Article **Journal-Agrarian and Natural Resource Economics**
June, 2019 Vol.3 No.4 11-18

Use and application of bioinformatics for the characterization of plant proteomes

Uso y aplicación de la bioinformática para la caracterización de proteomas vegetales

OSAWA-MARTÍNEZ, Eiko*†, MINJAREZ, Benito, MORALES-RIVERA, Moisés M. and MENA-MUNGUÍA, Salvador

Universidad de Guadalajara, Centro Universitario de Ciencias Biológicas y Agropecuarias (CUCBA), Doctorado BEMARENA Ciencias Agrícolas

ID 1st Author: *Eiko, Osawa-Martínez* / **ORC ID**: 0000-0001-7539-6044, **CVU CONACYT ID**: 724639

ID 1st Coauthor: *Benito, Minjarez* / **ORC ID**: 0000-0002-0974-4044, **CVU CONACYT ID**: 209055

ID 2nd Coauthor: *Moises M., Morales-Rivera* / **ORC ID**: 0000-0001-8579-0459, **CVU CONACYT ID**: 218482

ID 3rd Coauthor: *Salvador, Mena-Munguía* /**ORC ID**: 0000-0002-6423-4741, **CVU CONACYT ID**: 55746

DOI: 10.35429/JANRE.2019.4.3.11.18 Received March 21, 2019; Accepted June 30, 2019

Abstract

Proteomics and some other cutting-edge technologies have generated information clusters in sequencing and protein studies for plants, which can be used in other areas, such as food in quality control, pharmacological in allergens, characterizations of organisms in studies biological and agronomic for vegetables. The following is a description of the information that can be found in the databases (DB) and their interrelations with other specialized DB, of all the references to describe a protein. For this investigation we used a storage protein, Glutelin-2 in (*Zea mays*), we show some of the interrelated DB that can offer information for multiple studies of proteins in plants like UniProt KB and STRING-DB.

Protein, Glutelin-2, Proteomic, Data base, Description

Resumen

El uso de la proteómica y algunas otras tecnologías de punta como las plataformas bioinformáticas han generado cúmulos de información en secuenciación y estudios proteicos para diversos organismos vegetales. Cuyos resultados, pueden ser utilizados en otras áreas de investigación, como la alimenticia en el control de calidad, farmacológicas en los alergénicos, las caracterizaciones de organismos en estudios biológicos y agronómicos para vegetales. En el presente estudio describimos la información que se puede encontrar en las bases de datos (BD) públicas y sus interrelaciones a otras bases especializadas, de todas las referencias para describir una proteína y su relación con el proteoma general. Para lo cual, se utilizó una proteína de almacenamiento, llamada Glutelin-2 en granos de maíz (*Zea mays*), se muestran algunas de las BD interrelacionadas que pueden ofrecer información para múltiples estudios proteómico moleculares como UniProt KB y STRING-DB.

Proteína, Glutelin-2, Proteómica, Base de datos, Descripción

Citation: OSAWA-MARTÍNEZ, Eiko, MINJAREZ, Benito, MORALES-RIVERA, Moisés M. and MENA-MUNGUÍA, Salvador. Use and application of bioinformatics for the characterization of plant proteomes. Journal-Agrarian and Natural Resource Economics. 2019. 3-4: 11-18

* Correspondence to Author (email: eiko.osawa@cucba.udg.mx)
† Researcher contributing as first author.

Introduction

Food and its supply in many countries is complemented through imports, especially in the particular case of cereals; taking as an example the corn in Mexico, for the year 2016 it presented a deficit of 12.5 million tons (SAGARPA, 2016) which were imported, having as main origin the USA (González, 2018).

Another important aspect to consider in relation to food is that they regularly share the same geographical origin as the population that consumes them, which creates a greater root in the uses and customs of the food of the different agricultural products. Thus, the increase in the population and the industrialization of these products, increasingly generates an increase in cost and a decrease in their availability. Therefore, it is imperative to implement new techniques that promote greater performance in production rates, in addition to the need for an improvement in the quality of these, as well as ensuring that these foods are free of toxic substances such as metals heavy (Cd, Pb), which represent a high risk factor in the health of the population that consumes them (Chakrabarty et al., 2009, Cao et al., 2017).

