IET Systems Biology

Special Section: Selected Papers from the International Workshop on Computational Network Biology: Modeling, Analysis, and Control (CNB-MAC 2017)

SMILE: a novel procedure for subcellular module identification with localisation expansion

ISSN 1751-8849
Received on 9th December 2017
Accepted on 12th December 2017
E-First on 15th January 2018
doi: 10.1049/iet-syb.2017.0085
www.ietdl.org

Lixin Cheng[1], *Pengfei Liu*[1], *Kwong-Sak Leung*[1] ✉

[1]*Department of Computer Science & Engineering, Chinese University of Hong Kong, Ma Liu Shui, Hong Kong*

✉ *E-mail: ksleung@cse.cuhk.edu.hk*

Abstract: Computational clustering methods help identify functional modules in protein–protein interaction (PPI) network, in which proteins participate in the same biological pathways or specific functions. Subcellular localisation is crucial for proteins to implement biological functions and each compartment accommodates specific portions of the protein interaction structure. However, the importance of protein subcellular localisation is often neglected in the studies of module identification. In this study, the authors propose a novel procedure, subcellular module identification with localisation expansion (SMILE), to identify super modules that consist of several subcellular modules performing specific biological functions among cell compartments. These super modules identified by SMILE are more functionally diverse and have been verified to be more associated with known protein complexes and biological pathways compared with the modules identified from the global PPI networks in both the compartmentalised PPI and InWeb_InBioMap datasets. The authors' results reveal that subcellular localisation is a principal feature of functional modules and offers important guidance in detecting biologically meaningful results.

1 Introduction

Protein–protein interaction (PPI) resources are invaluable for proteomic and network related analysis, which have provided great insights for the mechanistic understanding of human diseases and drug design [1–9]. Computational clustering methods help identify functional modules in PPI networks since proteins usually cluster together to participate in the same biological pathways or specific functions [10–14]. However, cell compartmentalisation is overlooked by previous standard procedures, despite a large number of exciting results emerging from analyses at the global cellular level [15–18]. In fact, eukaryotic cells are composed of several subcellular compartments that enable the cell to implement various metabolic activities simultaneously, and proteins need to target appropriate compartment to interact with each other and to form compound functional complexes in the signalling pathways for specialised biological processes and functions [19, 20].

Cell compartmentalisation, the formation of cellular compartments, is physical and a vital regulator of several main biochemical processes in eukaryotic cells, which assign certain biomolecules in different partitions of the cell. Several properties and characteristics like intracellular pH and enzyme systems distinguish one compartment from the others [21]. Many works have observed that the interactions and functions of proteins are closely related to their localisation in the cell [22]. More importantly, localising in common compartments is of vital importance for proteins to interact with each other, at least transiently or conditionally. Accumulated experimental evidence suggested that translocation is an efficient regulation mode in cells and erroneous localisation may lead to disorders or even diseases [23, 24]. For instance, a transcription factor *P53* may be located to a nucleus to promote transcription of certain genes and thereby activating autophagy program upon stimulation, a cellular process of self-eating [25–28]. In contrast, when targeting at cytoplasm, however, *P53* plays an opposite role as a master repressor of the autophagy program [29, 30]. Moreover, localisation-based modulation can change cellular program completely. For instance, the protein *ATG5* is involved in several cellular processes including autophagy and apoptosis. The two cellular programs can be switched as the localisation of *ATG5* changes between mitochondria and cytoplasm [31]. These common examples in experimental biology cannot be fully figured out through analysing the global cellular network without a comprehensive study of analysing the localisation compartmentalised subnetworks. Therefore, identifying modules consisting of closely interacted proteins localised in a specific compartment is expected to generate more biologically meaningful results, as cells can naturally be decomposed into several compartments.

In the meanwhile, several algorithms have been designed for the identification of protein complexes in the bioinformatics community, although none of them take the cell compartmentalisation into consideration. ClusterONE (Cluster with Overlapping Neighbourhood Expansion) [16] strives to discover not only densely connected clusters with comparable accuracies but also possibly overlapping clusters. It executes a greedy growth algorithm to cluster networks from small seeds supervised by a fitness function concentrating on the cluster separability, which is formulated by the ratio between the number of internal interactions of a cluster and the number of all interactions linking the cluster. Then each generated cluster is statistically evaluated by a probability using Monte Carlo random interaction number of the clusters. Another clustering algorithm focusing on the explicit topological structure of protein complexes is finding low-conductance sets with dense interactions (FLCD) [32], a two-step algorithm considering both the internal and external connectivity of protein complexes. It first detects clusters with high separability and then the clusters with high edge density are detected as protein complexes. By mimicking Markovian random walk on networks, several other clustering algorithms were also developed, such as Markov Clustering (MCL), regularized Markov Clustering (R-MCL), and soft regularised Markov Clustering (SR-MCL) [12, 33]. MCL simulates many stochastic flows within a network by making the strong flows stronger and the weak ones weaker. After multiple iterations, the identified cluster come out with strong internal flows and separated by the boundary with no flows [33]. R-MCL, an improved version of MCL which is more accurate and less time consuming, scales much better to moderate sized networks by penalising the large clusters at each iteration. However, both R-MCL and SR-MCL can only identify non-overlapped clusters. To address this problem, another method SR-MCL was developed to achieve overlapped clusters by executing R-MCL multiple times [12].

Additionally, it is worth pointing out that proteins are interacting spatially to form a dynamic cellular network. Some proteins are localised in multiple compartments and may not directly interact with each other in the same compartment, but they still work towards similar cellular functionalities and hence should belong to the same modules. For instance, transmembrane receptor proteins tend to interact with cytoplasmic proteins as well as with extracellular ligands in signal transduction cascades [34]. Hence, highly overlapped protein modules, either from the same or different compartments, need to be merged to achieve the final super modules.

In this study, we introduce a novel procedure, subcellular module identification with localisation expansion (SMILE), to identify subcellular modules from each cell compartment with localisation extension. Theoretically, the identified super modules engage interactions with high confidence. Experimentally, our results demonstrate that SMILE outperforms the conventional clustering method with respect to protein complex detection and biological pathway annotation, especially for the novel modules exclusively identified by SMILE.

