

Essential proteins identification method based on four-order distances and subcellular localization information

Pengli Lu(卢鹏丽), Yu Zhong(钟雨), and Peishi Yang(杨培实)
Citation: Chin. Phys. B, 2024, 33 (**1**): 018903. DOI: 10.1088/1674-1056/acd7ca
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Identifying influential spreaders in social networks: A two-stage quantum-behaved particle swarm optimization with Lévy flight

Pengli Lu(卢鹏丽), Jimao Lan(揽继茂), Jianxin Tang(唐建新), Li Zhang(张莉), Shihui Song(宋仕辉), and Hongyu Zhu(朱虹羽)
Chin. Phys. B, 2024, 33 (**1**): 018901. DOI: 10.1088/1674-1056/acd3e0

Robustness of community networks against cascading failures with heterogeneous redistribution strategies

Bo Song(宋波), Hui-Ming Wu(吴惠明), Yu-Rong Song(宋玉蓉), Guo-Ping Jiang(蒋国平),Ling-Ling Xia(夏玲玲), and Xu Wang(王旭)
Chin. Phys. B, 2023, 32 (**9**): 098905. DOI: 10.1088/1674-1056/acd9c3

Identification of key recovering node for spatial networks

Zijian Yan(严子健), Yongxiang Xia(夏永祥), Lijun Guo(郭丽君), Lingzhe Zhu(祝令哲), Yuanyuan Liang(梁圆圆), and Haicheng Tu(涂海程)
Chin. Phys. B, 2023, 32 (**6**): 068901. DOI: 10.1088/1674-1056/acb75f

AG-GATCN: A novel method for predicting essential proteins

Peishi Yang(杨培实), Pengli Lu(卢鹏丽), and Teng Zhang(张腾)
Chin. Phys. B, 2023, 32 (**5**): 058902. DOI: 10.1088/1674-1056/acb9f9

Biased random walk with restart for essential proteins prediction

Pengli Lu(卢鹏丽), Yuntian Chen(陈云天), Teng Zhang(张腾), and Yonggang Liao(廖永刚)
Chin. Phys. B, 2022, 31 (**11**): 118901. DOI: 10.1088/1674-1056/ac7a17

Essential proteins identification method based on four-order distances and subcellular localization information

Pengli Lu(卢鹏丽)[1,†], Yu Zhong(钟雨)[1,2,‡], and Peishi Yang(杨培实)[1]

[1] *School of Computer and Communication, Lanzhou University of Technology, Lanzhou* 730050, *China*
[2] *School of Tianmen Vocational College, Tianmen* 431700, *China*

(Received 1 February 2023; revised manuscript received 28 April 2023; accepted manuscript online 23 May 2023)

Essential proteins are inseparable in cell growth and survival. The study of essential proteins is important for understanding cellular functions and biological mechanisms. Therefore, various computable methods have been proposed to identify essential proteins. Unfortunately, most methods based on network topology only consider the interactions between a protein and its neighboring proteins, and not the interactions with its higher-order distance proteins. In this paper, we propose the DSEP algorithm in which we integrated network topology properties and subcellular localization information in protein–protein interaction (PPI) networks based on four-order distances, and then used random walks to identify the essential proteins. We also propose a method to calculate the finite-order distance of the network, which can greatly reduce the time complexity of our algorithm. We conducted a comprehensive comparison of the DSEP algorithm with 11 existing classical algorithms to identify essential proteins with multiple evaluation methods. The results show that DSEP is superior to these 11 methods.

Keywords: protein–protein interaction (PPI) network, essential proteins, four-order distances, subcellular localization information

PACS: 89.75.–k, 87.23.Cc, 87.23.Ge **DOI:** 10.1088/1674-1056/acd7ca

1. Introduction

Essential proteins are an integral part of cell life and play an important role in maintaining the fundamental functions of organisms, whose absence would lead to the death or sterility of the organism.[1] Therefore, accurate identification of essential proteins has an important place in biomedical research. Traditional essential proteins identification relies on biological experimental methods, such as single gene knockout,[2] RNA interference[3] and conditional gene knockout,[4] which have a high degree of accuracy but often require a great deal of time and a high amount of money.

With the development of high throughput technologies, a huge amount of PPI network data has been mined, which makes it possible to use computational methods to identify essential proteins. Many methods based on the centrality of PPI network topology properties have also been proposed to identify essential proteins, such as degree centrality (DC),[5] betweenness centrality (BC),[6] eigenvector centrality (EC),[7] closeness centrality (CC),[8] subgraph centrality (SC),[9] local average center (LAC),[10] network centrality (NC),[11] and so on. However, there is a large amount of noisy data in the PPI network, which makes these centrality methods for identifying essential proteins less accurate. To overcome this problem, many researchers have tried to use biological information to enhance accuracy in identifying essential proteins. Currently, the main biological data used by researchers includes gene ontology (GO) terms,[12] subcellular localization information,[13] gene expression sequence,[14,15] and protein complexes information.[16] For example, Hsing *et al.*[12] developed a central protein classifier using GO terms to identify the central protein. The LDS method for identifying essential proteins was proposed by Li *et al.*,[17] which combines local fractal dimension and subcellular localization information. Li *et al.* and Tang *et al.* proposed the Pec[14] method and the Wdc[18] method using the topological properties of PPI network and gene expression data, respectively. Li *et al.*[16] proposed the UC method by considering the number of proteins that are present in the complexes. Lu *et al.*[19] proposed the CENC method for essential proteins identification studies by using protein complex information and the node clustering coefficients. Although these methods have yielded good results, it is still challenging to better combine the topological properties of PPI network and biological information to improve the performance of identifying essential proteins. As an optimization algorithm, the random walks method is also widely used to identify essential proteins. For example, Lei *et al.*[20] proposed the RWEP algorithm, which uses the random walks method to integrate the topological properties of PPI networks and four biological information to identify essential proteins, and obtained an excellent experimental result. In addition, the random walks method has been applied to link prediction,[21] recommender systems,[22] ranking,[23] community detection,[24] and transmission dynamics.[25]

