國立臺灣大學與中央研究院合辦

基因體與系統生物學碩士學位學程

碩士論文

Genome and Systems Biology Degree Program,

National Taiwan University and Academia Sinica

Master Thesis

DynaPho: 由磷酸化蛋白質體資料推論動態化生物訊息

DynaPho: inferring signaling dynamics from phosphoproteomics data

王建凱

Jian-Kai Wang

指導老師：阮雪芬 博士

Adviser: Hsueh-Fen Juan, Ph. D.

指導老師：歐陽彥正 博士

Adviser: Yen-Jen Oyang, Ph. D.

中華民國 104 年 7 月

July, 2015

國立臺灣大學（碩）博士學位論文
口試委員會審定書

DynaPho: 由磷酸化蛋白質體資料推論動態化生物訊息
DynaPho: inferring signaling dynamics from phosphoproteomics data

本論文係王建凱（學號 r02b48005 ）在國立臺灣大學基因體與系統生物學學位學程完成之碩（博）士學位論文，於民國 104 年 7 月 27 日承下列考試委員審查通過及口試及格，特此證明

口試委員：阮雪芬　歐陽彥正.（簽名）

（指導教授）

林振慶.

王禹超

基因體與系統生物學學位學程主任　生科院基因體與系統生物學學位學程主任郭明良　（簽名）

致謝

碩班生兩年的日子即將邁入尾聲，回首一看，進入實驗室也像是不久前的事，感謝阮雪芬老師在我剛進入 GSB 學程時，讓我加入這個大家庭，遇到許多優秀的學長與一起奮鬥的同學，也讓我在資源充足的環境下成長許多。感謝黃宣誠老師，老師的思考模式與對學生的用心與耐心，讓我體會到一個優秀的學者應該具備的素養。感謝許家郎學長，實在幸運能在加入實驗室後，能在學長的指導下完成本專題與研究，學長時常提出許多應該注意的細節與良好的建議，才能讓這份研究專題具有深度與廣度，雖然兩年的時間不長，但我也從學長的引導下學習到許多關於分析生物資料的方法與架構，實在感謝學長的教導。謝謝黃振綜學長，是一位能分析事物透徹且自我要求甚多的學長，可能同在類似的領域成長過，不僅在學界、產業界的資訊我們都能時常交換意見，學長對於許多事鞭辟入理的批判，確實也給我不少新的想法。感謝我的同期同學們，謝謝昭胤、子霆、士傑、瑋庭與曉婷，這段研究生的同甘共苦確實讓我們都比大學時期來的更加堅毅，對於我們心中的許多夢想也有了更真實與現實的規劃，很感謝有你們的相伴，讓我的碩士生活過的雖勞累但也快樂，也謝謝你們大家對於我的忍耐與照顧。對於逵悅，從一開始 GSB 百人教授中能選在同一實驗室就足以說明我們的緣分，儘管學界著實有不少的障礙要去跨過，相信並祝福你能完成自己的目標，也謝謝你這兩年的照顧。對於凱普，一位願意對自己的命運奮鬥的勇士，你的認真與開朗常是我相當敬佩的對象，儘管自己身體的不足，但仍努力地往前大步邁進，無畏地接受挑戰，是我打從心底尊敬的對象，祝福你的人生規劃能一切順心與順利。學妹允芃，雖然時常聽到妳對於研究的不安，但未雨綢繆也未必不好，面對新領域的挑戰，除了需要更多的努力外，也需要相對的自信心，要相信自己能夠有足夠的毅力與耐心來克服，相信妳可以做得很好。對於新進的碩一學弟妹們，直到碩班畢業才會有許多感觸，回首兩年來的日子，有苦、有累、但也有收穫與屬於自己的成長，永遠不要設限自己，相信你們大家都能有許多的收益，不僅是在研究成果，更是在心裡的建設與成熟，也願大家在未來研究上、工作上一切順利。

CONTENT

FIGUREURE CONTENT

TABLE CONTENT

摘要

細胞中蛋白質的磷酸化不僅調控許多生理生化反應，更在許多病理狀況中扮演關鍵角色。近年來，研究磷酸蛋白質體技術的躍昇，例如製造更高精確辨析的質譜儀及發掘磷酸化胜肽鍊的技術提升，足夠研究能以磷酸化胺基酸位置為主的磷酸蛋白質體。透過許多改良的技術產生大量磷酸化蛋白體的數據便急需一個更新或精進的計算方法或分析流程將這些大量數據轉成可理解並有價值的資訊。DynaPho 為一個以網頁操作為基礎的分析工具，透過多種演算法分析磷酸化資料和包含磷酸化位點的序列，並整合各種資料庫註解及解析磷酸訊號的動態變化。DynaPho 的組成為一個前處理模組與五個分析模組，包含 (1) 敘述性統計分析；(2) 磷酸化數據分群分析；(3) 趨勢性、時間性功能豐富性分析；(4) 磷酸酵素活化時間分析；(5) 蛋白質交互作用的網路分析。我們透過分析人類子宮頸癌細胞在細胞週期各階段的磷酸蛋白質體巨量資料來說明 DynaPho 的分析功能與流程。透過分析磷酸化的胺基酸序列，不僅找出細胞週期中關鍵的 CDK 家族，更進一步分析出酵素活化的時間變化表，如 CDK1，在第一階段成長期、合成期與第二階段成長期有活化的現象。透過蛋白質交互網路更可以綜觀細胞在細胞週期中的連續性訊息的傳遞，如 RanBP2-ErbB2 的傳遞路徑等。因為 DynaPho 可運用於分析磷酸化蛋白質體動態訊息的變化，能使我們更加瞭解複雜的生物系統。DynaPho 可以透過網址 http://dynapho.hchuang.info/ 免費地連結使用。

ABSTRACT

Protein regulatory phosphorylation controls normal and pathophysiological signaling activities in cell. Recently, great advances in phosphorproteomics, including high-accuracy mass spectrometry (MS) and phosphopeptide-enrichment techniques, have allowed identifying site-specific phosphorylation. Development of computational analysis methods is required to transform large-scaled phosphoproteome data into valuable information of biological relevance. DynaPho is a web-based tool for analyzing temporal phosphoproteomes. It combines several algorithms to analyze the phosphorylation profiles as well as sequence-content of phosphosites and integrates various databases to annotate and uncover the dynamics of phosphosignaling. DynaPho consists of five major analysis modules: (1) description and summary of phosphoproteomics data; (2) clustering of phosphorylation profile; (3) temporal functional enrichment; (4) generation of kinase activation profile; and (5) temporal protein interaction network. We illustrate DynaPho via an analysis of massive phosphoproteomics dataset of cell cycle on HeLa cell. Based on the phosphorylation profiles, these data were divided into eight clusters corresponding to different cell cycle stages. The analysis of kinase activation profile revealed CDK family play a major role in cell cycle signaling. For instance, CDK1 is activated in G1, synthesis and G2 stage. The temporal protein interaction network discoveries RanBP2-ErbB2 signaling across mitosis and G1 stage. DynaPho can reveal the dynamics of temporal phosphoproteomics data contributing to improved understanding of complexity of biological systems. DynaPho is freely available at http://dynapho.hchuang.info/.

CHAPTER 1 INTRODUCTION

Protein phosphorylation is one of the post-translation modifications of protein that is an important factor in cellular signaling systems. It is a transient reaction which temporarily alters protein activities or complex responses by the addition of a phosphate. Abnormal regulation of phosphorylation is related to disease formation and progression, including cancer. Several drugs have been invented to provide better ways in treatments by targeting protein phosphorylation, such as Fasudil and Icotinib [1-3]. Site of phosphorylation is crucial for protein function of its efficiency. For example, different phosphorylation sites in neuropeptide $NPFF_2$ contributed significantly different Ca^{2+} releasing rates [4]. In nature, the most commonly phosphorylated amino acids contain a hydroxyl group (~ 17% of total residues which are serine (~8.5%), Thr (~5.7%) and Tyr (~3.0%)). About 700,000 potential phosphorylation sites exist if it is assumed there are ~10,000 different proteins with ~400 amino acids in average in a typical eukaryotic cell [5]. The increasing number of identified phosphorylation sites raises fundamental questions about their nature and biological relevance.

Modern mass spectrometry has provided accurate identification, high resolution and precise quantification for high-throughput proteomics. [6] Large datasets obtained by these techniques have been promoted for the development of customized analysis pipelines and facilitate for interpretation. Most of these pipelines supported large data repositories which store experimental details, such as PeptideAtlas [7], Human Proteinpedia [8], and NCBI Peptidome [9]. In recent years, platforms that combine more specific repositories for phosphorylation data and their compatible analysis tools have emerged, such as PhosphositePlus [10], Phospho.ELM [11], Scansite [12], and PPSP [13]. Repositories which stored several types of post-translational modification data have also been developed, such as SysPTM [14], and PTMScout [15]. To interpret large data from high-throughput experiments or repositories, customized computational tools were developed, including myProMS [16], PrestOMIC [17], PeptideDepot [18], ProteoConnections [19], MASPECTRAS2 [20], and Qupe [21]. Most tools emphasized the processing of MS data, such as peptide identification and protein searching as well as spectra quantification and general issues of protein expression. Other tools which were specific to interpret phosphorylation activities included NetworKIN [22], NetPhorest [23], and KinomeXplorer [24]. These tools are only specialized in sequence-based kinase signaling network modeling. The tool SELPHI [25] performed the phosphorylation peptide-based correlation analysis to interpret downstream cellular signaling. However, such analysis pipeline leaves either huge fortune of biological messages unexplored or other temporal regulating information behind.

Phosphorylation, a signal system, causes the transient status of protein property to response the environment change. These signals change as time goes by and represent what the condition the cell has undergone (Figure 1). Here, we develop DynaPho (Dynamic Phosphorylation), a web-based analysis platform that facilitates the exploration of global phosphoproteome datasets. DynaPho performs a data-driven analysis pipeline and distinguishes itself from other phosphorylation data analysis by focusing on facilitating the interpretation of temporal biological information. Users can upload their data which are preprocessed using MS quantification software, such as MaxQuant [26], or are self-calculated datasets. The data must contain the accession name, the phosphorylation site sequence and more than three time-coursed data. DynaPho analyzes datasets by using the clustering algorithm to identify temporally co-expression sets of phosphorylation events among serial time, using GO term functional enrichment analysis to infer temporal signaling changes, identifying conserved phosphorylation motifs to potential kinases by PSSM (position-specific scoring matrix), revealing the temporal activation profile of these kinases, and mapping modulated phosphosites onto temporal protein interaction network. Since DynaPho integrates valuable information from plenty of resources, including databases and tools, it can help to provide detailed phosphoproteomics information.

To demonstrate the ability of DynaPho, we re-analyzed the phosphoproteomics dataset of the cell cycle on HeLa cell [27]. After DynaPho preprocesses the raw data, 14,703 phosphosites are identified among six continuous stages (mitosis, G1, G1/S, early S, late S and G2) with high confidence and quantified for analyzing. Eight co-expressed clusters are identified for dynamic phosphorylation profiles. Function enrichment of these clusters not only infers the same process with the original result but also reveals temporal signaling changes of these biological processes. Several key kinases in cell cycle, including CDK1 and CDK3, are identified by peptide sequence similarity analysis and their temporal activation profiles were also inferred. Furthermore, protein interaction network, including kinase-substrate network, can assist to present the temporal signaling profile among different proteins as well as infer signaling from RanBP2 to ErbB2 across mitosis and G1 stages. DynaPho performs series of analyses and strengthens the temporal resolution to interpret the cellular signaling and dynamic biological information on phosphoproteome dataset.

CHAPTER 2 MATERIALS AND METHODS

2.1 Position in analyzing MS data

Generally, a MS-based experiment starts from a sample preparation. It includes adding protease and phosphatase (PTPs) inhibitors, extracting proteins, making the reduction of proteins for unfolding to a linear form, digesting proteins to peptides, labeling peptides on Lys and Arg based on different conditions, making series fractionation to separating peptides and enriching phosphorylation peptides by TiO_2 microbeads or antibodies in advanced [28]. All collected fractions were separated on a reverse-phase liquid chromatography (LC) and then electrosprayed into a mass spectrometer. The searching engines, such as MaxQuant, can identify MS spectrum and map onto proteins on the basis of MS spectral databank, such as MassBank [29]. The input data of DynaPho is the output data of the searching engine. DynaPho is used to analyze downstream cellular signaling and interprets biological datasets (Figure 2).

2.2 Input data format

Basically, DynaPho accepted labeling phosphosite datasets. The data format for each submission must contain accession name, phosphorylation peptides (over seven amino acids) and at least three labeling ratios on series time (Figure 3). The labeling ratio is not allowed to be transformed. The null or not detected labeling ratio can be represented by “NA”, “na” or blank. On the other hand, the label-free dataset can be transformed into ratio-like one for the submission.

2.3 Phosphosites among six stages of cell cycle in HeLa cell as a case study

The cell cycle is a highly conserved process which results in the duplication of cell’s content and molecular components. The progress of cell cycle is governed by the complex network of signaling pathways and also abides by regular time periods. We used the phosphosite dataset from the Olsen *et al*. investigation [27], and found total 24,714 phosphorylation events (FDR < 1%). 20,443 events of which were specific to a phosphorylation residue with high confidence (class I sites). The phosphosites were measured on six synchronous (by Thymidine and Nocodazole) and continuous stages (mitosis, G1, G1/S, early S, late S and G2). Furthermore, total 20,443 phosphosites (class I type) were filtered by the rule of no ratio change on all stages, including one (zero in log2 scaled) or null value. In the final, 14,703 phosphorylation events, which at least one stage was the perturbation status, were further analyzed.

2.4 Collect databases

DynaPho contains several databases which are used in different analysis modules, including function annotation module, kinase activation time profile module and protein interaction network module. The Gene Ontology (GO) biological process database was downloaded on 02/10/2015. Both the motif matrix for position specific scoring matrix and the motif logo repository were downloaded respectively from PhosphoNetworks on 12/10/2014 and 12/18/2014. The protein interaction databases were collected from BioGrid [30], HPRD [31], InAct [32], CCSB [33] and MINT [34] on 03/20/2015. All the above databases can be downloaded on the webpage of Dyanpho by FTP, HTTP, or the origin source. The Uniport database was directly downloaded by the instruction of uniprot.org with MySQL core on 03/20/2015.

2.5 Architecture and Sequential analyzing flowchart

The architecture and workflow of DynaPho is presented in Figure 4. The analyzing module of DynaPho was composed of six modules, including data preprocessing, statistical analysis, profile clustering, function enrichment, kinase activation profile, and interaction network. The uploaded data is first preprocessed by both filtering and filling procedures to make it reasonable. The statistical analysis is better executed on the next step to present proportions of each phosphorylation sites and the ratio distribution from total labeling ratios. The profile clustering module groups co-expressed phosphorylation sites according to ratio changes on the series time. If phosphorylation events are in the same cluster, the trend or expression change on the series time is similar for the dynamic regulation. The function enrichment module annotates phosphorylation sites based on the same cluster or one specific time point from GO biological process database. DynaPho automatically extracts conserved patterns by uploading all phosphorylation sequences to motif-x [35]. Furthermore, the kinase is inferred by conserved patterns with their peptides based on the position specific scoring matrix which is collected from PhosphoNetworks [36]. Temporal kinase activation profile is established by fisher's exact test. The interaction network presents temporal protein-protein interactions for the dynamics of signaling changes. Customization input parameters of each analysis module and their default values are listed in Table 1. Public databases or services integrated in Dyanpho are listed in Table 2.

2.6 Filter and Fill data in data preprocessing

In data preprocessing, DynaPho first filtered missing values, including "NA", "na", null or not detection caused by anthropic error or mechanical limitation, on all labeling events in one phosphorylation site. A low proportion of missing value was tolerated and also filled by the machine learning algorithm. DynaPho provided users with customized threshold to filter phosphorylation events (delete entire phosphorylation site with its ratios). The following format shows how to filter phosphorylated site in raw data. The set of all phosphorylation sites is symbolized by I and $\forall_{\mathrm{i}} \in$ I. $\mathrm{J_i}$ is all $\frac{\mathrm{H}}{\mathrm{L}}$ ratios of phosphorylation site i and $\forall\, \mathrm{j_i} \in \mathrm{J_i}$. Total $\frac{\mathrm{H}}{\mathrm{L}}$ ratios of phosphorylation site i is $\mathrm{n_i}$. T is the threshold defined by user to filter phosphorylation events.

$$\text{retains, when } \frac{\mathrm{n}_{\nexists \mathrm{j_i}}}{\mathrm{n_i}} \leq \mathrm{T}\,;\text{otherwise deletes the phsophorylation event}$$

After DynaPho filters raw data, it fills all missing value in each phosphorylation events by giving a real number from the other phosphorylation events whose values are not empty on the same time point. DynaPho provides users with two machine learning methods to fill missing value, average and k nearest neighbors (KNN). The average method is achieved by the following mathematical formula. The value of phosphorylation site i on time point j is $\mathrm{V_{ij}}$. For all missing values, DynaPho filled it by averaging all the other values on the same time point.

$$\text{if } \mathrm{V_{ij}} = \text{ missing value}, \mathrm{V_{ij}} = \overline{\sum_{x=1}^{I} (\exists \mathrm{V_{xj}})}$$

K nearest neighbors (KNN) is a supervised classification method in data mining or machine learning and ‘k’ means the number (= 1,2,3, …, n, n is a positive integer) of data points close to obvious one based on the criterion, for example, distance, similarity, etc. In DynaPho, KNN filled one missing value in a specific time by first calculating euclidean distance with each one of the other data points whose value existed. Sort the distance in increasing order and average values from the first k members to fill the missing data. If total nearest member is less than k, average all the remaining. The k nearest neighbors defined by users is $\mathrm{k_d}$. The euclidean distance list of phosphorylation site i in increasing order is $\mathrm{d_{in}^{i}}$ and the number of member in the list is $\mathrm{n\left(d_{in}^{i}\right)}$.

$$\text{if } \mathrm{n\left(d_{in}^{i}\right)} \geq \mathrm{k_d}, \nexists \mathrm{V_{ij}} = \overline{\sum_{x=1}^{\mathrm{k_d}} \exists \left(\mathrm{d_{in}^{i}}\right)_{xj}} \text{ ; otherwise ; } \nexists \mathrm{V_{ij}} = \overline{\sum_{x=1}^{\mathrm{n(d_{in}^{i})}} \exists (\mathrm{d_{in}^{i}})_{xj}}$$

The vector of phosphorylated site i containing all ratios without time point j is $\vec{\mathrm{V}}_{j}^{i}$.

$$(\mathrm{d_{in}^{i}})_j = \left\{ \left\|\vec{\mathrm{V}}_j^i - \vec{\mathrm{V}}_j^k\right\| \text{ in increasing} \,\middle|\, i, k \in I \text{ and } k \neq i \right\}, \text{and the distance}$$

$$\left\|\vec{\mathrm{V}}_j^i - \vec{\mathrm{V}}_j^k\right\| = \sqrt[2]{\sum_{y=1;\, y \neq j}^{n_J} (\mathrm{V_{iy}} - \mathrm{V_{ky}})^2 \text{ ; if } \nexists \mathrm{V_{iy}} \text{ or } \nexists \mathrm{V_{ky}}, (\mathrm{V_{iy}} - \mathrm{V_{ky}})^2 = 0}$$

After average- or KNN-based filling procedure, the data preprocessing is complete due to no ambiguous value on each phosphorylation site.

2.7 Workflow and methods in basic statistics module

After the data is preprocessed, DynaPho presented users with the status of modified data and the analyzing flowchart of statistics module in DynaPho is shown on Figure 5. DynaPho automatically calculated proportion of each phosphorylation site (the ratio of serine, threonine and tyrosine) and a statistical distribution plot presents the number of total phosphorylation sites with their centrality degrees for the perturbation degree and more the analysis potential. If the distribution is more similar with normal distribution, the effect caused by the perturbation is less in cell and the analyzing potential is also less. DynaPho provides users with two parameters for centrality degree, including $\frac{\text{Inter−Quartile Range (IQR)}}{1.35}$ and the standard deviation. All processed labeling ratios in increasing order is $\mathrm{R_{in}}$. Here $\mathrm{V_i}$ is ratio value in $\mathrm{R_{in}}$.

$$\frac{\mathrm{IQR} = (\mathrm{R_{in}})_{75\%} - (\mathrm{R_{in}})_{25\%}}{1.35} \cong \sqrt[2]{\frac{1}{\mathrm{N_{R_{in}}}} \sum_{\mathrm{i=1}}^{\mathrm{R_{in}}} \|\mathrm{V_i} - \overline{\mathrm{R_{in}}}\|} = \mathrm{S.D.}$$

If the value of $\frac{\mathrm{IQR}}{1.35}$ is more similar with standard deviation, less potential for analyzing and less fluctuation. The module also provided users with a trend chart of interested phosphorylation sites selected manually. The statistical analysis module is the foundation stone of the other analysis modules.

2.8 Workflow and methods in profile clustering module

Most signaling events are temporal regulations so that similar dynamics from different phosphorylation sites implies similar biological functions or unified biological intentions. In profile clustering module (Figure 6), DynaPho first calculates the clustering number for different types of dynamic phosphorylation profiles (phosphorylation changes on sequential time) or receive the one from the user-defined by field knowledge. Auto detection method for the clustering number is composed of three calculations, inner sample z-scored normalization, matrix transformation and trend clustering (Figure 7). Three fixed parameters involved are inner z-scored standard deviation (1.1 by default), variation threshold for all labeling ratios in specific time (over 0.01 S.D. by default) and the number threshold for the member in the cluster (over 1% of all phosphorylation events by default). The z-scored normalization transforms each labeling ratios in each phosphorylation events into a z-scored matrix. In single phosphorylation site i, the transformed value from the labeling ratio j is $\mathrm{Z_{ij}}$ and the standard deviation of all ratios is σ_i.

$$\mathrm{Z_{ij}} = \frac{(\mathrm{V_{ij}} - \overline{\mathrm{V}}_\imath)}{\sigma_\mathrm{i}}\ (\mathrm{z - scored\ normalization})$$

Z-scored matrix is further transformed into a three-status matrix (1, -1, 0 for up, down or no change) in each phosphorylation events. The inner z-scored standard deviation is $\mathrm{Z_{SD}}$. The new status value of each labeling ratio is $\mathrm{S_{ij}}$.

$$\mathrm{if\,Z_{ij}} \geq \mathrm{Z_{SD}},\ \mathrm{S_{ij}} = 1;\mathrm{else\ if\,Z_{ij}} \leq -1 \times \mathrm{Z_{SD}},\ \mathrm{S_{ij}} = -1;\mathrm{else\ S_{ij}} = 0$$

The status matrix is further processed by filtering specific time if its standard deviation is less than the variation threshold for removing redundancy information (status diversity of the specific time is less then variation threshold). The standard deviation of each time points in status matrix is $\mathrm{F_i}$ and the variation threshold is $\mathrm{F_{SD}}$.

$$\mathrm{remained, if\,F_i} \geq \mathrm{F_{SD}};\ \ \mathrm{removed, if}\ SD_{T_j} < \mathrm{F_{SD}}\ \mathrm{the\ entire\ status\ in\ specific\ time}$$

In trend clustering, DynaPho collects all trends (status vectors on time-scaled) existing on the status matrix and counts phosphorylation events of each trends. If the member number of one group is more than number threshold (proportion), it is seemed as a cluster; otherwise, ignore it. The remained status of one phosphorylation event i is $\overrightarrow{\mathrm{V}_\imath}$ (a vector of all statuses). The member of all the other phosphorylation events whose vectors are equal to $\overrightarrow{\mathrm{V}_\imath}$ is $\mathrm{n}(\overrightarrow{\mathrm{V}_\imath})$. If $\mathrm{n}\left(\overrightarrow{\mathrm{V}_\imath}\right) \geq \mathrm{number\ threshold}$, then remained; otherwise, ignored the cluster.