2 Material and methods

2.1 Datasets

Two human PPI networks, ComPPI v1.1 [24] and InWeb_InBioMap 2016_09_12 [15], were employed to identify the functional modules in this study. ComPPI (compartmentalised PPI) is an online database which provides qualitative information on both the interactions among proteins and their localisations. With experimental evidence, the interactions in ComPPI are collected from nine high-quality PPI databases, i.e. the Drosophila Interactions Database (DroID), the Human Protein Reference Database (HPRD), the Matrix Database (MatrixDB), the Munich Information Center for Protein Sequences (MIPS), the Biological General Repository for Interaction Datasets (BioGRID), the Center for Cancer Systems Biology (CCSB), the Database of Interacting Proteins (DiP), the IntAct Molecular Interaction Database (IntAct), the Molecular INTeraction Database (MINT). We excluded the biological unlikely interactions with interaction scores <0.8, resulting in 16,053 proteins and a total of 193,691 corresponding interactions among six major cellular compartments, i.e. nucleus, cytosol, mitochondrion, secretory-pathway, membrane, and the extracellular compartment. The major compartments were defined in the ComPPI database, among which several minor secretory organelles are combined into one major compartment 'secretory-pathway', including Golgi apparatus, endoplasmic reticulum, endosome, peroxisome, lysosome, vacuole, and vesicles. The subcellular localisation annotations are coming from both experiments and computational predictions. The same as Veres *et al.*, localisation score is used to measure the probability of localisation for each protein, depending on the evidence type (experimental or predicted) and the number of sources. Only proteins with a high localisation score (>0.8) are retained for further study.

InWeb_InBioMap, or simply InWeb_IM for short, is the largest dataset of human PPIs at present. It has an extremely large coverage of PPIs (more than half a million) that are retrieved from eight orthology PPI datasets. 57% of the interactions have experimental evidence and the others were computational predicted. Similar to Veres *et al.* [24] suggested we assigned the localisation information to proteins and interactions by calculating the localisation score and interaction score, respectively. The interaction score distribution of ComPPI shows a majority of interactions score higher than 0.8 as shown in Figure S1, so we can end up with the same results with other thresholds less than it.

To evaluate the performance of SMILE on protein complex detection, we estimated the identified modules with two golden standards of a protein complex, i.e. a comprehensive resource of mammalian protein complexes and Protein Complex Database with a Complex Quality Index (PCDq) [35–37]. The latest versions of them were used and only protein complexes including five or more members were considered for further study.

To examine whether the identified modules are biologically meaningful, we used four pathway resources, Kyoto Encyclopedia of Genes and Genomes (KEGG) [38, 39], Protein ANalysis THrough Evolutionary Relationships (PANTHER) [40, 41], BioCarta (https://cgap.nci.nih.gov/Pathways /BioCarta_Pathways), and Reactome [42, 43], as the golden standards for function prediction. Pathway annotations of PANTHER were obtained from PANTHER Pathway 3.4.1 and the other data were collected from the curated gene sets of Molecular Signatures Database (MSigDB v5.2) [44, 45].

2.2 Global module identification

We used ClusterONE [16] to identify modules from the entire PPI network. ClusterONE strives to discover not only densely connected clusters with comparable accuracies but also possibly overlapping regions within a given network, a distinct advantage of ClusterONE. It plugs in Cytoscape [46] and executes a greedy growth algorithm to cluster networks from small seeds supervised by a fitness function. Each generated cluster is evaluated by a cohesiveness score, which is a ratio of the practical interaction number over the theoretical interaction number of the cluster, measuring how likely is a group of proteins to be a module (or cluster separability) [16]. Let V denote a cluster in the PPI network, $w^{\text{in}}(V)$ denotes the number of interactions contained within the cluster, $w^{\text{bound}}(V)$ denotes the number of interactions coming out of the cluster, and $p|V|$ is a penalty term aiming to model the uncertainty of unchecked interactions in the PPI network, the cohesiveness of V is defined as follows:

$$f(V) = \frac{w^{\text{in}}(V)}{w^{\text{in}}(V) + w^{\text{bound}}(V) + p|V|} \tag{1}$$

In this study, we used the default function parameters of ClusterONE and only the identified clusters with a size larger than ten were considered as modules, as small modules are usually more factorisable [47]. These clusters were defined as global modules since they were identified in the entire PPI network instead of the compartmentalised subnetworks.

2.3 Super module identification

As shown in Fig. 1, the procedure of SMILE works in three steps: first, constructing subcellular networks based on localisation annotation; second, identifying clusters with high cohesiveness from each subnetwork; and third, combining highly overlapped clusters.

(i) *Subnetwork construction:* Suppose the input PPI network is $G = (P, E)$, where P is the set of proteins and E is the set of interactions among proteins. Each protein P is annotated to one or more than one compartment. Based on the information of protein subcellular localisation, we extract subnetworks where proteins are localised in an identical compartment. For compartment C_i, we can define its corresponding subnetwork to be $G_i = (P_i, E_i)$, where P_i is the subset of proteins in the compartment C_i and E_i is the subset of interactions in the compartment C_i.

(ii) *Module identification in each subnetwork:* Calculate the cohesiveness of $f(V_i)$ (see (1)) in compartment C_i using ClusterONE, where V_i is the clusters in the subnetwork of C_i.

(iii) *Merging:* The highly overlapped modules identified from different subcellular networks are merged together to generate super modules. Let A and B represent two clusters of proteins from different subnetworks, respectively. Their similarity is measured by an overlap score, which is defined as

$$\omega(A, B) = \frac{|A \cap B|^2}{|A||B|} \tag{2}$$

where $|A|$ and $|B|$ are the sizes of the two clusters, respectively, and $|A \cap B|$ is the number of the overlapping proteins annotated with both subcellular locations.