Although these methods have good accuracy and have

†Corresponding author. E-mail: lupengli88@163.com
‡Corresponding author. E-mail: 13026326063@163.com

© 2024 Chinese Physical Society and IOP Publishing Ltd

contributed to the development of the field of bioinformatics, however, they usually only consider the interaction properties between nodes and neighboring nodes in a PPI network, while ignoring the interactions between nodes at higher-order distances. To break these limitations, in this paper a new method called DSEP is proposed that combines the four-order distances and subcellular localization information of a PPI network to identify essential proteins. In DSEP, we first construct a topological information matrix based on the four-order distances of PPI network, we then construct a subcellular localization information matrix based on the four-order distances, and we finally use the random walks method to combine the topological information matrix of PPI network and the subcellular localization information matrix to identify essential proteins. To evaluate the properties of the DSEP algorithm, we used three PPI network data sets (i.e., BioGRID, YHQ, and YMBD). The experiment revealed that DSEP has an outstanding performance in identifying essential proteins compared to 11 comparison methods.

2. Methods

The PPI network can be expressed as $G = (V(G), E(G))$, where $V(G) = \{v_1, v_2, \ldots, v_n\}$ denotes the set of vertices in the PPI network and $E(G) = \{e_1, e_2, \ldots, e_m\}$ denotes the set of edges. The distance between vertices v_i and v_j is denoted as $d_{i,j}$.

The DSEP algorithm for scoring proteins in PPI networks has the following three main steps, and the algorithm flowchart is shown in Fig. 1.

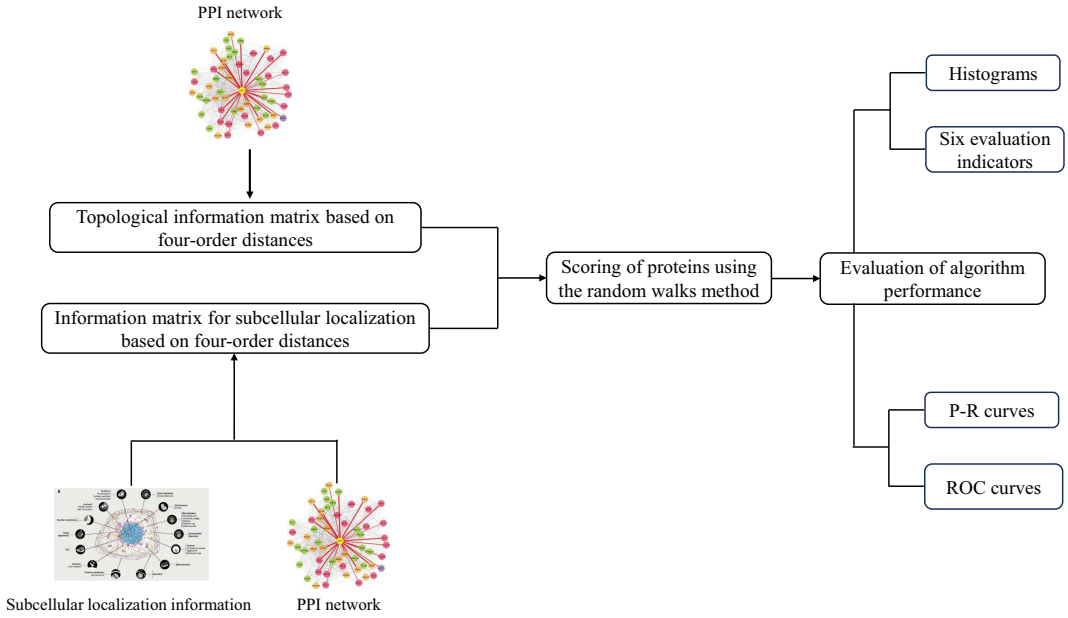


Fig. 1. Flowchart of DSEP.

Step 1 Calculate the four-order distances-based topological information matrix in a PPI network.

Step 2 Calculate the information matrix of subcellular localization based on four-order distances in a PPI network.

Step 3 Based on these two similarity matrices, the proteins are scored using the random walks method.

In Ref. [26], the authors report that the number of paths with distance k between nodes v_i and v_j in graph G is $A^k_{i,j}$ ($\boldsymbol{A}$ denotes the adjacency matrix of graph G). Therefore, we proposed a method to calculate the finite-order distance matrix of the graph G (see step 8 of Algorithm 1). In solving the distance matrix of the network, the idea of Dijkstra's algorithm is often used to solve the distance matrix, which has a time complexity of O (n^3), while our proposed DSEP algorithm only needs to use the four-order distances. With our proposed method, the time complexity of solving the finite-order distances can be reduced to O (n^2), which greatly reduces the time complexity of the DSEP algorithm.

2.1. Topological information matrix based on four-order distances

Most real networks are known to satisfy a power-law distribution. Based on this, Li *et al.*,[17] using the local fractal dimension, proposed the following formula to calculate the scores of nodes in the PPI network:

$$B_{v_i}(r) = C \cdot r^{D_{v_i}}, \tag{1}$$

where $B_{v_i}(r)$ denotes the sum of those nodes with distance from node v_i between 0 to r. D_{v_i} denotes the score of node v_i, and C denotes a constant.

Equation (1) can be expressed as $D_{v_i} = m \cdot \ln B_{v_i}(r)/\ln r$. In this equation, each order of distance of node v_i has the same importance for node v_i, which is inconsistent with the actual situation. Therefore, we define the following formula to cal-

culate the topological similarity of nodes in the PPI network:

$$B_{v_i} = c_1 \ln r_1 + \sum_{i=2}^{4} c_i \frac{\ln r_i}{\ln i}, \tag{2}$$

where c_i $(i = 1,2,3,4)$ denotes the coefficient of the i-th order distance of node v_i, and r_i $(i = 1,2,3,4)$ denotes the total number of nodes of the i-th order distance of node v_i. In this paper, we have mainly considered the four-order distances of v_i.

