Tropical Plants

ARTICLE

Open Access

https://doi.org/10.48130/TP-2023-0018

Tropical Plants **2023**, 2:18

SCCGs_Prediction: a machine learning tool for prediction of sulfur-containing compound associated genes

Authors

Shuang He, Liu E, Fei Chen*, Zhidong Li*

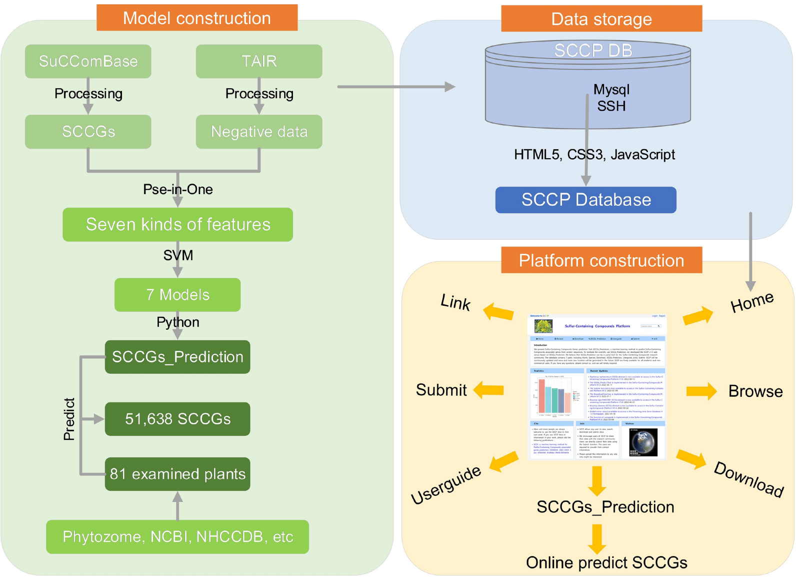
Correspondences

feichen@hainanu.edu.cn; m15132506079_1@163.com

In Brief

Developed a machine learning-empowered algorithm for predicting genes associated with sulfur-containing compounds. This algorithm was validated through enrichment analysis and literature data, and it was used for extensive predictions across 81 representative species, including bananas and cocoa. Additionally, we have established a platform for predicting genes related to sulfur-containing compounds.

Graphical abstract



Highlights

- Machine learning enabled predictive tool: We have developed a sulfur-containing compound-related gene prediction algorithm based on machine learning technology.
- Method is reliable: the reliability of the algorithm was further verified through enrichment analysis, literature data.
- Online service offered: to facilitate user access to the prediction algorithm created in this research, we have additionally provided an online gene prediction service related to sulfur-containing compounds.

Citation: He S, E L, Chen F, Li Z. 2023. SCCGs_Prediction: a machine learning tool for prediction of sulfur-containing compound associated genes. *Tropical Plants* 2:18 https://doi.org/10.48130/TP-2023-0018

Tropical
Plants

ARTICLE

Open Access

https://doi.org/10.48130/TP-2023-0018
Tropical Plants **2023**, 2:18

SCCGs_Prediction: a machine learning tool for prediction of sulfur-containing compound associated genes

Shuang He[1,2,3], Liu E[4], Fei Chen[1,2,3*] and Zhidong Li[1,2,3,4*]

[1] *College of Breeding and Multiplication, Sanya Institute of Breeding and Multiplication, Hainan University, Sanya 572025, China*
[2] *School of Tropical Agriculture and Forestry, Hainan University, Haikou 570228, China*
[3] *Hainan Yazhou Bay Seed Laboratory, Sanya 572024, China*
[4] *State Key Laboratory of Crop Genetics & Germplasm Enhancement, Key Laboratory of Biology and Genetic Improvement of Horticultural Crops (East China), Ministry of Agriculture and Rural Affairs of the PR China, Engineering Research Center of Germplasm Enhancement and Utilization of Horticultural Crop, Ministry of Education of the PR China, College of Horticulture, Nanjing Agricultural University, Nanjing Suman Plasma Engineering Research Institute, Nanjing 210095, China*
* Corresponding authors, E-mail: feichen@hainanu.edu.cn; m15132506079_1@163.com

Abstract

Sulfur-containing compounds (SCCs) are pivotal secondary metabolites widely distributed in plants, particularly within the Brassicaceae family. These compounds play crucial roles in human health and in interactions between plants and pests. In this groundbreaking study, we harnessed the extensive SuCComBase database, harvesting 1,285 protein sequences associated with sulfur-containing compounds. Employing the SVM algorithm, we pioneered the development of a predictive model for plant SCCGs, representing a novel computational approach based on sequence data. Remarkably, our SVM-Kmer model delivered exceptional performance metrics (F1score = 0.945, ACC = 0.938, AUC = 0.936). Building upon this achievement, we introduced the SCCGs_Prediction tool, a resource born of our model. Through this tool, we identified an astounding 51,638 SCCGs from a staggering 2,873,697 protein sequences spanning 81 different species. Intriguingly, our findings highlighted that the Brassicaceae and Papilionoideae subfamilies exhibit a notably higher prevalence of SCCGs compared to other plant families. In our commitment to facilitate enhanced utilization of the SCCGs_Prediction tool and the extensive plant SCCGs datasets, we have established the Sulfur-Containing Compounds Platform (SCCP). We firmly believe that the SCCP will serve as an invaluable resource hub, providing comprehensive information to the SCCs research community.

Citation: He S, E L, Chen F, Li Z. 2023. SCCGs_Prediction: a machine learning tool for prediction of sulfur-containing compound associated genes. *Tropical Plants* 2:18 https://doi.org/10.48130/TP-2023-0018

Introduction

Sulfur-containing compounds (SCCs) represent pivotal secondary metabolites that are widely distributed in the plant kingdom, with a notable prevalence in the Brassicaceae family[1] . The functional role of SCCs in mediating interactions between plants and pests is of paramount significance. Each plant family harbors its distinct set of chemical defenses, exemplified by thiopene in Asteraceae and glucosinolates in Brassicales[2]. Beyond their ecological significance, sulfur-containing compounds exhibit diverse therapeutic effects in humans, encompassing chemoprotective activities against cancer, endocrine system regulation, and improvements in sexual function[3−5]. Due to the multiple benefits of SCCs and the high incidence of cancer, researchers have increasingly shifted their focus to these compounds. Previous studies have revealed the complex regulation of SCC synthesis involving numerous genes.

To facilitate a more systematic investigation of sulfur-containing compounds (SCCs), the SuCComBase database was employed. It is the first and only manually curated resource dedicated to SCCs studies in plants (http://plant-scc.org, accessed on 15 June 2022). This comprehensive database serves as a repository for all molecular information pertaining to SCCs biosynthesis in *Arabidopsis thaliana*, encompassing SCCs biosynthetic pathway genes, proteins, and related data. By collating data from 224 papers, a total of 778 potential SCCs-related encoding genes (SCCGs) were identified, comprising 147 known *A. thaliana* sulfur-containing compounds associated genes and 92 putative sulfur-containing compounds associated genes[2]. The SuCComBase database presents a valuable resource for researchers, offering convenient access to a rich dataset that facilitates the systematic study of sulfur-containing compounds in plants.

To date, the identification of genes related to sulfur-containing compounds (SCCs) has predominantly relied on laborious and expensive biological experiments, which are also intricate in nature. Existing computational methods, such as BLAST+[6] and HMMER[7], have made significant strides in identifying homologous sequences; however, their performance remains suboptimal when it comes to identifying non-homologous sequences. Alternatively, within the realm of artificial intelligence, machine learning leverages statistical, probabilistic, and optimization methodologies to discern patterns within pre-existing data, thereby enabling the anticipation of novel data points[8]. This approach proves immensely valuable in the exploration of novel and pivotal genes associated with sulfur-containing compounds. In recent years, machine learning has demonstrated its effectiveness across a wide spectrum of biological disciplines[8−11]. Its prowess shines particularly in managing complex, multidimensional datasets that often exhibit high levels of noise and/or incompleteness. What sets machine learning apart is its capacity to work without imposing stringent assumptions about the underlying probability distribution and data generation process. Among the plethora of machine learning methods at our disposal, the Support Vector Machine (SVM) stands as a versatile and well-established choice with a track record of success across various bioinformatics challenges. For example, N'Diaye et al. effectively harnessed the SVM algorithm to unveil tissue-specific gene expression patterns in wheat[11]. The potential of machine learning holds the promise of transforming SCCs research by streamlining the identification of key genes, thus advancing our comprehension of these crucial compounds.