The clustering number is the parameter for clustering the phosphorylation profile by fuzzy c-means algorithm implemented in R named Mfuzz package [37]. Fuzzy c-means algorithm clusters phosphorylation events for similar profiles (similar status vectors). Fuzzy c-means algorithm is one kinds of soft clustering that one data vector is no more definitely belonging to one center but to use values ranging from 0 to 1 representing levels how it is related to centers. The relation between each data vector and each cluster center is a membership matrix. [38] Mfuzz first normalizes all labeling ratios in each phosphorylation events by z-scored with one standard deviation (similar with the auto-detection method). The C is a set of all m centers and c_i is a center in C with $i = 1 \dots m$. The V is a set of all n phosphorylation status vectors and v_j is a vector in V with $j = 1 \dots n$.

$$\mu_{c_i v_j} \text{ is the belonging level, and } \sum_{x=1}^{m} \mu_{c_x v_j} = 1, \text{ and } \forall j = 1 \dots n$$

For each status vector, the summary of all relationship levels with each center is 1. Initial step is to randomly generate the μ matrix fitting the above definition and it would be changed iteratively in order to find the optimized membership. The objective function F is defined as the following formula. The fuzzification value ranging from 1 to ∞ is M and the M value is 1.25 by default.

$$F_M = \sum_{j=1}^{n} \sum_{i=1}^{m} (\mu_{c_i v_j})^M \operatorname{dist}(v_j - c_i)^2$$

The distance function used in fuzzy c-means clustering is the Euclidean distance (the same in Mfuzz). If optimize the objective function, the function of each center with membership and status vector is the following formula.

$$c_i = \frac{\sum_{j=1}^{n} (\mu_{c_i v_j})^M v_j}{\sum_{j=1}^{n} (\mu_{c_i v_j})^M}$$

The membership value between each center and each data vector is the following.

$$\mu_{c_i v_j} = \frac{1}{\sum_{x=1}^{n} \left(\frac{\operatorname{dist}(x_j - c_i)}{\operatorname{dist}(x_j - c_x)} \right)^{\left(\frac{2}{M-1}\right)}}$$

In each iterator, calculate each center c_i ($\in C$) and then calculate new membership $\mu_{c_i v_j}$

between each center and each data vector. The converged condition of fuzzy c-means clustering is new value of objective function less than a threshold, or say it is much less than the previous value of objective function. In the Mfuzz, function mfuzz is main execution body and its output value, membership, presents the level how status vectors are related with the clustering centers. In each cluster, the member with low relationship is colored as blue; on the contrary, one with high relationship is colored as red. From the result calculated from fuzzy c-means clustering, clusters can be enriched with functions by gene ontology to interpret the signaling information.

2.9 Workflow and methods in function annotation module

The function of a set of proteins represents specific biological information in the cell and several mechanisms evidenced previously for responding to the perturbation collaboratively and effectively. For example, MyD88 and TRIP, activate different downstream phosphorylation signaling, ERK and JUK, to cooperatively respond to the infection in the beginning of the inflammation caused by lipopolysaccharide (LPS) [39]. Besides, opposite signaling taken by different sets of proteins may also achieve the same goal (one type of synergistic effects), for example, bufalin up-regulated DR4/DR5 and down-regulated Cbl-b for TRAIL-induced apoptosis in the breast cancer cell [40]. The cooperative mechanism is uncovered by the profile clustering module and the opposite one is revealed by filtering specific time profile. Users can directly process clusters from profile clustering module or set the threshold of fold change or standard deviation to the specific time profile for function enrichment analysis. In specific time profile analysis, the phosphorylation events are first filtered by fold change or standard deviation and then map onto uniprot session names without repeat ones. The function of selected proteins or clusters is enriched by GO analyzing and the hypergeometric test is used with the background selected by users from GO biological process database or the total non-repeated proteins (input). (Figure 8) The specific GO term is g. The protein number of the background is N. The number of non-repeated proteins is n. The number of proteins in the background with current GO term is M. The number of non-repeated proteins with current GO term is i.

$$\text{Hypergeometric Probability: } P(E = g) = \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

After hypergeometric testing, use Benjamini and Hochberg method to justify each p-values for false discovery rate (FDR) [41]. The total p-values of GO terms is m. The list of all m p-values in increasing order is P_{in}. The $\widetilde{P_{i}}$ is adjusted ith p-value in P_{in}.

$$\widetilde{P_i} = \min_{k=i\ldots m}\left\{\min\left(\frac{m}{k}P_k, 1\right)\right\}$$

The smaller adjusted p-value, the GO term is much possibly involved in the biological process. The function network analyzing links two related GO terms by the number of proteins involved in both (the node is GO term and the edge is intersected number of proteins) for the core activities. Furthermore, DynaPho provides users with the dynamics of biological progresses (GO terms) for deeply recovering complete signaling changes among temporal or cluster-based events in the cell. The dynamics is achieved and adjusted p-values is transformed by z-scored normalization.

2.10 Workflow and methods in kinase activation profile module

Kinases are keys in the phosphorylation signaling and also dynamic in temporal profiles. In this module, Dyanpho first finds conserved motif patterns by motif-x from phosphorylation sequences, compares the kinase PSSM with PhosphoNetworks databases evidenced in microarray platforms, clusters conserved patterns for similar kinases, and then constructs the temporal activation profile from these clusters by the fisher’s exact test (Figure 9). In the beginning, DynaPho collects three sets of phosphorylation sequences based on different central phosphosites, uploads each one of them into motif-x server with user-defined parameters, including occurences, significant threshold and reference (default values on Dyanpho is the same in motif-x, others parameters are also the same in motif-x but fixed and hidden in DynaPho) and then fetches the result. Motif-x conducts a statistical analysis by the binomial theorem possibility distribution to find successive significant residues. The core result contains several conserved motifs in different center phosphosites and their corresponding phosphorylation peptides. The PSSM is generated for each conserved motifs based on their sequence members and composed of each amino acid in x-axis, position relative to center phosphorylation site in y-axis (from -7 to 7 when the sequence length is 15) and percentages in the content. (Figure 10) The PSSM profile is further compared with the evidence-based database in PhosphoNetworks by pearson or spearman correlation analysis to discovery potential kinases. The PSSM database based on different phosphosites is P_{db}, p is one of them ($\forall p \in P_{db}$) and $p'$is further ranked in increasing order. The calculated PSSM based on conserved motifs is P_u, u is one of them ($\forall u \in P_u$) and u' is further ranked in increasing order. Total percentages in PSSM without the profile of center phosphosite are T_s (center phosphosite which is certainly high correlation causes the bias) and assume the length of phosphorylation sequence is 15, T_s is $(15 - 1) * 20 = 280$.

$$\mathrm{cor}(p,u)_{pearson} = \frac{\mathrm{covariance}(p,u)}{\sigma_p \sigma_u} = \frac{\sum_{i=1}^{T_s}(p_i - \overline{p})(u_i - \overline{u})}{\sqrt[2]{\sum_{i=1}^{T_s}(p_i - \overline{p})^2}\sqrt[2]{\sum_{i=1}^{T_s}(u_i - \overline{u})^2}}$$

$$\mathrm{cor}(p',u')_{spearman} = 1 - \frac{6\sum_{i=1}^{T_s}(d_i)^2}{T_s({T_s}^2 - 1)};\ \text{where } d_i = p'_i - u'_i$$

The correlation matrix reveals potential kinases involved but it is necessary to be simplified because different phosphorylation peptides possibly belong to the same kinases (due to short length of phosphorylation sequence). DynaPho automatically calculates clustering number by iteratively resampling to cluster conserved motifs and implements it by R package “clusterCons”. ClusterCons calculates the area under the curve (AUC) from the dataset in different clustering number, finds the largest change of AUC (the quantity ΔK), and then merges consensus clustering results from different algorithms [42]. Follow the instruction, clustering algorithms implemented in DynaPho is k-means, agglomerative nesting (agnes) and partitioning around medoids (pam) with fifty resampling. The clustering number determines groups of potential kinases by dividing the hierarchical clustering from conserved motifs in the correlation matrix. After DynaPho groups the conserved motifs into several clusters which kinase correlation profiles are similar, the fisher’s exact test is taken to analyze the correlation probabilities between the temporal profile and kinase clusters. In each time profile, DynaPho maps kinase clusters back to conserved motifs, extracts labeling ratios whose phosphorylation sequence belongs to these motifs and filters ratios by fold change or standard deviation to construct the contigency table. The number of labeling ratios crossing the threshold in the kinase cluster on the specific time is a and the other is c (not crossing the threshold). The number of labeling ratios crossing the threshold not in current kinase cluster on the specific time is b and the number of the other is d (not crossing the threshold). The n represents all number of labeling ratios ($n = a + b + c + d$). The C represents all clusters calculated from “clusterCons”, and c is one of cluster in C ($\forall c \in C$). The T represents all time profiles and t is one specific time ($\forall t \in T$).

$$P(\forall c, \forall t) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{(a!)(b!)(c!)(d!)(n!)}$$

The potential kinase activation or deactivation profile is generated by iteratively calculating p-values from up- or down-expression contigency table. The more significant p-value represents that high probabilities potential kinases in the specific cluster are possibly involved in the specific time. DynaPho also transforms p-values into $-1 \times \log 10$ scaled and colors them for the dynamic profile.

2.11 Workflow and methods in interaction network module

The phosphorylation signaling is composed of several proteins from upstream to downstream and achieved by their interactions to transfer chemical groups. However constraints to the experiment design, its results lose a part of important interaction messages and in some cases, transient interactions are the key of the signaling system. In interaction network module, DynaPho focuses on these two issues, one is to connect direct interactions based on the specific time and the other is to link intermediary interactions for simulating the transient signaling. DynaPho first filters labeling ratios based on fold change or standard deviation on each time profile map them back into a set of uniprot names. Two non-redundancy proteins are connected by searching evidenced-based interaction databases whether the interaction exists and further annotating their functions and types from Uniprot database, including kinase, transcription factor, phosphatase and the other protein (Figure 11). The intermediary interaction are established when two proteins do not interacted with each other, but they have the same hub protein which is over the threshold on the other time profile. After DynaPho constructs the interaction network, the global view of temporal phosphorylation signaling presents how the perturbation affects the phosphoproteome in the cell and how phosphorylation proteins influence another one to pass biological messages.

2.12 DynaPho implementation

DynaPho is constructed in LAMP (Linux 3.10.0_x86_64, Apache 2.4.6, MySQL 5.5.40, and PHP 5.4.16) system. It is composed of five sub-systems, including web interface, job deployment, task recording, base framework, and administration. The base framework subsystem as an information center stores all types of settings, the configuration, meta information, and used images. The setting and configuration are mainly related to available network location or the physical path. The meta information stores titles and details of each module. This subsystem also constructs the base framework of the web interface, including the composition of all webpages. In addition, it integrates jQuery EasyUI (http://www.jeasyui.com/) to achieve the tab-based operation. The base framework subsystem as a checkpoint examines the availability of the task ID (after a successful uploading and preprocessing raw data) or session ID (each analysis). Furthermore, it also checks the status of raw data and generates a unique ID to a task or a session.

The job deployment subsystem controls analyzing procedures of five modules. Each module contains three branches which are cooperative to achieve every analysis session. One is the task monitor which controls the progression of analysis (by two scripts progress.php and progress_body.php in each module). Second one is the collection of analyzing programs which combine several languages (Perl 5, Python 2.7.5, and R 3.1.1) and techniques (such as parallel computing, multi-task) into a hybridization computing for the better performance. The other is the web presentation which is specific to each analysis result. In addition, this subsystem deletes the task, which is not executed over seven days, by the job scheduling method.

The web interface subsystem integrates lots of resources, including Plotly (https://plot.ly/) used in statistics module, jQuery Flot (http://www.jqueryflottutorial.com/) used in statistics module, jQuery EasyUI (http://www.jeasyui.com/) used in the whole subsystem, and Cytoscape.js [43] used in functional enrichment module and interaction network module. The main architecture consists of html, CSS and javascript (including jQuery) for presenting analyzed results.

The recording subsystem stores available tasks, execution sessions, error (or warning) logs, and analyzed results. The subsystem is independent from the other three ones. It means that DynaPho allows users to execute analyses derived from different tasks.

The administration subsystem stores the contact information of users. The subsystem is operated under the authorization. DynaPho is a module-based platform so that it is potential for extending new analyzing module in the future. Detailed composition of DynaPho with their physical path is listed in Table 5.

.

CHAPTER 3 RESULTS

3.1 Data quality status monitors changes over the cell cycle

After DynaPho preprocesses raw data, 14,703 phosphorylation events (class I phosphosites) are remained and 5,740 ones were filtered. The proportion of each phosphosites to these events is presented in Figure 12. The proportion of serine, threonine and tyrosine are respectively 78%, 19% and 3%. The phosphosite number of serine, threonine and tyrosine are 11,526, 2,742 and 435 respectively (no ambiguous phosphosite exists). The ratio between three phosphosites is similar with previous study [5], but tyrosine-based peptides are discovered more than it. Different phosphosites play discrepancy roles in phosphorylation signaling; for instance, phosphorylation of tyrosine was stringently regulated than others. Its phosphorylation was related to cellular regulatory function and its signaling pathway which is the major role in complex organisms [28].

All labeling ratios from these phosphorylation events are further transformed into log2 scale in a distribution chart (Figure 13). There are total of 88,218 labeling ratios, 81,913 are under 2 S.D., 4,216 are between 2 S.D. and 3 S.D., and 2,089 are over 3 S.D. About seven percentages of total ratios is potential analyzing and distributes over 14,703 phosphorylation events. These ratios are under -3 (labeling change is 0.125) and over 4 (labeling change is 16) in log2 scale. The parameter (interquartile range (IQR) / 1.35) is about 0.541 and is dissimilar with standard deviation. These descriptive statistics parameters represent the analyzing potential in the cell cycle.

3.2 Dynamic phosphorylation profiles reveal unified biological information

Eight co-expression clusters are identified by the analysis of profile clustering module. (Figure 14) Eight clusters stand for eight different purposes and signaling systems in the cell cycle. More precisely, there are nine clusters because preprocessing procedure filters phosphorylation events which all labeling ratios are not changed. Ninth cluster may relate to housekeeping, homeostasis functions or is unallied to the cell cycle. The result of each cell cycle stage contains the dynamic phosphorylation profile, the member number of phosphorylation events in each cluster, and the analysis of biological process (GO) by function enrichment module. The table of each cluster presents biological processes whose p-values in log10 scale are the top. The member number among these eight clusters is similar except that fifth cluster contains 6,394 phosphorylation events (43.488%). The dynamic profile of this cluster is that labeling ratios are highly changed in mitosis stage, but no changed in the other ones. The biological process analysis of the cluster presents most phosphorylation events are involved in mitosis cell cycle process, including nucleic acid organization (chromatin organization) and cytoplasmic component organization (organelle organization, cytoskeleton organization, macromolecular complex assembly, cytoplasmic transport, protein complex assembly, single-organism intracellular transport). These processes coincide with not only mitosis stage but also the original article. Both show that about half the peptides are phosphorylated in mitosis phase.

The dynamic profile of each cluster perfectly coincides with each cell cycle stages do not include first and eighth cluster. The mitosis, G1, G1/S, early S, late S, and G2 stages are respectively to fifth, second, fourth, third, seventh, and sixth clusters. In function enrichment analyses of these six clusters, it is not hard to understand biological processes involved in specific cell cycle stage. For example, fourth cluster (in G1/S) is mainly related to synthesis processes that prepare for DNA replication, including chromosome organization, gene expression and chromatin organization. The dynamic profile of eighth cluster is a sub-group which should be a part of mitosis stage (fifth cluster) in the original article because the labeling is highly changed in mitosis stage but a lightly down changed on early S and late S stages. And its biological process analysis presents highly related to the mitosis stage and homeostasis (including regulation functions). It means that some phosphorylation events are up-regulated in mitosis stage but down-regulated in S stage for homeostasis or regulation purpose. The dynamic profile of the first cluster is that labeling changes are high in both mitosis and G1 stages. This cluster is also a sub-group which should be a part of G1 stage (second cluster) in the original article due to its result in Fig 3A. The biological process analysis indicates functions involved in both stages, including mitotic cell cycle process, microtubule-based process, and cytoskeleton organization. It implies that some phosphorylation events are in transition functions from mitosis to G1 stage. DynaPho strengthens the temporal resolution of phosphorylation events rather than transitional analyses. Also, DynaPho provides users more precise analysis algorithms to perform better clustering so that two more detailed co-expression sets are discovered.

3.3 Cellular signaling in temporal function profiles

After the analysis of function enrichment module, DynaPho summarizes core processes over all cycle stages in a functional network that nodes and edges are respectively biological processes (GO terms) and the proportion of intersection proteins. (Figure 15) Detailed biological processes with their adjusted p-values are listed in Table 4. In the functional network, each sub-network represents a set of biological processes for one or more cellular signaling (purposes). More precisely, these sub-networks correspond with stages in cell cycle. For example, the sub-networks located on the bottom and right are mitotic chromosome condensation, mitotic nuclear division and chromatin organization. These processes present biological functions involved in the mitosis stage. Besides, sub-network in the center is mainly for homeostasis and contains lots of biological processes.

DynaPho analyzes core and detailed processes which are presented respectively in a network and a list on each temporal profile (on each cell cycle stage). The list contains biological processes which their adjusted p-values in log10 scale are the top. (Figure 16A - F). It is easy to map the detailed processes into the summarized network on each temporal profile. For example, in Figure 16D, several biological processes, including cellular protein complex assembly, cellular protein localization, DNA packaging, and regulation of chromosome segregation are highly related to synthesis stage. These processes are further summarized into two functional networks located on the upper right and bottom.

DynaPho reinforces the dynamic analysis with a temporal heatmap profile. Summarize all biological processes with their adjusted p-values in a heatmap. The adjusted p-values are first transformed into log10 scale and further normalized by z-score. (Figure 17) In the temporal heatmap, DynaPho embraces all biological processes and their dynamic signaling among all cell cycle stages. For example, the processes which the regulation of chromosome segregation and regulation of microtubule polymerization or depolymerization are mainly involved in early S stage (time 4). The processes for nuclear envelope disassembly, chromatin organization and cytoskeleton organization are mainly in mitosis (time 1). DynaPho provides comprehensive analyses rather than transitional bioinformatics analyses.

3.4 Regulated phosphoproteome by potential kinases

DynaPho uncovers dynamic activation profiles of kinases after the analysis of kinase activation profile module. The phosphorylation sequences are sent to motif-x service based on different phosphosites and then DynaPho fetches several conserved motifs. (Table 3) Seventy-three serine-based motifs are found, twelve threonine-based ones are found, but no conserved motif exists when the phosphosite is tyrosine. Tyrosine-based motif is not found are the parameter settings (occurrences and significance) in motif-x due to maintenance of a low false positive rate.

The conserved motifs with their contribution sequences are valuable information for kinase sequence profiles. The evidenced sequence profiles maintained by PhosphoNetworks are further compared with ones generated by DynaPho from motif-x. The kinase similarity between PhosphoNetworks databases and sequence profiles from motif-x is presented in a heatmap. (Figure 18) The x-axis is the conserved motif whose phosphosite is serine or threonine, and the y-axis is the kinase. The text in white with grey background on the top of heatmap presents the cluster of conserved motifs whose kinase profiles are similar. The darker red represents higher similarity between conserved motifs and kinases; in the contrary, the light green represents lower similarity. Three sections on the heatmap are relatively higher similarity. The corresponding conserved motifs and kinases are listed in Figure 19.

In the top of Figure 19, the temporal profiles of kinase activation and deactivation are also presented (the same with Figure 11). The adjusted p-values are transformed into log10 scale and the value is further multiplied by -1. The cluster 1 contains sixteen kinases, including cell cycle-related CDK family. The dynamic profile presents that kinases in cluster 1 are potential activation in G1, early S and G2 stages. The result is similar to the previous study that CDK1 was involved in G1 and G2 stages [44]. The cluster 2 contains thirteen kinases and most of them are related with cellular homeostasis or regulation. For example, both AKT1 and PAK4 were involved in homeostasis functions [45, 46]. Therefore, the activation and deactivation profile of cluster 2 are not significantly changed. DynaPho strengthens the analysis of phosphoproteome on potential kinases rather than transitional pathway analyses.

3.5 Phosphorylation signaling in cell cycle by protein interaction network

In interaction network module, DynaPho links sequential phosphorylation events across all cell cycle stages for the comprehensive signaling. Networks on Figure 20 A, B, C, D, E, and F are respectively the interaction network on mitosis, G1, G1/S, early S, late S, and G2 stage. Four shapes, including triangle, rectangle, hexagon, and circle, are respectively transcription factor, phosphatase, kinase, and the other protein type. The nodes and edges are respectively proteins (in gene name) and interactions. The interaction is composed of two types that are linking in the same stage (solid line) and linking across different stages (dashed line). The color of protein stands for its labeling ratio. The proteins in grey represented their labeling ratios that are filtered in the stage.

The interaction means the signaling event; for example, in G1 stage (Figure 20B) kinase EGFR phosphorylates transcription factor STAT3. Such signaling was evidenced by previous study that EGFR-STAT pathway is involved in liver regeneration on G1/S stage [47]. In advanced, DynaPho links signaling information across cell cycle stages. For example, the interaction is linked between RanBP2 (Nup358) and ErbB2 across mitosis (Figure 20A) and G1 (Figure 20B) stage. In previous study, the cell membrane-embedded ErbB2 activates PI3K-signaling pathways which constitute important regulation in G1 stage. It migrates from the cell surface to the nucleus through endocytosis process by interacting with a nuclear pore protein RanBP2 as a traffic light [48-50]. DynaPho reinforces the analysis to construct a dynamic network across different cell cycle stages for further validation rather than the analysis in single stage.