Based on the equation defined above, we construct a four-order distances-based topological information matrix, defined as follows:

$$\boldsymbol{T}_{i,j} = \begin{cases} \dfrac{c_1 \ln r_1(i)}{r_1(i)}, & \text{if } d_{i,j} = 1, \\ c_{d_{i,j}} \dfrac{\ln r_{d_{i,j}}(i)}{\ln d_{i,j}} \\ \hline r_{d_{i,j}}(i), & \text{if } d_{i,j} = 2,3,4, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $r_{d_{ij}}(i)$ $(d_{i,j} = 1,2,3,4)$ denotes the total number of nodes in the PPI network with distance j to node v_i.

2.2. Information matrix for subcellular localization based on four-order distances

Subcellular localization information plays an indispensable role in the study of gene function, which represents the specific location of proteins or gene expression products in the cell. We analyzed the involvement of proteins in 11 intracellular locations and defined a matrix of information on the subcellular localization of proteins as follows:

$$S_{i,j} = \begin{cases} 1, & \text{if } v_i \text{ and } v_j \text{ in the same subcellular}, \\ 0, & \text{otherwise}. \end{cases} \tag{4}$$

Combining Eq. (4), we define the information matrix for subcellular localization based on four-order distances

$$\boldsymbol{L}_{i,j} = \begin{cases} \dfrac{a_1 \ln sr_1(i)}{sr_1(i)}, & \text{if } d_{i,j} = 1 \text{ and } S_{i,j} = 1, \\ a_2 \dfrac{\ln sr_2(i)}{\ln 2} \\ \hline sr_2(i), & \text{if } d_{i,j} = 2 \text{ and } S_{i,j} = 1, \\ a_3 \dfrac{\ln sr_3(i)}{\ln 3} \\ \hline sr_3(i), & \text{if } d_{i,j} = 3 \text{ and } S_{i,j} = 1, \\ a_4 \dfrac{\ln sr_4(i)}{\ln 4} \\ \hline sr_4(i), & \text{if } d_{i,j} = 4 \text{ and } S_{i,j} = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where $sr_{d_{ij}}(i)$ $(d_{i,j} = 1,2,3,4)$ denotes the number of proteins in the PPI network which are distant from node v_i as j and are jointly involved in subcellular processes.

2.3. DSEP algorithm

We propose the DSEP algorithm based on the random walks method. We chose this approach because of its ability to integrate well the topological properties of the network. The traditional random walks method only considers the topological properties of the network. In 2021, Ahmed *et al.*[27] proposed the EPD-RW method by combining four biological properties of proteins and eight topology-based centrality methods using the random walks method. The random walks are iteratively formulated as follows:

$$\boldsymbol{q}^{(t+1)} = \alpha \boldsymbol{P} \boldsymbol{q}^{(t)} + \beta \boldsymbol{B} \boldsymbol{q}^{(t)} + (1 - \alpha - \beta)\boldsymbol{e}, \tag{6}$$

where $\boldsymbol{P}$ denotes the proteins' topological similarity matrix, $\boldsymbol{B}$ denotes the biological property similarity matrix of the proteins, and $\boldsymbol{e}$ is an $1 \times n$-dimensional (n denotes the total amount of proteins) vector with value all $1/n$, which denotes the probability vector of returning to the start node in a random walk.

Equation (6) combines the biological and topological properties of proteins well but the probability of each restart is $1/n$, which does not reflect the actual situation effectively. Therefore, we propose the following equation for the random walks method:

$$\boldsymbol{q}^{(t+1)} = \alpha \boldsymbol{L}' \boldsymbol{q}^{(t)} + (1 - \alpha)\boldsymbol{T}', \tag{7}$$

where $\boldsymbol{L}'$ is the matrix obtained by normalizing the matrix $\boldsymbol{L}$ (defined by Eq. (8)), which represents the transfer probability matrix in the random walks method, and $\boldsymbol{T}'$ is the matrix evolved from the matrix $\boldsymbol{T}$ (the exact process can be seen in Eqs. (9) and (10)), which represents the probability of returning to the starting position in the random walks method.

In order not to destroy the distance information contained in the matrix, we normalize $\boldsymbol{L}$ according to a new normalization method as follows:

$$\boldsymbol{L}' = \frac{\boldsymbol{L}}{\text{MS-L}}. \tag{8}$$

where MS-L denotes the maximum row sum of matrix $\boldsymbol{L}$.

We obtained the scoring matrix $\boldsymbol{C}$ of the proteins based on the PPI network topology using the row sum of matrix $\boldsymbol{T}$. The definition is as follows:

$$\boldsymbol{C}_i = \sum_{j=1}^{n} T_{i,j}. \tag{9}$$

After that, we normalized $\boldsymbol{C}$ using the sigmoid function, so that the restart probability of each node is 0–0.5.

$$\boldsymbol{T}'_i = \frac{1}{1 + \mathrm{e}^{-\boldsymbol{C}_i}} - 0.5, \quad (i = 0,1,2,\ldots,n). \tag{10}$$

Algorithm 1 DSEP

Input:

The data of PPI network $G=(V,E)$;
The adjacency matrix $\boldsymbol{A}$ of the PPI network;
Subcellular localization information;
Parameters in Table 2; Stopping error $\varepsilon = 0.000001$;

Output:

The sorting vector q of the PPI network;
Calculate the matrices $\boldsymbol{A}^2$, $\boldsymbol{A}^3$, $\boldsymbol{A}^4$ through the matrix $\boldsymbol{A}$;
Initialize the four-order distances matrix $\boldsymbol{D}$ of the graph G to be a matrix with all values of $n \times n$ being zero;
for each $v_i, v_j \in V(G)$ **do**
    if $A_{i,j} = 1$
        $D_{i,j} = 1$
    elif $A^2_{i,j} \geq 1$
        $D_{i,j} = 2$
    elif $A^3_{i,j} \geq 1$
        $D_{i,j} = 3$
    elif $A^4_{i,j} \geq 1$
        $D_{i,j} = 4$
end for
Calculating the four-order distances-based topological information matrix $\boldsymbol{T}$ by Eq. (3);
Calculating the information matrix of subcellular localization based on four-order distances $\boldsymbol{L}$ by Eqs. (4) and (5);
Normalize the matrix $\boldsymbol{L}$ to obtain the matrix $\boldsymbol{L}'$ by Eq. (8);
Reconstruct matrix $\boldsymbol{T}$ to get matrix $\boldsymbol{T}'$ by Eqs. (9) and (10);
Initialize the vector $\boldsymbol{q}^{(t)} = (1, 1, \ldots, 1)$;
while $\|\boldsymbol{q}^{(t+1)} - \boldsymbol{q}^{(t)}\| \geq \varepsilon$ **do**
    Compute $\boldsymbol{q}^{(t+1)}$ by Eq. (7);
end while
$\boldsymbol{q} = \boldsymbol{q}^{(t+1)}$;
return $\boldsymbol{q}$

3. Results and discussion

3.1. Experimental data

To evaluate the performance of DSEP, we have performed a computational analysis of the PPI network of Saccharomyces cerevisiae, which is the most integral and reliably available single-cell network. We used data from three PPI networks, which are BioGRID,[28] YHQ,[29] and YMBD. The BioGRID data came from the BioGRID database,[28] the YHQ data was constructed by Yu *et al.*,[29] and the YMBD data comes from the Mark Gerstein Lab website. After removing duplicate edges and self-interaction from the data, the network data are shown in Table 1. The data for standard essential proteins are integrated from the following data sets: MIPS,[30] SGD,[31] DEG,[32] and SGDP.[33] Subcellular localization information data is downloaded from COMPARTMENTS database,[34] which contains 11 types of subcellular localization information.

3.2. Parameter settings

In this paper, all of our parameters are locally optimal solutions obtained in the various modules of Algorithm 1. For example, in the BioGRID data, we chose the parameters for the best performance case of identifying essential proteins, which was to obtain the specific values (proportions) of parameters c_1, c_2, c_3, c_4, here we set to 0–10 to traverse, respectively, and bring them into Eq. (3), find the row sums, and rank them in descending order. The parameters that we chose are shown in Table 2.

Table 1. The information contained in the three PPI networks: BioGRID, YHQ, YMBD.

Data set	Proteins	Interaction	Essential proteins
BioGRID	5616	52833	1199
YHQ	4743	23294	1108
YMBD	2559	11835	763

Table 2. Parameters in Algorithm 1.

Data sets	a_1	a_2	a_3	a_4	c_1	c_2	c_3	c_4	α
BioGRID	3	0	3	1	4	0	5	2	0.9
YHQ	2	1	0	4	1	1	0	8	1.0
YMBD	1	2	0	0	2	0	1	1	0.2

3.3. Comparative analysis with other methods

To validate the performance of our proposed DSEP algorithm, we comprehensively compared DSEP with 11 other algorithms that excelled in identifying essential proteins. The experiments showed that DSEP outperformed these compared methods. In the following comparison experiments, the compared methods are DC,[5] BC,[6] EC,[7] CC,[8] SC,[9] LAC,[10] NC,[11] UC,[16] Pec,[14] Wdc,[18] and CENC.[19]

3.3.1. Histograms

A histogram can visualize the performance difference of various methods in correctly identifying the number of essential proteins. First, we ranked the DSEP and 11 centrality methods in sorted descending order with their scores. After that, the top 100–600 selected proteins were used as the candidate essential protein set. Finally, the number of true essential proteins in the candidate essential protein set was derived by comparing the standard essential protein set. The results of the experiments are shown in Figs. 2–4.

Figure 2 shows the identification results of DSEP with 11 comparison methods on the BioGRID data set. DSEP identified 24, 39, 50, 46, 39, and 50 more essential proteins when compared to the best identification results of these 11 methods in the top 100–600, respectively.

Figure 3 shows the identification results of DSEP with 11 comparison methods on the YHQ data set. DSEP identified 30, 59, 77, 63, 76, and 87 more essential proteins when compared to the best identification results of these 11 methods in the top 100–600, respectively.

Figure 4 shows the identification results of DSEP with 11 comparison methods on the YMBD data set. DSEP identified 13, 21, 35, 47, 53, and 51 more essential proteins when compared to the best identification results of these 11 methods in the top 100–600, respectively.

This analysis shows that DSEP was significantly better at accurately identifying the number of essential proteins in the top 100–600 of the BioGRID, YHQ, and YMBD data sets than the 11 comparison methods.

