International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:8, 2007

A Probability based Pair Extension Method in Protein 2-DE Gel Image Analysis

Yanhua Jin, and Won Suk Lee

***Abstract*—**The two-dimensional gel electrophoresis method (2-DE) is widely used in Proteomics to separate thousands of proteins in a sample. By comparing the protein expression levels of proteins in a normal sample with those in a diseased one, it is possible to identify a meaningful set of marker proteins for the targeted disease. The major shortcomings of this approach involve inherent noises and irregular geometric distortions of spots observed in 2-DE images. Various experimental conditions can be the major causes of these problems. In the protein analysis of samples, these problems eventually lead to incorrect conclusions. In order to minimize the influence of these problems, this paper proposes a partition based pair extension method that performs spot-matching on a set of gel images multiple times and segregates more reliable mapping results which can improve the accuracy of gel image analysis. The improved accuracy of the proposed method is analyzed through various experiments on real 2-DE images of human liver tissues.

***Keywords*—**Proteomics, spot-matching, two-dimensional electrophoresis.

I. Introduction

PROTEOMICS is one of the most rapidly expanding fields currently expanding as well as developing in the field of biology. The main issues of proteomics studies focus on which proteins work inside a sample such as a cell, a tissue or a biological system, and how they interact with other under specific conditions. Expression proteomics aims to analyze the anomalous differences between the protein actions of a diseased sample and those of a healthy one. In order to separate the proteins of a sample, two analytical methods, namely 2-DE and non-2-DE, are predominately used. The former employs a two-dimensional electrophoresis technique while the latter mainly employs liquid chromatography-mass spectrometry (LC-MS) or specific affinity tags such as an isotope-coded affinity tag (ICAT) [1] as well as a mass-coded abundance tag (MCAT) [2]. Although the non-2-DE method is an emerging technique that can automate the process of protein separation, the 2-DE method is much more popular since it can accomplish the task in a more cost-effective way [3,4].

For a given sample, the results of the 2DE method are represented by a two-dimensional gel image. A 2-DE gel image may include hundreds of spots. Each spot corresponds to a protein inside its sample and has its own unique properties based on its position, shape and optical attitude within the image. Each spot has a unique identification number *spotID*. The 2-DE method first separates the proteins of a tissue based on their isoelectric points (pI) and then performs the second separation process based on their electric charges which are proportional to their molecular weights.

Considering the number of spots within a typical gel image, it is impossible to analyze the entire number of proteins of a gel image manually. Consequently, most biologists employ commercial software packages such as Melanie, Progenesis or PDQuest [5,6] for this purpose. These packages provide two major operations: *spot detection* and *spot-matching*. A spot-detection operation detects individual spots in a gel image by employing intensive image processing techniques like Laplacian, Gaussian and smooth-by-diffusion.

The purpose of a spot-matching operation performed on two different gel images is to associate the spots of one image to those of the other in order to link the same protein represented by the two gel images. In a 2-DE spot-matching operation, the one whose spots are more clearly separated is designated as a *reference gel image*. The other image is called as a *target gel image*. By performing a spot matching operation, a target gel image spot is pair-wisely associated with a reference gel image spot if the properties of these two spots are similar enough to be the same protein. The two pair-wisely associated spots are considered to be the same kind of protein and further analysis is continued based on the previously identified facts.

The accuracy of a spot-matching operation is totally dependent upon the expertise of a user to designate a single a reference gel image [7]. The matching result is greatly influenced by the quality of the reference gel image. The software package Progenesis developed by Nonlinear Dynamics provides another way to choose a reference gel image by statistically integrating the properties of several gel images [4]. Such a gel image is called a *virtual average gel image*. However, a virtual average gel image may lose important properties possessed by an individual gel image since the properties of the virtual average gel image are merely statistically averaged.

In order to solve the problems mentioned above, we propose

Manuscript received October 15, 2001. This work was supported in part by the Korea Science and Engineering Foundation(KOSEF) through the National Research Lab. Program funded by the Ministry of Science and Technology (No.R0A-2006-000-10225-0) and the MIC(Ministry of Information and Communication), Korea, Under the ITFSIP(IT Foreign Specialist Inviting Program) supervised by the IITA(Institute of Information Technology Assessment)

Yanhua Jin is with the Computer Science, Yonsei University, Shinchon-Dong, Seodaemun-Gu, Seoul, Korea (phone: 82-2-2123-2716; fax: 82-2-365-2579; e-mail: yienah@database.yonsei.ac.kr).

