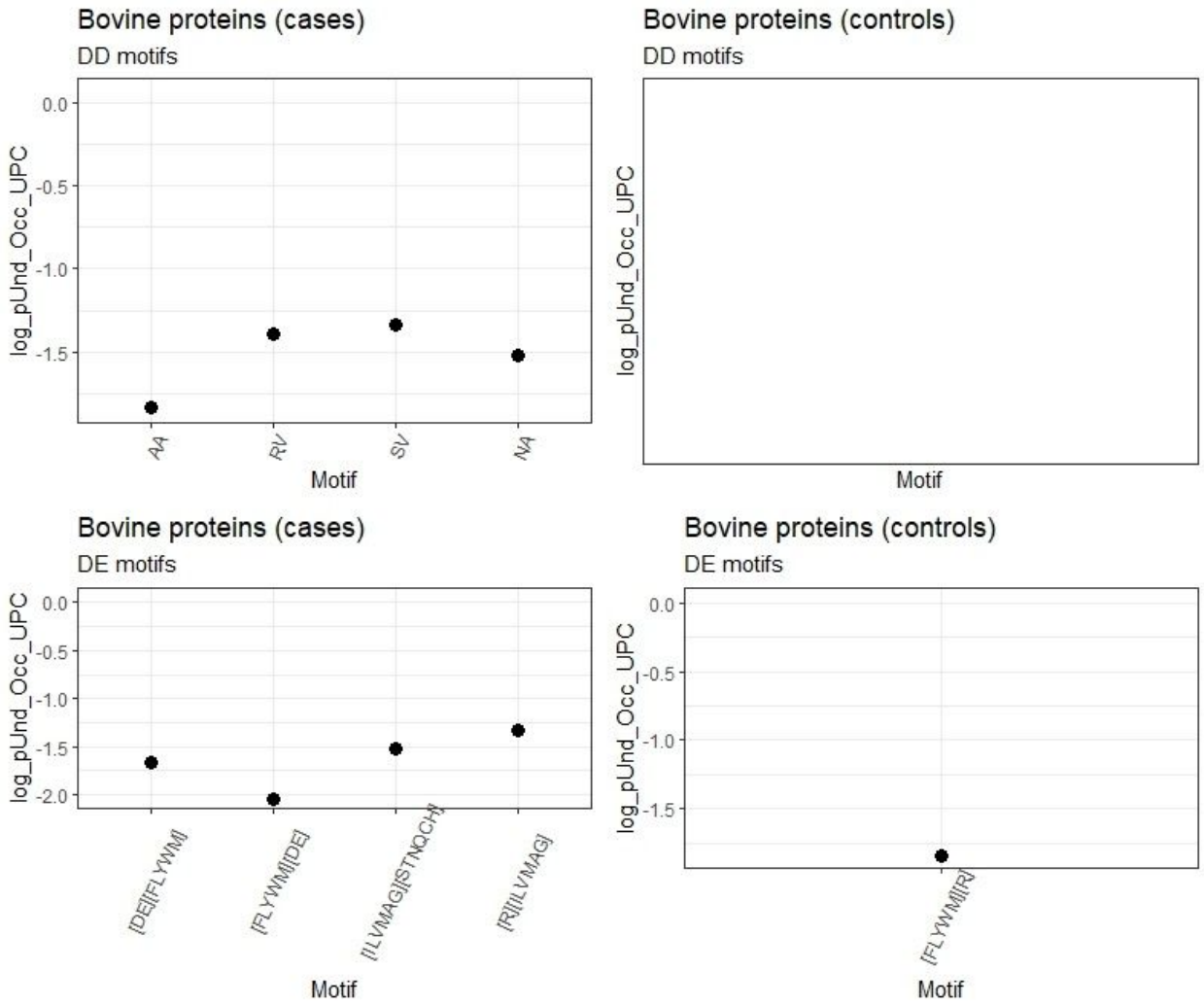

Figure 4. *SLiMProb runs and analysis of the bovine protein dataset*

As seen above, no dipeptide motifs were seen in the control dataset of bovine proteins as well. AA, RV, SV and NA were underrepresented in the cases, while Negatively charged, Chymotrypsin and Pepsin; Negatively-charged, trypsin; polar and non-polar,non-aromatic non-polar,non-aromatic (avoided in pepsin) were observed in the case. In the controls, only chymotrypsin and pepsin along with trypsin were observed.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was performed on the above obtained results to find out which motifs are most valuable when the data is clustered, which will give an idea of underrepresentation of the motifs. In PCA, the data is centered on the origin (Refer Figure 5, 6) and a line is tried to fit to it corresponding to one item (motif on the graph). If it does not fit, it is 'rotated' to such a point that it has a corresponding line (axis) that passes through the origin which fits the best in the case. The points on the graph are categorised according to the quantitative values, adn in this case, the self-calculated values in the Unrelated Protein Clusters (UPCs) like N_Occ_UPC, E_Occ_UPC, pUnd_Occ_UPC with the earlier Poisson statistics and the considered p-value taken in the calculations. PCA either measures the distance from the data points to the specific line or tries to find a line that minimises those distances. When these distances get larger, the lines fit better. In other words, it maximises the distance from the projected point of the origin and calculates the distance. This Is an overall explanation, adn in theory, a lot of calculation with the components such as squaring up the values and summation take place which can be explored in detail in (Abdi and Williams, 2010)). Below is the graph of the PCA analysis for the motifs underrepresented in the human protein sequence dataset.

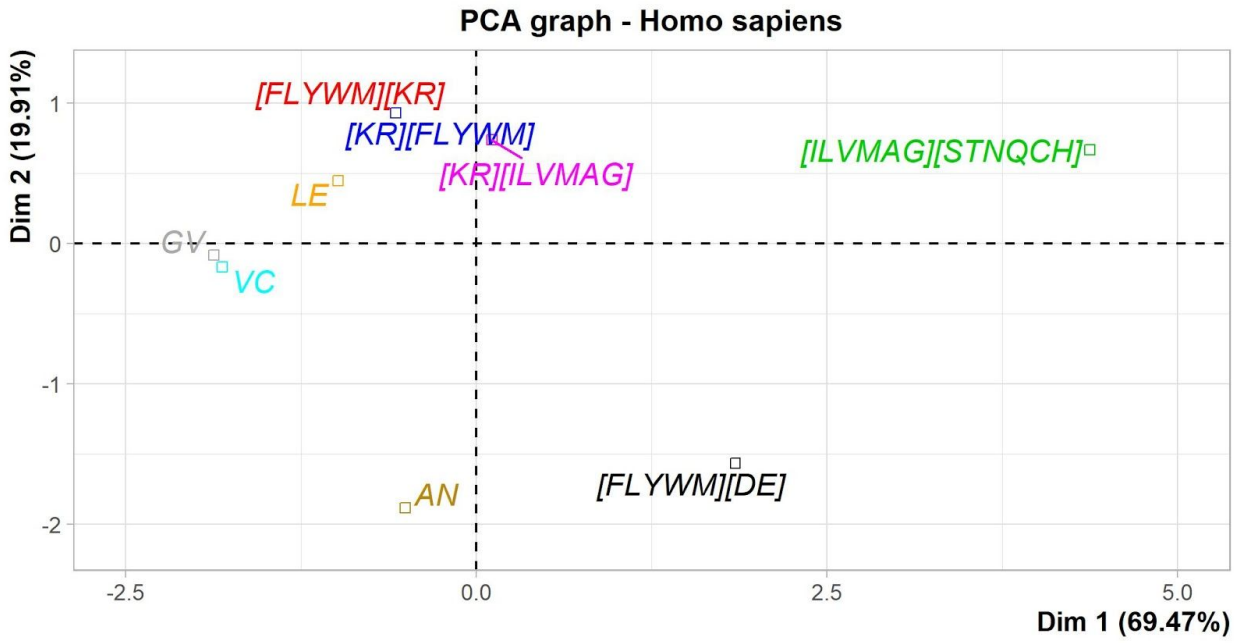

Figure 5. *PCA graph of underrepresented motifs in the human dataset*

Here, the motifs that are highly correlated cluster together while the axes (PC1/Dim1) and PC2/Dim2 are ranked in order of importance. Differences along Dim1 are more significant than those along axis Dim2. We can observed that the motifs [KR][FLYWM], [FLYWM][KR], [KR][ILVMAG] and LE were clustered closely, whiel CV and VC did the same nea the origin.AN, [FLYWM][DE] were in the negative axis of Dim2, but the latter showed a higher value along Dim1, which is of significant importance. [ILVMAG][STNQCH] was an outlier shown in the positive dimension along both the axes.

