BMC Bioinformatics

Research article

Mining frequent patterns for AMP-activated protein kinase regulation on skeletal muscle

Qingfeng Chen[1] and Yi-Ping Phoebe Chen*[1,2]

Address: [1]School of Engineering & Information Technology, Deakin University, Melbourne, Australia and [2]Australia Research Council (ARC) Centre in Bioinformatics, Australia

Email: Qingfeng Chen - qifengch@deakin.edu.au; Yi-Ping Phoebe Chen* - phoebe@deakin.edu.au

* Corresponding author

Published: 30 August 2006

BMC Bioinformatics 2006, **7**:394 doi:10.1186/1471-2105-7-394

Received: 28 March 2006
Accepted: 30 August 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/394

Abstract

Background: AMP-activated protein kinase (AMPK) has emerged as a significant signaling intermediary that regulates metabolisms in response to energy demand and supply. An investigation into the degree of activation and deactivation of AMPK subunits under exercise can provide valuable data for understanding AMPK. In particular, the effect of AMPK on muscle cellular energy status makes this protein a promising pharmacological target for disease treatment. As more AMPK regulation data are accumulated, data mining techniques can play an important role in identifying frequent patterns in the data. Association rule mining, which is commonly used in market basket analysis, can be applied to AMPK regulation.

Results: This paper proposes a framework that can identify the potential correlation, either between the state of isoforms of α, β and γ subunits of AMPK, or between stimulus factors and the state of isoforms. Our approach is to apply item constraints in the closed interpretation to the itemset generation so that a threshold is specified in terms of the amount of results, rather than a fixed threshold value for all itemsets of all sizes. The derived rules from experiments are roughly analyzed. It is found that most of the extracted association rules have biological meaning and some of them were previously unknown. They indicate direction for further research.

Conclusion: Our findings indicate that AMPK has a great impact on most metabolic actions that are related to energy demand and supply. Those actions are adjusted via its subunit isoforms under specific physical training. Thus, there are strong co-relationships between AMPK subunit isoforms and exercises. Furthermore, the subunit isoforms are correlated with each other in some cases. The methods developed here could be used when predicting these essential relationships and enable an understanding of the functions and metabolic pathways regarding AMPK.

Background

In recent years, there has been a tremendous growth in biological data and the emergence of new, efficient experimental techniques. A variety of genomic and proteomic databases are now publicly accessible over the Internet. However, it is widely recognized that the mere gathering of discrete data is insufficient for us to discover the potential correlations amongst them. The biological interpretation and analysis of these data are crucial. Such biological data not only provides us with a good opportunity for understanding living organisms, but also poses new chal-

lenges. This has led us to the development of a new method to analyze the data.

Protein kinases' regulation data can be a good foundation for understanding their structure, function, and expression. One goal, in terms of analyzing protein kinase regulation data, is to determine how an external stimulus might affect the catalytic subunit and regulatory subunit of protein kinases. Figure 1 presents Protein kinase *X* uses IRS_i, a regulatory subunit (a second protein molecule) to control the activity of a catalytic subunit ICS_j. Each subunit consists of several gene encoding isoforms. Another goal is to determine what isoforms are expressed or unchanged in expression as a result of certain conditions. AMP-activated protein kinase (AMPK) has recently emerged as a potential key signalling pathway, in the regulation of exercise-induced changes in glucose and lipid metabolism in skeletal muscle [1,2]. This enzyme is activated as a result of the alterations in cellular energy levels [3]. The activation of AMPK also exerts long-term effects at the level of both gene expression and protein synthesis, such as positive effects on glucose uptake of heart, food intake of hypothalamus, and negative effects on insulin secretion of pancreas and cholesterol synthesis of liver [4]. Hence, the investigation into the degree of activation and deactivation of subunit isoforms of AMPK will contribute to a greater understanding of AMPK and disease treatment [5].

Unfortunately, the traditional computational methods have been mainly used in sequence alignment [6], gene prediction [7], and microarray analysis. However, the efforts to develop robust methods to analyze AMPK regulation data lag behind the rate of data accumulation. Most of the current analysis rests on isolated discussion of single experimental results. Also, there has been a lack of systematical collection and interpretation of diverse AMPK regulation data. Besides, the existing approaches that seek to analyze biological data cannot cope with the AMPK regulation data that contains status messages of subunit isoforms and stimulus factors. This calls for the use of sophisticated computational techniques.

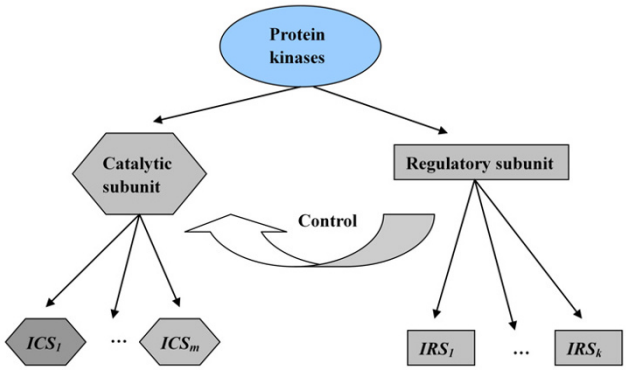

Figure 1
Catalytic subunit and Regulatory subunit of Protein kinases, in which *ICS* and *IRS* represent the isoform of catalytic subunit and regulatory subunit respectively.

Recently, data mining techniques have emerged as a means of identifying patterns and trends from large quantities of data. Among them, association rule mining is a popular summarization and pattern extraction algorithm to identify correlations between items in transactional databases [8]. Several attempts have been made to mine biological databases using association rule mining [9-11]. Earlier investigations mainly focus on discovering an association between the gene expression, genetic pathways and protein-protein interaction. However, not much work has been found to address AMPK regulation data. Hence, it is necessary to identify implicit, but potentially useful, frequent patterns from the AMPK regulation data.

An association rule is an implication of the form $X \Rightarrow Y$, where *X* and *Y* are itemsets. For example, as for AMPK regulation, *Y* and *X* can represent the subunit isoforms of AMPK that is highly expressed or unchanged in expression, and stimulus factors that describe the certain conditions such as the intensity and duration, respectively. A rule might be $\{\text{moderate intensity}\} \Rightarrow \{\alpha_{1a}|, \alpha_{2a}\uparrow\}$ where α_{1a} and α_{2a} represent the activity of α subunits of AMPK. It shows that α_{1a} usually has no change in expression, and α_{2a} is highly expressed in most experiments where the exercise intensity is moderate. Typically, this method requires a user to specify a minimum support threshold for the generation of all itemsets. However, without specific knowledge, it is difficult for users to set the support threshold. Therefore, it would be better to identify top-*N* interesting itemsets, instead of specifying a fixed threshold value for all itemsets of all sizes like [12].

This paper presents a framework by which to analyze the AMPK regulation data derived from the published experimental results. FPtree (Frequent Pattern Tree) algorithm [13] is used to extract frequent itemsets efficiently, and item constraints in the closed interpretation are proposed to specify general constraints on itemset generation. A number of rules of interest are discovered from mining AMPK data. A cursory analysis of some of the like reveals numerous potential associations between the states of subunit isoforms of AMPK, or between the stimulus factor and the state of isoforms, many of which make sense biologically. Those suggesting new hypotheses may warrant further investigation. If a data set from existing experiments has missing values for some subunit isoforms that are intentionally untested in corresponding experiments, these items are filtered out ahead of mining. Furthermore,

the items that are not tested by adequate experiments will be reserved in databases for future use.

This paper is organised as follows. In section 'Methods', the basic concepts used in this paper, are discussed. Section 'Mining AMPK' with item constraints, presents the procedure to find association rules from AMPK regulation data. Section 'Results' explains implementation of the algorithm to discover association rules using experiments. Section 'Discussion' refers to our methodology and future directions. Finally, this paper is concluded in section 'Conclusions'.

Methods

The basics of association rule mining

An association rule is an implication of the form, $A \Rightarrow B$, where A and B are itemsets, and $A \cap B = \emptyset$. The following criteria can be used to evaluate the association rule:

1. *support* for a rule $A \Rightarrow B$ is the percentage of transactions in D that contain $A \cup B$, and is defined as $supp(A \cup B)$; and

2. *confidence* for a rule $A \Rightarrow B$ is defined as $conf(A \Rightarrow B) = supp(A \cup B)/supp(A)$.