CHAPTER 4 DISCUSSION

DynaPho is the integrated analysis platform for the dynamic signaling. DynaPho is distinguishable from other phosphoproteome analysis pipelines by focusing on facilitating the interpretation of temporal biological signaling. On the timeline of interpreting MS data, many well-known tools, including myProMS [16], PrestOMIC [17], and PeptideDepot [18], were developed to solve problems for identification and quantification, or their derived issues, such as differential expression. Nearly a decade, several tools were developed to extract biological information, such as ProteoConnections [19]. Furthermore, tools aimed to specific issues were also developed; for example, NetPhorest [23] focused on the study of kinome, and KinomeXplorer [24] highlighted modeling kinase-substrate interactions. DynaPho takes another approach to discovery dynamic biological signaling and further to interpret them. DynaPho implements the trend detection and further soft clustering to group co-expression phosphorylation events on temporal profiles. These groups (clusters) are probably related in biological functions so that their biological processes are further analyzed by GO in the function enrichment module. Besides, the temporal changes of biological processes present when GO term is highly involved. DynaPho constructs an interaction network across the whole time (all cell cycle stages in the article) to present the complete signaling, including the one across different time points (stages in the article). Certainly, another service, named SELPHI [25], takes similar approaches to function enrichment module to infer the dynamics of pathways or similar networks. For instance, SELPHI constructs different types of networks, including kinase-kinase, and kinase-phosphatase. There are essential differences; for example, SELPHI focused on pathway comparisons among several experiments, or it conducted correlation analysis between different kinase phosphopeptides (or phosphatase phosphopeptides) and their associated phosphopeptides. In general, DynaPho is a brand-new analysis platform to comprehensively model dynamic signaling in the cell.

There are still several improvements. DynaPho is a data-driven analysis tool so that the status of original dataset is important. In the article, there is the bias or noise in raw data, even after preprocessing, because the average of total labeling ratios is slightly away from 0 and some of them are extremely high or low. The phosphorylation information is hardly unique or highly specific to one protein or sequence. This common phenomenon is caused by MS techniques or searching engines so that it makes the validation hard to proceed. It also makes users confused on ambiguous sequences or proteins. The bias or noise probably influences the construction of interaction network. General issues about GO term analysis are non-specific and redundancy information. These conditions also exist in DynaPho. Taking more GO terms into considerations is better for comprehensive analyses. In kinase activation module, sequence information does imply potential kinases, but similar sequence profile is not enough for similar functions of different kinases; for example, CDK1 and CDK2. The future work can focus how to analyze kinase activation profiles without redundancy sequence information. DynaPho integrates lots of resources and is implemented by plenty of programming techniques in order to provide users with better performance not only in the execution time but also in the web-based interaction. Hardware constraint or inadequate software skills also cause worse performance when extremely large scaled dataset is analyzed. DynaPho will continue to extend or enhance the phosphoproteome analysis by integrating or replacing analysis module for more services to non-bioinformatics experts in the future.

CHAPTER 5 CONCLUSIONS

DynaPho is a web-based and user-friendly platform to comprehensively analyze temporal phosphoproteome datasets. It consists of one preprocessing and five sequential analyzing modules to infer the dynamic phosphorylation signaling. In human HeLa cell cycle dataset, statistical analysis reveals that seven percentage of labeling ratios (6,305 ratios over 2 S.D.) are potential for analyzing and distribute over 14,703 phosphorylation events. After the analysis of profile clustering, eight co-expression profiles are identified. They are further analyzed by function enrichment module. It not only reveals unified biological information but also resolves more deep into the dynamic phosphorylation profiles. After the analysis of function enrichment, DynaPho summarizes core processes over all cycle stages in a functional network and also reveals detailed biological processes. Besides, DynaPho also embraces all biological processes and their dynamic signaling among all cell cycle stages in a heatmap. After the analysis of kinase activation profile module, DynaPho finds potential kinases and further presents the temporal profiles of both activation and deactivation. For instance, both AKT1 and PAK4 are involved in homeostasis functions and CDK1 involved in G1, S, and G2 stage. In the interaction network module, DynaPho links sequential phosphorylation events across all cell cycle stages for the comprehensive signaling, such as EGFR-STAT pathway in G1/S stage and the signaling from RanBP2 to ErbB2 in mitosis/G1 stage. DynaPho improves many shortages of traditional analyses and strengthens the analysis of phosphoproteome. The advancement of modern mass spectrometry technology and the integrity of bioinformatics analyses, to make the analysis of dynamic signaling cell is possible.

REFERENCES

1. Cohen, P., *The origins of protein phosphorylation.* Nat Cell Biol, 2002. **4**(5): p. E127-30.
2. Ying, H., et al., *The Rho kinase inhibitor fasudil inhibits tumor progression in human and rat tumor models.* Mol Cancer Ther, 2006. **5**(9): p. 2158-64.
3. Liu, D., et al., *Clinical pharmacokinetics of Icotinib, an anti-cancer drug: evaluation of dose proportionality, food effect, and tolerability in healthy subjects.* Cancer Chemother Pharmacol, 2014. **73**(4): p. 721-7.
4. Bray, L., et al., *Identification and functional characterization of the phosphorylation sites of the neuropeptide FF2 receptor.* J Biol Chem, 2014. **289**(49): p. 33754-66.
5. Ubersax, J.A. and J.E. Ferrell, Jr., *Mechanisms of specificity in protein phosphorylation.* Nat Rev Mol Cell Biol, 2007. **8**(7): p. 530-41.
6. Choudhary, C. and M. Mann, *Decoding signalling networks by mass spectrometry-based proteomics.* Nat Rev Mol Cell Biol, 2010. **11**(6): p. 427-39.
7. Deutsch, E.W., H. Lam, and R. Aebersold, *PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows.* EMBO Rep, 2008. **9**(5): p. 429-34.
8. Mathivanan, S., et al., *Human Proteinpedia enables sharing of human protein data.* Nat Biotechnol, 2008. **26**(2): p. 164-7.
9. Slotta, D.J., T. Barrett, and R. Edgar, *NCBI Peptidome: a new public repository for mass spectrometry peptide identifications.* Nat Biotechnol, 2009. **27**(7): p. 600-1.
10. Hornbeck, P.V., et al., *PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse.* Nucleic Acids Res, 2012. **40**(Database issue): p. D261-70.

11. Dinkel, H., et al., *Phospho.ELM: a database of phosphorylation sites--update 2011.* Nucleic Acids Res, 2011. **39**(Database issue): p. D261-7.

12. Obenauer, J.C., L.C. Cantley, and M.B. Yaffe, *Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs.* Nucleic Acids Res, 2003. **31**(13): p. 3635-41.

13. Xue, Y., et al., *PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory.* BMC Bioinformatics, 2006. **7**: p. 163.

14. Li, H., et al., *SysPTM: a systematic resource for proteomic research on post-translational modifications.* Mol Cell Proteomics, 2009. **8**(8): p. 1839-49.

15. Naegle, K.M., et al., *PTMScout, a Web resource for analysis of high throughput post-translational proteomics studies.* Mol Cell Proteomics, 2010. **9**(11): p. 2558-70.

16. Poullet, P., S. Carpentier, and E. Barillot, *myProMS, a web server for management and validation of mass spectrometry-based proteomic data.* Proteomics, 2007. **7**(15): p. 2553-6.

17. Howes, C.G. and L.J. Foster, *PrestOMIC, an open source application for dissemination of proteomic datasets by individual laboratories.* Proteome Sci, 2007. **5**: p. 8.

18. Yu, K. and A.R. Salomon, *PeptideDepot: flexible relational database for visual analysis of quantitative proteomic data and integration of existing protein information.* Proteomics, 2009. **9**(23): p. 5350-8.

19. Courcelles, M., et al., *ProteoConnections: a bioinformatics platform to facilitate proteome and phosphoproteome analyses.* Proteomics, 2011. **11**(13): p. 2654-71.

20. Ubaida Mohien, C., et al., *MASPECTRAS 2: An integration and analysis platform for proteomic data.* Proteomics, 2010. **10**(14): p. 2719-22.

21. Albaum, S.P., et al., *Qupe--a Rich Internet Application to take a step forward in the analysis of mass spectrometry-based quantitative proteomics experiments.*

Bioinformatics, 2009. **25**(23): p. 3128-34.

22. Linding, R., et al., *NetworKIN: a resource for exploring cellular phosphorylation networks.* Nucleic Acids Res, 2008. **36**(Database issue): p. D695-9.

23. Miller, M.L., et al., *Linear motif atlas for phosphorylation-dependent signaling.* Sci Signal, 2008. **1**(35): p. ra2.

24. Horn, H., et al., *KinomeXplorer: an integrated platform for kinome biology studies.* Nat Methods, 2014. **11**(6): p. 603-4.

25. Petsalaki, E., et al., *SELPHI: correlation-based identification of kinase-associated networks from global phospho-proteomics data sets.* Nucleic Acids Res, 2015. **43**(W1): p. W276-82.

26. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.

27. Olsen, J.V., et al., *Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis.* Sci Signal, 2010. **3**(104): p. ra3.

28. Sharma, K., et al., *Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling.* Cell Rep, 2014. **8**(5): p. 1583-94.

29. Horai, H., et al., *MassBank: a public repository for sharing mass spectral data for life sciences.* J Mass Spectrom, 2010. **45**(7): p. 703-14.

30. Chatr-Aryamontri, A., et al., *The BioGRID interaction database: 2015 update.* Nucleic Acids Res, 2015. **43**(Database issue): p. D470-8.

31. Keshava Prasad, T.S., et al., *Human Protein Reference Database--2009 update.* Nucleic Acids Res, 2009. **37**(Database issue): p. D767-72.

32. Hermjakob, H., et al., *IntAct: an open source molecular interaction database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D452-5.

33. Rual, J.F., et al., *Towards a proteome-scale map of the human protein-protein*

interaction network. Nature, 2005. **437**(7062): p. 1173-8.

34. Licata, L., et al., *MINT, the molecular interaction database: 2012 update.* Nucleic Acids Res, 2012. **40**(Database issue): p. D857-61.

35. Schwartz, D. and S.P. Gygi, *An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets.* Nat Biotechnol, 2005. **23**(11): p. 1391-8.

36. Hu, J., et al., *PhosphoNetworks: a database for human phosphorylation networks.* Bioinformatics, 2014. **30**(1): p. 141-2.

37. Kumar, L. and E.F. M, *Mfuzz: a software package for soft clustering of microarray data.* Bioinformation, 2007. **2**(1): p. 5-7.

38. J., C., Dunn, *A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters.* Journal of Cybernetics, 1974. **3**(3): p. 32-57.

39. Newton, K. and V.M. Dixit, *Signaling in innate immunity and inflammation.* Cold Spring Harb Perspect Biol, 2012. **4**(3).

40. Yan, S., et al., *Down-regulation of Cbl-b by bufalin results in up-regulation of DR4/DR5 and sensitization of TRAIL-induced apoptosis in breast cancer cells.* J Cancer Res Clin Oncol, 2012. **138**(8): p. 1279-89.

41. Yoav, B. and H. Yosef, *Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society, 1995. **57**(1): p. 289-300.

42. Simpson, T.I., J.D. Armstrong, and A.P. Jarman, *Merged consensus clustering to assess and improve class discovery with microarray data.* BMC Bioinformatics, 2010. **11**: p. 590.

43. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.

44. Enserink, J.M. and R.D. Kolodner, *An overview of Cdk1-controlled targets and*

processes. Cell Div, 2010. **5**: p. 11.

45. Altomare, D.A. and A.R. Khaled, *Homeostasis and the importance for a balance between AKT/mTOR activity and intracellular signaling.* Curr Med Chem, 2012. **19**(22): p. 3748-62.

46. Guo, Q., et al., *PAK4 kinase-mediated SCG10 phosphorylation involved in gastric cancer metastasis.* Oncogene, 2014. **33**(25): p. 3277-87.

47. Collin de L'hortet, A., H. Gilgenkrantz, and J.E. Guidotti, *EGFR: A Master Piece in G1/S Phase Transition of Liver Regeneration.* Int J Hepatol, 2012. **2012**: p. 476910.

48. Giri, D.K., et al., *Endosomal transport of ErbB-2: mechanism for nuclear entry of the cell surface receptor.* Mol Cell Biol, 2005. **25**(24): p. 11005-18.

49. Renner, O., et al., *Mst1, RanBP2 and eIF4G are new markers for in vivo PI3K activation in murine and human prostate.* Carcinogenesis, 2007. **28**(7): p. 1418-25.

50. Vadlakonda, L., M. Pasupuleti, and R. Pallu, *Role of PI3K-AKT-mTOR and Wnt Signaling Pathways in Transition of G1-S Phase of Cell Cycle in Cancer Cells.* Front Oncol, 2013. **3**: p. 85.

51. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

FIGURES

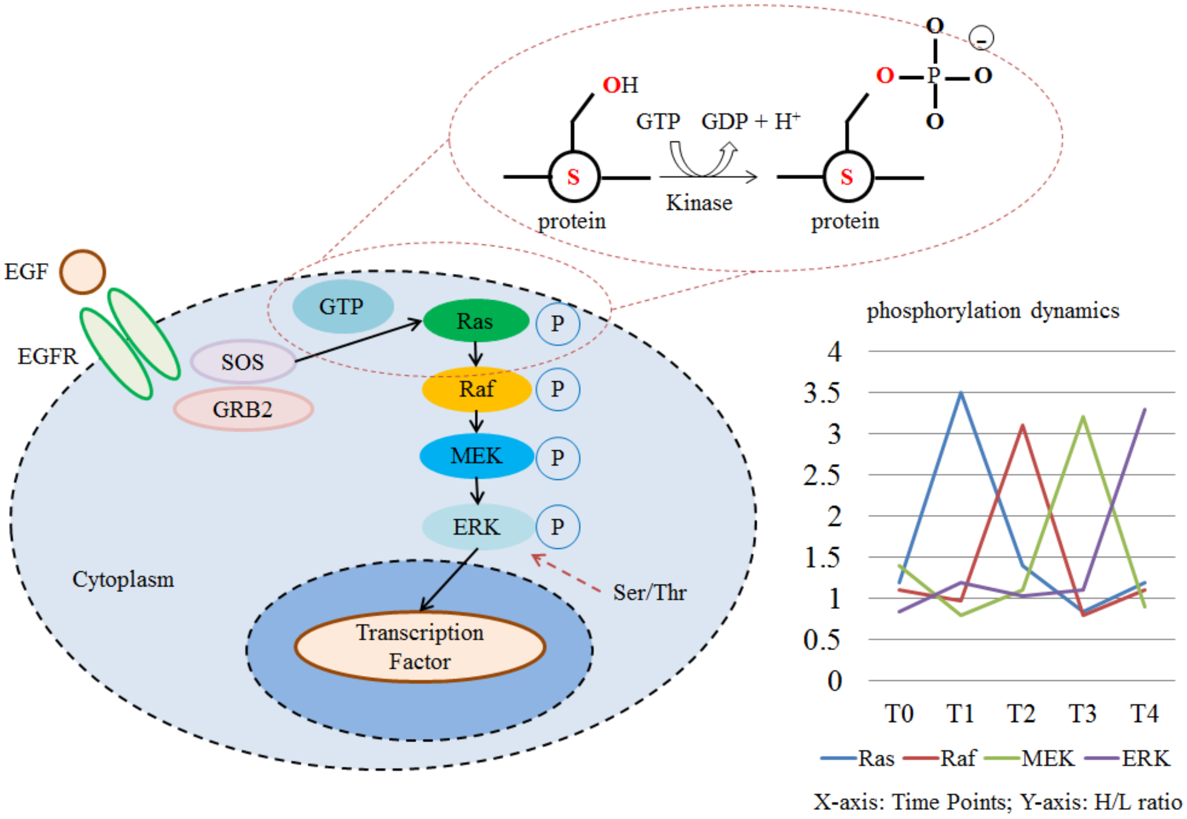


Figure 1 Dynamic signaling represents what conditions the cell had undergone.

The example RAS-RAF-MEK-ERK pathway is activated by epidermal growth factor (EGF). After the cell is simulated by EGF, kinases transfer a phosphate group from GTP to RAS protein. The signaling starts from RAS protein to extracellular-signal-regulated kinases (ERK). The phosphorylated ERK activates different types of transcription factors. The activated transcription factor starts downstream gene expression. If experiment on sequential time points under control and test conditions, it is highly possible to capture the dynamics of multiple proteins from MS data. These dynamics are the best interpreter what the cell had undergone.

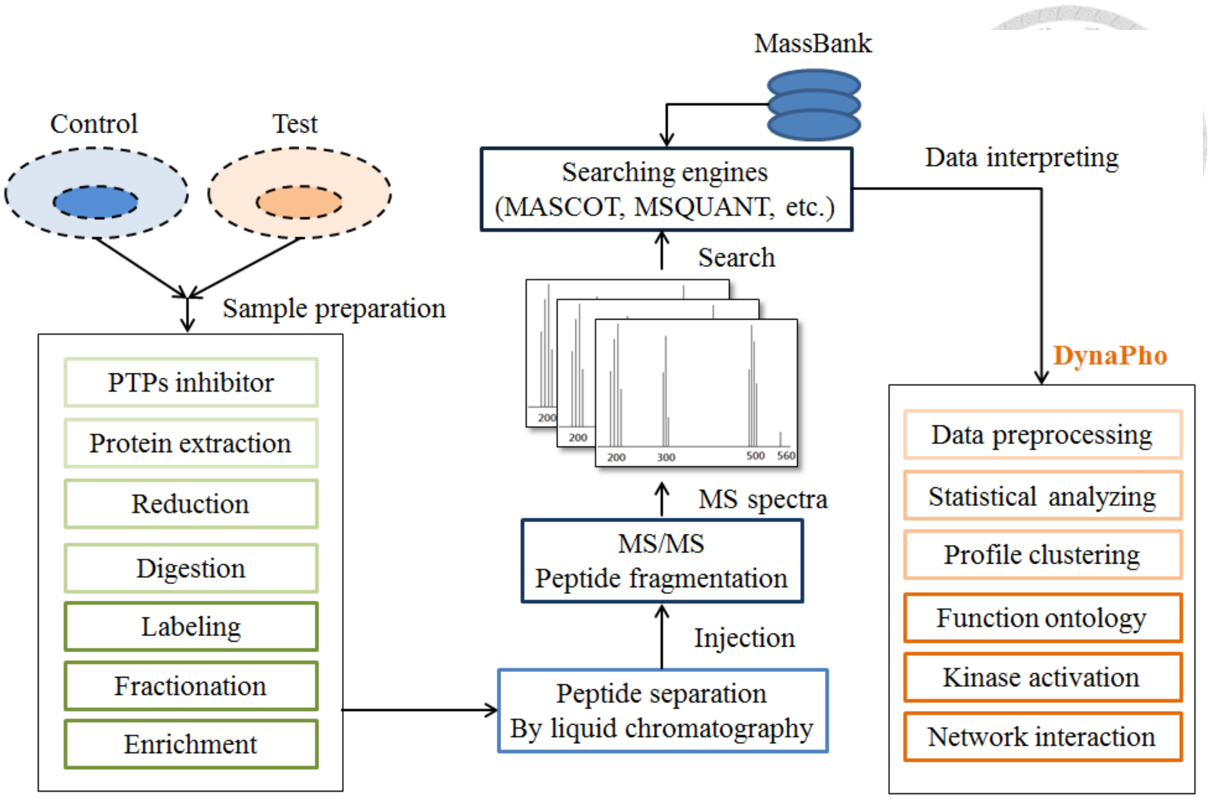
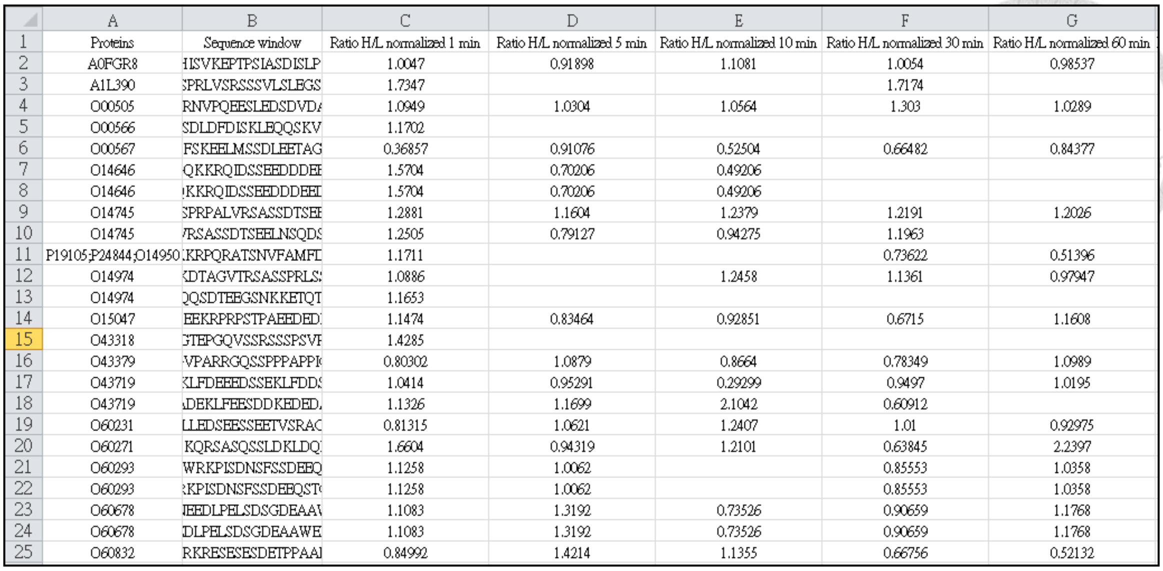


Figure 2 DynaPho interprets biological information on the downstream of analyzing.

The control and test samples are first prepared on multiple steps in order to get linear phosphorylation peptides as more as possible. The phosphorylation peptides are separated by liquid chromatography (LC) and then identified by mass spectrometry (MS). The spectrum of peptides and their corresponding proteins can be identified or mapped back by searching engine on the basis of spectral databases (for example MassBank). The raw data generated by the searching engine contains temporal labeling ratios, protein session names and phosphosite sequences, etc. Such raw data can be processed into the original data for DynaPho. DynaPho is located on the downstream of flow of MS data interpreting.



