MCDB: A comprehensive curated mitotic catastrophe database for retrieval, protein sequence alignment, and target prediction

Le Zhang, Lei Zhang, Yue Guo, Ming Xiao, Lu Feng, Chengcan Yang, Guan Wang, Liang Ouyang

Innovation Center of Nursing Research, West China Biomedical Big Data Center, State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, College of Computer Science, and Collaborative Innovation Center of Biotherapy, Sichuan University, Chengdu 610065, China
Nursing Key Laboratory of Sichuan Province, West China Hospital, Sichuan University, Chengdu 610041, China
Pera Corporation Ltd., Beijing 100025, China

Received 12 January 2021; received in revised form 12 March 2021; accepted 6 May 2021

Abstract Mitotic catastrophe (MC) is a form of programmed cell death induced by mitotic process disorders, which is very important in tumor prevention, development, and drug resistance. Because rapidly increased data for MC is vigorously promoting the tumor-related biomedical and clinical study, it is urgent for us to develop a professional and comprehensive database to curate MC-related data. Mitotic Catastrophe Database (MCDB) consists of 1214 genes/proteins and 5014 compounds collected and organized from more than 8000 research articles. Also, MCDB defines the confidence level, classification criteria, and uniform naming rules for MC-related data, which greatly improves data reliability and retrieval convenience. Moreover, MCDB develops protein sequence alignment and target prediction functions. The former can be used to predict new potential MC-related genes and proteins, and the latter can facilitate the identification of potential target proteins of unknown MC-related compounds. In short,
1. Introduction

Mitotic catastrophe (MC), originally proposed by Molz Lisa et al. in 1989, was named for the first time by the International Nomenclature Committee on Cell Death in 2012. MC is a form of programmed cell death caused by the deregulation of mitotic process as an intrinsic onco-suppressive mechanism. Studies have identified that DNA lesions, mitotic defects, failure of cytokinesis could cause MC, and even tumor cells are more susceptible to this mitotic abnormality than normal cells. At present, in addition to photo and proton radiotherapy, there are a variety of chemotherapeutic drugs that could exert anti-tumor effects by inducing MC, covering microtubule regulators, CHK1 inhibitors, PARPs inhibitors, WEE1 inhibitors, PLKs inhibitors, and so on. Currently, with the in-depth exploration of relevant studies on MC, the significance of MC in tumor prevention, treatment, drug resistance and radiosensitivity gradually developed, which has attracted widespread attention from chemists to biologists in the fields of medicinal chemistry, molecular biology, and bioinformatics.

Recently, the rapid increased MC-related data (genes, proteins, and compounds) is greatly promoting MC-related drug discovery, synthesis, and repositioning. Thus, a lot of commonly used public databases, such as Public Medicine (PubMed) database, Public Chemistry (PubChem) database, Universal Protein Resource (UniProt), and Protein Data Bank (PDB), consist of a large amount data of MC-related genes, proteins, and compounds. However, previous databases did not specifically establish a confidence level and classification criteria for MC-related data, resulting in data reliable ambiguity and retrieval inconvenience. Also, previous databases did not develop such uniform naming rules for these MC-related data that severely restricted their accessibility and usability. Lastly, previous databases neither provide MC-related tools/functions to discover new MC-related genes, proteins, and compounds, nor assist us to understand the MC-related biological functions, signal transduction mechanisms, and biological processes.

In order to overcome these previous shortcomings, we develop a Mitotic Catastrophe Database (MCDB) with three major innovations: (1) MCDB is the first comprehensive database for MC-related data curation, which not only consists of 1214 genes/proteins and 5014 compounds collected and organized from more than 8000 research articles, but also data upload function of which can update the curated MC-related data. (2) MCDB defines the confidence level, classification criteria, and uniform naming rules for MC-related data, which significantly improves data reliability and retrieval convenience. (3) MCDB offers protein sequence alignment and target prediction functions. The former can be used for predicting new potential MC-related genes and proteins, and the latter can facilitate the identification of potential target proteins of unknown compounds for MC.

In general, MCDB provides such a proprietary, standardized, and comprehensive data retrieval and analysis platform that not only can facilitate the design and discovery of MC-based anti-tumor drugs, but also promote tumor biomedical and clinical treatment research in the distant future.

2. Materials and methods

2.1. Website development

MCDB is developed on the cloud Linux server (CentOS 7.5.1804), which employs Nginx (version 1.18) and SQLite (version 3.32.3) as the web and database server, respectively. The back and front end of the website are respectively based on the Django framework (version 3.0.8) and Bootstrap framework (version 4.4.1). Here, we distributed the website on http://www.combio-lezhang.online/MCDB/index_html/to provide open access (Fig. 1).

