Database, 2022, 1–7
DOI: https://doi.org/10.1093/database/baac085
Original article

DGPD: a knowledge database of dense granule proteins of the Apicomplexa

Hang Hu†, Zhenxiao Lu†, Haisong Feng, Guojun Chen, Yongmei Wang, Congshan Yang* and Zhenyu Yue

School of Information and Computer, College of Animal Science and Technology, Anhui Provincial Engineering Laboratory for Beidou Precision Agriculture Information, Anhui Agricultural University, 130 Changjiangxilu, Hefei, Anhui 230036, P. R. China

* Corresponding author: Tel: +86 15652381499; Email: congshanyang@sina.cn
Correspondence may also be addressed to Zhenyu Yue. Email: zhenyuyue@ahau.edu.cn
†These authors contributed equally to this work.

Citation details: Hu, H., Lu, Z., Feng, H. *et al.* DGPD: a knowledge database of dense granule proteins of the Apicomplexa. *Database* (2022) Vol. 2022: article ID baac085; DOI: https://doi.org/10.1093/database/baac085

Abstract

Apicomplexan parasites cause severe diseases in human and livestock. Dense granule proteins (GRAs), specific to the Apicomplexa, participate in the maintenance of intracellular parasitism of host cells. GRAs have better immunogenicity and they can be emerged as important players in vaccine development. Although studies on GRAs have increased gradually in recent years, due to incompleteness and complexity of data collection, biologists have difficulty in the comprehensive utilization of information. Thus, there is a desperate need of user-friendly resource to integrate with existing GRAs. In this paper, we developed the Dense Granule Protein Database (DGPD), the first knowledge database dedicated to the integration and analysis of typical GRAs properties. The current version of DGPD includes annotated GRAs metadata of 245 samples derived from multiple web repositories and literature mining, involving five species that cause common diseases (*Plasmodium falciparum, Toxoplasma gondii, Hammondia hammondi, Neospora caninum* and *Cystoisospora suis*). We explored the baseline characteristics of GRAs and found that the number of introns and transmembrane domains in GRAs are markedly different from those of non-GRAs. Furthermore, we utilized the data in DGPD to explore the prediction algorithms for GRAs. We hope DGPD will be a good database for researchers to study GRAs.

Database URL: http://dgpd.tlds.cc/DGPD/index/

Introduction

Apicomplexan parasites include *Plasmodium falciparum*, *Toxoplasma gondii*, *Hammondia hammondi*, *Neospora caninum*, *Cystoisospora suis*, etc., causes diseases not only in animals but also in humans (1). Nearly, all creatures can be the host of the apicomplexan species (2). *P. falciparum* and *T. gondii* are the causative agents of two important human diseases: malaria and toxoplasmosis, respectively (3, 4). Toxoplasmosis, as one of the most important diseases, is also related to reproductive failure of sows (5). *N. caninum* engenders neosporosis causing infectious abortion in cattle worldwide (6). Thus, apicomplexan parasites have a great influence on human health and animal husbandry, resulting in public health problems and economic loss (7, 8).

Dense granule proteins (GRAs) are a category of immunocompetent proteins secreted by the apicomplexan parasites' secretory organelles known as dense granules. Most of the GRAs locate within the parasitophorous vacuole (PV) where the parasite multiplies and maintains intracellular parasitism in nearly all nucleated host cells, mainly by modifying the PV at the interface between the host cell and the parasite (9). Besides, several members of the GRAs also are secreted to nucleus or cytoplasm of infected host cells (10). The functions of these GRAs with different localization are also diverse, such as participating in the formation of tubular membrane (11), regulating signaling pathways in host cells (12) and affecting the transport of substances in the vacuolar membrane (13). Even so, the exact biological mechanisms of GRAs are not fully understood.

The traditional identification methods used to isolate parasite's dense granules were biochemical fractionation approaches, but the excessive parasite and/or host contamination limited its application (14). Recently, proximity-dependent biotin identification (BioID) technique has been widely used for GRAs screening, but there is also the problem of non-specific protein contamination (15). The vast workload has brought inconvenience to the experimental work and caused the waste of resources. The next-generation sequencing technology provides new ideas for peptide research and bioinformatics methods are commonly used by current researchers to discover new functional peptides. As a special class of proteins, different GRAs also share a few same features, which commonly play a role in GRAs identification (6). There are already two genomics resources