Won Suk Lee is with the Computer Science, Yonsei University, Shinchon-Dong, Seodaemun-Gu, Seoul, Korea (phone: 82-2-2123-2716; fax: 82-2-365-2579; e-mail: leewo@database.yonsei.ac.kr).

a probability based pair extension method based on the pairing results of multiple spot matching operations. More precisely, given a set of gel images, a spot-matching operation can be repeatedly performed for more than one reference gel image in order to cope with the noises as well as irregular geometric distortions of spots in each gel image. Each spot-matching operation produces its own pairs for each distinct protein with respect to its underlying reference gel image. If a gel image spot is more frequently associated with a specific protein, the possibility that the spot is truly corresponding to the protein increases. Consequently, the proposed method can eliminate possible mismatching of a single spot-matching operation and provide better reliability for further protein quantitative analysis represented within gel images.

The rest of this paper is organized as follows: Section 2 describes the problem in detail. Section 3 presents the proposed probability based pair extension method to improve the reliability of 2-DE protein analysis. In Section 4, the accuracy of the proposed method is analyzed through a series of experiments. The conclusions are discussed in Section 5.

II. Problem Definition

To denote a gel image and its spots in this paper, the notations in Table I are used. Given a set of n gel images $G = \{g_1, g_2, ..., g_n\}$, let a gel image $g_r \in G$ be a designated reference gel image and $g_i \in G (r \neq i)$ as a target gel image. If a spot s_r^x of g_r, denoted by $s_r^x \propto g_r$, is pair-wisely associated with a spot s_i^y of $g_i (r \neq i)$, there is a pair-wise association between these two spots and the set of two spots, i.e., (s_r^x, s_i^y), is called as a pair. Let $PA(g_r, g_j)(r < j)$ denote the set of all pairs between two gel images g_r and g_j, i.e., $(s_r^x, s_i^y) \in PA(g_r, g_j)$. Given two target gel images g_i and g_j, a spot $s_i^x \propto g_i$ is pair-wisely associated with a spot $s_j^y \propto g_j$, i.e., $(s_i^x, s_j^y) \in PA(g_i, g_j)$, if these two spots are pair-wisely associated with the same spot s_r^z of the reference gel image, i.e., $(s_r^z, s_i^x) \in PA(g_r, g_i)$ and $(s_r^z, s_j^y) \in PA(g_r, g_j)$. Fig. 1 shows the pair-wise associations produced by a spot-matching operation between the reference gel image g_r and each of *n-1* target gel images G-$\{g_r\}$. Given a reference gel image g_r, the set of all pair-wise associations found in G can be represented by a set of pairs $\Phi(g_r) = \bigcup_{k=1}^{k=n} PA(g_r, g_k) \cup \bigcup_{i=1}^{i=n-1} \bigcup_{j=2}^{j=n} PA(g_i, g_j)$ $(r \neq k, r \neq i \neq j)$. This set $\Phi(g_r)$ is called a pair transaction which is an atomic unit of pair information obtained by the same reference gel image g_r. Since a spot of a gel image is supposed to represent a distinct protein, every spot of the reference gel image g_r can only be pair-wisely associated with at most one spot of every target gel image in G-$\{g_r\}$. Consequently, all of the pairs in $\Phi(g_r)$ are one-to-one relationships between the spots of gel images in G as stated in the Property 1.

TABLE I
Notations of Gel Images and Spots

Notation	Meaning
n	Total number of gel images
G	A set of 2-DE gel images
g_i	The i^{th} gel image, $G = \{g_1, g_2, ..., g_n\}$
s_i^x	A x^{th} spot in the gel image g_i, $g_i = \{s_i^x, s_i^y, ..., s_i^z\}$
R	A set of reference gel images $R = \{r_1, r_2, ..., r_m\}$, $R \subset G$

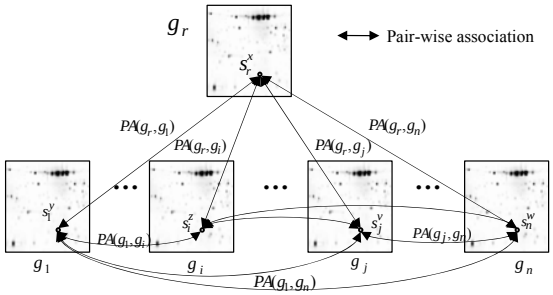


Fig. 1 Pair-wise associations between two spots

[Property 1] Uniqueness of a pair

For a pair $(s_i^x, s_j^y) \in PA(g_i, g_j)$ in $\Phi(g_r)$, none of these two spots should have a pair-wise association with any other spot s_j^v or $s_i^w (x \neq v, y \neq w)$, i.e., $(s_i^x, s_j^v) \notin PA(g_i, g_j)$ or $(s_i^w, s_j^y) \notin PA(g_i, g_j)$ in $\Phi(g_r)$. □

The set of spots that represent the same protein is called a protein class defined in Definition 1.