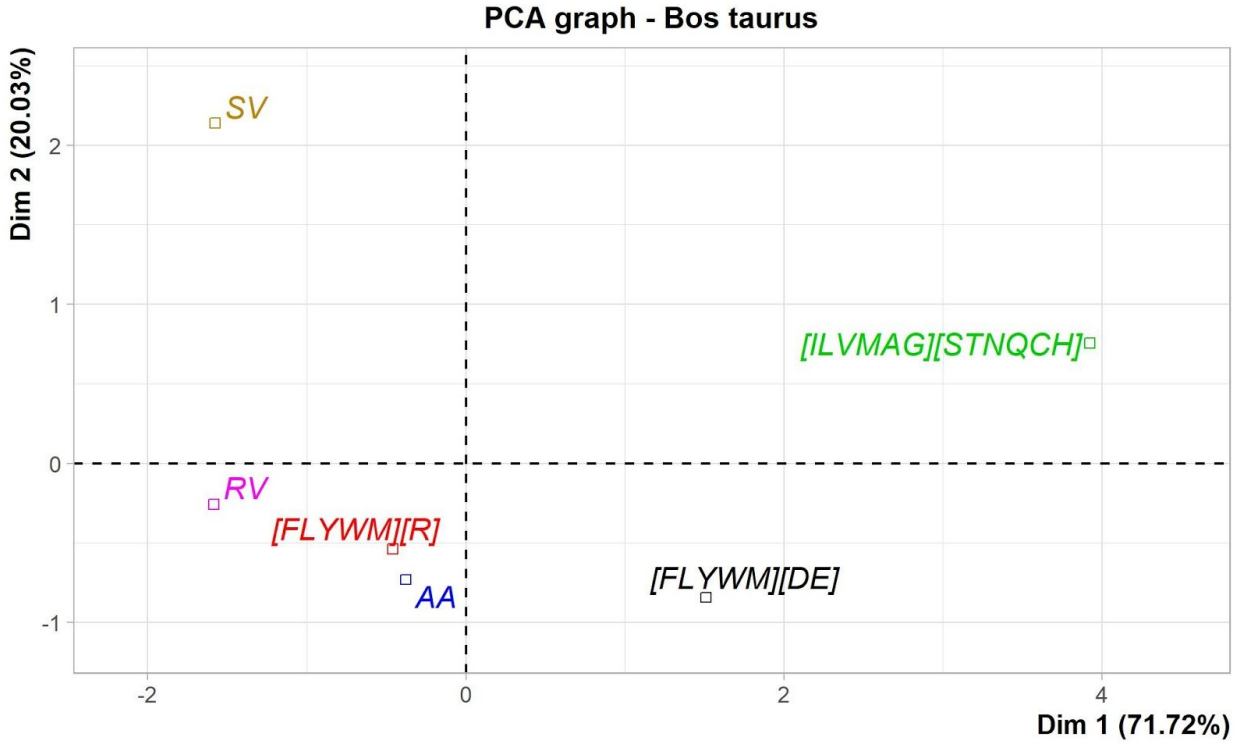

Figure 6. PCA graph of underrepresented motifs in the bovine dataset

[FLYWM][R], AA and RV were seen in closer to the origin in negative axes of both the dimension, while SV was observed as a positive outlier along Dim2 and negative laong Dim1. Again, in this case [FLYWM][DE] and [ILVMAG][STNQCH] were observed almost at the same position as seen in *Homo sapiens*, and this will be analysed upon in the discussion. These motifs have been significantly underrepresented in the same pattern in both the organisms, which shows a certain trend among motifs involved in milk digestion in both humans and cows.

DISCUSSION

The results show that in humans, more number of underrepresented motifs (6/9) are present in cases, the milk proteins (6/9) while the controls, or the secretory proteins have a lesser number of underrepresented motifs (3/9). The same trend is found in the bovine dataset, wherein 8 out of the 9 underrepresented motifs are found in cases, and just one in controls. For ease of discussion ,the under represented motifs are again tabulated below with their abbreviations.

Table 8. *Underrepresented motifs in Homo sapiens*

Dataset	Underrepresented Motifs
case_human	AN
case_human	GV
case_human	LE
case_human	VC
case_human	[FLYWM] [DE] Chymotrypsin & Pepsin (aromatic) + Negatively charged
case_human	[ILVMAG][STNQCH] Non-polar, non-aromatic + Polar
control_human	[FLYM][KR] Chymotrypsin & Pepsin (aromatic) + Trypsin
control_human	[KR][FLYM] Trypsin + Chymotrypsin & Pepsin (aromatic)
control_human	[KR][ILVMAG] Trypsin + Non-polar, non-aromatic

Table 9. *Underrepresented motifs in Bos taurus*

Dataset	Underrepresented Motifs
case_bovine	AA
case_bovine	RV
case_bovine	SV
case_bovine	NA
case_bovine	[DE][FLYWM] Negative + chymotrypsin and pepsin
case_bovine	[FLYWM][DE] Chymotrypsin and Pepsin + Negative
case_bovine	[ILVMAG][STNQCH] Non-polar, non-aromatic + polar
case_bovine	[R][ILVMAG] Non-polar + avoided in Pepsin
control_bovine	[FLYWM][R] Chymotrypsin and Pepsin + avoided in Pepsin P1

Firstly, the underrepresented SLiMs seen above represent a digestive enzyme, as described below the amino acid expression (Betts and Russell, 2003). The higher the number of underrepresented SLiMs, the lower the presence of those particular SLiMs is found in the dataset. It is seen that more SLiMs are avoided in the cases (Milk proteins) than in the controls (secreted proteins). This difference in motif depletion could be characterised by a higher selection pressure in milk proteins as compared to secreted proteins. In the case of milk digestion, if milk proteins were to be broken down quickly, they would show motifs that are hard to digest by the proteome. On the flipside, if the case was reversed, i.e. if there were a fewer number of depleted motifs in milk proteins, it would mean that no milk protein is hard to digest by the particular digestive enzymes.

For better interpretation of the results, let us study the properties of all the individual amino acids, as the motifs are highly dependent on the physicochemical properties of amino acids for their function (E. Davey et al., 2012). The epalantions are by the alphabetical order of the names of the amino acids.

Alanine (A)

Alanine substitutes with other small amino acids, and is one of the least active amino acids. It contains a normal Cβ carbon atom, and is hence generally hindered with respect to confirmation based on other amino acids. It is not specifically hydrophobic, and is nonpolar in nature. Hence, it is present in almost all non-critical protein functional activities. It plays a crucial role in sucrate recognition, importantly with interactions with other non-reactive atoms surrounding it (Betts and Russell, 2003).

Arginine (R)

It is a polar and a positively charged amino acid which prefers to substitute for other positive amino acids such as lysine. In certain cases, it also resists a change to other polar amino acids. It is considered an amphipathic amino acid as one of its side chains nearest to the backbone is long, hydrophobic and contains a carbon; while the end of the side chain is positively charged. arginine is involved in stabilising hydrogen bonds for greater protein stability, and it interacts with negatively charged non-protein atoms (Betts and Russell, 2003).