In this study, we have achieved a significant milestone by developing SCCGs_Prediction, a state-of-the-art machine-learning software meticulously designed for the identification of genes associated with sulfur-containing compounds using protein sequences. Powered by the robust SVM-Kmer model, which boasts impressive performance metrics (F1score = 0.945, ACC = 0.938, AUC = 0.936) honed through rigorous SVM algorithm training, our software represents a cutting-edge tool in the field. To ensure that researchers can harness the full potential of SCCGs_Prediction with ease, we have gone a step further and established a dedicated platform known as SCCP (www.sagsanno.top:8080/SCCP, accessed on 25 August 2022). This platform is exclusively tailored to cater to the intricate world of sulfur-containing compounds. We are confident that SCCP will be an invaluable resource, offering crucial insights and information to the SCCs research community.

Materials and methods

Datasets construction

The SuCComBase database stands as an invaluable resource, offering researchers unparalleled convenience and a wealth of data for the systematic exploration of sulfur-containing compounds. From this extensive repository, we extracted a comprehensive dataset comprising 147 confirmed, 92 potential, and 778 putative *A. thaliana* genes associated with sulfur-containing compounds, which served as our positive data reference[2]. For our negative data set, we turned to the TAIR database (www.arabidopsis.org, last accessed on June 16, 2022)[12]. Subsequently, we meticulously curated the data using Python scripts, removing sequences containing ambiguous amino acids (B, J, O, U, X, and Z) and those with a length below 50 residues. To further refine our dataset, we applied the CD-HIT program with a stringent threshold of 0.7 to eliminate redundant sequences. This meticulous data curation process resulted in a final dataset consisting of 1,285 sequences for our positive data and 8,494 sequences for our negative data, which were subsequently employed for training our classification model.

Selection of feature set

We harnessed the versatile Pse-in-One 2.0 software, to extract three distinct types of features essential for our analysis. These features encompassed the Kmer, Parallel Correlation Pseudo Amino Acid Composition (PC-PseAAC), and Auto-Cross Covariance (ACC) characteristics. Specifically, we configured the software to generate the Kmer feature with kmer = 2, the PC-PseAAC feature with $\lambda = 5$ and $\omega = 0.2$, and the ACC feature with LAG = 14. These feature extraction processes were executed utilizing the dedicated scripts nac.py, pse.py, and acc.py[13]. Leveraging these meticulously extracted features, we proceeded to construct the SCCGs_Prediction predictor, a pivotal component of our study.

Machine learning

Before our machine learning prediction model becomes operational, it must undergo a crucial training phase to fine-tune its parameters from an extensive array of possibilities. In this context, we harnessed the power of Support Vector Machines (SVM), an integral machine learning algorithm available through the auto-sklearn package, to construct our classification model. To optimize its performance, we meticulously fine-tuned critical hyperparameters, including cost, gamma, and kernel, employing an exhaustive grid search approach.

Evaluation method and metrics

In order to evaluate the performance of the classification model, we used the fivefold cross-validation, and three indicators including F1score, ACC, and AUC. The pROC package was used to calculate AUC values. Meanwhile, the F1score, Precision, Sensitivity and ACC were calculated using the following formulas:

$$\mathrm{F1score} = \frac{\mathrm{Precision} \times \mathrm{Sensitivity}}{\mathrm{Precision} + \mathrm{Sensitivity}}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

where TP, FP, FN, and TN represent true positive, false positive, false negative and true negative, respectively.

Large-scale predict SCCGs

We accessed protein sequences from 81 plants through publicly available databases (Supplemental Table S1)[12,14−30]. To ensure data integrity, we removed sequences that contained unidentified amino acids from our dataset. Concurrently, we employed the Pse-in-One 2.0 software to extract the Kmer feature (kmer = 2) from the refined dataset. Utilizing our meticulously trained SVM-Kmer model, we subsequently embarked on a comprehensive prediction endeavor, covering a diverse spectrum of sulfur-containing compounds across these 81 plant species.

Enrichment analysis and collection of literature data

NHCCDB database mining (http://tbir.njau.edu.cn/NhCCDbHubs/). GO Enrichment, and KEGG Enrichment tools were used for gene functional enrichment analysis. Genes related to sulfur compounds in *Brassica napus* and *Brassica rapa,* sourced from the PubMed database, were gathered and employed for testing a classification model.

Creation of the Sulfur-Containing Compounds Platform (SCCP)

The Sulfur-Containing Compounds Platform (SCCP: www.sagsanno.top:8080/SCCP or http://plants.hainanu.edu.cn/SCCP, accessed on 25 August 2022), was meticulously crafted within a Linux operating system environment, expertly hosted on an Apache Tomcat server. To create an intuitive and user-friendly interface, we artfully combined a range of programming languages and technologies in the front-end development, including Java, Python, JavaScript, and HTML scripts. On the back end, we implemented a robust data management system by housing and organizing all SCCP data within MySQL databases. This backend infrastructure ensures seamless data retrieval and storage. We are proud to mention that the SCCP website boasts cross-browser compatibility, offering accessibility through popular web browsers such as Firefox, Internet Explorer, and Google Chrome.

Results

SVM performance

In this investigation, we have harnessed the Support Vector Machine (SVM) for prediction, as SVM has demonstrated widespread success in the realm of bioinformatics. To train the classification model, 80% of the filtered dataset was utilized (Fig. 1).

Seven distinct types of features, namely SVM-ACC, SVM-Kmer, SVM-PC-PseAAC, SVM-Kmer-ACC, SVM-Kmer-PC-PseAAC, SVM-ACC-PC-PseAAC, and SVM-ACC-Kmer-PC-PseAAC, were employed in this study. To address the issue of data imbalance, we assigned different weights to the positive and negative datasets. Furthermore, we fine-tuned the cost, gamma, and kernel hyperparameters. Subsequently, the performance of the classification model was evaluated using 20% of the filtered datasets. The results, as presented in Table 1 and Supplemental Table S2, indicate that the SVM-Kmer model exhibited the most exceptional prediction performance, achieving a F1score of 0.945, ACC of 0.938, and AUC of 0.936. The SVM-Kmer-PC-PseAAC model closely followed, attaining an F1score of 0.944, ACC of 0.935, and AUC of 0.933.