[Definition 1] Protein class $p(s_r^x)$

Given the set of all pairs $\Phi(g_r)$ identified by a reference gel image g_r, a protein class $p(s_r^x)$ is defined by a set of spots each of which has a pair-wise association with the spot s_r^x as follows: $p(s_r^x) = \{s_r^x\} \cup \{s_i^y \mid (s_r^x, s_i^y) \in \Phi(g_r), 1 \leq \forall i \leq n, r \neq i\}$. □

A protein class is a set of spots representing the same protein. Let $PS(g_r)$ denote the set of protein classes that are found by a reference image g_r, i.e., $PS(g_r) = \{p(s_r^x) \mid \forall s_r^x \in g_r\}$. The two different protein classes $p(s_r^x)$ and $p(s_r^y)$ found by a spot matching operation should be disjoint, i.e., $p(s_r^x) \cap p(s_r^y) = \phi$. The maximal cardinality of a protein class is the number of gel images in G. Let $PS(g_r)$ denote the set of protein classes that are found by a reference image g_r, i.e., $PS(g_r) = \{p(s_r^x) \mid \forall s_r^x \in g_r\}$. Since each protein class should contain only one spot of the reference image, the maximum number of protein classes in $PS(g_r)$ is the number of spots in the reference gel image g_r. Furthermore, due to the characteristics of spot matching operation, a protein class should satisfy the

completeness property of a protein class. When $\left|p(s_r^x)\right| = k$, there are *(k-1)* pairs between the spot s_r^x and a spot in each of *(k-1)* target gel images. In addition, these *(k-1)* spots in each of the *(k-1)* target gel images should also be pair-wisely associated with each other, so that there are *(k-1)*C_2 pairs. Therefore, the total number of all pairs for the protein class $p(s_r^x)$ should be *(k-1)+(k-1)*C_2*=k*C_2*=k(k-1)/2*.

In order to minimize the influence of the noises and geometric distortions of spots in spot-matching operation, gel images of good quality can be designated as reference gel images. We consider the number of spots in a gel image to be an objective quality measure for selecting a reference gel image to eliminate the subjectivity. A spot-matching operation is repeatedly performed with respect to each of selected reference gel images. The set of pairs $\Phi(g_r)$ obtained by a reference gel image g_r can be regarded as a semantically atomic unit of information. Therefore, it is also called a pair transaction.

Given a set of gel images *G*, suppose a set of gel images $R = \{r_1, r_2 ... r_m\} (r_k \in G, 1 \le k \le n, R \subseteq G)$ is chosen to be reference gel images. Let a pair database *D* be all of m pair transactions, i.e., $D = \{\Phi(r_1), \Phi(r_2) ..., \Phi(r_m)\}$. Since the noises and geometric distortions of each reference gel image are different, the set of pairs produced by one reference gel image is not the same as that produced by another gel image. Therefore, the uniqueness property of a pair is no longer valid. A pair (s_i^x, s_j^y) is a *pure pair* if it satisfies the uniqueness of a pair. In other words, the pair (s_i^x, s_j^y) is the only pair concerning the two spots s_i^x and s_j^y. Otherwise, it is called a *contradicting pair*. Two contradicting pairs are related if they share the same spot from the same gel image. A set of related contradicting pairs grouped together is defined to be a contradicting pair set. As a result, all of identified contradicting pairs are split into a number of contradicting pair sets. According to the uniqueness property of a pair, a spot of one gel image should be pair-wisely associated with only one spot of another gel image regardless of a reference gel image. However, a contradicting pair set occurs when not all of *m* spot-matching operations agree on a certain pair. The support of a pair is defined by the fraction of the number of pair transactions that include the pair. Only those contradicting pairs whose supports are high enough and satisfy the uniqueness property are extracted to find a more reliable set of protein classes.

