Escola Tècnica Superior
d'Enginyeria Agronòmica i del Medi Natural

UNIVERSITAT POLITÉCNICA DE VALÈNCIA
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA AGRONÓMICA Y DEL MEDIO NATURAL
Grado en Biotecnología

De novo design of an ideal protein for feeding meat chickens from 0 to 21 days

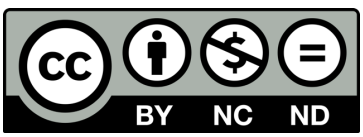
---

FINAL DEGREE PROJECT

Author: Clara María Lledó Morell
Academic year 2019/2020

Academic tutor: Prof. D. Juan José Pascual Amorós
Academic cotutor: Prof. Dña. María Cambra López

Valencia 6th July 2020


BY NC ND

De novo design of an ideal protein for feeding meat chickens from 0 to 21 days.

Abstract

The ideal protein concept is based on the idea that to achieve optimum performance and maximum growth at a given stage of life, birds need specific amounts and ratios of amino acids. However, these amino acids must be obtained from feed, and their utilization efficiency (proportion, digestion and metabolization) is generally low, leading to high losses of nitrogen in excreta. Ideally, it should be possible to synthesise a protein that will meet the requirements of all amino acids (without excess or defect) while being fully digested and metabolised. The aim of this work was to develop a novel process to design an ideal protein completely digestible and usable in broilers from 0 to 21 days. To do so, we conducted a research on the net requirements of amino acids and the functioning of the digestive system at enzymatic level of chickens of that age. From these data, we designed possible primary structures of the polypeptide. Then, tertiary structure and its physicochemical properties were predicted by means of computational methods. The obtained proteins were evaluated based on their digestibility, physicochemical characteristics and future projections of synthesis and production. This procedure can be applied for obtaining proteins in other vital stages of the animal based on the knowledge of its specific nutritional requirements. The sequential method proposed presents a first approach for the design of proteins in precision feeding and is susceptible to improvement and implementation with progress in the field of Protein Engineering and Design.

Keywords

Amino acids, proteins, enzymes, digestible, computational modelling, design.

Diseño *de novo* de una proteína ideal para la alimentación de pollos de carne de 0 a 21 días.

Resumen

El concepto de proteína ideal esta basado en la idea de que para alcanzar su rendimiento óptimo y máximo crecimiento en una determinada etapa vital las aves necesitan unas cantidades y ratios específicos de aminoácidos. Sin embargo, dichos aminoácidos deben obtenerlos de alimentos, cuya eficacia de utilización (proporción, digestión y metabolización) no es elevada, llevado a unas elevadas pérdidas de nitrógeno en las deyecciones. Lo ideal sería poder sintetizar una proteína que cubriera las necesidades de todos los aminoácidos (sin excesos, ni defectos) y que fuera totalmente digerida y metabolizada. El objetivo de este trabajo es desarrollar un proceso de diseño de una proteína ideal completamente digestible y aprovechable en pollos de engorde de 0 a 21 días. Para ello realizamos una investigación sobre los requerimientos netos de aminoácidos y el funcionamiento del sistema digestivo a nivel enzimático de los pollos de dicha edad. A partir de estos datos diseñamos posibles estructuras primarias del polipéptido. A continuación, predecimos su estructura terciaria y sus propiedades fisicoquímicas mediante métodos computacionales. Las proteínas obtenidas se valoran en base a su digestibilidad, características fisicoquímicas y proyecciones futuras de síntesis y producción. Este procedimiento es aplicable para la obtención de proteínas en otras etapas vitales del animal partiendo del conocimiento de las exigencias nutricionales específicas. El método secuencial planteado presenta una primera aproximación para el diseño de proteínas en alimentación de precisión y es susceptible de mejora e implementación con los avances en el campo de la lngeniería y Diseño de Proteínas.

Palabras clave

Aminoácidos, proteínas, enzimas, digestible, modelado computacional, diseño.

Disseny *de novo* d'una proteïna ideal per a l'alimentació de pollastres de carn de 0 a 21 dies.

Resum

El concepte de proteïna ideal esta basat en la idea que per a aconseguir el seu rendiment òptim i màxim creixement en una determinada etapa vital les aus necessiten unes quantitats i ràtios específics d'aminoàcids. Dits aminoàcids han d'obtenir-se d'aliments, l'eficàcia d'utilització dels quals (proporció, digestió i metabolització) no és elevada, portat a unes pèrdues considerables de nitrogen en les dejeccions. L'ideal seria poder sintetitzar una proteïna que cobrís les necessitats de tots els aminoàcids (sense excessos, ni defectes) i que fóra totalment digerida i metabolitzada. L'objectiu d'aquest treball és desenvolupar un procés de disseny d'una proteïna ideal completament digestible i aprofitable en pollastres d'engreixament de 0 a 21 dies. A conseqüència, realitzem una investigació sobre els requeriments nets d'aminoàcids i el funcionament del sistema digestiu a nivell enzimàtic dels pollastres de dita edat. A partir d'estes dades dissenyem possibles estructures primàries del polipèptid. A continuació, prediem la seua estructura terciària i les seues propietats fisicoquímiques per mitjà de mètodes computacionals. Les proteïnes obtingudes es valoren basant-se en la seua digestibilitat, característiques fisicoquímiques i projeccions futures de síntesi i producció. Este procediment és aplicable per a l'obtenció de proteïnes en altres etapes vitals de l'animal partint del coneixement de les exigències nutricionals específiques. El mètode seqüencial plantejat presenta una primera aproximació per al disseny de proteïnes en alimentació de precisió i és susceptible de millora i implementació amb els avanços en el camp de l'Enginyeria i Disseny de Proteïnes.

Paraules clau

Aminoàcids, proteïnes, enzims, digestible, modelat computacional, disseny.

Author: Clara María Lledó Morell
Academic tutor: Prof. D. Juan José Pascual Amorós
Academic cotutor: Prof. Dña. María Cambra López
Valencia 6th July 2020

ACKNOWLEDGEMENTS

I would first like to express my sincere thanks to my project tutor and cotutor Juanjo and María, who gave me the opportunity of participating in such a challenging project. For teaching and guiding me with patience and, most important, showing me how is to enjoy making science.

To my family, my greatest supporters in these four years. Specially my parents, who have taught me that hard work and effort pays off. For always being by my side and give me the strength to endure the most critical moments. To my little sister Maria José, for her unparalleled encouragement and her gift to make me laugh at any difficult situation. This accomplishment would not have been possible without you.

Finally, I also wish to thank all the people who have contributed directly or indirectly to this four-year road and this last research.

INDEX

INDEX OF TABLES

INDEX OF FIGURES

1. INTRODUCTION

1.1. The poultry meat sector

Poultry (for meat and eggs) is one of the most relevant sectors in the agricultural industry nowadays. It is based on the breeding, care and commercial exploitation of different domesticated birds being *Gallus gallus* (including chickens and hens) the most important specie from an economical point of view. According to FAO (2017) chickens reared for meat production accounted for 92% of the world's total poultry population, providing 89% of meat production. Chicken reared for meat are called broilers. There are different lines which have been genetically selected for their high growth rate, which has led to a considerable increase in meat production.

Worldwide, Europe ranks fourth and accounts for 12% of total chicken meat production. In Europe, Spain is the second-largest producer of chicken meat behind the United Kingdom. Within Spain, the chicken meat-producing sector accounts for 15% of the final livestock production and 6% of final agricultural production. In the last decade, the number of poultry farms has increased considerably from 14,252 in January 2010 to a total of 19,633 register in January 2020, being 40% of Spanish farms devoted to chicken (MAPA, 2018). In Spain, it should be noted that although the consumption of fresh chicken meat has followed a downward trend since 2012, export demand continues increasing, which explains the growing effort to increase meat production.

The relevance and growth of this sector in the last decades has been accompanied by modernization and constant search for improvements. Thus, the development of new high-performance commercial broiler lines, the use of precision feeding and the automation and refinement of production techniques, with the ultimate goal of lowering costs without affecting productivity, have been promoted. Feed is a major input in broiler production and feed costs can reach 70% of production costs. Modern farmers face the need to feed the animals to achieve maximum performance maintaining the minimum expense.

Feed cost is determined by its different nutrients (mainly energy and protein), as well as the main sources used to obtain these nutrients. In this respect, protein is one of the most expensive nutrients. However, average nitrogen (N) gain per unit of N intake in broiler intensive indoor production is 0.58, being higher to that obtained in free-range (0.49) and organic (0.38) production systems (Kratz et al., 2004). Thus, despite the fact that the broiler is one the most efficient animal in transforming protein into meat, almost half of the protein ingested by this animal is not retained and is excreted. The N lost in the excreta contributes to increasing the N environmental load and caused economic losses. Consequently, different scientific disciplines such as genetics, nutrition and biotechnology seek methods and alternatives to the current feeding model to improve the use of this macronutrient.

1.2. The ideal protein concept

Proteins are macromolecules involved in different functions of living organisms such as catalysis of metabolic reactions (enzymes), transport of molecules and metabolites, providing structural support to tissues, storage and energy functions, mechanisms of defence of the immune system and regulation of cellular functions (hormones) amongst others.

The protein requirement of an animal is determined by its amino acid (AA) requirements that hinge on animal's productive period, genetics, feed consumption and environmental factors

such as temperature. Insufficient protein input, to cover the specific requirements at a given time (related to animal's age and weight), results in a reduction or cessation of growth, starting with nonessential functions. If animals are fed below-requirement levels of amino acids over time, vital functions can be highly affected, as well.

Amino acids are the basic units that build proteins and same as for proteins, they have major roles in the body. In relation to their synthesis, AA can be classified into two groups: essential (EAA) and non-essential (NEAA). There are 11 AA that the broilers body is not able to synthesize by itself, known as EAA, therefore they should be provided in the diet. These are phenylalanine, lysine, threonine, tryptophan, leucine, isoleucine, valine, serine, arginine, histidine and methionine (Quentin et al., 2004). On the other hand, NEAA can be synthesized via body metabolic pathways. The NEAAs play a crucial role in several functions such as gene regulation, signalling, intestinal activity, etc. In any case, NEAAs must also be provided in the diet, since their supply cannot be exclusively based on their obtaining from ingested EAA. For example, cysteine and tyrosine have an indispensable role as precursors for polypeptides formation in cells from the intestinal mucosa and are produced from phenylalanine and methionine, therefore they are not traditionally considered EAA. However, not all animals can produce phenylalanine and methionine *de novo*, consequently, their inclusion in the diet is crucial for proper functioning of the intestine.

In this framework, it has long been considered that including the necessary amounts of EAA in the diet was sufficient to achieve optimal growth as the body could stock up on the necessary amounts of NEAA. However, it has been shown that this assumption is flawed and to achieve maximum growth performance and proper organism functioning the synthesis of NEAA from EAA is not enough. In conclusion, all animals have requirements of all AA to develop their maximum genetic potential and the classification of AA into essential and non-essential is purely descriptive (Wu, 2014).

When all EAA are provided to the animal, if there is one supply which does not meet the animal requirements it is called the limiting aminoacid and it prevents the animal from developing its maximal potential growth and performance because as long as this deficiency is not covered, the animal cannot use the other AA. Once this lack of the limiting aminoacid is covered in the feed then another aminoacid will become limiting (second limiting AA, third, fourth, and so on). The first limiting aminoacid is mostly determine by the type of diet. In poultry diets four essential AA dominate as limiting amino acids: lysine and threonine in most cereals, Methionine in legumes, and tryptophan in maize. First limiting AA are used as reference AA to ratio the needs of all the other AA in the animal diets. This method allows to calculate the exact requirement of the animal for every other EAA resulting in all AA being co-limiting for performance, what is called the ideal protein concept. In the broilers diet, as well as in swine (van Milgen and Dourmad, 2015), the most widely used AA for such purposes is lysine for three main reasons: (1) its use is practically limited to protein accretion and is not affected by other metabolic functions, (2) it does not interact with other AA and (3) its analysis is relatively simple and accurate.

Therefore, one basic idea of the ideal protein concept is that animals need AA in a certain balance to ensure optimum performance. Any absorbed AA which is in excess (due to an excessive inclusion or the lack of a limiting AA) will be oxidised and N will be excreted. Therefore, adjusting the dietary supply of all AA according to the ideal protein helps to maximise N utilisation (Lemme, 2003).

The ideal protein concept was first used in the late 1950s when Mitchell and Scott from Illinois University (United States of America) determined the optimal proportions of EAA in

poultry diets based on the AA composition of casein and egg. However, this first attempt failed due to an excess of EAA and not considering any NEAAs. In 1960, Scott and his team improved the considered requirements after studying EAA's content in the chicken carcass. It was in the 60-70s that NEAAs began to be included in studies about protein content optimization in broiler diets, leading to the drafting of different versions of their standard AA requirements. In spite of their differences, all versions shared common features: they included all EAAs and some AAs synthesized from them; they did not contain alanine, aspartic acid, glutamic acid nor serine; and they considered lysine as the reference AA (Wu, 2014).

Amino acid requirements can be determined through different methods, either with feeding experiments as the dose-response or deletion studies, or by mathematical modelling as the factorial approach based on the N balance by measuring of the net change in total body protein (Lemme, 2003). Thus, the ideal profile for each AA, related to each other by ratios emerges from these studies focused on the AA requirements in broilers. Although the needs of the bird can vary in different situations, the AA ratios remain relatively constant through age and therefore it is easy to adjust the quantities of the rest of AA from a single known reference AA value.

There are numerous studies which have determined lysine requirements in broilers at different ages (Wecke et al., 2016). Combining this information together with the 'Ideal Amino Acid Ratio' (IAAR), the ideal protein profile is designed, which provides the right amount and proportion of AA needed for maximum growth and optimal performance without any deficiencies or excesses. Thanks to this balance it is possible to reduce the amount of protein in feed, optimize its use and reduce N excretions derived from the deamination of over-consumed AA and its excretion as uric acid in the urine. As indicated above, this can contribute to reducing feed costs and limiting the amount of raw material imports needed for feed can mitigate deforestation and reduce its transportation.