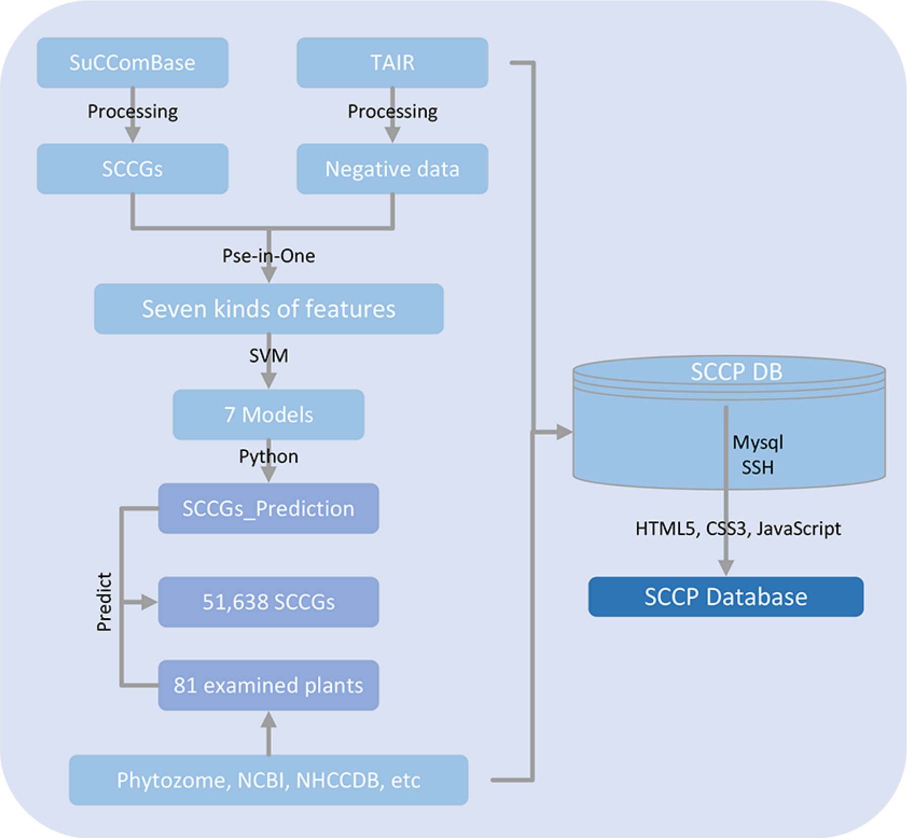


Fig. 1 Flowchart showing that SCCP database was created in this study as a dataset for training of machine learning algorithms.

Table 1. The prediction performance of the SVM model.

Methods	Number of features	F1score	ACC	AUC
SVM-ACC	100	0.904	0.906	0.895
SVM-Kmer	400	0.945	0.938	0.936
SVM-PC-PseAAC	25	0.808	0.831	0.911
SVM-Kmer-ACC	500	0.922	0.923	0.910
SVM-Kmer-PC-PseAAC	425	0.944	0.935	0.933
SVM-ACC-PC-PseAAC	125	0.916	0.917	0.907
SVM-ACC-Kmer-PC-PseAAC	525	0.921	0.923	0.911

Plant SCCGs prediction tool

We have developed a total of seven machine learning models utilizing the SVM algorithm. The evaluation results demonstrate that the SVM-Kmer model (F1score = 0.945, ACC = 0.938, AUC = 0.936) exhibits the highest performance, closely followed by the SVM-Kmer-PC-PseAAC model (F1score = 0.944, ACC = 0.935, AUC = 0.933). Leveraging the superior SVM-Kmer model, we have constructed a user-friendly prediction tool named 'SCCGs_Prediction' (www.sagsanno.top:8080/SCCP, accessed on 25 August 2022). This tool enables users to efficiently identify protein sequences encoded by plant sulfur-containing compound-associated genes on a large scale. The 'SCCGs_Prediction' tool holds great promise for researchers who have been conducting laborious wet experiments to identify genes associated with sulfur compounds in plants, significantly streamlining their efforts in this domain.

Identification of SCCGs in 81 plants

In this study, we successfully identified a total of 51,638 SCCs-related encoding genes from 2,873,697 protein sequences across 81 species (Supplemental Table S2). The investigated species encompassed 12 lower plants and 69 higher plants. The higher plants were further categorized into 49 eudicots, 16 monocots, and four other higher plants. Specifically, the examined species comprised 13 kinds of vegetables (*Asparagus officinalis*, *Beta vulgaris*, *Brassica juncea*, *Brassica oleracea*, *B. rapa*, *Capsicum annuum*, *Cicer arietinum*, *Citrullus lanatus*, *Cucumis melo*, *Cucumis sativus*, *Cucurbita maxima*, *Daucus carota*, *Raphanus raphanistrum*), 10 kinds of fruit trees (*Actinidia chinensis*, *Ananas comosus*, *Citrus grandis*, *Coffea canephora*, *Juglans regia*, *Malus domestica*, *Musa acuminata*, *Musa nana* Lour., *Phoenix dactylifera*, *Vitis vinifera*), four medicinal plants (*Leersia perrieri*, *Marchantia polymorpha*, *Panax ginseng*, *Spirodela polyrhiza*), and 13 kinds of ornamental plants (*Amaranthus hypochondriacus*, *Aquilegia coerulea*, *Capsella grandiflora*, *Chrysanthemum nankingense*, *Cynara cardunculus*, *Helianthus annuus*, *Ipomoea nil*, *Kalanchoe fedtschenkoi*, *Lupinus angustifolius*, *Phalaenopsis equestris*, *Rosa chinensis*, *Theobroma cacao*, *Trifolium pratense*). Upon analysis, the average number of SCCGs for each category was determined to be 690.29 for vegetables, 624.40 for fruit trees, 547.00 for medicinal plants, and 715.46 for ornamental plants (Fig. 2a). Notably, ornamental plants exhibited a higher average count of SCCGs compared to vegetables, fruit trees, and medicinal plants.

The average count of SCCGs was determined to be 637.50, with approximately 61.45% (51 species) of the examined species containing more than 500 SCCGs. The average proportion of SCCGs relative to the total number of genes in each species was found to be 1.75%. Among the investigated species, *Medicago truncatula* exhibited the highest SCCGs percentage at 7.61%, while *Cyanidioschyzon Merolae* and *Dunaliella Salina* demonstrated the lowest SCCGs percentages at 0.24% (Supplemental Table S2).

Validation of the constructed prediction algorithm using the predictive results from *B. rapa*. In this study, we identified 1,325 genes associated with sulfur-containing compounds in the *B. rapa* ssp. *chinensis* whole genome. The subsequent Gene Ontology (GO) enrichment analysis revealed the top 15 enriched GO terms (Fig. 3), including 'heme binding', 'oxygen binding', 'oxidoreductase activity, acting on paired donors', 'monooxygenase activity', 'indoleacetic acid biosynthetic process', 'glucosinolate biosynthetic process', 'oligopeptide transport', 'identical protein binding', 'serine-type endopeptidase activity', 'flavin adenine dinucleotide binding', 'negative regulation of catalytic activity', 'coumarin biosynthetic process', 'tryptophan catabolic process', 'pyridoxal phosphate binding', and 'ATP biosynthetic process'. This analysis revealed that processes such as 'indoleacetic acid biosynthetic process', 'glucosinolate biosynthetic process', and 'tryptophan catabolic process' are linked to sulfur-containing compounds or genes related to sulfur-containing compounds. The enrichment analysis further underscores the reliability of our prediction tool.

To validate the predictive capabilities of our algorithm on other species, we retrieved ten genes related to sulfur-containing compounds in *B. napus* and *B. rapa* from the PubMed database. These genes include *BnaC02g41790D*[31], *BnaA09g10030D*[31], *BnaA02g29380D*[31], *BnaA01g06540D*[31], *Bra029966*[32], *Bra016787*[32], *Bra011761*[32], *Bra006830*[32], *Bra011759*[32], *Bra029248*[32]. The prediction results indicate that *BnaC02g41790D*, *BnaA09g10030D*, *BnaA02g29380D*, *BnaA01g06540D*, *Bra029966*, *Bra016787*, *Bra011761*, *Bra006830*, *Bra011759*, and *Bra029248* are all associated with sulfur-containing compounds. This outcome demonstrates that the prediction tool developed in this study can indeed accurately identify other species sulfur-containing compound-related genes. This validation further strengthens the reliability and robustness of our prediction model.

