pathogens

MDPI

Article

ProtozoaDB 2.0: A *Trypanosoma Brucei* Case Study

Rodrigo Jardim [1], Diogo Tschoeke [2,3] and Alberto M. R. Dávila [1,*]

[1] Computational and Systems Biology Laboratory, Oswaldo Cruz Institute, Fiocruz, Rio de Janeiro 21040-900, Brazil; rodrigo_jardim@fiocruz.br

[2] Microbiology Laboratory, Rio de Janeiro Federal University, Rio de Janeiro 21941-901, Brazil; diogoat@gmail.com

[3] Nucleus in Ecology and Socio-Environmental Development of Macaé (NUPEM), Rio de Janeiro Federal University, Macaé, Rio de Janeiro 21941-901, Brazil

* Correspondence: davila@fiocruz.br; Tel.: +55-21-3865-8132

Received: 20 June 2017; Accepted: 16 July 2017; Published: 20 July 2017

Abstract: Over the last decade new species of Protozoa have been sequenced and deposited in GenBank. Analyzing large amounts of genomic data, especially using Next Generation Sequencing (NGS), is not a trivial task, considering that researchers used to deal or focus their studies on few genes or gene families or even small genomes. To facilitate the information extraction process from genomic data, we developed a database system called ProtozoaDB that included five genomes of Protozoa in its first version. In the present study, we present a new version of ProtozoaDB called ProtozoaDB 2.0, now with the genomes of 22 pathogenic Protozoa. The system has been fully remodeled to allow for new tools and a more expanded view of data, and now includes a number of analyses such as: (i) similarities with other databases (model organisms, the Conserved Domains Database, and the Protein Data Bank); (ii) visualization of KEGG metabolic pathways; (iii) the protein structure from PDB; (iv) homology inferences; (v) the search for related publications in PubMed; (vi) superfamily classification; and (vii) phenotype inferences based on comparisons with model organisms. ProtozoaDB 2.0 supports RESTful Web Services to make data access easier. Those services were written in Ruby language using Ruby on Rails (RoR). This new version also allows a more detailed analysis of the object of study, as well as expanding the number of genomes and proteomes available to the scientific community. In our case study, a group of prenyltransferase proteinsalready described in the literature was found to be a good drug target for Trypanosomatids.

Keywords: protozoa; information extraction; *Trypanosoma brucei*; trypanosomatids; ProtozoaDB

1. Introduction

Over the last decade new species of Protozoa were sequenced and deposited in GenBank [1–4]. The availability of the primary genome sequence is a good starting point for the community to contribute further analyses (e.g., identification and functional annotation of coding sequences as well as comparative genomics analysis) in order to infer new information on the biology of these organisms. Analyzing large amounts of data generated by genomics experiments, especially using Next Generation Sequencing (NGS), is not a trivial task. The ongoing NGS technology makes the sequencing of more and more eukaryote genomes a reality, giving rise to new paradigms (either for the development and improvement of semi-automatic analysis/annotation systems for this huge amount of data, or for an object-view concept where raw reads are the main, fixed object, and assemblies with their annotations take a role of dynamically changing and modifying views of the object [5]).

The processes involved in the sequencing and preparation of genomic information can be represented in a similar way as the life cycle of software (Figure 1). The first step is data acquisition that can be performed by: (i) downloading from public databases; and (ii) sequencing across multiple

platforms, like Sanger and/or NGS (Illumina, Ion Torrent, Nanopore and/or Pacific BioScience). The second step, called pre-processing, formats and stores genomic data for subsequent use. The third step refers to the use of a number of computational tools to transform raw data into knowledge. The fourth and last step is distributing and making this information available to the community for further analysis and inferences.

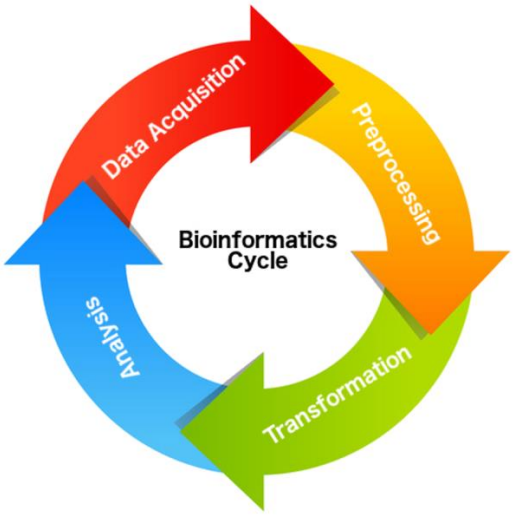

Figure 1. Example for data lifecycle in bioinformatics. The lifecycle begins with data acquisition, through data pre-processing, data transformation, and, finally, the analysis of the results (information) generated by this process.

Therefore, in order to facilitate information extraction [6], we developed the ProtozoaDB [7] database system, which in its first version included five protozoan genomes (*Entamoeba histolytica*, *Leishmania major*, *Plasmodium falciparum*, *Trypanosoma cruzi*, and *T. brucei*) and a set of tools for searching and analyzing data, including phylogeny inference. In the present study we present a new version of ProtozoaDB called ProtozoaDB 2.0 (http://protozoadb.biowebdb.org) that, according to the above description, fits into the third and final steps of the bioinformatics cycle: transforming raw data into information followed by distribution and availability. The development of new generation databases as ProtozoaDB is being encouraged by the community, especially in the context of the BioCreative initiative [8] and reviewed by Krallinger et al. (2008) [9].

The system has been fully remodeled to allow for new tools and a more expanded view of data, using advanced computational techniques and providing a wider range of information for users. Now with the genomes of 22 pathogenic Protozoa, this new version includes analyses such as: (i) similarities with other databases (*Homo sapiens*, model organisms, Conserved Domains Database—CDD and Protein Data Bank—PDB); (ii) visualization of the metabolic pathways of Kyoto Encyclopedia of Genes and Genomes—KEGG [10]; (iii) protein structures by PDB [11]; (iv) homology studies, using results from OrthoMCL [12], KEGG Orthology (KO), and OrthoSearch [13]; (v) the search for related publications at PubMed; (vi) superfamily classification [14]; and (vii) phenotype inferences based on comparisons with model organisms, particularly with *Saccharomyces cerevisae*.