In addition, it is worth highlighting the importance in the content of amino acids or the presence of traces of other foods such as flours (Colgrave et al., 2015) Thus, quality controls, health requirements, adequate control of Herbicide or heavy metal contamination indexes, or the identification of genetically modified organisms, are aspects that must be detected in a short time, with precision and efficiency, using a reduced sample of the product to be analyzed, in addition to using techniques that do not increase the final cost of the product (Arvanitoyannis and Vlachos, 2009).

For this purpose, it is that recently the latest technologies and highly precise tools have been implemented in the characterization of the different agricultural products such as proteomics and bioinformatics. Thus, proteomics is the technique that is responsible for the description of proteins expressed by a genome at a given time and conditions (Wilkins et al., 1996). Whose technique can be applied in the characterization of cereals and legumes, in different parts of the plant or under different planting and nutritional conditions.

In addition, that its use can be applied in the study of other foods such as honey, milk, flours or fruits, among other materials; Also, it is important to note that these techniques can be coupled to other tools such as liquid chromatography and mass spectrometry by increasing the identification of the protein content of small sample quantities (micrograms) with great precision (Wilkins et al., 1996).

Thus, proteomics based on bioinformatics and mass spectrometry (MS) are the methodologies most used in the identification, quantification and characterization of proteins and can be a useful tool when it comes to the study of different cereal varieties as for example, corn (Zea mays) (PPDB; http://ppdb.tc.cornell.edu (Sun et al. 2009; Chen et al. 2017), rice (Oryza sativa) http: //gene64.dna. affrc.go.jp/RPD/ (Komatsu et al. 2004) and wheat http://www.wheatgenome.org/ (Vu et al., 2017) In addition, these devices are supported and powered by different bases of international (BD) data; where, free of charge, the sequences of the peptides and the name of the protein detected in the analysis and characterization can be located according to their function and / or class even when they come from complex samples such as total or simpler homogenized from a band extracted by conve electrophoresis national, where the results obtained are characterized by being highly accurate, reliable and in a very short time.

As described above, we propose the use of these techniques for the generation of new methodologies in the generation of data with potential in the identification of important properties for quality controls and the selection of a material for food, industrialization or other potential uses such as medical and / or pharmacological. Therefore, the present work has the purpose of showing part of the information contained in the BDs for proteins within plant organisms and their application in the In silico characterization of said crops.

Justification

The molecular proteomic characterization of organisms or materials in a precise way, with high levels of confidence, quickly and with a low amount of sample, is an attractive methodology and an important source of information in the identification of the quality of a food such as cereals , oils or flours, as well as cultivars that are imported for food or industry.

ISSN 2524-2091

Also, it is important to indicate that the proteomic characteristics in the case of study materials allow the generation of new sources of important information for a more detailed and detailed description, in addition to conducting interdisciplinary studies that focus on analyzing different aspects of the organism of interest as well as the characterization and genomic and proteomic behavior of the plant under very specific conditions.

Therefore, the use of internationally available BD can contribute to the generation of new knowledge and information necessary for the different objectives in an investigation, in relation to protein and nutritional content.

Problem

The use of BD and bioinformatics platforms in the description of the quality and protein content of a crop quickly, accurately and effectively are key properties for the characterization and selection of the main products for human, agricultural or industrial consumption. Therefore, the consultation of information available in the BD, can contribute to complement important properties for decision making in the characterization or destination of the product or object analyzed.

Hypothesis

The use of the information obtained from the BD and its analysis through bioinformatics platforms contributes to the description of a protein or the complete proteome of a culture**.**

Objectives

Identify the potential of bioinformatics tools and BD in the in silico description of a protein and its relationship with the rest of the proteome in corn grains.

General objective

Contribute to the description of the corn proteome.

Specific objectives

Identify and describe the information necessary for a protein through the BD linked to bioinformatics platforms in proteomic analyzes.

Theoretical framework

In almost all the world the lack or low availability of food is complemented through imports which increases the final cost of them. For example, in Mexico for the year 2016, 12.5 million tons of corn were imported for national consumption only (Agri-food Outlook for corn 2016). This is even more complex when we highlight that cereals such as wheat, corn, sorghum or rice; They are the largest source of protein consumed primarily by the population of developing and emerging countries. These crops are also deficient in some essential amino acids, such as lysines and tryptophan, which negatively impacts the development and nutrition of the population, especially people with low incomes and vulnerable populations such as children, older adults and pregnant women (FAOSTAT, 2014).