Comparative analyses of SCCGs in representative plants

In higher plants, the average count of SCCGs was observed to be 735.87, contrasting with a lower average of 71.92 in lower plants (Fig. 2b). Remarkably, the number of SCCGs in higher

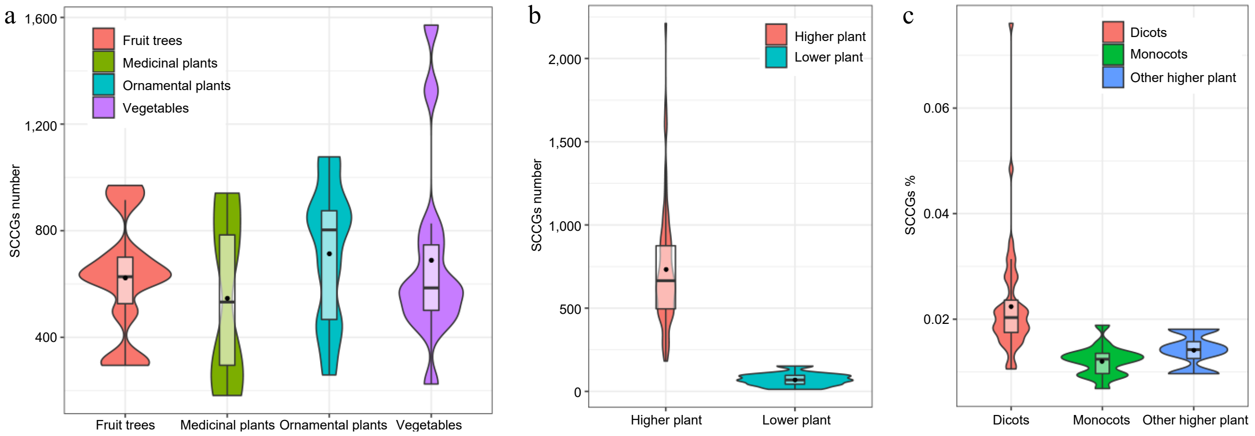


Fig. 2 Using Violin plots to compare the number of SCCGs among different plant categories. (a) Comparison of SCCGs number between Fruit trees and medicinal plants, ornamental plants, and vegetables. (b) Comparison of SCCGs number between higher plants and lower plants. (c) Comparison of SCCGs number between dicots, monocots, and other higher plant species.

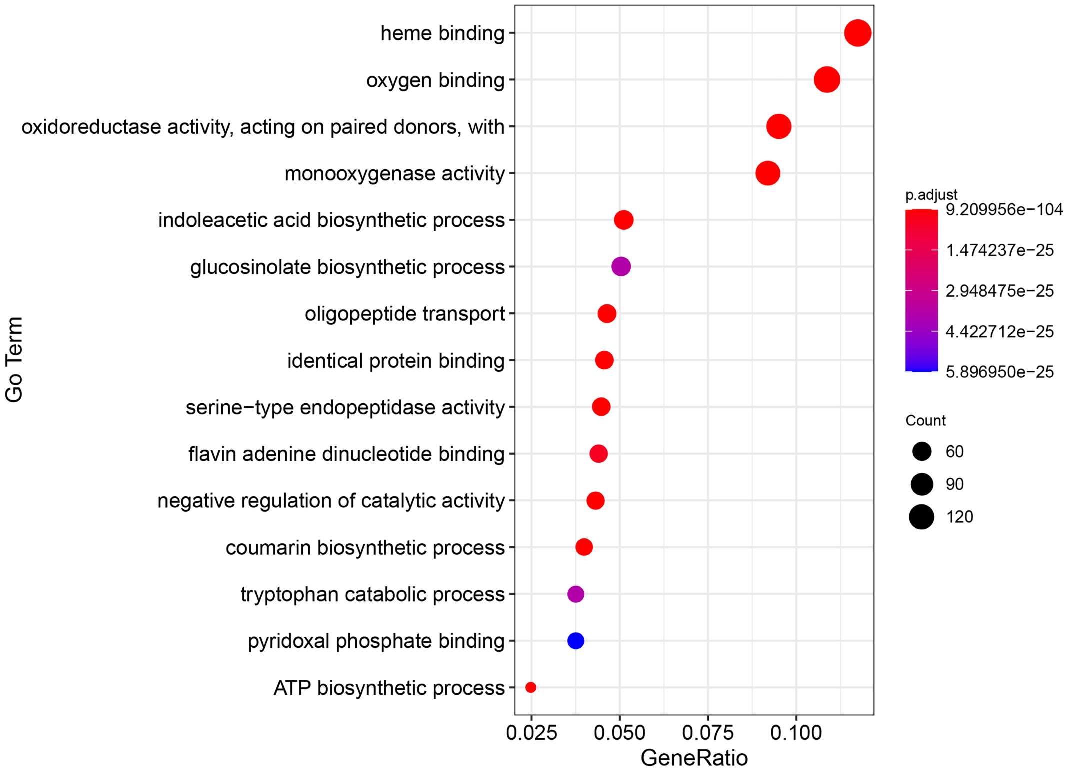


Fig. 3 The top 15 GO enrichment items of genes related to sulfur-containing compounds in *B. rapa*.

plants surpasses that in lower plants by more than tenfold. This significant difference in SCCG abundance between higher and lower plants may be attributed to the occurrence of whole-genome duplication and whole-genome triplication events, which are common in most higher plants. These genomic events likely contributed to the expansion and diversification of SCCGs in higher plants compared to their lower plant counterparts.

Among the top 20 species exhibiting a higher percentage of sulfur-containing compound-associated genes, all of them belonged to eudicots plants (Supplemental Table S2). This phenomenon suggested that eudicots plants might contain a higher proportion of sulfur-containing compounds associated genes than other higher plants (Fig. 2c).

Among the top 10 species with a lower percentage of sulfur-containing compound-associated genes, the majority (nine) belonged to lower plants (Supplemental Table S2). Notably, the sole higher plant in this category was *Musa nana Lour.* (banana), a tropical fruit. In banana, only 295 sulfur-containing compound-associated genes were detected from a total of 43,041 genes in the whole genome, representing a mere 0.69% of all genes (Supplemental Table S2). This observation suggests that gene losses of sulfur-containing compound-associated genes may have occurred more frequently than gene duplications in banana. Moreover, the two species with the lowest percentage of sulfur-containing compound-associated genes, *Cyanidioschyzon Merolae* and *Dunaliella Salina*, both belong to lower plants.

Sulfur-containing compounds are pivotal secondary metabolites that exhibit widespread occurrence in plants, with particular significance in the Brassicaceae family. Intriguingly, among the top 10 species with a higher percentage of sulfur-containing compound-associated genes, the majority (eight) belong to the Brassicaceae family (Supplemental Table S2). These species include *Arabidopsis thaliana*, *Capsella rubella*, *Boechera stricta*, *Eutrema salsugineum*, *Capsella grandiflora*, *Barbarea vulgaris*, *Arabidopsis helleri*, and *Schrenkiella parvula*. This observation suggests that the Brassicaceae family may harbor a higher proportion of genes associated with sulfur-containing compounds compared to other plant families. Notably, *Medicago truncatula*, belonging to the Papilionoideae subfamily, exhibits the highest percentage of sulfur-containing compound-associated genes. In *Medicago truncatula*, a total of 1077 SCCGs were detected from the 14,158 genes in the whole genome, accounting for 7.61% of all genes (Supplemental Table S2). The presence of another species, *Lupinus angustifolius*, from the Papilionoideae subfamily among the top 10 species with a higher percentage of SCCGs also supports the notion that the Papilionoideae subfamily might indeed possess a higher proportion of genes associated with sulfur-containing compounds. Further investigation is warranted to explore the unique biochemical and ecological roles of sulfur-containing compounds in the Brassicaceae and Papilionoideae.