According to the support-confidence framework [14], a rule $A \Rightarrow B$ is of interest if $supp(A \cup B) \geq minsupp$ and $conf(A \Rightarrow B) \geq minconf$. In this article, the conventional association rule mining is extended for analyzing AMPK regulation data. Suppose $E = \{E_1, ..., E_n\}$ is a set of experiments. Each experiment consists of an *eid* (experiment identifier) and two itemsets, $E_i = (eid, S_i, ST_i)$; $ST_i \Rightarrow S_i$ is treated as an initial rule. Let $I = \{x \mid x \in S_i \cup ST_i, 1 \leq i \leq n\}$ be a set of items, and $A \subseteq I$ and $B \subseteq I$ be itemsets. A rule $A \Rightarrow B$ has support, s, in the set of experiments if s% of experiments contains A and B. The rule has confidence, c, if c% of experiments containing A, also contain A and B.

Definition of problem

For regulating critical biological processes such as memory, hormone responses, and cell growth, living organisms rely on a family of enzymes called *protein kinases*. In particular, AMPK is activated in skeletal muscle in response to exercise, phosphorylating target proteins along diverse metabolic pathways, such as glucose uptake and fatty-acid oxidation. The response to environmental demands is accomplished by signal transduction by which an extracellular signal interacts with receptors at the cell surface, activating factors in signaling pathways and ultimately sustaining muscle performance by activating skeletal muscle remodeling genes. Furthermore, recent findings point to the AMPK pathway as a potential target for therapeutic strategies to restore metabolic balance to patients, such as type 2 diabetic and obesity patients. Thus, AMPK pathway is a particularly challenging problem in bioinformatics.

Recent studies [1,2] have shown that AMP-activated protein kinase (AMPK) plays an important role in the regulation of exercise-induced changes, which occur in glucose and lipid metabolism in skeletal muscle, and gene expression and protein synthesis. AMPK contains *catalytic subunit isoforms* α_1 and α_2 and *regulatory subunit isoforms* β_1, β_2, γ_1, γ_2 and γ_3, and the regulation controlled by subunits is shown in Figure 2. The subunit isoforms congregate together to perform the functions of AMPK. Therefore, interpretation and analysis of AMPK regulation data and identification of potential associations from the data, may lead to a better understanding of its structure, function and expression. However, only limited studies have been conducted to analyze this valuable data, and the majority of these studies have focused on isolated experiments. In this article, a framework is proposed to discover association rules from a collection of AMPK regulation data.

Given a set of experimental results, E, each result is described by a set of items. The items can be classified into two categories:

1. *Stimulus item* (ST) represents a parameter that is used to measure stimuli. For example, intensity, load and duration are generally used to measure the exercise stimuli. +++, ++, + and - indicate that the stimulus is "*intense*", "*high*", "*moderate*" and "*low*", respectively.

2. *State item* (S) represents isoforms of α, β and γ subunits of AMPK. ↑, | and ↓ are used to indicate "*highly expressed*", "*expressed*" and "*no change*" in expression, respectively.

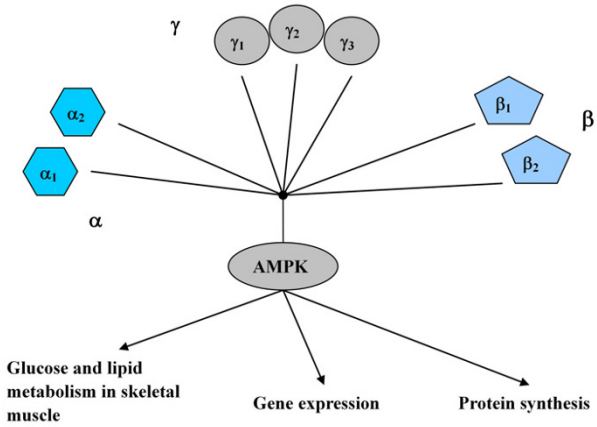

Figure 2
Subunit isoforms of AMPK and its functions.

Items collected from different experiments contain a great deal of hidden information that may be meaningful. For example, in an observed experiment based on the training of moderate intensity treadmill, β_2 and γ_2 isoforms of AMPK are found to be highly expressed in terms of activation in white quadriceps [15]. The initial rule for this experiment can be written as $A = \{$*moderate intensity treadmill*$\} \Rightarrow B = \{\beta_{2a}\uparrow, \gamma_{2a}\uparrow\}$. It is hypothetically derived from an experiment. Nevertheless, in practice, the association rules have to be mined from an experimental data set.

The best known strategy to mining frequent itemsets is Apriori [14], which lives on the essential assumption that all itemsets have a uniform minimum support. However, in reality, the minimum support is not uniform. For example, rules containing *coffee* and *milk* or *coffee* and *sugar* usually have higher support than rules containing *coffee* and *tea*. The occurrence of a large itemset is inherently smaller than that of a small itemset in accordance with probability [16]. On the other hand, it is still troublesome for users to specify an optimal support threshold for all itemsets of all sizes. If the threshold is set too small, there may be too many results for the users. This may be time consuming during the computational phase and result in extra efforts to sort out answers of interest. If it is too large, there may be only a small number of results. In that case, some useful results can be missed. Thus, the constraint-based mining technique has been highlighted in recent work [12]; it provides a flexible way for users to specify a set of constraints and allow them to search and control the interesting frequent patterns. For example, in AMPK regulation, we may only want to know the relationships among items of different attributes, such as the activity of α_1*isoform* and *stimulus factors*, e.g. intensity and duration. Their explanation will be described in the next two sections.

Mining AMPK regulation with item constraints

Deriving initial items

In general, mining of association rules starts with generating all frequent itemsets. The mined data set can be derived from different experimental results. The experimental conditions have to be considered in order to find out correct association rules. To discover frequent itemsets, initial items from experimental data, including experimental conditions such as moderate intensity and treadmill training, are to be derived. Also, the item derived from AMPK regulation data includes not only its status measurement but also item name such as α_{1a} | and $\beta_{2a}\uparrow$.

In AMPK regulation data, *activity*, *protein expression* and *phosphorylation* are used as testing indexes for α, whereas only *protein expression* is used for β and γ. The following steps can be used to generate initial items from experimental data:

- Generate initial rules $ST \Rightarrow S$ from experimental data where S and ST represent *state itemset* and *stimulus itemset*, respectively and

- Derive a set of initial items $I = S \cup ST$.

It is illustrated in the form of $\{x_1, x_2, ..., x_n\}$ where $x_i \in I$. Let the experimental universe be $EID = \{E_1, E_2, E_3, E_4\}$. Table 1 represents an example of initial items derived from different experiments. Each row in Table 1 corresponds to an experiment for AMPK regulation. "-" represents that this item is untested in the corresponding experiments. There are four initial rules, namely, 1) $\{\phi^+, \phi^-\} \Rightarrow \{\alpha_{1a}\downarrow, \alpha_{1p}\uparrow\}$, 2) $\{\phi^+, \phi^-\} \Rightarrow \{\alpha_{1a}\downarrow, \alpha_{1e}|, \alpha_{1p}|, \beta_{1e}|, \beta_{2e}|\}$, 3) $\{\phi^+, \phi^+\} \Rightarrow \{\alpha_{1a}\downarrow, \alpha_{1e}|, \beta_{1e}|\}$ and 4) $\{\phi^{++}, \phi^+\} \Rightarrow \{\alpha_{1a}|, \alpha_{1e}\downarrow, \alpha_{1p}|, \beta_{1e}|, \beta_{2e}|\}$ where subscripts *a*, *p* and *e* represent *activity*, *phosphorylation* and *protein expression*, respectively. Therefore, $\forall A \subseteq \{\alpha_{1a}\downarrow, \alpha_{1a}|, \alpha_{1e}|, \alpha_{1e}\downarrow, \alpha_{1p}\uparrow, \alpha_{1p}|, \beta_{1e}|, \beta_{2e}|, \phi^+, \phi^{++}, \phi^-, \phi^+\}$ is an itemset and $\forall \gamma \in \{\alpha_{1a}\downarrow, \alpha_{1a}|, \alpha_{1e}|, \alpha_{1e}\downarrow, \alpha_{1p}\uparrow, \alpha_{1p}|, \beta_{1e}|, \beta_{2e}|, \phi^+, \phi^{++}, \phi^-, \phi^+\}$ can be treated as an initial item.

In order to determine association rules, the frequent itemsets need to be sorted out from the obtained initial items.

Specifying item constraints

FP-tree algorithm [13] constructs a frequent pattern tree that has an extended prefix-tree structure and stores quantitative information about frequent patterns. It can generate a complete set of frequent patterns. The typical data format for using this algorithm is shown in Table 1. We notice that not every item does occur in all experiments in this Table. Our method marks these items with *filtering symbols* so as to filter out the itemsets that contain such items during the itemset generation. Actually, if an itemset is not tested in most of the experiments, it will not be considered when finding association rules for the time being, instead it will be reserved for future use.