2.2. Data collection and preprocessing

Fig. 2 describes how to collect and organize MC-related 1214 genes/proteins and MC-related 5014 compounds from more than 8000 articles and Gene Ontology (GO) knowledge. As shown in Fig. 2A, we manually identified 188 MC-related genes/proteins from 1012 original publications by mining PubMed with the keyword mitotic catastrophe. Meanwhile, we identified 1126 MC-related genes/proteins in the GO knowledgebase by analyzing the items involved in the mitosis biological process. By intersecting the results from GO and PubMed databases, we have 1214 genes/proteins. Next, we unified gene names, protein names, synonyms, UniProt accessions by the standards of UniProt. For example, when the same gene/protein has multiple names and abbreviations, we annotated its gene name and protein name by the standard of UniProt, whereas putting its other names and abbreviations in the synonyms field. When different genes/proteins share the same name, we confirmed these genes/proteins by corresponding UniProt accessions.

As shown in Fig. 2B, we obtained 5014 compounds from more than 7000 published articles by text mining. Next, we used ChemBioDraw Ultra (version 14.0.0.117) software to draw compound structures, and then obtained the Simplified Molecular Input Line Entry Specification (SMILES) expression of compounds by SMILES conversion. Subsequently, we calculated the International Chemical Identifier (InChI) and the International Chemical Identifier hash (InChI Key) by using the International Union of Pure and Applied Chemistry (IUPAC) standard for SMILES.
2.3. Data confidence level and classification criteria

To improve the data reliability for MC-related genes and proteins, we define the confidence level criteria for MC-related genes and proteins in Table 1 and Fig. 2A. The first level (94 proteins, 7.74%), with literature support and GO items association; the second level (89 proteins, 7.33%), with literature support but GO items are not mentioned; the third level (106 proteins, 8.73%), with GO items association and homologous enzymes have been reported in the literature; the fourth level (925 proteins, 76.20%), GO items are indicated to be associated with mitosis which may be involved in MC.

To solve the unnecessary duplication and confusion caused by inconsistent descriptions of the compound’s effect on protein in literature, we defined classification criteria for compounds in Table 2 and Fig. 2B. Inhibitor, the molecule that can inhibit the function of proteins, but does not induce conformational change in the protein, includes inhibitor, antagonist, inactivator, destabilizing agent, and so on; activator, the molecule that can activate the function of proteins, but does not induce conformational change in the protein, includes activator, agonist, enhancer, stabilizing agent, and so on; allosteric regulator, the molecule that can induce conformational change in the protein is recognized as allosteric regulators.

2.4. The BLAST+ for protein sequence alignment

BLAST+ is performed to compare the similarity of multiple protein sequences. After the user submits the sequence in FASTA format, multiple sequence comparisons will be automatically carried out. Each entry in the library will be paired with the sequence submitted to calculate attributes including Length, Score, Expect, Identities, Positives, Gaps, which are detailed in Table 3. Generally, the Threshold of Expect should be less than or equal to $1 \times 10^{-5}$, in which $1 \times 10^{-5}$ matches would be expected to occur by chance.

2.5. The similarity for compound

Because similar molecular structures usually have the same/similar target proteins and similar biological properties, we employ the Tanimoto coefficient to compute the two-dimension similarity score between different compound structures, described by Eq. (1):

$$T_\text{Tanimoto} = \frac{X \cap Y}{X \cup Y}$$

Here, $X$ and $Y$ respectively represent the binary data of the compound molecular fingerprint. The value range of Tanimoto coefficient is $[0,1]$, and the threshold is generally set to $0.85$.

We also employ SHAFTS to compute the three-dimension Hybrid Score between different structures, described by Eqs. (2)–(6):

$$V_{AB} = \sum_{i \in A} \sum_{j \in B} \int d\vec{r} p_i(\vec{r}) p_j(\vec{r})$$

$$= \sum_{i \in A} \sum_{j \in B} p_i p_j \exp \left( -\frac{\gamma_i \gamma_j}{\gamma_i + \gamma_j} \right) \left( \frac{\pi}{\gamma_i + \gamma_j} \right)^{3/2}$$

Here, $i$ and $j$ respectively represent the atom of A and B, $d_{ij}$ is the interatomic distance between atom $i$ and $j$, $\gamma$ is the width of a Gaussian relevant with van der Waals radii.

$$\text{ShapeScore} = \frac{V_{AB}}{V_A V_B}$$

The final ShapeScore is normalized to the range between 0 and 1.

$$F_{AB} = \sum_{f \in F} \sum_{i \in A} \sum_{j \in B} \exp \left( -2.5 \left( \frac{d_{ij}}{R_f} \right)^2 \right)$$
Here \( i \) and \( j \) respectively represent the feature point of A and B with the same type \( f \), \( d_{ij} \) is the distance between point \( i \) and \( j \), and \( R_f \) is the overlap tolerance with a default value of 0.8 Å.