Asparagine (N)

Being a polar amino acid, it mostly substitutes for other polar residues. It often substitutes for asprate, which contains an oxygen atom in place of the amino group in asparagine, hence the name. Being polar and hydrophilic, it is often exposed to an aqueous environment with it being present on the surface of proteins. The polar side-chain of asparagine is essential for interacting with other polar/charged atoms. Often, it plays a similar role to aspartate in some proteins (Betts and Russell, 2003).

Aspartate (D)

Aspartate, or Aspartic acid, is a negatively-charged, polar amino acid. Like asparagine, it also tends to be represented on the surface of proteins and exposed to an aqueous environment (hydrophilic). When buried within the protein, aspartates are often involved in salt-bridges where they bond with a positively charged amino acid to create stable

hydrogen bonds, leading to stability of the protein. It has a negative charge, and hence interacts with positive non-protein atoms like zinc (Betts and Russell, 2003).

Cysteine (C)

Cysteine can tolerate minute substitution with other small amino acids, but does not prefer to substitute with a major amino acid. Its role depends on the cellular location of the protein it is housed in. Although, it is mostly found in extracellular proteins where it is involved in the formation of disulphide bonds when it undergoes oxidation to do the same. This provides stability to the protein structure and makes cysteine an abundant amino acid found in protein binding sites (Betts and Russell, 2003).

Glutamine (Q)

Gulatiemine is a polar amino acid, which prefers to substitute for other polar residues. It differs from the amino acid Glutamate in such a way that it contains an amino group in the place of the oxygen atom in Glutamate. Similar to asparagine and aspartate, it prefers to be on the surface of proteins being open to an aqueous environment around it. It is also involved in protein binding sites and interaction with other polar or charged atoms (Betts and Russell, 2003).

Glutamate (E)

Glutamate, or Glutamic Acid, is a negatively charged, polar amino acid. It also prefers to be on the surface of proteins and exposed to an aqueous environment. Alike Aspartate, it also is invoked in stabilising the hydrogen bonds in a protein when present within the

protein by being active in salt-bridge formation.Being negatively charged, it too interacts with positively charged cations such as zinc. Its role is very similar to that of aspartate (Betts and Russell, 2003)

Glycine (G)

Unlike all amino acids, Glycine contains a hydrogen atom on its either slides, rather than a carbon chain. Hence, a greater level of conformational flexibility is observed in glycine and a glycine residue can be found in places of the protein structure where other amino acids find it virtually impossible to exist. Glycine plays a distinct functional role, and uses its backbone which is free of a side-chain to bind to phosphates. This makes glycine an important amino acid while doing observational studies (Betts and Russell, 2003).

Histidine (H)

Histidine is a polar amino acid which has a pKa near to that of physiological pH. Hencem it is quite helpful in moving proteins on and off its side chains, thus affecting its charge. Due to this, it is quite ambiguous in nature, adn could be found buried in a protein core or exposed to a solvent. It is the most common amino acid found in protein active or binding sites and often acts in tandem with cysteine (Betts and Russell, 2003).

Isoleucine (I)

Isoleucine is an aliphatic and hydrophobic amino acid which is often found buried in the hydrophobic cores of the protein. Its side chain is very non-reactive, and is directly

involved in protein function as well as substrate recognition. Isoleucine is involved in binding and recognition of hydrophobic ligands such as lipids (Betts and Russell, 2003).

Leucine (L)

Like isoleucine, leucine is also an aliphatic and a hdyophic amino acid which is buried in the protein's hydrophobic cores. It prefers being with alpha helices as compared to beta strands, Its side chain is non-reactive, and is rarely involved in protein function, unlike isoleucine. It plays a role is suctaret recognition, and is also involved in binding and recognition of hydrophobic ligands (Betts and Russell, 2003).

Lysine (K)

Lysine is a positively charged, polar amino acid that is considered to be amphipathic. A part of its side chain nearest to the backbone is long and hydrophobic in nature and contains a carbon atom. Contrary, the end of the slide chain i s positively charged. Hence, lysine prefers to be on the external surfaces of proteins where they pair with a negatively charged major amino acids to create stable hydrogen bonds and is involved in protein stability (Betts and Russell, 2003)

Methionine (M)

It is a hydrophobic amino acid which can be categorised with other aliphatic amino acids. By being buried in the hydrophobic cores of the protein, ir prefers to be substituted by other hydrophobic amino acids. The side chain of methionine is

non-reactive and is rarely involved primarily in protein function. It also plays a role in binding and recognition of hydrophobic ligands (Betts and Russell, 2003).

Phenylalanine (F)

Like the couple of amino acids stated above, phenylalanine is also an aromatic and hydrophobic amino acid whose side chain is aromativ. Hence, it is involved in stacking interactions with other aromatic chains. Its side chain is non-reactive and hence phenylalanine is rarely involved directly in protein function. Hydrophobic amino acids may be involved in interactions with non-protein ligands containing aromatic groups via stacking interactions with the side chain (Betts and Russell, 2003)

Proline (P)

Proline is the only amino acid whose side chain is connected to the protein backbone twice, and it forms a five-membered ring which contains a nitrogen atom. Hence, proline cannot take up a lot of the main chain conformations which can be easily done by other amino acids. As a result, proline is found in very tight turns in protein structures. The side of of proline, due to its complicated nature, is very non-reactive and due to this, proline is really involved in protein active or binding sites (Betts and Russell, 2003)

Serine (S)

Serine can be present on the interior of the protein as well as on the surface, as it is very small in size as compared to other amino acids. Due to its small size, it is often found within the tight turns of the proteins, where it effectively mimics proline. Serine is

also found in functional centres of the protein and its reactive hydroxyl group is able to form hydrogen bonds with numerous polar substrates (Betts and Russell, 2003).

Threonine (T)

Although slightly polar, it generally substitutes with other polar amino acids.Threonine differs from serine in the way it has a methyl group in place of the hydrogen atom, found in serine. Benign found in both interior and exterior of the protein, it contains only one non-hydrogen substituent. Its backbone is a lot more bulky as compared to other amino acids and hence it is more restricted in the manner in which the main chain can adopt conformations. Theronien is often found in functional centres of the protein, , and plays the role of phosphorylation within intracellular proteins along with signal transduction with protein kinases (Betts and Russell, 2003).

Tryptophan (W)

Tryptophan is an aromatic and hydrophobic amino acid which subtitles with other amino acids of the same type. It is buried in the hydrophobic cores of the protein, and contains an aromatic side chain which helps with tryptophan benign involved in stacking interactions with other aromatic side-chains. It contains nitrogen atoms in the aromatic ring and hence is more reactive than ohenlyalaine which also has one. Tryptophan is involved in interactions with non-protein ligands and contains aromatic groups (Betts and Russell, 2003).

Tyrosine (Y)

It is an aromatic and a hydrophobic (partial) amino acid, and is buried in the hydrophobic cores of the protein as above. On its side chain, cytosine contains a reactive hydroxyl group which helps with it being involved in interaction with other non-protein atoms. Phosphorylation is carried out by tyrosine in intracellular proteins (Betts and Russell, 2003)

Valine (V)

Valine is an aliphatic and hydrophobic amino acid which substitutes with other amino acids of the same type. Like other similar amino acids, it prefers to be buried in the hydrophobic cores of the protein due to its nature. Its side chain is non-reactive and hence is rarely involved directly in protein function. Valien plays a role in substrate recognition and like other hydrophobic amino acid, it can be involved in binding and recognition of other hydrophobic ligands such as lipids (Betts and Russell, 2003)

Discussion of underrepresented motifs:

Cases: Homo sapiens

The dipeptide motifs AN, GV, LE and VC were seen to be underrepresented in the milk proteins found in humans. Here, valine is found in two cases, which is an aliphatic and a hydrophobic amino acid. Other amino acids include alanine (hydrophobic), asparagine (neutral-polar), glycine (hydrophobic), leucine (hydrophobic), glutamate (acidic, charged) and cysteine (neutral-polar). A majority of these are hydrophobic, and are

found buried in the deep hydrophobic core of proteins or within the lipid portion of the membrane (Betts and Russell, 2003). Their log_pUnd_Occ_UPC values were between -1.5 and -1.9, which show the statistical power of their said nature. The digestive enzymes [FLYWM][DE] (chymotrypsin & Pepsin (aromatic) + Negatively charged) and [ILVMAG][STNQCH] (Non-polar, non-aromatic + Polar) were found with log values -2.61 and -2.32 respectively, showing their lower statistical power as compared to the dipeptide motifs. Chymotrypsin facilitates the cleavage of peptide bonds via hydrolysis and pepsin is an endopeptidase which breaks down huge chunks of proteins into smaller peptides thus helps in digestion with the aid of non-polar enzymes(Coring, 1980; Janiak, 2016; Whitcomb and Lowe, 2007). Almost all of the amino acids involved in these enzymes are hydrophobic, indicating a certain trend in their nature.