Suppose $support(X, E)$ denotes the support of X in experiment E and $f(X, E)$ represents the occurrence of itemset X in E, it can be defined as:

$$support(X, E) = |f(X, E)|/|E| \quad (1)$$

Table 1: An Example of Experimental Database

EID	Items						
	α_1			β		ϕ	ϕ
E_1	$\alpha_{1a}\downarrow$	-	$\alpha_{1p}\uparrow$	-	-	ϕ^+	ϕ^-
E_2	$\alpha_{1a}\downarrow$	$\alpha_{1e}\|$	$\alpha_{1p}\|$	$\beta_{1e}\|$	$\beta_{2e}\|$	ϕ^+	ϕ^-
E_3	$\alpha_{1a}\downarrow$	$\alpha_{1e}\|$	-	$\beta_{1e}\|$	-	ϕ^+	ϕ^+
E_4	$\alpha_{1a}\|$	$\alpha_{1e}\downarrow$	$\alpha_{1p}\|$	$\beta_{1e}\|$	$\beta_{2e}\|$	ϕ^{++}	ϕ^+

where $f(X, E) = \{E_i \in E \mid E_i \text{ contains } X\}$. For example, *support*($\alpha_{1e}|$, E) in Table 1 is 2/4 or 50%.

X is a high occurrence itemset if *support*(X, E) ≥ *minoccur* (*minimum occurrence*). To avoid missing valuable initial itemsets, we assume *minoccur* = 2/|E|, which is the lower bound of minimum support corresponding to the support requirement of at least 2 experiments. Otherwise, it is called a *low occurrence itemset*. Consequently, a collection of *initial interesting itemsets* can be obtained in light of the initialized *filtering symbols* and given *minimum occurrence*.

To extract frequent itemsets, users usually need to specify a fixed minimum support but this method has been criticized due to its difficulty [12,13,16]. As described above, it is quite subtle to set a minimum support: a too small threshold may result in the output of many redundant patterns, whereas a too big one may generate no answer or miss interesting knowledge. In addition, the probability of occurrence of a larger size itemset is inherently much smaller than that of a smaller size itemset [12]. A more flexible option is to allow users to specify different thresholds in accordance with different itemsets [12]. Consequently, the *top-N* interesting itemsets are returned as answers.

Suppose a ***k-itemset*** denotes a set of items containing k items. We have the following two definitions.

Definition 1. ***Top-N* interesting *k*-itemsets**. Suppose a ***k*-itemset** is sorted in descending order by their supports. Let s be the support of the *Nth k-itemset* in the sorted list. The ***top-N* interesting *k*-itemsets** represent the set of *k-itemsets* whose supports are equal to or larger than s.

From the observation, it is possible that there are multiple itemsets that have the same support s. The ***top-N* interesting *k*-itemsets** may contain more than N itemsets. In this extreme case, it will be reported to the user, rather than returning all of them [12].

Definition 2. ***Top-N* interesting itemsets** is the union of the ***top-N* interesting *k*-itemsets** where $1 \le k \le k_{max}$ and k_{max} is the upper bound of the size of itemsets we would like to find. An itemset is of interest if, and only if, it is in the ***top-N* interesting itemsets**.

EXAMPLE 3.1. In Table 1, we specify $k_{max} = 7$ because the maximal tested items in every experiment are 7. The **interesting *1*-itemsets** is $\{\alpha_{1a}\downarrow, \alpha_{1e}|, \alpha_{1p}|, \beta_{1e}|, \beta_{2e}|, \phi^+, \phi^-, \phi^+\}$ after filtering $\alpha_{1a}|$, $\alpha_{1e}\downarrow$, $\alpha_{1p}\uparrow$ and ϕ^{++} by formula (1). As a result, ***top-1* interesting 1-itemsets** = $\{\alpha_{1a}\downarrow\}$ and ***top-2* interesting 1-itemsets** = $\{\alpha_{1a}\downarrow, \beta_{1e}|, \phi^+\}$.

From the observation, it is clearly impractical to enumerate all ***top-N* interesting itemsets**. Although several ***top-N*** mining algorithms [12,13,16,17] already exist, they focus, without exception, on finding all ***top-N* itemsets**. However, this may be time-consuming and can lead to many useless or uninteresting itemsets. For example, itemsets from {*training*, *glycogen*, *diabetes*, *nicotinic acid*, *intensity*, *duration*} rather than their subsets are of interest because a part of the testing indexes cannot correctly describe an experiment. Our method is to partition a set of items, I, into several bins, where each bin B_j contains a subset of items in I. We use item constraint in a similar way to the enumeration-based specification defined in [16] so that the items in a bin are not distinguished in terms of the specification. For example, $B_1 = \{\beta_{1e}, \beta_{2e}\}$ and $B_2 = \{\alpha_{1a}, \alpha_{1e}\}$ represent that the user is interested in protein expression of β, rather than α and γ, and only *activity* and *protein expression* of α, rather than *phosphorylation*. The constraint can be expressed in the following brief formula:

$$IC_i(B_{i1}, \ldots, B_{im}) = N_i \quad (2)$$

where $B_{ij} \cap B_{ik} = \varnothing$, $1 \le j, k \le m$, $j \ne k$. N_i is the number of itemsets satisfying IC_i in terms of a derived support described in Section 'Results'. It explicitly defines what particular items we focus on and which should be presented in the identification of frequent patterns.

The concept of **closed interpretation** in [16] is adopted as well. An itemset $X \in I$ satisfies a constraint IC_i in the closed interpretation if X contains exactly one item from each B_{ij} in IC_i, and these items are completely different. Suppose IC_j is an item constraint, and $|IC_j| = k$. As for close interpretation, a collection of *k-itemsets* can be generated. These itemsets are sorted in light of their support in descending order. Let the N_jth greatest support be θ. Consequently, all *k-itemsets*, with support not less than θ, are interesting due to the constraint. We call them **top-N_j interesting itemsets** of IC_j in the closed interpretation. Usually, users can specify several item constraints. In this scenario, we want to find **top-N_i interesting itemsets** for each IC_i in the closed interpretation.

Suppose, for example, we partition the items in Table 1 into seven bins: $B_1 = \{\alpha_{1a}\downarrow, \alpha_{1a}|\}$, $B_2 = \{\alpha_{1e}|, \alpha_{1e}\downarrow\}$, $B_3 = \{\alpha_{1p}\uparrow, \alpha_{1p}|\}$, $B_4 = \{\beta_{1e}|\}$, $B_5 = \{\beta_{2e}|\}$, $B_6 = \{\phi^+, \phi^{++}\}$, $B_7 = \{\phi^-, \phi^+\}$. Let $IC_1(B_1, B_2, B_3) = 2$, $IC_2(B_4, B_5) = 3$ and $IC_3(B_6, B_7) = 2$. Consider itemset $X_1 = \{\alpha_{1a}\downarrow, \alpha_{1e}|\}$, $X_2 = \{\beta_{1e}|, \beta_{2e}|\}$ and $X_3 = \{\phi^+, \phi^-\}$, they correspond to bin patterns (B_1, B_2, B_3), (B_4, B_5) and (B_6, B_7), respectively. Therefore, we say that X_1 satisfies IC_1 in the closed interpretation;X_2 satisfies IC_2 in the closed interpretation; and X_3 satisfies IC_3 in the closed interpretation.

Until now, we have not referred closely to how the filtering symbols are specified. The untested items in experiments need to be pruned before going on to set up item constraints. Our approach converts the initial items of experiments into the format of non-negative integer. Consequently, each bin can contain several non-negative integers but any two bins must be disjoint. Table 2 is a conversion of Table 1, in which 8, 12, 29 and 30 correspond to the untested items regarding α_{1e}, α_{1p}, β_{1e} and β_{2e}, respectively. Therefore, the bins defined in the last paragraph can be converted into $B_1 = \{1, 2\}$, $B_2 = \{5, 6\}$, $B_3 = \{10, 11\}$, $B_4 = \{26\}$, $B_5 = \{29\}$, $B_6 = \{91, 92\}$, and $B_7 = \{100, 101\}$. These filtering symbols are reserved in the data set and will be skipped when finding patterns from the *initial interesting itemsets*.

EXAMPLE 3.2. In Table 2, the non-negative integers 8, 12, 27 and 30 represent the specified filtering symbols. They correspond to the untested items with respect to α_{1e}, α_{1p}, β_{1e} and β_{2e}, respectively.

Discovering association rules

The obtained **top-N_i interesting itemsets** for item constraints in the closed interpretation can then be used to construct association rules. There are two approaches to identify association rules, depending on which antecedents or consequents of rules the item constraint *IC* are defined. Let X and X' be top interesting itemsets from *IC* and *IC*', respectively, and $X' \subset X$.