FeatureScore = \( \frac{F_{AB}}{\sqrt{F_A F_B}} \)  

The final FeatureScore is normalized to the range between 0 and 1.

HybridScore = ShapeScore + FeatureScore  

HybridScore is the sum of ShapeScore and FeatureScore. HybridScore is scaled to the range between 0 and 2, and compounds have a certain similarity when HybridScore is equal or greater than 1.

2.6. The ratio for compounds and structures

We usually will focus on such proteins that can form compounds or have structures, since they are very useful for pharmacological and pharmacochemical research. Here, we define the ratios for compounds (\( \text{Ratio}_c \)) and structures (\( \text{Ratio}_s \)) to help us to show the distribution of these types of proteins in MCDB.

| Table 1 | Confidence levels of genes and proteins. |
|---------|------------------------------------------|
| Confidence level | Confidence level criterion | Number of genes and proteins |
| 1       | The gene or protein is both supported by literature and associated with GO item. | 94 |
| 2       | The gene or protein is supported by literature but not associated with GO item. | 89 |
| 3       | The gene or protein is associated with GO item, while gene or protein is not supported by literature but their homologous enzymes have been reported in the literature. | 106 |
| 4       | The gene or protein is only associated with GO item. | 925 |
3.1. Data statistics

As shown in Fig. 2, MCDB collected and organized MC-related 1214 genes/proteins and 5014 compounds from more than 8000 original articles and Gene Ontology (GO) knowledge. And due to the inconvenience of retrieval and use caused by irregular abbreviations and duplicate names of genes and proteins, we have unified all gene names, protein names, synonyms, UniProt accessions based on standards of UniProt. When the same gene/protein has multiple names and abbreviations, we annotated its gene name and protein name by the standard of UniProt, whereas putting its other names and abbreviations in the synonyms field. When different genes/proteins share the same name, we confirmed these genes/proteins by corresponding UniProt accessions.

To improve the MC-related genes and proteins data reliability, storage and retrieval convenience, these genes and proteins are divided into four confidence levels according to the evidence reliability of the correlation with MC (Fig. 2 and Table 1). The first confidence level account for 94 entries (7.74%), the second level account for 89 entries (7.33%), the third confidence level account for 106 entries (8.73), and the fourth confidence level is the most with 925 entries (76.20%) (Fig. 3 and Table 1). Also, MCDB collected 5014 targeted compounds for the first three confidence levels of MC-related genes and compounds. And in order to eliminate the unnecessary duplication and confusion caused by inconsistent descriptions of the compound’s effect on protein from different articles, we divided these compounds into 3 classifications as Table 2. Fig. 3 shows that there are 4630 inhibitors, 319 activators, and 65 allosteric regulators, which account for 92.34%, 6.36%, and 1.3% of the total compound entries, respectively.

3.2. Data visualization

The data visualization module is designed to visualize the data, which can visually show users the statistics and distribution of MC-related genes, proteins, and compounds in the database. After clicking the “Data Visualization” link of the homepage, users can not only use this module to visualize the data distribution for MCDB curated MC-related genes/proteins, and compounds under different confidence levels or classifications, but also obtain the data distribution for such proteins that can form compounds or have structures.

Fig. 4A shows the data distribution for MC-related genes and proteins under different confidence levels, which is described in Table 1 and Fig. 2A. Here, the first, second, third, and fourth confidence levels have 94 entries (7.7%), 89 entries (7.33%), 106 entries (8.73%), and 925 entries (76.20%), respectively. Fig. 4B shows the data distribution for MC-related compounds under different classifications, which is described in Table 2 and Fig. 2B. Here, the number (percentage) of the inhibitor, activator, and allosteric regulator is 4630 entries (92.34%), 319 entries (6.36%), and 65 entries (1.30%), respectively.

As previous studies, researchers always focus on such proteins that can form compounds or have structures. Fig. 4C and D shows distributions of genes and proteins with compounds and structures respectively. For example, after we input “30” and click the “Submit” button at the top of Fig. 4C, the bottom of Fig. 4C show Ratio_c discussed by Eq. (7) for the top 30 proteins that can form compounds. The purple box of Fig. 4C shows the
corresponding gene name “GSK3B” and its 263 compounds, when the mouse is on this circle.