ProtozoaDB source code was completely rewritten in another programming language and with more elaborated techniques. It now uses a framework for developing Web applications known as Rails (http://rubyonrails.org/). It was developed in layers, allowing for the separation of the business object code of the pages displayed to users, making maintenance easier and consequently access to its pages lighter and faster. Furthermore, there is a specific layer to deal with data to be fetched from other sources. The Ruby language, suitable for the use of Rails, was adopted for this version together with BioRuby library [15], enabling the development of pages with less code and better reuse of functions. ProtozoaDB 2.0 was also implemented using concepts of Object Orientation and

Design Patterns. This made the application lighter, safer, and simpler to maintain. The use of the JQuery library made possible for the web pages to work with Asynchronous Javascript and XML (AJAX), creating a friendlier user interface. Now it is possible to view all the information provided by ProtozoaDB 2.0 on one page. The new system uses the concept of Web Services to access all internal and external databases. Thus, the application focuses only on usability and user-friendly information. All databases are queried simultaneously allowing a response time considered to be satisfactory for the application. ProtozoaDB 2.0 allows queries by several methods, including: Genbank Identifier (GI), Accession Number, Description, Blast, Motif, and Phenotype. With all information in one place, it is now possible to infer information on the biology and biological systems of the protozoan species studied. Additionally, ProtozoaDB 2.0 now has information inferred from phenotypes. Orthology analysis helps to transfer phenotype information based on genotypes. According to [16] it is also possible to transfer functional information based on similar phenotypes and a specialized database called PhenomicDB was developed using this concept [17].

2. Results

2.1. Protozoa Genomic Data

ProtozoaDB 2.0 provides descriptive, quantitative, qualitative, and comparative information on the genomes and proteins of 22 protozoan species (Table 1), thus allowing a more detailed analysis of each organism including the inference of relationships between them. The new version contains: (i) 193,559 genes; (ii) 218,100 proteins; (iii) 26,101 homologous groups (21,119 orthologous groups and 4982 paralogous groups) obtained by OrthoMCL analysis (Figure 2); and (iv) 195 phenotypes inferred by crossing information with the *Saccharomyces* Database.

Table 1. List of organism species loaded in ProtozoaDB 2.0.

Organism/Strain
Babesia bovis T2Bo
Crypstosporidium parvum Iowa II
Cryptosporidium hominis TU502
Cryptosporidium muris RN66
Entamoeba dispar SAW760
Entamoeba histolytica HM1:IMSS
Giardia lamblia ATCC 50803
Leishmania braziliensis MHOM BR 75 M2904
Leishmania infantum JPCM5
Leishmania major Friedlin
Plasmodium berghei ANKA
Plasmodium chabaudi chabaudi AS
Plasmodium falciparum 3D7
Plasmodium knowlesi strain H
Plasmodium vivax SaI 1
Plasmodium yoelii yoelii 17XNL
Theileria annulata Ankara
Theileria parva Muguga
Toxoplasma gondii ME49
Trichomonas vaginalis G3
Trypanosoma brucei treu927
Trypanosoma cruzi CL Brener

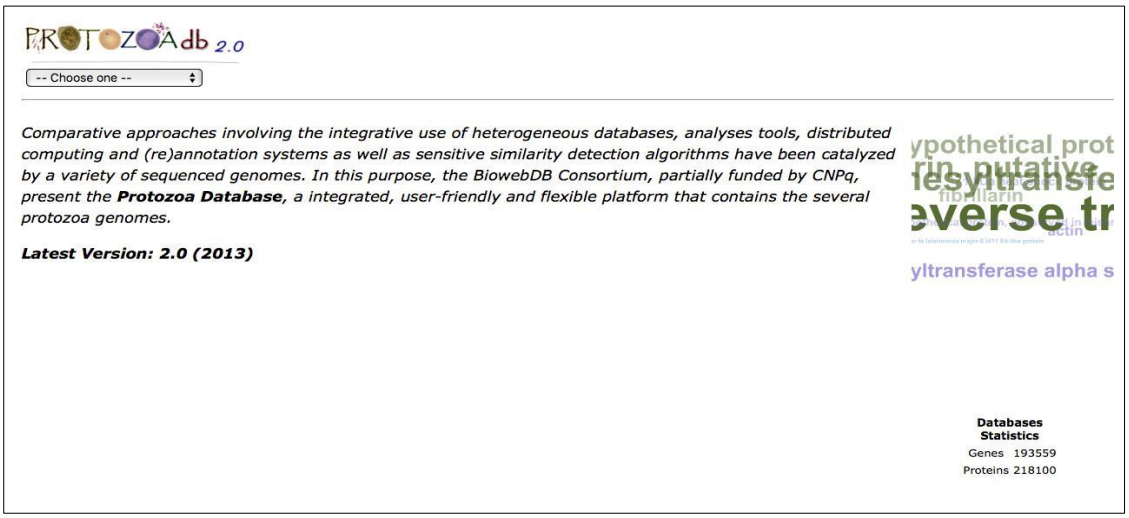

Figure 2. The front page of ProtozoaDB 2.0 displaying database statistics, the search field, and the tag's cloud.

2.2. Proteome

The information about the proteins of 22 different Protozoa is complemented by the results obtained by real-time queries, performed in several remote databases through the use of Web Services. Two similarity analyses are performed using BLAST [18] and FASTA [19] against PDB [11]. The FASTA similarity results facilitate a visual comparison of the protein 3D structures, while the BLAST results also allow users to select any or all hits, as well as to retrieve and export their sequence in FASTA format (Figure 3). Conserved domains analyses use CDD [20]. As a plus, a similarity analysis against the human proteome is also performed. All this information is displayed showing the top 10 results (Figure 4).