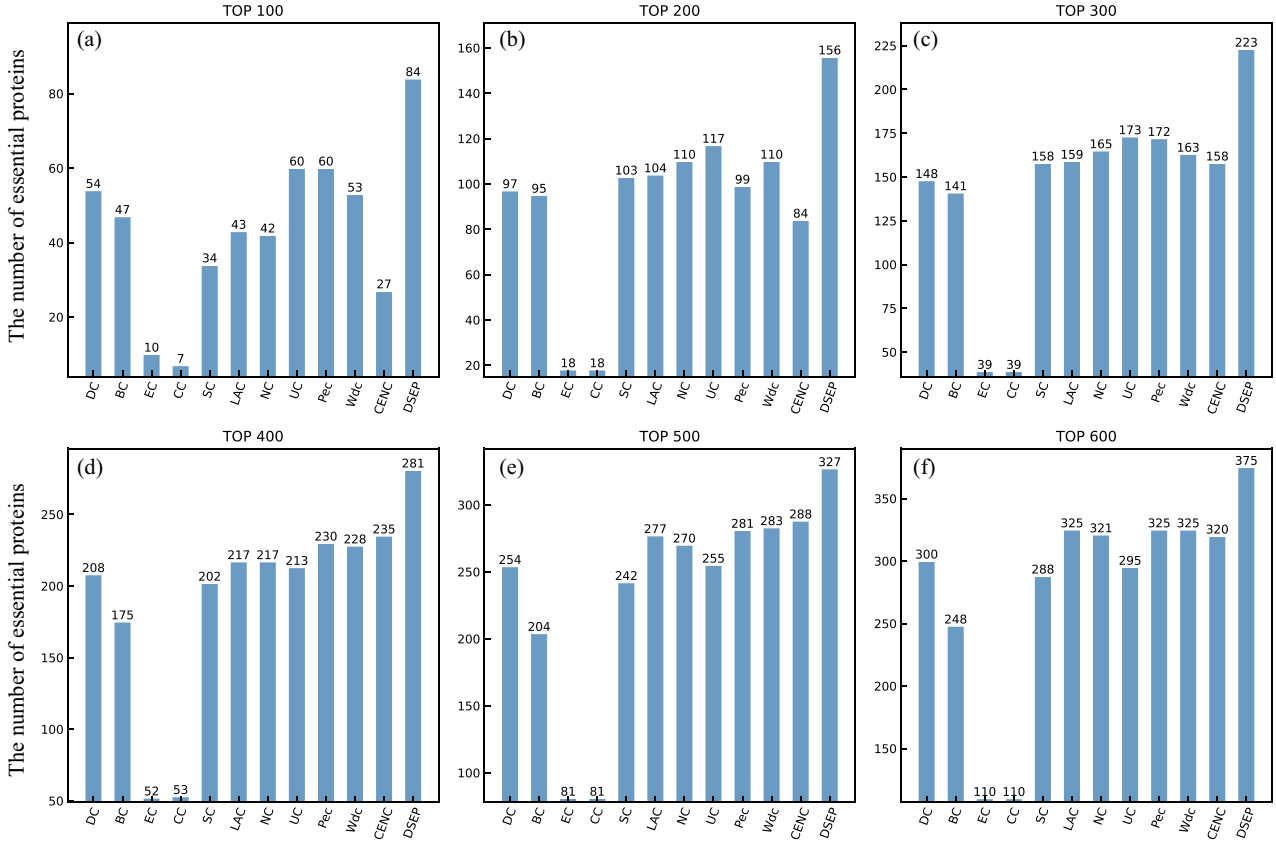


Fig. 2. Comparison of the number of essential proteins accurately identified by DSEP with 11 other methods on the BioGRID data set.

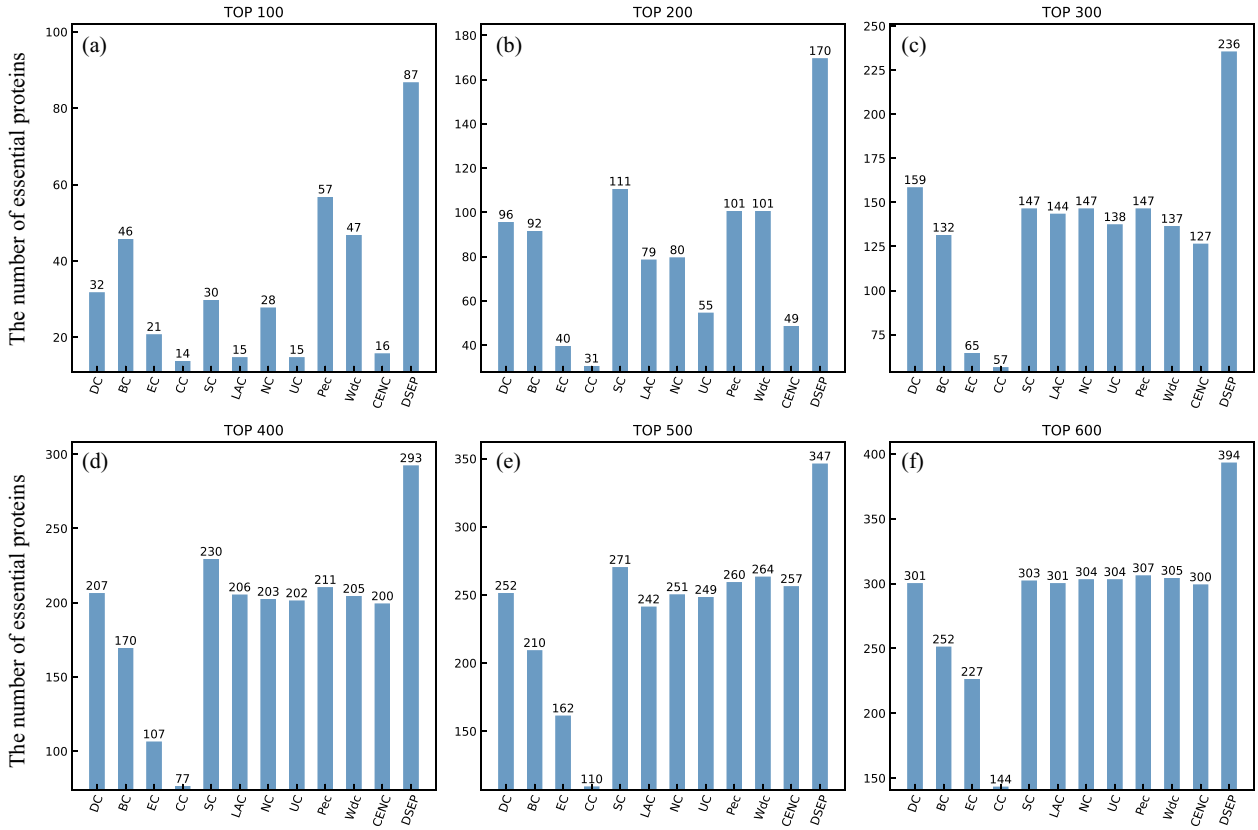


Fig. 3. Comparison of the number of essential proteins accurately identified by DSEP with 11 other methods on the YHQ data set.