Controls: Homo sapiens

No peptide motifs from the DD dataset were found to be underrepresented in the controls, which are a set of secretory proteins found in and around the stomach. However, [FLYWM][KR] (*Chymotrypsin & Pepsin (aromatic)* +*Trypsin*), [KR][FLYWM] (*Trypsin* + *Chymotrypsin & Pepsin (aromatic)*), [KR][ILVMAG] (*Trypsin* + *Non-polar, non-aromatic*) were found to be underrepresented in this dataset. The first two enzymes are inverses of each other, as explained in the Materials and Methods section, and most of the constituent amino acids involved are either hydrophobic and basic (Betts and Russell, 2003). Trypsin breaks down proteins that haven't been digested in the stomach, in the small intestine and chymotrypsin cleaves peptide bonds via hydrolysis (Janiak, 2016)

Cases: Bos taurus

AA, NA, RV, SV were found to be depleted in the milk proteins in the cow. These primarily include alanine and valine, both of which show the common property of hydrophobicity, the latter being hydrophobic as well. Their logp values, as in the human dataset case, tended to be between 1.3 and 1.8. The digestive enzymes [DE][FLYWM] (*Negative + chymotrypsin and pepsin*), [FLYWM][DE] (C*hymotrypsin and Pepsin + Negative*), [ILVMAG][STNQCH] (*Non-polar, non-aromatic + polar*), [R][ILVMAG] (*Non-polar + avoided in Pepsin*) were found in the same. The properties of chymotrypsin and trypsin are the same in cows as explained above, and these two enzymes were found to be common in both the organisms. In fact, teh motif [ILVMAG][STNQCH] (*Non-polar, non-aromatic + polar*) was a common outlier in both humans and cows in the PCA graph, suggesting its statistical significance in affecting the digestion of milk in the body.

Controls: Bos taurus

While no peptides from the DD were found underrepresented in the controls again after the same results as in the humans, the digestive enzyme [FLYWM][R] (*Chymotrypsin and Pepsin + avoided in Pepsin P1*) was found to be depleted. Note the absence of the enzyme in pepsin in the P1 state, hunting its probable behavior different from that of Pepsin found in the normal state.

***Homo sapiens*:** From Figure 5, we can observe the distribution of the avoided motifs in *Homo sapiens* via the PCA graph. All of the motifs in the control dataset along with LE (case) are clustered together, suggesting a similar pattern in their behaviour towards milk digestion, i.e. breaking down of milk proteins. While CV and VC were almost close the the X-axis, the motif [FLYWM][DE] (C*hymotrypsin and Pepsin + Negative*) was found in the fourth quadrant alone, and tending towards the positive values of PC1/Dim1, hunting its strong statistical significance in the process. [ILVMAG][STNQCH] (*Non-polar, non-aromatic + polar*) was seen as an extremely impactful outlier in this case, and the same was observed in *Bos taurus* as well.

***Bos taurus*:** The results seen here were statistically different than those in humans in some cases of the PCA plot. AA, RV and [FLYWM][R] (*Chymotrypsin and Pepsin + avoided in Pepsin P1*) were obe=sevred in the third quadrant, where negative values of both Dim1 and Dim2 indicated their lesser significance, added to that their clustered results. Again, in this case [FLYWM][DE] (C*hymotrypsin and Pepsin + Negative*) and [ILVMAG][STNQCH]] (*Non-polar, non-aromatic + polar*) were observed almost at the same position as seen in *Homo sapiens*. These two motifs are of significant importance in this experiment, as according to the milk digestion conundrum, chymotrypsin and pepsin are the key digestive enzymes.

Brines and Brock (1983) conducted experiments with these two enzymes in study with relation to milk, particularly lactoferrin in humans and bovine. They showed that neither of the two enzymes have any effect of antimicrobial activity with regards to lactoferrin,

but they had an impact on the iron-binding capacity of purified lactoferrin in the lab. It was seen that lactoferrin in cows was more resistant to digestion as compared to humans, and this may be due to the evolutionary development for the survival of the lactoferrin in the gut of the infant.

Khaldi et al. (2014), while experimenting on the enzymes that are involved in breast milk digestion showed that the major cleavage of proteins in human milk is at the cleavage sites of trypsin and plasmin, one of these two being our outlier motif. The below figure extracted from their work shows the similar bioactivity of trypsin and chymotrypsin shows their close range of work function, with pepsin coming close by.

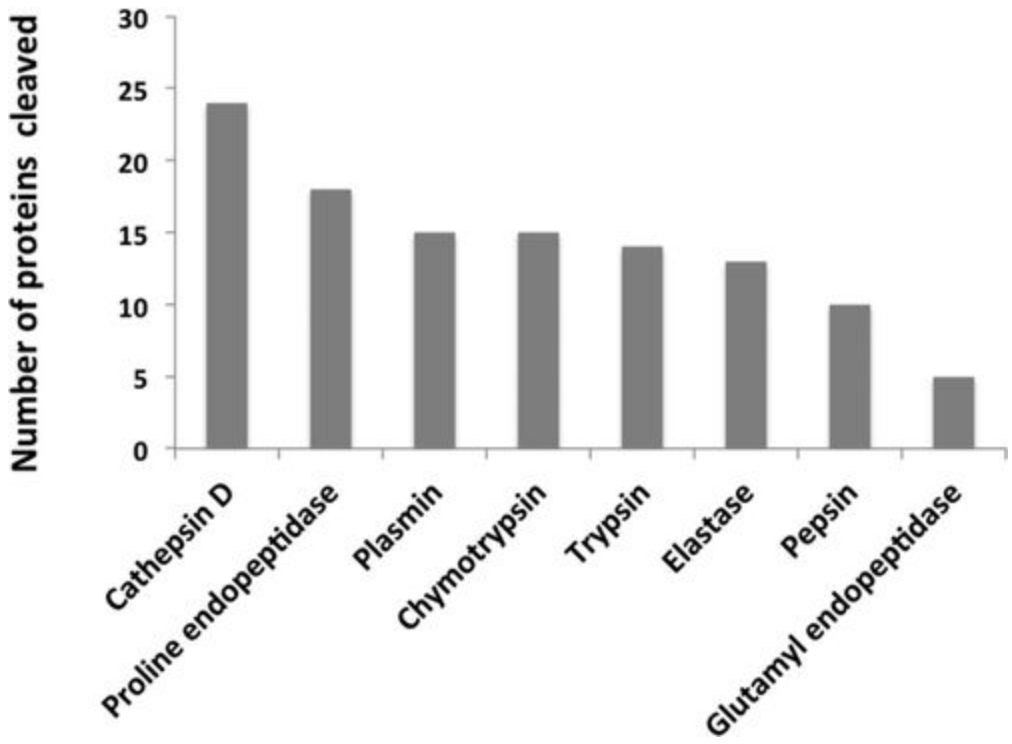

Figure 7. *Representation of the total number of cleaved milk proteins per predicted enzyme (Khaldi, Nora, et al. "Predicting the Important Enzymes in Human Breast Milk Digestion." Journal of Agricultural and Food Chemistry, vol. 62, no. 29, 10 July 2014, pp.*

Moreover, the hypothesis, which stated that there would be a greater number of underrepresented SLiMs present in the milk proteins as compared to the secreted proteins, has also been proved true to some extent. This may be because the presence of underrepresented cleavage sites in milk proteins is essential for optimal growth and nutrition of the neonate in both humans and cows. The reasons behind the advantages and the necessity for the mother's milk in both the organisms has been discussed in detail in the Introduction, and our results tend to show similar outcomes. This is necessary for the development of the infant's immune system and formation of the required quantity of antibodies.