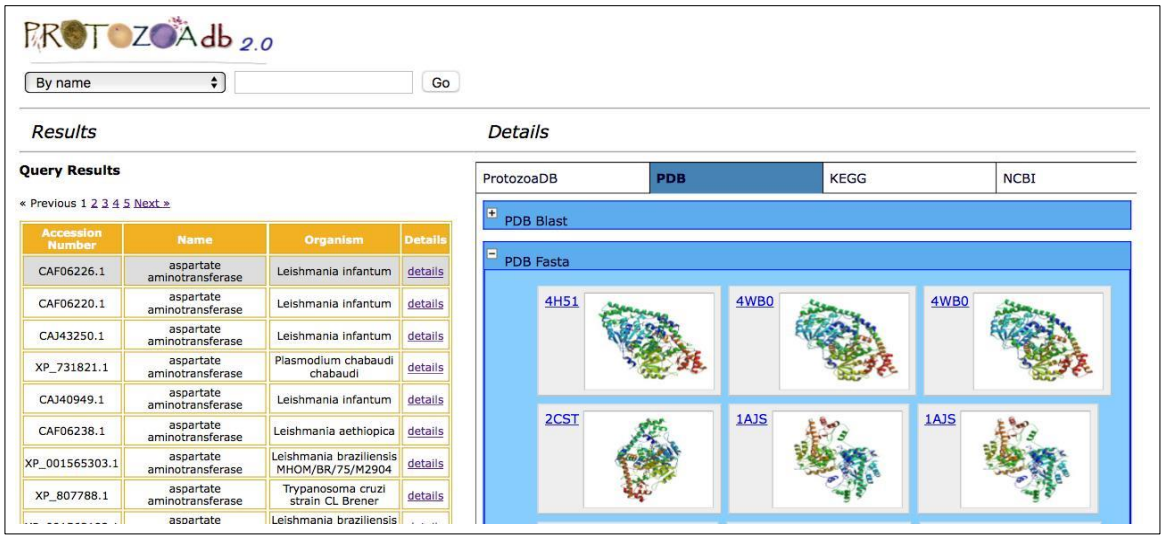

Figure 3. The similarity results against PDB: 2D (Blast) and 3D (Fasta). The figure shows sequence similarities displayed in 3D with the aspartate aminotransferase of *L. major*. Clicking on each figure of the system shows the complete information in the remote website.

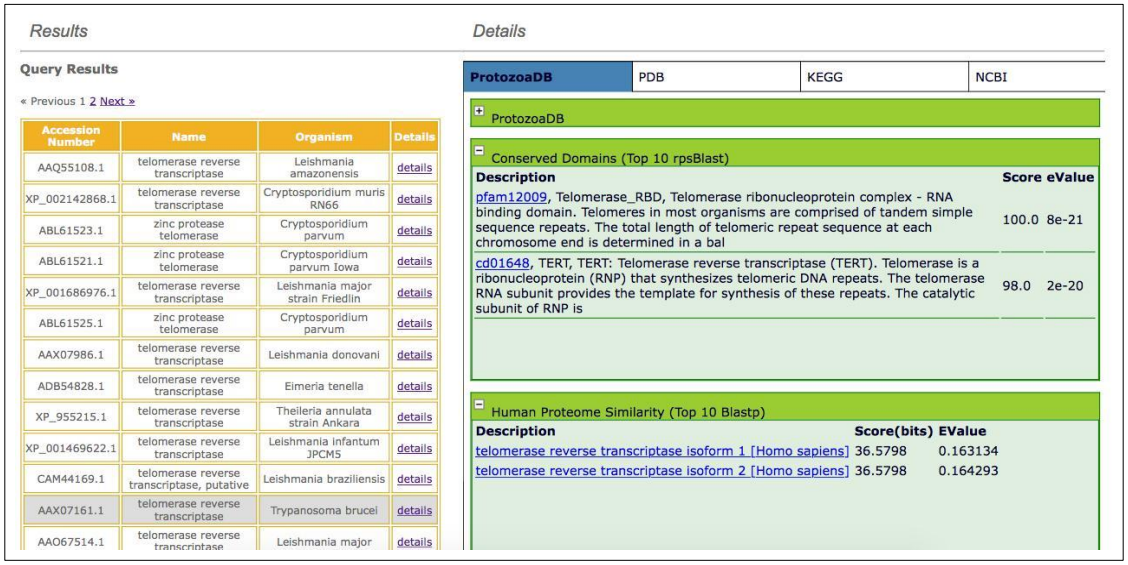

Figure 4. The results for similarity searches against *Homo sapiens* proteome and the Conserved Domains Database. Only the top ten results are shown. Clicking on links in blue opens a new window in the remote website.

2.3. Homology

The results of the preliminary analyses of homology among the 22 Protozoa are available for queries. The orthologous groups were inferred by the methodology implemented in OrthoMCL (Figure 5) and OrthoSearch, using either a Blast-based or Hmmer-based algorithm, respectively.

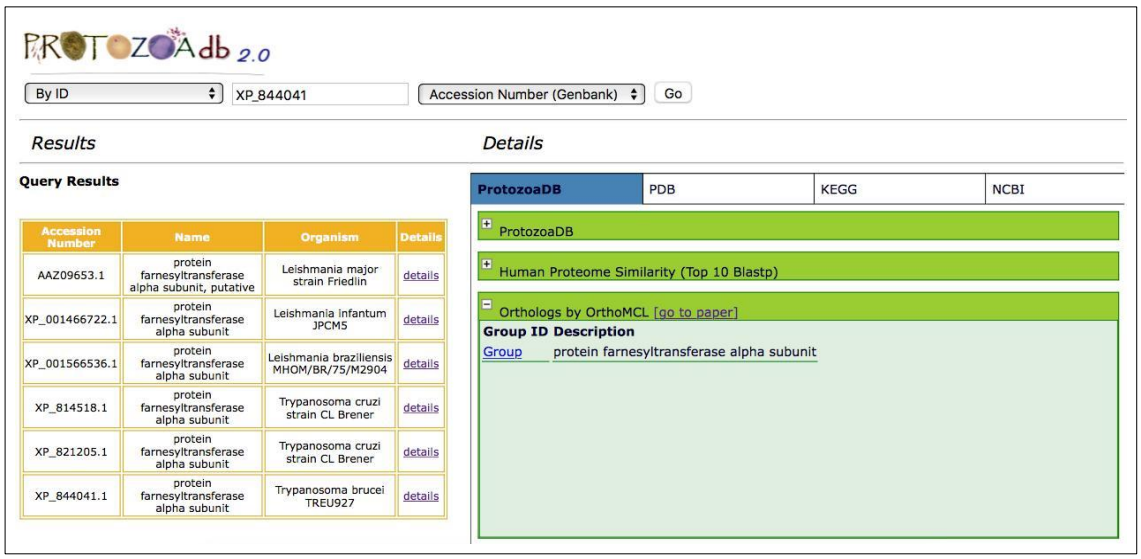

Figure 5. Orthologous groups inferred using OrthoMCL methodology. Clicking the "Group" link shows all proteins of that group that are shown in the left panel (Query Results).

2.4. Metabolic Pathways

The system performs a web service-based query to retrieve metabolic maps available on KEGG, showing the involvement of a given protein in that pathway (Figure 6).