Received 16 July 2022; Revised 24 August 2022; Accepted 7 September 2022

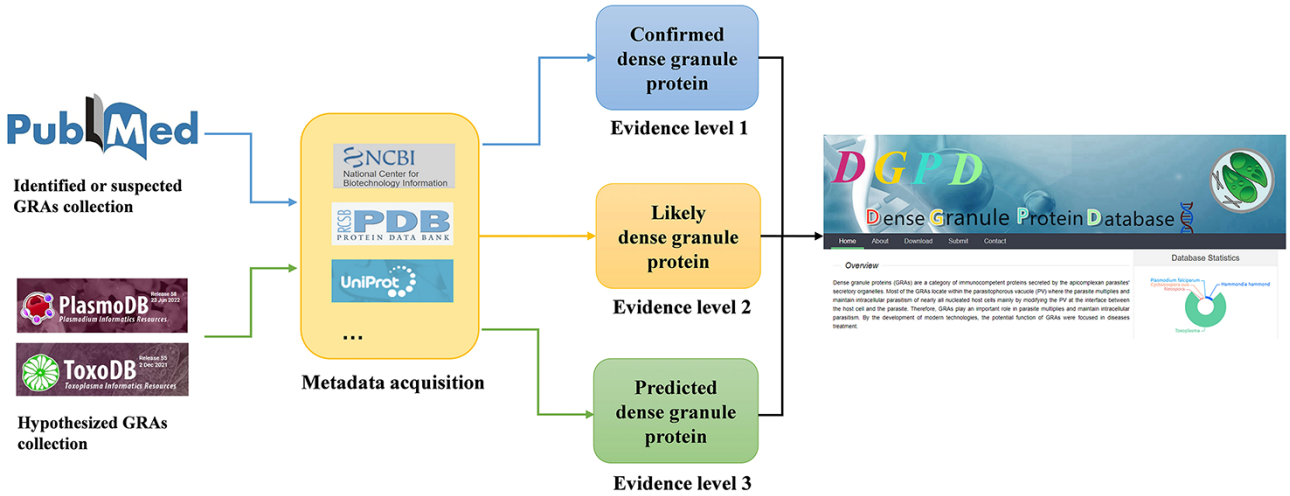

Figure 1. Workflow for data curation in DGPD database. Experimentally validated GRAs are classified as the group of 'confirmed GRAs' (with blue arrows). Highly suspected GRAs existing in the main text or attachment of literature are included in the group of 'likely GRAs' (with orange arrows). Homologous proteins of known dense granule proteins in PlasmoDB and ToxoDB database are included in the group of 'predicted GRAs' (with green arrows).

(PlasmoDB and ToxoDB) for *Plasmodium* and *Toxoplasma*, including GRAs of *P. falciparum*, *T. gondii* and other species (16, 17). However, there is no integrated database for GRAs at present, bringing about difficult for researchers to analyze functional characteristics of GRAs and to develop prediction tools.

By the development of modern technologies, many studies report that GRAs have potential applications in different aspects of functions (18, 19). But little has been done to build a golden benchmark GRA dataset in this research field, thus there is an urgent need for a dedicated database. Here, we integrate rich GRAs data to develop Dense Granule Protein Database (DGPD) to address these problems. Furthermore, we also use the available data to investigate the GRAs prediction algorithms. Comprehensive information about molecular weight, intron, signal peptide and signal peptide, etc., are available at http://dgpd.tlds.cc/DGPD/index/.

In this study, our main contributions are summarized as follows:

- We integrate GRAs about five separate species from Apicomplexans. To the best of our knowledge, this is the first time to collect and study biological information for GRAs in depth.
- We explore and analyze the baseline characteristics of GRAs in DGPD through comparison with non-GRAs, finding that GRAs tend to have fewer introns and more transmembrane domains.
- This database can be a publicly available gold-standard benchmark data set for the development and evaluation of methods for predicting novel GRAs.

Material and methods

Design of database

The DGPD construction and analysis include the following steps as shown in Figure 1.

(1) Search and collect GRAs related literature from PubMed.

(2) Sort out the specific information and relevant features of genes in the literature, according to database design and requirements.

(3) Manually screen positive and negative samples, following by feature engineering.

(4) Develop prediction models for GRAs.

(5) Build the website and complete relevant tests.

Acquisition of protein data

The core idea of this work is to analyze existing data on GRAs, and our first task is to collect the biological information. Most of the GRAs in the database are acquired in relevant literatures, and a small portion come from our previous study. At first, we searched the scientific literatures about GRAs from PubMed with a set of keywords, such as 'dense granule protein', 'GRA', 'TgGRA', 'NcGRA' and so on. At this step, more than 1200 articles were obtained. Then, we preferentially selected two model organisms, *P. falciparum* and *T. gondii*, as well as their similar species as literature screening strategies. After literatures extracting, we removed papers without full-text and others unrelated to GRAs during this process. Next, we browsed the articles based on title and abstracts, then downloaded the full-text PDF version. For each protein, we extracted the corresponding metadata including a brief description and correlation property from papers. Herein, the detailed protein information were downloaded from databases of PlasmoDB, ToxoDB, NCBI, Uniprot and PDB by the gene login number in the literature.