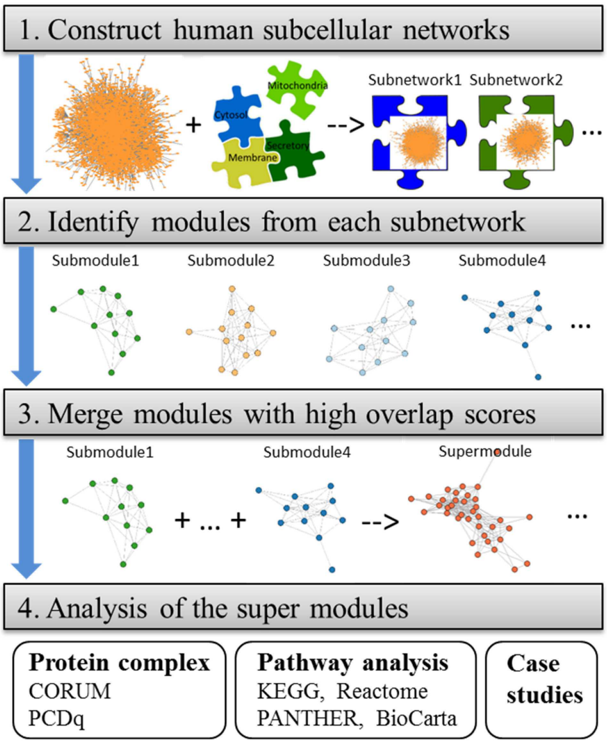


Fig. 1 *Flowchart of the main steps for super module identification*

```
Input: A protein interaction network $G = (P, E)$, a protein subcellular localization list L, and threshold of overlapping score $\omega_0$.
Output: A set of super modules S.
1. $M = \{\}$
2. for $i$ in $L$ do
3.     $G_i = (P_i, E_i)$ // Resulting subcellular network $G_i$
4.     $M_i = ClusterONE(G_i)$ // Resulting subcellular modules from location i
5.     $M = M \cup M_i$
6. end for
7. $O = OverlapScore(M)$ // Resulting similarity matrix O measuring the overlap score of each pair of subcellular modules
8. $G = Graph(O > \omega_0)$
9. $S = DFS(G)$ // Resulting the connected components
```

Fig. 2 *Algorithm: SMILE procedure*

Lastly, we combine subcellular modules identified from different subnetworks. Other than the module compartment, the proteins may localise in some other compartments, considering the multi-localisation property of proteins. We first calculate the overlap scores for each pair of subcellular modules and constructs an adjacency matrix in which each row (or column) represents a module. Two modules are overlapped if the overlap score is larger than a given threshold ω_0. The threshold is set as 0.5 by default, which implies >70% of the members of the two modules being compared are overlapped if they have the same size. Then, we create a graph from the adjacency matrix and split it into connected subgraphs (or components) using depth-first search, where nodes correspond to modules and edges denote the overlapping relationship among all modules. Finally, the modules in each subgraph are merged together and defined as a super module. Mathematically, the procedure of SMILE is also shown in the box of algorithm (see Fig. 2). The identified super modules are essentially all connected components of subcellular modules.

2.4 Section headings

To evaluate the performance of SMILE, we used three quality measurements [11, 13, 16] to compare the results of golden standard complexes with the global modules (modules found using ClusterONE), the super modules (modules found by our method SMILE), and novel modules (modules in a super module but do not in global module). Given r predicted and s reference complexes, let t_{ij} denote the number of proteins that exist in both predicted complex i and reference complex j, v_i and w_j represent the number of proteins in predicted complex i and reference complex j, respectively. Then the three measurements, Sn (sensitivity), PPV (positive predictive value), and Acc (accuracy), are defined as follows:

$$\text{Sn} = \frac{\sum_{j=1}^{s} \max_{i=1,\ldots,r} t_{ij}}{\sum_{j=1}^{s} w_j} \tag{3}$$

$$\text{PPV} = \frac{\sum_{i=1}^{r} \max_{j=1,\ldots,s} t_{ij}}{\sum_{i=1}^{r} \sum_{j}^{s} t_{ij}} \tag{4}$$

$$\text{Acc} = \frac{\sum_{j=1}^{s} \max_{i=1,\ldots,r} t_{ij}}{\sum_{i=1}^{r} v_i} \tag{5}$$

Essentially, Acc is the geometric mean of Sn and PPV. Using the three measurements, we evaluated the global modules, super modules and novel modules with two reference sets CORUM and PCDq (see Section 2.1), respectively.

2.5 Evaluation of module biological relevance

The hypergeometric test was adopted to evaluate whether a module, M, is overrepresented within a biological pathway, X. The probability of observing at least t proteins annotated by X with size T is defined as

$$P = \sum_{i=t}^{n} \frac{\binom{N-T}{n-i}\binom{T}{i}}{\binom{N}{n}} \tag{6}$$

where N is the total number of proteins in the given PPI network and n is the size of the module M. The outputting P value is then adjusted by the Benjamini & Hochberg method for false discovery rate control. The pathway is said to be enriched in the module M at a significance level if the adjusted $P < 0.05$.

The overrepresentation score (ORS) [13, 16] was used to evaluate the biological relevance of the identified modules in pathways. We say a module is biologically meaningful if it is significantly enriched in any biological pathway. Given a set of identified modules, ORS is calculated as the ratio of the number of biologically meaningful modules over the size of the module set, given as

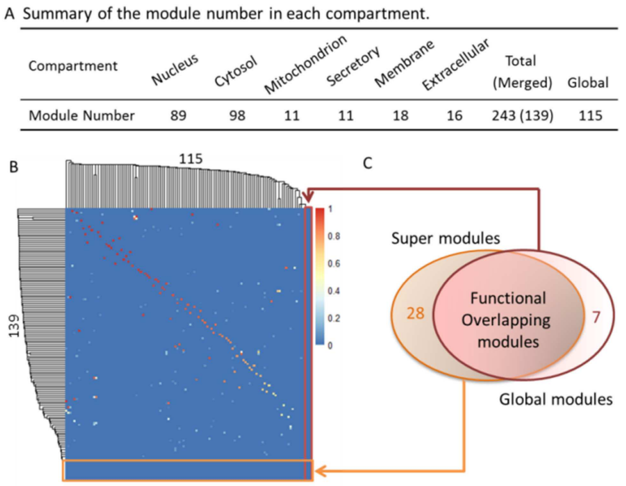
$$\text{ORS} = \frac{\sum_{i}^{U} \text{sgn}\left(\sum_{j}^{V} \text{sgn}\left(P_{\text{cutoff}} - P_{M_iX_j}\right) - 1\right)}{U} \tag{7}$$

where U is the total number of identified modules and V equals the number of pathways. $P_{M_iX_j}$ represents the adjusted P value for module M and pathway X, while P_{cutoff} represents the threshold of the P value of hypergeometric test. ORS ranges from 0 to 1, where 1 represents the case that all the identified modules are significantly associated with reference pathways.