This problem basically turns out to be similar to the problem of finding frequent itemsets in the descriptive data mining although additional constraints should be checked. However, most algorithms for finding frequent itemsets are optimized to a data set with a large number of transactions occurred by a relatively small number of items. On the contrary, a database of pair transactions discussed in this paper has a huge number of items, i.e., pairs, with a small number of pair transactions. Due to these reasons, the conventional algorithms [8,9,10] are not efficient to be employed.

III. Probability Based Pair Extension Algorithm

The problem of finding frequent itemsets can be stated as follows: Given a set of items $I = \{i_1, i_2, ..., i_n\}$, let *D* denote a set of transactions. Each transaction *T* is a subset of *I*, i.e., *T*□*I*. A set of items is defined as a frequent itemset if the ratio of the number of transactions containing the set of items over the total number of transactions in *D* is greater than or equal to a user-specified minimum support S_{min}. An itemset composed of *n* items is called an *n*-itemset. The terms in the conventional definition of finding frequent itemsets can be mapped to their counterparts as follows: A pair database *D* can be regarded as a set of pair transactions $\Theta = \{\Phi(r_1), \Phi(r_1) ..., \Phi(r_m)\}$. A pair transaction can be regarded to be a transaction *T*. A distinct pair can be regarded as an item, so that the set of all distinct pairs can be regarded as the set of items *I*. An *n*-pair set is a set of *n* pairs each of which shares at lease one common spot with another pair of the *n* pairs. For example, two pairs $(s_i^x, s_j^y), (s_j^y, s_r^z)$ are a 2-pair set while $(s_i^x, s_j^y), (s_j^w, s_r^z)$ is not. A frequent *n*-pair set is a set of *n* pairs whose supports are greater than or equal to a S_{min}. In a pair database, not all of the frequent *n*-pair sets are useful. As mentioned above, a protein class of length *k* should contain exactly *k(k-1)/2* pairs in a pair database *D*. Therefore, the completeness property should be checked in order to represent a protein class by an *n*-pair set. Such an *n*-pair set satisfying the completeness property is called as a complete *n*-pair set. Therefore, the distinct spots of a frequent complete *n*-pair set are a true protein class with respect to the given value of S_{min}. A complete *n*-pair set can be explained more clearly when a pair database is regarded as a graph. A distinct spot of an *n*-pair set is regarded as a node and a pair-wise association between its two spots is considered as an edge between the corresponding nodes of the two spots. A complete *n*-pair set in a pair database is corresponding to a clique which is a complete graph whose edges connect every pair of nodes.

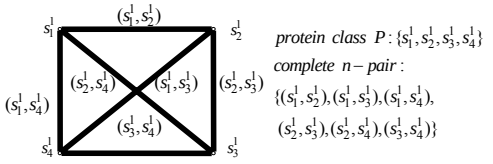


Fig. 2 An example of complete *n*-pair set

Fig. 2 shows an example of a complete *n*-pair set which is corresponding to a protein class. For a protein class $P = \{s_1^1, s_2^1, s_3^1, s_4^1\}$ of 4 spots in the figure, when each of its spots is regarded as a node, there should be six edges in order to be a complete graph. These edges are corresponding to (s_1^1, s_2^1), (s_1^1, s_3^1), (s_1^1, s_4^1), (s_2^1, s_3^1), (s_2^1, s_4^1) and (s_3^1, s_4^1). However, a set of two pairs $Q = \{(s_1^1, s_2^1), (s_2^1, s_3^1)\}$ is not a complete 2-pair set since *Q* has three distinct spots but it does not contain *3(3-1)/2=3* pairs.

Input : parameters: δ_{min}, pair database D

```
two − phase pair − wise_extension(D, s_min){
1)  F = find Frequent pairs(D, S_min);
2)  P = φ; PC = φ;
3)  for each pair p_i ∈ F
4)      P = partition(F, P, p_i);
5)  PC = PC ∪ pw_extension(P); }
6)  partition(F, P, p_i){
7)    P = P ∪ {p_i};
8)      for each pair p_j
9)        if p_i ∩ p_j ≠ φ
10)         P = P ∪ {p_j};
11)     F = P − F;
12) return P; }
13) pw_extension(P){
14)   FP = φ;
15)   for each pair in p = (s_i^x, s_j^z) ∈ P, do
16)     pc = {s_i^x, s_j^z};
17)     S = {s_1^{z_1}, ..s_k^{z_k}., s_j^z} where(s_i^x, s_k^{z_k}) ∈ P
18)       for all p' = {s_i^x, s_l^{z_l}} ∈ P
19)         for each s_q^{z_q} ∈ pc
20)           if (s_l^{z_l}, s_q^{z_q}) ∈ P, pc+ = {s_l^{z_l}}
21)           else break;
22)   if pc ∉ FP, FP+ = {pc}
23) return FP; }
```