We conducted a comparison using the widely adopted homologous sequence search software, Blast+. Using the Blast+ software, we identified a total of 2846 sulfur compound-related genes from the *B. rapa* genome, with the following parameters: E-value ≤ 10^{-5}, Identity > 60%, and Score > 150. The algorithm developed in our research identified 1325 sulfur compound-related genes in the *B. rapa* genome, of which 824 were found to be common between the two algorithms (Fig. 4a). Our research algorithm also uncovered 501 sulfur compound candidate genes that were not identified by the Blast+ tool. Further KEGG enrichment analysis of these 501 candidate genes revealed a significant association with the 'sulfate transporter 3' pathway, which is related to sulfur compounds (Fig. 4b). Compared to the Blast+ software, our research algorithm has predicted three novel candidate genes for sulfate transporter 3, namely, *BraC03g047110.1*, *BraC09g049990.1*, and *BraCxxg010000.1*.

Platform construction of sulfur-containing compounds

By leveraging the SCCGs_Prediction tool and a comprehensive dataset comprising 51,638 encoding genes associated with sulfur-containing compounds, we successfully established the Plant Sulfur-Containing Compounds Platform (SCCP: www.saggsanno.top:8080/SCCP, accessed on 25 August 2022). The SCCP is thoughtfully designed to empower researchers with functionalities to predict, download, and search for encoding genes related to sulfur compounds in plants (Fig. 5). This platform encompasses seven core tools, namely Home, Browse, Download, SCCGs_Prediction, Submit, Userguide, and Link. We are committed to continuous updates and enhancements of SCCP, ensuring it serves as a comprehensive community resource for advancing research in the domain of plant sulfur-containing compounds.

Tool SCCGs_Prediction

From Table 1, it is evident that the SVM-Kmer classification model achieved the highest performance with an F1score of 0.945, ACC of 0.938, and AUC of 0.936. The SVM-Kmer-PC-PseAAC model secured the second position with an F1score of 0.944, ACC of 0.935, and AUC of 0.933. Leveraging the SVM-Kmer and SVM-Kmer-PC-PseAAC models, we have developed an online service using Java, HTML5, and JavaScript scripts. This service allows users to predict encoding genes associated with sulfur-containing compounds by simply uploading the amino acid sequence in FASTA format, selecting the preferred classification model, and submitting the task. The prediction results can be conveniently viewed and downloaded directly from the results interface (Fig. 6).

Tool Browse

In this research, we have successfully identified 51,638 sulfur-containing compounds associated genes (SCCGs) from a comprehensive pool of 2,873,697 gene sequences encompassing 81 different species. To enhance user accessibility to these datasets, the plant SCCGs datasets have been diligently organized and stored within the Browse module. For each species, we have compiled detailed information comprising gene identification (ID), coding sequences (CDS), and protein sequences (PEP). Researchers can effortlessly access the desired information by selecting the corresponding species option. To illustrate the browsing results, we have employed *B. rapa* as a representative example. Users can also conveniently download the sequences of their interest through the dedicated download module. This streamlined interface is intended to foster a user-friendly experience and expedite research in the realm of sulfur-containing compounds in plants.

Tool Download

The SCCGs_Prediction tool and plant SCCGs datasets are readily accessible through the Download module, which comprises two main sections: Forecasting Tool and SCCGs. Within the Forecasting Tool section, users can download not only the SCCGs_Prediction prediction tool but also the SVM-Kmer, SVM-Kmer-PC-PseAAC models, as well as the positive and negative datasets. On the other hand, the SCCGs part encompasses data from 81 species, comprising 12 lower plants and 69 higher plants, with a collective total of 51,638 genes associated with sulfur-containing compounds. Researchers can efficiently acquire the pertinent data they need through the convenient and user-friendly interface offered by the Download module.

Userguide and Submit tool

We provide instructions to help users use the SCCP website better, faster, and more easily. Frequently asked questions are also provided at the bottom of the home page. To encourage users to share the data related to sulfur-containing compounds, we added the Submit function in the SCCP database. We believe that the SCCP database will be useful for all researchers studying the gene associated with sulfur-containing compounds.

Discussion

Sulfur-containing compounds (SCCs) are significant secondary metabolites extensively found in plants, playing a crucial role in plant-pest interactions. Furthermore, SCCs have shown diverse therapeutic effects in humans, including

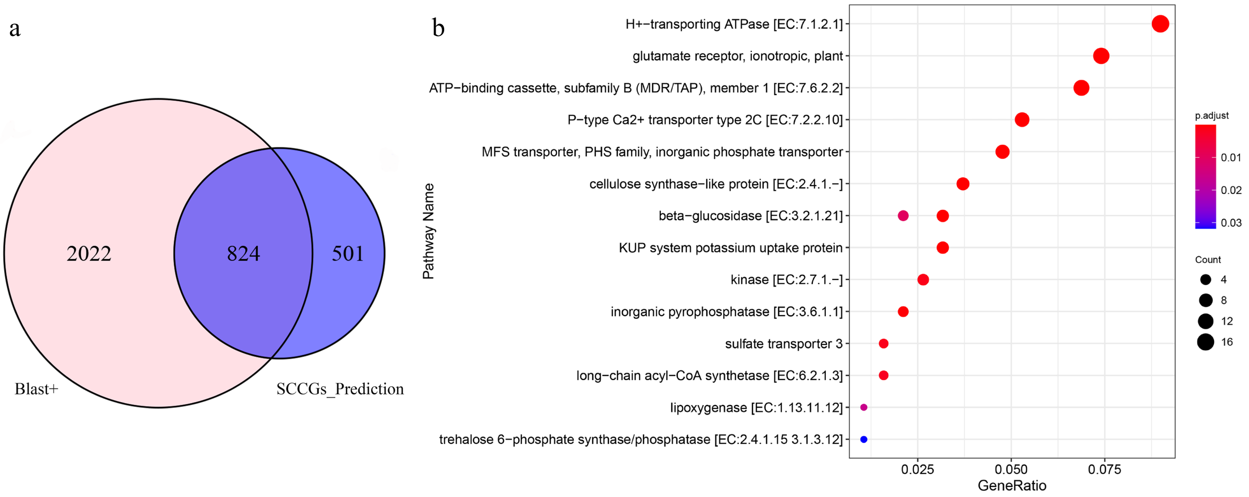
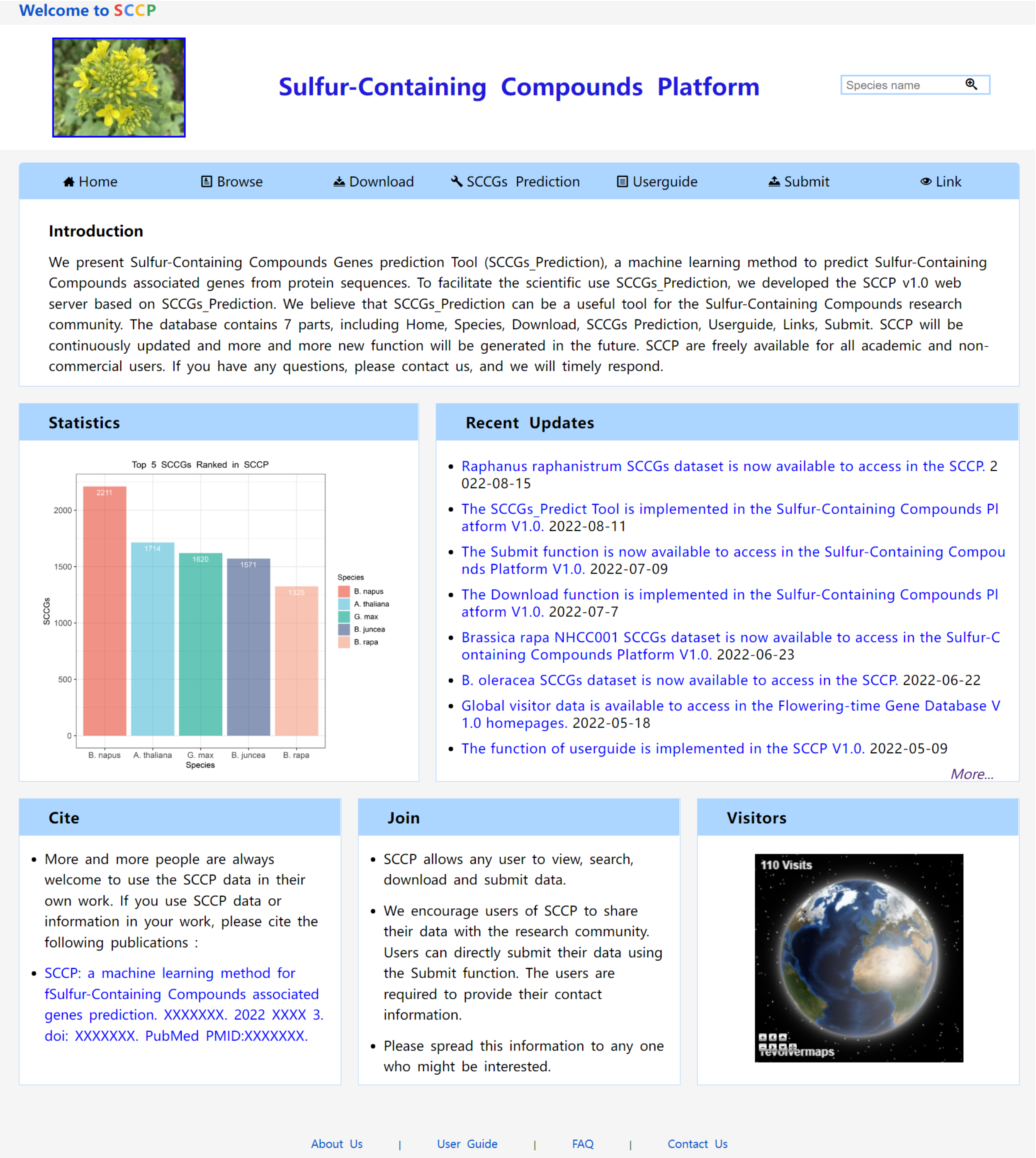