	A	B	C	D	E	F	G
1	Proteins	Sequence window	Ratio H/L normalized 1 min	Ratio H/L normalized 5 min	Ratio H/L normalized 10 min	Ratio H/L normalized 30 min	Ratio H/L normalized 60 min
2	A0FGR8	HISVKEPTPSIASDISLP	1.0047	0.91898	1.1081	1.0054	0.98537
3	A1L390	SPRLVSRSSSVLSLEGS	1.7347			1.7174	
4	O00505	RNVPQEESLEDSDVDA	1.0949	1.0304	1.0564	1.303	1.0289
5	O00566	SDLDFDISKLEQQSKV	1.1702				
6	O00567	FSKEELMSSDLEETAG	0.36857	0.91076	0.52504	0.66482	0.84377
7	O14646	QKKRQIDSSEEDDDEE	1.5704	0.70206	0.49206		
8	O14646	KKRQIDSSEEDDDEED	1.5704	0.70206	0.49206		
9	O14745	SPRPALVRSASSDTSEE	1.2881	1.1604	1.2379	1.2191	1.2026
10	O14745	VRSASSDTSEELNSQDS	1.2505	0.79127	0.94275	1.1963	
11	P19105;P24844;O14950	KRPQRATSNVFAMFD	1.1711			0.73622	0.51396
12	O14974	KDTAGVTRSASSPRLS	1.0886		1.2458	1.1361	0.97947
13	O14974	QQSDTEEGSNKKETQT	1.1653				
14	O15047	EEKRPRPSTPAEEDED	1.1474	0.83464	0.92851	0.6715	1.1608
15	O43318	GTEPGQVSSRSSSPSVP	1.4285				
16	O43379	VPARRGQSSPPPAPPI	0.80302	1.0879	0.8664	0.78349	1.0989
17	O43719	KLFDEEEDSSEKLFDDS	1.0414	0.95291	0.29299	0.9497	1.0195
18	O43719	ADEKLFEESDDKEDED	1.1326	1.1699	2.1042	0.60912	
19	O60231	LLEDSEESSEETVSRAC	0.81315	1.0621	1.2407	1.01	0.92975
20	O60271	KQRSASQSSLDKLDQ	1.6604	0.94319	1.2101	0.63845	2.2397
21	O60293	WRKPISDNSFSSDEEQ	1.1258	1.0062		0.85553	1.0358
22	O60293	RKPISDNSFSSDEEQST	1.1258	1.0062		0.85553	1.0358
23	O60678	NEEDLPELSDSGDEAAV	1.1083	1.3192	0.73526	0.90659	1.1768
24	O60678	DLPELSDSGDEAAWE	1.1083	1.3192	0.73526	0.90659	1.1768
25	O60832	RKRESESESDETPPAAI	0.84992	1.4214	1.1355	0.66756	0.52132

Figure 3 The basic format of the upload file

The format accepted by DynaPho is a table prepared by processing the raw phosphosite data from MSQuant or manually generating from non-labeling datasets. Constraints on the dataset include more than fifteen phosphorylated events. Each one contains more than one uniprot accession name, more than seven amino acids on phosphorylation sequence and more than three labeling ratios. The uniprot accession name or phosphorylation sequences can be multiple in the same column and be separated by a semicolon (;).

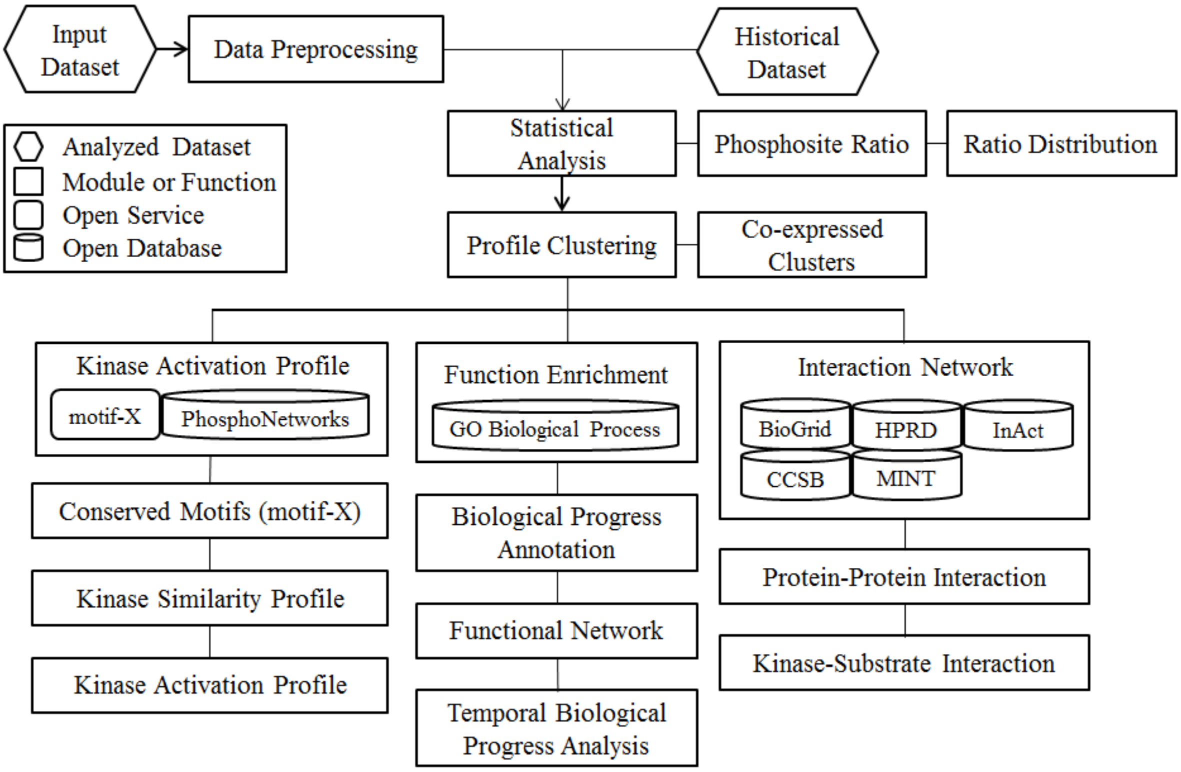


Figure 4 The architecture and workflow of Dyanpho

Users can start analyzing data from new upload file or the historical one. Each new upload file must be preprocessed first. Suggested analysis flow starts from statistical analysis, profile clustering, function enrichment, kinase activation profile and then interaction network. Crossing analysis also exists in DynaPho, the result from profile clustering module can be further analyzed by function enrichment module.

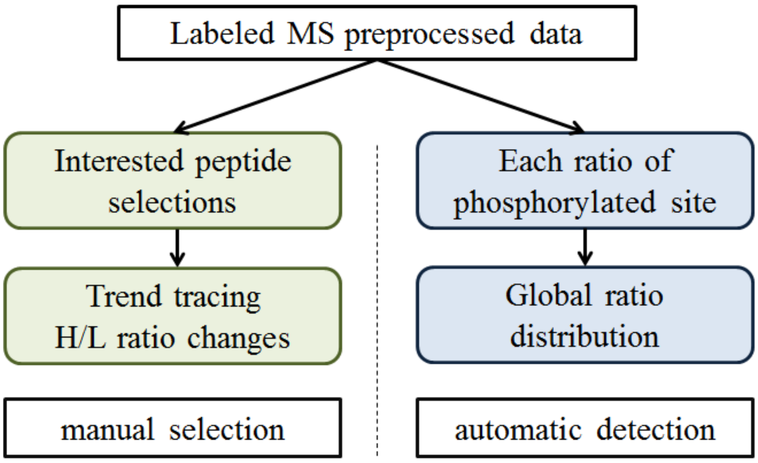

Labeled MS preprocessed data
Interested peptide selections
Trend tracing H/L ratio changes
manual selection
Each ratio of phosphorylated site
Global ratio distribution
automatic detection

Figure 5 The analyzing flow of statistics module

The statistics module is composed of two separated analyses, one is the proportion of each phosphorylated site and the distribution chart of total labeling ratios, and the other one is plotting labeling ratio changes of interested phosphorylation events selected by users.

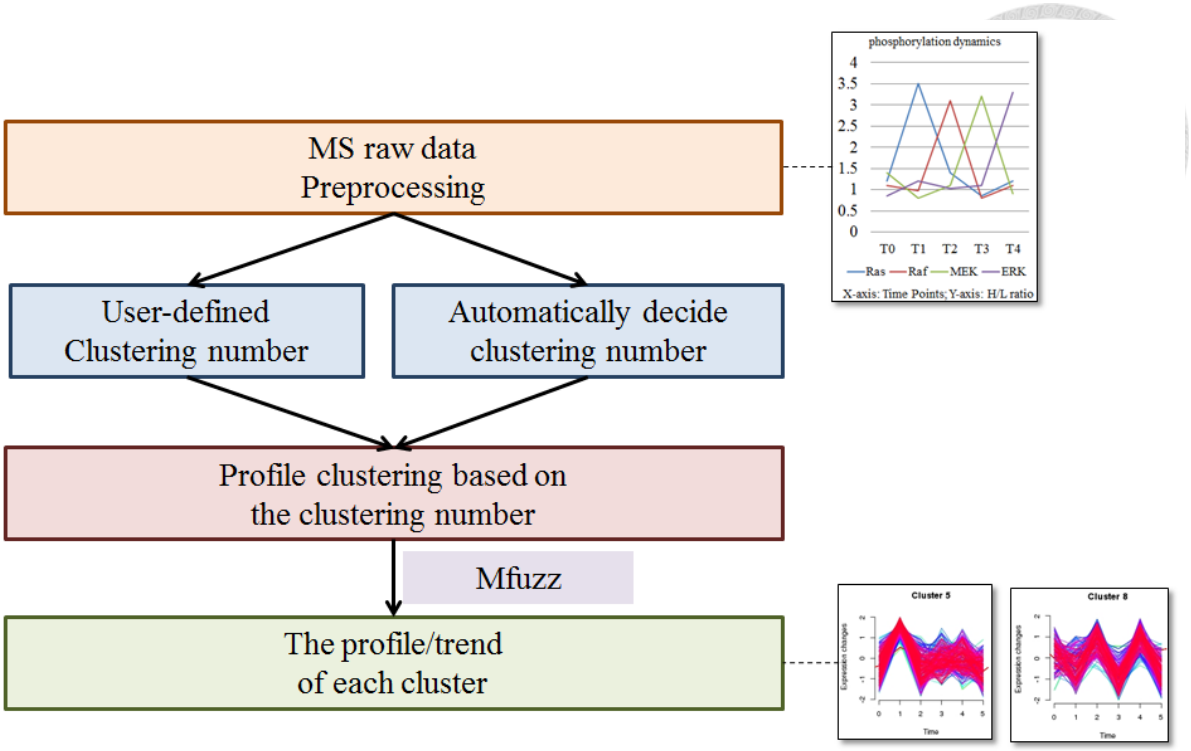
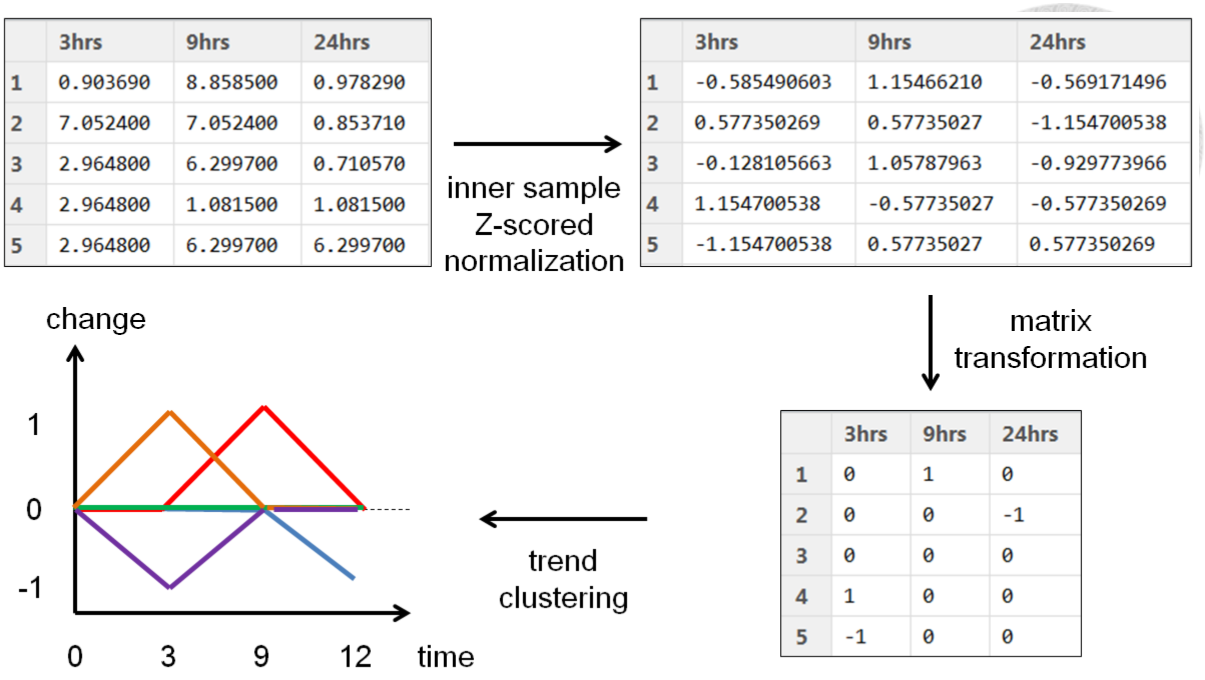


Figure 6 The analyzing flow of profile clustering module

The flow consists of two steps, generating clustering number and clustering the co-expression phosphorylation events. The clustering number is determined either by users or the detection algorithm in Dyanpho. Fuzzy c-means clustering takes the clustering number as a parameter and clusters phosphorylation events whose trends are similar.



	3hrs	9hrs	24hrs
1	0.903690	8.858500	0.978290
2	7.052400	7.052400	0.853710
3	2.964800	6.299700	0.710570
4	2.964800	1.081500	1.081500
5	2.964800	6.299700	6.299700

	3hrs	9hrs	24hrs
1	-0.585490603	1.15466210	-0.569171496
2	0.577350269	0.57735027	-1.154700538
3	-0.128105663	1.05787963	-0.929773966
4	1.154700538	-0.57735027	-0.577350269
5	-1.154700538	0.57735027	0.577350269

	3hrs	9hrs	24hrs
1	0	1	0
2	0	0	-1
3	0	0	0
4	1	0	0
5	-1	0	0

Figure 7 The example of auto detection method for determining clustering number

Parameter of the example is the same with defaults (inner z-scored S.D. is 1.1, variation threshold in specific time is 0.01 S.D. and the number threshold is 1%). There are five different trend profiles for five clusters labeled with different colors in the example.

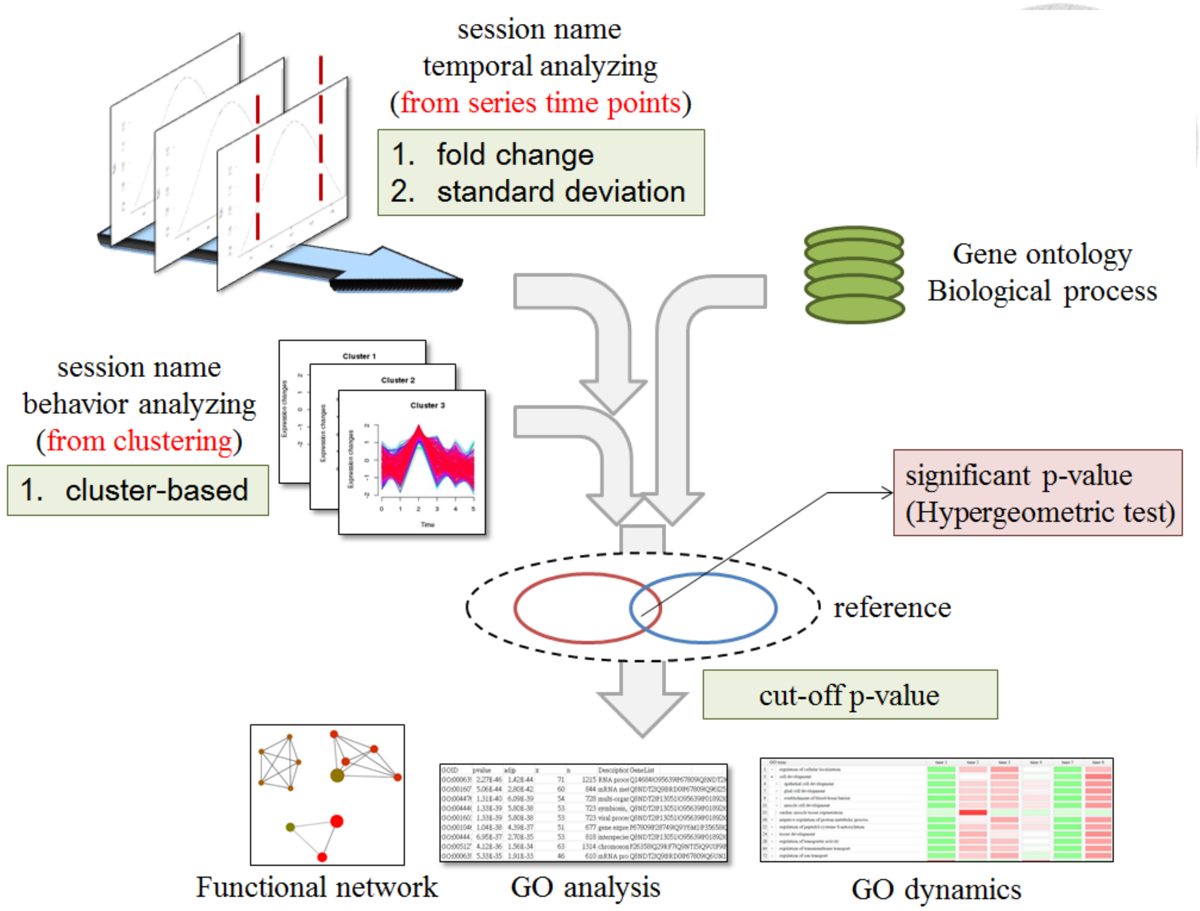


Figure 8 The analyzing flow of function enrichment module

The non-repeated protein session names are selected from temporal analyzing by fold change or standard deviation or from the cluster calculated by profile clustering module. Selected proteins and total uniprot proteins are analyzed by the hypergeometric test with the biological processes database of Gene Ontology. The function enrichment network and the dynamics of biological processes are further analyzed to present core and detailed functions.

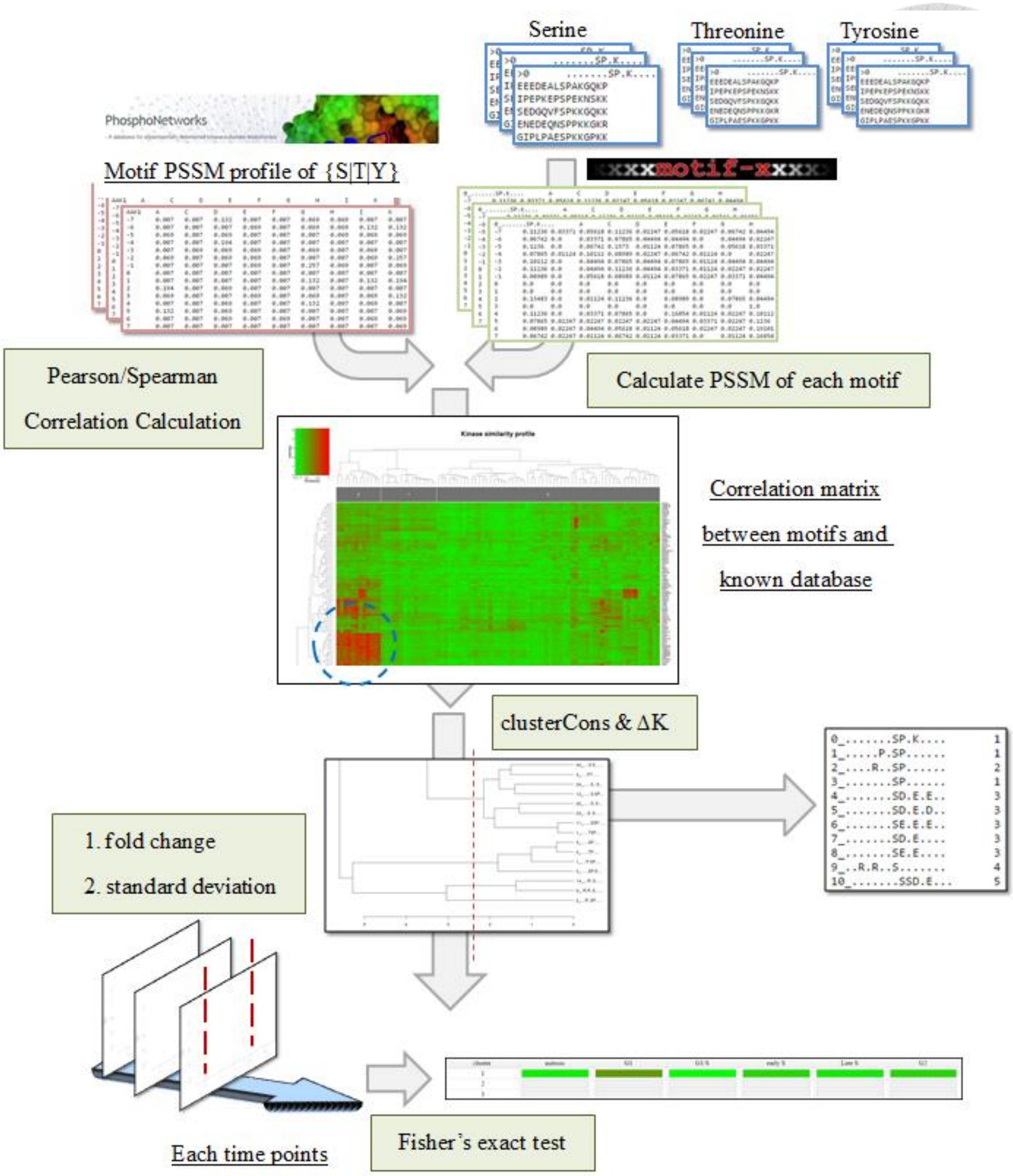
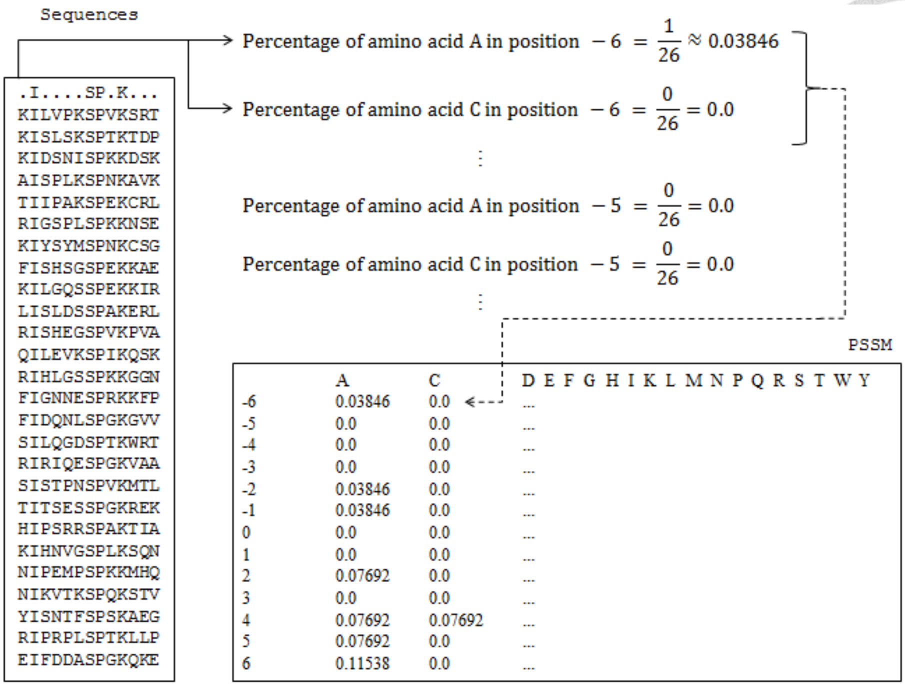


Figure 9 The analyzing flow of kinase activation profile module

All phosphorylation sequences are separated into 3 sets based on the center phosphosite. Each set is sent to motif-x separately and then DynaPho fetches the conserved motif information. DynaPho further generates a PSSM table for each conserved motif. The correlation between PSSMs and PhosphoNetworks databases shows potential kinases. Conserved motifs are reduced into smaller clusters by clusterCons algorithm. Temporal profiles of both kinase activation and deactivation are generated by fisher's exact test.