Fig. 3 Pair-wise extension algorithm

Unlike the conventional problem of finding frequent itemsets, an item in a pair database is a pair of two spots representing the same protein, which imposes an additional constraint. As mentioned above, a spot s_i^x in a protein class of k spots should have *(k-1)* pair-wise associations with the other *(k-1)* spots. Therefore, these *(k-1)* pairs must share the same common spot s_i^x in its pair transaction $\Phi(g_r)$. Obviously, another constraint is the uniqueness property of a pair mentioned in Property 1. These constraints make the problem be more difficult than the conventional problem of finding frequent itemset.

The proposed probability based pair extension algorithm is illustrated in Fig. 3. Its main idea is to partition the set of all frequent pairs into a number of smaller disjoint sets according to their pairing relationships among spots, so that the search space of extending protein classes can be reduced. Two distinct protein classes should not share any common spot. Therefore, this mutually exclusive property of two distinct protein classes makes it possible to partition the set of pairs. In the first phase, all the contradicting pairs are deleted and frequent pairs are found (line 1). Consequently, these frequent pairs are divided into a number of disjoint partitions by the sub-routine *partition()* based on their pairing relationships among the spots of them (line 6-11). Only the pairs of an individual partition can possibly form a complete n-pair set. Given a minimum support S_{min}=0.4, Fig. 7 illustrates how the *partition()* procedure is performed. In a pair database D in Fig. 4-(a), all of its frequent pairs are found as shown in Fig. 4-(b). These pairs satisfy the uniqueness property and their supports are greater than or equal to S_{min}=0.4 respectively. In Fig. 4-(c), the frequent pairs $(s_1^1,s_2^1),(s_1^1,s_3^1),(s_2^1,s_3^1)$ are considered as the first partition *Part[1]* since each of them share at least one common spot with another pair. For example, (s_1^1,s_2^1) and (s_1^1,s_3^1) share spot s_1^1; (s_1^1,s_2^1) and (s_2^1,s_3^1) share spot s_2^1. Similarly, the pairs $(s_4^2,s_5^2),(s_4^2,s_6^2),(s_5^2,s_6^2)$ are formed as the second partition *Part[2]*.

$\Phi(g_1)=\{(s_1^1,s_2^1),(s_1^1,s_3^1),(s_2^1,s_3^1)\}$	(s_1^1,s_2^1)
$\Phi(g_2)=\{(s_1^1,s_2^1),(s_1^1,s_3^1),(s_2^1,s_3^1)\}$	(s_1^1,s_3^1)
$\Phi(g_3)=\{(s_1^1,s_2^1),(s_1^1,s_3^1),(s_2^1,s_3^3)\}$	(s_2^1,s_3^1)
$\Phi(g_4)=\{(s_4^2,s_5^2),(s_4^2,s_6^2),(s_5^2,s_6^2)\}$	(s_4^2,s_5^2)
$\Phi(g_5)=\{(s_4^2,s_5^2),(s_4^2,s_6^2),(s_5^2,s_6^2)\}$	(s_4^2,s_6^2)
	(s_5^2,s_6^2)

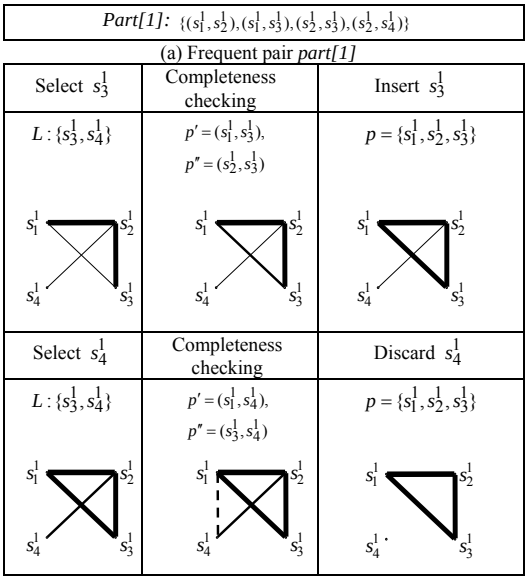
(a) A pair database D (b) Frequent pairs