Noteworthy, most homologous proteins possess identical or similar functions (20). Therefore, we also searched homologous GRAs of other species in PlasmoDB and ToxoDB by experimentally validated GRAs, and then brought them into DGPD. The above method was used for data collection of other species, except *Plasmodium* and *Toxoplasma*. Many studies proved that some typical characteristics play a part in the biological function of protein, such as the presence or absence of signal peptides (21), domains (22), the number of intron (23), etc. Hence, we focused on collecting the

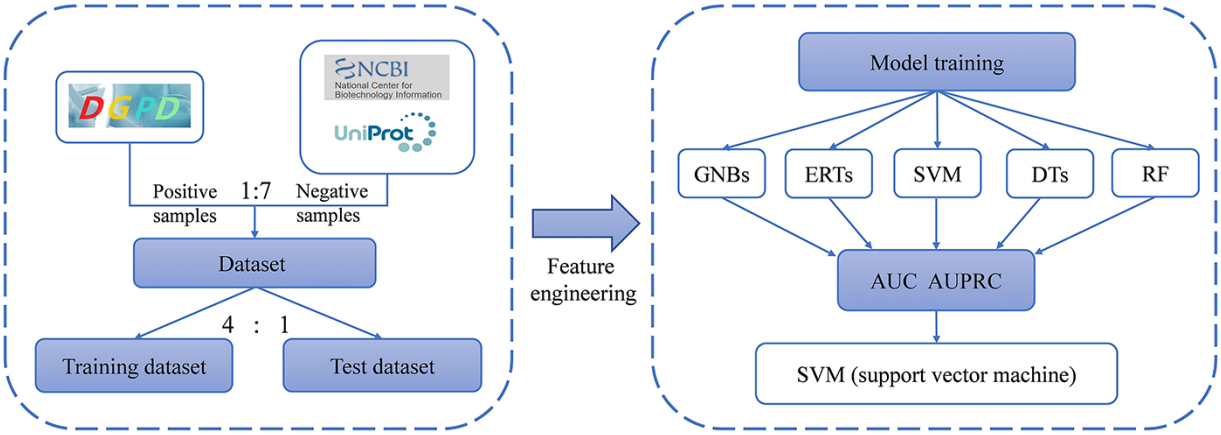

Figure 2. The workflow of prediction models for identifying GRAs.

protein characteristics that contribute to the GRAs research for improving the availability of database.

Data integration and processing

GRAs in the DGPD database contain the searchable names of gene and species, signal peptide (SignalP), intron, transmembrane domains (TMHMM), molecular weight and their level of evidence. For sharing information, data standardization and annotation are essential. Therefore, the collected data need to be processed into format that users can access.

Apart from the data directly mined from the literature, we acquired the GRAs in some existing resources. For example, ToxoDB is a database closely related to *Toxoplasma*, which has developed into maturity gradually since its initial release (24). In the ToxoDB website, the number of exons was obtained from the gene Model section and subtracted one to get the number of introns. From the Protein Feature and Properties section, the molecular weight of protein was obtained. Whether the protein contains signal peptides would be determined by the 'yes' and 'no' under the 'Has SignalP' column. For collecting the TMHMM and domain information, we need to observe the table from the attributes and protein browser section, confirming whether the domain exists through the descriptions under the track of InterPro Domains: one or more bands indicate presence and vice versa. Besides, the number of purple band under the track of transmembrane domains equals to the TMHMM number of the protein.

Combining information obtained from the related literatures, we divided the data into three evidence levels based on their source. The GRAs which has experimentally evidence are categorized as the highest level of confidence (Level 1). The ones that are documented in the main body or supplementary materials of the articles are the highly suspected GRAs (Level 2). Homologous proteins collected by the identified GRAs usually have the same function with them, regarding as the predicted GRAs (Level 3). We also added a download function to the DGPD database; biomedical researchers could explore, visualize and intuitively analyze these data.

GRAs prediction

To identify the authenticity of the data and demonstrate the usefulness of the DGPD, we constructed a binary-classification model to distinguish GRAs. The workflow for developing prediction algorithm is shown in Figure 2. We first built the training datasets from DGPD and ToxoDB. And, then feature engineering was performed to extract sequence features. Finally, five machine learning models were compared in predicting GRAs, as elaborated below. We obtained 245 protein information from the DGPD database as positive samples for the prediction experiment. For negative samples, from the ToxoDB database, we retrieved 15 621 proteins in the five parasite species as the positive ones. To increase the likelihood that proteins are not GRAs and to retain sufficient proteins for this dataset, we only used proteins with the description 'unspecified product' or 'hypothetical protein' in the ToxoDB database. A total of 2826 proteins were collected in this manner. After deduplication, we obtained 1706 proteins as the putative non-GRAs. With these constraints, the final dataset contained 245 and 1706 proteins.

iLearn is an integrated platform and meta-learner for modeling of DNA, RNA and protein sequence data (25). And, we utilized the protein sequences to extract a variety of protein features by ilearn, including CTD (composition/transition/distribution), CKSAAP (composition of k-spaced amino acid pairs), SOCNumber (sequence-order-coupling number), CTDD (distribution) and CTDC (composition) (26). In this paper, we chose the CTD features, which denoted the distribution pattern of some particular amino acids.

Five classical machine learning algorithms, i.e. decision trees, random forest, extremely randomized trees, Gaussian naïve Bayes and support vector machine (SVM), were selected to develop the classifiers. We adopted two evaluation metrics, the area under precision–recall curve (AUPRC) (27) and the area under ROC curve (AUC) (28) to evaluate the overall performance in the prediction experiment. Furthermore, as known GRAs are much less than non-GRAs, we used AUPR as the primary metric, which punishes false positive more in the evaluation process (29, 30). And, other metrics are also calculated, including recall, specificity, precision, ACC and

Table 1. Statistics in DGPD

Species	Level 1	Level 2	Level 3	Total
Toxoplasma gondii	66	26	80	172
Hammondia hammondi	11	0	18	29
Plasmodium falciparum	8	11	0	19
Neospora caninum	16	0	0	16
Cystoisospora suis	9	0	0	9

F1-score for comparing different machine learning methods for constructing prediction models.