For these reasons the optimization of protein intake in diets is a current topic in poultry production world based on the pursuit of providing the ideal protein to the animal. However, there exists no traditional feed source with the ideal protein profile. The efficiency of the use of ingested dietary protein by broilers depends on the digestibility and the AA content and balance relative to the animal requirements. In the field of nutrition this efficiency is usually referred as apparent digestibility coefficient and in chicken it is around 50%, although it is tightly bound to the food source (Bryan et al, 2019) among other factors. Several efforts have been made to overcome the poor protein digestibility of traditional feeding methods. Increasing the crude protein (CP) content has proof to entail negative effects in chicken health, environmental and production cost. (Esmail, 2016). On the other hand, low CP level diets with the addition of crystalline AA neither constitute a suitable solution because it reduces chicken growth performance. Supposedly, crystalline aminoacids are absorbed faster causing an imbalance of AA availability to support protein synthesis (Bryan et al., 2019). In this way, the “perfect” diet in terms of protein supply could be feeding with low level inclusions of highly purified and digestible proteins.

Protein digestibility depends not only on the molecular features of the protein but also on the action of enzymes (i.e. proteases) involved in the digestion process. This process is known as the ‘Protein digestion dynamics’ (Figure 1). In order to allow absorption by enterocytes in the small intestinal mucosa, proteins must be broken down into dipeptides, tripeptides or free amino acids, the specificity of enzymes and their enzyme:substrate ratio will determine the final level of protein hydrolysis achieved.

In poultry, the first enzyme responsible for the initiation of protein digestion is pepsin. It is released by the chief cells in the proventriculus, which auto-activates in an acidic environment

and cleaves full dietary proteins producing smaller peptides that enter the duodenum. Further hydrolysis is performed by pancreatic proteases, mainly trypsin and chymotrypsin endopeptidases (Recoules et al., 2019), as well as elastase exopeptidase and prolidase dipeptidase. (Recoules et al., 2017). Each enzyme has a particular specificity for a substrate cleaving peptide bonds differently, however the activation of some enzymes relies on others such as the activation of chymotrypsin and most pancreatic proteases which is dependent on the presence of trypsin. The final stage of protein digestion occurs at the brush border membrane of the small intestinal mucosa by carboxypeptidases and aminopeptidases (Erickson and Kim, 1990). Particularly in chickens, this last step if performed by carboxypeptidases A and B (Zelikson et al., 1971), and aminopeptidase N (Jamadar et al., 2003).

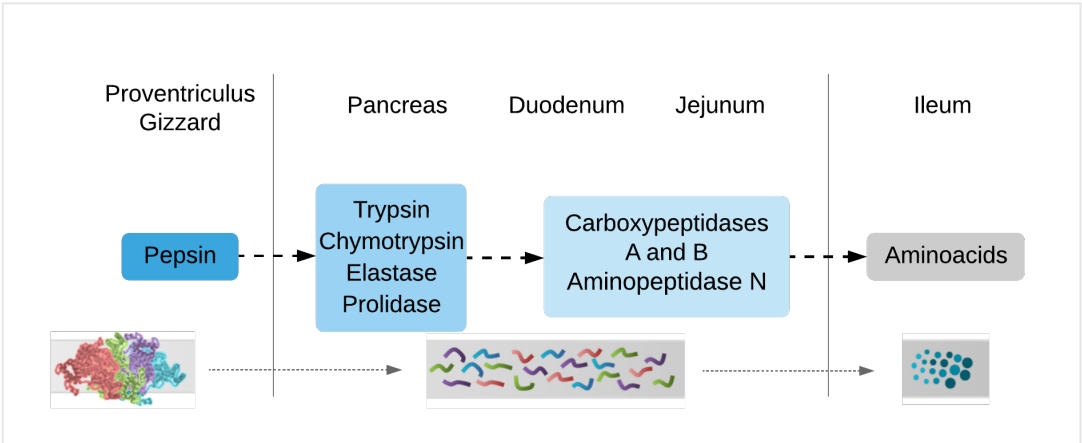


Figure 1. Protein digestion dynamics workflow in chicken.

Our objective is thus to design the exact polypeptide in terms of AA composition and fitting the ideal characteristics for the avian digestion process, a novel issue that certainly implies biotechnology. Until recently, biotechnology has contributed to the field of animal nutrition scope in four main ways: i. biosynthesis of nutrients, including the own synthetic AA and some vitamins; ii. biosynthesis of additives addressed to animal efficiency and health, including enzymes, antibiotics and phages; ii. value-addition to feedstuffs by bioprocessing; iv. biosynthesis of gut microbiota modulators for improving animal performance, including probiotics and prebiotics.

However, the development and search of the Ideal Protein concept opens a new field of application for Biotechnology together with Protein engineering. At present, biotechnology is already in the protein production scope. The manufacture of protein products is nowadays a booming industry, mainly devoted to therapeutic proteins, diagnostic/analytical proteins, and industrial (bulk) proteins. In contrast to AA, chemical synthesis is not a viable option for protein production given its size and complexity, and that is why living cells and their cellular machinery are used as manufacturing factories (Thermo Fisher Scientific, n.d.) increasing the hinder at large-scale production. One of the main challenges of massive protein production, with room for improvement, is the product yield optimization, consequently continuous efforts are made to develop new organisms and look for the most suitable operational conditions to increase the yield. One of the promising ideas that will contribute to the productivity improvement is the use of industrial biodegradable wastes as a substrate for protein production (Spalvins et al., 2018), an economical and strategy that fits perfectly with the objective of this work, which aims at developing the perfect protein with the lowest possible environmental impact.

1.3. Computational protein design and protein structure prediction

Computational protein design (CPD) is a multidisciplinary field established between basic and applied sciences. It can be considered as an evolution of the non-computational *in vitro* protein design including *in silico* methods to design a protein from aminoacidic level to the complex and functional structure conformation. The field has numerous applications in medicine and biotechnology highlighting rational drug design.

The very beginning of CPD field was in the '60s when Christian B. Anfinsen developed the thermodynamic hypothesis supported with his research about ribonuclease A folding (Anfinsen and Haber, 1961; Anfinsen et al., 1961) and many other investigations in the protein field. He hypothesised that the three-dimensional structure of a protein in a normal environment is the one with the lowest overall Gibbs free energy that depends on interatomic interactions. Therefore, he stated that the final conformation of a protein is determined by its aminoacidic sequence through a merely physical mechanism. Due to this postulate, later in 1972, Anfinsen was awarded with half Nobel prize for chemistry.

From then on, the field has experimented an enormous evolution, especially in the last decades. One of the main drivers of such a dramatic improvement in this field in the last years is the advance in computing power and technological breakthroughs, creation of new software and the automatization of experimental procedures that have sped up the obtention of experimental data leading to the growth of protein data bank (PDB). In 2019, 10,581 protein structures were released in PDB almost triplicating the total number of entries available in 2009 (PDB Statistics: Protein-only Structures Released Per Year, https://www.rcsb.org/stats/growth/growth-protein).

Computational protein design is strongly linked with the field of Protein Structure Prediction (PSP), also raised from Anfinsen's research, which aims to elucidate the folding of a protein from the knowledge of its primary structure. The framework of both disciplines is usually overlapped sharing common methods and facing similar limitations, as a matter of fact, PSP can be considered one stage in the CPD or as the target of CPD methodology itself.

As a multidisciplinary field, CPD can pursue different goals: i. protein folding or inverse folding, above mentioned as PSP, ii. design of specific interactions, iii. search for proteins with stability for a particular environment, iv. optimization of natural existing proteins by synthetic biology and v. negative design. Among all the objectives previously mentioned, the problem of protein folding is the most recurrent and one of the biggest challenges that brings together biologists, chemists, engineers and physicists.

There are four levels in the structure of proteins: the simplest level corresponds to the sequence and order of AA in the polypeptide chain giving rise to the primary structure. Amino acids establish joints by hydrogen bridges between them, acquiring certain positions in space that give rise to patterns that form the secondary structure. The most common structural motifs between proteins are alpha helixes and beta sheets, followed by loop regions. Loop regions are irregular structures that act as interconnectors among other motifs of the secondary structure. The tertiary structure is the absolute spatial arrangement of polypeptide atoms through interactions of lateral chains and disulfide bonds between residues. The last level of organization occurs in proteins made up of more than one subunit, such as haemoglobin consisting of four subunits that form two heterodimers (Marengo-Rowe, 2006). The different polypeptides establish non-covalent joints and acquire their quaternary structure necessary to perform their biological functions.

In the scope of PSP, CPD can be applied at different levels. Size is one of the main structural features for classification. The smallest elements are rotamers and conformers determined by the polypeptide side chains. Nowadays, it is possible to find all feasible conformations for each side-chain in rotamer libraries, e.g. Dunbrak rotamer library. The next level is dedicated to the backbone and its critical flexibility that determines the geometrical constraints of the final structure. Working with fragments is another possibility, due to complexity and difficulties in finding proper homology levels between proteins it is possible to combine fragments of different proteins from data banks to solve the overall structure of the problem protein. Lastly, the biggest level implies geometrical general characteristics and application of analytical equations for universal features that can be useful in guiding *de novo* protein design.

Since the beginning, PSP discipline has been divided into two schools of thought: the physics-based and the one based on evolution principles. The stream based on physical principles is rooted in the thermodynamic hypothesis and consequently the methods simulate all possible conformations seeking for the lowest energy state. The number of candidate structures to each aminoacidic chain is gigantic and it constitutes the protein energy landscape. Despite all the advances in the field, nowadays searching in such a vast space is a limitation to overcome but several approaches have been developed as optimized searching algorithms and specific models that increase the success of such methods. On the other hand, the school of thought based on evolution principles includes most methods used in PSP. It relies on the proven scientific theory which claims that evolutionary related proteins preserve common structural features, notwithstanding mutations acquired during divergence. Consequently, many methods rely on the alignment and comparison of a problem protein with other known as templates to infer a final conformation.

The on-going progress in PSP and the emergence of new methods are compiled in the biennial Critical Assessment of Protein Prediction (CASP), a world-wide experiment founded by Mould and co-workers in 1994. Every two years, researchers have the opportunity to present their prediction methods for objective assessment by double-blind resolution of the structure of an aminoacidic sequence. Subsequently, the result presented by the participants is compared with the Nuclear Magnetic Resonance (NMR) spectroscopy or crystallography analysis of the problem protein. In this way, the results and methodology applied by each research group is ranked and all the new data regarding the field is gathered and published in the PSP Center web page (https://predictioncenter.org/index.cgi).

On this basis, computational methods for structure prediction can be divided in four main groups, according to Floudas (2007):

- Comparative modelling, also known as template-based or homology modelling, begins with the selection of a similar protein of known structure as a template, followed by the alignment with the target ending with the modelling to adjust mutations and gaps. The accuracy of the predictions depends on the homology level with target-template; it is considered that from a 30% identity the resulting conformation is mostly successful. Thus, such methods are exclusive for proteins that have a certain degree of homology with others from the PDB. For those proteins without global structural similarity, template-free methods, which do not involve a single template, are the choice.
- Fold recognition or threading works on the basic premise that the number of sequences is much greater than folds, therefore it tries to find among all possible conformations the appropriate for a given sequence despite lacking a homologous template (Rost et al., 1997).

- Ab initio approach mixes databases with physic principles. It is mainly performed as a fragment assembly method, where several templates are selected to perform multiple alignments with the target, allowing to elucidate a structure from the combination of structural features from different known proteins (Bujnicki, 2006).
- Finally, the prediction without previous data exclusively based on physical principles only relies on the Anfinsen dogma and the search of the lowest free energy structure.

The results offered by template-free methods are less reliable yet suitable for each and every protein, that is why they should be always considered.

While it is true that each method has its own mechanism, they all share a set of fundamental steps. The general scheme in predicting protein structures begins obtaining all possible 3D structures from a linear AA chain. Once we have the infinite possibilities, the guided simulation is launched through the landscape guided by a certain energy function. After the search, several candidate structures are generated, among which we will have to determine the final native structure. Because the structures obtained are simplified versions, after determining the shaping must be rebuilt to give higher complexity to both the backbone and the side chains. Once this step has been completed, the model is finished with a refinement that increases the quality of the results (Deng et al., 2018).

But how accurate is the final model when using computational design or prediction of structures? Unlike experimentally obtained structures, where the certainty of the structure can be estimated from the experiment, models need to assess their validity through other alternatives. Such is the task of Model Quality Assessment (MQA) methods, that evaluate the model on the basis of features like solvent exposure, local-side chain and backbone interactions and bonding, molecular environment and secondary structure, and geometric packing. For this purpose, these methods include energy functions, statistical analysis, stereochemical tests and the help of machine learning techniques. The relevance of the aforementioned issue is demonstrated by the creation of a separate category in CASP to evaluate specifically the MQA methods available.

CPB and the PSP problem has been nourishing from the development bioinformatics. Bioinformatics has contributed by providing new tools, servers and software aimed to deal with all aspects of protein modelling. Nowadays a world of possibilities opens up when it comes to modelling informatic tools, from highly complex software to on-line free user-friendly options. The choice is not simple as the software must be tailored to the research specific purpose and previous knowledge about the suitable method and the operating mechanism of the program is necessary as the manual intervention required depends on every single server.

MODELLER software is used for comparative or homology design from an alignment provided by the user and it includes additional tasks to perform during the modelling (Šali and Blundell, 1993). I-TASSER software is an on-line server that predicts 3D-conformation and biological function of an AA sequence, it provides highly accurate results and has been ranked Top 1 server in CASP experiment several editions (Yang and Zhang, 2015). SWISS-MODEL generates a conformational model of a structure, but it is limited by the existence of a homologous in PDB. Conversely, X-RAPTOR is useful for elucidation of structures of non-homology proteins. Higher complexity and deep understanding are needed to perform ROSSETA software. It works as a template-free method based on Monte Carlo algorithm (Zhang and Chou, 1992). Its varied functional modules as Rosetta Design, Rosetta Docking, Rosetta Fragments and others (Kaufmann et al., 2010) allow performing a diverse set of modelling tasks.

The programs and servers mentioned above are just a minor example of the vast collection available, and in the same way, there are also tools devoted to protein optimization and validation (Samant et al., 2014).

In addition, within the framework of bioinformatics and proteins, tools based on the enzymatic activity of some proteins have been developed, from which we can predict the cleavage site and potential substrates on the basis of experimental data available. Online servers as PoPS (Boyd et al.,2005), PeptideCutter (Gasteiger et al., 2005) and CaSPredictor (Garay-Malpartida et al, 2005) are used as *in silico* methods to predict and model proteases digestion of known peptides. *In silico* analysis with PeptideCutter have proven to be useful in the prediction of peptides generated by proteases (Chica and Manuela, 2017) before the *in vitro activity* confirmation. It has been used with different objectives mostly Proteomics studies (Chong et al., 2010). In the present work, PeptideCutter will be applied to simulate the gastrointestinal digestion of the proteins, following the applications described in Cavatorta et al. (2010) and Yang et al. (2019).