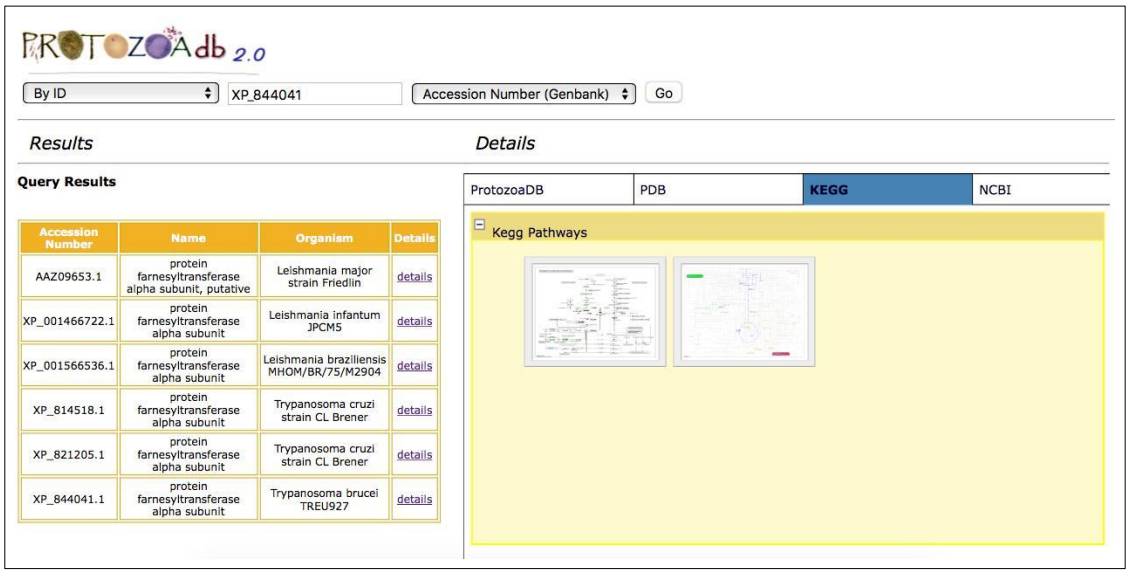

Figure 6. Metabolic pathways from KEGG. The figure shows all metabolic pathways that include aspartate aminotrasnferase. Clicking on a map opens a new window in a remote web site (KEGG).

2.5. Phenotypes

ProtozoaDB 2.0 allows web service-based queries through the phenotypes mapped from the *Saccharomyces* Database [21], retrieving proteins from the 22 Protozoa that could potentially provide such features. This information was made possible by mapping the proteome of the 22 species with information from the KEGG orthologous groups (Kegg Orthology—KO) as part of the "transformation" step described in the introduction (Figure 7).

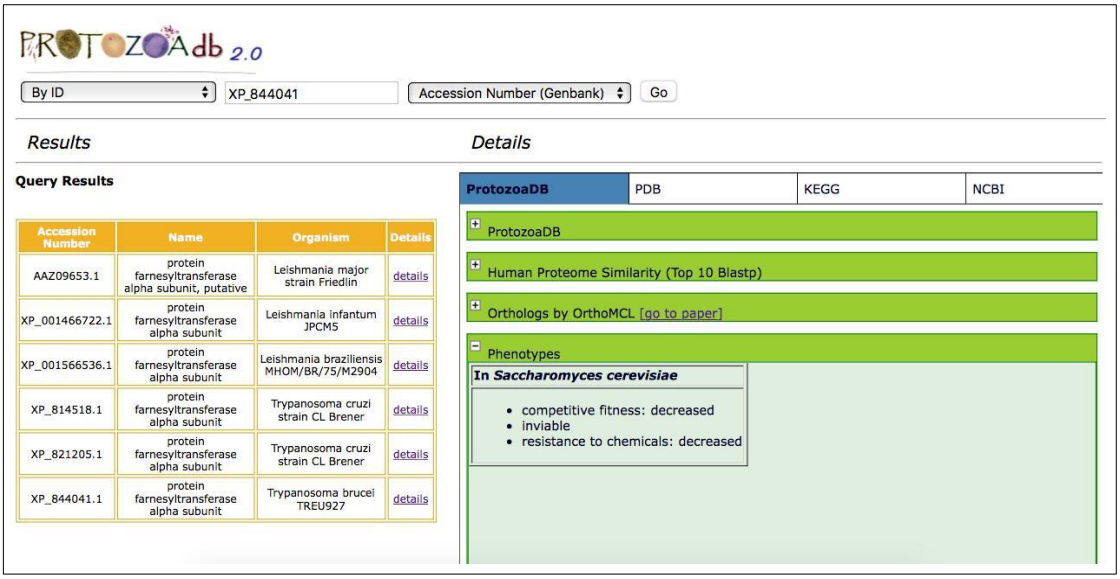

Figure 7. Phenotypes found by orthology with *Saccharomyces cerevisiae* for farnesyltransferase alpha protein subunit.

2.6. How to Search

The new system retrieves the information through various search engines. Based on the previous version, the system searches for the description of the protein or part of the description, Accession

Number, Genbank Identifier (GI), and organism name. In addition to these mechanisms, this new version also allows query by phenotype and similarity (Blast).

2.7. How to Search Using Our Web Service

In addition we also made a set of web service functions available to retrieve all information available in our system. The page http://services.biowebdb.org/howtouse contains the information about how to use available services including source code examples. Functions to search Protozoa proteins by Accession Number, Genbank Identifier, description (annotation), organism, phenotype, and Blast, as well as details of protein analyses like orthologous groups, similarity results, KEGG pathways, and phenotypes, are available for queries with our web services.

2.8. Information Extraction—T. brucei Case Study

To demonstrate the usefulness of ProtozoaDB 2.0 for information extraction, a case study was conducted using phenotypes in the Kinetoplastea species. Through the search field system the option Phenotypes was chosen and the keyword "inviable" used with the Kinetoplastea subset database. This phenotype may indicate (depending on the experiment) a situation of impossibility for the survival of the organism [21]. Based on orthology with *Saccharomyces cerivisae*, the system returns a list containing a wide range of proteins that show this phenotype. From the obtained list, the first hit meeting the following requirements was chosen: (i) low similarity with the human proteome; (ii) high similarity with the bacterial species; and (iii) a pathway available in KEGG. The chosen hit was XP_844041.1 protein farnesyltransferase (PFT) alpha subunit from *Trypanosoma brucei*, because of the high similarity to the bacterial prenyltransferase group (Figure 8).