For another example, after we input “30” and click the “Submit” button at the top of Fig. 4D, the bottom of Fig. 4D shows the Ratio_s discussed by Eq. (8) for the top 30 proteins that have structures. The purple box of Fig. 4D shows the corresponding gene name “INS” and its 284 structures, when the mouse is on this circle.

3.3. MC-related Gene and Protein Search

The MC-related Gene and Protein Search module is designed to facilitate users to query MC-related genes and proteins, and provides users with standardized information, including UniProt accession, gene name, GO identifier, GO term, protein name, synonyms, confidence level, and PubMed identifier (PMID) of MC relevant literature. And we also provide a hyperlink for data download so that users can obtain all the data of MC-related genes and proteins. This module can help users determine which genes and proteins are related to MC and obtain standardized detailed information of the MC-related genes and proteins, so as to further study the MC-related biological processes and molecular mechanisms they participated in. After clicking the “MC-related Gene and Protein Search” link of the homepage (Fig. 1), we can query the information for MC-related genes and proteins by inputting the UniProt Accession or Gene name (Fig. 5A).

For example, after we input Gene Name “BRD4” in the search box and click the “Submit” button in Fig. 5A and B shows its UniProt Accession is “O60885”, Gene Name is “BRD4”, GO Identifier is “GO:0043123”, GO Term is “positive regulation of I-kappaB kinase/NF-kappaB signaling”, Protein Name is “Bromodomain-containing protein 4”, Synonyms is “HUNK1”, Confidence Level is “1” (described in Table 1), and MC-PMID is “31199520,19596781”. When the mouse hovers over the help marks next to the “GO Identifier”, “Synonyms”, “Confidence

3097

Figure 3

Data statistics of MCDB.

Figure 4

Data visualization. (A) The data distribution for MC-related genes and proteins under different confidence levels; (B) The data distribution for MC-related compounds under different classifications; (C) The data distribution for the top 30 proteins that can form compounds. The Y-axis is the Ratio_c defined by Eq. (7); (D) The data distribution for the top 30 proteins that have structures. The Y-axis is the Ratio_s defined by Eq. (8). Here, the circles with different colors represent different confidence levels.
3.4. MC-related Compound Search

The MC-related Compound Search module is to facilitate users to query MC-related compounds with detailed information, including UniProt accession, gene name, confidence level, SMILES, InChI, InChI Key, molecular formula, molecular weight, classification. And a hyperlink for data download was also provided in this module, so that users can obtain all the data of MC-related compounds. This module can help users determine if compounds are related to MC and obtain detailed information about the MC-related compounds, so as to further understand what roles of these compounds play to regulate MC and MC-related genes/proteins. After clicking the “MC-related Compound Search” link of the homepage (Fig. 1), we can query the information for MC-related compounds by inputting SMILES, InChI, InChI Key, UniProt Accession, or Gene Name (Fig. 6A). Additionally, users can fill in the desired number before query submission; otherwise, ten records will be displayed per page by default.

For example, after inputting SMILES “O=C1(CO) (CO)N2CCC1CC2” and clicking the “Submit” button, Fig. 6B shows its UniProt Accession is “P04637”, Gene Name is “TP53”, Confidence Level is “1” (described in Table 1), SMILES is “O=C1(C)(CO)N2CC(C1)CC2”, InChI is “1S/C9H15NO3/c11-5-9(6-12)8(13)7-1-3-10(9)4-2-7/h7,11-12H,1-6H2”, InChI Key is “RFBVBRVVOPAFSS-UHFFFAOYSA-N”, Molecular Formula is “C9H15NO3”, Molecular Weight is “185.22”, Classification is “activator” (described in Table 2). When the mouse hovers over the help marks next to the “Confidence Level” or “Classification”, MCDB will automatically detail the related explanations. Also, users can enter P04637 (UniProt accession) or TP53 (gene name) to query all inhibitors, activators, and allosteric regulators. Notably, the structures can be displayed as Fig. 6C, and the SMILES, InChI, and InChI Key of each compound in this module can also increase the convenience for users to use external software or links for extended queries and research. Additionally, users can click the sorting markers next to “Molecular Weight” or “Classification” to sort the retrieval results.

3.5. PDB search

The PDB Search module is to facilitate users to query structure information of MC-related proteins with UniProt accession, gene name, confidence level, PDB code, released date, method. Users can obtain the structural information of MC-related proteins by PDB Search module to facilitate further biophysics research and structures-based drug design and discovery for protein. After clicking the “PDB Search” link of the homepage (Fig. 1), we can query the structures for MC-related proteins by inputting UniProt Accession (Fig. 7A).