In other contexts, the need for strict quality controls for the various foods offered in the market and the identification of those products with high potential in the diversification of their uses, is another aspect of interest such as the capacity of pharmacological uses, industrial, alternative energy sources or the possibility of being classified and labeling them as possible risk factors and / or allergens, is attractive and necessary for a better offer to the final consumer.

Thus, proteomic analyzes can show the characteristics of these products and raise the processes of quality control and selection of the different agricultural products for human, animal and industrial consumption (Carrera et al., 2017). In addition, the characterization of food through proteomics based on mass spectrometry represents the possibility of an alternative tool in the typing of food accurately and reliably. What would have a favorable impact on food biosecurity and compliance with international safety standards. (Korte and Brockmeyer, 2017; Ortea et al., 2016).

On the other hand, the biological BDs are the repositories of all the information of the genomes and proteins investigated by a large number of researchers and institutions both public and private around the world. In addition, the BDs are integrated by a set of data belonging to the same context, whose objective is to organize the information, structuring it in specific registers and where many of them are available for free to the general public.

ISSN 2524-2091

It should also be noted that, each record is composed of specific fields for each peptide, lipid, chemical nucleotide sequence, etc. Where, you have to specify a piece of information called value (in the case of proteins the registration key, or compound name). In order to access the data, programs called BD management systems are developed that make it easier to consult the information contained, which allows a report to be generated through graphics maintaining the integrity of the data. (Rodríguez, 2013). Thus, the biological BD can be focused on flat, animals, viruses, fungi or any other organism (presented as a list of text-type information), but have relational fields, (Leagues) that connect with other databases thus concentrating , more information about the search (Uniprot KB Consortium, 2014).

For the particular case of proteins, which are the central biomolecules responsible for all cellular functions in the living organism, they represent the central focus of proteomic studies, further expanding the panorama of the metabolic, physiological and / or pathological processes of an organism. at a given time or circumstances (Chandramouli and Qian 2009). Thus, proteomics is becoming an indispensable tool for global phenotypic characterization at the molecular level that provides information on the identification of genes and the role of dynamic post-translational modifications (PTM) and protein interactions to other biomolecules, linking the genotype and its metabolisms with the phenotype and functionality (Hu et al ,. 2015).

In addition, different tools have emerged exponentially increasing peptide analyzes thus strengthening the information available in the BDs; One such technique is tandem mass spectrometry (MS / MS). Whose information generated, can be analyzed by the use of various BDs both for complex protein sequences and their post-translational modifications, or be subjected to homology and function analysis through different software. Although, it is necessary to indicate that sometimes the results obtained may present some ambiguity and this may be related to different factors, such as the methodology used in the realization of the project or the incompatibility of the reagents for its analysis, which generates false positives or the loss of the analyzed sample.

Research Methodology

The mapping for the identification of proteomes of specific organisms by mass spectrometry is based on several BD characterizing the different spectrograms obtained in relation to the mass and load of peptide sequences contained in the sample (Abián et al., 2008). In addition, the development of proteomic meta-analyzes represent an invaluable generation of knowledge of the proteome of an organism at specific times and circumstances, which in turn provides a broader picture of the physiological and / or pathological processes of said experimental analyzes and their impact on the metabolism of the organism and finally the impact on the observed phenotype. These studies represent the enrichment of the biological BD and strengthen, in turn, the information contained in the different bioinformatics platforms, thus allowing to administer and take advantage of all this information more efficiently.

It should be noted that, this work is part of the research project: Typing of MR 2008 corn and its progenitors of the BEMARENA Postgraduate Program in Agricultural Sciences of the University of Guadalajara of CUCBA.

For which, the Glutelin-2 corn endosperm storage protein was randomly selected. For this selection, the use of the UniProt KB protein BD was necessary. Identifying proteins characterized in previous works carried out by various researchers and contained in said database, to verify their presence in the plant. A list of proteins was displayed, selecting it based on the coincidences with the name of the protein and the plant species. So this Glutelin-2 protein was chosen with the access code P04706**.**

Materials and methods:

For the present study, two databases were used mainly; which are online and are freely accessible. In the case of direct reference generation analysis and general corn information analysis, we used the NCBI BD (National Center for Biotechnology Information). For the specific case of the information related to protein level we work with the UniProt KB BD using the access code P04706.