Results and discussion

Statistics of database

Presently, DGPD provides 245 GRAs covering five typical species: *T. gondii* (70.2%), *C. suis* (3.7%), *H. hammondi* (11.8%), *N. caninum* (6.5%) and *P. falciparum* (7.8%). Some important protein metadata were supplemented in DGPD, such as the protein sequences, intron, thnum, etc. In particular, we labeled each protein with an evidence level based on its source to ensure the data credibility, including 110 confirmed GRAs (Level 1), 37 likely GRAs (Level 2) and 98 predicted GRAs (Level 3). Table 1 shows the detailed database statistics. In DGPD, the indistinct GRAs or those whose functions/features are unclear exist in the group of 'likely' or 'predicted'. We also welcome users to contact us through the Submit Panel or email provided at the webpage when finding novel GRAs. And, the request will be validated. Additionally, we will constantly collect the experimentally proven GRAs and DGPD will be periodic updated.

Implementation of database website

DGPD provides a user-friendly interactive web and users can browse, search and download the data. We adopt Django frame to coordinate MySql database for back-end setup of DGPD. LayUI, an open-source web framework is used to

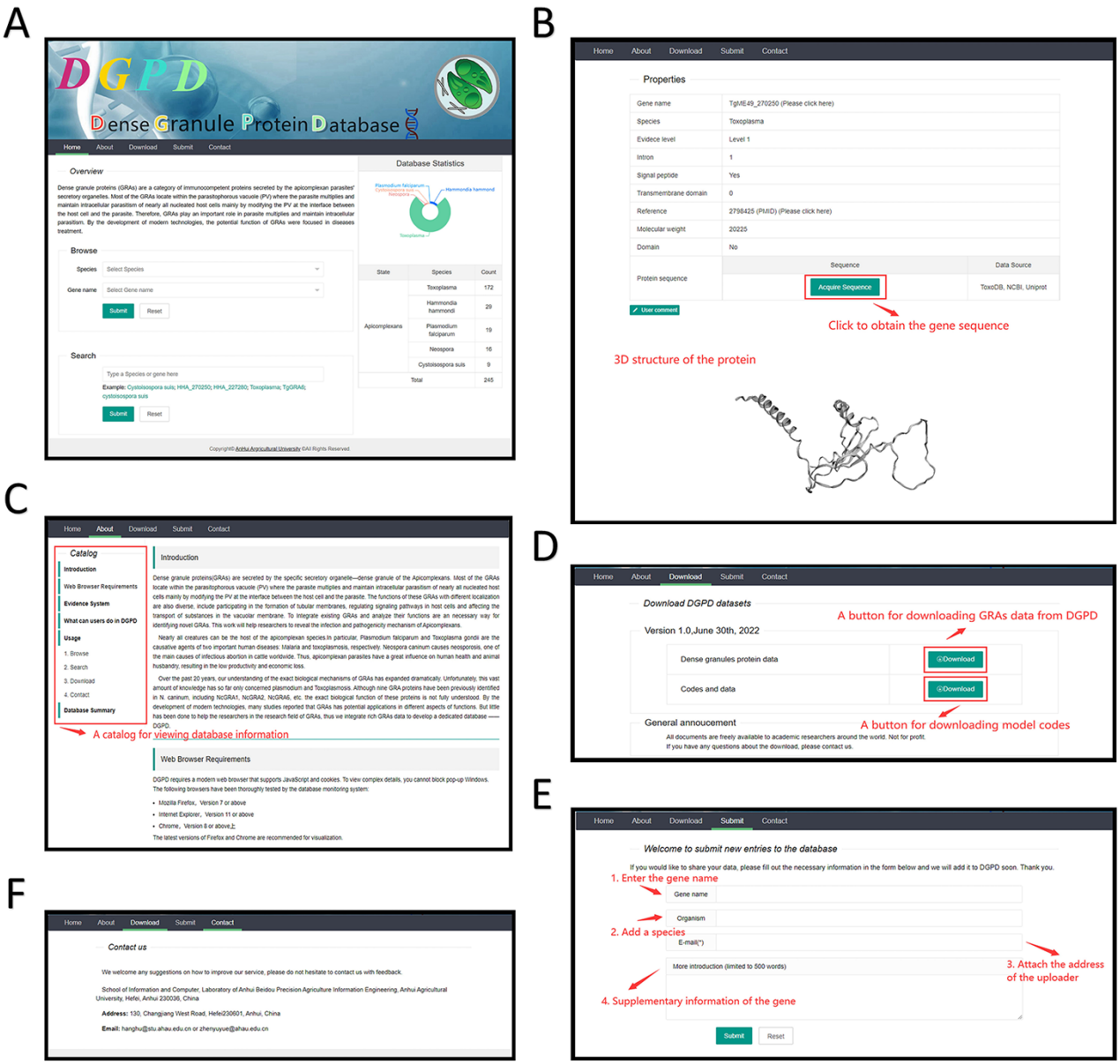

Figure 3. A web-interface of DGPD database. (A) Panel of GRA repository. A statistics visualization is displayed on the right. The gene information can be viewed by submitting keywords in search bar. (B) Panel of gene information. Detailed information of gene that users search is visualized on this panel. (C) Panel of database introduction and help. Users will receive help and brief introduction for database functions. The catalog is displayed on the top left. (D) Download panel. All data are available through this panel. (E) Data submission panel. The novel GRAs information is allowed to submit in this panel. (F) Contact panel. The different contact ways is provided for user to communicate with us.

construct the front-end panel. The DGPD homepages consist of five panels. And, Figure 3 shows the details of web.

Home Panel to search\browse proteins

In this panel, users can browse the desired proteins by selecting the species or gene names. It also allows users to use the specified data to search (e.g. organism name) (Figure 3A). After submitting specific search criteria, the webpage will redirect to gene browsing page (Figure 3B) with data message (e.g. gene sequence). Users could click on the hyperlinks of genes or PMIDs to reach the detailed information from the corresponding NCBI pages. In addition, we provide a 3D graph for each specific protein to help visualize the information on the tertiary and secondary structure of GRAs.

About panel to introduce the database

The catalog in the panel makes users utilize DGPD reasonably (Figure 3C). We adopted different tabs to facilitate users to view helps and messages about the database, such as the brief introduction, the web browser requirements and the database usages.

Download panel to obtain the protein data