1. if $supp(X)/supp(X') \geq minconf$, $X' \rightarrow X - X'$ is a rule of interest.

2. if $supp(X)/supp(X - X') \geq minconf$, $X - X' \rightarrow X'$ is a rule of interest.

In addition, those rules that do not make sense biologically are pruned while mining AMPK regulation data. Hence, the valid rules should conform to the rules as given below:

- *Rule1*: $A \Rightarrow B$, $A \subseteq S$ and $B \subseteq S$ and

- *Rule2*: $A \Rightarrow B$, $A \subseteq ST$ and $B \subseteq S$.

Table 2: A Converted Experiment Dataset

EID	Items						
	α_1				β	ϕ	ϕ
E_1	1	8	11	27	30	91	100
E_2	1	6	10	26	29	91	100
E_3	1	6	12	26	30	91	101
E_4	2	5	10	26	29	92	101

If the antecedent or consequent of a rule contains either both state items and stimulus items, or only stimulus items, it will be pruned. Obviously, the relation between only stimulus items such as *intensity* and *time* is not meaningful in biology at all. Besides, state items actually rely on stimulus items; therefore, they should not be included in the antecedent and consequent of a rule simultaneously. These kinds of rules are uninteresting and need to be pruned, and the search space can be reduced.

Results

Pre-processing of experimental data

FP-tree algorithm is applied to generate all frequent itemsets and find association rules that make sense biologically. We experimented on the published experimental data of AMPK regulation, in which AMPK is activated by endurance training in human skeletal muscle. The data is collected through searching the NCBI database and surely ranges over all the openly available data. Originally, the data contained 17 attributes, 46 items and 45 experiments [see Additional file 1]. After conversion into the format of non-negative integer, there are 57 items owing to the insert of filtering items [18]. In addition, initial items from experiments are divided into state items and stimulus items. Each item corresponds to an attribute/value pair.

AMPK consists of state items:*catalytic subunit* (α) and *regulatory subunit* (β, γ). The isoforms (α_1, α_2) of α are measured by *activity*, *phosphorylation* and *protein expression* but the isoforms of β and γ (β_1, β_2, γ_1, γ_2, γ_3) are measured only by *protein expression*. As a result, there are 11 state items including all isoforms of AMPK, α_{1a}, α_{1e}, α_{1p}, α_{2a}, α_{2e}, α_{2p}, β_{1e}, β_{2e}, γ_{1e}, γ_{2e} and γ_{3e}. Six stimulus indexes, including *training*, *glycogen*, *diabetes*, *nicotinic acid*, *intensity* and *duration*, are considered.

Table 3 is a random example of the activity and expression of AMPK in skeletal muscle, in which *N*, *I* and *D* represent *nicotinic acid*, *intensity* and *duration*, respectively. Each column corresponds to an experiment. The subscript *a* and *e* means *activity*, and *protein expression*, respectively. 1 and 5 represent *no change* in corresponding items; 2 and 6 represent *expressed*; 50 and 51 represent *trained* and *untrained*, respectively; 60 represents *normal glycogen*; 70 represents

Table 3: Activity and expression of α_{1a} and α_{1e} of AMPK in skeletal muscle

α_{1a}	α_{1e}	*Training*	*Glycogen*	*Diabetes*	*N*	*I*	*D*
2	5	50	60	70	80	92	101
1	5	51	60	70	80	90	103
1	6	51	60	70	80	91	100
1	6	51	60	70	80	91	101

normal nicotinic acid; 100, 101 and 103 represent the duration under 20 minutes, between 20 and 60 minutes and above 90 minutes, respectively; 90, 91 and 92 represent the *maximal oxygen uptake* (VO(2)max) under 50%, VO(2)max between 65% and 75%, and VO(2)max between 80% and 100%, respectively.

To generate constraint specification, we group the items from the same attribute into a bin, yielding 17 bins for the 17 attributes. Although each bin corresponds to only one attribute, each attribute can contain several state measurements. Thus, each bin can have more than one item. We organise these items into seventeen disjoint non-negative integer intervals so that each bin B_i contains the items matching integers in the *i*th interval. Table 4 shows the specified intervals, and the details can be found in [18]. We specify item constraint IC_i in the closed interpretation [see Additional file 1]. Suppose BV_i represents a bin variable.

$$IC_i(BV_1,..., BV_k) = N_i,\ 0 \le k \le k_{\max} \quad (3)$$

$k_{\max}$ is the upper bound of the size of itemsets that users want to find. We specify N_i as the number of itemsets that satisfies constraint IC_i with supports $\ge \theta_i(BV_1, ..., BV_k)$, where

$$\theta_i(BV_1,...,BV_k) = \begin{cases} 0.044 & if\ \lambda^{k-1} \times S(BV_1) \times \cdots \times S(BV_k) \le 0.044 \\ 1 & if\ \lambda^{k-1} \times S(BV_1) \times \cdots \times S(BV_k) > 1 \\ \lambda^{k-1} \times S(BV_1) \times \cdots \times S(BV_k) & otherwise \end{cases} \quad (4)$$

where $S(BV_i)$ is the smallest support of the items in the bin BV_i. λ is an integer larger than 1 and used to slow down the decrease of $S(BV_1) \times ... \times S(BV_k)$ in case of large k [16].

Rule generation

Unlike traditional methods, our approach starts with pruning items of low occurrence, rather than generating *frequent itemsets* directly. *FP-tree* algorithm is implemented on the obtained initial items to generate initial interesting itemsets. The process is repeated until all initial interesting itemsets are extracted. The procedure of finding association rules consists of four phases:

1. Generate *initial interesting itemsets* from *initial items*.

2. Sort out the itemsets containing no filtering symbols.

3. Set up bin B_j and item constraint IC_i.

4. Identify association rules using IC_i.

The value of N_i is determined by formula (3) and (4). We assign λ with 5 like [12]. $k_{\max} = 9$ is set using the principle of rule generation in section 3.3 because the itemsets containing more than 9 items are not interesting according to the specified maximum item constraints $IC(B_9, B_{10}, B_{11}, B_{12}, B_{13}, B_{14}, B_{15}, B_{16}, B_{17})$. There are 47 item constraints specified in a file for the itemsets in the closed interpretation such as $IC(B_{12}, B_{13}, B_{14}, B_{15}, B_{16}, B_{17})$ and $IC(B_1, B_4)$. The constraints are classified into two categories:

1. Contain all items of stimulus factors at least; and

2. Contain only items of AMPK subunits.

They enumerate the items in a bin in the closed interpretation, on the basis that the user is interested in only the items with respect to specification, rather than all possible combinations of items.

Based on *Definition 2*, the obtained itemsets, due to item constraints, need to be sorted in light of their supports in descending order, by which we can generate the *top-*N_i interesting itemsets for each corresponding bin IC_i. Therefore, the mining will focus on finding out association rules on the basis of these *top-N* interesting itemsets. Table 5 shows the results of *initial interesting itemsets*, *itemsets without filtering symbols*, *sorted itemsets in bins in the closed interpretation* and *top-N interesting itemsets in bins*. From the observation, the search space is greatly reduced as a result of the use of filtering symbols and item constraints.

Table 4: The bins and corresponding interval

B_i	Attibute	Interval	B_i	Attibute	Interval
B_1	α_{1a}	[1, 4]	B_{10}	γ_{2e}	[34,37]
B_2	α_{1e}	[5, 8]	B_{11}	γ_{3e}	[38, 40]
B_3	α_{1a}	[9 12]	B_{12}	training	[50, 51]
B_4	α_{1a}	[13, 16]	B_{13}	glycogen	[60, 62]
B_5	α_{1e}	[17, 20]	B_{14}	diabetes	[70, 71]
B_6	α_{2p}	[21, 24]	B_{15}	nicotinic acid	[80, 81]
B_7	α_{1e}	[25, 27]	B_{16}	intensity	[90, 93]
B_8	α_{2e}	[28, 30]	B_{17}	duration	[100, 103]
B_9	γ_{1e}	[31, 33]			

Table 5: The result of itemset generation

Itemset	Number
Initial interesting itemset	964519
Itemset without filtering symbols	14985
Sorted itemset in bins	97
Top-N interesting itemset in bins	97