One important factor noticed was that the amino acid Proline (P) was nowhere to be seen in the results. Proline is an amino acid which is very hard to degrade in the body (Wu et al., 2011). Hence, there should be a lot of proline-containing motifs found to be avoided as the more the better, given the complex process of proline breakdown. Maybe this was because of the small sample size of the datasets of both the proteins and the motifs, as well as the statistics considered which filtered the majority of the outputs. This had to be done for focus-based calculations and interpretation.

CONCLUSION

The motifs avoided in the process of milk digestion was studied in humans and cows, as such studies give an insight to the mechanisms involved in what goes behind giving optimum nutrition to the neonate. SLiMProb, with its ability to calculate the underrepresented/overrepresented motifs statistically with various calculations of the expected and actual occurrences of motifs is a great tool to do the same. Here, motifs avoided in both the major milk proteins and the secreted proteins were identified and the statically significant ones were considered and analysed. These findings help fiddn the occurrence of certain dipeptides and enzymes found which could be checked for optimal digestion of milk in the neonate's body.

This project was an introductory experiment of what could be done at the large-scale level with the 'omic' data which is growing exponentially by the day, and computational tools like SLiMProb are benign developed to find out the hidden messages in them which are getting sophisticated and refined with time as well. In the long run, this could

help in understanding food digestion better and commercially, develop better patient-specific foods involving peptides which are suited for an individual's body.

REFERENCES

- Abdi, H., Williams, L.J., 2010. Principal component analysis. WIREs Comput. Stat. 2, 433–459. https://doi.org/10.1002/wics.101
- Adkins, Y., Lönnerdal, B., 2001. Binding of Transcobalamin II by Human Mammary Epithelial Cells, in: Newburg, D.S. (Ed.), Bioactive Components of Human Milk, Advances in Experimental Medicine and Biology. Springer US, Boston, MA, pp. 469–477. https://doi.org/10.1007/978-1-4615-1371-1_58
- Afroze, S., Meng, F., Jensen, K., McDaniel, K., Rahal, K., Onori, P., Gaudio, E., Alpini, G., Glaser, S.S., 2013. The physiological roles of secretin and its receptor. Ann. Transl. Med. 1, 29. https://doi.org/10.3978/j.issn.2305-5839.2012.12.01
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
- Betts, M.J., Russell, R.B., 2003. Amino Acid Properties and Consequences of Substitutions, in: Bioinformatics for Geneticists. John Wiley & Sons, Ltd, pp. 289–316. https://doi.org/10.1002/0470867302.ch14
- Brines, R.D., Brock, J.H., 1983. The effect of trypsin and chymotrypsin on the in vitro antimicrobial and iron-binding properties of lactoferrin in human milk and bovine colostrum. Unusual resistance of human apolactoferrin to proteolytic digestion. Biochim. Biophys. Acta 759, 229–235.

https://doi.org/10.1016/0304-4165(83)90317-3

- Bundgaard, J.R., Vuust, J., Rehfeld, J.F., 1995. Tyrosine O-sulfation promotes proteolytic processing of progastrin. EMBO J. 14, 3073–3079.
- Clare, D.A., Swaisgood, H.E., 2000. Bioactive Milk Peptides: A Prospectus1. J. Dairy Sci. 83, 1187–1195. https://doi.org/10.3168/jds.S0022-0302(00)74983-6
- Corring, T., 1980. The adaptation of digestive enzymes to the diet : Its physiological significance. Reprod. Nutr. Dév. 20, 1217–1235. https://doi.org/10.1051/rnd:19800713
- Di Leo, G., Sardanelli, F., 2020. Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. Eur. Radiol. Exp. 4, 18. https://doi.org/10.1186/s41747-020-0145-y
- Drucker, D.J., 2003. Glucagon-like peptides: regulators of cell proliferation, differentiation, and apoptosis. Mol. Endocrinol. Baltim. Md 17, 161–171. https://doi.org/10.1210/me.2002-0306
- E. Davey, N., Roey, K.V., J. Weatheritt, R., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., J. Gibson, T., 2012. Attributes of short linear motifs. Mol. Biosyst. 8, 268–281. https://doi.org/10.1039/C1MB05231D
- Edwards, R.J., Davey, N.E., Shields, D.C., 2007. SLiMFinder: A Probabilistic Method for Identifying Over-Represented, Convergently Evolved, Short Linear Motifs in Proteins. PLOS ONE 2, e967. https://doi.org/10.1371/journal.pone.0000967
- Edwards, R.J., Palopoli, N., 2015. Computational Prediction of Short Linear Motifs from Protein Sequences, in: Zhou, P., Huang, J. (Eds.), Computational

Peptidology, Methods in Molecular Biology. Springer, New York, NY, pp. 89–141. https://doi.org/10.1007/978-1-4939-2285-7_6

· Edwards, R.J., Paulsen, K., Aguilar Gomez, C.M., Pérez-Bercoff, Å., 2020. Computational Prediction of Disordered Protein Motifs Using SLiMSuite, in: Kragelund, B.B., Skriver, K. (Eds.), Intrinsically Disordered Proteins: Methods and Protocols, Methods in Molecular Biology. Springer US, New York, NY, pp. 37–72. https://doi.org/10.1007/978-1-0716-0524-0_3

· Goldman, A.S., 2000. Modulation of the Gastrointestinal Tract of Infants by Human Milk. Interfaces and Interactions. An Evolutionary Perspective. J. Nutr. 130, 426S-431S. https://doi.org/10.1093/jn/130.2.426S

· Goldman, A.S., Chheda, S., Garofalo, R., 1997. Spectrum of immunomodulating agents in human milk. Int. J. Pediatr. Hematol. 4, 491–497.

· Guo, M., Hendricks, G.M., 2008. Chemistry and Biological Properties of Human Milk. Curr. Nutr. Food Sci. 4, 305–320. https://doi.org/10.2174/157340108786263667

· Hamosh, M., n.d. Enzymes in Human Milk 40.

· Hendricks, G.M., Guo, M., 2014. Bioactive components in human milk, in: Human Milk Biochemistry and Infant Formula Manufacturing Technology. Elsevier, pp. 33–54. https://doi.org/10.1533/9780857099150.1.33

· Holton, T.A., Vijayakumar, V., Dallas, D.C., Guerrero, A., Borghese, R.A., Lebrilla, C.B., German, J.B., Barile, D., Underwood, M.A., Shields, D.C., Khaldi, N., 2014. Following the Digestion of Milk Proteins from Mother to Baby. J. Proteome Res. 13, 5777–5783. https://doi.org/10.1021/pr5006907