All GRAs data in the DGPD are open-source. Researchers can obtain the detailed protein information by clicking the download button (Figure 3D).

Submit panel to upload the new data

In recent years, the correlational study of GRAs has developed rapidly. Many novel GRAs continue to emerge in this field. To ensure effectiveness of DGPD, we will update the database regularly. Furthermore, we welcome users to provide protein information related with new GRAs by submit panel (Figure 3E). After information submitting, we will review it, and the feedback will be sent to the submitted email address.

Contact panel to stay in touch with us

Address and email of us are listed in this panel (Figure 3F). We hope that more researchers will contribute valuable comments to our database. We would like to encourage users to communicate with us on relevant topics and issues.

Exploration of GRAs characteristics

Generally, gene-related attributes can be obtained from some dominating characteristics. For instance, the number of intron affects the gene expression (31) and the TMHMM number influences the transport of proteins (32). Thus, we carried out characteristic analysis on the curated dataset. We found that the median of intron number closes to 0 in identified GRAs. In contrast, for the negative samples, the median of intron number closes to 3 (Figure 4A). For example, the intron number of TGME49_227280 protein (GRA3) is 0 and TGGT1_209200 protein (non-GRAs) is 11. Excessive introns can cause the dysregulation of gene products expression, and GRAs possessing low-level intron number may avoid aberrant expression (33). Refer to TMHMM, the median of its quantity in GRAs usually close to 2. In contrast, the TMHMM numbers of negative samples is usually lower than that of the positive ones, and the median of TMHMM number is 0 for negative samples (Figure 4B). For instance, the TMHMM number of TgME49_268900 (GRA10) is 2 and TGME49_208760 (non-GRAs) is 0. TMHMM is essential for transmembrane proteins, mostly composed of hydrophobic amino acids (32). As secreted proteins that is mostly the type I transmembrane proteins, GRAs usually contain a variable number of TMHMM that may affect the structure of intravacuolar network membrane (9). Signal peptides play an important influence in the protein translocation (34). We also investigate the signal peptide pattern of GRAs by comparing them with putative non-GRAs. The bar plots in Figure 4C show that GRAs in DGPD are more inclined to contain a signal peptide than non-GRAs (P-values < 2.2e−16, Fisher's exact test). These results may provide new ideas for dense granule protein discovery.

Development of prediction model for GRAs

Here, to demonstrate how to use the data in DGPD, we conduct a case study to develop machine learning-based prediction models for GRAs. In this task, after generating the dataset containing positive and negative samples depicted in the above, the CTD feature descriptors are extracted (35). Then, the full dataset and quantified features are used to fit the models. We optimize the existing model framework and maximize the mean AUC, AUPRC and other evaluation