ISSN 2524-2091

Kind of investigation

Descriptive applied research, in order to identify the sources to obtain information in the description of a protein belonging to a plant.

- The need to characterize organism and its protein content, to allocate them to different uses.
- The identification of food quality or its origin.
- Obtain sufficient and efficient information on the characterization of a protein.
- Identify the scope of the information obtained in the BD interconnected to MS.

Results

For this work, the registry for Zea mays Gluterlin-2 was selected, with the UniProt access code KB P04706. From the access code it is possible to access the information contained in the databases and thus know the characteristics of the protein, but it is also possible to do so by means of the exact name of the protein or the name of the gene.

Once on the page and with any of the protein data of interest mentioned above, the search is done in the UniProt KB database, which allows us to access the most updated and reliable information worldwide, as we It provides not only information of the protein of interest but also offers us information contained in other platforms with possible relevance for our investigations;

Among which we highlight the interaction with other proteins, biochemical data such as its isoelectric point, (pI) its molecular weight, its secondary structure, third-dimensional modeling (3D) homology with other proteins, post-translational modifications; In addition to the most relevant works for physiological and metabolic issues within the organisms to study. For this investigation we find the following information:

The protein name for the access code P04706 for corn is Gluterlin-2, where until now, the gene has not been described (N / A), and its subcellular location is the vacuole, (Figure 1A).

Its function is also described which classifies it as a seed storage protein and with a molecular function of nutrient reservoir. It also gives us other alternative names used to identify it as 27 kDa zein or alcohol soluble reduced glutelin. On the other hand, it is possible to know them different post-translational modifications characterized until the moment of the consultation and their implication in the final corn phenotype. In addition, it shows specific information on protein domains highlighting the peptide sequence, location within the protein and function of each domain.

On the other hand, through the STRING platform we can know the functional protein association networks. Where the different labels (Figure 1A, colored lines) of the diagram, redirect us to the characteristics of each of them in this platform https://string-db.org/. For our case study, we see in Figure 1A a diagram interconnecting between the proteins most strongly associated by previous work on our proteins of interest by means of lines of different colors. Each line has a specific meaning of interaction, for example, the union of two proteins by a yellow line indicates that both proteins have been mentioned in an article already published.

Thus, in our case, Gluterlin-2 has been mentioned with Zeina beta in at least two scientific articles such as the one carried out by Yao et al., In 2016, where they indicate that both proteins are essential for morphology in the corn endosperm. It also allows us to identify which are the biological and molecular processes where our protein of interest has been reported, so we can see that in the case of the interaction network shown in Figure 2B the main molecular function is that of nutrient reservoir where five Of the 18 proteins present in the diagram participate in this function and gives us a statistical value of 7.18e-13. In addition, it is possible to know the secondary structure of the proteins contained in the interactome, which can show us the specific domains for each protein in order to identify their functions. Returning to the UniProt KB platform, in the sequence record we find, cross references, in addition to links to other related bases, where we can find the pI: 8.40, average mass 23 689 Da, (https: //web.expasy. org).

ISSN 2524-2091

For this protein that belongs to a plant a specific database for the organism in this case for the corn grain (Zea mays) (https://www.maizegdb.org), 3D protein modeling (Figure 2 A), (https://www.proteinmodelportal.org), allergenic studies (http://www.allergome.org).

Protein-related articles, in the field of Display Publications in the UniProt KB registry, a text mining system with additional bibliography for more proteins, eGenPub and protein alignment to indicate homology with other proteins, https: // npsa- prabi.ibcp.fr/. (Figure 3).

This amount of information shows us how we can identify important data that help us clarify or expand the relationship of the protein under study with the biological, physiological or molecular functions within the plant, thus giving an idea of its manifestation at the phenotypic level within the species under study.

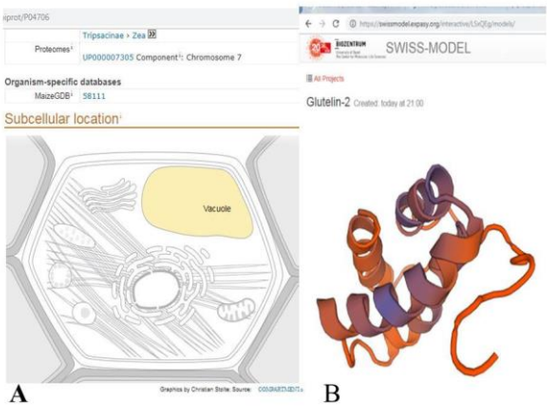

Figure 1 UniProtKB, protein information under study. Registration of the Glutelin-2 protein in UniProtKB P04706, showing the subcellular location (A) and the 3D structure (B)

The information or fields of each record may vary according to the protein investigated, for this case the Glutelin-2 (Zea mays) record appears with experimental evidence, which provides more information in the fields of this protein's record..