· Janiak, M.C., 2016. Digestive enzymes of human and nonhuman primates. Evol. Anthropol. Issues News Rev. 25, 253–266. https://doi.org/10.1002/evan.21498
· Khaldi, N., Vijayakumar, V., Dallas, D.C., Guerrero, A., Wickramasinghe, S., Smilowitz, J.T., Medrano, J.F., Lebrilla, C.B., Shields, D.C., German, J.B., 2014. Predicting the Important Enzymes in Human Breast Milk Digestion. J. Agric. Food Chem. 62, 7225–7232. https://doi.org/10.1021/jf405601e
· Kieffer, T.J., Habener, J.F., 1999. The glucagon-like peptides. Endocr. Rev. 20, 876–913. https://doi.org/10.1210/edrv.20.6.0385
· Kim, D.-K., Yun, S., Son, G.H., Hwang, J.-I., Park, C.R., Kim, J.I., Kim, K., Vaudry, H., Seong, J.Y., 2014. Coevolution of the spexin/galanin/kisspeptin family: Spexin activates galanin receptor type II and III. Endocrinology 155, 1864–1873. https://doi.org/10.1210/en.2013-2106
· Koldovsky: Growth factors and cytokines in milk - Google Scholar [WWW Document], n.d. URL https://scholar.google.com/scholar_lookup?title=Growth%20Factors%20and%20Cytokines%20in%20Milk&publication_year=1999&author=O.%20Koldovsky&author=A.%20Goldman (accessed 1.10.21).
· Le, T.T., Deeth, H.C., Larsen, L.B., 2017. Proteomics of major bovine milk proteins: Novel insights. Int. Dairy J., 25th Anniversary of the International Dairy Journal 67, 2–15. https://doi.org/10.1016/j.idairyj.2016.11.016
· Lecker, S.H., Goldberg, A.L., Mitch, W.E., 2006. Protein degradation by the ubiquitin-proteasome pathway in normal and disease states. J. Am. Soc. Nephrol. JASN 17, 1807–1819. https://doi.org/10.1681/ASN.2006010083

· Lee, W.-Y., Wang, C.-J., Lin, T.-Y., Hsiao, C.-L., Luo, C.-W., 2013. CXCL17, an orphan chemokine, acts as a novel angiogenic and anti-inflammatory factor. Am. J. Physiol. Endocrinol. Metab. 304, E32-40. https://doi.org/10.1152/ajpendo.00083.2012

· Lönnerdal, B., 2003. Nutritional and physiologic significance of human milk proteins. Am. J. Clin. Nutr. 77, 1537S-1543S. https://doi.org/10.1093/ajcn/77.6.1537S

· Ma, J.-N., Currier, E.A., Essex, A., Feddock, M., Spalding, T.A., Nash, N.R., Brann, M.R., Burstein, E.S., 2004. Discovery of novel peptide/receptor interactions: identification of PHM-27 as a potent agonist of the human calcitonin receptor. Biochem. Pharmacol. 67, 1279–1284. https://doi.org/10.1016/j.bcp.2003.11.008

· Mirabeau, O., Perlas, E., Severini, C., Audero, E., Gascuel, O., Possenti, R., Birney, E., Rosenthal, N., Gross, C., 2007. Identification of novel peptide hormones in the human proteome by hidden Markov model screening. Genome Res. 17, 320–327. https://doi.org/10.1101/gr.5755407

· Neduva, V., Russell, R.B., 2006. Peptides mediating interaction networks: new leads at last. Curr. Opin. Biotechnol. 17, 465–471. https://doi.org/10.1016/j.copbio.2006.08.002

· Newburg, D.S., 2001. Bioactive Components of Human Milk, in: Newburg, D.S. (Ed.), Bioactive Components of Human Milk, Advances in Experimental Medicine and Biology. Springer US, Boston, MA, pp. 3–10. https://doi.org/10.1007/978-1-4615-1371-1_1

· Patton, S., Huston, G.E., 1986. A method for isolation of milk fat globules. Lipids 21, 170–174. https://doi.org/10.1007/BF02534441
· Pickering, L.K., Morrow, A.L., Ruiz-Palacios, G.M., Schanler, R.J., 2013. Protecting Infants through Human Milk: Advancing the Scientific Evidence. Springer Science & Business Media.
· Pisabarro, M.T., Leung, B., Kwong, M., Corpuz, R., Frantz, G.D., Chiang, N., Vandlen, R., Diehl, L.J., Skelton, N., Kim, H.S., Eaton, D., Schmidt, K.N., 2006. Cutting edge: novel human dendritic cell- and monocyte-attracting chemokine-like protein identified by fold recognition methods. J. Immunol. Baltim. Md 1950 176, 2069–2073. https://doi.org/10.4049/jimmunol.176.4.2069
· Smithers, G.W., Copeland, A.D., 1998. 1997 International Whey Conference.
· The UniProt Consortium, 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43, D204–D212. https://doi.org/10.1093/nar/gku989
· Tremblay, L., Laporte, M.F., Léonil, J., Dupont, D., Paquin, P., 2003. Quantitation of Proteins in Milk and Milk Products, in: Fox, P.F., McSweeney, P.L.H. (Eds.), Advanced Dairy Chemistry—1 Proteins: Part A / Part B. Springer US, Boston, MA, pp. 49–138. https://doi.org/10.1007/978-1-4419-8602-3_2
· Van Roey, K., Uyar, B., Weatheritt, R.J., Dinkel, H., Seiler, M., Budd, A., Gibson, T.J., Davey, N.E., 2014. Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation. Chem. Rev. 114, 6733–6778. https://doi.org/10.1021/cr400585q
· Wang, X., Robbins, J., 2014. Proteasomal and lysosomal protein degradation and heart disease. J. Mol. Cell. Cardiol., Protein Quality Control, the Ubiquitin

Proteasome System, and Autophagy 71, 16–24. https://doi.org/10.1016/j.yjmcc.2013.11.006

- Weisstein, E.W., n.d. Bonferroni Correction [WWW Document]. URL https://mathworld.wolfram.com/BonferroniCorrection.html (accessed 1.14.21).
- Whitcomb, D.C., Lowe, M.E., 2007. Human Pancreatic Digestive Enzymes. Dig. Dis. Sci. 52, 1–17. https://doi.org/10.1007/s10620-006-9589-z
- Wu, G., Bazer, F.W., Burghardt, R.C., Johnson, G.A., Kim, S.W., Knabe, D.A., Li, P., Li, X., McKnight, J.R., Satterfield, M.C., 2011. Proline and hydroxyproline metabolism: implications for animal and human nutrition. Amino Acids 40, 1053–1063.
- Ziegler, E.E., Fomon, S.J., Nelson, S.E., Rebouche, C.J., Edwards, B.B., Rogers, R.R., Lehman, L.J., 1990. Cow milk feeding in infancy: Further observations on blood loss from the gastrointestinal tract. J. Pediatr. 116, 11–18. https://doi.org/10.1016/S0022-3476(05)90003-6

APPENDICES

Table 10. *Significantly avoided motifs in Homo sapiens - Major Milk Proteins (Cases) - DD motif dataset*

	Dataset	RunID	Masking	Run Time	Seq Num	UPNum	AANum	Motif	Pattern	IC	N_Occ	E_Occ	p_Occ	pUnd_Occ	N_Seq	E_Seq	p_Seq	pUnd_Seq	N_UPC	E_UPC	p_UPC	pUnd_UPC	Split N	N_Occ_UPC	E_Occ_UPC	Ratio	pUnd_Occ_UPC	log_pUnd_Occ_UPC
1	case_human	2011 12-20:53	FreqNoMask	0:00:02	6	5	2054	AN	AN	2	1	7.01	1	0.007	1	3.45	0.99	0.14	1	3.24	0.99	0.17	1	1	6.583304	0.151899	0.01049	-1.97923
2	case_human	2011 12-20:53	FreqNoMask	0:00:02	6	5	2054	GV	GV	2	1	5.44	1	0.028	1	2.13	0.93	0.37	1	2.09	0.93	0.38	1	1	5.33784	0.187342	0.030461	-1.51625
3	case_human	2011 12-20:53	FreqNoMask	0:00:02	6	5	2054	LE	LE	2	8	15.7	0.99	0.026	5	4.66	0.6	0.68	4	3.91	0.7	0.65	1	6.4	13.17318	0.485836	0.023382	-1.63112
4	case_human	2011 12-20:53	FreqNoMask	0:00:02	6	5	2054	VC	VC	2	1	5.48	1	0.027	1	2.45	0.96	0.3	1	2.41	0.96	0.31	1	1	5.390531	0.18551	0.029138	-1.53554