The *top-N* interesting itemsets are then used to identify frequent patterns based on the defined criteria in subsection 'Discovering Association Rules', by which some uninteresting rules are pruned. We vary the minimum confidence starting from 0.4 to 1 by increasing 0.1 each time. Figure 3 shows the result of frequent patterns. We observe that there is no sharp drop in rule output when setting the minimum confidence from 0.7 to 1 in comparison with 0.6. Therefore, we select the results by 0.7 in contrast to the results by 0.6 in the following analysis. There are 74 and 51 association rules by minimum confidence 0.6 and 0.7, respectively. From these rules, we can find many potential correlations between itemsets. The rules by 0.7 are classified into the form of Rule1 (40 rules) and the form of Rule2 (11 rules) in terms of the definition in subsection 'Discovering Association Rules'. Nevertheless, the rules need to be pruned because some of them can overlap with each other. Suppose R_i: $A_i \rightarrow B_i$ and R_j: $A_j \rightarrow B_j$ are two rules. The pruning complies with the principles below:

- *Delete* R_j, if $A_i = A_j$ and $B_i \supseteq B_j$

- *Replace* R_i and R_j with $A_i \vee A_j \rightarrow B_i$, if $B_i = B_j$, $A_i \not\subset A_j$ and $A_j \not\subset A_i$

Finally, we obtain 32 rules after pruning. For example, 29 $\rightarrow$ 32 are removed due to 29 $\rightarrow$ 6 32; 14 $\rightarrow$ 1 and 13 $\rightarrow$ 1 are replaced by 14 $\vee$ 13 $\rightarrow$ 1 where 1, 6, 13, 14, 29 and 32 denote $\alpha_{1a}\downarrow$, $\alpha_{1e}|$, $\alpha_{2a}\downarrow$, $\alpha_{2a}|$, $\beta_{2e}|$ and $\gamma_{2e}|$, respectively. For convenience, we reconvert the integers into readable symbols in section 'Methods'. Table 6 shows some selected association rules mined from the AMPK regulation data set, in which *t*, *I*, *T*, *ng*, *lg*, *hg*, nc, *ut* and *nd* represent *time*, *intensity*, *trained*, *normal glycogen*, *low glycogen*, *high glycogen*, *untrained* and *normal diabetes*, respectively.

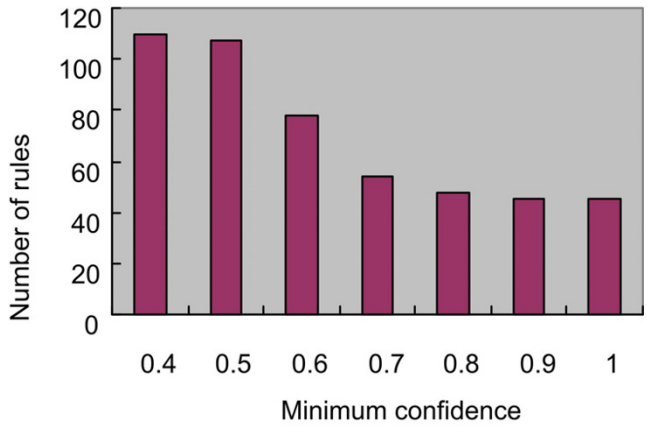

Figure 3
The frequent patterns for AMPK regulation data.

Performance comparison and feature discovery

Our miner (eFP) extends the FPgrowth* method [19] to identify the potential correlation from AMPK regulation. Table 7 presents a performance comparison between our miner and algorithms FPgrowth [13], dEclat [20] and MAFIA [21]. In the comparison, we use a synthetic sparse dataset *T40I10D100K* http://www.almaden.ibm.com/cs/quest/syndata.html and a dense dataset *connect-4* http://www.cs.sfu.ca/~wangk/ucidata/dataset/connect-4/.

In the FI (frequent itemset) mining, the comparison shows that eFP outperforms FPgrowth, dEclat and MAFLA algorithms, and can still run for a small minimum support. It is observed that both dElcat and MAFIA suffer from high memory consumption and low speed when a small minimum support is used. eFP has almost the same running time as FPgrowth in case of the dense dataset or very low levels minimum support, but shows a speedup when a sparser dataset is applied. eFP uses the compact FPtree [19], it not only spends less time on constructing and traversing the trees, than the time on TID-array intersections (dEclat) and bitvector *and* operations (MAFLA), but also needs less main memory space for storing FP-trees than that for storing diffsets or bitvectors. Thus eFP runs faster than the other three algorithms, and it performs well, even when giving very small minimum support.

In addition, our framework includes some novel features, by which to efficiently identify the association rules of interest. Preprocessing of experiment data is used to con-

Table 6: Selected association rules from AMPK regulation data set

	Association rule
1	$\{t^{+}\ ng\ I^{+}\ nc\ ut\ nd\} \vee \{hg\ t^{+}\ nc\ nd\ I^{+}\ T\} \vee \{t^{++}\ nc\ I^{+}\ ut\ ng\ nd\} \rightarrow \{\alpha_{1a}\downarrow\}$
2	$\{lg\ T\ nd\ nc\ I^{+}\ t^{+}\} \rightarrow \{\alpha_{1a}\downarrow\alpha_{2a}|\}$
3	$\{I^{++}\ T\ t^{+}\ nd\ nc\ ng\} \rightarrow \{\alpha_{2a}|\}$
4	$\{I^{++}\ t^{-}\ ut\ nc\ nd\ ng\} \rightarrow \{\alpha_{2a}\uparrow\}$
5	$\{I^{++}\ t^{+}\ nc\ nd\ ng\ T\} \rightarrow \{\alpha_{2p}|\alpha_{1p}|\}$
6	$\{\alpha_{1a}\uparrow\} \rightarrow \{\alpha_{2a}\uparrow\}$
7	$\{\alpha_{1e}|\} \rightarrow \{\alpha_{2e}|\}$
8	$\{\alpha_{1p}\uparrow\} \rightarrow \{\alpha_{1a}\downarrow\alpha_{2a}\uparrow\alpha_{2p}\uparrow\}$
9	$\{\beta_{2e}|\beta_{1e}|\} \vee \{\gamma_{2e}\uparrow\gamma_{1e}|\gamma_{3e}|\} \rightarrow \{\alpha_{2a}\uparrow\alpha_{1a}\downarrow\}$
10	$\{\gamma_{1e}|\gamma_{3e}|\} \rightarrow \{\gamma_{2e}\uparrow\beta_{2e}|\beta_{1e}|\}$
11	$\{\alpha_{1p}|\} \rightarrow \{\alpha_{2p}|\}$
12	$\{\alpha_{1e}|\alpha_{2e}\downarrow\} \rightarrow \{\beta_{1e}\downarrow\beta_{2e}|\}$

Table 7: Performance comparison in mining FI (frequent itemsets)

Miner	Data set	Minimum support (%)	CPU Time(s)
eFP	*T40I10D100K*	$0.05 \leq minsupp \leq 2$	[4.5, 390]
FPgrowth	*T40I10D100K*	$0.05 \leq minsupp \leq 2$	[9.5, 390]
dEclat	*T40I10D100K*	$0.25 \leq minsupp \leq 2$	[7.9, 400]
MAFIA	*T40I10D100K*	$0.75 \leq minsupp \leq 2$	[60, 5936]
eFP	*connect-4*	$20 \leq minsupp \leq 90$	[0.04, 740]
FPgrowth	*connect-4*	$20 \leq minsupp \leq 90$	[0.04, 740]
dEclat	*connect-4*	$20 \leq minsupp \leq 60$	[1, 97]
MAFIA	*connect-4*	$20 \leq minsupp \leq 80$	[8.1, 593]

vert the qualitative data into quantitative data (non-negative integers). Particularly, filtering symbols are used to assist in pruning the untested items. They facilitate the identification of frequent itemsets. The constraint-based mining technique, namely top-*N* interesting itemsets, provides a flexible way for users to specify a set of constraints and allow them to search and control the interesting frequent patterns. The experiments discover a number of interesting frequent patterns that make sense biologically. They not only demonstrate former experiments but also predict some potential results that were unknown previously. These can benefit to the understanding of the signalling pathway of AMPK in regulating metabolism, and disease diagnosis and treatment.

Interpretation

AMP-activated protein kinase (AMPK) is a protein kinase which is ubiquitously expressed and functioned as a stress sensor of the cellular energy status. A functional AMPK exists as a heterotrimetric complex comprising one catalytic subunit α and two regulatory subunits β and γ. In mammals, each subunit is encoded by multiple genes (α_1, α_2, β_1, β_2, γ_1, γ_2, γ_3) [22]. AMPK activity is regulated not only by cellular AMP/ATP ratio, but also by upstream kinases. The skeletal muscle is one of most energy turnover tissue in mammals. AMPK was found to play an important role in the regulation of muscle metabolism. We have selected 45 experiments [18] in relevant bio/medical literatures in Medline by searching the NCBI using keywords "*AMPK and human skeletal muscle*" (103 papers) and "*AMPK and endurance training*" (16 papers), respectively.