Part[1]	Part2
(s_1^1,s_2^1)	(s_4^2,s_5^2)
(s_1^1,s_3^1)	(s_4^2,s_6^2)
(s_2^1,s_3^1)	(s_5^2,s_6^2)

(c) *partition()* process

Fig. 4 An example of *paritition()*

In the second phase, the sub-routine *pw_extension()* is independently performed for each partition to find out a set of frequent complete n-pair sets each of which is corresponding to a protein class (line 13-23). For the *part[1]* in Fig. 5-(a), pair (s_1^1,s_2^1) is considered as an initial protein class $pc=\{s_1^1,s_2^1\}$ which is extended further. Since the frequent pairs (s_1^1,s_3^1), (s_2^1,s_3^1) and (s_2^1,s_4^1) share the spots s_1^1 and s_2^1 of pc respectively, their remaining spots s_3^1 and s_4^1 can be candidates for the extension of pc. Therefore, the two spots form the extendable spot set L. In order to extend one more spot to a protein class, $|pc|$ additional pairs are required to satisfy the completeness property. The spot s_3^1 is added to the protein class $pc=\{s_1^1,s_2^1\}$ since the pairs $(s_1^1,s_3^1),(s_2^1,s_3^1)$ are all frequent pairs in the pair database. Therefore, the protein class is extended to $pc=\{s_1^1,s_2^1,s_3^1\}$. However, for the spot s_4^1, since both of its pairs (s_1^1,s_4^1) and (s_3^1,s_4^1) are not frequent, it cannot be added to pc for the further extension. When all of the extendable spots are checked, the protein class becomes $pc=\{s_1^1,s_2^1,s_3^1\}$. Therefore, when all of these two partitions are considered, two protein classes $\{s_1^1,s_2^1,s_3^1\}$ and $\{s_4^2,s_5^2,s_6^2\}$ are finally formed.



(a) Frequent pair *part[1]*

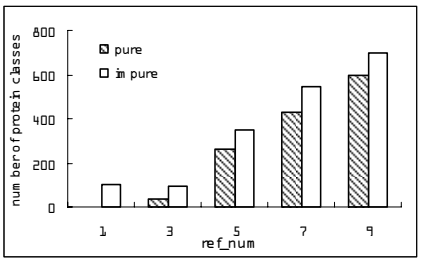
(b) *pw_extension()* process

Fig. 5 An example of *pw_extension()*

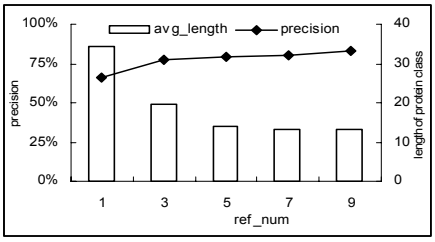
IV. Experimental Results

In this section, the performance of the proposed method is analyzed by a data set containing 53 gel images of human liver samples of hepato-cellular carcinoma (HCC) patients. The average number of spots in a gel image and its standard deviation is 871 spots and 304 spots respectively. The maximum and the minimum number of spots in a gel image are 1609 spots and 257 spots respectively. Among the 53 gel images, 10 different gel images are selected to be used as reference gel images. Generally, a gel image of good quality is selected as a reference gel image by a well-trained analyst. However, this approach may be somewhat subjective. To avoid this factor, the quality of a gel image is measured by the total number of spots in the gel image. In addition, ImageMaster 2D Platinum is employed to perform the spot detection and matching operations of the gel image. Among the protein classes with more than 20 spots are randomly chosen and they are examined manually to determine the correctness of each spot.

In this experiment of Fig. 6, the 10 selected gel images are employed as reference gel images to perform 10 independent spot matching operations to populate a pair database. Fig. 6-(a) shows the characteristics of pure and impure protein classes obtained by the proposed probability based pair extension algorithm. In this figure, the term 'pure' denotes the number of protein classes without any incorrect spot while the term 'impure' does the number of protein classes with at least one incorrect spot. When S_{min}=0.1, i.e., only one reference gel image with the largest number of spots is employed, no pure protein class is identified. The number of impure protein classes is decreased as S_{min} is increased. On the other hand, when S_{min} is increased from 0.1 to 0.5, the number of pure protein classes is enlarged. This is because more incorrectly matched pairs are discarded by the proposed algorithm as S_{min} is increased. On the contrary, when S_{min} is closer to 1 from 0.5, the number of pure protein classes is decreased. Fig. 6-(b) shows the average length of identified protein classes and their average precision together. The term 'precision' of a protein class denotes the ratio of correct spots over the total number of spots in that protein class. It is the counter part of the impurity of a protein class.