Table 11. *Significantly avoided motifs in Homo sapiens - Major Milk Proteins (Cases) - DE motif dataset*

	Dataset	RunID	Masking	Run Time	Seq Num	UPNum	AANum	Motif	Pattern	IC	N_Occ	E_Occ	p_Occ	pUnd_Occ	N_Seq	E_Seq	p_Seq	pUnd_Seq	N_UPC	E_UPC	p_UPC	pUnd_UPC	N_Occ_UPC	E_Occ_UPC	Ratio	pUnd_Occ_UPC	log_pUnd_Occ_UPC
1	case_human	2011 17-00:35	FreqNoMask	0:00:02	6	5	2054	[FLYWM][DE]	[FLMWY][DE]	1.231378	32	51.3	1	0.003	5	5.79	0.98	0.48	4	4.81	0.99	0.47	25.6	42.6171	0.600698	0.002492	-2.6034
2	case_human	2011 17-00:35	FreqNoMask	0:00:02	6	5	2054	[ILVMAG][STNQCH]	[AGILMV][CHNQST]	0.803792	146	183.1	1	0.003	6	6	1	0.61	5	5	1	0.62	121.6667	152.5833	0.797378	0.004716	-2.32643

Table 12. Significantly avoided motifs in *Homo sapiens* - Secreted Proteins (Controls) - DE motif dataset

	Dataset	RunID	Masking	RunTime	SeqNum	UPNum	AANum	Motif	Pattern	IC	N_Occ	E_Occ	p_Occ	pUnd_Occ	N_Seq	E_Seq	p_Seq	pUnd_Seq	N_UPC	E_UPC	p_UPC	pUnd_UPC	N_Occ_UPC	E_Occ_UPC	Ratio	pUnd_Occ_UPC	log_pUnd_Occ_UPC
1	control_human	2011 17-0 0:38	FreqNoMask	0:00:01	6	4	807	[FLYWM][KR]	[FLMWY][KR]	1.23 1378	16	26.3	0.99	0.021	6	5.84	0.85	0.63	4	3.87	0.88	0.65	10.6 6667	17.4 2825	0.61 2033	0.04 0103	-1.3 9682
2	control_human	2011 17-0 0:38	FreqNoMask	0:00:01	6	4	807	[KR][FLYWM]	[KR][FLMWY]	1.23 1378	16	26.3	0.99	0.021	6	5.84	0.85	0.63	4	3.87	0.88	0.65	10.6 6667	17.4 2825	0.61 2033	0.04 0103	-1.3 9682
3	control_human	2011 17-0 0:38	FreqNoMask	0:00:01	6	4	807	[KR][ILVMAG]	[KR][AGILMV]	1.17 0518	26	39.6	0.99	0.014	6	5.97	0.97	0.61	4	3.97	0.97	0.63	17.3 3333	26.3 3367	0.65 8219	0.03 5975	-1.4 44

Table 13. *Significantly avoided motifs in Bos taurus - Major Milk Proteins (Cases) - DD motif dataset*

	Dataset	RunID	Masking	RunTime	SeqNum	UPNum	AANum	Motif	Pattern	IC	N_Occ	E_Occ	p_Occ	pUnd_Occ	N_Seq	E_Seq	p_Seq	pUnd_Seq	N_UPC	E_UPC	p_UPC	pUnd_UPC	Split N	N_Occ_UPC	E_Occ_UPC	Ratio	pUnd_Occ_UPC	log_pUnd_Occ_UPC
1	case_bovine	2011 12-2 0:47	FreqNoMask	0:00:02	6	5	2085	AA	AA	2	5	13.2	1	0.009	3	3.67	0.84	0.5	3	3.48	0.83	0.54	1	5	12.5 1662	0.39 9469	0.01 4666	-1.8 3368
2	case_bovine	2011 12-2 0:47	FreqNoMask	0:00:02	6	5	2085	NA	NA	2	1	5.54	1	0.026	1	2.88	0.98	0.22	1	2.78	0.98	0.23	1	1	5.34 7639	0.18 6998	0.03 0211	-1.5 1984
3	case_bovine	2011 12-2 0:47	FreqNoMask	0:00:02	6	5	2085	RV	RV	2	1	5.25	0.99	0.033	1	2.7	0.97	0.25	1	2.57	0.97	0.27	1	1	4.99 7222	0.20 0111	0.04 0521	-1.3 9232
4	case_bovine	2011 12-2 0:47	FreqNoMask	0:00:02	6	5	2085	SV	SV	2	5	8.98	0.94	0.12	4	4.25	0.76	0.58	3	3.73	0.89	0.49	1	3.75	7.88 1271	0.47 5812	0.04 5908	-1.3 3811

Table 14. *Significantly avoided motifs in Bos taurus - Major Milk Proteins (Cases) - DE motif dataset*

	Dataset	RunID	Masking	Run Time	Seq Num	UPNum	AANum	Motif	Pattern	IC	N_Occ	E_Occ	p_Occ	pUnd_Occ	N_Seq	E_Seq	p_Seq	pUnd_Seq	N_UPC	E_UPC	p_UPC	pUnd_UPC	N_Occ_UPC	E_Occ_UPC	Ratio	pUnd_Occ_UPC	log_pUnd_Occ_UPC
1	case_bovine	2011 17-0 0:33	FreqNoMask	0:00:01	6	5	2085	[DE][FLYWM]	[DE][FLMWY]	1.23 1378	41	54.8	0.98	0.032	5	5.93	1	0.46	4	4.94	1	0.45	32.8	45.6 5126	0.71 849	0.02 1322	-1.6 7118
2	case_bovine	2011 17-0 0:33	FreqNoMask	0:00:01	6	5	2085	[FLYWM][DE]	[FLMWY][DE]	1.23 1378	37	54.8	1	0.007	6	5.93	0.93	0.62	5	4.94	0.94	0.63	30.8 3333	45.6 5126	0.67 541	0.00 9106	-2.0 4067
3	case_bovine	2011 17-0 0:33	FreqNoMask	0:00:01	6	5	2085	[ILVMAG][STNQCH]	[AGILMV][CHNQST]	0.80 3792	159	186. 7	0.98	0.021	6	6	1	0.61	5	5	1	0.62	132. 5	155. 5833	0.85 1634	0.02 9665	-1.5 2775
4	case_bovine	2011 17-0 0:33	FreqNoMask	0:00:01	6	5	2085	[R][ILVMAG]	R[AGILMV]	1.40 1896	21	27.2	0.9	0.14	4	4.75	0.89	0.49	3	4.06	0.95	0.42	15.7 5	23.2 4884	0.67 7453	0.04 7012	-1.3 2779

Table 15. *Significantly avoided motifs in Bos taurus - Secreted Proteins (Controls) - DE motif dataset*