We try to find out potential relationships between the AMPK isoform expression and activity, and association between subunits. As mentioned above, this relationship is critical to assist biologists in further predicting AMPK pathways. These are usually hidden in experimental data and these pathways are not easy to identify by traditional statistics. From the experimental results in this study, we have selected 12 association rules to describe AMPK regulation in skeletal muscle, shown in Table 6. For example, as for rule 1 in Table 6, the items from the rule are mentioned together in 27% of the relevant experiments, but only 2% or 0% for those items in other rules, such as $\{ng\ t^{-}\ nc\ nd\ I^{++}\ T\}$ and $\{\alpha_{1a}\downarrow\}$ (2%), $\{ng\ t^{++}\ nc\ nd\ I^{++}\ T\}$ and $\{\alpha_{1a}\downarrow\}$ (0%), $\{ng\ t^{+}\ nc\ nd\ I^{++}\ T\}$ and $\{\alpha_{1a}|\}$ (2%), $\{ng\ t^{+}\ nc\ nd\ I^{+}\ ut\}$ and $\{\alpha_{1a}|\}$ (0%), $\{ng\ t^{+}\ nc\ nd\ I^{++}\ ut\}$ and $\{\alpha_{1a}\uparrow\}$ (2%) and $\{ng\ t^{++}\ nc\ nd\ I^{++}\ ut\}$ and $\{\alpha_{1a}\uparrow\}$ (0%). In other words, the items in rule 1 occur much more frequently than those from other rules under the same indexes. The remaining rules can be explained in the same way. These can be secondary evidence to support the fact that our chosen rules are more important than the other ones.

The rules by 0.7 (minimum confidence) in Table 6 are a subset of the originally obtained rules [see Additional file 2]. They not only show that in most of the experiments where stimuli were imposed the isoforms of AMPK can be activated, but also represent that specific correlations between the states of isoforms exist. For example, rule 4 in Table 6 states that in most of the cases where α_{1a} was highly expressed, α_{2a} was highly expressed too. The rest of the rules in Table 6 can be interpreted in a similar manner. In particular, the rules that are not previously known will be highlighted.

Looking at Table 6 and the supplementary rules, it's seen that the expression of α, β and γ varies from diverse exercise stimuli. In particular, skeletal muscle expresses predominantly the α_2 and β_2 subunits. Referring to the rules by 0.6, we see β_1, β_2, γ_1 and γ_3 except γ_{2e} are also activated (interacting with α) in response to endurance training. This is due to β and γ that are two regulatory subunits, they are activated only when they are associated with α subunit. However, these rules have a slightly lower confidence than the cutoff threshold (70%), which would help explain their absence in the rules by 0.7. Synthetically, these factors/subunits/isoforms α_{1a}, α_{2a}, α_{1p}, α_{2p}, α_{1e}, α_{2e}, β_{1e}, β_{2e}, γ_{1e} and γ_{3e} are actively involved in regulating the metabolism in response to energy demand and supply in skeletal muscle. We seek more experiments to ascertain the characteristic of γ_{2e} (in white muscle). In addition, we see a number of rules that state the latent correlation between the isoforms of AMPK. Most of them are new and

make sense biologically. Furthermore, we can predict some latent associations that are selected for biologists to test in future experiments.

From our observations, most of the found rules except rule 2, 3, 4, 5 were unknown previously. They are important due to the identification of innovative correlations between AMPK subunits, which imply meaningful information for understanding their association and regulation in signalling pathways. Recent experimental results also demonstrate this [4,23,24]. We also view rule 1 as a type of specific rule using disjunctive normal form in the antecedent. This rule is able to integrate the knowledge from three sub-rules and lead to new knowledge using comparison amongst and between them. The details can be seen below.

Looking at rule 1 in Table 6, the activation of α_1-AMPK usually has no change if the exercise intensity is not high enough and the duration is not extra long. Regardless of the status of training, glycogen, diabetes and nicotinic acid, α_1 activity remains unchanged. Although some experiments suggested that AMPK can be substantially activated during maximal sprint-type exercise in humans [25-27]. Unfortunately, this rule cannot be generated at this stage due to insufficient experimental support. Rule 1 is a novel point since we only have seen α_1 activity changes in very few experiments. In addition, it is innovative because the disjunction normal form of antecedents may simultaneously integrate information from multiple experiments. Thus, we determine that α_1 activity cannot be significantly affected by the status of training, glycogen, diabetes and nicotinic acid because α_2 instead of α_1 predominantly localized in skeletal muscle. Such comparison cannot be achieved through traditionally isolated experiments. This can be tested in future experiments and aids in understanding the properties of α_1. Thus, the regulation of AMPK in skeletal muscle is probably relied on α_2 rather than α_1. Furthermore, AMPK may be a critical mediator or exercise-induced changes in glucose uptake and fatty acid oxidation in human skeletal muscle, and the AMPK α_2-containing heterotrimer is possible to be the predominant complex responsible for the regulation in both healthy and diabetic subjects.

Looking at rules 2, 3 and 4 in Table 6, α_{2a} is expressed in skeletal muscles from untrained to trained individuals. Nevertheless, α_{2a} is only highly expressed in skeletal muscle from well-trained individuals in both moderate-intensity and high-intensity training. Importantly, α_2 activity is probably activated more easily in skeletal muscle from untrained individuals than well-trained individuals at the same relative intensity. These associations are in accordance with that of [1,28-30]; α_{1a} is usually unchanged below high-intensity training, which is in agreement with the results of [29,31]. A recent study also supports these rules [32]. These findings reveal that α_2 and α_1, may play different physiological roles in mediating metabolic events in skeletal muscle. Actually, AMPK α_2 is predominantly localized in skeletal muscle, heart and liver, whereas α_1 is widely distributed. In addition, the acute activation of α_2-AMPK during exercise may result in not only a significant increased glucose disposal in muscle but also decreased malonyl-CoA contents, which might ameliorate insulin resistance and improve glycemia [1]. These may explain the above rules. Indeed, AMPK signalling was reduced by either overexpressing a kinase-dead α_2-AMPK construct or knocking out the catalytic α_2-isoform [33,34]. Although an increase in α_2 activity of subjects with type 2 diabetes was found during acute exercise, accompanying a significant decrease in blood glucose concentrations [1], we need more results to establish this rule.

Looking at rule 5, α-AMPK phosphorylation on the α_1 and α_2 are increased in skeletal muscle from well-trained individuals with high-intensity and moderate duration training. This is because the phosphorylation of α_1 and α_2 is more relative to cell regulation by environment stress. In contrast, the expression of α_2 tends to be gene regulation. This suggests that AMPK activity or upstream kinase(s) is being regulated by training, which is in agreement with the results of [3,35,36]. A newly published literature also defends this rule [37].

Looking at rules 6 and 7, both AMPK α_1 and α_2 expressed in skeletal muscle. There is no clear evidence that α_2 activity is leaded by α_1 since they are actually independent with each other. In [33], α_1 protein expression was increased 2–3-fold in α_2 knockout mice, while α_2 protein level in α_1 knockout mice is comparable with that observed in WT mice. Since α_1 activity is usually unchanged while α_2 activity is easily increased under the same exercises. α_1 activity might change only under higher intensity exercises. In contrast, α_2 activity will certainly go up under the same condition (higher intensity). That is why we get Rule 6.

Looking at rule 8, in comparison with α_2, α_1 activity mostly remains unchanged, notwithstanding the high increase of phosphorylation of α_1 because α_2 phosphorylation should be higher and contributed to higher α_2 activity after exercise. Actually, the phosphorylation of α is usually presented as a ratio of phosphorylated α divided by total α. Although α_1 phosphorylation may slightly increases, this cannot cause the increase of α_1 activity. As to our knowledge, no literature in Medline previously points out the relationship between phosphorylation and expression of α_2. The experimental results in [38] also match our computational discovery.

Rule 9 in Table 6 shows that α activity is co-expressed with expression of β or γ. α_2 activity rather than α_1 activity appears to correlate with expression of β and γ subunits. Co-expression of α subunits with β or γ subunits modestly increases kinase activity accompanied by the formation of α/β or α/γ heterodimers. In addition to binding of each noncatalytic subunit to the α subunit, β and γ subunits bind to each other, possibly resulting in a more stable heterotrimeric complex [39]. The increase in kinase activity associated with expression of this heterotrimer is due both to an increase in enzyme-specific activity (units/enzyme mass) and to an apparent enhanced α subunit expression. Co-expression of the noncatalytic β or γ subunits is required for optimal activity of the α catalytic subunits. This may explain the possible and/or necessary co-expression of α with β and γ subunits. In the same way, we can explain Rule 10, which represents that if the expression of γ_1 and γ_3 are significantly increased, the expression of β_1, β_2 and γ_2 will increase too.