	A	C	D E F G H I K L M N P Q R S T W Y
-6	0.03846	0.0	...
-5	0.0	0.0	...
-4	0.0	0.0	...
-3	0.0	0.0	...
-2	0.03846	0.0	...
-1	0.03846	0.0	...
0	0.0	0.0	...
1	0.0	0.0	...
2	0.07692	0.0	...
3	0.0	0.0	...
4	0.07692	0.07692	...
5	0.07692	0.0	...
6	0.11538	0.0	...

Figure 10 The example of generating the PSSM table of each conserved motifs

The conserved motif, ‶.I....SP.K...″, is obtained from motif-x. The following 26 sequences are members contributing to the motif and each one is composed of 13 different amino acids. Start from the center phosphosite, the number -6 to -1 and 1 to 6 are relative sequence positions on both sides of it. The x-axis in PSSM consists of total amino acids and y-axis is the relative position. The number in PSSM is the proportion of the amino acid in current position.

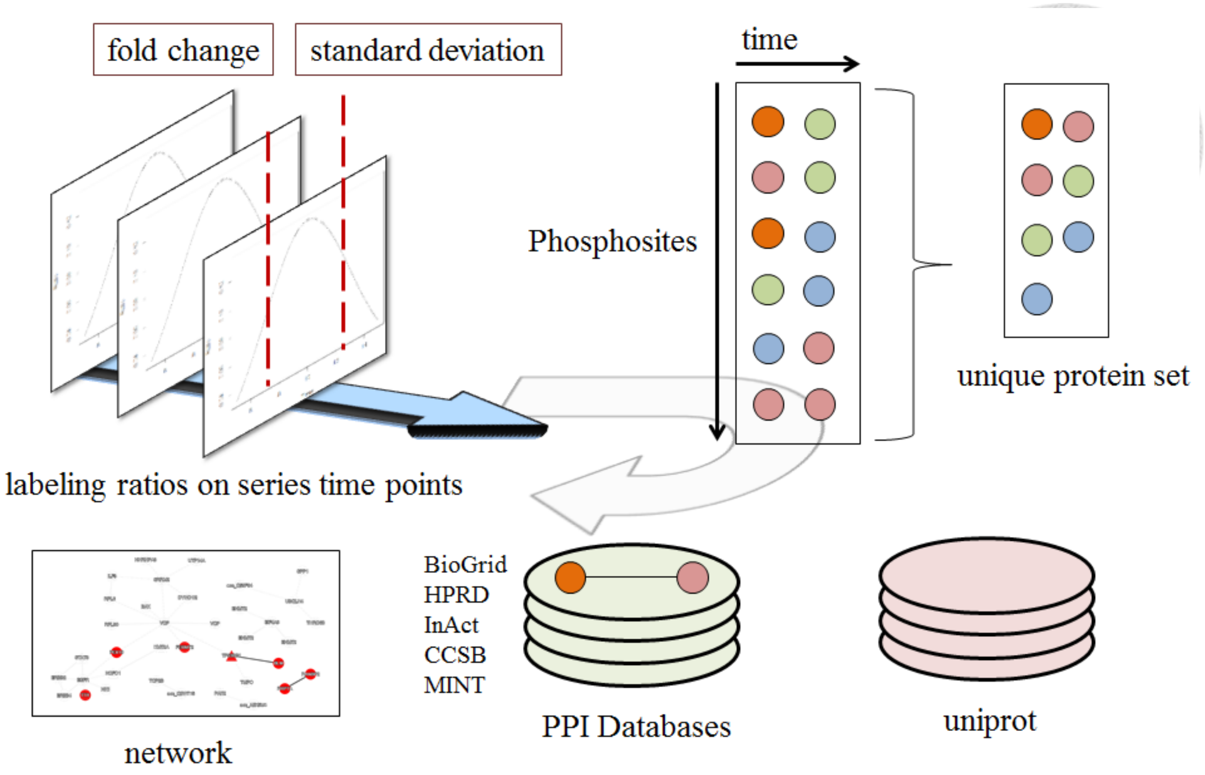


Figure 11 The analyzing flow of interaction network module

Phosphorylation events are filtered by the standard deviation or the fold change. Map all phosphorylation sequences back into proteins and prevent repeated ones. These proteins construct a interaction network. If two proteins are not interacted in the specific time, DynaPho links both them with intermediary proteins which are connected to each one but are not significant expression in the current time.

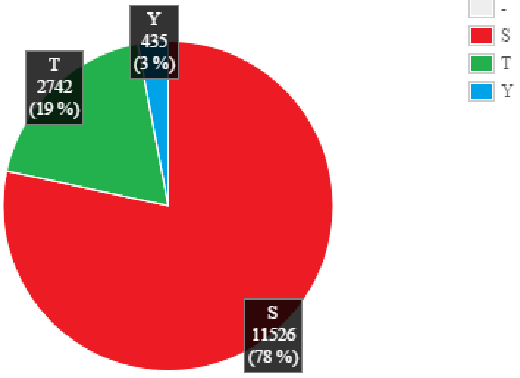


Figure 12 The number and proportion of each phosphorylation sites

The pie chart presents the proportion of three phosphosite, serine (S), threonine (T) and tyrosine (Y) with their numbers in the sequence pool. The dashed '-' presents the number of phosphosites which are not S, T or Y in the sequence.

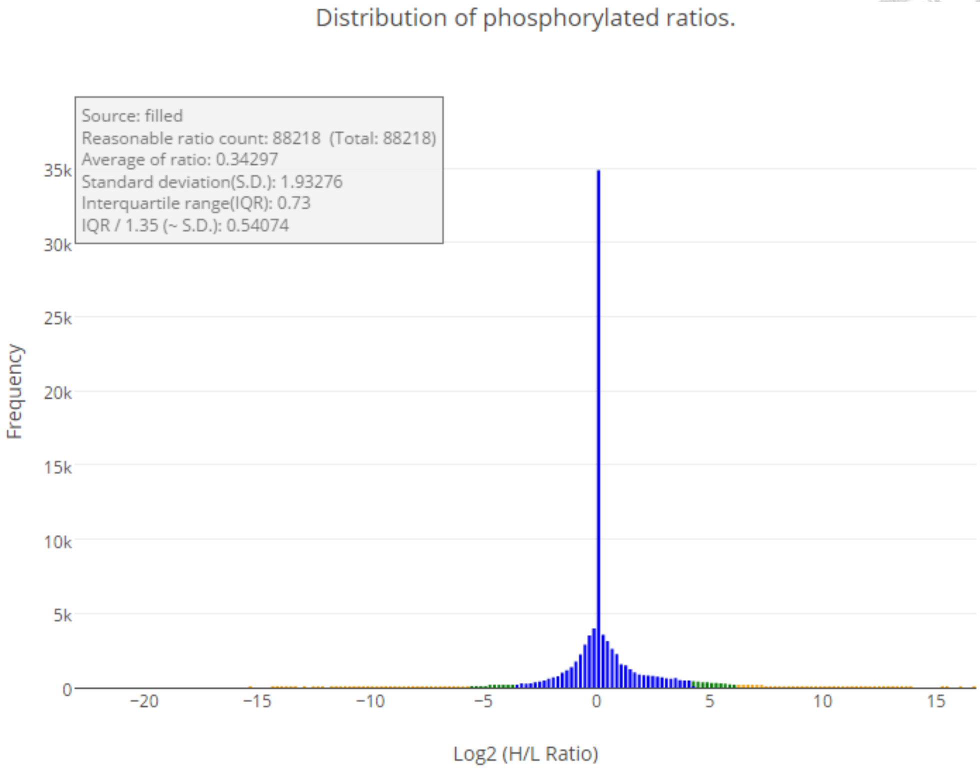
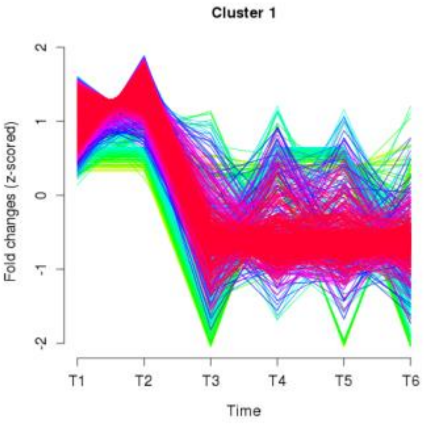


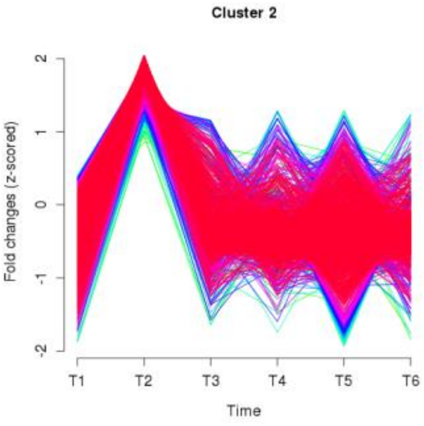
Figure 13 The distribution of all labeling ratios in log2 scaled

The space of each column in x-axis is 0.2. Blue, green and orange respectively stand for ratios in 2 S.D., more than 2 S.D. and less than 3 S.D., and over 3.S.D. The plot is generated by R script with Plotly.



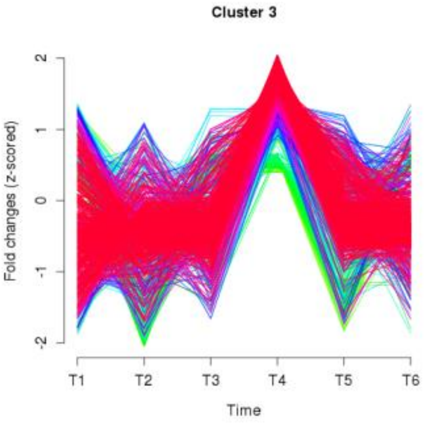
1,099 (7.475 %)

Biological process (GO)	-log10 (adj. P)
gene expression	40.22
mRNA metabolic process	40.22
cell cycle process	37.92
microtubule-based process	35.42
RNA processing	31.04
cellular macromolecular complex assembly	30.92
cytoskeleton organization	30.07
mitotic cell cycle process	27.78
regulation of organelle organization	25.14
protein complex assembly	22.57
multi-organism cellular process	21.23



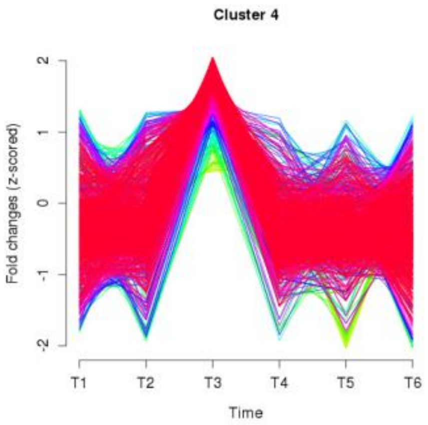
1,746 (11.875 %)

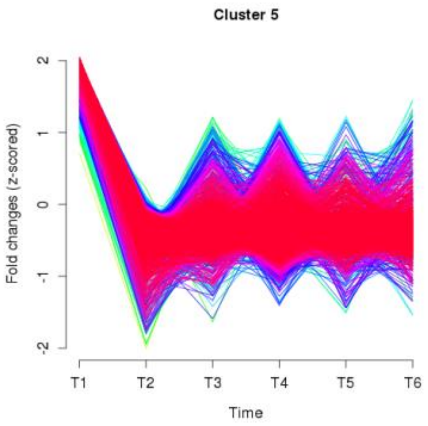
Biological process (GO)	-log10 (adj. P)
cell cycle process	82.48
mitotic cell cycle process	67.10
RNA processing	62.59
microtubule-based process	58.10
cytoskeleton organization	50.07
mRNA metabolic process	49.92
chromosome organization	49.03
cellular macromolecular complex assembly	46.63
gene expression	42.10
protein complex assembly	40.05
mitotic cell cycle	38.28

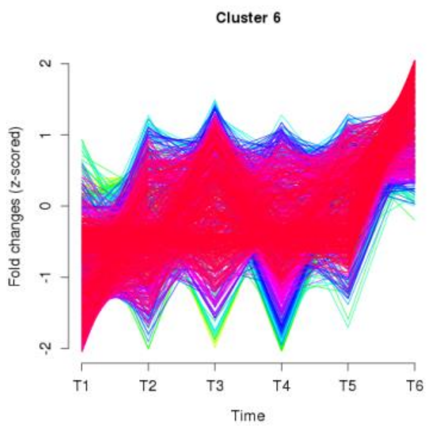


1,010 (6.869 %)

Biological process (GO)	-log10 (adj. P)
mitotic cell cycle process	63.37
cytoskeleton organization	41.92
chromosome organization	36.68
muscle cell cellular homeostasis	31.00
microtubule cytoskeleton organization	27.92
negative regulation of phosphorus metabolic process	27.80
cellular macromolecular complex assembly	27.19
regulation of cell cycle	25.40
regulation of peptidyl-cysteine S-nitrosylation	25.37

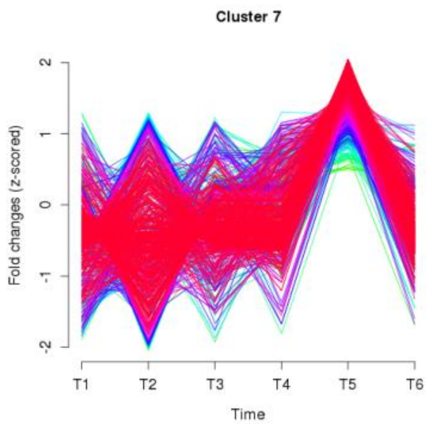
	Biological process (GO)	-log10 (adj. P)
 1,146 (7.794 %)	RNA processing	38.79
	cell cycle process	35.04
	chromosome organization	34.13
	mRNA metabolic process	30.47
	cellular response to DNA damage stimulus	30.17
	gene expression	27.20
	symbiosis, encompassing mutualism through parasitism	21.61
	multi-organism cellular process	21.55
	cytoskeleton organization	20.97
	vesicle-mediated transport	18.62
	chromatin organization	18.45

	Biological process (GO)	-log10 (adj. P)
 6,394 (43.488 %)	mRNA metabolic process	117.77
	chromosome organization	117.69
	RNA processing	116.73
	cell cycle process	112.33
	mitotic cell cycle process	110.88
	gene expression	98.87
	chromatin organization	96.81
	symbiosis, encompassing mutualism through parasitism	89.97
	regulation of organelle organization	75.14
	cytoskeleton organization	73.88
	cellular macromolecular complex assembly	73.58
	negative regulation of cellular macromolecule biosynthetic process	72.61
	negative regulation of RNA metabolic process	70.15
	cytoplasmic transport	59.24
	mRNA transport	53.99
	protein complex assembly	53.17
	cellular response to DNA damage stimulus	52.07
	transcription from RNA polymerase II promoter	43.76
	single-organism intracellular transport	43.31



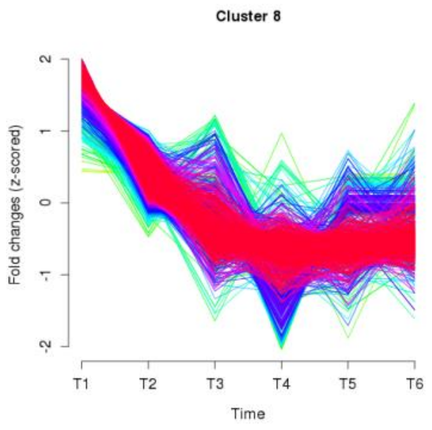
1,151 (7.828 %)

Biological process (GO)	-log10 (adj. P)
regulation of peptidyl-cysteine S-nitrosylation	72.10
olfactory nerve structural organization	69.49
establishment of glial blood-brain barrier	68.86
regulation of skeletal muscle contraction by regulation of release of sequestered calcium ion	68.70
positive regulation of sodium ion transmembrane transporter activity	66.44
regulation of voltage-gated calcium channel activity	63.30
neurotransmitter receptor metabolic process	62.75
cardiac muscle cell action potential	62.26
nucleus localization	60.99



771 (5.244 %)

Biological process (GO)	-log10 (adj. P)
cell cycle process	43.80
chromosome organization	33.98
cellular response to DNA damage stimulus	31.21
mRNA metabolic process	27.12
RNA splicing	25.77
gene expression	21.32
regulation of cell cycle	20.84
regulation of organelle organization	19.35



1,386 (9.427 %)

Biological process (GO)	-log10 (adj. P)
mitotic cell cycle process	74.06
cellular macromolecular complex assembly	71.74
RNA processing	71.13
muscle cell cellular homeostasis	68.61
regulation of peptidyl-cysteine S-nitrosylation	68.48
olfactory nerve structural organization	68.30
establishment of glial blood-brain barrier	67.59
regulation of skeletal muscle contraction by regulation of release of sequestered calcium ion	67.26

T1	T2	T3	T4	T5	T6
mitosis	G1	G1/S	early S	Late S	G2

Figure 14 Co-expression clustering of dynamic phosphorylation profiles

Fuzzy c-means is a soft clustering algorithm, trends of phosphorylation events which are colored in the light green stand for the outlier of the cluster. On the contrary, ones which are colored in darker red mean the core of the cluster. The number under the plot presents the number of phosphorylation events in the cluster. The “adj. P” is the abbreviation of adjusted p-value.

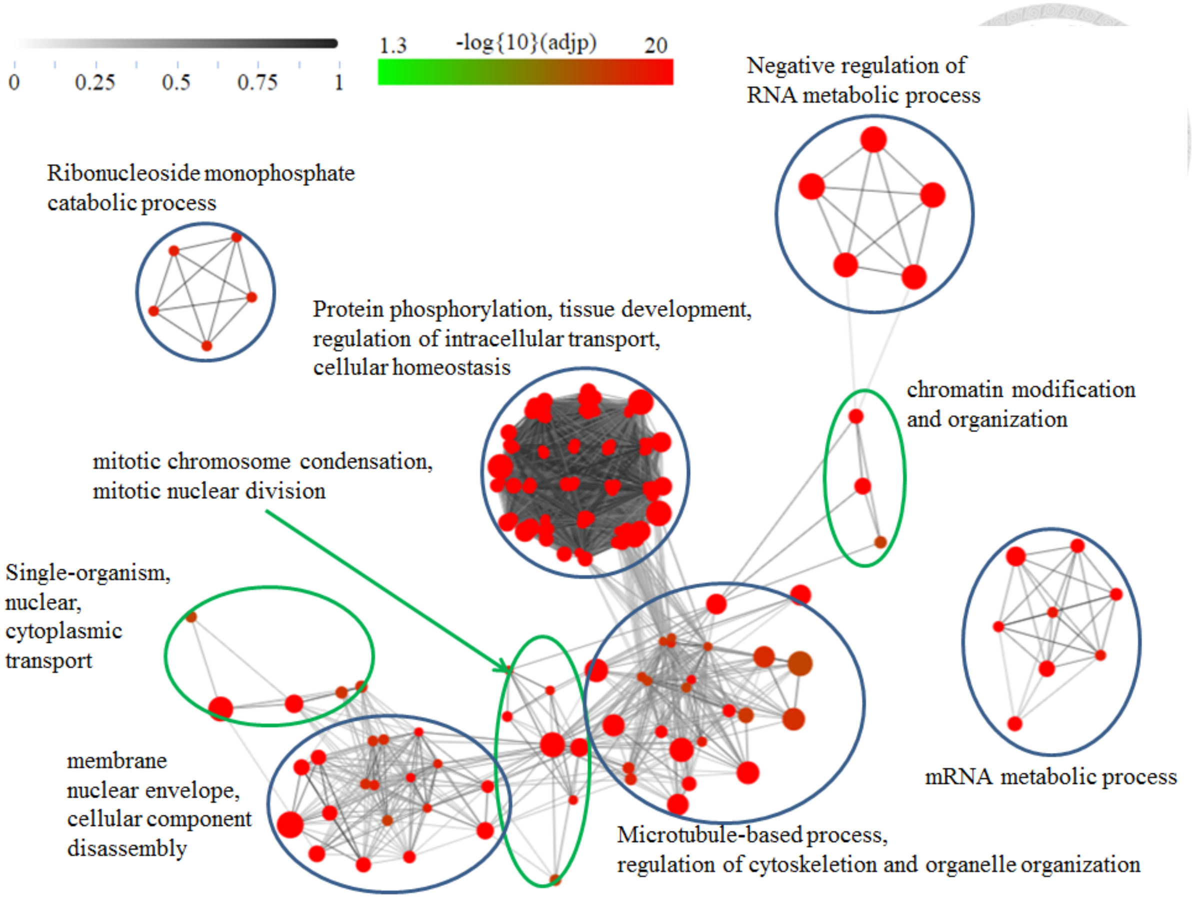
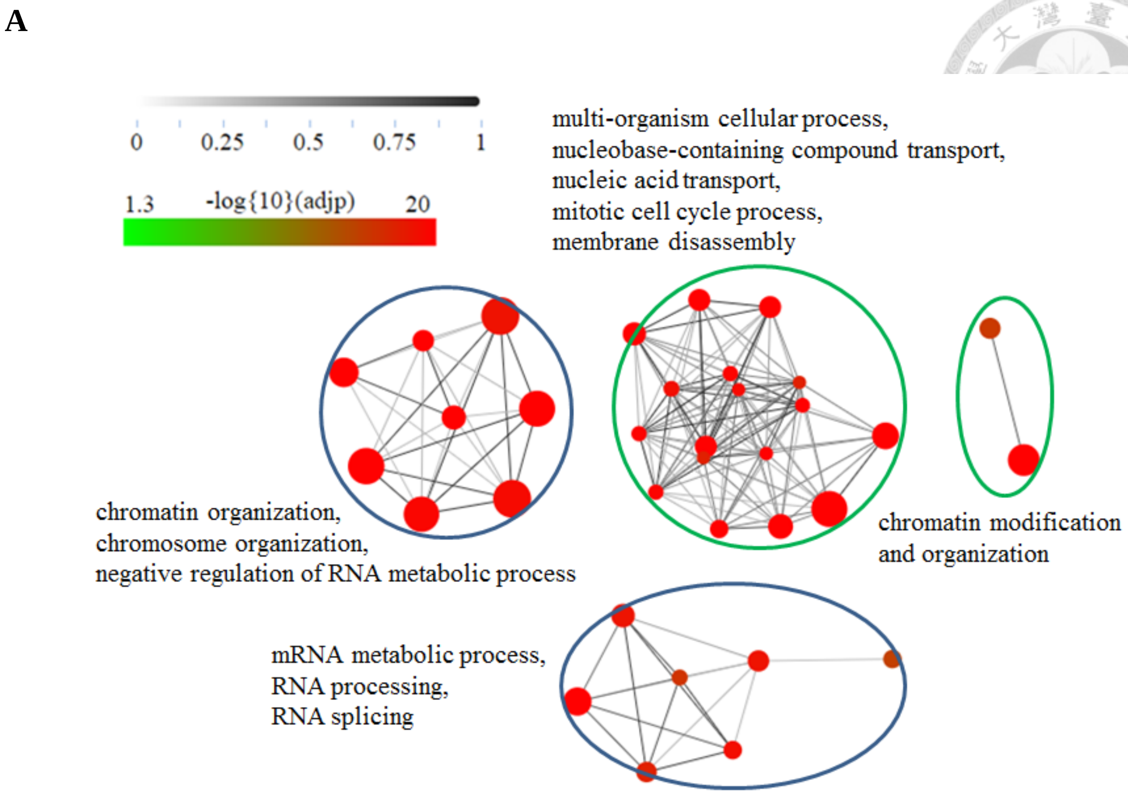