2. OBJECTIVES

The main objective of the present study was to develop a novel process for the design and structure modelling of an ideal protein completely digestible and usable to meet nutritional requirements of broiler chickens from 0 to 21 days of age. The designed protein will provide the exact amount of each AA without any deficiency or excess for optimal performance and maximal growth of the animal.

To achieve the overall goal, specific objectives are addressed:

- To design protein primary structures containing the minimal AA quantities that can be fully digested by enzymes from avian digestive system.

- To predict and model secondary and tertiary conformations of already designed polypeptides.

- To examine and validate different protein quality evaluation measures to assess candidate proteins and select the most optimal based on digestibility and reliability of the predicted structure.

3. MATERIALS AND METHODS

3.1. Sequence design (primary structure)

3.1.1. Amino acid composition and Minimal ideal protein profile

De novo protein design of a protein that fully meets the requirements of broilers, should be designed based on the net amino acid requirement at each age of the animal (maintenance needs + growth requirements). Although the UPV Animal Feeding research group is working to obtain this information, it is not yet available. For this reason, the present work has been based on the use of the closest available information to these net requirements, which corresponds to the true ileal digestible AA.

Furthermore, total AA requirements can change with age. However, to define the frequency of these AA in the novel protein to be designed, only the relative requirements (with respect to lysine) are needed. These relative requirements do not change as much with age and they may not even differ too much from the net ones. In any case, the present work will focus on the design of a *de novo* protein to cover the requirements for true ileal digestible AA in broilers from 0 to 21 days.

Table 1 shows different ideal AA profiles for broilers from 0 to 21 days, based on the available literature and recognized international nutritional guidelines: Canadian NRC (National Research Council, 1994), Spanish FEDNA (Fundación Española para el Desarrollo de la Nutrición Animal (Santomá and Mateos, 2018), Dutch CVB (Veeroederbureau,2008), Brazilian Tablas Brasileñas para Aves y Cerdos (Rostagno et al., 2017), North American Texas AM University (Wu, 2014), as well as amino acidic content analysis in 10-day chicken's meat (Wu, 2014).

Among all the collected data, Texas AM University recommendations (Wu, 2014) was selected to calculate the total quantity of each single AA to construct the "minimal ideal protein" containing a total of 108 amino acids (Table 1), and thereafter for larger versions of the protein. The main reasons were that these recommendations are not far from the current recommendations for most of the AA provided by FEDNA and NRC (main national and international benchmarks). Moreover, it is one of the only ones that provided recommendations for the 20 AA and it has been derived from true ileal digestibility AA contents, accounting for the proportion of AA in the whole body of broilers.

3.1.2. Evaluation of avian digestive enzymes function

A deep and comprehensive study of the avian digestive system and enzymes action was conducted. The choice of enzymes was based on experimental data from Recoules et al. (2017) about *in vivo* digestion of vegetable proteins in broiler. Thereupon pepsin, trypsin, chymotrypsin, elastase, prolidase, carboxypeptidase A and B and aminopeptidase were characterized in terms of substrate specificity and activity based on the available literature (Table 2). Because there is little information available regarding chicken digestive enzymes other species were included in the research ensuring that both considered enzymes from each specie present substantial homology and the main characteristics were conserved.

Table 1. Amino acids requirements for chickens from 0 to 21 days.

AMINOACIDS	Mw (g/mol)	Mw (g/molecule)	NRC 1994[1]	FEDNA 2018 [2]	CVB 2018 [3]	BRASIL 2018 [4]	Texas AM University (Wu,2014) [4]	10 days Chicken meat	Texas AM - FEDNA	TAMU/ 10d meat	Sequence x 1
Lysine	146,19	2,43E-22	100	100	100	100	100	100	0	1,00	6
Alanine	89,09	1,48E-22					102	108		0,94	6
Arginine	174,20	2,89E-22	114	105	105	108	105	111	0	0,95	7
Asparagine	132,12	2,19E-22					56	59,3		0,94	4
Aspartate	133,10	2,21E-22					66	70,1		0,94	4
Cystein	121,16	2,01E-22	36	34	35	33	32	24,4	-2	1,31	2
Glutamate	147,13	2,44E-22					178	135		1,32	11
Glutamine	146,15	2,43E-22					128	82,1		1,56	8
Glycine	75,07	1,25E-22	66,5	73,5	84	85,5	176	187	102,5	0,94	11
Histidine	155,16	2,58E-22	32			37	35	34,3	3	1,02	2
Isoleucine	131,17	2,18E-22	73	67	66	67	67	58,4	0	1,15	4
Leucine	131,17	2,18E-22	101	107		107	109	113	2	0,96	7
Methionine	149,21	2,48E-22	45	40	38	39	40	30,7	0	1,30	3
Phenylalanine	165,19	2,74E-22	65,5			63	60	56,6	-5,5	1,06	4
Proline	115,13	1,91E-22	54,5				184	195,6	129,5	0,94	12
Serine	105,09	1,75E-22	47,5	52,5	60	61,5	69	73,2	16,5	0,94	4
Threonine	119,12	1,98E-22	72	65	65	65	67	59	2	1,14	4
Tryptophan	204,23	3,39E-22	18	17	16	17	16	18,9	-1	0,85	1
Tyrosine	181,19	3,01E-22	56,5			52	45	43,3	-11,5	1,04	3
Valine	117,15	1,95E-22	82	78,5	80	77	77	68	-1,5	1,13	5
										Nº AAs:	108
										Mw(g/mol):	12492,88

Mw: Molecular weight, NRC: (National Research Council, 1994), FEDNA: Fundación Española para el Desarrollo de la Nutrición Animal, CVB: (Veeroederbureau,2008) , BRASIL: Tablas Brasileñas para Aves y Cerdos (Rostagno et al., 2017) , TAMU/10d meat: Texas AM amino acid calculations divided by amino acid amount on ten days chicken meat , AAs: amino acids.

[1]calculated from total amount in diet, [2]calculated from real faecal digestibility, [3]calculated from apparent faecal digestibility, [4] calculated from true ileal digestibility

Table 2. Main proteases from the avian digestive system involved in protein digestion

Enzyme	E.C.	Type	Substrate especificity [1] [2]	References
Pepsin	3.4.23.1	Endopeptidase	– P1 exerts the greatest influence on the cut – Cleavage occurs over 40% of the time after Phe or Leu and more than 30% Met (P1) – No cleavage after His, Lys, Pro, and Arg (P1) – Aromatic residues are favoured at position P1' – Pro is forbidden at the P2 position – His, Lys, and Arg are disfavoured at the P3 position – Do not cleave dipeptides	Baudys and Kostka (1983); Wang et al. (1995); Hamuro et al. (2008).
Trypsin	3.4.21.4	Pancreatic endopeptidase	– Cuts preferentially after Lys and Arg (P1) – Pro at P1' hinders the cleavage – Arg, Ile, Leu, Lys or Phe at P2 decreases the activity 2 to 16-fold – Pro at position P3 decreases activity 3 to 9-fold.	Zelikson et al. (1971); Olsen et al. (2004); Rodriguez et al. (2008).
Chymotrypsin	3.4.21.1	Pancreatic endopeptidase	– Preference cleaves after Trp, Tyr, Phe and Met (P1) – Cuts with less preference Leu and His on P1 – Cut after Trp is prevented by Met or Pro in P1' – Cut after Met is prevented by Tyr in P1' – Cut after His is prevented by Met, Trp or Asp in P1' – Pro at P1' blocks the cleavage	Zelikson et al. (1971); Schellenberg et al. (1991);
Elastase	3.4.21.36	Pancreatic endopeptidase	– Preference cleavage after Ala, Leu, Gly, Val or Ile (P1)	Guyonnet et al. (1999);
Prolidase	3.4.13.9	Pancreatic dipeptidase	– Acts on X-Pro dipeptides – Cleaves Pro imino-terminal	Davis and Smith (1957);
Carboxypeptidases (CBPA, CBPB)	3.4.17.1 3.4.17.2	Exopeptidase	– Splits C-terminal of di- and tripeptides – CBPA cleaves aromatic AA or long aliphatic side chains: Ile. Leu, Phe, Tyr, and Trp – CBPB cuts after Arg and Lys	Barrett et al. (2012).
Aminopeptidase N	3.4.11.20	Exopeptidase	– Splits N-terminal of di- and tripeptides – Preferentially hydrolyses Leu followed by Ala, Phe, Tyr and Gly (P1') – Pro or Val residue at the P-1 or P-2 position are poorly or not hydrolysed	Gal-Garber and Uni (2000); Damle et al. (2010)

[1]Three letter aminoacid code: Ala: Alanine, Arg: Arginine, Asn: Asparagine, Asp: Aspartic acid, Cys: Cysteine, Glu: Glutamic acid, Gln: Glutamine, Gly: Glycine, His: Histidine, Ile: Isoleucine, Leu: Leucine, Lys: Lysine, Met: Methionine, Phe: Phenylalanine, Pro: Proline, Ser: Serine, Thr: Threonine, Trp: Tryptophan, Tyr: Tyrosine, Val: Valine.
[2]General nomenclature of cleavage site positions by Schechter and Berger (1968): cleavage site between P1-P1', incrementing the numbering in the N-terminal direction of the cleavage (P2, P3, P4, etc..) and on the carboxyl side in the same way (P1', P2', P3' etc.).

3.1.3. Approaches for primary structure modelling

From the 108 AA described in Table 1, a first starting protein sequence was generated with randomized order of residues using RandSeq, a free-access program on the ExPASy online portal (SIB Bioinformatics Resource Portal). This sequence was named "blank" and was used as a control for comparison purposes in the present work. The online tool is frequently used to build randomly scrambled peptide libraries from a specific AA composition for different purposes (Grishaeva and Bogdanov, 2013; Singh et al., 2019). From the original random sequence, several primary structures were designed following three different approaches (Figure 2). These approaches are described below:

1. Approach 1 - Sequential optimization exclusively based on Peptide Cutter software information about enzymes performance (ExPASy Bioinformatics Portal, Swiss Institute of Bioinformatics); step-by-step per individual enzyme.
2. Approach 2 - Sequential optimization based on data on Peptide Cutter software information updated with our own collected data in the 'Avian Digestive enzymes study' (see 1.2; Table 2) about substrate specificity of each enzyme; step-by-step per individual enzyme.
3. Approach 3 - Non-sequential direct optimization considering all enzymes at once on the overall sequence.

In all three approaches, the optimization was made-by-hand, changing residue positions in order to maximize the number of cleavages by the enzymes in the linear polypeptide chain.

During sequential optimization (Approaches 1 and 2) several runs of PeptideCutter digestion were performed. Beginning with pepsin, each run included one more enzyme following the theoretical order of action in the chicken protein digestion mechanism. In contrast, Approach 3 included several runs of PeptideCutter with all the possible enzymes at once, simulating digestion with all enzymes as a pool. The performance of Approach 3 included several assumptions:

- The substrates of each enzyme are different and the ideal sequence to maximize the cleavage efficiency of one of them can reduce the ulterior efficiency of another, so that a global vision allows optimizing and ensuring that the most of each enzyme is obtained.
- Although *in vivo* enzymes act sequentially many cuts are redundant, therefore the final result from making a pool or doing following a step-wise procedure will be similar.

It was not possible to simulate digestion with all digestive proteins as some of them were not available in the program. Prolidase, carboxypeptidases and aminopeptidase N digestion simulation was manually performed.

After the manual optimization approaches (1 to 3), three different primary sequences were obtained named as 'Round 1', 'Round 2' and 'Round 3' (Figure 2). Subsequently, complete digestion of these three sequences (rounds) plus the original blank random sequence were simulated to show the maximum digestibility of the linear polypeptides and asses the most convenient approach. Once the best methodology was chosen, it was used to obtain four extra sequences. The process of each sequence will begin with the generation of a random sequence and its subsequent optimization, so that finally we will obtain four new sequences with the ideal AA profile but different primary structures. These four new sequences will be considered as candidates for being the ideal protein.

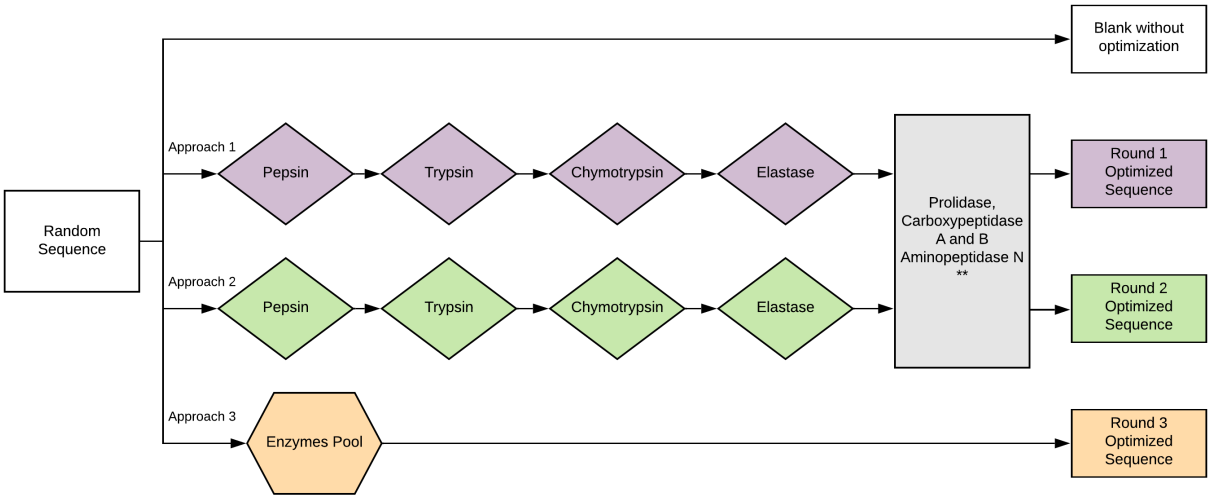


Figure 2. Optimization approaches flow chart. Top-down: Blank Sequence (white), sequential optimization based on Peptide Cutter enzymes (purple), sequential optimization based on Peptide Cutter plus additional data on enzymes (green), all at once optimization (orange). **Grey box: unavailable enzymes on PeptideCutter server.