Rule 11 does not make sense in biology because there is no relationship between the phosphorylation of α_1 and α_2. Rule 12 shows the protein expression of α_1 is co-expressed with the expression of β_2. This may imply that the tight relation between γ_3 and α_2, γ_3 and β_2, γ_1 and both α_1 and α_2, and γ_1 and β_2. The new experimental results in [23,24] defend these rules. There are 12 theoretically possible AMPK heterotrimetric complexes. But in human skeletal muscle, there are only three detectable combinations exist $\alpha_2\beta_2\gamma_1 >> \alpha_2\beta_2\gamma_3 = \alpha_1\beta_2\gamma_1$ [24]. Our results predict that there maybe $\alpha_1\beta_1\gamma_1$ complex in human skeletal muscle as well.

Furthermore, we can predict some latent associations from the derived association rules. For example, if the training is not of high-intensity, we can predict deductively α_1 activity in terms of rule 1 in Table 6. Besides, the newly found association rules can be used in the design of future experiments. For example, α_{2a} was not highly activated at rest status. If we want to activate it, we can regulate the intensity or duration of exercise as indicated in rules 2, 3 and 4 in Table 6. Therefore, the identified association rules may play important roles in three aspects:

- demonstrating former experiments via matching more experimental results;

- predicting potential results based on existing conditions; and

- guiding the design of future experiments.

From our experiments, we demonstrated that association rules can not only discover important biological patterns but also be used to reduce the cost of labor, resources and other associated activities. Experiments can be conducted based on the derived association rules, thereby reducing the number of experiments. For example, if α_2 activity is increased with exercise in one experiment, we can predict that α_{1a} possibly has no change under the same condition based on rules 1 and 2 in Table 6. Similarly, if α_{1a} is highly increased with high intensity training, we can predict that α_{2a} is possibly highly increased either in light of rule 6. Therefore, it can save the experimental time by avoiding extra (unnecessary) stimuli. For example, rule 4 in Table 6 implies that *high-intensity* training can result in high expression of α_{2a}. If we want to observe that α_{2a} is highly expressed in experiments, we can purposefully handle *high-intensity* stimuli rather than *intense-intensity*. Consequently, the rules are beneficial to understand the signalling pathway of AMPK in regulating metabolism and its potential benefits to disease treatment.

Discussion

In this study, we have proposed a framework by which to identify association rules of interest, either between the state of isoforms of α, β and γ subunits of AMPK, or between stimulus factors and the state of isoforms, from AMPK regulation data. In particular, the item constraints are applied in the closed interpretation to the itemset generation. We have shown how to specify a threshold in terms of the amount of results instead of a fixed threshold for all itemsets.

Our approaches start with collecting hidden data from publications in Medline. The collected experimental data is qualitative and does not correspond with the data mining softwares [19]. Besides, we have shown that untested items in some experiments may result in many unrelated or even inaccurate rules. If the untested items are not pruned, it seemed to cause many inaccurate results, and the implementation of software became less efficient when identifying frequent patterns [8]. Consequently, we marked the untested items using the filtering symbols, which facilitate the pruning of frequent itemsets and avoid the generation of irrelevant frequent itemsets. To meet the criteria of softwares in [19], it is needed to conduct data preprocessing and convert the qualitative items into the form of quantitative items (non-negative integers), which benefit to the execution of software.

The traditional association rule mining typically requires a user to specify a minimum support threshold for the generation of all itemsets. It has been argued that without specific knowledge, it is difficult for users to obtain an optimal support threshold. It was showed that a better way is to identify top-*N* interesting itemsets, instead of specifying a fixed threshold value for all itemsets of all sizes like [12,13,16]. However, a major problem is that not all top-*N* itemsets are interesting. Some correlations

might not make sense biologically at all. The item constraints in the closed interpretation are applied to the itemset generation. For example, in AMPK regulation, the relationships between the activity of *subunit isoforms* and *stimulus factors* such as intensity and duration are interesting but the relationships between *stimulus factors* are not. It provides a flexible way for users to specify a set of constraints and allow them to search and control the interesting frequent patterns.

In our selection of association rules, we adopt 0.7 as the minimum confidence because there is no sharp drop in rule output in contrast with 0.6. Although this helps us focus on the interpretation of significant rules, some rules that have a close confidence to 0.7 might be ignored. Extending the interpretation to those less important rules is therefore desirable, but it would require more computational resources and collaboration with biologists.

Bayesian network, a graphic model, has been widely used to identify metabolic pathways and construct genetic networks due to its statistical significance and inherited advantage in handling the information with uncertainty. Nevertheless, as an assumption-driven method, it relies on the quality and extent of the prior beliefs and is only as useful as this prior knowledge is reliable. However, the hidden and insufficient AMPK regulation data in the publications of Medline prevents us from obtaining reliable prior knowledge. Furthermore, some potential AMPK pathways are undetermined and might be ignored from the assumption. Considering the above difficulties, the authors turn to data mining, a data-driven method, in this study.

We have focused our attention here on human skeletal muscle. Also, our methods are eligible for other organisms, such as mouse, where the experimental indexes or criteria may be quite different. A modification with respect to data preprocessing may be adopted according to the criteria. Another interesting question is whether our methods can be used to explore the AMPK regulation on other tissues and organs, such as adipose tissue, heart and liver [4]. Intuitively, it is necessary to classify the data into different groups because different compositions may play a similar role on different tissues or organs. On the other hand, studying the potential metabolic pathways between AMPK regulations on different tissues or organs should be useful in disease prediction, diagnosis and treatment. We plan to seek answers for these questions in our future work. The results can enable biologists to understand AMPK pathways and extend it to the other kinases to form a full kinase interaction mapping.

Conclusion

AMPK has emerged as an important energy sensor in the regulation of cell metabolism. Recent experiments reveal that physical exercises are closely linked with AMPK activation in skeletal muscle. This paper proposes a framework by which to identify association rules of interest from the published experimental data.

Unlike the conventional methods, the measurement of items from AMPK regulation data is taken into account. In addition, the items that have low occurrence in existing experiments are pruned prior to mining. Furthermore, we apply item constraints in the closed interpretation to the itemset generation so that a threshold is specified on the amount of results rather than a fixed threshold, thereby reducing the search space vastly.

Our framework was demonstrated by mining realistic AMPK regulation data set with respect to skeletal muscle. Many of the found rules make sense biologically, others suggesting new hypotheses that may warrant further investigation. Particularly, some of them were unknown previously. Moreover, they help us understand the characteristics of AMPK and relevant disease treatment. It is thus promising in the interpretation of AMPK regulation data.

Authors' contributions

QFC conceived of and carried out the AMPK regulation studies, applied the data mining method to AMPK data sets, interpreted biologically the experimental results and wrote the manuscript. YPPC supervised the project and suggested ways of improving the study, and participated in writing the manuscript. Both authors read and approved the final version of the manuscript.

Additional material

Additional file 1

An implementation of the FP-tree-based algorithm. The files consist of the AMPK regulation data with respect to the human skeletal muscle, and the applied item constraints in this study.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-7-394-S1.zip]

Additional file 2

Our results from mining AMPK regulation data set regarding human skeletal muscle. The results provided present the interesting frequent patterns with respect to the AMPK pathways.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-7-394-S2.zip]

Acknowledgements

The work reported in this paper was partially supported by the Australian Research Council's Discovery Project Grant DP0559251 and ARC Grant LX0560616. Authors would like to thanks Prof. Limsoon Wong (National University of Singapore), Dr. ZP Chen (Protein Chemistry & Metabolism Unit, St Vincent's Institute), anonymous reviewers and associate editor for their time, input and useful feedback to increase the quality of this paper.