3 Results

3.1 Super modules and novel modules

We used ClusterONE to predict functional modules from the ComPPI network. As shown in Fig. 3, a total of 115 modules are



A Summary of the module number in each compartment.

Compartment	Nucleus	Cytosol	Mitochondrion	Secretory	Membrane	Extracellular	Total (Merged)	Global
Module Number	89	98	11	11	18	16	243 (139)	115

Fig. 3 *Overview of the module numbers and the comparison between super modules and global modules in ComPPI*

(a) Number of identified subcellular modules (243), super modules (139), and global modules (115), ***(b)*** Heatmap of the overlap scores between super modules and global modules. It is essentially an adjacent matrix between the two module sets, in which red represents high score while blue represents low score, ***(c)*** Venn diagram for the super modules and global modules. The super modules involve 28 novel modules and the global modules contain seven unique modules

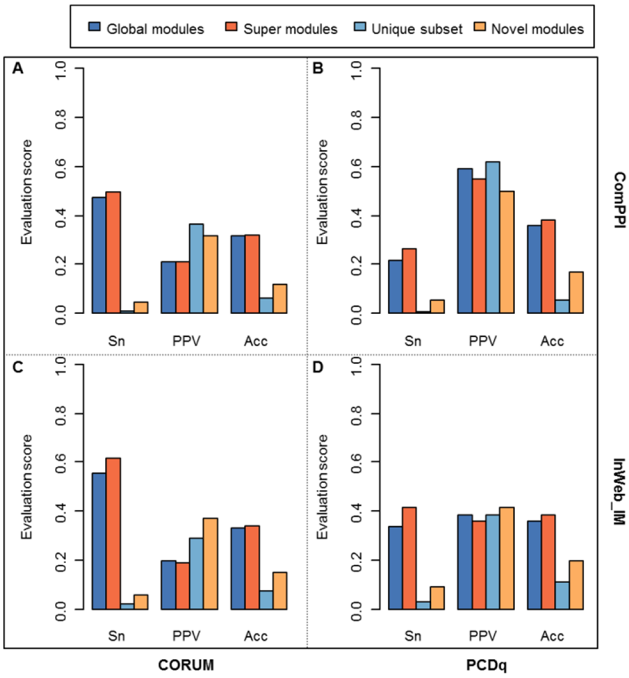


Fig. 4 *Performance of the protein complex prediction. Two PPI networks and two complex references were used for evaluation*

(a) Results using ComPPI network and CORUM reference, ***(b)*** Results using ComPPI network and PCDq reference, ***(c)*** Results using InWeb_IM network and CORUM reference, ***(d)*** Results using InWeb_IM network and PCDq reference. Sn, sensitivity; PPV, positive predictive value; Acc, accuracy

identified and defined as global modules. Using SMILE, on the other hand, we identified 89, 98, 11, 11, 18, and 16 individual modules from nucleus, cytosol, mitochondrion, secretory-pathway, membrane, and extracellular, respectively. These subcellular modules were then merged to generate super modules with larger size if they share substantial module members (see Methods and Figure S2). Eventually, we obtained 139 super modules and 28 out of them have no functional overlap with the global ones (overlap score <0.25). For simplicity, hereafter the super modules not functionally overlapped with the global modules were called as *novel modules*. Likewise, the modules involved in the global module set but not included in the super module set are defined as *global unique modules*. For the InWeb_IM network, as shown in Figure S3, 261 super modules and 158 global modules were captured using SMILE and ClusterONE, respectively. Among the super modules, around one-third of them (82) are novel modules. Note that here the overlap score threshold ω_0 is set as 0.5 by default since we have no preference to merge the modules with high or low overlap for the human PPI datasets. However, for species with quite complete PPI networks, such as yeast, a threshold of 0.75 is suggested to guarantee only highly overlapped modules are merged. Additionally, a series of the overlap score thresholds were used to study the parameter sensitivity of ω_0. Please refer to Supplementary Table S2 for more details.

3.2 Performance comparison for protein complex identification

We found SMILE outperforms the conventional procedure of ClusterONE based on two protein complex reference sets, CORUM and PCDq, on two PPI networks ComPPI and InWeb_IM. Three measurements, Sn, PPV, and Acc, were used to assess the quality of the identified module sets with respect to protein complex prediction. As shown in Fig. 4, it is clear that the super modules generated by SMILE have the highest Acc score and cover more proteins clustered into the reference complexes or modules on both networks. For the ComPPI network, SMILE consistently gets higher Sn and Acc scores based on both references and can identify a comparable proportion of matched complexes in CORUM. Although ClusterONE has a higher PPV than SMILE when using PCDq as the golden standard, SMILE is able to detect more matched protein complexes (76 versus 68, Table S1). For the InWeb_IM network, we can achieve the similar result for the performance of SMILE, which consistently achieves the highest scores of Sn and Acc and a comparable PPV. Strikingly, the exclusively identified novel modules tend to match more protein complexes from both CORUM and PCDq. In particular, 0.3701 and 0.4139 of the novel modules are highly associated with reference complexes from CORUM and PCDq, respectively, whereas the figures are only 0.2885 and 0.3820 for the ClusterONE unique modules. It makes an opposite result in the ComPPI dataset for PPV instead, the reason might be that the coverage of ComPPI is much lower than InWeb_IM and therefore a larger number of unpredicted interactions have yet to be addressed. For more details please refer to Table S1.