3.1.4. Improvement and refinement

All the final sequences obtained were subjected to a manual refinement step to increase their digestibility. In other words, increasing the number of free-AA in the final sequence by adding some extra specific AA that will break the remnants dipeptides. Such extra AA were chosen under two criteria: being the target AA of various digestive enzymes and having been rounded down in the proposed minimum ideal protein.

After this step, a total of six primary structures (Round 2, Round 3 and the four extra sequences) that can be fully digested based on *in silico* methods were assessed as candidates of the perfect protein.

Additionally, it was studied the possibility of producing the ideal protein increasing the length of its sequence up to ten-fold. Larger proteins sequences were compared to the minimal length (x1) considering how they fit chickens nutritional requirements, as well as the possible advantages and disadvantages from production point of view.

3.2. Structure prediction

Two on-line servers were used to predict the folding of the different sequences: I-TASSER and QUARK, both developed in the Yang Zhang Lab at the University of Michigan, Ann Arbor.

- **I-TASSER**

Iterative Threading ASSEmbly Refinement is a hierarchical protocol for structure and functional prediction of AA sequences. It comprises three consecutive steps: threading, fragment assembly and iteration (Figure 3) (Yang and Zhang, 2015; Zhang, 2007). First, from the AA sequence, the program searched homologous templates from proteins in PDB library using a simple Profile-Profile Alignment (PPA) approach. The score of alignment of the problem protein with the template residues was calculated as described in Equation 1:

$$Score(i,j) = \sum_{k=1}^{20} F_{query}(i,k) P_{template}(j,k) + c_1 \delta\left(s_{query}(i), s_{template}(j)\right) + c_2$$

(Equation 1)

where F_{query} (i, k) is the frequency of the k^{th} amino acid at the i^{th} position of the multiple alignments; $P_{template}$(j, k) is the sum of the log-odds profile of template sequence in the PSI-BLAST search; S_{query}(i) is the secondary structure prediction from PSIPRED for the i^{th} residue of the query sequence; and $S_{template}$(j) is the secondary structure assignment by DSSP for the j^{th} residue of the template. The weight factor c_1 is an adjustable parameter for balancing the profile term and the secondary structure matches, and the shift constant c_2 avoids the alignment of unrelated regions in the local alignment.

Next, aligned fragments of templates were translated to construct several full-length models while the unaligned regions were built from scratch by *ab initio* modelling using Monte Carlo simulation (MC). The assembled models by MC were clustered by SPICKER and used to construct a representative model for each group. Further on, these models were refined by full-atomic simulations in order to obtain the lowest-free-energy conformation. The server provided 5 top-ranked models based on the confident score calculation (C-score) calculated as described in Equation 2:

$$C - score = ln\left(\frac{M/M_{tot}}{\langle \mathrm{RMSD} \rangle} * \frac{1}{N}\sum_{i=1}^{N}\frac{Z_i}{Z_{cut,i}}\right)$$

(Equation 2)

where M/M_{tot} is the number of structure decoys in the SPICKER cluster divided by the total number of decoys from the I-TASSER simulations; 〈RMSD〉 is the average root-mean-square deviation of the decoys to the cluster centroid; and Z_i /$Z_{cut,i}$ is the normalized Z -score of the best template gene, rated by the threading program.

C-score value varied from -5 to 2 and is correlated with the proposed structure quality. A C-score >-1.5 can be considered as a reliable model. Finally, functional aspects were obtained from matching with proteins in BioLiP library (http://zhanglab.ccmb.med.umich.edu/BioLiP/).

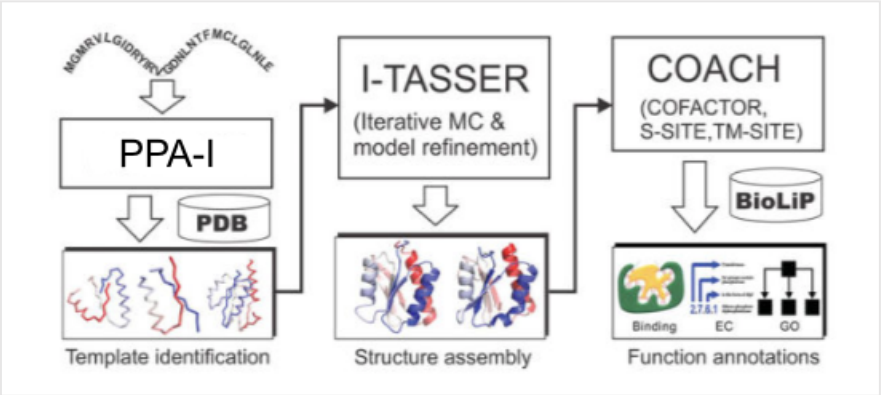


Figure 3. The I-TASSER protocol for protein structure and function prediction. Adapted from Yang and Zhang (2016). PPA-I: Profile-Profile threading alignment, PDB: Protein data bank, COACH: function annotation program, BioLip: database for ligand-protein binding interactions.

We support the election of this server on the world-wide CASP experiments, where I-TASSER has been ranked as best server for structure prediction in the latest editions CASP7, CASP8, CASP9, CASP10, CASP11, CASP12 and CASP13; and also ranked as the best for function prediction in CASP9. Along the CASP contest, it has been shown that methods based on the combination of different techniques for the structural prediction provide better results. In this framework, I-TASSER is one example of a composite approach that has demonstrated to provide high accuracy models. Moreover, while many servers are limited to structural elucidation, I-TASSER goes one step further and provides information about the possible biological functions of the model protein including ligand binding sites, Enzyme Commission (EC) and Gene Ontology (GO). The timing was a relevant factor in the conduction of the project, I-TASSER output generation takes ~36 hours for a medium-size protein (~200 AA) although it depends on the number of previous jobs submitted to queue and the query protein prediction difficulty. Compared to most online-servers available I-TASSER provides the best quality/time relation.

- **QUARK**

Quark was chosen as secondary server, it is based on *ab initio* folding, construction of protein structures by fragment assembly from unrelated proteins (W. Zhang, et al., 2016). The pipeline is described in Figure 4. It began with a compilation of continuously distributed fragments of 1 to 20 AA. 4000 structures were generated at each position based on gapless threading between the problem fragment and a library of 6023 high-resolution PDB structures. Next, a distance profile containing low-range interactions was obtained. In the next step, MC simulations were applied to obtain full-length models under a physics- and knowledge-based potential assembly. Finally, the decoys from the simulations were clustered by SPICKER and ranked by the size of the clusters.

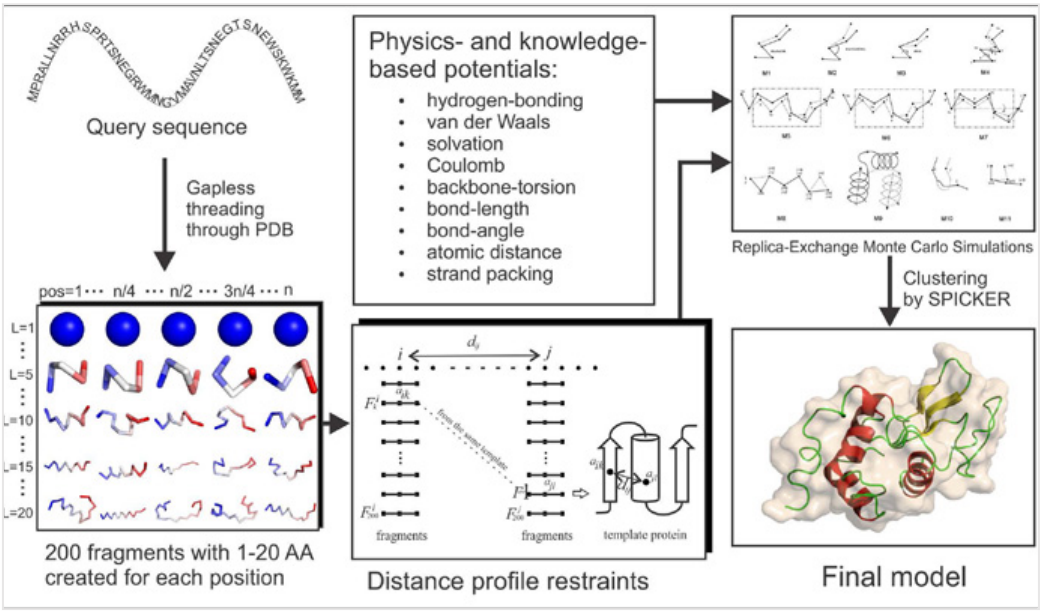


Figure 4. Flowchart of QUARK structure assembly simulations (Zang et al., 2016)

QUARK has also played successfully in CASP experiments ranked as the leader server in Free-modelling (FM) in CASP9 and CASP10. It provided a visual output which contains the information about secondary structure prediction, but unlike I-TASSER it does not provide information on biological function. It is considered as the best prediction method for short (<200 AA) hard targets, in other words, for that short proteins without significant homologous templates in PDB. The time consumption of QUARK is higher than I-TASSER, depending on the structure complexity and the queue length of prior submitted jobs it takes from 2 to 5 days. While it is not the fastest *ab initio* modelling software, time is forsaken in favour of model accuracy (Yousef et al., 2019).

3.3. Assesment and comparisons of the predicted structures

The I-TASSER predictions for each structure were benchmarked on the basis of different criteria:

1. Quality and reliability of the Top 1 model based on three main parameters:

- C-score value:

Confidence score that asses the estimated accuracy of the proposed model ranging from -5 to 2, increasing with high confidence. It is calculated based on the significance of threading template alignments and the convergence parameters of the structure (see above).

- TM-score:

Measuring of the structural similarity between the predicted model relative to the native structures based on C-score as described in Equation 3. Being a TM-score > 0.5 similar topology and TM-score< 0.3 random similarity.

$$TM - score = Max\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right]$$

(Equation 3)

where L_N is the length of the native structure; L_T is the length of the peptides to the template structure; d_i is the distance between the i^{th} pair of aligned residues; d_0 is a scale to normalize the match difference; and 'Max' means the maximum value after optimal spatial superposition (Zhang and Skolnick, 2004a).

- Cluster density:

The number of structure decoys from MC Simulation in the SPICKER cluster. Where higher cluster density means higher occurrence of the structure in the simulation trajectory and a better-quality model, calculates as Equation 4 shows:

$$cluster\ density = \frac{M}{(M_{tot} \times \langle RMSD \rangle)} \quad \text{(Equation 4)}$$

where M is the number of decoys in the cluster; Mtot is the total number of decoys used by SPICKER; and <RMSD> is the average Root Mean Square Deviation (RMSD) of the decoys to the cluster centroid (Zhang and Skolnick, 2004b).

2. Quantity and type of structural motifs: number of -helices (H) and ß sheets (S) obtained from the 'Secondary Structure Prediction' visual output, Considering the relation between digestibility and protein conformation based on experimental data (Carbonaro et al., 2012).

3. Predicted function using COFACTOR and COACH which includes Ligand-Binding Site, EC number and GO term prediction. All three evaluated with its own C-score calculation in the range of [0,1], where a C-score of higher value signifies high confidence and *vice-versa*. Specific consideration is taken to GO-score, where a GO-score > 0.5 is considered reliable gene homology.

After I-TASSER results assessment, QUARK predictions were used to support and complement the secondary structure predictions from I-TASSER. Finally, the conclusions of each structure were compared to select the best option to meet the final objective of the project

3.4. Final model evaluation

Once the ideal protein was selected its model was validated using different tools. The structure stereochemical evaluation was carried out by Ramachandran Plot via PROCHECK (Laskowski et al., 1993) measuring the torsion angles among the residues that make up the protein in the model which allows to determine which aminoacids are permitted or not in each position. Further validation of the protein folding energy was evaluated by using ProSA server (Wiederstein and Sippl, 2007) widely used to figure out possible errors in 3D protein models comparing the model with already registered structures of proteins with same sequence length in the PDB.

4. RESULTS AND DISCUSSION

During the experimental procedure, many decisions were made through trial-and-error until defining the optimal methodology to design the ideal protein balancing results quality and time invested.

4.1. Sequence design (primary structure)

4.1.1. Sequence construction (approaches assessment)

The three different approaches used for the obtention of the primary structure of our polypeptide were initially evaluated based on the digestion level of the final chain, as well as the difficulty and time devoted to the optimization.

Figure 5 shows the final sequences after simulated digestion following each approach. The final number of peptides and free AA are described in Table 3. The blank sequence digestion only reached the theoretical release of around 48% of the AA that composed the polypeptide. The rest were found mainly as dipeptides, tripeptides and three notorious resistant polypeptides of more than three residues. Comparing these products of digestion with the ones from the three different approaches, where only some dipeptides remained undigested, it is clear that manual optimization of the sequences proved to be a useful procedure to optimize enzyme digestion and AA release in the avian digestive tract. Amino acid release using Rounds 1 to 3 ranged from 91 to 93% (Table 3).

BLANK: MPEENQSPPTQV-K-A-T-R-FQH-L-EN-I-F-Q-L-CV-G-G-A-NPV-EH-L-K-EG-Y-R-E-F-G-G-QPPGS-PR-EG-DDTSQ-I-PN-L-L-PR-I-PQ-EM-I-R-QA-S-K-K-DA-W-D-P-L-A-K-G-R-V-Y-G-EM-EG-K-F-E-R-A-G-C-L-T-P-Y-V

Round 1: M-E-P-F-Q-K-S-K-QV-N-P-L-H-E-F-N-R-V-C-P-L-TG-C-K-Q-K-A-EV-N-P-L-EH-E-I-G-S-P-Y-G-E-R-A-G-Q-P-F-T-K-G-E-P-L-D-R-G-T-P-L-Q-I-Q-R-V-DM-N-P-W-T-I-M-E-P-L-S-I-Q-R-A-S-P-Y-Q-K-A-A-D-P-F-G-D-R-V-G-E-L-G-R-G-E-Y-A-G-E

Round 2: M-H-Q-K-V-E-P-F-E-P-L-S-R-A-Q-P-F-EV-Q-R-G-C-P-L-G-N-L-V-H-N-K-E-I-G-E-P-Y-G-T-R-S-R-G-Q-P-F-E-P-L-G-D-P-W-Q-K-G-Q-P-L-S-I-Q-R-D-R-N-I-M-N-K-A-T-K-T-I-A-E-P-L-D-Y-SA-D-P-F-V-G-E-P-L-EA-M-G-E-K-G-EA-Q-Y-G-CR-V-T

Round 3: M-E-P-F-V-N-P-F-A-E-P-L-Q-K-T-R-H-N-I-Q-P-L-V-G-G-A-V-E-P-L-N-K-EG-Q-Y-Q-R-G-E-P-F-G-QH-G-S-R-EG-Q-I-N-P-L-M-E-P-L-S-R-T-I-Q-I-D-R-QA-S-K-D-K-A-D-W-D-P-L-A-T-K-G-S-R-V-E-Y-G-M-V-C-K-G-E-P-F-E-R-A-G-C-P-L-T-P-Y-E

Figure 5. Results from *in silico* digestion of initial designed sequences for primary structure with different approaches. Approaches: Blank, Sequence without optimization; Round 1, obtained by sequential optimization based on Peptide Cutter enzymes; Round 2, obtained by sequential optimization based on Peptide Cutter plus additional data on enzymes; Round 3, obtained by all at once optimization. One letter amino acid code: A: Alanine, C: Cysteine, D: Aspartic acid, E: Glutamic acid, F: Phenylalanine, G: Glycine, H: Histidine, I: Isoleucine, K: Lysine, L: Leucine, M: Methionine, N: Asparagine, P: Proline, Q: Glutamine, R: Arginine, S: Serine, T: Threonine, V: Valine, W: Tryptophan, Y: Tyrosine.