For example, after inputting UniProt Accession “Q969H0”, the first row of Fig. 7B shows that UniProt Accession is “Q969H0”, Gene Name is “FBXW7”, Confidence Level is “2” (described in Table 1), PDB Code is “2OVP”, Released Date is “2007-04-24”, Method is “X-RAY DIFFRACTION 2.9 Å”. It is noted that “Released Date” refers to the released time of the protein structure and the “Method” refers to the way used to analyze the protein structure. Additionally, Fig. 7B shows the PDB Codes such as “2OVP”, “2OVR”, “2OVQ”, “5V4B”, “5IBK” have hyperlinks. Once users click these hyperlinks, they will go to the PDB24 website for further research. In this module, users can sort the results according to the priority of “PDB Code” or “Released Date”.

Figure 5

MC-related Gene and Protein Search. (A) The search interface; (B) The result page for MC-related Gene and Protein.
3.6. Protein sequence alignment

The Protein Sequence Similarity Alignment module is to compare the protein sequence submitted by users with protein sequences in the database. Users can comprehensively consider the results by both “Confidence Levels” defined in Table 1 and “Scores” shown in Fig. 8. Since the proteins with similar sequences may have similar functions and structures, users can use protein sequence...
alignment results to carry out protein structure simulation, function prediction, evolutionary research, especially the prediction of novel potential MC-related genes and proteins. After clicking the “Protein Sequence Alignment” link of the homepage (Fig. 1), we can compute the similarity between the submitted protein sequence and the MC-related protein sequences of MCDB by inputting the protein sequence into Protein Sequence Alignment module (Fig. 8A).

For example, after submitting the sequence of CDK12 in Fig. 8A and B shows the protein sequence alignments with key attributes (Table 3) with decreasing order of attribute “Score”. The first row of Fig. 8B shows that UniProt Accession is “Q14004”, Gene Name is “CDK13”, Confidence Level is “3”, Length is “1512”, Score is “803 bits (2074)”, Expect is “0.0”, Identities is “408/596 (68%)”, Positives is “466/596 (78%)”, Gaps is “31/596 (5%)”. When the mouse hovers over the help marks next to the “Confidence Level”, “Length, Score”, “Expect”, “Identities”, “Positives”, or “Gaps”, MCDB will automatically detail the related explanations. And users can sort results according to the priority of “Score”, “Expect”, “Identities”, “Positives”, or “Gaps”, which could help us better analyze the alignment results. Also, Fig. 8B shows that each UniProt Accession has a hyperlink. Once we click the hyperlink of UniProt Accession, MCDB can automatically run the “MC-related Gene and Protein Search” module to retrieve the detailed gene and protein information. Additionally, users can click the “download” button to have full data for further research (Fig. 8B).

### 3.7. Target prediction

The Target Prediction module is to compare the compound submitted by users with compounds in the database to predict the potential targets. This module provides two target prediction options for users. One is Tanimoto Score\(^4\), which is used to predict targets based on molecular fingerprints, the other is SHAFTS, which is used to predict targets based on 3D structure\(^5\).

The Target Prediction module not only can help users identify potential target proteins of unknown compounds and predict their impacts on MC, but also facilitate the ligand-based drug design and discovery based on MC. After clicking the “Target Prediction” link of the homepage (Fig. 1), we can compute the similarity between the submitted compound and the curated MC-related compounds of MCDB by inputting SMILES into Target Prediction module (Fig. 9A).

For example, after we submit an example of SMILES and choose the SHAFTS(3D), the first row of Fig. 9B shows that Uniprot Accession is “P42345”, Gene Name is “MTOR”, Classification is “inhibitor”, Confidence Level is “2” (described in Table 1), Hybrid Score is “1.045”, Shape Score is “0.7931”, Feature Score is “0.2521”, SMILES is “CCOC(O)C1=CC=CC=CC=NCC2=CC=C(O)OC3=CC=C(O)OC4=CC=C(O)OC”. When the mouse hovers over the help marks next to the “Classification”, “Confidence Level”, “Hybrid Score”, “Shape Score”, or “Feature Score”, MCDB will automatically detail the related explanations. And users can sort the results according to the priority of “Uniprot Accession”, “Gene Name”, “Classification”, “Confidence Level”, “Hybrid Score”, “Shape Score”, or “Feature Score”, MCDB will automatically detail the related explanations. And users can sort the results according to the priority of “Uniprot Accession”, “Gene Name”, “Classification”, “Confidence Level”, “Hybrid Score”, “Shape Score”, or “Feature Score”. Figure 8  Protein Sequence Alignment. (A) The submit interface; (B) The result page.
“Hybrid Score”, “Shape Score”, or “Feature Score”, which can help us better analyze the prediction results.