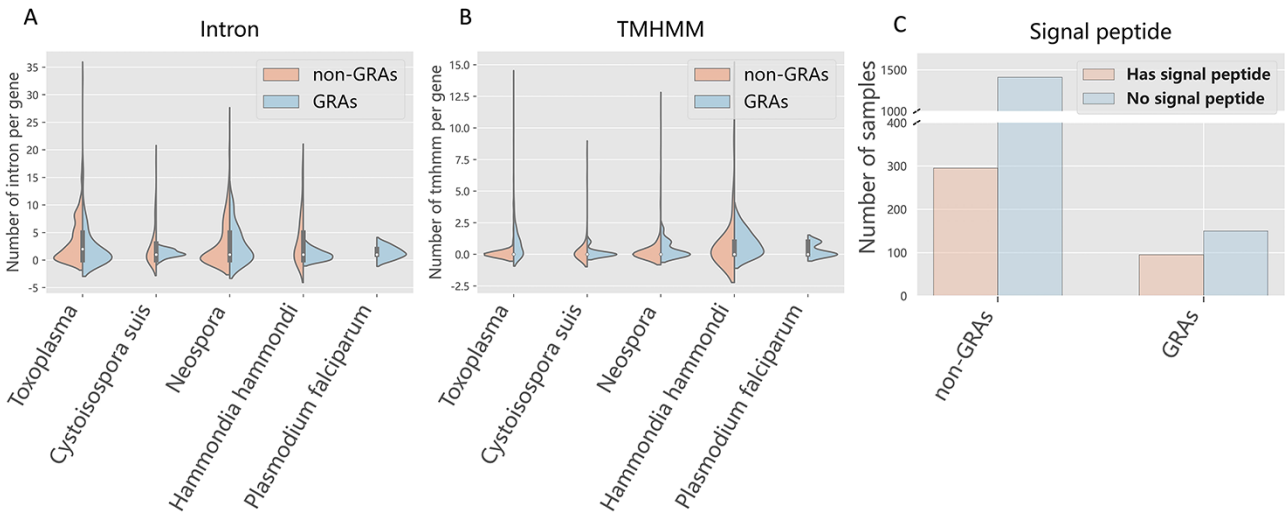

Figure 4. Feature analysis between positive and negative samples across species. Orange and blue represent GRAs and non-GRAs, respectively.

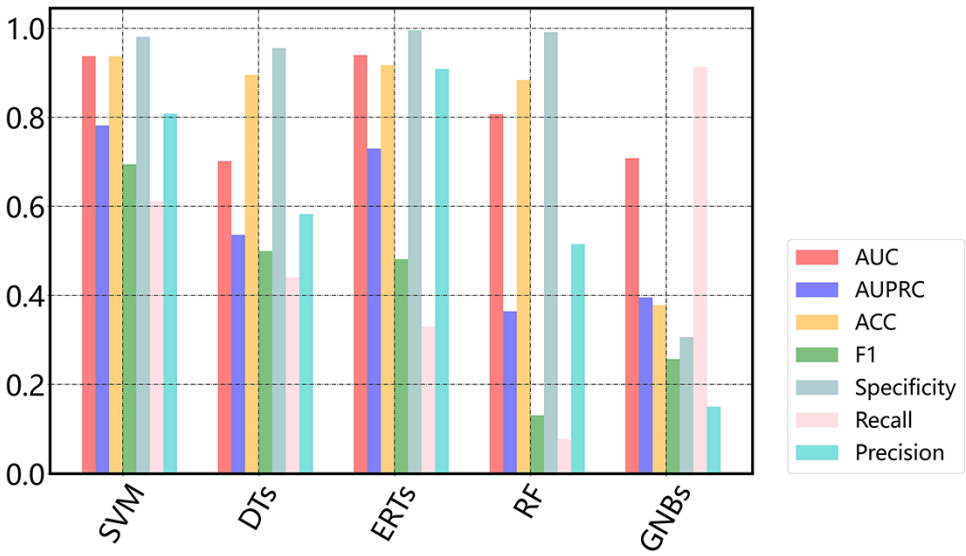

Figure 5. Performance of different machine learning-based models.

metrics by using 5-fold cross-validation (36). SVM, a machine learning method based on statistical learning theory for small sample set, is a common choice in the binary-classification problems (37, 38). Herein, the performance of SVM consistently outperforms other compared models on most metrics (Figure 5). The average AUC and AUPRC of the best model are 0.9372 and 0.7815, respectively. Thus, we select SVM to establish the prediction model and make the source codes with training data available in the Download page. In addition, we conduct hyperparameter optimization and found it had a small influence on the model performance. Figure 5 illustrates the performance of different classification algorithms on the dataset with 5-fold cross-validation.

In practice, non-GRAs always dominate over true GRAs, even a small false positive rate will result in a large number of false positive predictions. As can be seen in Figure 5, the AUC value is generally higher than AUPRC for each machine learning algorithm. This might be partly due to that the data imbalance causes the prediction bias to negative samples. This bias means more samples are classified into non-GRAs, resulting in a higher specificity that leads to a relatively high AUC. On the contrary, these false positives further lead to a lower precision, which is the crucial factor that makes the AUPRC decreases (29).

Conclusion

Dense granule proteins have been demonstrated to play a major role in multiple complex diseases caused by apicomplexan parasites. Thus, deepening the research on GRAs is essential to understand and treat the diseases. However, this research field is progressing slowly owing to difficulties in GRAs data collection and mining. In this paper, we integrate rich existing GRAs data to build the first repository, DGPD. It contains abundant proteins across five representative categories, consisting primarily of the basic information and additional annotations. Users can utilize the DGPD database information to conduct targeted research on existing GRAs and further on understanding the action mechanism.