References

1. Musi N, Fujii N, Hirshman MF, Ekberg I, Froberg S, Ljungqvist O, Thorell A, Goodyear LJ: **AMP-activated protein kinase (AMPK) is activated in muscle of subjects with type 2 diabetes during exercise.** *Diabetes* 2001, **50(5):**921-927.
2. Carlson D, Kim KH: **Regulation of Hepatic Acetyl Coenzyme A Carboxylase by Phosphorylation and Dephosphorylation.** *Journal of Bio Chem* 1973, **248(1):**378-380.
3. Kemp BE, Mitchelhill KI, Stapleton D, Michell BJ, Chen ZP, Witters LA: **Dealing with energy demand: The AMP-activated protein kinase.** *Trends Biochem Sci* 1999, **24(1):**22-25.
4. Hardie DG: **AMP-activated protein kinase: the guardian of cardiac energy status.** *J Clin Invest* 2004, **114:**465-468.
5. Musi N, Goodyear LJ: **AMP-activated protein kinase and muscle glucose uptake.** *Acta Physiologica* 2003, **178(4):**337-345.
6. Altschul ST, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of Molecular Biology* 1990, **215(3):**403-410.
7. Durbin R, Eddy S, Krogh A, Mitchison G: **Biological sequence analysis: probabilistic models of proteins and nucleic acids.** Cambridge University Press; 1998.
8. Zhang CQ, Zhang SC: **Association Rule Mining: Models and Algorithms.** LNAI 2307, Springer-Verlag; 2002.
9. Doddi S, Marathe A, Ravi SS, Torney DC: **Discovery of Association Rules in Medical Data.** *Med Inform Internet Med* 2001, **26(1):**25-33.
10. S , Bamidis PD, Maglaveras N, Pappas C: **Mining Association Rules from Clinical Databases: An Intelligent Diagnostic Process in Healthcare.** *Medinfo* 2001, **10(2):**1399-1403.
11. Chen YP Ed: **Bioinformatics Technologies.** Springer SCI; 2005:396.
12. Yin_Ling Cheung , Ada Wai-Chee Fu : **Mining Frequent Itemsets without Support Threshold: With and Without Item Constraints.** *IEEE Transaction on Knowledge and Data Engineering* 2004, **16(9):**1052-1069.
13. Han J, Pei J, Yin Y: **Mining frequent patterns without candidate generation.** *Proceedings of the ACM SIGMOD International Conference on Management of Data* 2000:1-12.
14. Agrawal R, Imielinshki T, Swami A: **Mining Association Rules between Sets of Items in Large Databases.** *Proceeding of ACM-SIGMOD International Conference on Management of Data* 1993:207-216.
15. Durante PE, Mustard KJ, Park SH, Winder WW, Hardie DG: **Effects of Endurance Training on Activity and Expression of AMP-activated Protein Kinase Isoforms in Rat Muscles.** *Am J Physiol Endocrinol Metab* 2002, **283(1):**178-186.
16. Wang K, He Y, Han J: **Pushing Support Constraints into Association Rules Mining.** *IEEE Transaction on Knowledge and Data Engineering* 2003, **15(3):**642-658.
17. Gong G, Tan Kian-Lee, Tung KH, Xu X: **Mining top-K covering rule groups for gene expression data.** *Proceedings of the 2005 ACM SIGMOD international conference on Management of data* 2005:670-681.
18. Chen QF: 2006 [http://www.deakin.edu.au/~qifengch/ampk/skele_muscle/ampktab.xls].
19. Gösta Grahne and Zhu J: **Efficiently Using Prefix-trees in Mining Frequent Itemsets.** *Proceeding of the First IEEE ICDM Workshop on Frequent Itemset Mining Implementations FIMI'03) Melbourne* 2003.
20. Zaki M, Gouda K: **Fast vertical mining using diffsets.** In *Proceedings of ACM SIGKDD'03* Washington, DC:316-325.
21. Burdick D, Calimlim M, Gehrke J: **MAFIA: A maximal frequent itemset algorithm for transactional databases.** *Proceedings of ICDE'01* :443-452.
22. Hardie DG, Scott JW, Pan DA, Hudson ER: **Management of cellular energy by the AMP-activated protein kinase system.** *FEBS Lett* 2003, **546(1):**113-120.
23. Barnes BR, Zierath JR: **Role of AMP-Activated Protein Kinase in the Control of Glucose Homeostasis.** *Current Molecular Medicine* 2005, **5(3):**341-348.
24. Eijnde BO, Derave W, Wojtaszewski JF, Richter EA, Hespel P: **AMP kinase expression and activity in human skeletal muscle: effects of immobilization, retraining, and creatine supplementation.** *J Appl Physiol* 2005, **98(4):**1228-1233.
25. McConell GK, Lee-Young RS, Chen ZP, Septo NK, Huynh NN, Stephens TJ, Canny BJ, Kemp BE: **Short-term exercise training in humans reduces AMPK signalling during prolonged exercise independent of muscle glycogen.** *J Physiol* 2005, **568(2):**665-676.
26. Frosig C, Jorgensen SB, Hardie DG, Richter EA, Wojtaszewski JF: **5' -AMP-activated protein kinase activity and protein expressed are regulated by endurance training in human skeletal muscle.** *Physitol Endocrinol Metab* 2004, **286(3):**411-417.
27. Chen ZP, McConell GK, Michell BJ, Snow RJ, Canny BJ, Kemp BE: **AMPK signaling in contracting human skeletal muscle: acetyl-CoA carboxylase and NO synthase phosphorylation.** *Physiol Endocrinol Metab* 2000, **279(5):**1202-1206.
28. Yu M, Stepto NK, Chibalin AV, Fryer LG, Carling D, Krook A, Hawley JA, Zierath JR: **Metabolic and Mitogenic Signal Transduction in Human Skeletal Muscle after Intense Cycling Exercise.** *J Physiol* 2003, **546(2):**327-335.
29. Wojtaszewski JF, Nielsen P, Hansen BF, Richter EA, Kiens B: **Isoform-specific and Exercise Intensity-dependent Activation of 5'-AMP-activated Protein Kinase in Human Skeletal Muscle.** *J Physiol* 2000, **528(1):**221-226.
30. Nielsen JN, Mustard KJ, Graham DA, Yu H, MacDonald CS, Pilegaard H, Goodyear LJ, Hardie DG, Richter EA, Wojtaszewsji JF: **5' -AMP-activated protein kinase activity and subunit expression in exercise-trained human skeletal muscle.** *Appl Physiol* 2003, **94(2):**631-641.
31. Fujii N, Hayashi T, Hirshman MF, Smith JT, Habinowski SA, Kaijser L, Mu J, Ljungqvist O, Birnbaum MJ, Witters LA, Thorell A, Goodyear LJ: **Exercise Induces Isoform-specific Increase in 5'AMP-activated Protein Kinase Activity in Human Skeletal Muscle.** *Biochem Biophys Res Commun* 2000, **273(3):**1150-1155.
32. Wadley GD, Lee-Young RS, Canny BJ, Wasuntarawat C, Chen ZP, Hargreaves M, Kemp BE, McConell GK: **Effect of exercise intensity and hypoxia on skeletal muscle AMPK signaling and substrate metabolism in humans.** *Am J Physiol Endocrinol Metab* 2005, **290(4):**694-702.
33. Jørgensen SB, Viollet B, Andreelli F, Frøsig C, Birk JB, Schjerling P, Vaulont S, Richter EA, Wojtaszewski JFP: **Knockout of the α_2 but not α_1 5'-AMP-activated protein kinase isoform abolishes 5-Aminoimidazole-4-carboxamide-1-β-4-ribofuranoside but not contraction-induced glucose uptake in skeletal muscle.** *J Biol Chem* 2004, **279(2):**1070-1079.
34. Kemp BE, Stapleton D, Campbell DJ, Chen ZP, Murthy S, Walter M, Gupta A, Adams JJ, Katsis F, van Denderen B, Jennings IG, Iseli T, Michell BJ, Witters LA: **AMP-activated protein kinase, super metabolic regulator.** *Biochem Soc Trans* 2003, **31:**162-168.
35. Clark SA, Chen ZP, Murphy KT, Aughey RJ, McKenna MJ, Kemp BE, Hawley JA: **Intensified exercise training does not alter AMPK signaling in human skeletal muscle.** *Physitol Endocrinol Metab* 2003, **286(5):**737-743.
36. Roepstorff C, Vistisen B, Donsmark M, Nielsen JN, Calbo H, Green KA, Hardie DG, Wojtaszewski JF, Richter EA, Kiens B: **Regulation of hormone sensitive lipase activity and Ser563 and Ser565 phosphorylation in human skeletal muscle during exercise.** *Physiology* 2004, **560(2):**551-562.
37. Coffey VG, Zhong Z, Shield A, Canny BJ, Chibalin AV, Zierath JR, Hawley JA: **Early signaling responses to divergent exercise stimuli in skeletal muscle from well-trained humans.** *FASEB Journal* 2006, **20(1):**190-192.
38. Hurst D, Taylor EB, Cline TD, Greenwood LJ, Compton CL, Lamb JD, Winder WW: **AMP-activated protein kinase kinase activity and phosphorylation of AMP-activated protein kinase in contracting muscle of sedentary and endurance-trained rats.** *Am J Physiol Endocrinol Metab* 2005, **289(4):**710-715.
39. Dyck JRB, Gao G, Widmer J, Stapleton D, Fernandez CS, Kemp BE, Witters LA: **Regulation of 5' -AMP-activated Protein Kinase**

Activity by the Noncatalytic β and γ Subunits. *J Biol Chem* 1996, **271(30):**17798-17803.