For another example, after we submit an example of SMILES and choose the Tanimoto Coefficient(2D), the first row of Fig. 9C shows that UniProt Accession is “P61073”, Gene Name is “CXCR4”, Classification is “inhibitor”, Confidence Level is “2”, Similarity Score is “0.5469”, SMILES is “C(CC(CCNCCCN)CNCCCN)”. Here, users can sort results according to the priority of “Uniprot Accession”, “Gene Name”, “Classification”, “Confidence Level”, or “Similarity Score”, which can help us better analyze the prediction results.

3.8. Upload

The Upload module is designed for the constantly increasing experimental data, users can upload MC-related genes, proteins and compounds which are not included in the database. And then, we can maintain the database according to the influx of large amounts of MC-related data. After clicking the “Upload” link of the homepage (Fig. 1), the Upload module allows users to upload the candidate MC-related genes, proteins, and compounds onto MCDB. We employ Fig. 10A to upload the candidate MC-related gene and protein data onto MCDB by filling the UniPort Accession, Gene Name, Evidence, and Description. We employ Fig. 10B to upload the candidate MC-related compound onto MCDB by filling SMILES, Uniprot Accession, Gene Name, Evidence, and Description. The MCDB curator will validate uploaded data periodically. And then, the confirmed data will be saved in MCDB by assigning a corresponding confidence and classification level, which are described in Tables 1 and 2.

4. Discussion and conclusions

Currently, MC-related data is increasing very rapidly because of its importance in tumor prevention, treatment, and drug resistance. However, to the best of our knowledge, there is no such a specialized database that can facilitate the retrieval and analysis for these MC-related data. Therefore, we build up a Mitotic Catastrophe Database (MCDB) for extensive MC research (Fig. 1).

As indicated by Fig. 2, we already collected such a large amount of MC-related data from more than 8000 articles that could offer users a comprehensive MC-related database for further
Moreover, the upload module of MDCB (Fig. 10) could help us periodically update the rapidly increased MC-related data. Compared to previous public databases that curate MC-related data, MCDB defines confidence level and classification (Tables 1 and 2) for MC-related data, which not only can effectively reduce the duplication and confusion resulting from inconsistent descriptions, but also can help us comprehensively understand the reliability and availability of similarity score for Protein Sequence Alignment and Target Prediction modules (Figs. 8 and 9).

Described by Figs. 4–10, MCDB offers seven functional modules to search, update, and analyze the MC-related data. And the Protein Sequence Alignment module (Fig. 8) can contribute to protein structure simulation, function prediction, evolutionary research, and especially for new potential MC-related protein prediction by computing the similarity between protein sequences. The Target Prediction module (Fig. 9) can help us to explore the potential target proteins of unknown compounds, their possible effects on MC prediction, and ligand-based drug design by computing the similarity between compounds.

Although MCDB already supplements MC’s knowledge gaps in the database field and offer specific MC-related online service, it can neither automatically update MC-related data, nor real-time compute the similarity between MC-related compounds.

To overcome these shortcomings, we will employ natural language processing technology to improve manual update efficiency and use high-performance computing to reduce the processing time for similarity computing. Finally, we will make MDCB as a highly integrated web-based MC-related data platform by integrating more advanced bioinformatics applications and algorithms in the future.

Acknowledgments

This work was supported by grants from National Natural Science Foundation of China (Grant Nos. 81803755 and 81922064), National Science and Technology Major Project (Grant No. 2018ZX10201002, China), China Postdoctoral Science Foundation (2018M640926 and 2020M673221), and Sichuan University Postdoctoral Research and Development Foundation (2020SCU12062 and 2020SCU12056, China).

Author contributions

Guan Wang and Liang Ouyang conceived the project and supervised the project. Le Zhang and Lei Zhang Zhu summed up the literature and drafted the manuscript. Yue Guo and Ming Xiao were involved in drawing the figures. Lu Feng and Chengcan Yang collected and organized the genes, proteins and compounds. Guan Wang, Le Zhang and Lei Zhang proofread the genes, proteins and compounds. Guan Wang and Liang Ouyang revised the manuscript. All authors approved the final manuscript.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