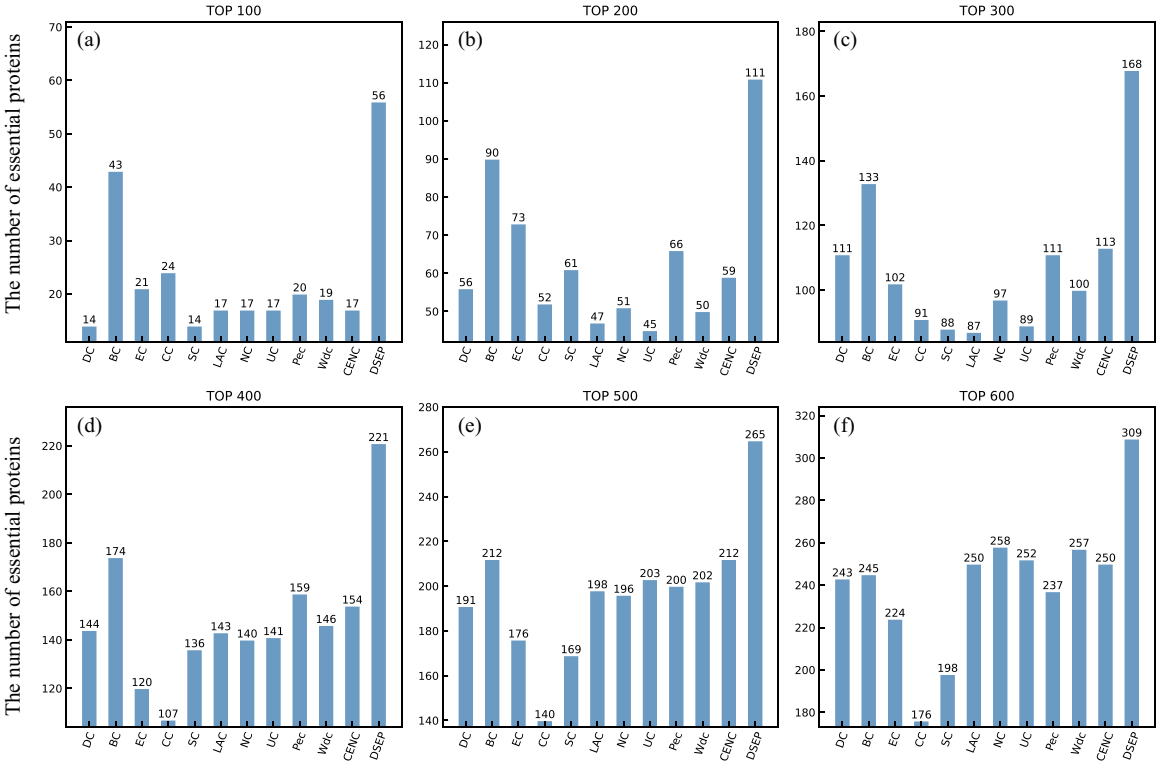


Fig. 4. Comparison of the number of essential proteins accurately identified by DSEP with 11 other methods on the YMBD data set.

3.3.2. Six evaluation indicators

To further assess the performance of our algorithm, we have compared DSEP with 11 other methods in a comprehensive manner under six evaluation metrics. These six evaluation metrics are sensitivity (SN), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), F-measure (F), and accuracy (ACC),[16] which are defined as follows:

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{11}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}}, \tag{12}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{13}$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}, \tag{14}$$

$$\mathrm{F} = \frac{2\times\mathrm{SN}\times\mathrm{PPV}}{\mathrm{SN}+\mathrm{PPV}}, \tag{15}$$

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \tag{16}$$

where TP indicates the amount of protein positively predicted as essential, TN indicates the amount of protein positively predicted as non-essential, FP indicates the amount of non-essential protein falsely predicted as essential, and FN indicates the amount of essential protein falsely predicted as non-essential.

We ranked the scores of the individual proteins in descending order and selected the top 20% of the proteins as the candidate essential protein set and the other 80% of the proteins as the candidate non-essential protein set. From comparing the standard essential protein sets, we were able to obtain the amount of correctly identified essential proteins in the candidate essential protein sets. Among these six statistical metrics, the higher the value of the metric indicates the better the execution of the algorithm. The comparison results of DSEP algorithm with 11 methods in these six indicators are shown in Table 3, and the experiments show that our algorithm is better than these comparison methods.

3.3.3. The *P*–*R* curves

The *X*-axis of the precision–recall curve (*P*–*R* curve) represents the recall and the *Y*-axis represents the precision. In the *P*–*R* curve, the area enclosed by the curve and the axes is represented by AUPR. When the algorithm's AUPR value is higher, the algorithm has a better performance. Figures 5–7 show the *P*–*R* curves of our DSEP algorithm and other 11 comparison methods on three different data sets (i.e., BioGRID, YHQ, and YMBD), and it is shown from the AUPR values that our algorithm performs better. The formulas for the Precision and Recall are defined as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{17}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{18}$$

3.3.4. ROC curves

The ROC curve is also a good way to evaluate an algorithm's good or bad performance. The area enclosed by the

ROC curve and the coordinate axes is called AUC, and similar to AUPR, the higher the AUC value, the better the algorithm performs in the ROC curve, as shown in Figs. 8–10, our algorithm outperforms the other 11 methods in the ROC curve.

Table 3. Comparison of the DSEP algorithm with the other 11 methods in the six statistical indicators on three different data sets.