Table 3. Number of peptides and free aminoacids from *in silico* digestion of initial design sequences.

Sequence	Dipeptides	Tripeptides	>3 residues polypeptides	Free AA
BLANK	14 (25.9%)	2 (5.6%)	3 (20.4%)	52 (48.1%)
Round 1	5 (9.3%)	0	0	98 (90.7%)
Round 2	4 (7.4%)	0	0	100 (92.6%)
Round 3	4 (7.4%)	0	0	100 (92.6%)

Values in brackets indicate the percentage represented in the total sequence.
Approaches: Blank, Sequence without optimization; Round 1, obtained by sequential optimization based on Peptide Cutter enzymes; Round 2, obtained by sequential optimization based on Peptide Cutter plus additional data on enzymes; Round 3, obtained by all at once optimization.

Looking into the results from the three different rounds, the most digestible sequence based on the action of the chicken digestive enzymes are those obtained by Rounds 2 and 3, because there is a slight improvement in the number of free amino acids compared to Round 1 (Figure 5). These results were the expected as Round 2 was performed analogously to Round 3, with updated information on enzymes performance including the expansion of specific substrates. Likewise, Round 3 provided a broader perspective of the overall enzymes digestion (assumption referred above), what was assumed to be an advantage for optimization.

Both Rounds 2 and 3, considered as the best candidates/approaches, were time-consuming. The time invested to complete each approach from the random sequence to the final primary structure was considered as a relevant factor. Nevertheless, as illustrated in Figure 2, the number of steps required to fulfil Round 3 were considerably lower. This fact is supported by researchers experience during the development of the present experiment.

In summary, the approach followed in Round 3 to obtain the primary structure was the most suitable for AA sequence optimization based on digestion results with avian enzymes and, in consequence, it was used to generate four extra sequences (Rounds 3.1, 3.2, 3.3 and 3.4) in the improvement and refinement procedures addressed to obtain a complete digestible protein.

4.1.2. Improvement and Refinement

Round 2, Round 3 and the extra sequences generated in the last step (Figure 6) presented the same digestibility result pattern (93% of free AA release and four remnant dipeptides, Table 3). In order to improve the sequences by breaking the four remnant dipeptides, four extra AA were included in the composition taking into account the following considerations:

- Choosing those AA that are a frequent target for digestive enzymes in chickens (arginine, isoleucine, leucine, lysine, phenylalanine, tryptophan and tyrosine).
- The calculated AA requirements were round off to exact numbers and thus the ones rounded down could imply a risk of shortage. Therefore, it was more advisable to use those AA that had been rounded down. That was the reason why isoleucine and lysine, were chosen for the additional provision. Special attention was paid on lysine, due to its role as first limiting and as reference AA, being worthwhile to ensure its minimum requirement.
- Lysine, sulphur amino acids (methionine and cysteine), Arginine and tryptophan have been described as the first limiting amino acids in chicken diets. Once their requirements were met, methionine and cysteine were avoided for being sulphur-containing amino

acids (risk of disulphide bonds, that difficult digestive enzymes efficiency), while arginine and tryptophan were chosen for extra inclusion.

-

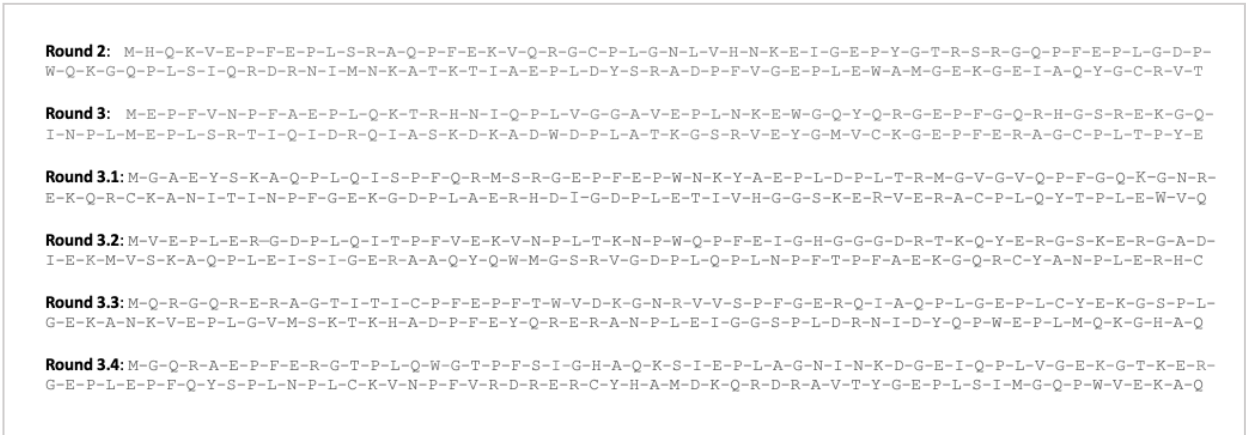


Figure 6. Final 112 residues sequences improve for complete digestion. Round 2: obtained by sequential optimization based on Peptide Cutter plus additional data on enzymes; Round 3, Round 3.1, Round 3.2, Round 3.3 and Round 3.4: obtained by all at once optimization.

All things considered, the addition of the four extra AA resulted in increasing only slightly the amount of isoleucine by a 25.00%, lysine by 16.67% and arginine by 14.29%, but tryptophan by 100%. The most considerable increase is that of tryptophan since the initial sequence contained only one AA and in the refinement step its amount was doubled. Studying the nature and biological functions of tryptophan, no evidence indicating that an excess in the diet could result in considerable adverse effects on the animal was found. By contrast, tryptophan can be beneficial in as it participates in different biological essential functions (Mund et al., 2020). Tryptophan works as a precursor of different hormones including serotonin, which is related to diminishing stress before slaughter and can thus have a direct effect on production performance and final meat quality (Bai et al., 2017). Moreover, several experiments have shown the direct role of Tryptophan in humoral and cellular immune response (Emadi et al., 2011; Mund et al., 2020). Therefore, tryptophan addition was considered as safe and beneficial. Emadi et al. (2011), in a study on infectious bursal disease in chickens, suggested that increasing two times the NRC level of tryptophan along with 2.5 times the NRC level of arginine in chickens basal diet could have positive effects for immune response. In this way, Wang et al. (2014) concluded that increasing 1.5 fold dietary tryptophan recommendations can improve chicken welfare and relieve oxidative stress.

Regarding the other three extra AA, their amount was increased in much lower percentage compared with tryptophan. Therefore, we assumed that *a priori* there does not seem to be a counterproductive effect on the animal. This assumption is supported by experimental studies on the effects of excess lysine (Ghoreyshi et al., 2019), isoleucine (Farran et al., 2003) and arginine (Ebrahimi et al., 2014) in the diet, showing that when added above requirements, chicken growth was not affected.

The decision to add extra AA involves moving away in greater or lesser extent from established chickens net requirements, and the effects of the imbalances can only be determined with certainty in *in vivo* trials. An alternative to the addition of AA is the removal of some of them, obtaining a protein with a shorter sequence while providing the removed AA in an isolated form through supplementation. Nowadays although all AA are commercially available synthetically produced or in the form of crystalline AA, only DL-methionine, L-lysine, L-threonine, L-arginine and L-tryptophan are used in animal feeding.

Despite the large availability of AA, this strategy presents few constraints. Firstly, although the market price of AA has decreased in the last decade thanks to the development of biotechnology, synthetic AA are expensive and its supplementation is not economically profitable. Additionally, digestive dynamics of supplemental and protein-bound AA are different, considering rates and sites of absorption along the digestive tract (Selle and Liu, 2019), which could cause a lack of synchronization in the absorption to support protein synthesis leading to a reduction in chicken performance as reported by Bryan et al. (2019).

4.1.3. Protein length

After the definition of the best approach for sequence optimization and the sequence refinement using the minimum sequence to meet the broilers requirements, we considered increasing the size by repeating the minimum sequence. These proteins could equally meet the requirements of these animals, and perhaps could be obtained without the need for extra AA. Therefore, the defined overall procedure described previously was also used to obtain larger sequences from this minimum amino acidic sequence obtained (results not provided).

Figure 7 shows the number of extra AA required to be expressed as a % of the AA requirements (standardised to a single size, x1) to obtain a full digestible protein in function of the sequence length. Our results showed that the number of standardised extra AA required to be included in the improvement and refinement steps increased linearly as the length of the sequence increased (from x1 to x4 sizes) and then, from x4 and beyond, remained more or less constant. In consequence, it seems that the best fitting sequence to meet the chicken's requirements is the protein designed with the minimum number of sequence repetitions (i.e. size x1).

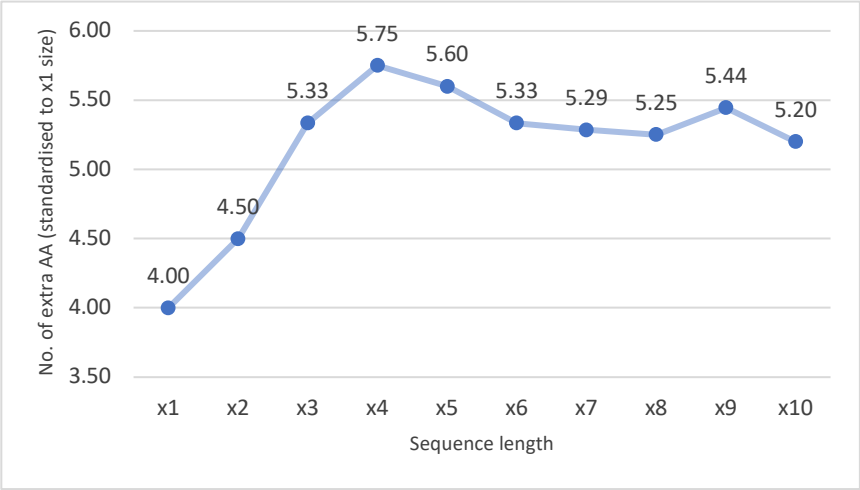


Figure 7. Number of extra aminoacids (standardised to x1 size) needed for complete digestion according to protein length.

As regards protein size, a likely event during protein production is denaturing and aggregation of proteins. This leads to the formation of inclusion bodies inside the expression system, mainly caused by cell stress produced by the overexpression of heterologous proteins (Kopito, 2000; Villaverde and Carrió, 2003). The probability of aggregation increases when the produced proteins are larger than those typical produced by the host, as in mammalian protein expression in *Escherichia coli*. While aggregation eases protein recovery by simply breaking the cells and centrifuging, it also reduces possible toxicity to the host. The refolding step is a high

inefficient process and the downstream operations in order to recover a soluble and well-folded protein can be very complicated and expensive (Clark, 2001). Whether a protein will form inclusion bodies or not in a specific expression vector and how difficult it will be to recover its native folding is very uncertain and difficult to predict. Therefore, a soluble protein is always the preferable choice, and that is why the ideal protein is designed with the minimum size (x1).

In addition, during protein synthesis errors occur in the translation process, leading to wrongly placed AA on an estimated ratio of 6×10^{-4} to 5×10^{-3} per AA incorporated (Zaher and Green, 2009). This implies that as the size of the proteins increases, so does the probability of encountering errors in the AA chain. This fact does not always imply a detrimental effect, but in some cases, it may affect the subsequent folding and functionality of the protein. From the point of view of obtaining the ideal protein, these errors are very relevant since the amount of AA that make up the protein have been defined exactly to cover the minimum needs of the chicken and a mismatch would not meet the requirements and therefore the protein would not achieve the overall objective for which it was created. Therefore, the production of the ideal protein in its smallest form is beneficial as it also diminishes the possibility of AA misincorporations.

4.2. Protein Structure Prediction

4.2.1. Evaluation of the secondary structure models

- Benchmarking I: Model quality and accuracy

The final sequences (Round 2, Round 3, Round 3.1, Round 3.2, Round 3.3 and Round 3.4) were provided as input in the I-TASSER on-line server which developed a Top 1 model for each sequence with its statistical parameters. Three main parameters (C-score, TM-score and cluster density) were considered to predict the absolute or relative quality of each protein model in order to select the best quality 3D model. The server also provided the root-mean-square deviation of atomic positions (RMSD) calculation but although this index can give an explicit concept of modelling errors, it was not considered as a main parameter for the conformation assessment because, in some cases, an isolated local error on template-sequence alignment can cause large RMSD value even though the global topology might be adequate (Kufareva and Abagyan, 2011).

Table 4 shows the values of the main statistics used to evaluate the quality and reliability of the secondary structure models obtained with the different protein sequences. Protein structure predicted using Round 3.3 sequence showed the most reliable and quality model, as indicated by their C-score (accuracy) and TM-score (similarity to native structures) values higher than the other sequences (-1.08 and 0.58, respectively). However, Round 3.1 also had significant high C and TM scores (-1.99 and 0.48). The cluster density, related with higher occurrence and quality, was also higher for Round 3.3 sequence compared with the rest (0.42 in Round 3.3 *vs.* 0.24 on average in the rest).