1. Molz L, Booher R, Young P, Beach D. CDC2 and the regulation of mitosis: six interacting mcs genes. Genetics 1989;122:773–82.
2. Mc Gee MM. Targeting the mitotic catastrophe signaling pathway in cancer. Mediat Inflamm 2015;2015:146282.
3. Ke B, Tian M, Li J, Liu B, He G. Targeting programmed cell death using small-molecule compounds to improve potential cancer therapy. Med Res Rev 2016;36:983–1035.
4. Sia J, Szymy R, Hau E, Gee HE. Molecular mechanisms of radiation-induced cancer cell death: a primer. Front Cell Dev Biol 2020;8:41.
5. Vitovcova B, Srarkova V, Rudolf K, Rudolf E. Biology of glioblastoma multiforme—exploration of mitotic catastrophe as a potential treatment modality. Int J Mol Sci 2020;21:5324.
6. Arnst KE, Banerjee S, Chen H, Deng S, Hwang DJ, Li W, et al. Current advances of tubulin inhibitors as dual acting small molecules for cancer therapy. Med Res Rev 2019;39:1398–426.
7. Warren NHJ, Eastman A. Comparison of the different mechanisms of cytotoxicity induced by checkpoint kinase I inhibitors when used as
single agents or in combination with DNA damage. Oncogene 2020;39:1389–401.

8. Hatchi E, Livingston DM. Opening a door to PARP inhibitor-induced lethality in HR-proficient human tumor cells. Cancer Cell 2020;37:139–40.

9. Deneka AY, Einarson MB, Bennett J, Nikonova AS, Elmekawy M, Zhou Y, et al. Synthetic lethal targeting of mitotic checkpoints in HPV-negative head and neck cancer. Cancers (Basel) 2020;12:306.

10. Zhang Z, Zhang G, Kong C. Targeted inhibition of Polo-like kinase 1 by a novel small-molecule inhibitor induces mitotic catastrophe and apoptosis in human bladder cancer cells. J Cell Mol Med 2017;21:758–67.

11. Liu L, Charville GW, Cheung TH, Yoo B, Santos PJ, Schroeder M, et al. Impaired Notch signaling leads to a decrease in p53 activity and mitotic catastrophe in aged muscle stem cells. Cell Stem Cell 2018;23:544–56.e4.

12. Drápeľa S, Khirsariya P, van Weerden WM, Fedr R, Suchanová T, Bázová D, et al. The CHK1 inhibitor MU380 significantly increases the sensitivity of human docetaxel-resistant prostate cancer cells to gemcitabine through the induction of mitotic catastrophe. Mol Oncol 2020;14:2487–503.

13. Yang L, Shen C, Pettit CJ, Li T, Hu AJ, Miller ED, et al. Wee1 kinase inhibitor AZD1775 effectively sensitizes esophageal cancer to radiotherapy. Clin Cancer Res 2020;26:3740–50.

14. Forey R, Poveda A, Sharma S, Barthe A, Padioleau I, Renard C, et al. Mec1 is activated at the onset of normal S phase by low-dNTP pools impeding DNA replication. Mol Cell 2020;78:396–410.e4.

15. Zhang L, Fu C, Li J, Zhao Z, Hou Y, Zhou W, et al. Discovery of a ruthenium complex for the theranosis of glioma through targeting the mitochondrial DNA with bioinformatic methods. Int J Mol Sci 2019;20:4643.

16. Duan Y, Liu W, Tian L, Mao Y, Zhang S. Using game theory to investigate the epigenetic control mechanisms of embryo development: comment on: “Epigenetic game theory: how to compute the epigenetic control of maternal-to-zygotic transition” by Qian Wang et al. Philos Trans R Soc Lond B Biol Sci 2017;372:140–2.

17. Zhang L, Zhang S. Using game theory to investigate the epigenetic control mechanisms of embryo development: comment on: “Epigenetic game theory: how to compute the epigenetic control of maternal-to-zygotic transition” by Qian Wang et al. Phys Life Rev 2017;20:140–2.

18. Red Hat Enterprise Linux. CentOS Linux. 2018. Available from: https://www.centos.org.

19. F5 incorporation. Nginx. 2020. Available from: https://www.nginx.com.

20. SQLite Consortium. SQLite release 3.32.3 on 2020-06-18. 2020. Available from: https://www.sqlite.org.

21. Django Software Foundation. Django 3.0.8 release notes. 2020. Available from: https://djangoproject.com.

22. Bootstrap Team. Bootstrap 4.1.1. 2019. Available from: https://getbootstrap.com.

23. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.

24. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–15.

25. Deng Y, Zhu L, Cai H, Wang G, Liu B. Autophagic compound database: a resource connecting autophagy-modulating compounds, their potential targets and relevant diseases. Cell Prog 2018;51:e12403.

26. Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, et al. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29:3094–3095.

27. Xiao M, Yang X, Yu J, Zhang L. CGIDLA: developing the Web Server for CpG island related density and LAUPs (lineage-associated underrepresented permutations) study. IEEE ACM Trans Comput Biol Bioinform 2020;17:2148–54.