The ORSs (see (7)) are calculated for the three types of modules of the two networks, respectively, among four pathway resources, KEGG, PANTHER, BioCarta, and Reactome. The composite score is the sum of ORS for the four resources.

3.3 Performance comparison for biological pathway annotation

Then, we examined the biological relevance of the detected modules by performing overrepresentation analysis of pathway associations. As shown in Table 1, super modules, especially the novel ones, consist of more proteins in the biological pathways on all the four resources on both PPI networks of ComPPI and InWeb_IM. Specifically, in ComPPI, 106 out of 139 super modules (76.26%) are significantly over represented in the KEGG pathways, while this number is dropped to 83 (72.17%) for the modules identified using ClusterONE. For the novel identified modules, 75% of them are significantly enriched in the KEGG pathways, which is also higher than the figure of global modules and its unique subset. In particular, Fig. 5 illustrates the KEGG pathway, ‘SNARE interactions in vesicular transport’ (hsa04130), comprise a significant proportion of proteins that are predicted as members of super modules. Specifically, the proteins in two super modules, module 26 and module 51, were mapped to the KEGG pathways [38, 39] using the KEGG Mapper facility (http://www.genome.jp/kegg/mapper.html). The proteins involved in modules 26 and 51 are marked in yellow and red, respectively, while proteins contained in both modules are marked in orange. It is clear that all the members of module 26 and six proteins in module 51 are the important components of the KEGG PATHWAY:

Table 1 Performance comparison for pathway annotation from four resources

Module set	KEGG	PANTHER	BioCarta	Reactome	Comp score
ComPPI					
global	0.7217	0.2081	0.313	0.9391	2.1819
unique	0.5714	0.2857	0.1428	0.7143	1.7142
super	**0.7626**	0.3165	0.3165	**0.9496**	2.3452
novel	0.75	**0.3929**	**0.3929**	0.9286	**2.4644**
InWeb_IM					
global	0.7975	0.4177	0.3228	0.9241	2.4621
unique	0.6667	0.3333	0.2667	0.8	2.0667
super	0.8314	0.4444	0.3716	0.9579	2.6053
novel	**0.8537**	**0.4878**	**0.3902**	**0.9634**	**2.6951**

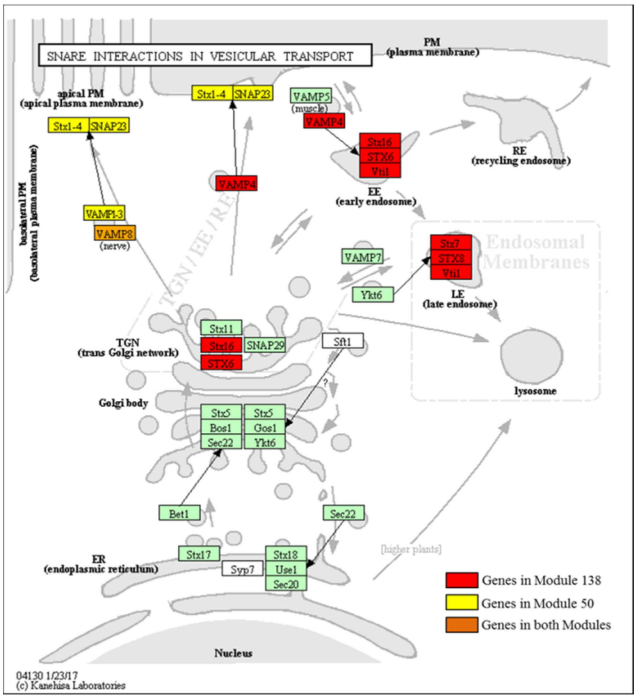


Fig. 5 *SNARE interactions in vesicular transport pathway in KEGG. Genes clustered in modules 50, 138, and both, are marked in yellow, red and orange, respectively. Module 50 is a super module while module 138 is a novel module*

hsa04130, despite the size of the two modules are merely 10 and 15. Strikingly, module 26 is a novel module that is exclusively identified using SMILE, implying its powerful ability in mining pathway relevant information.

The tendency is even more apparent for the InWeb_IM network and other pathway resources. SMILE identified modules, especially the novel modules among them, consistently obtain the highest ORS (Table 1). The composite score is the sum of all ORSs among the four pathway resources and it was used to compare the overall performance of pathway annotation. As expected, the novel modules and super modules consistently obtained the highest and the second highest composite score, both of which are higher than that of the global modules and its unique subset.

3.4 Super modules tend to have more biological relevance

The novel modules identified by SMILE are not only likely to over-represent in specific biological pathways, but also tend to involve in gene families. As shown in Fig. 6*a*, the novel module exclusively identified by SMILE in ComPPI consists of 23 proteins and 15 out of them belong to the Bcl-2 family, an apoptosis regulator filled with evolutionarily related proteins. The proteins in this family supervise mitochondrial outer membrane permeabilisation and usually work as promoters or suppressors of apoptosis [48, 49]. Fig. 6*b* shows another novel module that is enriched with SNARE proteins. Nine out of ten proteins in the module are involved in the protein family of SNARE. The SNARE proteins play a key role as the mediator of vesicle fusion, the fusion of vesicles and their target compartments, among an assortment of others. These results reveal that the SMILE-identified novel modules are highly associated with biological functions, especially the compartment related functions.