Table 4. Quality and reliability traits for the secondary and tertiary structure of the Top 1 I-TASSER predicted models for the different protein sequences.

Protein Sequence	Statistics parameters			
	C-score	TM-score	RMSD	Cluster density
Round 2	-4.24	0.27±0.08	14.1±3.8	0.0212
Round 3	-3.30	0.35±0.12	11.6±4.5	0.0525
Round 3.1	-1.99	0.48±0.15	8.4±4.5	0.1945
Round 3.2	-4.10	0.28±0.09	13.7±0.09	0.0281
Round 3.3	-1.08	0.58±0.14	6.4±3.9	0.4161
Round 3.4	-4.26	0.26±0.08	14.1±3.8	0.2300

C-score: confident score, TM-score: template modelling score, RMSD: Root mean square deviation.

The models presented as 'Top 1' by I-TASSER, i.e. those the program suggests as the most reliable models of each sequence, are those that present the highest values in the parameters mentioned above. Theoretically, this implies a higher reliability in predicting the actual 3D protein conformation. This assumption is consistent with many studies in which the predicted structure has been experimentally validated using computational methods before the native structure is available. For example, Kemege et al. (2011) studied a *Chlamydia trachomatis* protein CT296 in which I-TASSER was used for *in silico* structure prediction and was subsequently validated by X-ray chromatography. In this study, from the five models predicted by I-TASSER, Model 1 had the highest C-score (considerably higher value than the remaining four models) and X-ray crystallography showed significant overall structural similarity with the Model 1 I-TASSER predicted structure.

Only in some cases the differences between the C-scores of the five I-TASSER proposed models are not very significant and the Top 1 Model cannot be selected immediately. S.Zhang et al. (2016) studied *Aspergillus niger* N5-5 tannase and showed how I-TASSER Model 2 may be the closest-native structure model although its C-score value was slightly lower than Top 1 Model C-score. In these situations, models must be studied on the basis of other characteristics.

For all the sequences in this experiment the Top 1 Model were selected undoubtfully as they presented much higher C-score values than the other four proposed models (data not provided). Using the same criteria, a general valuation of all proteins in Table 4 can be done considering that, among all the Top 1 models of the different sequences presented by I-TASSER, the one with the highest values of quality and reliability parameters will represent the best and most reliable prediction among all the possible sequences.

Obtaining an accurate model for the secondary structure is essential, since this structure will be closely related to protein physical (solubility, aggregation, secretion ability) and functional characteristics. As previously mentioned, protein solubility is very relevant regarding production processes, mainly affecting cell excretion and recovery downstream processes as it is directly related with the aggregation phenomena. Solubility is mainly defined as a function of solvent characteristics (pH, salt concentration, temperature, etc.) and protein structure. Solubility starts to be defined from the primary structure, each AA has its own and different solubility and water affinity depending on their molecular nature (Schein, 1990). The contribution of each AA to the polypeptide sequence, as well as their distribution, guides protein folding in order to acquire the most stable structure by burying the non-polar AA and exposing the soluble residues to interact with water in the molecular surface. This can directly affect

overall protein solubility (Trevino et al., 2007). Moreover, secondary structural motifs can also affect peptide solubility. Solubility can increase with the ratio of α-helix to β-sheet in *in vitro* experiments (Bai et al., 2016). Therefore, a model with higher reliability would allow us to have more solid information about the possible behavior and properties of the protein in order to anticipate future complications and develop accurate experimental trials and production plans.

All these models have been obtained by I-TASSER, a homology modelling server that uses threading. Therefore, the quality is tightly determined by the existence of protein templates in PDB with significant sequence similarity to the problem sequence. In other words, sequences without homologous in PDB will be more difficult to model and the result will be based in less evidence leading to overall lower reliability results. In fact, as it can be seen in Figure 8, the secondary sequence predicted with the most reliable model obtained by I-TASSER (Round 3.3), includes many coil regions. These coil regions between α-helices and β-sheets may be true or a consequence of a lack of homologous information in the PDB.

(a)

```
                    20        40        60        80        100
                    |         |         |         |         |
Sequence    MQRGQRERAGTITICPFEPFTWVDKGNRVVSPFGERQIAQPLGEPLCYEKGSPLGEKANKVEPLGVMSKTKHADPFEYQRERANPLEIGGSPLDRNIDYQPWEPLMQKGHAQ
Prediction  CCCCCCCCCCSSSSCCCCCCSSSCCCCSSSCCCCHHHHHCCCCCCCSSCCCCCCCCCCCCCCCCSSSCCCCCCCHHHHHHHHCCCCCCCCCCCCCCCCCCCCHHHHHCCCCC
Conf.Score  975521116615862677705560698563663123330325785440358850211345354312424545880666663078411278755565677756277763357 9
            H:Helix; S:Strand; C:Coil
```

(b)

Figure 8. Predicted secondary and tertiary structure of sequence Round 3.3 by I-TASSER. (a) 3D structure cartoon model. In pink α-helices, in yellow β-sheets and in white coil regions. (b) Secondary predicted structure. H: α-helices, S: β-sheets, C: Coil regions.

Besides this, a unique protein, whose structure greatly differed from any known homologous, could be advantageous. Firstly, it could decrease the mimic phenomena that could lead to problems in the organism chosen for future production (whether bacteria, yeasts or any other choice organism). Secondly, it is more probable that, if it does not belong to any protein family, it will not have any relevant function itself.

- Benchmarking II: Structural secondary motifs

In terms of protein digestibility and solubility, it is crucial to consider the occurrence of two main structural patterns: α-helices and β-sheets. Carbonaro et al. (2011) studied *in vitro* the structure-digestibility relationship of different proteins from animal and plant origin and quantified the different structural motifs. Their results, consistent with other experiments (Gabriel et al., 2008; Yang et al., 2016), showed an inversely proportional decrease in hydrolysis

degree (HD) to the number of ß-sheets. The main explanation lies in the hydrophobic character of these structures that promotes aggregation and protein-protein interaction.

Figure 9 shows the percentage of overall sequence covered by α-helices, β-sheet and random coil regions from I-TASSER models. This figure shows that the secondary structure of the protein simulated in Round 3.3 had the higher amount of β structures among all the candidates, and therefore it could be the less digestible protein. Following the same criteria, the secondary structure of Round 3.1 protein could have a higher HD as its structure contained the lowest percentage of β-sheet and simultaneously the highest number of α-helices among all the models. Regarding the amount of coil regions, as their conformational prediction is more intricate, these regions could raise unexpected folds in a future production phase. This is why defined structures and α-helices are preferable. The structure of the protein simulated in Round 3.1 presented one of the lowest percentages of coil regions.

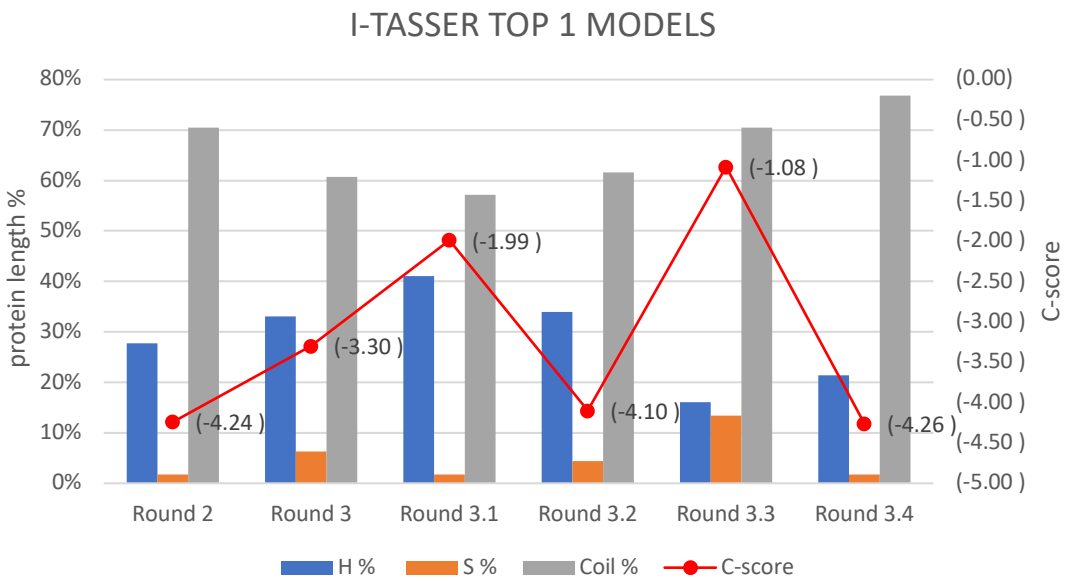


Figure 9. Secondary structure motifs as percentage of total protein length in I-TASSER predicted Top 1 Models. H: α-helices, S: β-sheets.

The sequences were submitted to QUARK to evaluate its predicted models compared to I-TASSER. QUARK results shown in Figure 10 support the secondary structure of predicted Top 1 Model from I-TASSER in all the sequences, in fact QUARK outputs showed an increase in the confidence score of single AA in the secondary structure prediction.

Moreover, in Round 2, Round 3, Round 3.2, Round 3.3 and Round 3.4 the QUARK models placed few extra motifs on defined coil regions in the I-TASSER prediction, which can be visualized when Figures 9 and 10 are compared. where a decrease in coil sequence length percentage was observed. The only model that remained constant for both software predictions was Round 3.1.

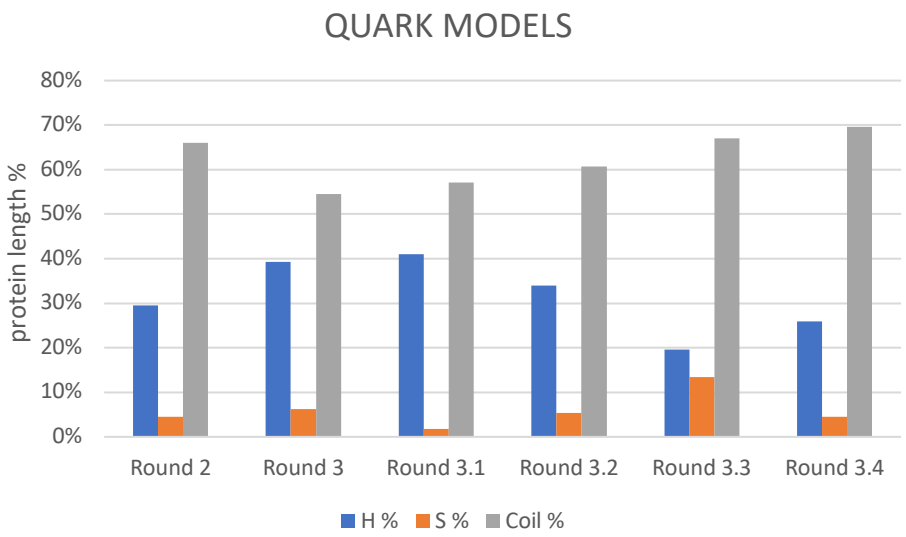


Figure 10. Secondary structure motifs as percentage of total protein length QUARK predicted models. H: α-helices, S: β-sheets

- Benchmarking III: Predicted Biological Function

As previously mentioned, another aspect to be taken into account when producing a new heterologous protein in an expression system, is that this protein does not have biological functions that could affect the normal functioning of the host organism.

Function prediction is based on 3 subsections: Ligand-binding sites, EC and GO. Table 5 shows the values of the parameters used for function prediction from the I-TASSER predicted Top 1 Models for each sequence. Virtually all proteins have negligible confidence scores below 0.5 regarding Ligand Binding Site prediction and EC. This implies that none of the proteins is predicted to bind ligand neither to belong to an already existing enzyme classification. GO-scores of some of the proteins clearly differed from the rest. The most notable case was Round 3.3 which consensus prediction of GO terms values were above 0.5 threshold. This indicated molecular function, biological process and cellular localization. Likewise, protein simulated in Round 3.1 showed a GO-score of biological process prediction lightly above the cut-off, corresponding to cellular catabolic process (GO:0044248). Nevertheless, given the low values of the other two prediction terms (molecular function and cellular component), it could be considered as a protein with low interaction with host metabolism.

Table 5. Function prediction parameters from Top 1 I-TASSER predicted models for the final sequences.

Protein Sequence	Function prediction parameters				
	Ligand binding site C-score	EC C-score	Molecular function GO- score*	Biological process GO-score*	Cellular component GO- score*
Round 2	0.10	0.085	0.37	0.07	0.37
Round 3	0.07	0.084	0.10	0.32	0.10
Round 3.1	0.11	0.153	0.47	0.53	0.15
Round 3.2	0.09	0.066	0.24	0.37	0.07
Round 3.3	0.17	0.279	0.71	0.81	0.56
Round 3.4	0.09	0.065	0.47	0.07	0.07

GO: gene ontology, EC: enzyme commission.
*Consensus terms

Despite I-TASSER is among the most reliable software for the forecast of proteins function and it was also ranked as the best for function prediction in CASP9, it must be noted that the prediction accuracy will never be absolute. Nowadays, there are many available methods for predicting function mainly based on protein sequence, three-dimensional structure or genome annotations (Watson et al., 2005). Nonetheless, some targets are still complicated to tackle by a single strategy and certain methods are only appropriate for a specific protein type. For example, COFACTOR consistently outperforms simple homology-based analysis so it depends on template availability and its accuracy will be lower for proteins with novel structures. The information provided by different methods can concur or can be different. However, inferred results will be never conclusive and the only way to verify protein's function is experimentally. Machine learning methods can provide clues about the future activity of proteins but only *in vivo* studies of the molecule and its cellular context, considering full complexity interactions and pathways, can provide absolute certainty.

As we are not looking for a functional protein, it is better to avoid proteins prone to present any specific role. Functional proteins can be toxic when overexpressed in the host organism and, depending on the function of the expressed recombinant protein, it can have detrimental effects on the proliferation and differentiation of the host cell used as expression system, diminishing protein production yield. As toxicity is such a common phenomenon, there are many different strategies to overcome it (Ahmad et al., 2018). One of them is based on removing protein activity by expressing them as unfolded peptides that form inclusion bodies. As previously mentioned, this strategy entails an additional cost and difficulty in protein production as it requires downstream processes to purify and isolate the proteins (Mustafa et al., 2019) and then to return the protein to its native structure and function (Vallejo and Rinas, 2004). This problem is avoided with completely dysfunctional proteins that will not be toxic to the host cell though they create metabolic burden for the host cell.