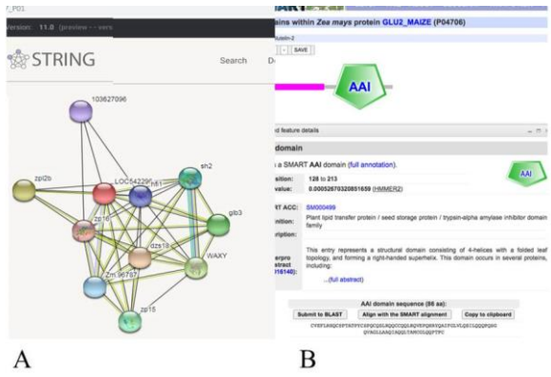

Figure 2 (A) Databases interrelated with the *Zea mays* Glutelin-2 protein registry (STRING) Domains (B) SMART

Conclusions

The databases provide relevant and in many cases updated and specific information for vegetable proteins, it is important to interact with other databases, to provide data that can help identify important characteristics that contribute to the study in different areas, as well as specific of a vegetable protein.

The identification of important characteristics such as its pI, its molecular weight, the interaction with other proteins, its secondary structure, shows us a vision of the properties of the vegetable protein under study, an overview of the physiological, pathological and molecular processes that can contribute in the development of interaction studies as in the proteomic meta-analysis of plant species; In addition, the use of tools such as mass spectrometry that allows to identify with good precision the presence of proteins from other species present in the sample, can contribute to food quality control, under specific circumstances or at a given time or place in the development of crop, also contributing to clarify part of the phenotype observed in a vegetable.

We must also consider that it is only an approach to accurate information, but it does not stop being valuable for the studies that are going to be carried out with vegetable proteins in any of the types of analysis. Therefore, it is important to keep the information on protein recognition up to date and homogenize criteria when naming proteins for all bases and keep the vectors among them updated. In this way complement the studies and research not only in the biological and agronomic area but also in the quality control of food and pharmacological or medical areas.

ISSN 2524-2091

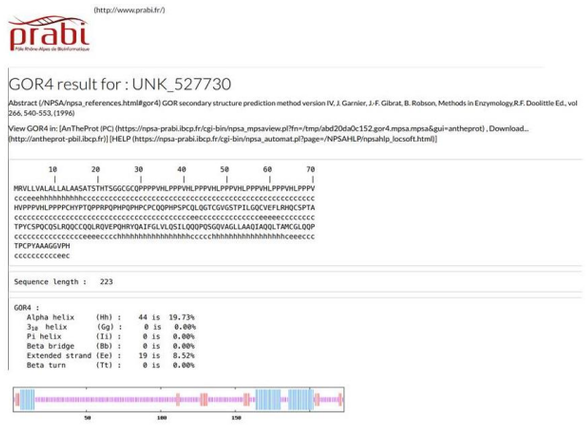

Figure 3 Protein alignment to indicate some functional relationship

References

Arvanitoyannis I. S. y Vlachos A. (2009) *Maize Authentication: Quality Control Methods and Multivariate Analysis (Chemometrics)*, Critical Reviews in Food Science and Nutrition, 49:6, 501-537, DOI: 10.1080/10408390802068140.

Cao, Z., Mou, R., Cao, Z., Lin, X., Ma, Y., Zhu, Z., & Chen, M., (2017), *Quantitation of glutathione S-transferases in rice (Oryza sativa L.) roots exposed to cadmium by liquid chromatography-tandem mass spectrometry using isotope-labeled wing peptides as an internal standard.* Plant methods, 13, 64. doi:10.1186/s13007-017-0214-2.

Carrera M., Cañas B., Gallardo J. M., (2017), *Advanced proteomics and systems biology applied to study food allergy*, Elsevier, Current opinion in food science. 22:9-16. https://doi.org/10.1016/j.cofs.2017.12.001.