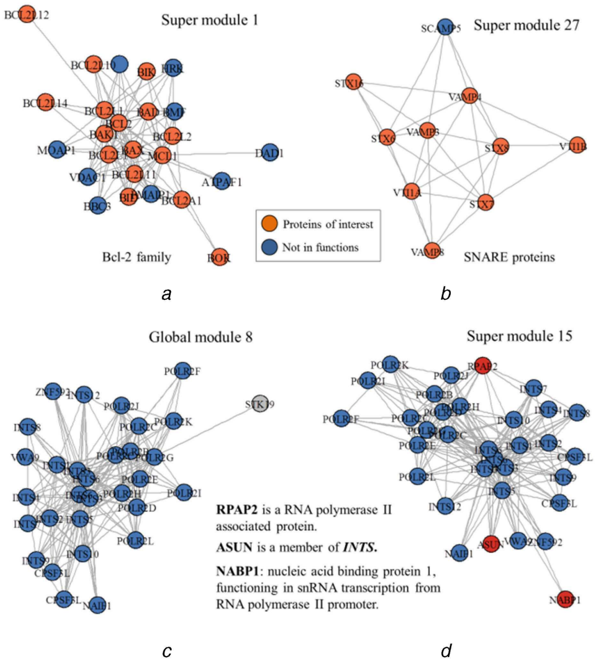


Fig. 6 *Super modules identified using SMILE are more biologically relevant*

(a) Novel module involved in Bcl-2 family, the proteins in the module are localised in nucleus and cytosol, ***(b)*** Novel module enriched SNARE protein and all its proteins are localised in the extracellular region. Proteins of interest are marked in orange, ***(c-d)*** Comparison of a global module and its corresponding super module. Blue nodes denote the common proteins involved in both modules, red nodes denote proteins exclusively detected in the super module, and grey nodes denote the global module unique proteins

Also, SMILE outperforms ClusterONE with respect to the functionally overlapping modules, the super modules that share a large fraction of proteins with global modules. Fig. 6*c* and *d* illustrates a super module and its corresponding highly overlapped global module captured from the ComPPI network. The super module covers all the members of the global module except STK19, a protein with unknown specific function. These common proteins are mainly involved in two protein complexes, POLR2 (RNA polymerase II) and INTS (Integrator complex), which are in charge of regulating RNA polymerase II and RNA processing. POLR2 is an enzyme that can promote the transcription of DNA to synthesise the precursors of mRNA, snRNA, and microRNA [50, 51]. INTS is a highly conserved nuclear complex that usually interacts with the C-terminal tail of the largest subunit of the POLR2 complex to promote 3′-snRNA processing [52]. Interestingly, three more proteins exclusively detected in the super module, ASUN, RPAP2, and NABP1, are closely related to RNA polymerase II. More specifically, ASUN is a member of the complex INTS, RPAP2 is an RNA polymerase II associated protein, and NABP1 functions in snRNA transcription from RNA polymerase II promoter [52, 53]. Overall, SMILE is superior in identifying super modules, either overlapped with global modules or not, with higher biological relevance and functional significance.

3.5 Application to MCL

By default, SMILE use ClusterONE to detect functional modules in a given biological network, because ClusterONE is an efficient algorithm that allows identification of overlapping modules and its plugin in Cytoscape is user-friendly. However, SMILE can be easily applied to other clustering algorithms such as MCL [33], which identifies modules in networks using a mathematical bootstrapping procedure. Hence, MCL was adopted for module identification on both PPI datasets, although it cannot handle overlapping modules. In ComPPI, 28 and 234 modules were identified using MCL and SMILE-MCL (MCL under the strategy of SMILE), respectively; among them, 200 modules were exclusively selected using SMILE-MCL while the counterpart is only 2 for MCL. Importantly, based on CORUM and PCDq references, results of MCL revealed that complexes detected with the SMILE strategy consistently have higher scores on all the evaluation scores including Sn, PPV, and ACC. The two algorithms are comparable when comparing pathway ORSs. For the ComPPI dataset, the ORS of SMILE-MCL for KEGG is less than that of MCL, but SMILE-MCL outperforms MCL on the other three pathway references. For the PPI dataset of InWeb_IM, almost all the ORSs of SMILE-MCL are to some extent less than the scores of MCL. The reason is that only 38 modules are identified using MCL, while the number is 246 for the SMILE-MCL modules, indicating that it is not a powerful way to detect modules merely using MCL.

4 Conclusions

Considering the importance of cell compartmentalisation, we propose a novel procedure SMILE for identifying functional protein modules, which first predict modules separately from each cell compartment, and then compound the highly overlapped ones to generate super modules. These super modules derived by SMILE demonstrated better correspondence with known protein complexes on two databases and biological pathways in four resources than the results of conventional procedures.

Although the dataset used in this study has integrated several available data sources to improve data coverage and quality, the method is limited to those proteins with subcellular localisation information. This limitation can be partially addressed using prediction tools, but in the future, much more work is needed to improve the accuracy of these tools.

Taking the protein subcellular location information into account is the major part of the SMILE procedure and it is a general transformation that universally helps existing complex prediction algorithms perform better, although only ClusterONE and MCL were compared in this study. Specifically, ClusterONE outperforms MCL based on the reported performance evaluation scores in the main manuscript and the Supplementary tables. As shown in Table S3, the super modules identified using ClusterONE consistently obtain the highest evaluation scores except for the pathway composite score of the ComPPI data. For the ComPPI dataset, the accuracy scores of ClusterONE are 0.3212 and 0.3978 for the two references CORUM and PCDq, respectively, both of which are much higher than the other accuracy scores calculated from MCL. The pathway composite score of ClusterONE is 2.3452, which is also comparable to that of MCL (2.3547). Better yet, for the InWeb_IM dataset, the ClusterONE induced super modules to show the best performance amongst the others regardless of the clustering strategies and algorithms.

Not limited to ClusterONE and MCL, SMILE is also applicable to other module identification algorithms depending on users' preference, since it provides more meaningful biological data by evaluating how within a compartment or cross-compartment protein interactions altered or propagated within proteomic datasets. Furthermore, SMILE can be easily applied to other types of network studies to capture modules with multiple components like lncRNA, miRNA, and mRNA [54–57]. In future studies, we will provide more computational procedures to both the coding and non-coding molecules to build a more comprehensive picture of how compartmentalised networks can interact.

5 Acknowledgments

This work was supported by the Chinese University of Hong Kong Direct Grant and Research Grants Council of Hong Kong GRF grant [414413].