With that in mind, protein Round 3.3 was not considered as the best candidate because its functional predictions made it prone to cause host cell toxicity and future problems when producing the molecule. In this framework, Round 3.1 could be considered as a protein with low interaction with host metabolism.

4.2.2. Summary and final protein evaluation

From all the information obtained in this experiment (related to digestive release of AA, the precision and reliability of the proposed secondary structure and the possible interaction with the host) our aim was to define which of the models could be the most suitable to design a completely digestible protein that covered all nutritional requirements of broilers and that theoretically will allow us to produce it efficiently. Considering quality and reliability assessment of the structure obtained, the best model could be the sequence proposed in Round 3.3. However, based on the relationship between secondary structure – DH, Round 3.3 protein could be expected to have a low digestibility potential due to its higher richness in β-sheet structures.

Regarding the digestibility potential based on structural motifs, the best protein was Round 3.1, as it contained the highest number of α-helices and the lowest of β-sheet. Moreover, it was ranked second in terms of reliability in the basis of C-score and TM-score showing an acceptable quality level. In addition, the scores related to potential biological functions that could affect the normal functioning of the host are clearly better in the protein obtained in Round 3.1 than in Round 3.3.

The rest of the models seemed to have a lower reliability due to their higher proportion of coil regions. The coil regions correspond mostly to unaligned fragments, this happens because these sequences have no significant homologs in the PDB and therefore less evidence and templates to construct the final conformation.

In view of the above, there is no point in selecting a reliable model if the desired protein characteristics are not met. We prefer to have a medium-quality prediction model of a protein that meets all the requirements. In conclusion, our choice was the protein sequence obtained in Round 3.1, considering its theoretical structural digestibility, meaningful quality of the predicted model, and reduced biological and functional predictions.

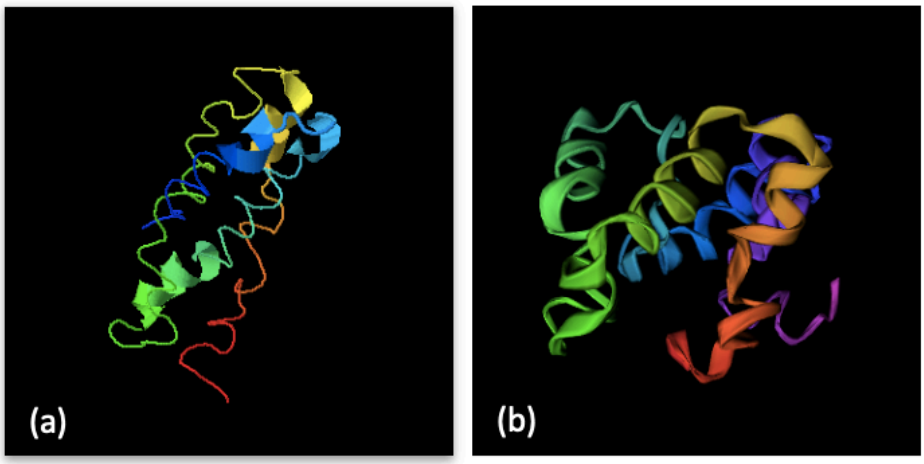


Figure 11. Round 3.1 protein 3D structure cartoon model. (a) I-TASSER model, (b) QUARK model.

Once the model corresponding to Round 3. 1 protein (Figure 11) has been selected, it was validated. Validation of three-dimensional models is a core aspect on PSP before structure deposition on PDB. It is also an essential subject in structural biology. Therefore, validation tools based in different criteria have been developed as PROCHECK and ProSA, which were used in this experiment.

Firstly, PDB file for Round 3.1 model 1 was downloaded and submitted to PROCHECK online server where Ramachandran plot is developed for the structure validation (Figure 12 (a)). Ramachandan plot which shows the overall residue by residue/structural geometry analysis, demonstrated that 43.7% of residues appeared in favoured regions, although other 48.3% of residues were laid in the allowed region; and 8.0% of residues fell in disallowed regions. These percentages from Ramachandran analysis of protein structure demonstrate that our displayed protein is predicted to be quite stable with a minor amount of AA in the disallowed regions. Compared with Beg et al. (2018), the residue percentage in totally favoured region is lower but the residues in not allowed regions is similar and is within the range of acceptable values. Therefore, *a priori* it is not a model of unstable nature.

Protein folding energy was carried by ProSA server by comparing the model with structures of same size in relation to the AA chain registered in PDB. A Z-score value for the model indicates the overall quality. Protein Round 3.1 had a Z-score of −1.25, as shown in Figure 12 (b) plot. This value is within the range of scores typically found for native proteins of similar size experimentally determined by Nuclear Magnetic Resonance (NMR) spectroscopy.

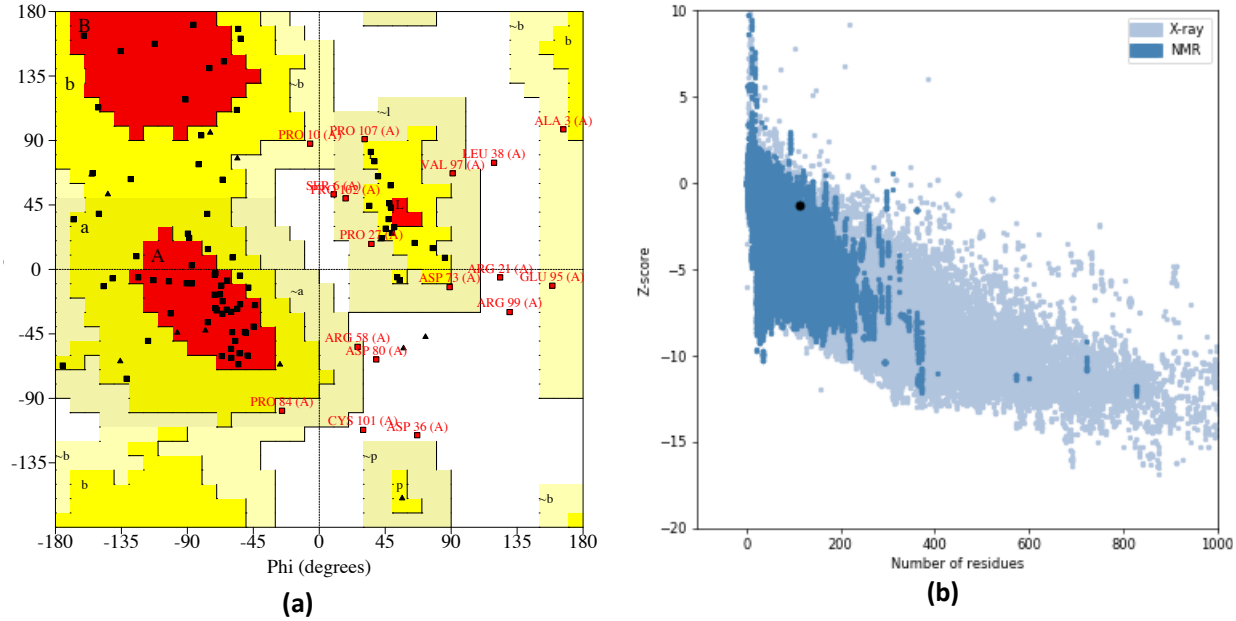


Figure 12. Model validation plots. (a) Ramachandran plot of Round 3.1 protein predicted structure model from PROCHECK. Dark dots represent amino acids, and red zones A, B, and L represent the most favoured regions. (b) Z-score of input protein Round 3.1 using ProSA. ProSA-web z-scores of all protein chains in PDB determined by X-ray crystallography (light blue) or Nuclear Magnetic Resonance spectroscopy (dark blue) with respect to their length.

In light of the results obtained in the validation of the Round 3.1 model, we can conclude that this is a reliable model as it fits with existing data from other proteins of similar size. According to its geometry, it is relatively stable, although there is room for improvement. Consequently, we considered this choice as the *de novo* ideal protein.

5. CONCLUSIONS

The development of a novel process for the design and structure modelling of an ideal protein for broiler chickens from 0 to 21 days of age has been successfully achieved. From our research, the following main conclusions can be drawn:

- Amino acidic sequences with full *in silico* digestibility, using the less time-consuming method, were obtained through a non-sequential direct optimization considering all enzymes at once on the overall sequence.
- The protein primary structure composed exclusively by the AA net requirements of chickens from 0 to 21 age did not ensure complete protein digestion by proteases actions, despite all efforts on sequence optimization. Therefore, adding extra AA, initially rounded down, was necessary to obtain a perfect substrate based on digestion dynamics.
- The most adequate size to produce the ideal protein was the minimal (x1, 112 AA), as it is the most proximal to the exact ideal AA profile and it can be beneficial regarding biological synthesis and industrial production.
- Computational predicted models can be used to predict the most reliable future protein structure, being the α -helices, the most digestible motifs.
- Secondary and tertiary structure modelling can give clues about the future protein functional features and behaviour.

Finally, for our main goal, the ideal protein must be chosen combining structural digestibility, quality and reliability of the predicted model and reduced biological and functional predictions. In any case, more studies will have to be done in the future to improve the definition of this protein, taking into account the characteristics of the potential hosts, before carrying out the first pilot tests aimed at its biosynthesis.

6. REFERENCES

AHMAD, I., NAWAZ, N., DARWESH, N. M., UR RAHMAN, S., MUSTAFA, M. Z., KHAN, S. B., and PATCHING, S. G. (2018). Overcoming challenges for amplified expression of recombinant proteins using Escherichia coli. *Protein Expression and Purification, 144*, 12-18.

ANFINSEN, C. B., and HABER, E. (1961). Studies on the reduction and re-formation of protein disulfide bonds. *Journal of Biological Chemistry*, *236*(5), 1361-1363.

ANFINSEN, C. B., HABER, E., SELA, M., and WHITE JR, F. H. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, *47*(9), 1309-1314.

BAI, M., QIN, G., SUN, Z., and LONG, G. (2016). Relationship between molecular structure characteristics of feed proteins and protein *in vitro* digestibility and solubility. *Asian-Australasian Journal of Animal Sciences, 29*(8), 1159.

BARRETT, A. J., WOESSNER, J. F., and RAWLINGS, N. D. (Eds.). (2012). *Handbook of proteolytic enzymes* (Vol. 1). Elsevier.

BAUDYŠ, M., and KOSTKA, V. (1983). Covalent structure of chicken pepsinogen. *European Journal of Biochemistry*, *136*(1), 89-99.

BEG, M., THAKUR, S. C., and MEENA, L. S. (2018). Structural prediction and mutational analysis of Rv3906c gene of Mycobacterium tuberculosis H37Rv to determine its essentiality in survival. *Advances in Bioinformatics,* 2018.

BOYD, S. E., PIKE, R. N., RUDY, G. B., WHISSTOCK, J. C., and DE LA BANDA, M. G. (2005). PoPS: a computational tool for modeling and predicting protease specificity. *Journal of Bioinformatics and Computational Biology*, *3*(03), 551-585.

BRYAN, D. D. S. L., ABBOTT, D. A., VAN KESSEL, A. G., AND CLASSEN, H. L. (2019). *In vivo* digestion characteristics of protein sources fed to broilers. *Poultry science*, *98*(8), 3313-3325.

BUJNICKI, J. M. (2006). Protein-structure prediction by recombination of fragments. *Chembiochem*, *7*(1), 19-27.

CARBONARO, M., MASELLI, P., and NUCARA, A. (2012). Relationship between digestibility and secondary structure of raw and thermally treated legume proteins: a Fourier transform infrared (FT-IR) spectroscopic study. *Amino acids*, *43*(2), 911-921.

CAVATORTA, V., SFORZA, S., AQUINO, G., GALAVERNA, G., DOSSENA, A., PASTORELLO, E. A., and MARCHELLI, R. (2010). In vitro gastrointestinal digestion of the major peach allergen Pru p 3, a lipid transfer protein: molecular characterization of the products and assessment of their IgE binding abilities. *Molecular Nutrition and Food Research*, *54*(10), 1452-1457.

CHICA, S., and MANUELA, E. (2017). Desarrollo de metodologías alternativas para la evaluación de proteínas precursoras de péptidos bioactivos. (Trabajo de fin de grado). Universitat Politècnica de València.

CHONG, F. C., TAN, W. S., BIAK, D. R. A., LING, T. C., and TEY, B. T. (2010). Modulation of protease activity to enhance the recovery of recombinant nucleocapsid protein of Nipah virus. *Process Biochemistry*, *45*(1), 133-137.

CLARK, E. D. B. (2001). Protein refolding for industrial processes. *Current Opinion in Biotechnology, 12*(2), 202-207.

DAMLE, M., HARIKUMAR, P., and JAMDAR, S. (2010). Chicken intestine: A source of aminopeptidases. *Science Asia*, *36*, 137-141.

DAVIS, N. C., and SMITH, E. L. (1957). Purification and some properties of prolidase of swine kidney. *Journal of Biological Chemistry, 224(*1), 261-275.

DENG, H., JIA, Y., and ZHANG, Y. (2018). Protein structure prediction. *International Journal of Modern Physics B*, *32*(18), 1840009.

EBRAHIMI, M., ZARE SHAHNEH, A., SHIVAZAD, M., ANSARI PIRSARAEI, Z., TEBIANIAN, M., RUIZ-FERIA, C. A., ... and MOHAMADNEJAD, F. (2014). The effect of feeding excess arginine on lipogenic gene expression and growth performance in broilers. *British Poultry Science, 55*(1), 81-88.

EMADI, M., JAHANSHIRI, F., KAVEH, K., HAIR-BEJO, M., IDERIS, A., and ALIMON, A. R. (2011). Nutrition and immunity: the effects of the combination of arginine and tryptophan on growth performance, serum parameters and immune response in broiler chickens challenged with infectious bursal disease vaccine. *Avian Pathology, 40*(1), 63-72.

ERICKSON, R. H., and KIM, Y. S. (1990). Digestion and absorption of dietary protein. *Annual Review of Medicine*, *41*(1), 133-139.

ESMAIL, S. H. (2016). Understanding protein requirements. Retrieved 6 May 2020, from https://www.poultryworld.net/Nutrition/Articles/2016/11/Understanding-protein-requirements-2914798W/

FARRAN, M. T., BARBOUR, E. K., and ASHKARIAN, V. M. (2003). Effect of excess leucine in low protein diet on ketosis in 3-week-old male broiler chicks fed different levels of isoleucine and valine. *Animal Feed Science and Technology, 103*(1-4), 171-176.