28. Zhang L, Bai W, Yuan N, Du Z. Comprehensive benchmarking applications for detecting copy number variation. PLoS Comput Biol 2019;15:e1007069.

29. Zhang L, Dai Z, Yu J, Xiao M. CpG-island-based annotation and analysis of human housekeeping genes. Briefings Bioinform 2021;22:515–25.

30. Fassler J, Cooper P. BLAST glossary. 2011. Available from: https://BLAST.NCBI.nih.gov/Blast.cgi.

31. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–15.

32. Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, et al. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29:3094–3095.

33. Xiao M, Yang X, Yu J, Zhang L. CGIDLA: developing the Web Server for CpG island related density and LAUPs (lineage-associated underrepresented permutations) study. IEEE ACM Trans Comput Biol Bioinform 2020;17:2148–54.

34. Zhang L, Bai W, Yuan N, Du Z. Comprehensive benchmarking applications for detecting copy number variation. PLoS Comput Biol 2019;15:e1007069.

35. Zhang L, Dai Z, Yu J, Xiao M. CpG-island-based annotation and analysis of human housekeeping genes. Briefings Bioinform 2021;22:515–25.

36. Fassler J, Cooper P. BLAST glossary. 2011. Available from: https://BLAST.NCBI.nih.gov/Blast.cgi.

37. Casamassima C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinf 2009;10:421.

38. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.

39. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–15.

40. Zhang L, Bai W, Yuan N, Du Z. Comprehensive benchmarking applications for detecting copy number variation. PLoS Comput Biol 2019;15:e1007069.

41. Zhang L, Dai Z, Yu J, Xiao M. CpG-island-based annotation and analysis of human housekeeping genes. Briefings Bioinform 2021;22:515–25.

42. Fassler J, Cooper P. BLAST glossary. 2011. Available from: https://BLAST.NCBI.nih.gov/Blast.cgi.

43. Casamassima C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinf 2009;10:421.
50. Butina D. Unsupervised data base clustering based on daylight’s fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets. *J Chem Inf Comput Sci* 1999;39:747–50.

51. Dunkel M, Günther S, Ahmed J, Wittig B, RJNar Preisner. SuperPred: drug classification and target prediction. *Nucleic Acids Res* 2008;36: W55-W9.

52. Xu HY, Zhang YQ, Liu ZM, Chen T, Lv CY, Tang SH, et al. ETCM: an encyclopaedia of traditional Chinese medicine. *Nucleic Acids Res* 2019;47: D1976–82.

53. Liu X, Jiang H, Li H. SHAFTS: a hybrid approach for 3D molecular similarity calculation. 1. Method and assessment of virtual screening. *J Chem Inf Model* 2011;51:2372–85.

54. Capitani G, Duarte JM, Baskaran K, Bliven S, Somody JC. Understanding the fabric of protein crystals: computational classification of biological interfaces and crystal contacts. *Bioinformatics* 2016;32:481–9.

55. Pires DEV, Veloso WNP, Myung Y, Rodrigues CHM, Silk M, Rezende PM, et al. EasyVS: a user-friendly web-based tool for molecule library selection and structure-based virtual screening. *Bioinformatics* 2020;36:4200–2.

56. Wang JH, Zhao LF, Wang HF, Wen YT, Jiang KK, Mao XM, et al. GenCl iP 3: mining human genes’ functions and regulatory networks from PubMed based on co-occurrences and natural language processing. *Bioinformatics* 2020;36:1973–5.

57. Shi LX, Huang ZX, Hu N, Yang ZZ, Zhang L. Integrating semantic query function into D-NetWeaver. *J Med Imag Health In* 2015;5:982–6.

58. Jiang B, Struthers A, Sun Z, Feng Z, Zhao X, Zhao K, et al. Employing graphics processing unit technology, alternating direction implicit method and domain decomposition to speed up the numerical diffusion solver for the biomedical engineering research. *Int J Numer Meth Bio* 2011;27:1829–49.

59. Jiang BN, Dai WZ, Khalil A, Carey M, Zhou XB, Zhang L. Novel 3D GPU based numerical parallel diffusion algorithms in cylindrical coordinates for health care simulation. *Math Comput Simulat* 2015;109:1–19.

60. Zhang L, Jiang B, Wu Y, Strouthos C, Sun PZ, Su J, et al. Developing a multiscale, multi-resolution agent-based brain tumor model by graphics processing units. *Theor Biol Med Model* 2011;8:46.

61. Zhang L, Jiang B, Wu Y, Strouthos C, Sun PZ, Su J, et al. Developing a multiscale, multi-resolution agent-based brain tumor model by graphics processing units. *Theor Biol Med Model* 2011;8:46.