(a) Number of pure and impure protein classes

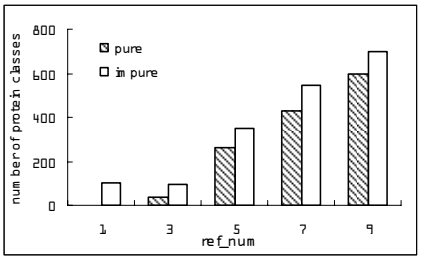


(b) Average length of impure protein classes and precision

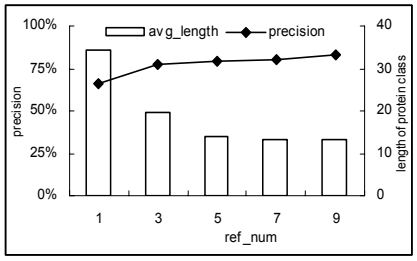
Fig. 6 The effects of reference gel images (ref_num=10)

Fig. 7 shows the performance of the proposed method as the number of reference gel images is increased. Among the 10 selected reference gel images, gel images are selected in the decreasing order of the number of spots. In this experiment, 5 different pair databases are generated by varying the number of reference gel images. The value of S_{min} is fixed to 0.5. In Fig. 7-(a), the number of pure protein classes is compared with that of impure protein classes. As shown in this figure, not only the number of pure protein classes but also the number of impure protein classes is increased. Therefore, in order to get more correct protein classes, it is necessary to increase the number of reference gel images. In Fig. 7-(b), the average length and precision of identified protein classes are shown. The average length of impure protein classes is decreased when more reference gel images are used. Therefore, when more reference gel images are used, the more reliable protein classes can be found. The precision is beyond 80% when more than 5 reference gel images are used.

In Fig. 8, the number of pairs obtained by the virtual average gel image of Progenesis is compared. The term 'avg_gel' denotes the result of the averaged gel image generated by Progenesis. The terms 'multi_ref_pure' and 'multi_ref_pp' mean the number of pure pairs and purified pairs found by the proposed method respectively. As shown in Fig. 8, more than five gel images are averaged for a virtual gel image, most spots of a target gel image fail to be matched with those of the virtual average gel image because of inordinate aggregation. Therefore, the use of virtual average gel image as a reference gel image in a spot matching operation may lead to unreliable



(a) Number of pure and impure protein classes



(b) Average length of identified protein classes and precision

Fig. 7 The effects of reference gel images (ref_num=10)

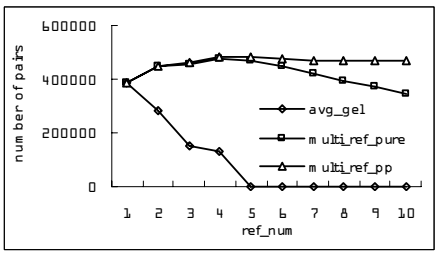
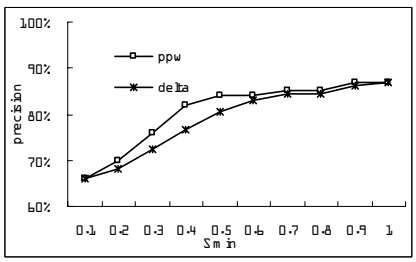
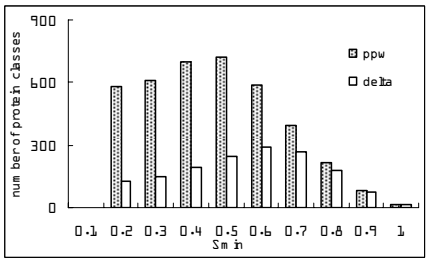


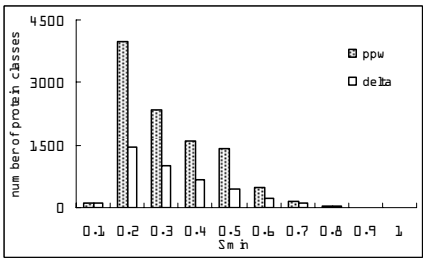
Fig. 8 Performance evaluation on multiple reference gel images



(a) Precision of protein classes



(b) Number of pure protein classes



(c) Number of impure protein classes

Fig. 9 Two-phase pair-wise extension vs δ-purification (ref_num=10)

results in this case. Additionally, even though the number of reference gel images is increased, the number of purified pairs remains almost the same. Therefore, a reasonable number of reference gel image is enough to get more correct pairs, it is not necessary to increase the number of reference gel images as much as possible.