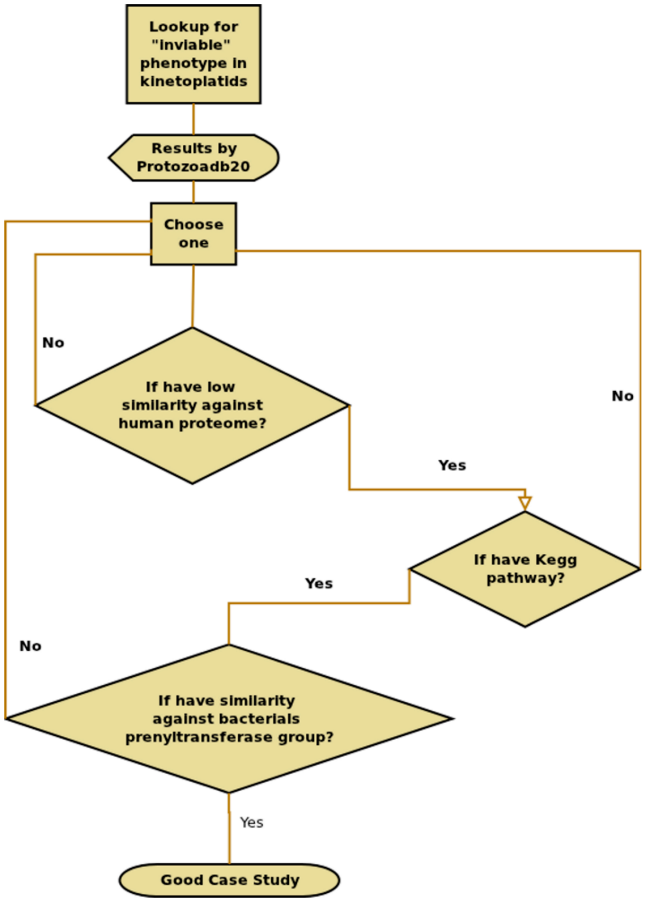

Figure 8. Flow chart showing the options and choices for the identification of a good case study.

Farnesyltransferase alpha subunit is a protein of the prenyltransferase group [22]. Pfam Farnesyltransferase and geranylgeranyltransferase are classified in the same family because of the CaaX motif present in both of them [23]. Figure 9 shows the PPTA family with 795 species of which 23% (223/795) are Metazoa that share this motif.

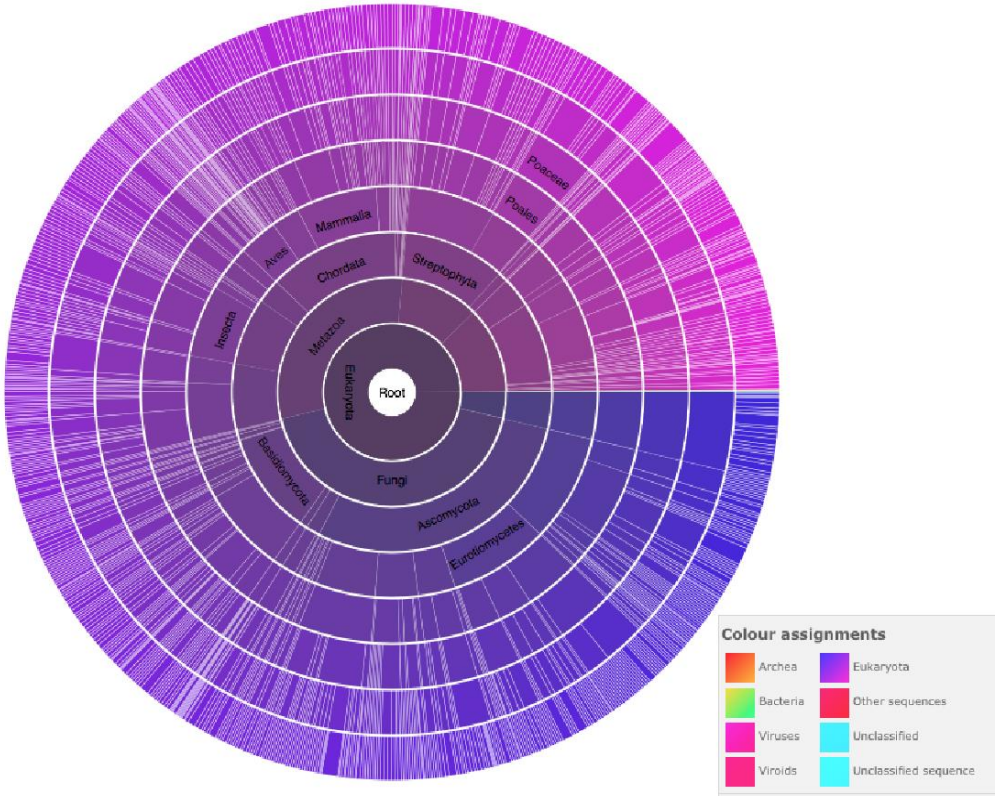

Figure 9. Distribution of prenyltransferase family across species, provided by PFAM. PFAM is part of EMBL-EBI and is provided as OpenScience [24].

2.9. Comparison with Another Information Extraction Tool

We performed a comparison of ProtozoaDB with EuPathDB [25] to evaluate the similarities and differences between these two information extraction tools (Table 2).

Table 2. Comparison with another information extraction tool (n/a = not available functionalities).

Functionalities	ProtozoaDB 2.0	EuPathDB
Blast similarities against *Homo sapiens*	Available	n/a
Blast similarities against model organisms	Available	n/a
Blast similarities against protozoa species	Available	Available
Blast similarities against CDD	Available	n/a
Blast similarities against PDB	Available	n/a
Similarities against Intepro Domains	n/a	Available
KEGG metabolic pathways	Available	n/a
Protein structures by PDB	Available	n/a
Homology study: OrthoMCL	Available	Available
Homology study: KEGG orthologous	Available	n/a
Homology study: OrthoSearch	Available	n/a
Publications at PubMed	Available	n/a
Phenotype Search	Available	n/a
SNP Characteristics Search	n/a	Available
Genomic Position Search	n/a	Available

3. Discussion

The previous version of ProtozoaDB contained only five pathogenic protozoa and some basic analyses. ProtozoaDB 2.0 increased over 17 protozoan species, totalizing 22 genomes and proteomes. New analyses were added in this new version, such as: homology analysis among the 22 organisms, using two different approaches; and phenotype inferences through orthology with the model organism. Furthermore, to allow for more comprehensive information about these organisms, several queries were performed in real time in third party (remote) sites, retrieving information about the proteome of organisms.