6 References

[1] Robinson, C.V., Sali, A., Baumeister, W.: 'The molecular sociology of the cell', *Nature*, 2007, **450**, (7172), pp. 973–982

[2] Yu, H., Braun, P., Yildirim, M.A., *et al.*: 'High-quality binary protein interaction map of the yeast interactome network', *Science*, 2008, **322**, (5898), pp. 104–110

[3] Vidal, M., Cusick, M.E., Barabasi, A.L.: 'Interactome networks and human disease', *Cell*, 2011, **144**, (6), pp. 986–998

[4] Koh, G.C., Porras, P., Aranda, B., *et al.*: 'Analyzing protein-protein interaction networks', *J. Proteome Res.*, 2012, **11**, (4), pp. 2014–2031

[5] Jiang, W., Zhang, Y., Meng, F., *et al.*: 'Identification of active transcription factor and mirna regulatory pathways in Alzheimer's disease', *Bioinformatics*, 2013, **29**, (20), pp. 2596–2602

[6] Hao, D., Li, C., Zhang, S., *et al.*: 'Network-based analysis of genotype-phenotype correlations between different inheritance modes', *Bioinformatics*, 2014, **30**, (22), pp. 3223–3231

[7] Qin, J., Li, M.J., Wang, P., *et al.*: 'Proteomirexpress: inferring microrna and protein-centered regulatory networks from high-throughput proteomic and mRNA expression data', *Mol. Cell Proteomics*, 2013, **12**, (11), pp. 3379–3387

[8] Lievens, S., Van der Heyden, J., Masschaele, D., *et al.*: 'Proteome-scale binary interactomics in human cells', *Mol. Cell Proteomics*, 2016, **15**, (12), pp. 3624–3639

[9] Cheng, L., Fan, K., Huang, Y., *et al.*: 'Full characterization of localization diversity in the human protein interactome', *J. Proteome Res.*, 2017, **16**, (8), pp. 3019–3029

[10] Dittrich, M.T., Klau, G.W., Rosenwald, A., *et al.*: 'Identifying functional modules in protein-protein interaction networks: an integrated exact approach', *Bioinformatics*, 2008, **24**, (13), pp. i223–i231

[11] Li, X., Wu, M., Kwoh, C.K., *et al.*: 'Computational approaches for detecting protein complexes from protein interaction networks: a survey', *BMC Genomics*, 2010, **11**, (Suppl 1), p. S3

[12] Shih, Y.K., Parthasarathy, S.: 'Identifying functional modules in interaction networks through overlapping Markov clustering', *Bioinformatics*, 2012, **28**, (18), pp. i473–i479

[13] Wang, Y., Qian, X.: 'Functional module identification in protein interaction networks by interaction patterns', *Bioinformatics*, 2014, **30**, (1), pp. 81–93

[14] Pizzuti, C., Rombo, S.E.: 'Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods', *Bioinformatics*, 2014, **30**, (10), pp. 1343–1352

[15] Li, T., Wernersson, R., Hansen, R.B., *et al.*: 'A scored human protein-protein interaction network to catalyze genomic interpretation', *Nat. Methods*, 2017, **14**, (1), pp. 61–64

[16] Nepusz, T., Yu, H., Paccanaro, A.: 'Detecting overlapping protein complexes in protein-protein interaction networks', *Nat. Methods*, 2012, **9**, (5), pp. 471–472

[17] Rhee, D.Y., Cho, D.Y., Zhai, B., *et al.*: 'Transcription factor networks in drosophila melanogaster', *Cell Rep.*, 2014, **8**, (6), pp. 2031–2043

[18] Zhang, T., Tan, P., Wang, L., *et al.*: 'Rnalocate: a resource for RNA subcellular localizations', *Nucleic Acids Res.*, 2016, **45**, (D1), pp. D135–D138

[19] An, S., Kumar, R., Sheets, E.D., *et al.*: 'Reversible compartmentalization of de novo purine biosynthetic complexes in living cells', *Science*, 2008, **320**, (5872), pp. 103–106

[20] Hao, N., O'Shea, E.K.: 'Signal-dependent dynamics of transcription factor translocation controls gene expression', *Nat. Struct. Mol. Biol.*, 2011, **19**, (1), pp. 31–39

[21] Casey, J.R., Grinstein, S., Orlowski, J.: 'Sensors and regulators of intracellular pH', *Nat. Rev. Mol. Cell Biol.*, 2010, **11**, (1), pp. 50–61

[22] Park, S., Yang, J.S., Shin, Y.E., *et al.*: 'Protein localization as a principal feature of the etiology and comorbidity of genetic diseases', *Mol. Syst. Biol.*, 2011, **7**, p. 494

[23] Barabasi, A.L., Oltvai, Z.N.: 'Network biology: understanding the cell's functional organization', *Nat. Rev. Genet.*, 2004, **5**, (2), pp. 101–113

[24] Veres, D.V., Gyurko, D.M., Thaler, B., *et al.*: 'ComPPI: a cellular compartment-specific database for protein-protein interaction network analysis', *Nucleic Acids Res.*, 2015, **43**, (Database issue), pp. D485–D493

[25] Fuchs, Y., Steller, H.: 'Programmed cell death in animal development and disease', *Cell*, 2011, **147**, (4), pp. 742–758

[26] Li, Y., Zhuang, L., Wang, Y., *et al.*: 'Connect the dots: a systems level approach for analyzing the mirna-mediated cell death network', *Autophagy*, 2013, **9**, (3), pp. 436–439

[27] Maiuri, M.C., Zalckvar, E., Kimchi, A., *et al.*: 'Self-eating and self-killing: crosstalk between autophagy and apoptosis', *Nat. Rev. Mol. Cell Biol.*, 2007, **8**, (9), pp. 741–752

[28] Marino, G., Niso-Santano, M., Baehrecke, E.H., *et al.*: 'Self-consumption: the interplay of autophagy and apoptosis', *Nat. Rev. Mol. Cell Biol.*, 2014, **15**, (2), pp. 81–94

[29] Tasdemir, E., Chiara Maiuri, M., Morselli, E., *et al.*: 'A dual role of P53 in the control of autophagy', *Autophagy*, 2008, **4**, (6), pp. 810–814