Data sets	Methods	SN	SP	PPV	NPV	F-measure	Accuracy
	DC	0.415	0.848	0.449	0.830	0.432	0.749
	BC	0.376	0.836	0.406	0.818	0.390	0.731
	EC	0.248	0.811	0.281	0.784	0.263	0.682
	CC	0.248	0.811	0.281	0.784	0.263	0.682
	SC	0.381	0.840	0.415	0.820	0.397	0.735
BioGRID	LAC	0.436	0.855	0.473	0.836	0.454	0.759
	NC	0.445	0.857	0.481	0.838	0.462	0.762
	UC	0.386	0.848	0.432	0.823	0.408	0.742
	Pec	0.409	0.846	0.442	0.828	0.425	0.746
	Wdc	0.440	0.856	0.476	0.837	0.457	0.760
	CENC	0.385	0.850	0.433	0.823	0.408	0.743
	DSEP	0.472	0.864	0.509	0.846	0.490	0.774
	DC	0.404	0.856	0.489	0.808	0.443	0.741
	BC	0.330	0.831	0.399	0.785	0.361	0.704
	EC	0.392	0.857	0.483	0.805	0.432	0.739
	CC	0.284	0.822	0.352	0.771	0.315	0.685
	SC	0.394	0.856	0.482	0.806	0.433	0.739
YHQ	LAC	0.431	0.867	0.524	0.817	0.473	0.756
	NC	0.426	0.864	0.517	0.816	0.467	0.753
	UC	0.402	0.858	0.491	0.808	0.442	0.742
	Pec	0.380	0.841	0.449	0.799	0.412	0.724
	Wdc	0.422	0.863	0.511	0.814	0.463	0.751
	CENC	0.419	0.860	0.504	0.813	0.458	0.748
	DSEP	0.457	0.866	0.537	0.824	0.493	0.762
	DC	0.261	0.832	0.413	0.713	0.320	0.653
	BC	0.282	0.830	0.430	0.718	0.340	0.659
	EC	0.239	0.820	0.376	0.704	0.292	0.639
	CC	0.194	0.801	0.306	0.686	0.238	0.611
	SC	0.224	0.814	0.354	0.698	0.274	0.630
YMBD	LAC	0.270	0.836	0.427	0.716	0.331	0.659
	NC	0.266	0.835	0.422	0.715	0.326	0.657
	UC	0.274	0.838	0.434	0.718	0.336	0.662
	Pec	0.270	0.823	0.410	0.713	0.325	0.651
	Wdc	0.274	0.836	0.431	0.717	0.335	0.660
	CENC	0.280	0.834	0.434	0.719	0.341	0.661
	DSEP	0.359	0.860	0.538	0.747	0.431	0.704

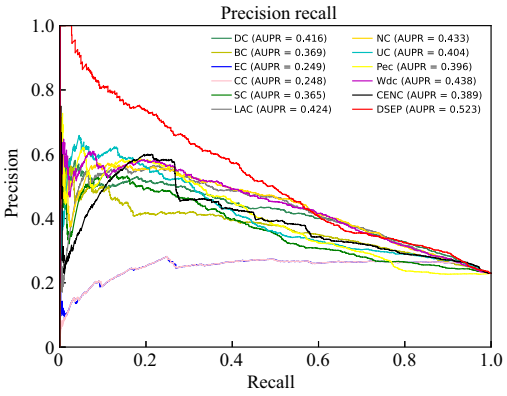


Fig. 5. The *P*–*R* curve of the DSEP algorithm and other 11 methods in the BioGRID data set.

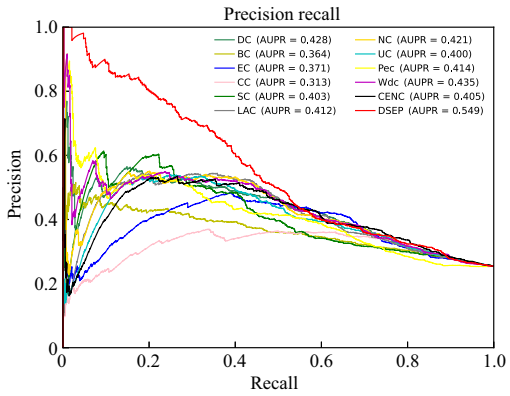


Fig. 6. The *P*–*R* curve of the DSEP algorithm and other 11 methods in the YHQ data set.

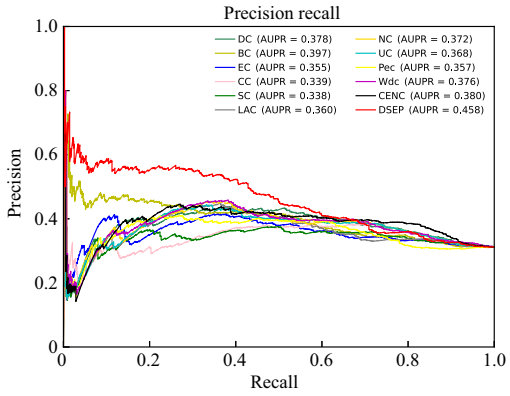


Fig. 7. The *P*–*R* curve of the DSEP algorithm and other 11 methods in the YMBD data set.

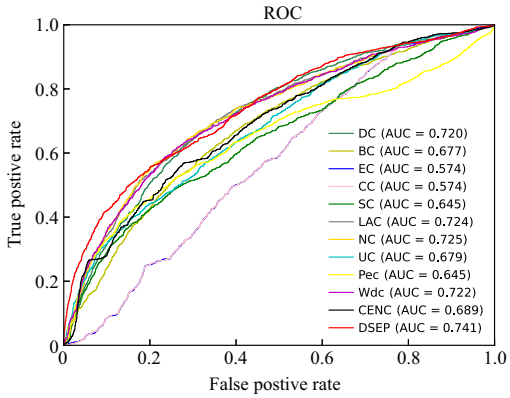


Fig. 8. ROC curves of the DSEP algorithm and other 11 methods in the BioGRID data set.

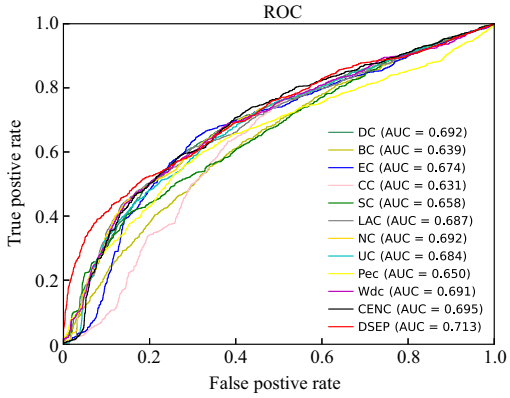


Fig. 9. ROC curves of the DSEP algorithm and other 11 methods in the YHQ data set.