There are some other databases containing Protozoa species [25,26]; however, ProtozoaDB is the first database and web server that provides "all-in-one" information about comparative genomics of 22 species.

The use of web services allows for a flexible system that: (i) integrates a range of related information; (ii) has direct access to information in their original (remote) sources; and (iii) does not use local storage data from third parties (remote databases) that could imply their periodic update. These advantages allow our system to be always updated, since most of the information is queried directly in source databases through web services. The use of web services is already a practice in bioinformatics, since a number of research groups are using this technology, e.g., BioSWR [27] and BOWS [28].

Using an AJAX-based framework enables ProtozoaDB 2.0 to perform all queries through web services while simultaneously making the response time queries quite suitable for online analysis. AJAX framework is used for modern web sites, including those related to health [29,30].

The new search engines, particularly through BLAST, allow researchers to query the ProtozoaDB 2.0 data directly by the protein or gene of interest, viewing several pieces of information. Thus, it is possible to find a potential drug target by just browsing through the system and using all the information provided.

3.1. T. Brucei Case Study

Farnesyltransferase is one enzyme of the prenyltransferase group, which attaches a 15-carbon isoprenoid farnesyl group to proteins with CAAX motif: a four-amino acid sequence at the carboxyl terminus of a protein [31]. Farnesylation is a type of prenylation, a post-translational modification of proteins [32], which binds a isoprenyl group (15-carbon isoprenoid) to a cysteine residue. In other words, protein farnesylation involves protein farnesyltransferase (PFT) that catalyzes the attachment of the farnesyl group from farnesyl pyrophosphate (FPP) to cysteine SH of the C-terminal sequence motif CAAX, where C is cysteine and usually, but not always, an aliphatic residue. The terminal amino acid is determinant of farnesylation because FTase is preferentially active on protein substrates with CAAX [19]. This is an important process to mediate protein-protein interactions and protein-membrane interactions [31,33].

Prenylation (farnesylation) and subsequent modifications are essential for correct membrane targeting and cellular functioning of a number of proteins in eukaryotic cells such as Ras superfamily GTPases [34]. The farnesyltransferase enzyme is heterodimeric and has two subunits: alpha (α) and beta (β). The α subunit consists of a double layer paired alpha helices piled up in parallel, whichpartly enfold the beta subunit like a mantle.

As shown in Figure 9, prenyltransferase alpha subunit is present in various eukaryote species and several studies show that this protein is potentially a good drug target for trypanosomatids [33], especially because inhibitors have potent activity against cultured forms and are less toxic to mammalian cells than parasite cells. Besides that, PFT inhibitors have been developed as antimalarial agents [35].

3.2. Comparison between ProtozoaDB 2.0 and EupathDB

Both information extraction tools evaluated have several features that allow a wider analysis on the organism studied. EuPathDB allows a more comprehensive view of the characteristics of the protein investigated, whereas ProtozoaDB 2.0 focuses its analysis to infer and/or confirm the

functional annotation of a given protein, based on its primary annotation deposited in Genbank. Furthermore, ProtozoaDB 2.0 also allows a view of the biological role played by the protein in biological systems, including information on related literature. Through ProtozoaDB 2.0 it is possible to re-annotate some of the proteins identified as "hypothetical" through similarity-based programs as well as SuperFamily-based classification. Finally, using the tools provided by ProtozoaDB 2.0, it is also possible to infer potential drug targets, as described in our case study.

4. Materials and Methods

4.1. Web Services

ProtozoaDB 2.0 supports RESTful (REpresentational State Transfer) [36] Web Services to make data access easier. Available in http://services.biowebdb.org/howtouse, these services were written in Ruby language using Ruby on Rails (RoR), allowing access to the information about proteome of Protozoa available in ProtozoaDB 2.0 web application.

4.2. New Source Code

The source code was rewritten in Ruby using Ruby on Rails, which allowed the development of three layers: view, with web pages based on Asynchronous JavaScript and XML (AJAX); model, with search algorithms in remote web services; and the controller, which is an interface between view and model.

4.3. Data Acquisition

The primary dataset of the genome and proteome of 22 Protozoa species (Table 1), including families *Babesiidae*, *Cryptosporidiidae*, *Entamoebidae*, *Hexamitidae*, *Plasmodium*, *Theileriidae*, *Trichomonadidae*, and *Trypanosomatidae*, was downloaded from Genbank [1] in Flat File format (GBFF).

4.4. Preprocessing

Primary data were stored locally in a server with a Database Manage Systems (DBMS) through the Genomics Unified Schema (GUS) version 3.5 [37] framework. The chosen DBMS was the PostgreSQL, version 8.4. PostgreSQL is a open source object-relational database system [38].

4.5. Transformation

Some analyses were performed to incorporate new information to the primary dataset. Inferences about homologies were performed on Protozoa data and the results locally stored, namely: (i) orthology inference using OrthoMCL [12]; (ii) OrthoSearch [13], which uses the Hidden Markov Model (HMM) through HMMER (version 3.0) and best reciprocal hits; and (iii) the BLAST-based similarity results.

Phenotypes were retrieved from the *Saccharomyces* Database [21] and stored locally. Mappings between (i) KEGG Orthology (KO), (ii) *Saccharomyces cerevisae* proteins, and (iii) Protozoa proteins were performed and stored locally, aiming to infer phenotypes in Protozoa.

4.6. Analysis

Web services technologies are used to access several databases worldwide to complement existing information (Table 3), among them:

(1) Similarity against the human proteome: the system performs a query, through the web service; the human proteome is locally stored in the database, updated every six months, and returns the top ten hits, independent of the score or e-value.
(2) Search for conserved domains: by running RPSBlast against the Conserved Domain Database (CDD) [39], the system returns the top ten results independent of the score or e-value.

(3) Superfamily classification: the Superfamily database [14,40] has structural, functional, and evolutionary information of proteins from different genomes, including Protozoa. The new system performs a query through the web service, retrieving graphical information on the classification of superfamily.

(4) Similar protein structure: the system retrieves information from 2D and 3D similarity by performing BLAST [18] and FASTA [19] directly in the Protein Data Bank (PDB) [41,42].

(5) Metabolic pathways: the system performs a query for the KEGG Pathway database [10] to retrieve the metabolic pathways where a given protein participates, showing the maps and their interactions with other proteins participating in that pathway.