In Fig. 9, the performance of the proposed method is compared with the δ-purification algorithm [11] which has been proposed for the same purpose. The term 'ppw' denotes the proposed probability based pair extension algorithm and the term 'delta' does the δ-purification algorithm. The pair database used in Fig. 7 is used. As shown in Fig. 9-(a), the precision of the proposed algorithm is slightly higher than the δ-purification method. However, in Fig. 9-(b), the number of pure and impure protein classes obtained by the proposed algorithm is much larger. Especially, when S_{min} in the range of [0.2, 0.5], the number of pure protein classes is 2 times larger.

V. Conclusion

The 2-DE method is a cost-effective way to analyze the same proteins of a sample. However, due to the possible noise and geographic distortions of spots in a protein gel image, 2-DE gel image analysis should contain a considerable number of error spots. In order to minimize the influence of these noise and geographic distortions of spots in a 2-DE gel image spot matching operation, several gel images of good quality can be designated as reference gel images to perform multiple independent spot-matching operations which form a pair database of pair transactions. The probability based pair extension method proposed in this paper finds the protein classes of a pair database. A set of experiments is performed to justify the effectiveness of proposed extension method. The experimental results show there is no pure protein class found by one reference gel image alone. However, more pure protein classes can be found when the number of reference gel images is increased. Additionally, the precision of an identified protein class is increased when S_{min} is enlarged.

Acknowledgment

This work was supported by the Korea Science and Engineering Foundation(KOSEF) through the National Research Lab. Program funded by the Ministry of Science and Technology (No.R0A-2006-000-10225-0) and by the MIC(Ministry of Information and Communication), Korea, Under the ITFSIP(IT Foreign Specialist Inviting Program) supervised by the IITA(Institute of Information Technology Assessment).

References

[1] Rabilloud T., "Two-dimensional gel electrophoresis in proteomics: Old, old fashioned, but it still climbs up the mountains", Proteomics, 2002, Vol. 2, pp.3-10.

[2] Mann, M., Hendrickson, R.C.; Pandey, A., "Analysis of Proteins and proteomes by mass spectrometry", Annu. Rev. Biochem. 2001, 70, pp.437-473.

[3] Steven P. G, Garry L. C., Y. Zhang, Y. Rochon and R. Aebersold. "Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology". PNAS, 2000, Vol. 97, No. 17, pp.9390-9395

[4] Rosengren A.T., Salmi J.M., Aittokallio T., Westerholm J.j Lahesmaa R., Nyman T.A and Nevalaien O.S., "Comparison of PDQuest and Progenesis software package in the analysis of two-dimensional electrophoresis gels", Proteomics, 2003, Vol. 3, pp.1936-1946.

[5] Rman B., Cheung A. and Martern M.R., "Quantitative comparison and evaluation of two commercially available, two-dimensional

electrophoresis image analysis software package, Z3 and Melanie", Electrophoresis, 2001, Vol.23, pp.283-291.
[6] Aittokallio T., Salmi J., Nyman TA., Nevalainen OS, "Geometrical distortions in two-dimensional gels applicable correction methods", Journal o f Chromatography, 2005, Vol.815/1-2, pp.25-37.
[7] Gustafsson J.S., Blomberg A. and Rudemo M., "Warping two-dimensional electrophoresis gel images to correct for geometric distortions of the spot pattern", Electrophoresis, 2002, Vol.23, pp.1731-1744.
[8] R. Agrawal, R. Srikant, "Fast algorithms for Mining Association Rules", 1994, VLDB conference, Santiago, Chile.
[9] J. Han, J. Pei, J. Yin, "Mining Frequent Patterns without Candidate Generation", Proc. of SIGMOD, 2000, pp. 1-12.
[10] E. Omiecinski, "Alternative interest measures for mining associations", IEEE Trans. Knowledge and Data Engineering, 2003.
[11] Y. Jin, J. E. Shim and W. S. Lee, "Error Spot Filtering Method in Protein 2-DE Image Spot-Matching Operation", IADIS International Conference, 2007.