With the development of science and technology, more novel GRAs and efficient prediction algorithms will, respectively, be discovered and developed. Our current work may still have flaws, such as the shallow study of machine learning algorithm in the GRAs prediction. By the advent of deep learning, the GRAs prediction capabilities will be strengthened. In the future, we will always focus on the latest research results in the GRAs field, and incorporate more new GRAs and species into DGPD. In view of the exact biological function of GRAs are still controversial, we will be devoted to exploring the biological problems related to GRAs. We hope that DGPD would be a unique platform for further investigation of GRAs function and action mechanism in the disease treatment.

Funding

This work was supported by the grants from the Science and Technology Department of Anhui Province (Natural Science Young Foundation of Anhui, 2008085QC136, 2008085QF293), the National Natural Science Foundation of China (62102004), the Natural Science Young Foundation of Anhui Agricultural University (2019zd12) and the Introduction and Stabilization of Talent Project of Anhui Agricultural University (yj2019-32).

Conflict of interest

None declared.

References

1. Mercier, C., Adjogble, K.D.Z., Däubener, W. *et al.* (2005) Dense granules: are they key organelles to help understand the parasitophorous vacuole of all apicomplexa parasites? *Int. J. Parasitol.*, **35**, 829–849. 10.1016/j.ijpara.2005.03.011.
2. Egea, P.F. (2020) Crossing the vacuolar rubicon: structural insights into effector protein trafficking in apicomplexan parasites. *Microorganisms*, **8**, 865. 10.3390/microorganisms8060865.