0
0.25
0.5
0.75
1
1.3
-log{10}(adjp)
20
Negative regulation of RNA metabolic process
Ribonucleoside monophosphate catabolic process
Protein phosphorylation, tissue development, regulation of intracellular transport, cellular homeostasis
chromatin modification and organization
mitotic chromosome condensation, mitotic nuclear division
Single-organism, nuclear, cytoplasmic transport
membrane nuclear envelope, cellular component disassembly
mRNA metabolic process
Microtubule-based process, regulation of cytoskeletion and organelle organization

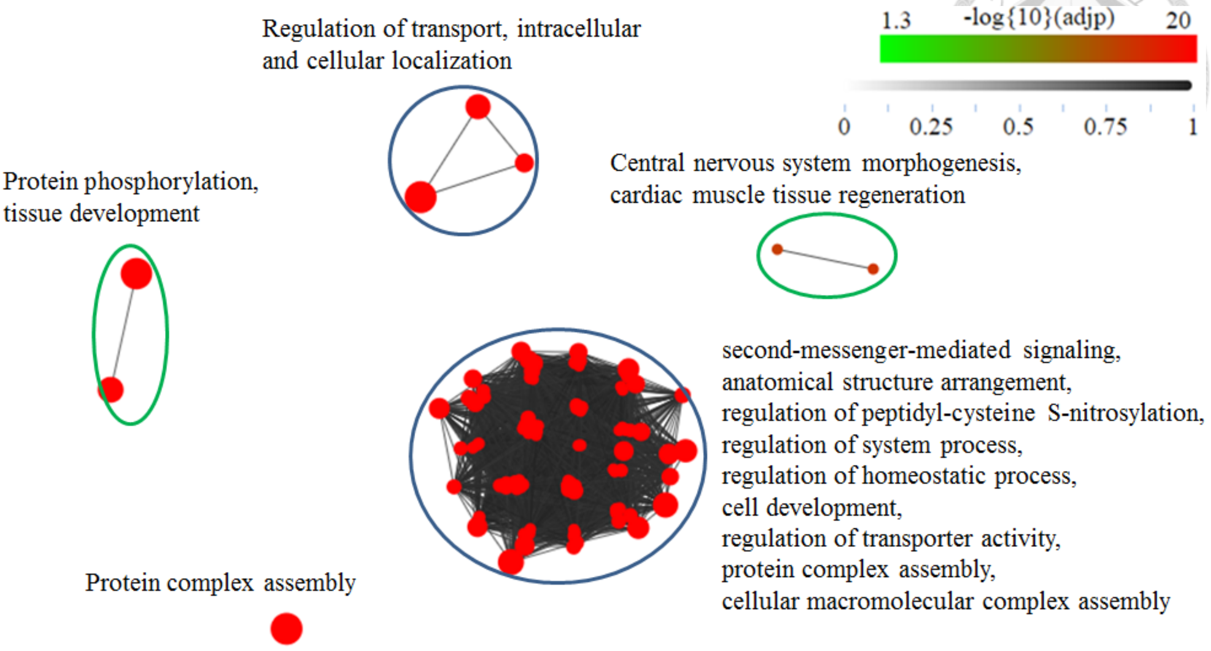
Figure 15 The summary of core functions over all cell cycle stages

The node and edge in the network respectively represents a GO term and the proportion of joint proteins. The size of each GO term is directly proportional to the background protein frequency. Current layout in the enrichment network analysis is implemented that parameter of similarity and style is respectively 0.3 and Cose.

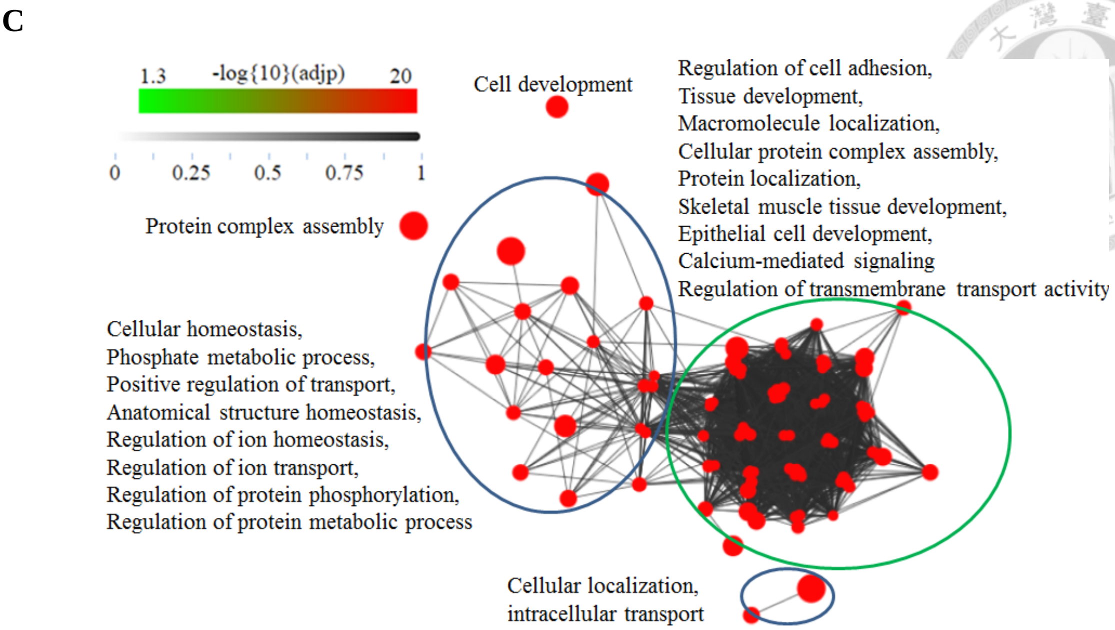


Biological process (GO)	-log10 (adj. P)
chromosome organization	37.35
mitotic cell cycle process	37.32
chromatin organization	26.32
cytoskeleton organization	24.80
negative regulation of RNA metabolic process	22.41
symbiosis, encompassing mutualism through parasitism	21.73
viral process	21.73
multi-organism cellular process	21.64
interspecies interaction between organisms	21.36
nuclear envelope organization	21.22
negative regulation of RNA biosynthetic process	21.01
mRNA transport	20.46
nucleic acid transport	19.79
mitotic nuclear envelope disassembly	19.44
negative regulation of cellular macromolecule biosynthetic process	19.34
RNA splicing	19.05
nucleobase-containing compound transport	19.02
negative regulation of gene expression	18.76
mRNA metabolic process	18.58
membrane disassembly	17.70

B

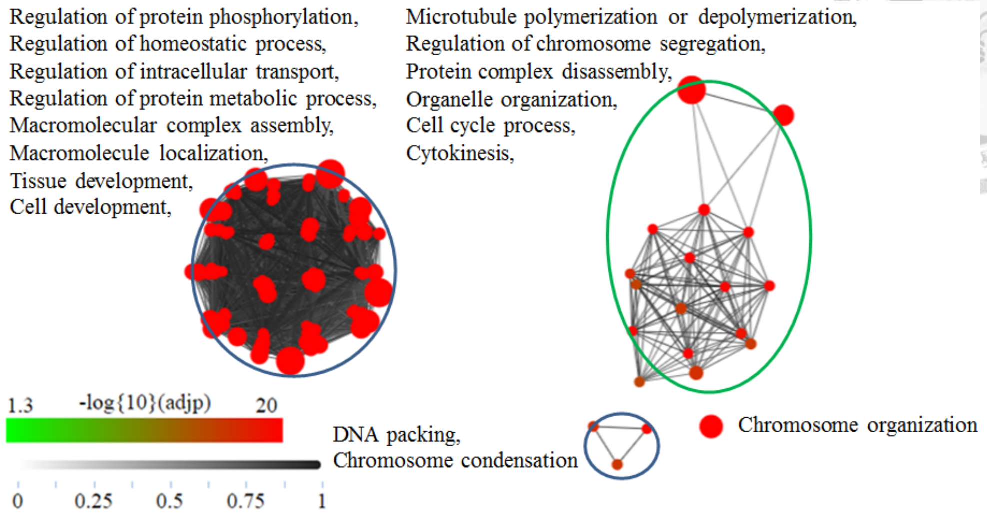

Regulation of transport, intracellular and cellular localization
1.3
-log{10}(adjp)
20
0
0.25
0.5
0.75
1
Central nervous system morphogenesis, cardiac muscle tissue regeneration
Protein phosphorylation, tissue development
second-messenger-mediated signaling, anatomical structure arrangement, regulation of peptidyl-cysteine S-nitrosylation, regulation of system process, regulation of homeostatic process, cell development, regulation of transporter activity, protein complex assembly, cellular macromolecular complex assembly
Protein complex assembly

Biological process (GO)	-log10 (adj. P)
regulation of peptidyl-cysteine S-nitrosylation	97.67
olfactory nerve structural organization	96.73
regulation of skeletal muscle contraction by regulation of release of sequestered calcium ion	96.73
establishment of glial blood-brain barrier	96.18
positive regulation of sodium ion transmembrane transporter activity	94.74
establishment of blood-nerve barrier	94.06
neurotransmitter receptor metabolic process	90.64
regulation of voltage-gated calcium channel activity	90.17
regulation of ryanodine-sensitive calcium-release channel activity	86.68
negative regulation of peptidyl-serine phosphorylation	85.52
nucleus localization	85.37
cardiac muscle cell action potential	85.11
myotube cell development	79.89
muscle cell cellular homeostasis	78.33
positive regulation of cell-matrix adhesion	74.82
muscle fiber development	71.91
skeletal muscle tissue development	70.07
regulation of intracellular transport	66.67
cellular protein complex assembly	50.32
cellular protein localization	45.40

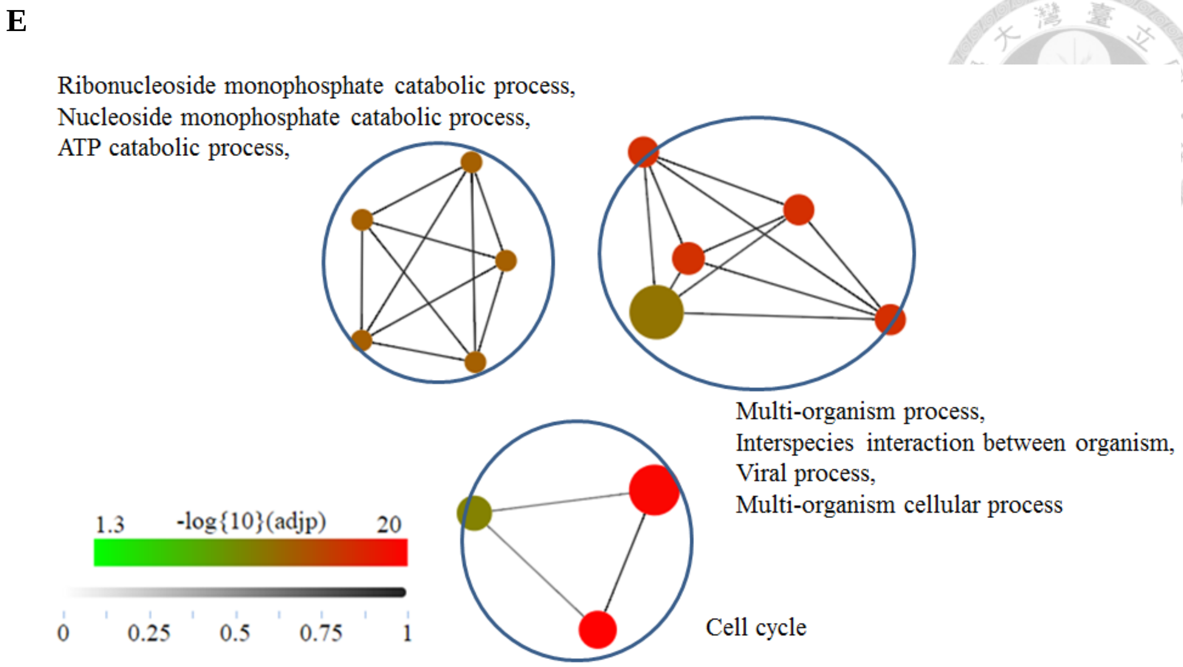

C
1.3
-log{10}(adjp)
20
0
0.25
0.5
0.75
1
Protein complex assembly
Cell development
Regulation of cell adhesion,
Tissue development,
Macromolecule localization,
Cellular protein complex assembly,
Protein localization,
Skeletal muscle tissue development,
Epithelial cell development,
Calcium-mediated signaling
Regulation of transmembrane transport activity
Cellular homeostasis,
Phosphate metabolic process,
Positive regulation of transport,
Anatomical structure homeostasis,
Regulation of ion homeostasis,
Regulation of ion transport,
Regulation of protein phosphorylation,
Regulation of protein metabolic process
Cellular localization,
intracellular transport

Biological process (GO)	-log10 (adj. P)
negative regulation of peptidyl-cysteine S-nitrosylation	102.36
olfactory nerve structural organization	101.33
regulation of skeletal muscle contraction by regulation of release of sequestered calcium ion	101.33
negative regulation of peptidyl-serine phosphorylation	100.88
establishment of glial blood-brain barrier	100.82
positive regulation of sodium ion transmembrane transporter activity	99.49
neurotransmitter receptor metabolic process	95.11
regulation of voltage-gated calcium channel activity	94.77
regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion	94.04
nucleus localization	89.93
regulation of ryanodine-sensitive calcium-release channel activity	88.83
myotube cell development	84.43
muscle cell cellular homeostasis	82.85
positive regulation of cell-matrix adhesion	79.33
muscle fiber development	76.42
receptor metabolic process	72.28
regulation of intracellular transport	69.80
positive regulation of cell-substrate adhesion	68.77
regulation of ion transmembrane transport	62.10
cellular protein complex assembly	55.30
cellular macromolecular complex assembly	48.73
cellular protein localization	48.47

D

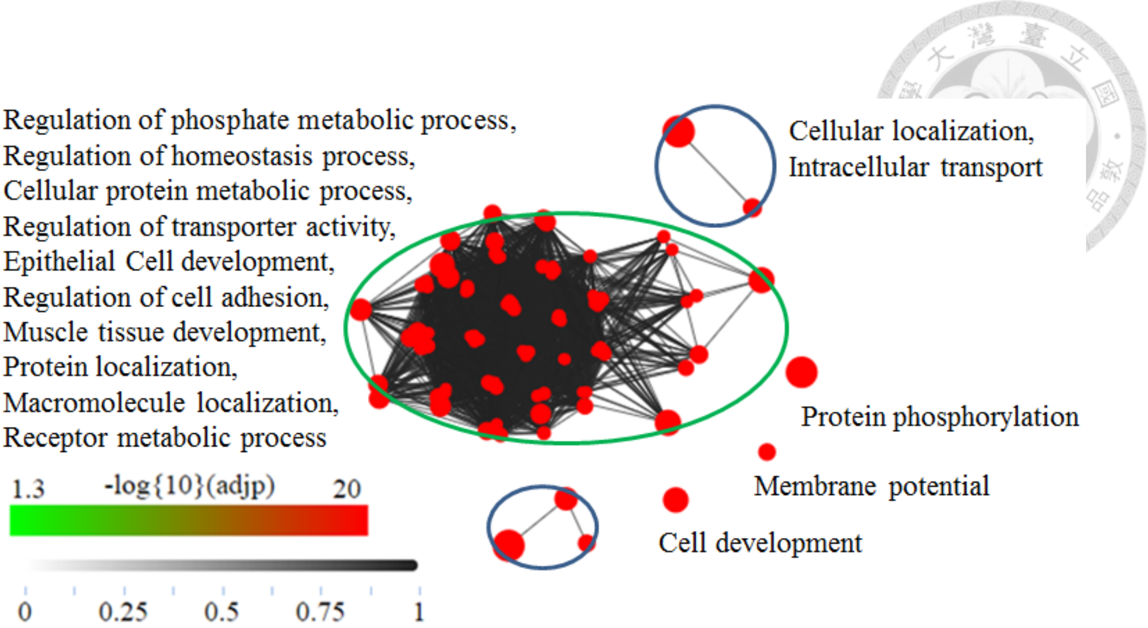

Regulation of protein phosphorylation,
Regulation of homeostatic process,
Regulation of intracellular transport,
Regulation of protein metabolic process,
Macromolecular complex assembly,
Macromolecule localization,
Tissue development,
Cell development,
Microtubule polymerization or depolymerization,
Regulation of chromosome segregation,
Protein complex disassembly,
Organelle organization,
Cell cycle process,
Cytokinesis,
1.3
-log{10}(adjp)
20
0
0.25
0.5
0.75
1
DNA packing,
Chromosome condensation
Chromosome organization

Biological process (GO)	-log10 (adj. P)
negative regulation of peptidyl-cysteine S-nitrosylation	87.48
olfactory nerve structural organization	86.66
regulation of skeletal muscle contraction by regulation of release of sequestered calcium ion	86.46
establishment of glial blood-brain barrier	86.15
positive regulation of sodium ion transmembrane transporter activity	84.45
neurotransmitter receptor metabolic process	80.45
regulation of voltage-gated calcium channel activity	79.73
regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion	79.00
nucleus localization	77.60
myotube cell development	69.78
positive regulation of cell-matrix adhesion	66.78
muscle fiber development	61.79
cellular protein complex assembly	47.64
cellular protein localization	41.61
regulation of intracellular transport	38.35
mitotic cytokinesis	28.54
chromosome organization	28.49
negative regulation of microtubule depolymerization	20.11
metaphase/anaphase transition of mitotic cell cycle	20.10
mitotic chromosome condensation	19.57
negative regulation of organelle organization	16.51
DNA packaging	16.08
regulation of chromosome segregation	15.11


E
Ribonucleoside monophosphate catabolic process,
Nucleoside monophosphate catabolic process,
ATP catabolic process,
Multi-organism process,
Interspecies interaction between organism,
Viral process,
Multi-organism cellular process
Cell cycle
1.3
-log{10}(adjp)
20
0
0.25
0.5
0.75
1

Biological process (GO)	-log10 (adj. P)
mitotic cell cycle process	25.54
cell cycle process	19.60
multi-organism cellular process	16.63
symbiosis, encompassing mutualism through parasitism	16.63
viral process	16.63
interspecies interaction between organisms	16.48
ATP catabolic process	13.09
purine nucleoside monophosphate catabolic process	13.03
purine ribonucleoside monophosphate catabolic process	13.03
ribonucleoside monophosphate catabolic process	13.03
nucleoside monophosphate catabolic process	12.98
DNA conformation change	10.23
single-organism intracellular transport	10.01

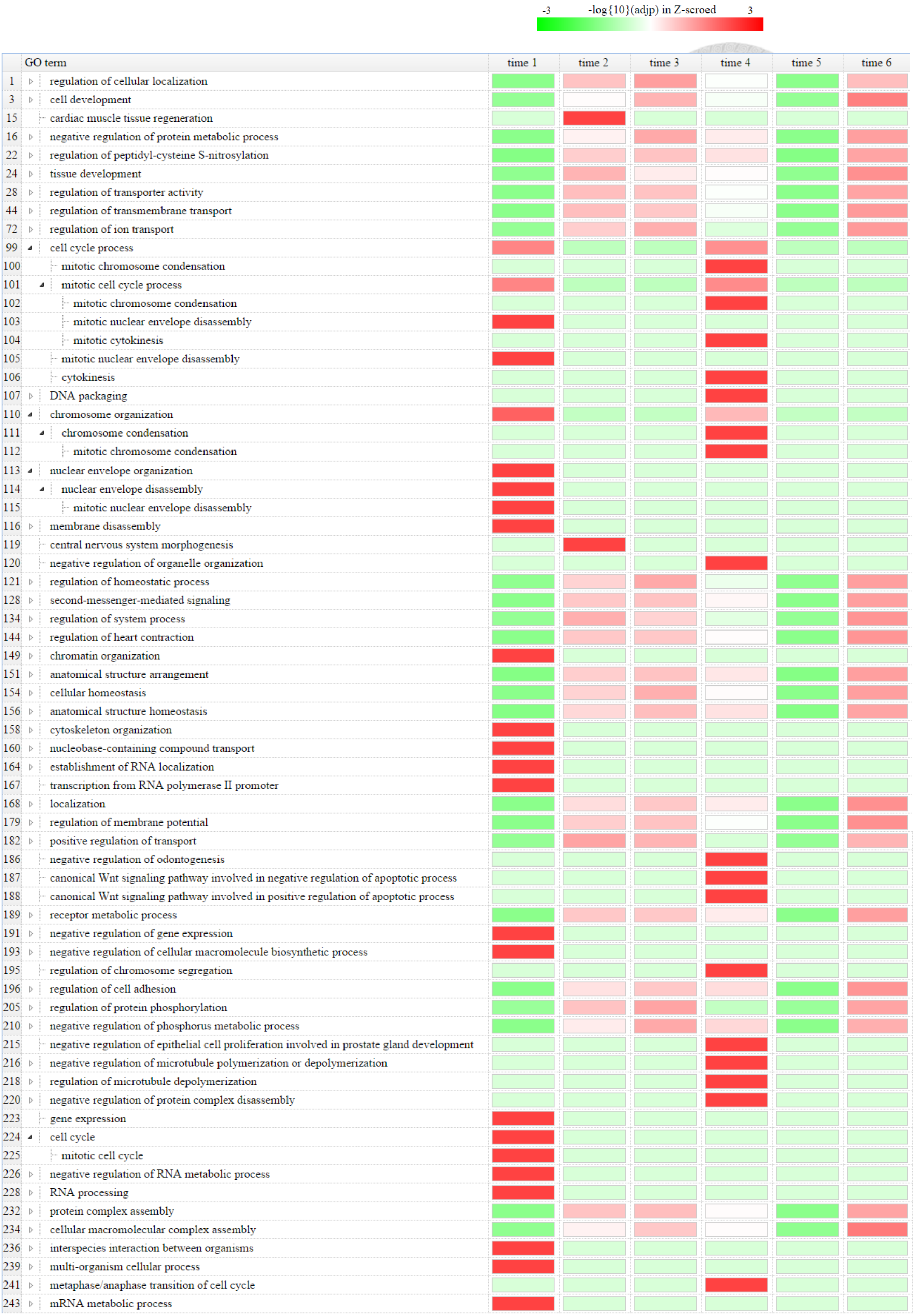
F


Regulation of phosphate metabolic process,
Regulation of homeostasis process,
Cellular protein metabolic process,
Regulation of transporter activity,
Epithelial Cell development,
Regulation of cell adhesion,
Muscle tissue development,
Protein localization,
Macromolecule localization,
Receptor metabolic process
Cellular localization,
Intracellular transport
Protein phosphorylation
Membrane potential
Cell development
1.3
-log{10}(adjp)
20
0
0.25
0.5
0.75
1

Biological process (GO)	-log10 (adj. P)
negative regulation of peptidyl-cysteine S-nitrosylation	118.64
olfactory nerve structural organization	117.50
regulation of skeletal muscle contraction by regulation of release of sequestered calcium ion	117.50
positive regulation of sodium ion transmembrane transporter activity	115.35
establishment of blood-nerve barrier	114.39
neurotransmitter receptor metabolic process	110.53
regulation of voltage-gated calcium channel activity	110.22
regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion	109.43
negative regulation of peptidyl-serine phosphorylation	105.29
nucleus localization	104.91
cardiac muscle cell action potential	102.21
myotube cell development	99.01
muscle cell cellular homeostasis	97.33
positive regulation of cell-matrix adhesion	93.62
skeletal muscle tissue development	91.14
muscle fiber development	90.52
cellular protein complex assembly	72.85
regulation of intracellular transport	64.73
cellular protein localization	61.08

Figure 16 Function enrichment analyses on the temporal profile

A, B, C, D, E, F respectively represents the cell cycle stage on mitosis, G1, G1/S, early S, late S and G2. Each function enrichment analysis is presented by both a summary network and a table. The table lists biological processes whose p-values in log transformation are the top. Execution parameters in function enrichment module are the same with default settings without late S stage (E). The enrichment analysis of late S stage is processed by parameters that S.D. and p-value respectively are 2.0 and 1e-10. The network layout on A, B, C, D, E, F is cose style and similarities (for edge) of A, B, C, D, E, F are respectively 0.3, 0.9, 0.96, 0.65, 0.3, and 0.97.