Fig. 4 Comparative analysis of sulfur compound-related genes unearthed by Blast+ and SCCGs_Prediction tools. (a) Common and differential genes detected by Blast+ and SCCGs_Prediction tools. (b) KEGG enrichment analysis of 501 specifically identified by SCCGs_Prediction tool.



Welcome to SCCP

Sulfur-Containing Compounds Platform

Introduction

We present Sulfur-Containing Compounds Genes prediction Tool (SCCGs_Prediction), a machine learning method to predict Sulfur-Containing Compounds associated genes from protein sequences. To facilitate the scientific use SCCGs_Prediction, we developed the SCCP v1.0 web server based on SCCGs_Prediction. We believe that SCCGs_Prediction can be a useful tool for the Sulfur-Containing Compounds research community. The database contains 7 parts, including Home, Species, Download, SCCGs Prediction, Userguide, Links, Submit. SCCP will be continuously updated and more and more new function will be generated in the future. SCCP are freely available for all academic and non-commercial users. If you have any questions, please contact us, and we will timely respond.

Statistics

Recent Updates

- Raphanus raphanistrum SCCGs dataset is now available to access in the SCCP. 2022-08-15
- The SCCGs_Predict Tool is implemented in the Sulfur-Containing Compounds Platform V1.0. 2022-08-11
- The Submit function is now available to access in the Sulfur-Containing Compounds Platform V1.0. 2022-07-09
- The Download function is implemented in the Sulfur-Containing Compounds Platform V1.0. 2022-07-7
- Brassica rapa NHCC001 SCCGs dataset is now available to access in the Sulfur-Containing Compounds Platform V1.0. 2022-06-23
- B. oleracea SCCGs dataset is now available to access in the SCCP. 2022-06-22
- Global visitor data is available to access in the Flowering-time Gene Database V1.0 homepages. 2022-05-18
- The function of userguide is implemented in the SCCP V1.0. 2022-05-09

More...

Cite

- More and more people are always welcome to use the SCCP data in their own work. If you use SCCP data or information in your work, please cite the following publications :
- SCCP: a machine learning method for fSulfur-Containing Compounds associated genes prediction. XXXXXXX. 2022 XXXX 3. doi: XXXXXXX. PubMed PMID:XXXXXXX.

Join

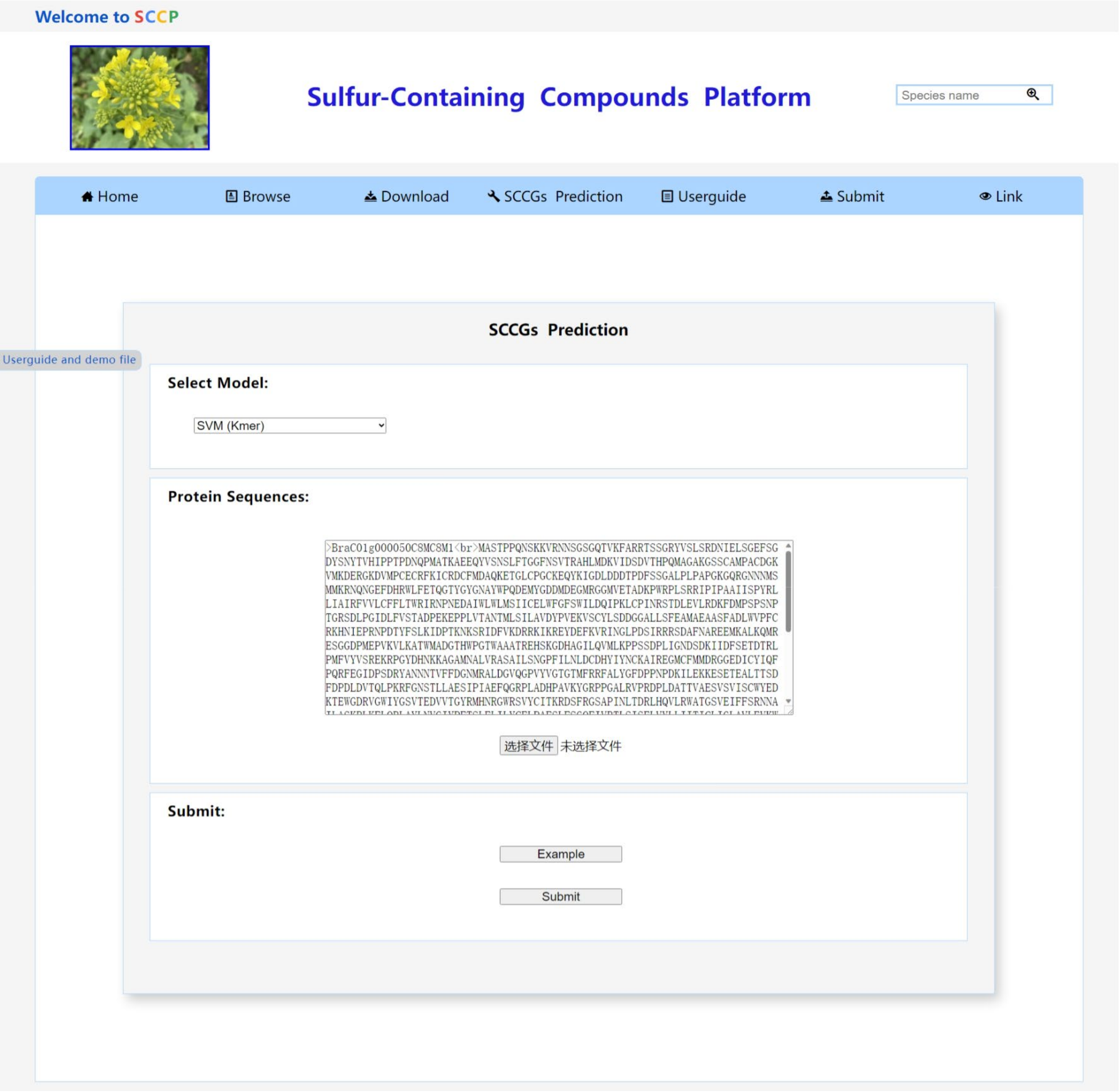
- SCCP allows any user to view, search, download and submit data.
- We encourage users of SCCP to share their data with the research community. Users can directly submit their data using the Submit function. The users are required to provide their contact information.
- Please spread this information to any one who might be interested.