[30] Tasdemir, E., Maiuri, M.C., Galluzzi, L., *et al.*: 'Regulation of autophagy by cytoplasmic P53', *Nat. Cell Biol.*, 2008, **10**, (6), pp. 676–687

[31] Yousefi, S., Perozzo, R., Schmid, I., *et al.*: 'Calpain-mediated cleavage of Atg5 switches autophagy to apoptosis', *Nat. Cell Biol.*, 2006, **8**, (10), pp. 1124–1132

[32] Wang, Y., Qian, X.: 'Finding low-conductance sets with dense interactions (FLCD) for better protein complex prediction', *BMC Syst. Biol.*, 2017, **11**, (Suppl 3), p. 22

[33] Enright, A.J., Van Dongen, S., Ouzounis, C.A.: 'An efficient algorithm for large-scale detection of protein families', *Nucleic Acids Res.*, 2002, **30**, (7), pp. 1575–1584

[34] Papin, J.A., Hunter, T., Palsson, B.O., *et al.*: 'Reconstruction of cellular signalling networks and analysis of their properties', *Nat. Rev. Mol. Cell Biol.*, 2005, **6**, (2), pp. 99–111

[35] Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., *et al.*: 'CORUM: the comprehensive resource of mammalian protein complexes', *Nucleic Acids Res.*, 2008, **36**, (Database issue), pp. D646–D650

[36] Ruepp, A., Waegele, B., Lechner, M., *et al.*: 'CORUM: the comprehensive resource of mammalian protein complexes--2009', *Nucleic Acids Res.*, 2010, **38**, (Database issue), pp. D497–D501

[37] Kikugawa, S., Nishikata, K., Murakami, K., *et al.*: 'PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from H-invitational protein-protein interactions integrative dataset', *BMC Syst. Biol.*, 2012, **6**, (Suppl 2), p. S7

[38] Kanehisa, M., Furumichi, M., Tanabe, M., *et al.*: 'KEGG: new perspectives on genomes, pathways, diseases and drugs', *Nucleic Acids Res.*, 2017, **45**, (D1), pp. D353–D361

[39] Kanehisa, M., Goto, S.: 'KEGG: kyoto encyclopedia of genes and genomes', *Nucleic Acids Res.*, 2000, **28**, (1), pp. 27–30

[40] Mi, H., Huang, X., Muruganujan, A., *et al.*: 'Panther version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements', *Nucleic Acids Res.*, 2017, **45**, (D1), pp. D183–D189

[41] Thomas, P.D., Campbell, M.J., Kejariwal, A., *et al.*: 'Panther: a library of protein families and subfamilies indexed by function', *Genome Res.*, 2003, **13**, (9), pp. 2129–2141

[42] Croft, D., O'Kelly, G., Wu, G., *et al.*: 'Reactome: a database of reactions, pathways and biological processes', *Nucleic Acids Res.*, 2011, **39**, (Database issue), pp. D691–D697

[43] Joshi-Tope, G., Gillespie, M., Vastrik, I., *et al.*: 'Reactome: a knowledgebase of biological pathways', *Nucleic Acids Res.*, 2005, **33**, (Database issue), pp. D428–D432

[44] Liberzon, A.: 'A description of the molecular signatures database (MSigDB) web site', *Methods Mol. Biol.*, 2014, **1150**, pp. 153–160

[45] Liberzon, A., Birger, C., Thorvaldsdottir, H., *et al.*: 'The molecular signatures database (MSigDB) hallmark gene set collection', *Cell Syst.*, 2015, **1**, (6), pp. 417–425

[46] Shannon, P., Markiel, A., Ozier, O., *et al.*: 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome Res.*, 2003, **13**, (11), pp. 2498–2504

[47] Dong, J., Horvath, S.: 'Understanding network concepts in modules', *BMC Syst. Biol.*, 2007, **1**, p. 24

[48] Marquez, R.T., Xu, L.: 'Bcl-2:Beclin 1 complex: multiple, mechanisms regulating autophagy/apoptosis toggle switch', *Am. J. Cancer Res.*, 2012, **2**, (2), pp. 214–221

[49] Pattingre, S., Tassa, A., Qu, X., *et al.*: 'Bcl-2 antiapoptotic proteins inhibit Beclin 1-dependent autophagy', *Cell*, 2005, **122**, (6), pp. 927–939

[50] Huang, L., Guan, R.J., Pardee, A.B.: 'Evolution of transcriptional control from prokaryotic beginnings to eukaryotic complexities', *Crit. Rev. Eukaryot Gene. Expr.*, 1999, **9**, (3–4), pp. 175–182

[51] Sims, R.J.3rd, Mandal, S.S., Reinberg, D.: 'Recent highlights of RNA-polymerase-II-mediated transcription', *Curr. Opin. Cell Biol.*, 2004, **16**, (3), pp. 263–271

[52] Baillat, D., Hakimi, M.A., Naar, A.M., *et al.*: 'Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II', *Cell*, 2005, **123**, (2), pp. 265–276

[53] Gray, K.A., Yates, B., Seal, R.L., *et al.*: 'Genenames.org: the HGNC resources in 2015', *Nucleic Acids Res.*, 2015, **43**, (Database issue), pp. D1079–D1085

[54] Cheng, L., Lo, L.Y., Tang, N.L., *et al.*: 'CrossNorm: a novel normalization strategy for microarray data in cancers', *Sci. Rep.*, 2016, **6**, pp. 18898

[55] Cheng, L., Wang, X., Wong, P.K., *et al.*: 'ICN: a normalization method for gene expression data considering the over-expression of informative genes', *Mol. Biosyst.*, 2016, **12**, (10), pp. 3057–3066

[56] Yi, Y., Zhao, Y., Li, C., *et al.*: 'Raid V2.0: an updated resource of RNA-associated interactions across organisms', *Nucleic Acids Res.*, 2017, **45**, (D1), pp. D115–D118

[57] Hu, X., Wu, Y., Lu, Z.J., *et al.*: 'Analysis of sequencing data for probing RNA secondary structures and protein-RNA binding in studying posttranscriptional regulations', *Brief Bioinf.*, 2016, **17**, (6), pp. 1032–1043