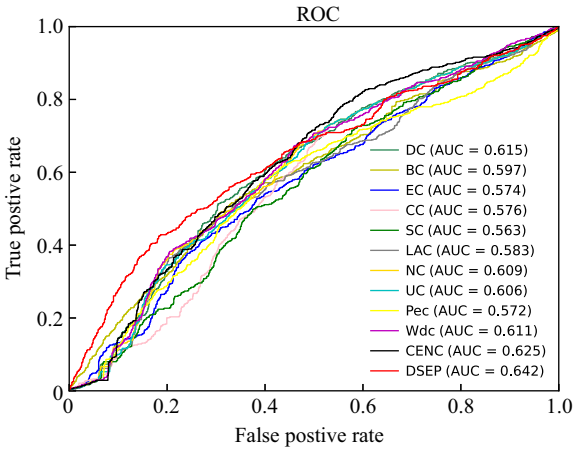


Fig. 10. ROC curves of the DSEP algorithm and other 11 methods in the YMBD data set.

4. Conclusion

In this paper, we propose an essential proteins identification algorithm based on four-order distances and a method to calculate finite-order distance combined with some knowledge of graph theory, which can effectively reduce the time complexity of solving the distance between nodes. To verify the performance of DSEP, we did comparative experiments on BioGRID, YHQ, and YMBD data sets using various evaluation methods against the existing 11 methods. The results of the experiments show that DSEP has an excellent performance.

Acknowledgments

Project supported by the Gansu Province Industrial Support Plan (Grant No. 2023CYZC-25), the Natural Science Foundation of Gansu Province (Grant No. 23JRRA770), and the National Natural Science Foundation of China (Grant No. 62162040).

References

[1] Acencio M L and Lemke N 2009 *BMC. Bio.* **10** 290
[2] Giaever G, Chu A M, Ni L, *et al.* 2002 *Nature* **418** 387
[3] Roemer T, Jiang B, Davison J, *et al.* 2003 *Mole. Micr.* **50** 167
[4] Cullen L M and Arndt G M 2005 *Immu. Cell. Bio.* **83** 217
[5] Freeman L C 1978 *Soc. Net.* **1** 215
[6] Joy M P, Brock A, Ingber D E, *et al.* 2005 *J. Bio. Biot.* **2005** 96
[7] Bonacich P 1987 *Am. J. Soc.* **92** 1170
[8] Wuchty S and Stadler P F 2003 *J. Theo. Bio.* **223** 45
[9] Estrada E and Rodriguez-Velazquez J A 2005 *Phy. Rev. E* **71** 056103
[10] Li M, Wang J, Chen X, *et al.* 2011 *Comp. Bio. Chem.* **35** 143
[11] Wang J X, Li M, Wang H, *et al.* 2012 *IEEE ACM. TCBB.* **9** 1070
[12] Hsing M, Byler K G and Cherkasov A 2008 *BMC. Sym. Bio.* **2** 80
[13] Peng X Q, Wang J X, Wang J, *et al.* 2015 *PloS One* **10** e0130743
[14] Li M, Zhang H, Wang J, *et al.* 2012 *BMC Sys. Bio.* **6** 15
[15] Xiao Q, Wang J, Peng X, *et al.* 2015 *BMC Geno.* **16** S1
[16] Li M, Lu Y, Niu Z, *et al.* 2015 *IEEE ACM. TCBB.* **14** 370
[17] Shen L, Zhang J, Wang F, *et al.* 2022 *Genes* **13** 173
[18] Tang X W, Wang J X, Zhong J C, *et al.* 2013 *IEEE ACM. TCBB.* **11** 407
[19] Lu P L and Yu J J 2020 *Inter. J. Mode. Phy. B* **34** 2050090
[20] Lei X, Yang X and Fujita H 2019 *Know. Bas. Sys.* **167** 53
[21] Zhou Y, Wu C and Tan L 2021 *Physica A* **570** 125783
[22] Park H, Jung J and Kang U 2017 2017 *IEEE International Conference on Big Data (Big Data)* pp. 756–765
[23] Jung J, Jin W, Sael L, *et al.* 2016 2016 *IEEE 16th International Conference on Data Mining (ICDM)* pp. 973–978
[24] Bahadori S, Moradi P and Zare H 2021 *Appl. Inte.* **51** 3561
[25] Bestehorn M, Riascos A P, Michelitsch T M, *et al.* 2021 *Cont. Mech. Ther.* **33** 1207
[26] Biggs N, Biggs N L and Norman B 1993 *Algebraic graph theory*, 2nd (Cambridge:Cambridge university press) pp. 9–10
[27] Ahmed N M, Chen L, Li B, *et al.* 2021 *Soft. Comp.* **25** 8883
[28] Stark C, Breitkreutz B J, Chatr-Aryamontri A, *et al.* 2011 *Nucl. Aci. Res.* **39** D698
[29] Yu H, Greenbaum D, Lu H X, *et al.* 2004 *TRENDS. Gene.* **20** 227
[30] Mewes H W, Frishman D, Mayer K F X, *et al.* 2006 *Nucl. Aci. Res.* **34** D169
[31] Cherry J M, Adler C, Ball C, *et al.* 1998 *Nucl. Aci. Res.* **26** 73
[32] Zhang R and Lin Y 2009 *Nucl. Aci. Res.* **37** D455
[33] Winzeler E A, Shoemaker D D, Astromoff A, *et al.* 1999 *Science* **258** 901
[34] Binder J X, Pletscher-Frankild S, Tsafou K, *et al.* 2014 *Database* bau012