Visitors

About Us | User Guide | FAQ | Contact Us

Fig. 5 Home page of the SCCP website.

chemoprotective properties against cancer[33,34], enhancement of the immune system[35], reduction in the risk of diabetes[36], etc. Currently, the identification of SCCs genes primarily relies on labor-intensive biological experiments and high-throughput omics technologies, incurring substantial costs and time. While BLAST+[6] and HMMER[7] tools have been effective in identifying homologous sequences, their efficiency in recognizing non-homologous sequences remains limited. To address this challenge, we introduce a novel computational approach in this study, aimed at identifying encoding genes associated with sulfur-containing compounds. This innovative methodology holds the potential to streamline and speed up gene discovery in the context of SCCs research. Comparing the existing tools for predicting sulfur-containing compounds, HMS-S-S[37] and HMSS2[38], both are constructed using the Hidden Markov Model (HMM) algorithm and are used to predict sulfur-containing compounds in prokaryotes. In contrast, SCCGs_Prediction is constructed using the Support Vector Machine (SVM) algorithm and is primarily focused on predicting sulfur-containing compounds in plants, which are eukaryotes. These two tools complement each other and contribute to advancing research on genes related to sulfur-containing compounds.



SCCGs Prediction

Select Model:

SVM (Kmer)

Protein Sequences:

```
>BraC01g000050C8MC8M1<br>MASTPPQNSKKVRNNSGSGQTVKFARRTSSGRYVSLSRDNIELSGEFSG
DYSNYTVHIPPTPDNQPMATKAEEQYVSNSLFTGGFNSVTRAHLMDKVIDSDVTHPQMAGAKGSSCAMPACDGK
VMKDERGKDVMPCECRFKICRDCFMDAQKETGLCPGCKEQYKIGDLDDDTPDFSSGALPLPAPGKGQRGNNNMS
MMKRNQNGEFDHRWLFETQGTYGYGNAYWPQDEMYGDDMDEGMRGGMVETADKPWRPLSRRIPIPAAIISPYRL
LIAIRFVVLCFFLTWRIRNPNEDAIWLWLMSIICELWFGFSWILDQIPKLCPINRSTDLEVLRDKFDMPSPSNP
TGRSDLPGIDLFVSTADPEKEPPLVTANTMLSILAVDYPVEKVSCYLSDDGGALLSFEAMAEAASFADLWVPFC
RKHNIEPRNPDTYFSLKIDPTKNKSRIDFVKDRRKIKREYDEFKVRINGLPDSIRRRSDAFNAREEMKALKQMR
ESGGDPMEPVKVLKATWMADGTHWPGTWAAATREHSKGDHAGILQVMLKPPSSDPLIGNDSDKIIDFSETDTRL
PMFVYVSREKRPGYDHNKKAGAMNALVRASAILSNGPFILNLDCDHYIYNCKAIREGMCFMMDRGGEDICYIQF
PQRFEGIDPSDRYANNNTVFFDGNMRALDGVQGPVYVGTGTMFRRFALYGFDPPNPDKILEKKESETEALTTSD
FDPDLDVTQLPKRFGNSTLLAESIPIAEFQGRPLADHPAVKYGRPPGALRVPRDPLDATTVAESVSVISCWYED
KTEWGDRVGWIYGSVTEDVVTGYRMHNRGWRSVYCITKRDSFRGSAPINLTDRLHQVLRWATGSVEIFFSRNNA
```

Submit:

Fig. 6 The SCCGs_Prediction tool page of the SCCP website.

In this study, we utilized the SVM algorithm and employed seven distinct features, namely SVM-ACC, SVM-Kmer, SVM-PC-PseAAC, SVM-Kmer-ACC, SVM-Kmer-PC-PseAAC, SVM-ACC-PC-PseAAC, and SVM-ACC-Kmer-PC-PseAAC, to train the classification model. Among these, the SVM-Kmer model demonstrated the most outstanding performance, achieving an impressive F1score of 0.945, ACC of 0.938, and AUC of 0.936. Leveraging the power of the SVM-Kmer model, we successfully developed the SCCGs_Prediction tool, enabling the identification of protein sequences encoded by plant sulfur-containing compound-associated genes. Importantly, our computational approach represents the first of its kind in predicting sulfur-containing compound-associated genes solely based on protein sequences. This pioneering method has effectively filled an international gap in the related field, opening new avenues for further research and exploration in this domain.

Moreover, we conducted large-scale predictions of SCCGs from 81 species, comprising 12 lower plants and 69 higher plants. Through a comprehensive analysis, we successfully identified a total of 51,638 SCCGs from the vast dataset encompassing 2,873,697 protein sequences of these 81 species. Notably, the abundance of SCCGs detected in higher plants significantly surpasses that in lower plants. This disparity may be attributed to the prevalent occurrence of whole-genome duplication and whole-genome triplication events in most higher plants, leading to an expansion of SCCGs in these species[39,40]. The significance of SCCs as crucial secondary metabolites is evident, particularly in the Brassicaceae family[41]. Most Brassicaceae species have demonstrated a higher percentage of sulfur-containing compound-associated genes compared to other species. Among the top 10 species exhibiting a higher percentage of SCCGs, two species belong to the Papilionoideae subfamily. We speculate that other species within the Papilionoideae subfamily might also possess a higher proportion of SCCGs. Notably, *Musa nana Lour*, a higher plant, accounts for merely 0.69% of all genes. This phenomenon suggests that gene losses of SCCGs may have outpaced gene duplications in this particular species.

Leveraging the SCCGs datasets and the SCCGs_Prediction program, we have successfully established the Sulfur-Containing Compounds Platform (SCCP: www.sagsanno.top:8080/SCCP, accessed on 25 August 2022), aiming to facilitate scientists in accessing plant sulfur-containing compound-associated genes datasets and predicting novel sulfur-containing compound-associated genes. As new genome sequences are unveiled in the future, we are committed to continuously identifying sulfur-containing compound-associated genes from these datasets, further enriching our database with comprehensive and up-to-date information.

Undoubtedly, this database will serve as a valuable and indispensable resource, catering to the needs of researchers across various disciplines and enhancing the collective understanding of sulfur-containing compounds and their role in the intricate world of plants.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: Li Z, He S, Chen F; experiments: Li Z; manuscript preparation: He S, Chen F, Li Z, E L. All authors reviewed the results and approved the final version of the manuscript.

Data availability

All data generated or analyzed during this study are included in SCCP (www.sagsanno.top:8080/SCCP or http://plants.hainanu.edu.cn/SCCP).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (32172614), National Natural Science Foundation of China-CG joint foundation (3181101517), Hainan Province Science and Technology Special Fund (ZDYF2023XDNY050). Authors thank the anonymous reviewers for their invaluable comments and suggestions. We thank a number of users for reporting bugs as well as the editor and anonymous reviewers for their valuable comments on this work.

Conflict of interest

The authors declare that they have no conflict of interest. Fei Chen is the Editorial Board member of *Tropical Plants* who was blinded from reviewing or making decisions on the manuscript. The article was subject to the journal's standard procedures, with peer-review handled independently of this Editorial Board member and the research groups.