-3 -log{10}(adjp) in Z-scroed 3

	GO term	time 1	time 2	time 3	time 4	time 5	time 6
1	regulation of cellular localization	🟩	🟥	🟥	⬜	🟩	🟥
3	cell development	🟩	⬜	🟥	⬜	🟩	🟥
15	cardiac muscle tissue regeneration	🟩	🟥	🟩	🟩	🟩	🟩
16	negative regulation of protein metabolic process	🟩	🟥	🟥	🟥	🟩	🟥
22	regulation of peptidyl-cysteine S-nitrosylation	🟩	🟥	🟥	🟥	🟩	🟥
24	tissue development	🟩	🟥	🟥	⬜	🟩	🟥
28	regulation of transporter activity	🟩	🟥	🟥	⬜	🟩	🟥
44	regulation of transmembrane transport	🟩	🟥	🟥	⬜	🟩	🟥
72	regulation of ion transport	🟩	🟥	🟥	🟩	🟩	🟥
99	cell cycle process	🟥	🟩	🟩	🟥	🟩	🟩
100	mitotic chromosome condensation	🟩	🟩	🟩	🟥	🟩	🟩
101	mitotic cell cycle process	🟥	🟩	🟩	🟥	🟩	🟩
102	mitotic chromosome condensation	🟩	🟩	🟩	🟥	🟩	🟩
103	mitotic nuclear envelope disassembly	🟥	🟩	🟩	🟩	🟩	🟩
104	mitotic cytokinesis	🟩	🟩	🟩	🟥	🟩	🟩
105	mitotic nuclear envelope disassembly	🟥	🟩	🟩	🟩	🟩	🟩
106	cytokinesis	🟩	🟩	🟩	🟥	🟩	🟩
107	DNA packaging	🟩	🟩	🟩	🟥	🟩	🟩
110	chromosome organization	🟥	🟩	🟩	🟥	🟩	🟩
111	chromosome condensation	🟩	🟩	🟩	🟥	🟩	🟩
112	mitotic chromosome condensation	🟩	🟩	🟩	🟥	🟩	🟩
113	nuclear envelope organization	🟥	🟩	🟩	🟩	🟩	🟩
114	nuclear envelope disassembly	🟥	🟩	🟩	🟩	🟩	🟩
115	mitotic nuclear envelope disassembly	🟥	🟩	🟩	🟩	🟩	🟩
116	membrane disassembly	🟥	🟩	🟩	🟩	🟩	🟩
119	central nervous system morphogenesis	🟩	🟥	🟩	🟩	🟩	🟩
120	negative regulation of organelle organization	🟩	🟩	🟩	🟥	🟩	🟩
121	regulation of homeostatic process	🟩	🟥	🟥	🟩	🟩	🟥
128	second-messenger-mediated signaling	🟩	🟥	🟥	⬜	🟩	🟥
134	regulation of system process	🟩	🟥	🟥	🟩	🟩	🟥
144	regulation of heart contraction	🟩	🟥	🟥	⬜	🟩	🟥
149	chromatin organization	🟥	🟩	🟩	🟩	🟩	🟩
151	anatomical structure arrangement	🟩	🟥	🟥	🟥	🟩	🟥
154	cellular homeostasis	🟩	🟥	🟥	⬜	🟩	🟥
156	anatomical structure homeostasis	🟩	🟥	🟥	🟥	🟩	🟥
158	cytoskeleton organization	🟥	🟩	🟩	🟩	🟩	🟩
160	nucleobase-containing compound transport	🟥	🟩	🟩	🟩	🟩	🟩
164	establishment of RNA localization	🟥	🟩	🟩	🟩	🟩	🟩
167	transcription from RNA polymerase II promoter	🟥	🟩	🟩	🟩	🟩	🟩
168	localization	🟩	🟥	🟥	🟥	🟩	🟥
179	regulation of membrane potential	🟩	🟥	🟥	⬜	🟩	🟥
182	positive regulation of transport	🟩	🟥	🟥	🟩	🟩	🟥
186	negative regulation of odontogenesis	🟩	🟩	🟩	🟥	🟩	🟩
187	canonical Wnt signaling pathway involved in negative regulation of apoptotic process	🟩	🟩	🟩	🟥	🟩	🟩
188	canonical Wnt signaling pathway involved in positive regulation of apoptotic process	🟩	🟩	🟩	🟥	🟩	🟩
189	receptor metabolic process	🟩	🟥	🟥	🟥	🟩	🟥
191	negative regulation of gene expression	🟥	🟩	🟩	🟩	🟩	🟩
193	negative regulation of cellular macromolecule biosynthetic process	🟥	🟩	🟩	🟩	🟩	🟩
195	regulation of chromosome segregation	🟩	🟩	🟩	🟥	🟩	🟩
196	regulation of cell adhesion	🟩	🟥	🟥	🟥	🟩	🟥
205	regulation of protein phosphorylation	🟩	🟥	🟥	🟩	🟩	🟥
210	negative regulation of phosphorus metabolic process	🟩	🟥	🟥	🟥	🟩	🟥
215	negative regulation of epithelial cell proliferation involved in prostate gland development	🟩	🟩	🟩	🟥	🟩	🟩
216	negative regulation of microtubule polymerization or depolymerization	🟩	🟩	🟩	🟥	🟩	🟩
218	regulation of microtubule depolymerization	🟩	🟩	🟩	🟥	🟩	🟩
220	negative regulation of protein complex disassembly	🟩	🟩	🟩	🟥	🟩	🟩
223	gene expression	🟥	🟩	🟩	🟩	🟩	🟩
224	cell cycle	🟥	🟩	🟩	🟩	🟩	🟩
225	mitotic cell cycle	🟥	🟩	🟩	🟩	🟩	🟩
226	negative regulation of RNA metabolic process	🟥	🟩	🟩	🟩	🟩	🟩
228	RNA processing	🟥	🟩	🟩	🟩	🟩	🟩
232	protein complex assembly	🟩	🟥	🟥	⬜	🟩	🟥
234	cellular macromolecular complex assembly	🟩	🟥	🟥	⬜	🟩	🟥
236	interspecies interaction between organisms	🟥	🟩	🟩	🟩	🟩	🟩
239	multi-organism cellular process	🟥	🟩	🟩	🟩	🟩	🟩
241	metaphase/anaphase transition of cell cycle	🟩	🟩	🟩	🟥	🟩	🟩
243	mRNA metabolic process	🟥	🟩	🟩	🟩	🟩	🟩

Figure 17 The temporal profile of biological processes

Time points 1, 2, 3, 4, 5, 6 indicate mitosis, G1, G1/S, early S, late S, and G2 stage, respectively. The adjusted p-values in the same biological process are transformed by z-scored normalization. The collapsed biological processes mean that all of their inherited ones show the same dynamics. Execution parameters in function enrichment module are the same with default settings.

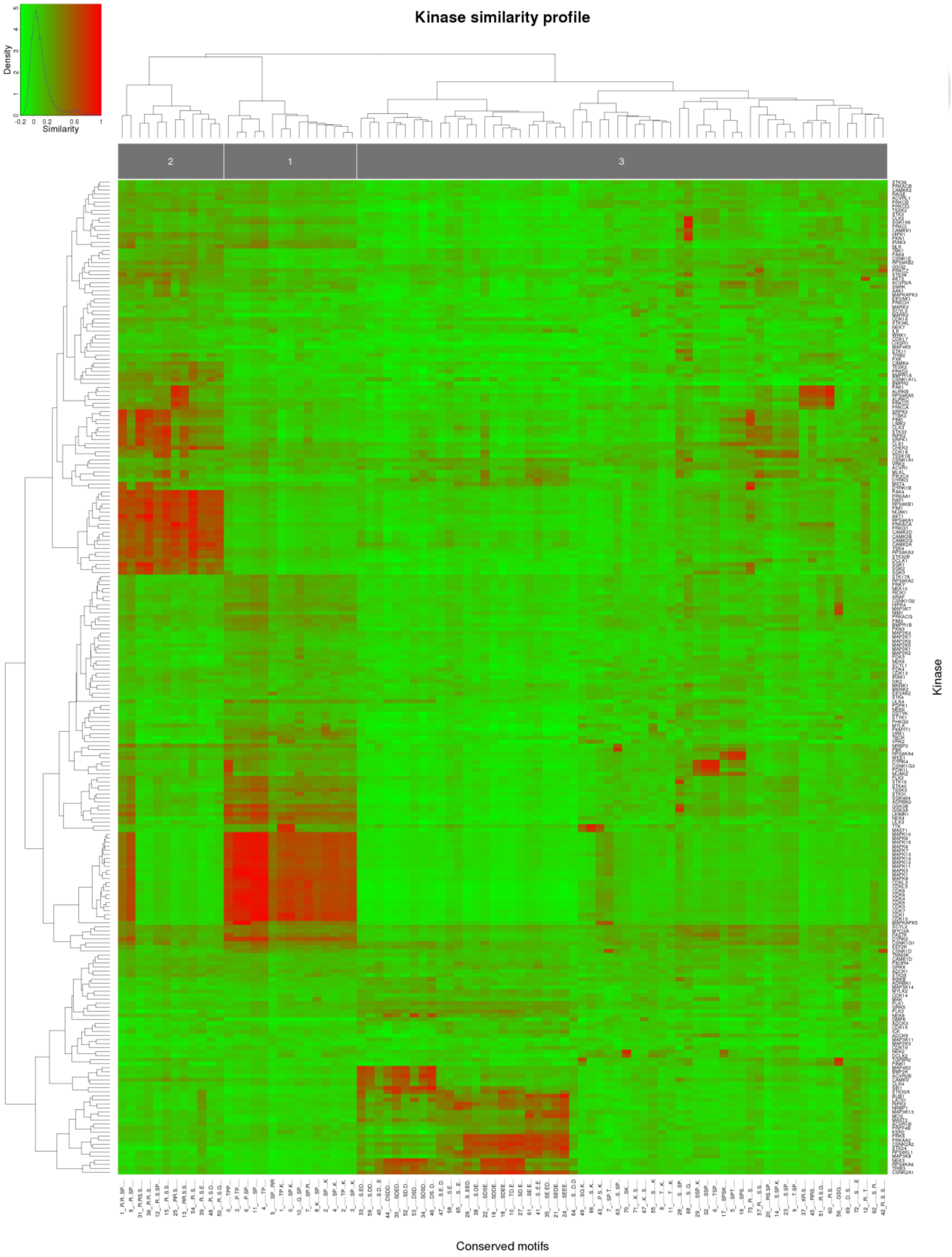
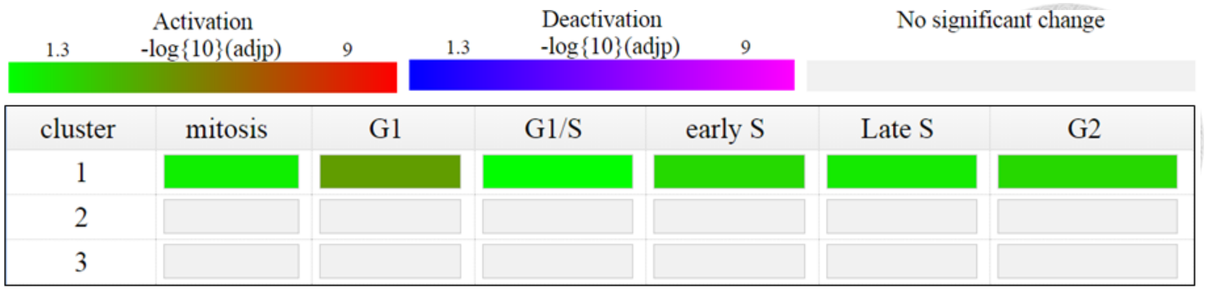


Figure 18 Conserved motifs imply potential kinases

The number on the top of heatmap is the cluster determined by clusterCons. Execution parameters in the analysis of kinase activation profile are the same with default settings.


Activation
-log{10}(adjp)
1.3
9
Deactivation
-log{10}(adjp)
1.3
9
No significant change

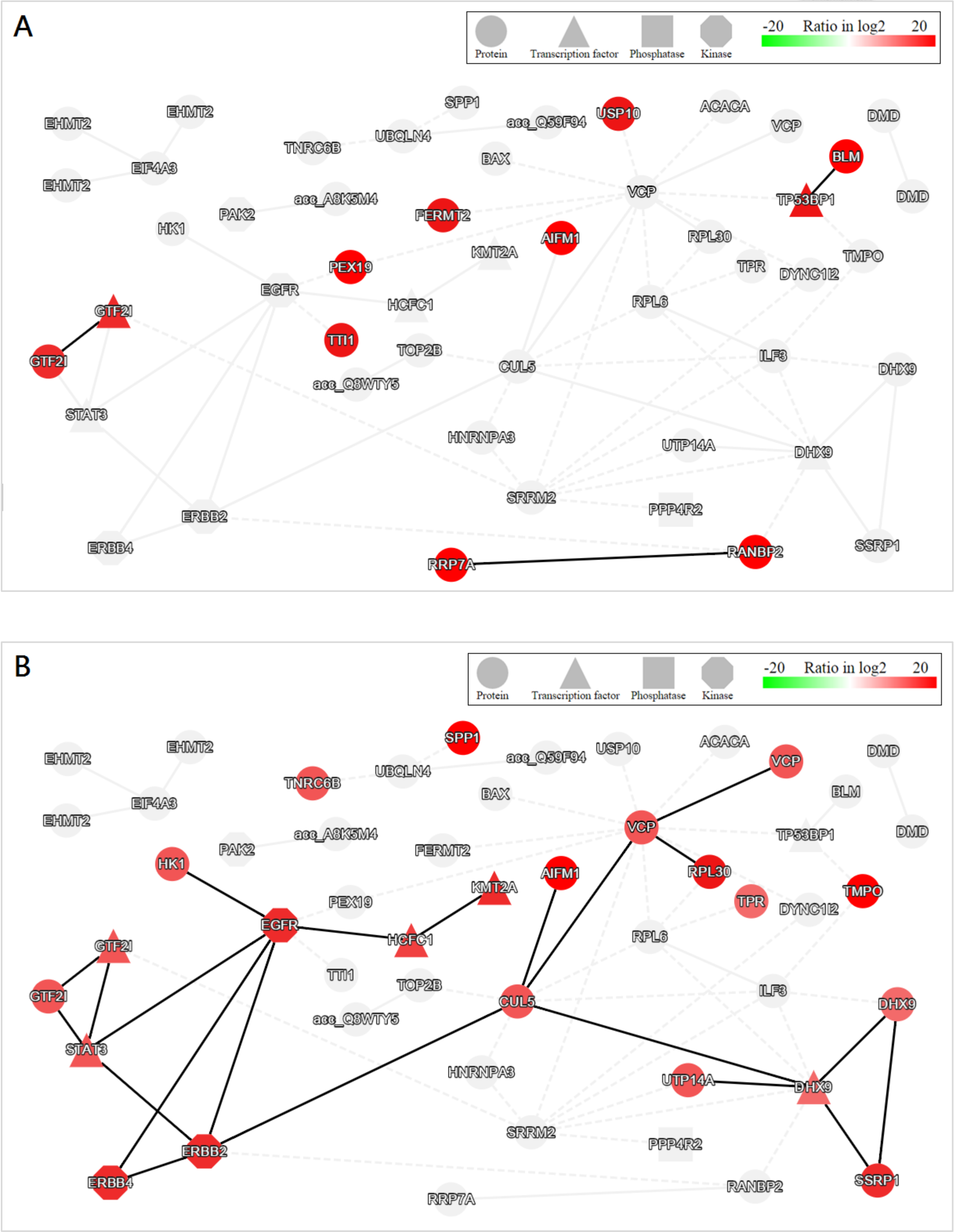
cluster	mitosis	G1	G1/S	early S	Late S	G2
1	🟩	🟩	🟩	🟩	🟩	🟩
2						
3						

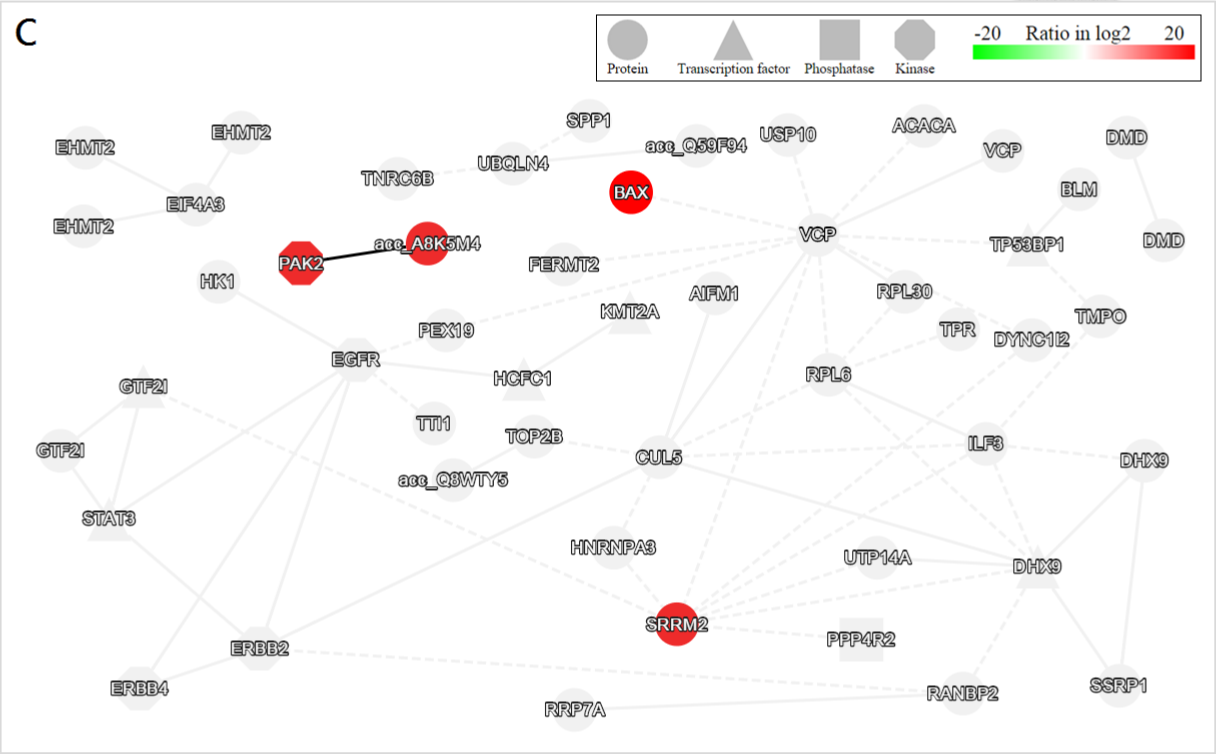
Cluster	1	2	3
Kinase	CDK1, CDK10, CDK3, CDK4, CDK5, CDK6, CDK7, CDK9, CDKL3, CDKL5, MAPK10, MAPK15, MAPK3, MAPK6, MAPK8, MAPK9	AKT1, CAMK2A, CAMK2B, NUAK1, PAK4, PIM1, PRKAA1, PRKACA, PRKG1, RAF1, RPS6KA1, RPS6KA3, RPS6KB1	-
Conserved motifs	0_......SP.K... 2_......SP....K 3_......SP...K. 4_......SP..K.. 5_......SP...RR 6_....P.SP..... 7_......SP.R... 8_K.....SP..... 10_...G..SP..... 11_......SP..... 0_......TPP.... 1_......TP.K... 2_......TP...K. 3_....P.TP..... 4_......TP.....	1_.R.R..SP..... 9_...R..SP..... 12_...R..S.SP... 13_...RR.S.S.... 15_...R..S.S.... 25_...RR.S...... 31_.R.RS.S...... 36_.R.R..S...... 39_...R..S.E.... 48_...R..S.D.... 50_...R..S.G.... 54_...R..S......	14_......S.SP.K. 16_......SDDE... 17_....SPSK..... 18_......SDEE... 19_....SPS...... 20_.....RS.SP... 21_......SEDE... 22_......SDSE... 23_......S.SP... 24_......SEEE... 26_......S..EED. 27_......SD.E... 28_......S...SP. 29_......SSP..K. 30_......SDED... 32_......SSP.... 33_......S.ED... 34_......SDSD... 35_......SE.ED.. 37_...KR.S...... 38_......S.DE... 40_......S.D...E 41_......S..E.E. 42_R..S..S...... 43_....P.S..K...