	Dataset	RunID	Masking	Run Time	Seq Num	UPNum	AANum	Motif	Pattern	IC	N_Occ	E_Occ	p_Occ	pUnd_Occ	N_Seq	E_Seq	p_Seq	pUnd_Seq	N_UPC	E_UPC	p_UPC	pUnd_UPC	N_Occ_UPC	E_Occ_UPC	Ratio	pUnd_Occ_UPC	log_pUnd_Occ_UPC
1	control_bovine	2011 17-0 0:36	FreqNoMask	0:00:01	6	4	715	[FLYWM][R]	[FLMWY]R	1.46 2756	7	14.9	0.99	0.019	5	5.01	0.74	0.61	3	3.72	0.97	0.49	4.2	11.0 6347	0.37 9628	0.01 4471	-1.8 3951

```
Call:
PCA(X = human.pca, quali.sup = c(1, 2), graph = FALSE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8
Variance               5.741   1.553   0.533   0.153   0.019   0.000   0.000   0.000
% of var.             71.766  19.414   6.659   1.918   0.240   0.002   0.000   0.000
Cumulative % of var.  71.766  91.180  97.840  99.758  99.998 100.000 100.000 100.000

Individuals
                  Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
1                | 2.405 | -0.550  0.586  0.052 | -2.238 35.834  0.866 |  0.065  0.087  0.001 |
2                | 2.253 | -2.017  7.870  0.801 | -0.084  0.051  0.001 |  0.990 20.448  0.193 |
3                | 1.348 | -1.085  2.276  0.647 |  0.131  0.123  0.009 |  0.156  0.507  0.013 |
4                | 2.174 | -1.942  7.295  0.798 | -0.199  0.282  0.008 |  0.952 18.882  0.192 |
5                | 3.035 |  2.000  7.743  0.434 | -1.999 28.589  0.434 | -0.979 19.990  0.104 |
6                | 6.116 |  6.003 69.733  0.963 |  0.880  5.539  0.021 |  0.767 12.257  0.016 |
7                | 1.721 | -1.057  2.163  0.377 |  1.241 11.024  0.520 | -0.544  6.178  0.100 |
8                | 1.721 | -1.057  2.163  0.377 |  1.241 11.024  0.520 | -0.544  6.178  0.100 |
9                | 1.434 | -0.296  0.169  0.042 |  1.026  7.533  0.512 | -0.861 15.473  0.361 |

Variables
                   Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
N_Occ            | 0.950 15.725  0.903 | 0.267  4.589  0.071 | 0.155  4.515  0.024 |
E_Occ            | 0.960 16.046  0.921 | 0.252  4.101  0.064 | 0.118  2.616  0.014 |
pUnd_Occ         |-0.766 10.227  0.587 | 0.431 11.961  0.186 | 0.389 28.469  0.152 |
N_Occ_UPC        | 0.951 15.758  0.905 | 0.243  3.788  0.059 | 0.188  6.626  0.035 |
E_Occ_UPC        | 0.964 16.184  0.929 | 0.212  2.888  0.045 | 0.160  4.827  0.026 |
Ratio            | 0.645  7.245  0.416 | 0.538 18.633  0.289 |-0.512 49.259  0.262 |
pUnd_Occ_UPC     |-0.703  8.610  0.494 | 0.684 30.113  0.468 |-0.123  2.837  0.015 |
log_pUnd_Occ_UPC |-0.765 10.205  0.586 | 0.610 23.926  0.372 | 0.067  0.852  0.005 |

Supplementary categories (the 10 first)
                  Dist    Dim.1   cos2 v.test    Dim.2   cos2 v.test    Dim.3   cos2 v.test
case             | 0.785 |  0.402 0.262  0.671 | -0.585 0.555 -1.877 |  0.325 0.171  1.781 |
control          | 1.570 | -0.803 0.262 -0.671 |  1.170 0.555  1.877 | -0.650 0.171 -1.781 |
[FLYWM][DE]      | 3.035 |  2.000 0.434  0.835 | -1.999 0.434 -1.604 | -0.979 0.104 -1.341 |
[FLYWM][KR]      | 1.721 | -1.057 0.377 -0.441 |  1.241 0.520  0.996 | -0.544 0.100 -0.746 |
[ILVMAG][STNQCH] | 6.116 |  6.003 0.963  2.505 |  0.880 0.021  0.706 |  0.767 0.016  1.050 |
[KR][FLYWM]      | 1.721 | -1.057 0.377 -0.441 |  1.241 0.520  0.996 | -0.544 0.100 -0.746 |
[KR][ILVMAG]     | 1.434 | -0.296 0.042 -0.123 |  1.026 0.512  0.823 | -0.861 0.361 -1.180 |
AN               | 2.405 | -0.550 0.052 -0.230 | -2.238 0.866 -1.796 |  0.065 0.001  0.089 |
GV               | 2.253 | -2.017 0.801 -0.842 | -0.084 0.001 -0.068 |  0.990 0.193  1.357 |
LE               | 1.348 | -1.085 0.647 -0.453 |  0.131 0.009  0.105 |  0.156 0.013  0.214 |
```

Figure 8. *Principal Component Analysis - Homo sapiens*

```
Call:
PCA(X = bovine.pca, quali.sup = c(1, 2), graph = FALSE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
Variance               6.007   1.433   0.505   0.045   0.010   0.000
% of var.             75.087  17.912   6.309   0.561   0.128   0.002
Cumulative % of var.  75.087  93.000  99.308  99.869  99.998 100.000

Individuals
                 Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
1                | 1.164 | -0.489  0.569  0.176 | -1.008 10.129  0.749 | -0.037  0.040  0.001 |
2                | 1.892 | -1.699  6.865  0.807 | -0.288  0.828  0.023 |  0.758 16.274  0.161 |
3                | 2.241 | -2.028  9.778  0.819 |  0.254  0.645  0.013 |  0.876 21.721  0.153 |
4                | 3.181 | -2.089 10.376  0.431 |  2.195 48.033  0.476 | -0.967 26.485  0.092 |
5                | 2.400 |  1.630  6.318  0.461 | -1.418 20.057  0.349 | -0.998 28.176  0.173 |
6                | 5.385 |  5.241 65.330  0.947 |  1.133 12.798  0.044 |  0.493  6.867  0.008 |
7                | 1.080 | -0.567  0.764  0.275 | -0.868  7.511  0.646 | -0.124  0.437  0.013 |

Variables
                  Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
N_Occ           | 0.945 14.876  0.894 | 0.291  5.915  0.085 | 0.146  4.236  0.021 |
E_Occ           | 0.959 15.318  0.920 | 0.254  4.497  0.064 | 0.123  2.996  0.015 |
pUnd_Occ        |-0.546  4.955  0.298 | 0.746 38.812  0.556 |-0.379 28.400  0.143 |
N_Occ_UPC       | 0.943 14.802  0.889 | 0.294  6.042  0.087 | 0.155  4.788  0.024 |
E_Occ_UPC       | 0.958 15.281  0.918 | 0.256  4.580  0.066 | 0.127  3.206  0.016 |
Ratio           | 0.881 12.919  0.776 | 0.215  3.216  0.046 |-0.413 33.811  0.171 |
pUnd_Occ_UPC    |-0.791 10.424  0.626 | 0.576 23.185  0.332 | 0.145  4.175  0.021 |
log_pUnd_Occ_UPC|-0.828 11.425  0.686 | 0.444 13.753  0.197 | 0.305 18.389  0.093 |

Supplementary categories
                  Dist    Dim.1   cos2 v.test    Dim.2   cos2 v.test    Dim.3   cos2 v.test
case_bovine     | 1.610 | -1.576  0.958 -1.819 |  0.288  0.032  0.681 |  0.157  0.010  0.627 |
case_human      | 3.450 |  3.436  0.991  2.172 | -0.143  0.002 -0.185 | -0.253  0.005 -0.551 |
control_bovine  | 1.080 | -0.567  0.275 -0.231 | -0.868  0.646 -0.725 | -0.124  0.013 -0.175 |
[FLYWM][DE]     | 2.400 |  1.630  0.461  0.665 | -1.418  0.349 -1.185 | -0.998  0.173 -1.404 |
[FLYWM][R]      | 1.080 | -0.567  0.275 -0.231 | -0.868  0.646 -0.725 | -0.124  0.013 -0.175 |
[ILVMAG][STNQCH]| 5.385 |  5.241  0.947  2.138 |  1.133  0.044  0.946 |  0.493  0.008  0.693 |
AA              | 1.164 | -0.489  0.176 -0.200 | -1.008  0.749 -0.842 | -0.037  0.001 -0.053 |
RV              | 2.241 | -2.028  0.819 -0.827 |  0.254  0.013  0.213 |  0.876  0.153  1.233 |
SV              | 3.181 | -2.089  0.431 -0.852 |  2.195  0.476  1.834 | -0.967  0.092 -1.362 |
```

Figure 9. Principal Component Analysis - *Bos taurus*

Note: *Refer to the .zip file attached along with the project for detailed information on all the datasets and the PCA results*