Supplementary Information accompanies this paper at (https://www.maxapress.com/article/doi/10.48130/TP-2023-0018)

Dates

Received 5 August 2023; Accepted 10 October 2023; Published online 25 October 2023

References

1. Harun S, Rohani ER, Ohme-Takagi M, Goh HH, Mohamed-Hussein ZA. 2021. ADAP is a possible negative regulator of glucosinolate biosynthesis in *Arabidopsis thaliana* based on clustering and gene expression analyses. *Journal of Plant Research* 134:327−39
2. Harun S, Abdullah-Zawawi MR, A-Rahman MRA, Muhammad NAN, Mohamed-Hussein ZA. 2019. SuCComBase: a manually curated repository of plant sulfur-containing compounds. *Database* 2019:baz021
3. Nowicki D, Rodzik O, Herman-Antosiewicz A, Szalewska-Pałasz A. 2016. Isothiocyanates as effective agents against enterohemorrhagic *Escherichia coli*: insight to the mode of action. *Scientific Reports* 6:22263
4. Rungapamestry V, Duncan AJ, Fuller Z, Ratcliffe B. 2007. Effect of cooking brassica vegetables on the subsequent hydrolysis and metabolic fate of glucosinolates. *The Proceedings of the Nutrition Society* 66:69−81
5. Prieto MA, López CJ, Simal-Gandara J. 2019. Glucosinolates: molecular structure, breakdown, genetic, bioavailability, properties and healthy and adverse effects. *Advances in Food and Nutrition Research* 90:305−50
6. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421
7. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, et al. 2018. HMMER web server: 2018 update. *Nucleic Acids Research* 46:W200−W204
8. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, et al. 2018. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics & Proteomics* 15:41−51
9. Li Z, Tang W, You X, Hou X. 2022. LSAP: a machine learning method for leaf-senescence-associated genes prediction. *Life* 12:1095
10. Meher PK, Mohapatra A, Satpathy S, Sharma A, Saini I, et al. 2021. PredCRG: a computational method for recognition of plant circadian genes by employing support vector machine with Laplace kernel. *Plant Methods* 17:46
11. N'Diaye A, Byrns B, Cory AT, Nilsen KT, Walkowiak S, et al. 2020. Machine learning analyses of methylation profiles uncovers tissue-specific gene expression patterns in wheat. *The Plant Genome* 13:e20027
12. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, et al. 2012. The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* 40:D1202−D1210
13. Liu B, Liu F, Wang X, Chen J, Fang L, et al. 2015. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Research* 43:W65−W71
14. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* 40:D1178−D1186
15. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, et al. 2022. Database resources of the national center for biotechnology information. *Nucleic Acids Research* 50:D20−D26
16. Gupta P, Naithani S, Tello-Ruiz MK, Chougule K, D'Eustachio P, et al. 2016. Gramene database: navigating plant comparative genomics resources. *Current Plant Biology* 7−8:10−15
17. Yu J, Zhao M, Wang X, Tong C, Huang S, et al. 2013. Bolbase: a comprehensive genomics database for *Brassica oleracea*. *BMC Genomics* 14:664
18. Li Z, Li Y, Liu T, Zhang C, Xiao D, et al. 2022. Non-heading Chinese cabbage database: an open-access platform for the genomics of *Brassica campestris* (syn. *Brassica rapa*) ssp. chinensis. *Plants* 11:1005
19. Zheng Y, Wu S, Bai Y, Sun H, Jiao C, et al. 2019. Cucurbit Genomics Database (CuGenDB): a central portal for comparative and

functional genomics of cucurbit crops. *Nucleic Acids Research* 47:D1128−D1136

20. Brown AV, Conners SI, Huang W, Wilkey AP, Grant D, et al. 2021. A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Research* 49:D1496−D1501
21. Jayakodi M, Choi BS, Lee SC, Kim NH, Park JY, et al. 2018. Ginseng Genome Database: an open-access platform for genomics of *Panax ginseng*. *BMC Plant Biology* 18:62
22. Sakai H, Naito K, Takahashi Y, Sato T, Yamamoto T, et al. 2016. The *Vigna* genome server, '*VigGS*': a genomic knowledge base of the genus *Vigna* based on high-quality, annotated genome sequence of the azuki bean, *Vigna angularis* (Willd.) Ohwi & Ohashi. *Plant & Cell Physiology* 57:e2
23. Yu HJ, Baek S, Lee YJ, Cho A, Mun JH. 2019. The radish genome database (RadishGD): an integrated information resource for radish genomics. *Database* 2019:baz009
24. Plomion C, Aury JM, Amselem J, Leroy T, Murat F, et al. 2018. Oak genome reveals facets of long lifespan. *Nature Plants* 4:440−52
25. Wei T, van Treuren R, Liu X, Zhang Z, Chen J, et al. 2021. Whole-genome resequencing of 445 *Lactuca* accessions reveals the domestication history of cultivated lettuce. *Nature Genetics* 53:752−60
26. Wang X, Wu J, Liang J, Cheng F, Wang X. 2015. *Brassica* database (BRAD) version 2.0: integrating and mining Brassicaceae species genomic resources. *Database* 2015:bav093
27. Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, et al. 2014. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* 345:950−53
28. Byrne SL, Erthmann PØ, Agerbirk N, Bak S, Hauser TP, et al. 2017. The genome sequence of *Barbarea vulgaris* facilitates the study of ecological biochemistry. *Scientific Reports* 7:40728
29. Droc G, Martin G, Guignon V, Summo M, Sempéré G, et al. 2022. The banana genome hub: a community database for genomics in the Musaceae. *Horticulture Research* 9:uhac221
30. Zhou Y, Qiao Y, Ni Z, Du J, Xiong J, et al. 2021. GDS: a genomic database for strawberries (*Fragaria* spp.). *Horticulturae* 8:41
31. Tang Y, Zhang G, Jiang X, Shen S, Guan M, et al. 2023. Genome-wide association study of glucosinolate metabolites (mGWAS) in *Brassica napus* L. *Plants* 12:639
32. Feng X, Ma J, Liu Z, Li X, Wu Y, et al. 2022. Analysis of glucosinolate content and metabolism related genes in different parts of Chinese flowering cabbage. *Frontiers in Plant Science* 12:767898
33. Gamet-Payrastre L. 2006. Signaling pathways and intracellular targets of sulforaphane mediating cell cycle arrest and apoptosis. *Current Cancer Drug Targets* 6:135−45
34. Zhang XF, Liu PY, Zhang SJ, Zhao KL, Zhao WX. 2022. Principle and progress of radical treatment for locally advanced esophageal squamous cell carcinoma. *World Journal of Clinical Cases* 10:12804−11
35. Miękus N, Marszałek K, Podlacha M, Iqbal A, Puchalski C, et al. 2020. Health benefits of plant-derived sulfur compounds, glucosinolates, and organosulfur compounds. *Molecules* 25:3804
36. Fuentes F, Paredes-Gonzalez X, Kong AN T. 2015. Dietary glucosinolates sulforaphane, phenethyl isothiocyanate, indole-3-carbinol/3,3'-diindolylmethane: antioxidative stress/inflammation, Nrf2, epigenetics/epigenomics and *in vivo* cancer chemopreventive efficacy. *Current Pharmacology Reports* 1:179−96
37. Tanabe TS, Dahl C. 2022. HMS-S-S: a tool for the identification of Sulphur metabolism-related genes and analysis of operon structures in genome and metagenome assemblies. *Molecular Ecology Resources* 22:2758−74
38. Tanabe TS, Dahl C. 2023. HMSS2: an advanced tool for the analysis of sulphur metabolism, including organosulphur compound transformation, in genome and metagenome assemblies. *Molecular Ecology Resources* 23:1930−45
39. Bell L, Chadwick M, Puranik M, Tudor R, Methven L, et al. 2020. The *Eruca sativa* genome and transcriptome: a targeted analysis of sulfur metabolism and glucosinolate biosynthesis pre and postharvest. *Frontiers in Plant Science* 11:525102
40. Liao N, Hu Z, Miao J, Hu X, Lyu X, et al. 2022. Chromosome-level genome assembly of bunching onion illuminates genome evolution and flavor formation in *Allium* crops. *Nature Communications* 13:6690
41. Yan X, Chen S. 2007. Regulation of plant glucosinolate metabolism. *Planta* 226(6):1343−52