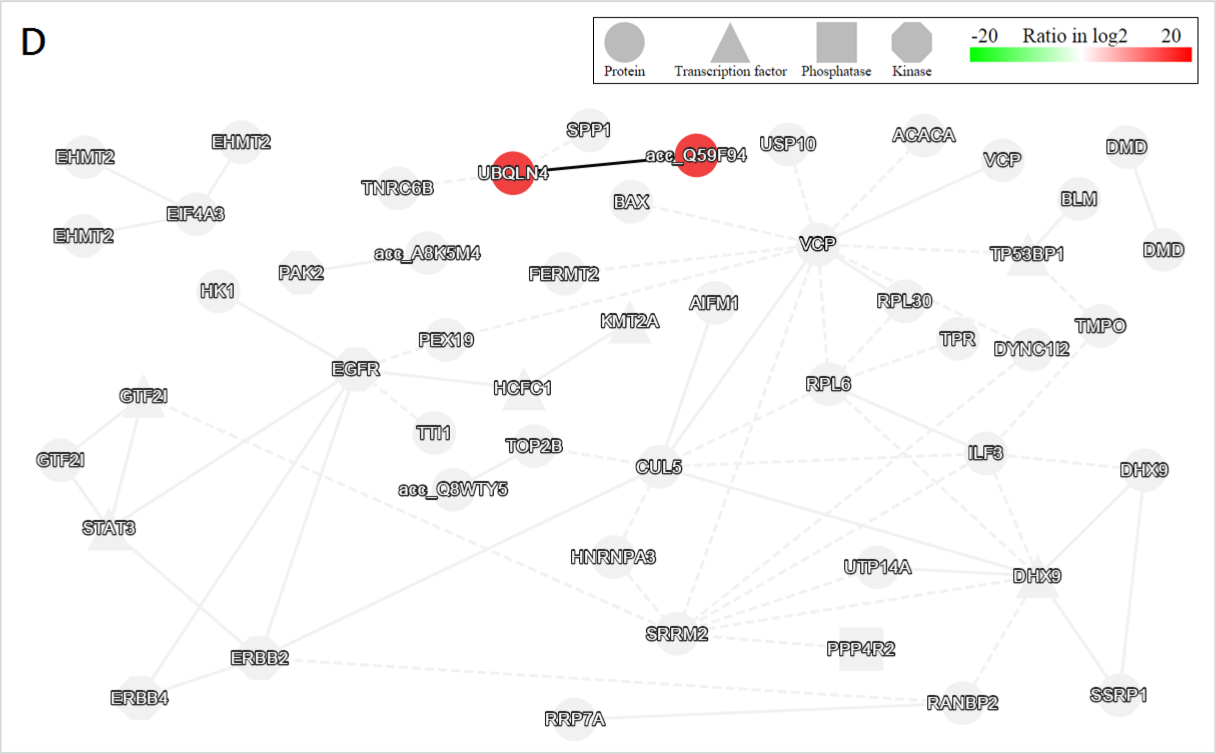
```
44_.....DSDD....
45_....RRS......
46_.....DS..D...
47_......S.E..D.
49_......SG.K...
51_....R.S.G....
52_......SD.D...
53_.....DSD.....
55_......S.....K
56_.....GSG.....
57_R.....S.S....
58_......S.E....
59_......S.DD...
60_....R.S......
61_......SE.E...
62_......S..R...
63_......S..SP..
64_......S...D.D
65_......S...E..
66_......S..K...
67_..K...S......
68_......S..G...
69_...D..S......
70_......SK.....
71_...K..S......
72_......S.....E
73_.R....S......
 5_....SPT......
 6_......TSP....
 7_...SP.T......
 8_......T...K..
 9_......T.SP...
10_......TD.E...
11_......T....K.
12_..R...T......
```

Figure 19 Dynamics to both kinase activation and deactivation time profiles

The number on each conserved motifs is the fetching order from motif-x.


A
Protein
Transcription factor
Phosphatase
Kinase
-20 Ratio in log2 20
EHMT2
EHMT2
SPP1
acc_Q59F94
USP10
ACACA
VCP
DMD
TNRC6B
UBQLN4
BAX
BLM
EIF4A3
EHMT2
acc_A8K5M4
VCP
TP53BP1
DMD
PAK2
FERMT2
HK1
AIFM1
RPL30
KMT2A
TMPO
PEX19
TPR
DYNC1I2
EGFR
HCFC1
RPL6
GTF2I
TTI1
TOP2B
ILF3
GTF2I
CUL5
DHX9
acc_Q8WTY5
STAT3
HNRNPA3
UTP14A
DHX9
SRRM2
PPP4R2
ERBB2
ERBB4
RANBP2
SSRP1
RRP7A
B
Protein
Transcription factor
Phosphatase
Kinase
-20 Ratio in log2 20
EHMT2
EHMT2
SPP1
acc_Q59F94
USP10
ACACA
VCP
DMD
TNRC6B
UBQLN4
BAX
BLM
EIF4A3
EHMT2
acc_A8K5M4
VCP
TP53BP1
DMD
PAK2
FERMT2
HK1
AIFM1
RPL30
KMT2A
TMPO
PEX19
TPR
DYNC1I2
EGFR
HCFC1
RPL6
GTF2I
TTI1
TOP2B
ILF3
GTF2I
CUL5
DHX9
acc_Q8WTY5
STAT3
HNRNPA3
UTP14A
DHX9
SRRM2
PPP4R2
ERBB2
ERBB4
RANBP2
SSRP1
RRP7A


C
Protein
Transcription factor
Phosphatase
Kinase
-20
Ratio in log2
20
EHMT2
EHMT2
EHMT2
EIF4A3
HK1
PAK2
acc_A8K5M4
TNRC6B
UBQLN4
SPP1
acc_Q59F94
BAX
USP10
ACACA
VCP
VCP
DMD
DMD
BLM
TP53BP1
FERMT2
AIFM1
KMT2A
RPL30
TMPO
TPR
DYNC1I2
PEX19
EGFR
HCFC1
GTF2I
GTF2I
TTI1
TOP2B
acc_Q8WTY5
CUL5
RPL6
ILF3
DHX9
DHX9
STAT3
HNRNPA3
UTP14A
SRRM2
PPP4R2
ERBB2
ERBB4
RRP7A
RANBP2
SSRP1


D
Protein
Transcription factor
Phosphatase
Kinase
-20
Ratio in log2
20
EHMT2
EHMT2
EHMT2
EIF4A3
HK1
PAK2
acc_A8K5M4
TNRC6B
UBQLN4
SPP1
acc_Q59F94
BAX
USP10
ACACA
VCP
VCP
DMD
DMD
BLM
TP53BP1
FERMT2
AIFM1
KMT2A
RPL30
TMPO
TPR
DYNC1I2
PEX19
EGFR
HCFC1
GTF2I
GTF2I
TTI1
TOP2B
acc_Q8WTY5
CUL5
RPL6
ILF3
DHX9
DHX9
STAT3
HNRNPA3
UTP14A
SRRM2
PPP4R2
ERBB2
ERBB4
RRP7A
RANBP2
SSRP1

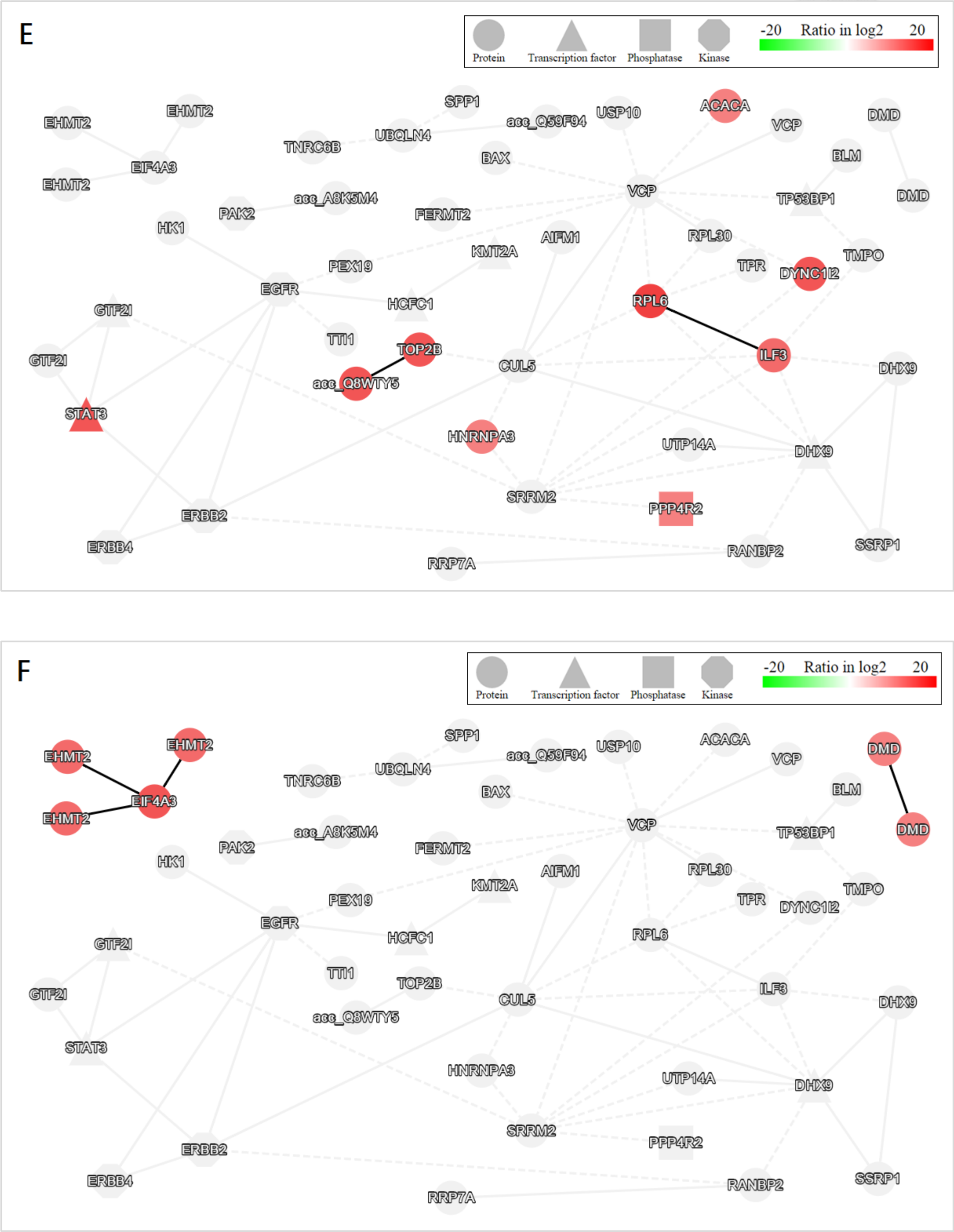


Figure 20 Biological signaling presented by the protein-protein interaction network

The text of each node is the gene name of the protein, if one with the prefix of “acc_”, the protein has not been assigned a gene name yet, but to present its uniprot accession name. The parameter of S.D. as threshold is 1.5.

TABLES

Table 1 Analysis customization input parameters and their default values.

All results in the article are executed on these parameters.

Analysis module	Parameters (default)	Description
Data preprocessing	NA ratio: 30% KNN: k=5	The upload file must be preprocessed first. The filling procedure can use either KNN or average.
Statistical analysis	NONE	View the quality of processed data.
Profile clustering	Number: auto-detection Method: mfuzz	The clustering number can be determined by the user or by auto-detection.
Functional enrichment	Filter: 3.0 S.D. Reference: GO database Cutoff p-value: 1e-15	The reference can be GO biological process database or the sample input.
Kinase activation profile	Occurences: 20 motif-x p-value: 0.000001 Reference: Human Correlation: pearson Cutoff Similarity: 0.5 Filter: 2.0 S.D. Profile p-value: 0.05	The p-value in motif-x must be float type, not scientific notation. All the other parameters not showed are the same with defaults of motif-x. The “background” in motif-x is IPI human proteome. The correlation method can be either pearson or spearman.
Interaction network	Filter: 2.0 S.D.	Construct dynamic network relied on protein interaction information.

Table 2 Public databases integrated in Dyanpho

Database/Service	Update	Description
motif-x [35]	NONE	Online service finds conserved patterns from a large sequence dataset by an iterative statistical approach.
PhosphoNetworks [36]	12/10/2014	Protein microarray-based database contains 4,191 proteins and 3,656 kinase-substrate relationships by performing 289 human phosphorylation reactions.
Gene Ontology [51]	02/10/2015	In DynaPho, only biological process database is involved.
BioGrid [30]	03/20/2015	Online dataset searches 44,978 publications for 826,051 proteins and genetic interactions from well-known model organism species
HPRD [31]	03/20/2015	Online database contains 30,047 proteins and 41,327 protein-protein interactions from existing literature.
InAct [32]	03/20/2015	Online database contains 526,612 protein-protein interactions from existing 13,562 literatures.
CCSB [33]	03/20/2015	A protein-protein interaction database for a number of different organisms.
MINT [34]	03/20/2015	A database stores data on functional interactions between proteins and contains 4,568 interactions and 782 indirect or genetic interactions.

Table 3 Conserved motifs from motif-x services

Serine (S)		Threonine (T)	Tyrosine (Y)
0_......SP.K...	1_.R.R..SP.....	0_......TPP....	–
2_......SP....K	3_......SP...K.	1_......TP.K...	
4_......SP..K..	5_......SP...RR	2_......TP...K.	
6_....P.SP.....	7_......SP.R...	3_....P.TP.....	
8_K.....SP.....	9_...R..SP.....	4_......TP.....	
10_...G..SP.....	11_......SP.....	5_....SPT......	
12_...R..S.SP...	13_...RR.S.S....	6_......TSP....	
14_......S.SP.K.	15_...R..S.S....	7_...SP.T......	
16_......SDDE...	17_....SPSK.....	8_......T...K..	
18_......SDEE...	19_....SPS......	9_......T.SP...	
20_.....RS.SP...	21_......SEDE...	10_......TD.E...	
22_......SDSE...	23_......S.SP...	11_......T....K.	
24_......SEEE...	25_...RR.S......	12_..R...T......	
26_......S..EED.	27_......SD.E...		
28_......S...SP.	29_......SSP..K.		
30_......SDED...	31_.R.RS.S......		
32_......SSP....	33_......S.ED...		
34_......SDSD...	35_......SE.ED..		
36_.R.R..S......	37_...KR.S......		
38_......S.DE...	39_...R..S.E....		
40_......S.D...E	41_......S..E.E.		
42_R..S..S......	43_....P.S..K...		
44_.....DSDD....	45_....RRS......		
46_.....DS..D...	47_......S.E..D.		
48_...R..S.D....	49_......SG.K...		
50_...R..S.G....	51_....R.S.G....		
52_......SD.D...	53_.....DSD.....		
54_...R..S......	55_......S.....K		
56_.....GSG.....	57_R.....S.S....		
58_......S.E....	59_......S.DD...		
60_....R.S......	61_......SE.E...		
62_......S..R...	63_......S..SP..		
64_......S...D.D	65_......S...E..		
66_......S..K...	67_..K...S......		
68_......S..G...	69_...D..S......		
70_......SK.....	71_...K..S......		
72_......S.....E	73_.R....S......		

Table 4 Detailed biological processes with their adjusted p-values on all stages

Biological process (GO)	-log10 (adj. P)
mitotic cell cycle process	72.40
regulation of peptidyl-cysteine S-nitrosylation	69.64
olfactory nerve structural organization	69.50
establishment of glial blood-brain barrier	68.83
regulation of skeletal muscle contraction by regulation of release of sequestered calcium ion	68.57
muscle cell cellular homeostasis	67.00
positive regulation of sodium ion transmembrane transporter activity	66.40
regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion	64.54
chromosome organization	63.36
neurotransmitter receptor metabolic process	62.78
regulation of voltage-gated calcium channel activity	61.39
nucleus localization	59.10
cardiac muscle cell action potential	58.69
regulation of ryanodine-sensitive calcium-release channel activity	56.79
myotube cell development	51.37
regulation of intracellular transport	51.00
protein complex assembly	49.94
positive regulation of cell-matrix adhesion	47.71
skeletal muscle tissue development	44.13
cellular macromolecular complex assembly	42.67
DNA conformation change	31.55
mRNA metabolic process	31.44
viral process	30.60
RNA splicing	28.26
negative regulation of RNA metabolic process	27.30
gene expression	27.18
single-organism intracellular transport	27.16
microtubule cytoskeleton organization	26.38
regulation of organelle organization	25.74
regulation of cell cycle	23.36
regulation of catabolic process	21.37
cellular component disassembly	20.46
membrane organization	20.40
cellular response to DNA damage stimulus	20.20

mitotic nuclear envelope disassembly	20.05
mitotic cytokinesis	19.76
mitotic chromosome condensation	18.22
ATP catabolic process	18.22
mRNA transport	17.97
purine ribonucleoside monophosphate catabolic process	17.88
positive regulation of microtubule polymerization	17.86
mitotic cell cycle phase transition	17.81
cell migration	16.96
negative regulation of protein depolymerization	16.52
metaphase/anaphase transition of mitotic cell cycle	16.38
mitotic nuclear division	15.49
establishment of protein localization to membrane	15.11

Table 5 Detailed composition of DynaPho

directory	description
/administrator	An independent subsystem stores the contact information from users. The operation must be authorized. This subsystem is constructed by LAMP.
/config	It stores lots of information about configuration and setting, such as the network location, SELinux policy. Besides, scripts related to system maintenance or the availability of the task are also stored here. For example, scripts in Python work with the job scheduling in Linux to delete tasks which no more analysis was submitted over seven days.
/core	Scripts are related to main framework of web interface, preprocessing module, and task recording subsystem.
/databases	It contains three key databases, which are used in function enrichment, kinase activation time profile, and interaction network module. In addition, a part of GO databases is independent constructed on MySQL.
/function	It contains scripts which are available to access MySQL databases. MySQL core stores biological process databases of GO. Scripts are designed in parallel computing and multi-task techniques. It also contains webpages for browser presentation.
/kinase	It contains scripts which are available to access PhosphoNetworks databases. Furthermore, scripts in PHP can automatically link motif-x service for the conserved motif analysis. It also contains webpages for browser presentation.
/libs	It contains CSS and javascript libraries (including jQuery, Cytoscape, etc.) which are used in webpage presentation. It also contains documents; for example, posters.
/network	It contains scripts which are available to access summarized protein interaction databases. It also contains webpages for browser presentation.
/profile	It contains scripts which are related to auto-detection method and mfuzz. Scripts are also designed in parallel computing and multi-task techniques. It also contains webpages for browser presentation.
/results	It is the task- and session-based directory. The hierarchy is the order which starts from task ID, analysis module, session ID, to analysis results.
/statistics	Scripts in R can automatically link Plotly service for the ratio distribution analysis. Two main analyses are executed on the uploading procedure. It also contains webpages for browser presentation.
/upload	Scripts are related to preprocessing and designed in parallel computing.
/index.php	It is an access to DynaPho service. It includes the framework of web interface and checks whether the task or the session is available or not.

APPENDIX

2014 Translational and Systems Biology Symposium

DynaPho: inferring signalling dynamics from phosphoproteomics data

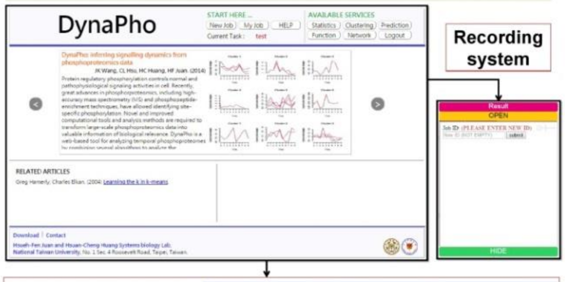
Jian-Kai Wang[1], Chia-Lang Hsu[2], Hsuan-Cheng Huang[3], Hsueh-Fen Juan[1,2,4]

[1]Genome and Systems Biology Degree Program, National Taiwan University,
[2]Department of Life Science, National Taiwan University
[3]Institute of Biomedical Informatics, Center for Systems and Synthetic Biology, National Yang-Ming University
[4]Institute of Molecular and Cellular Biology, National Taiwan University
Email: r02b48005@ntu.edu.tw, hsuancheng@ym.edu.tw and yukijuan@ntu.edu.tw

ABSTRACT

Protein regulatory phosphorylation controls normal and pathophysiological signaling activities in cell. Recently, great advances in phosphorproteomics, including **high-accuracy mass spectrometry** (**MS**) and phosphopeptide-enrichment techniques, have allowed identifying **site-specific** phosphorylation. Novel and improved **computational tools and analysis methods** are required to transform **large-scale** phosphoproteomics data into valuable information of biological relevance. **DynaPho** is a **web-based tool** for analyzing temporal phosphoproteomes by combining several algorithms to analyze the phosphorylation profiles as well as sequence-content of phosphosites and various databases to uncover the dynamics of phosphosignaling. DynaPho consists of five major analysis modules: (1) **basic statistics analysis**; (2) **phosphorylation profile clustering**; (3) time-dependent **functional annotation**; (4) **kinase activation profile**; (5) construction of **kinase-substrate interaction**. DynaPho is **freely** available at http://dynapho.hchuang.info/.

ARCHITECTURE / METHODS


DynaPho
START HERE ...
New Job
My Job
HELP
Current Task : test
AVAILABLE SERVICES
Statistics
Clustering
Prediction
Function
Network
Logout
RELATED ARTICLES
Greg Hamerly, Charles Elkan. (2004) Learning the k in k-means
Download | Contact
Hsueh-Fen Juan and Hsuan-Cheng Huang Systems biology Lab, National Taiwan University, No. 1 Sec. 4 Roosevelt Road, Taipei, Taiwan.
Recording system
Result
OPEN
Job ID: (PLEASE ENTER NEW ID)
submit
HIDE

Preprocessing System	Data status checking modules K nearest neighbors filling modules
Statistics System	Ratio of each phosphorylation sites Expression changed of sites based on time
Phosphoylated Profile Clustering	Distance-based clustering Trend of sets from clustering results
Phosphorylated Sites Functional Annotation	Annotation based on kinase consensus motifs, protein domains, binding motifs, etc.
Kinase Activation Profile	From annotations, kinase activation profile enrichment with databases is possible
Kinase Substrate Interaction Network	From kinase profile, network-like regulation of interaction show signalling dynamics

STRENGTH

- Friendly user interface
- Profile abundance changes
- Upstream to downstream signalling regulations
- Network-like scales between kinase and substrate

RESULTS

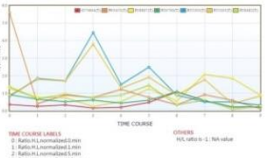


Figure.1 Snapshot of Dynapho
Friendly interface to make users easily capture view to the dynamic trend of interested phosphorylated sites.

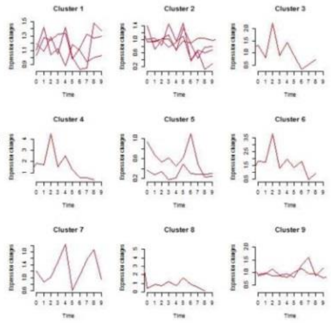


Figure.2 Results of phosphorylated profile clustering

Use the automatic clustering method, G-means, and soft clustering, Mfuzz, to capture the trends in group.

SUMMARY

- Dynapho is the first and integrated web-based tools interpreting dynamic phosphorylation signalling.
- Dynapho provides services analyzing from upstream raw data into downstream valuable information, including profile clustering, functional annotation, kinase-substrate network, etc. to make analyzing phosphoproteome more easily.

REFERENCES

- Chia-Wei Hu, Miao-Hsia Lin, *et al*. (2012) Journal of proteome research.
- Mathieu Courcelles, Se´bastien Lemieux, *et al*. (2011) Proteomics.

POSTER

http://ppt.cc/5DqF

WEBSITE

http://ppt.cc/VfPD