FLOUDAS, C. A. (2007). Computational methods in protein structure prediction. *Biotechnology and Bioengineering*, *97*(2), 207-213.

GABRIEL, I., QUILLIEN, L., CASSECUELLE, F., MARGET, P., JUIN, H., LESSIRE, M., ... and BURSTIN, J. (2008). Variation in seed protein digestion of different pea (Pisum sativum L.) genotypes by cecectomized broiler chickens: 2. Relation between in vivo protein digestibility and pea seed characteristics, and identification of resistant pea polypeptides. *Livestock Science, 113*(2-3), 262-273.

GAL-GARBER, O., and UNI, Z. (2000). Chicken intestinal aminopeptidase: partial sequence of the gene, expression and activity. *Poultry Science*, *79*(1), 41-45.

GARAY-MALPARTIDA, H. M., OCCHIUCCI, J. M., ALVES, J., and BELIZÁRIO, J. E. (2005). CaSPredictor: a new computer-based tool for caspase substrate prediction. *Bioinformatics*, *21*(1), i169-i176.

GASTEIGER, E., HOOGLAND, C., GATTIKER, A., WILKINS, M. R., APPEL, R. D., and BAIROCH, A. (2005). Protein identification and analysis tools on the ExPASy server. *The proteomics protocols handbook* (pp. 571-607). Humana press.

GHOREYSHI, S. M., OMRI, B., CHALGHOUMI, R., BOUYEH, M., SEIDAVI, A., DADASHBEIKI, M., ... and SANTINI, A. (2019). Effects of dietary supplementation of l-carnitine and excess lysine-methionine on growth performance, carcass characteristics, and immunity markers of broiler chicken. *Animals, 9*(6), 362.

GRISHAEVA, T. M., and BOGDANOV, Y. F. (2013). On the origin of synaptonemal complex proteins. Search for related proteins in proteomes of algae, lower fungi, mosses, and protozoa. *Russian Journal of Genetics: Applied Research*, *3*(6), 481-486.

GUYONNET, V., TŁUSCIK, F., LONG, P. L., POLANOWSKI, A., and TRAVIS, J. (1999). Purification and partial characterization of the pancreatic proteolytic enzymes trypsin, chymotrypsin and elastase from the chicken. *Journal of Chromatography A*, *852*(1), 217-225.

HAMURO, Y., COALES, S. J., MOLNAR, K. S., TUSKE, S. J., and MORROW, J. A. (2008). Specificity of immobilized porcine pepsin in H/D exchange compatible conditions. *Rapid Communications in Mass Spectrometry*, *22*(7), 1041-1046.

JAMADAR, V. K., JAMDAR, S. N., DANDEKAR, S. P., and HARIKUMAR, P. (2003). Purification and characterization of aminopeptidase from chicken intestine. *Journal of Food Science*, *68*(2), 438-443.

KAUFMANN, K. W., LEMMON, G. H., DELUCA, S. L., SHEEHAN, J. H., and MEILER, J. (2010). Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry*, *49*(14), 2987-2998.

KEMEGE, K. E., HICKEY, J. M., LOVELL, S., BATTAILE, K. P., ZHANG, Y., and HEFTY, P. S. (2011). Ab initio structural modeling of and experimental validation for Chlamydia trachomatis protein CT296 reveal structural similarity to Fe (II) 2-oxoglutarate-dependent enzymes. *Journal of Bacteriology, 193*(23), 6517-6528.

KOPITO, R. R. (2000). Aggresomes, inclusion bodies and protein aggregation. *Trends in Cell Biology, 10*(12), 524-530.

KRATZ, S., HALLE, I., ROGASIK, J., and SCHNUG, E. (2004). Nutrient balances as indicators for sustainability of broiler production systems. *British Poultry Science*, *45*(2), 149-157.

KUFAREVA, I., and ABAGYAN, R. (2011). Methods of protein structure comparison. *In Homology Modeling* (pp. 231-257). Humana Press.

LASKOWSKI, R. A., MACARTHUR, M. W., MOSS, D. S., and THORNTON, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, *26*(2), 283-291.

LEMME, A. (2003). The “Ideal Protein Concept” in broiler nutrition 1. Methodological aspects-opportunities and limitations. *Degussa AG Amino News*, *4*(1), 7-14.

MARENGO-ROWE, A. J. (2006). Structure-function relations of human hemoglobins. *Proceedings (Baylor University. Medical Center)*, *19*(3), 239–245.

MINISTERIO DE AGRICULTURA Y PESCA, ALIMENTACIÓN Y MEDIO AMBIENTE. DIRECCIÓN GENERAL DE PRODUCCIONES Y MERCADOS AGRARIOS. (2018). El sector de la avicultura en cifras. Principales indicadores económicos 2017. MAPAMA; pp.77.

MUND, M. D., RIAZ, M., MIRZA, M. A., RAHMAN, Z. U., MAHMOOD, T., AHMAD, F., and AMMAR, A. (2020). Effect of dietary tryptophan supplementation on growth performance, immune response and anti-oxidant status of broiler chickens from 7 to 21 days. *Veterinary Medicine and Science, 6*(1), 48-53.

MUSTAFA, A. D., KALYANASUNDRAM, J., SABIDI, S., SONG, A. A. L., ABDULLAH, M., RAHIM, R. A., and YUSOFF, K. (2019). Recovery of recombinant Mycobacterium tuberculosis antigens fused with cell wall-anchoring motif (LysM) from inclusion bodies using non-denaturing reagent (N-laurylsarcosine). *BMC Biotechnology, 19*(1), 27.

NATIONAL RESEARCH COUNCIL. (1994). Nutrient Requirements of Poultry Ninth Revised Edition. *National Academy Press.* Washington DC.

OLSEN, J. V., ONG, S. E., and MANN, M. (2004). Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular and Cellular Proteomics*, *3*(6), 608-614.

ORGANIZACIÓN DE LAS NACIONES UNIDAS PARA LA ALIMENTACIÓN Y AGRICULTURA. (2017). Producción y productos avícolas. Retrieved from http://www.fao.org/poultry-production-products/production/es/

QUENTIN, M., BOUVAREL, I., BASTIANELLI, D., and PICARD, M. (2004). Quels" besoins" du poulet de chair en acides aminés essentiels? Une analyse critique de leur détermination et de quelques outils pratiques de modélisation. *Productions Animales*, *17*(1), 19-34.

RECOULES, E., LESSIRE, M., LABAS, V., DUCLOS, M. J., COMBES-SOIA, L., LARDIC, L., ... and RÉHAULT-GODBERT, S. (2019). Digestion dynamics in broilers fed rapeseed meal. *Scientific Reports*, *9*(1), 1-11.

RECOULES, E., SABBOH-JOURDAN, H., NARCY, A., LESSIRE, M., HARICHAUX, G., LABAS, V., ... and RÉHAULT-GODBERT, S. (2017). Exploring the *in vivo* digestion of plant proteins in broiler chickens. *Poultry Science*, *96*(6), 1735-1747.

RODRIGUEZ, J., GUPTA, N., SMITH, R. D., and PEVZNER, P. A. (2008). Does trypsin cut before proline?. *Journal of Proteome Research*, *7*(01), 300-305.

ROST, B., SCHNEIDER, R., and SANDER, C. (1997). Protein fold recognition by prediction-based threading. *Journal of Molecular Biology*, *270*(3), 471-480.

ROSTAGNO, H. S., TEXEIRA ALBINO, L. F., HANNAS, M. I., LOPES DONZELE, J., SAKOMURA, N., PERAZZO, F. G., and DE OLIVEIRA BRITO, C. (2017). Tablas Brasileñas para Aves y Cerdos. (4th ed.). Viçosa: Departamento de Zootecnia.

ŠALI, A., and BLUNDELL, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, *234*(3), 779-815.

SAMANT, L. R., SANGAR, V. C., and CHOWDHARY, A. (2014). Online servers and offline tools for protein modelling, optimization and validation: a review. *International Journal of Pharmaceutical Sciences Review and Research*, *28*, 123-127.

SANTOMÁ, G., and MATEOS, G.G. (2018). Necesidades nutricionales en avicultura, Normas FEDNA. *Fundación Española para el Desarrollo de la Nutrición Animal (2nd ed.).*

SCHECHTER, I., and BERGER, A. (1968). On the active site of proteases. III. Mapping the active site of papain; specific peptide inhibitors of papain. *Biochemical and Biophysical Research Communications*, *32*(5), 898-902.

SCHEIN, C. H. (1990). Solubility as a function of protein structure and solvent components. *Biotechnology, 8*(4), 308-317.

SCHELLENBERGER, V., BRAUNE, K., HOFMANN, H. J., and JAKUBKE, H. D. (1991). The specificity of chymotrypsin: a statistical analysis of hydrolysis data. *European Journal of Biochemistry*, *199*(3), 623-636.

SELLE, P. H., and LIU, S. Y. (2019). The relevance of starch and protein digestive dynamics in poultry. *The Journal of Applied Poultry Research, 28*(3), 531-545.

SINGH, M., KUMAR, V., SIKKA, K., THAKUR, R., HARIOUDH, M., MISHRA, D. P., ... and SIDDIQI, M. I. (2019). Computational design of biologically active anti-cancer peptides and their interactions with heterogeneous POPC/POPS lipid membrane. *Journal of Chemical Information and Modeling*, *60* (1), 332-341.

SPALVINS, K., ZIHARE, L., and BLUMBERGA, D. (2018). Single cell protein production from waste biomass: comparison of various industrial by-products. *Energy Procedia*, *147*, 409-418.

THERMO FISHER SCIENTIFIC. Overview of Protein Expression Systems. (n.d.). Retrieved from https://www.thermofisher.com/nl/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-protein-expression-systems.html

TREVINO, S. R., SCHOLTZ, J. M., and PACE, C. N. (2007). Amino acid contribution to protein solubility: Asp, Glu, and Ser contribute more favorably than the other hydrophilic amino acids in RNase Sa. *Journal of Molecular Biology, 366*(2), 449–460.

VALLEJO, L. F., and RINAS, U. (2004). Strategies for the recovery of active proteins through refolding of bacterial inclusion body proteins. *Microbial Cell Factories, 3*(1), 11.

VAN MILGEN, J., and DOURMAD, J. Y. (2015). Concept and application of ideal protein for pigs. *Journal of Animal Science and Biotechnology*, *6*(1), 15.

VEEROEDERBUREAU, C. (2008). Table booklet feeding of poultry: Feeding standards, feeding advices and nutritional values of feed ingredients. *CVB Series*, *45,* 11-12.

VILLAVERDE, A., and CARRIÓ, M. M. (2003). Protein aggregation in recombinant bacteria: biological role of inclusion bodies. *Biotechnology Letters, 25*(17), 1385-1395.

WANG, B., MIN, Z., YUAN, J., ZHANG, B., and GUO, Y. (2014). Effects of dietary tryptophan and stocking density on the performance, meat quality, and metabolic status of broilers. *Journal of Animal Science and Biotechnology, 5*(1), 44.

WANG, K., GAN, L., LEE, I., and HOOD, L. (1995). Isolation and characterization of the chicken trypsinogen gene family. *Biochemical Journal*, *307*(2), 471-479.

WATSON, J. D., LASKOWSKI, R. A., and THORNTON, J. M. (2005). Predicting protein function from sequence and structural data. *Current Opinion in Structural Biology, 15*(3), 275-284.

WECKE, C., PASTOR, A., and LIEBERT, F. (2016). Validation of the lysine requirement as reference amino acid for ideal in-feed amino acid ratios in modern fast growing meat-type chickens. *Open Journal of Animal Sciences*, *6*(3), 185-194.

WIEDERSTEIN, M., and SIPPL, M. J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research*, *35*(2), W407-W410.

WU, G. (2014). Dietary requirements of synthesizable amino acids by animals: a paradigm shift in protein nutrition. *Journal of Animal Science and Biotechnology*, *5*(1), 34.

YANG, J., and ZHANG, Y. (2015). Protein structure and function prediction using I-TASSER. *Current Protocols in Bioinformatics*, *52*(1), 5-8.

YANG, X., CHEN, K., LIU, H., ZHANG, Y., and LUO, Y. (2019). Purification and identification of peptides with high angiotensin-I converting enzyme (ACE) inhibitory activity from honeybee pupae (Apis mellifera) hydrolysates with in silico gastrointestinal digestion. *European Food Research and Technology*, *245*(3), 535-544.

YANG, Y., WANG, Z., WANG, R., SUI, X., QI, B., HAN, F., LI, Y., and JIANG, L. (2016). Secondary structure and subunit composition of soy protein in vitro digested by pepsin and its relation with digestibility. *BioMed Research International*, 2016.

YOUSEF, M., ABDELKADER, T., and EL-BAHNASY, K. (2019). Performance comparison of ab initio protein structure prediction methods. *Ain Shams Engineering Journal*, *10*(4), 713-719.

ZAHER, H. S., and GREEN, R. (2009). Quality control by the ribosome following peptide bond formation. *Nature, 457*(7226), 161-166.

ZELIKSON, R., EILAM-RUBIN, G., and KULKA, R. G. (1971). The chymotrypsinogens and procarboxypeptidases of chick pancreas. *Journal of Biological Chemistry*, *246*(19), 6115-6120.

ZHANG, C. T., and CHOU, K. C. (1992). Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. *Biophysical Journal, 63*(6), 1523–1529.

ZHANG, S., CUI, F. C., CAO, Y., and LI, Y. Q. (2016). Sequence identification, structure prediction and validation of tannase from *Aspergillusniger* N5-5. *Chinese Chemical Letters, 27*(7), 1087-1090.

ZHANG, W., YANG, J., HE, B., WALKER, S. E., ZHANG, H., GOVINDARAJOO, B., ... and ZHANG, Y. (2016). Integration of QUARK and I-TASSER for ab initio protein structure prediction in CASP11. *Proteins: Structure, Function, and Bioinformatics*, *84*, 76-86.

ZHANG, Y. (2007). Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins: Structure, Function, and Bioinformatics*, *69*(S8), 108-117.

ZHANG, Y., and SKOLNICK, J. (2004a). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, *57*(4), 702-710.

ZHANG, Y., and SKOLNICK, J. (2004b). SPICKER: a clustering approach to identify near-native protein folds. *Journal of Computational Chemistry*, *25*(6), 865-871.