(6) Literature: finally, the system performs two queries in Pubmed (http://www.ncbi.nlm.nih.gov/pubmed) to retrieve the original publication of the protein and other publications having relevance to the organism and the product.

Table 3. Web services accessed.

Information	URL
ProtozoaDB	http://services.biowebdb.org/howtouse
PDB	http://www.rcsb.org/pdb/software/rest.do
Kegg	http://www.kegg.jp/kegg/docs/keggapi.html
Pubmed (NCBI)	http://www.ncbi.nlm.nih.gov/books/NBK55693/
Superfamily	http://supfam.cs.bris.ac.uk/SUPERFAMILY/web_services.html

5. Conclusions

ProtozoaDB 2.0 allows a more detailed analysis of the object of study, and expands the number of genomes and proteomes available to the scientific community. In our case study, a group of protein prenyltransferases was found by just browsing through the results provided by the web service-based tools, developed for this new version. This protein is already described in the literature as a good drug target for trypanosomatids for the following reasons: (i) its inhibitors have potent activity against cultured forms of these parasites and these inhibitors are more toxic against parasite cells than mammalian cells; (ii) for T. brucei PFT (TbPFT), the substrate specificities and inhibitor selectivity are distinct from mammalian PFT; and (iii) efforts of the pharmaceutical industry to develop small molecule inhibitors of mammalian PFTs for anti-cancer purposes creates an abundance of compounds that can be screened for selective activity against parasites. We were able to identify this potential drug target using only an "In Silico"-based strategy and the information available in public databases integrated was into ProtozoaDB 2.0.

Acknowledgments: National Council for Scientific and Technological Development (CNPq), Coordination for the Improvement of Higher Education (CAPES), Oswaldo Cruz Foundation (Fiocruz) for financial supports.

Author Contributions: All authors contributed equally to this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Benson, D.A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Sayers, E.W. GenBank. *Nucleic Acids Res.* **2013**, *41*, D36–D42. [CrossRef] [PubMed]
2. Brayton, K.A.; Lau, A.O.T.; Herndon, D.R.; Hannick, L.; Kappmeyer, L.S.; Berens, S.J.; Bidwell, S.L.; Brown, W.C.; Crabtree, J.; Fadrosh, D.; Feldblum, T.; Forberger, H.A.; et al. Genome Sequence of *Babesia bovis* and Comparative Analysis of Apicomplexan Hemoprotozoa. *PLoS Pathog.* **2007**, *3*, e148. [CrossRef] [PubMed]
3. Heidel, A.J.; Lawal, H.M.; Felder, M.; Schilde, C.; Helps, N.R.; Tunggal, B.; Rivero, F.; John, U.; Schleicher, M.; Eichinger, L.; et al. Phylogeny-wide analysis of social amoeba genomes highlights ancient origins for complex intercellular communication. *Genome Res.* **2011**, *21*, 1882–1891. [CrossRef] [PubMed]

4. Fritz-Laylin, L.K.; Prochnik, S.E.; Ginger, M.L.; Dacks, J.B.; Carpenter, M.L.; Field, M.C.; Kuo, A.; Paredez, A.; Chapman, J.; Pham, J.; et al. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell* **2010**, *140*, 631–642. [CrossRef] [PubMed]
5. Munõz, J.F.; Gallo, J.E.; Misas, E.; McEwen, J.G.; Clay, O.K. The eukaryotic genome, its reads, and the unfinished assembly. *FEBS Lett.* **2013**, *587*, 2090–2093. [CrossRef] [PubMed]
6. Kordjamshidi, P.; Roth, D.; Moens, M.F. Structured learning for spatial information extraction from biomedical text: Bacteria biotopes. *BMC Bioinform.* **2015**, *16*, 129. [CrossRef] [PubMed]
7. Dávila, A.M.R.; Mendes, P.N.; Wagner, G.; Tschoeke, D.A.; Cuadrat, R.R.C.; Liberman, F.; Matos, L.; Satake, T.; Ocaña, K.A.C.S.; Triana, O.; et al. ProtozoaDB: Dynamic visualization and exploration of protozoan genomes. *Nucleic Acids Res.* **2008**, *36*, D547–D552.
8. BioCreative, VI. Available online: http://www.biocreative.org (accessed on 16 July 2017).
9. Krallinger, M.; Valencia, A.; Hirschman, L. Linking genes to literature: Text mining, information extraction, and retrieval applications for biology. *Genome Biol.* **2008**, *9*. [CrossRef] [PubMed]
10. Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **2012**, *40*, D109–D114. [CrossRef] [PubMed]
11. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef] [PubMed]
12. Li, L.; Stoeckert, C.J.; Roos, D.S.; Christian, J.S., Jr. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* **2003**, *13*, 2178–2189. [CrossRef] [PubMed]
13. Da Cruz, S.M.S.; Batista, V.; Silva, E.; Tosta, F.; Vilela, C.; Cuadrat, R.; Tschoeke, D.; Dávila, A.M.R.; Campos, M.L.M.; Mattoso, M. Detecting distant homologies on protozoans metabolic pathways using scientific workflows. *Int. J. Data Min. Bioinform.* **2010**, *4*, 256–280. [CrossRef] [PubMed]
14. Wilson, D.; Pethica, R.; Zhou, Y.; Talbot, C.; Vogel, C.; Madera, M.; Chothia, C.; Gough, J. SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* **2009**, *37*, D380–D386. [CrossRef] [PubMed]
15. Goto, N.; Prins, P.; Nakao, M.; Bonnal, R.; Aerts, J.; Katayama, T. BioRuby: Bioinformatics software for the Ruby programming language. *Bioinform. Oxf. Engl.* **2010**, *26*, 2617–2619. [CrossRef] [PubMed]
16. Groth, P.; Weiss, B.; Pohlenz, H.D.; Leser, U. Mining phenotypes for gene function prediction. *BMC Bioinform.* **2008**, *9*, 136. [CrossRef] [PubMed]
17. Groth, P.; Pavlova, N.; Kalev, I.; Tonov, S.; Georgiev, G.; Pohlenz, H.D.; Weiss, B. PhenomicDB: A new cross-species genotype/phenotype resource. *Nucleic Acids Res.* **2007**, *35*, D696–D699. [CrossRef] [PubMed]
18. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
19. Pearson, W.R. Rapid and sensitive sequence comparison with {FASTP} and {FASTA}. *Methods Enzymol.* **1990**, *183*, 63–98. [PubMed]
20. Marchler-Bauer, A.; Lu, S.; Anderson, J.B.; Chitsaz, F.; Derbyshire, M.K.; DeWeese-Scott, C.; Fong, J.H.; Geer, L.Y.; Geer, R.C.; Gonzales, N.R. CDD: A Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **2011**, *39*, D225–D229. [CrossRef] [PubMed]
21. Cherry, J.M.; Hong, E.L.; Amundsen, C.; Balakrishnan, R.; Binkley, G.; Chan, E.T.; Christie, K.R.; Costanzo, M.C.; Dwight, S.S.; Engel, S.R. Saccharomyces Genome Database: The genomics resource of budding yeast. *Nucleic Acids Res.* **2012**, *40*, D700–D705. [CrossRef] [PubMed]
22. Maurer-Stroh, S.; Washietl, S.; Eisenhaber, F. Protein prenyltransferases. *Genome Boil.* **2003**, *4*, 212. [CrossRef]
23. Finn, R.D.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **2016**, *44*, D279–D285. [CrossRef] [PubMed]
24. European Bioinformatics Institute. Available online: http://www.ebi.ac.uk/about/terms-of-use (accessed on 16 July 2017).
25. Aurrecoechea, C.; Barreto, A.; Basenko, E.Y.; Brestelli, J.; Brunk, B.P.; Cade, S.; Crouch, K.; Doherty, R.; Falke, D.; Fischer, S. EuPathDB: The eukaryoticpathogen genomics database resource. *Nucleic Acids Res.* **2017**, *45*, D581. [CrossRef] [PubMed]