3. Hill, D. and Dubey, J.P. (2002) *Toxoplasma gondii*: transmission, diagnosis and prevention. *Clin. Microbiol. Infect.*, **8**, 634–640. 10.1046/j.1469-0691.2002.00485.x.
4. Feleke, S.M., Reichert, E.N., Mohammed, H. *et al.* (2021) *Plasmodium falciparum* is evolving to escape malaria rapid diagnostic tests in Ethiopia. *Nat. Microbiol.*, **6**, 1289–1299. 10.1038/s41564-021-00962-4.
5. Dubey, J.P., Lago, E.G. Gennari, S.M. *et al.* (2012) Toxoplasmosis in humans and animals in Brazil: high prevalence, high burden of disease, and epidemiology. *Parasitology*, **139**, 1375–1424. 10.1017/S0031182012000765.
6. Yang, C., Wang, C., Liu, J. *et al.* (2021) Biotinylation of the *Neospora caninum* parasitophorous vacuole reveals novel dense granule proteins. *Parasit. Vectors*, **14**, 521. 10.1186/s13071-021-05023-7.
7. Dessì, G., Tamponi, C., Pasini, C. *et al.* (2022) A survey on *Apicomplexa* protozoa in sheep slaughtered for human consumption. *Parasitol. Res.*, **121**, 1437–1445. 10.1007/s00436-022-07469-9.
8. Schares,G., Globokar Vrhovec, M., Tuschy, M. *et al.* (2021) A real-time quantitative polymerase chain reaction for the specific detection of *Hammondia hammondi* and its differentiation from *Toxoplasma gondii*. *Parasit. Vectors*, **14**, 78. 10.1186/s13071-020-04571-8.
9. Rome, M.E., Beck, J.R., Turetzky, J.M. *et al.* (2008) Intervacuolar transport and unique topology of GRA14, a novel dense granule protein in *Toxoplasma gondii*. *Infect. Immun.*, **76**, 4865–4875. 10.1128/IAI.00782-08.
10. Achbarou, A., Mercereau-Puijalon, O., Sadak, A. *et al.* (1991) Differential targeting of dense granule proteins in the parasitophorous vacuole of *Toxoplasma gondii*. *Parasitology*, **103**, 321–329. 10.1017/S0031182000059837.
11. Travier, L., Mondragon, R., Dubremetz, J.-F. *et al.* (2008) Functional domains of the *Toxoplasma* GRA2 protein in the formation of the membranous nanotubular network of the parasitophorous vacuole. *Int. J. Parasitol.*, **38**, 757–773. 10.1016/j.ijpara.2007.10.010.
12. Braun, L., Brenier-Pinchart, M.-P., Yogavel, M. *et al.* (2013) A *Toxoplasma* dense granule protein, GRA24, modulates the early immune response to infection by promoting a direct and sustained host p38 MAPK activation. *J. Exp. Med.*, **210**, 2071–2086. 10.1084/jem.20130103.
13. Heaslip, A.T., Nelson, S.R. and Warshaw, D.M. (2016) Dense granule trafficking in *Toxoplasma gondii* requires a unique class 27 myosin and actin filaments. *Mol. Biol. Cell*, **27**, 2080–2089. 10.1091/mbc.E15-12-0824.
14. Petry, F. and Harris, J.R. (1999) Ultrastructure, fractionation and biochemical analysis of *Cryptosporidium parvum* sporozoites. *Int. J. Parasitol.*, **29**, 1249–1260. 10.1016/S0020-7519(99)00080-6.
15. Kimmel, J., Kehrer, J., Frischknecht, F. *et al.* (2022) Proximity-dependent biotinylation approaches to study apicomplexan biology. *Mol. Microbiol.*, **117**, 553–568. 10.1111/mmi.14815.
16. Harb, O.S. and Roos, D.S. (2020) ToxoDB: functional genomics resource for toxoplasma and related organisms. In: Tonkin CJ (ed). Vol. 2071, *Toxoplasma Gondii, Methods in Molecular Biology*. Springer, New York, pp. 27–47.
17. Aurrecoechea, C., Brestelli, J., Brunk, B.P. *et al.* (2009) PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.*, **37**, D539–D543. 10.1093/nar/gkn814.
18. Fox, B.A., Sanders, K.L., Rommereim, L.M. *et al.* (2016) Secretion of rhoptry and dense granule effector proteins by nonreplicating *Toxoplasma gondii* uracil auxotrophs controls the development of antitumor immunity. *PLoS Genet.*, **12**, e1006189. 10.1371/journal.pgen.1006189.
19. Mercer, H.L., Snyder, L.M., Doherty, C.M. *et al.* (2020) *Toxoplasma gondii* dense granule protein GRA24 drives MyD88-independent p38 MAPK activation, IL-12 production and induction of protective immunity. *PLoS Pathog.*, **16**, e1008572. 10.1371/journal.ppat.1008572.
20. Overington, J.P. (1992) Comparison of three-dimensional structures of homologous proteins. *Curr. Res. Struct. Biol.*, **2**, 394–401. 10.1016/0959-440X(92)90231-U.
21. Mercier, C. and Cesbron-Delauw, M.-F. (2015) Toxoplasma secretory granules: one population or more? *Trends Parasitol.*, **31**, 60–71. 10.1016/j.pt.2014.12.002.
22. Quevillon, E., Silventoinen, V., Pillai, S. *et al.* (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120. 10.1093/nar/gki442.
23. Davis, E.O., Thangaraj, H.S., Brooks, P.C. *et al.* (1994) Evidence of selection for protein introns in the recAs of pathogenic mycobacteria. *EMBO J.*, **13**, 699–703. 10.1002/j.1460-2075.1994.tb06309.x.
24. Gajria, B., Bahl, A., Brestelli, J. *et al.* (2007) ToxoDB: an integrated *Toxoplasma gondii* database resource. *Nucleic Acids Res.*, **36**, D553–D556. 10.1093/nar/gkm981.
25. Chen, Z., Zhao, P., Li, F. *et al.* (2020) iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief. Bioinf.*, **21**, 1047–1057. 10.1093/bib/bbz041.
26. Chen, Z., Zhao, P., Li, F. *et al.* (2018) iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, **34**, 2499–2502. 10.1093/bioinformatics/bty140.
27. Ozenne, B., Subtil, F. and Maucort-Boulch, D. (2015) The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J. Clin. Epidemiol.*, **68**, 855–859. 10.1016/j.jclinepi.2015.02.010.
28. Lobo, J.M., Jiménez-Valverde, A. and Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.*, **17**, 145–151. 10.1111/j.1466-8238.2007.00358.x.
29. Davis, J. and Goadrich, M. (2006) The relationship between precision–recall and ROC curves. In: *Proceedings of the 23rd international conference on Machine learning— ICML'06*. ACM Press, Pittsburgh, Pennsylvania, pp. 233–240.
30. Yue, Z., Chu, X. and Xia, J. (2021) PredCID: prediction of driver frameshift indels in human cancer. *Brief. Bioinf.*, **22**, bbaa119. 10.1093/bib/bbaa119.
31. Buchman, A.R. and Berg, P. (1988) Comparison of intron-dependent and intron-independent gene expression. *Mol. Cell. Biol.*, **8**, 4395–4405. 10.1128/mcb.8.10.4395-4405.1988.
32. Käll, L., Krogh, A. and Sonnhammer, E.L.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036. 10.1016/j.jmb.2004.03.016.
33. Grabski, D.F., Broseus, L., Kumari, B. *et al.* (2021) Intron retention and its impact on gene expression and protein diversity: a review and a practical guide. *Wiley Interdiscip. Rev. RNA*, **12**, e1631. 10.1002/wrna.1631.
34. Choo, K.H., Tan, T.W. and Ranganathan, S. (2005) SPdb—a signal peptide database. *BMC Bioinform.*, **6**, 249. 10.1186/1471-2105-6-249.
35. Charoenkwan, P., Nantasenamat, C., Hasan, M.M. *et al.* (2022) StackDPPIV: a novel computational approach for accurate prediction of dipeptidyl peptidase IV (DPP-IV) inhibitory peptides. *Methods*, **204**, 189–198. 10.1016/j.ymeth.2021.12.001.
36. Fushiki, T. (2011) Estimation of prediction error by using K-fold cross-validation. *Stat. Comput.*, **21**, 137–146. 10.1007/s11222-009-9153-8.
37. Huang, S., Cai, N., Pacheco, P.P. *et al.* (2018) Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics*, **15**, 41–51. 10.21873/cgp.20063.
38. Mohammadi, M., Rashid, T.A., Karim,S.H.T. *et al.* (2021) A comprehensive survey and taxonomy of the SVM-based intrusion detection systems. *J. Netw. Comput. Appl.*, **178**, 102983. 10.1016/j.jnca.2021.102983.