Chakrabarty D, Trivedi PK, Misra P, Tiwari M, Shri M, Shukla D, et al., (2009), *Comparative transcriptome analysis of arsenate and arsenite stresses in rice seedlings*. Quemosfera ; 74 : 688–702. doi: 10.1016 / j.chemosphere.2008.09.082.

Chandramouli K. & Qian P. Y. (2009), *Proteomics: challenges, techniques and possibilities to overcome biological sample complexity*. Human genomics and proteomics: HGP, 2009, 239204. doi:10.4061/2009/239204.

Chen, C., Huang, H., & Wu, C. H., (2017), *Protein Bioinformatics Databases and Resources.* Methods in molecular biology, (Clifton, N.J.), 1558, 3-39.

Colgrave M., Goswami h., Byrne K., Blundell M., Howitt C., and TannerG.,(2015*), Proteomic Profiling of 16 Cereal Grains and the Application of Targeted Proteomics To Detect Wheat Contamination*, J. Proteome Res., 2015, 14 (6), pp 2659–2668 DOI: 10.1021/acs.jproteome.5b00187.

Combet C., Blanchet C., Geourjon C., and Deleage G., (2000*) Network Protein Sequence Analysis*, 25(3)291:147-150. https://www.gob.mx/cms/uploads/attachment/file/200637/Panorama_Agroalimentario_Ma_z_2016.pdf Consultado mar. 2018.

Hu J, Rampitsch Ch. And Bykova N.V. (2015), *Advances in plant proteomics toward improvement of crop productivity and stress resistance.* Front. PlantSci.6:209. doi: 10.3389/fpls.2015.00209.

Komatsu, S., Kojima, K., Suzuki, K., Ozaki, K., & Higo, K., (2004), *Rice Proteome Database based on two-dimensional polyacrylamide gel electrophoresis: its status in 2003*. Nucleic acids research, 32(Database issue), D388-92.

Korte R., and Brockmeyer J., (2017), *Novel mass spectrometry approaches in food proteomics*, Elsevier 96:99-106,https://doi.org/10.1016/j.trac.2017.07.010.
La jornada, Economía. Gónzalez S. https://www.jornada.com.mx/ultimas/2018/06/21/incrementan-importaciones-de-maiz-en-mexico-4833.html

Ortea I., O´connor G., Maquet A., (2016), *Review on proteomics for food authentication*, Elsevier, 147:212-225, http://dx.doi.org/10.1016/j.jprot.2016.06.033

SAGARPA (2016), *Panorama agroalimentario maíz 2016, Dirección de investigación y evaluación económica, sectorial*.

Sun, Q., Zybailov, B., Majeran, W., Friso, G., Olinares, P. D., & van Wijk, K. J. (2008). PPDB, *The Plant Proteomics Database at Cornell*. Nucleic acids research, 37(Database issue), D969-74.

The UniProt Consortium (2016). UniProt: the universal protein knowledgebase. *Nucleic acids research*, *45*(D1), D158-D169.

UniProt Consortium (2014), *UniProt: a hub for protein information.* Nucleic acids research, 43 (Database issue), D204-12.

Wilkins M.R., Sanchez J.C., Gooley A.A., Appel R.D., Humphery-Smith I., Hochstrasser D.F., Williams K.L.,(1996), *Progreso con proyectos de proteoma: por qué todas las proteínas expresadas por un genoma deben identificarse y cómo hacerlo*. Biotechnol Genet Eng Rev. 13 : 19–50.

Yao D, Qi W, Li X, Yang Q, Yan S, Ling H, Wang G, Wang G, Song R. Maize opaque10 Encodes a Cereal-Specific Protein That Is Essential for the Proper Distribution of Zeins in Endosperm Protein Bodies. PLoS Genet. 2016 Aug 19;12(8).

Acknowledgments

We want to thank CONACYT for the doctoral grant 724639/634015 by Estela Eiko Osawa Martínez.

Annex: URL of the bases

STRING http://string-db.org,

UNIPROT http://www.uniprot.org

SWISSPROT http://web.expasy.org/docs/swiss-prot_guideline.html,

SWISS-MODEL
http://swissmodel.expasy.org/interactive,

MAÍZ MAIZEDB
https://www.maizegdb.org/data_center/gene_product?id=58111
Identificación de dominios.
http://smart.embl-heidelberg.de/smart/set_mode.cgi?NORMAL=1

Alineamiento proteínas
https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_multalin.html