26. Anwar, T.; Gourinath, S. Deep Insight into the Phosphatomes of Parasitic Protozoa and a Web ResourceProtozPhosDB. *PLoS ONE* **2016**, *11*, e0167594. [CrossRef] [PubMed]
27. Repchevsky, D.; Gelpi, J.L. BioSWR—Semantic Web Services Registry for Bioinformatics. *PLoS ONE* **2014**, *9*, e107889. [CrossRef] [PubMed]
28. Velloso, H.; Vialle, R.A.; Ortega, J.M. BOWS (bioinformatics open web services) to centralize bioinformaticstools in web services. *BMC Res. Notes* **2015**, *8*, 206. [CrossRef] [PubMed]
29. Papastergiou, A.; Tzekis, P.; Hatzigaidas, A.; Tryfon, G.; Ioannidis, D.; Zaharis, Z.; Kampitaki, D.; Lazaridis, P. A web-based melanoma image diagnosis support system using topic map and AJAX technologies. *Inform. Health Soc. Care* **2008**, *33*, 99–112. [CrossRef] [PubMed]
30. Yeung, D.; Boes, P.; Ho, M.W.; Li, Z. A Web application for the management of clinical workflow inimage-guided and adaptive proton therapy for prostate cancer treatments. *J. Appl. Clin. Med. Phys.* **2015**, *16*, 5503. [CrossRef] [PubMed]
31. Ayong, L.; DaSilva, T.; Mauser, J.; Allen, C.M.; Chakrabarti, D. Evidence for prenylation-dependenttargeting of a Ykt6 SNARE in Plasmodium falciparum. *Mol. Biochem. Parasitol.* **2011**, *175*, 162–168. [CrossRef] [PubMed]
32. Shen, M.; Pan, P.; Li, Y.; Li, D.; Yu, H.; Hou, T. Farnesyltransferase and geranylgeranyltransferase I:Structures, mechanism, inhibitors and molecular modeling. *Drug Discov. Today* **2015**, *20*, 267–276. [CrossRef] [PubMed]
33. Buckner, F.S.; Eastman, R.T.; Nepomuceno-Silva, J.L.; Speelmon, E.C.; Myler, P.J.; Van Voorhis, W.C.; Yokoyama, K. Cloning, heterologous expression, and substrate specificities of protein farnesyltransferasesfrom Trypanosoma cruzi and Leishmania major. *Mol. Biochem. Parasitol.* **2002**, *122*, 181–188. [CrossRef]
34. Brunner, T.B.; Hahn, S.M.; Gupta, A.K.; Muschel, R.J.; McKenna, W.G.; Bernhard, E.J. Farnesyltransferase inhibitors: An overview of the results of preclinical and clinical investigations. *Cancer Res.* **2003**, *63*, 5656–5668. [PubMed]
35. Shen, Y.; Qiang, S.; Ma, S. The Recent Development of Farnesyltransferase Inhibitors as Anticancer andAntimalarial Agents. *Mini-Rev. Med. Chem.* **2015**, *15*, 837–857. [CrossRef] [PubMed]
36. Fielding, R.T. Architectural Styles and the Design of Network-Based Software Architectures. Ph.D Thesis, University of California, Irvine, CA, USA, 2000.
37. Davidson, S.B.; Crabtree, J.; Brunk, B.; Schug, J.; Tannen, V.; Overton, C.; Stoeckert, C. K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources. *IBM Syst. J.* **2001**, *40*, 512–531. [CrossRef]
38. PostgreSQL. Available online: http://www.postgres.org (accessed on 16 July 2017).
39. Marchler-Bauer, A.; Zheng, C.; Chitsaz, F.; Derbyshire, M.K.; Geer, L.Y.; Geer, R.C.; Gonzales, N.R.; Gwadz, M.; Hurwitz, D.I.; Lanczycki, C.J. CDD: Conserved domains and protein three-dimensional structure. *Nucleic Acids Res.* **2013**, *41*, D348–D352. [CrossRef] [PubMed]
40. De Lima Morais, D.A.; Fang, H.; Rackham, O.J.L.; Wilson, D.; Pethica, R.; Chothia, C.; Gough, J. SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res.* **2011**, *39*, D427–D434. [CrossRef] [PubMed]
41. Berman, H.M.; Battistuz, T.; Bhat, T.N.; Bluhm, W.F.; Bourne, P.E.; Burkhardt, K.; Feng, Z.; Gilliland, G.L.; Iype, L.; Jain, S. The Protein Data Bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2002**, *58*, 899–907. [CrossRef]
42. Berman, H.M.; Kleywegt, G.J.; Nakamura, H.; Markley, J.L. The Protein Data Bank at 40: Reflecting on the past to prepare for the future. *Struct. (Lond. Engl. 1993)* **2012**, *20*, 391–396. [CrossRef] [PubMed]