Cui et al. BMC Bioinformatics 2014, 15(Suppl 12):S8
http://www.biomedcentral.com/1471-2105/15/S12/S8

BMC Bioinformatics

RESEARCH Open Access

An improved independent component analysis model for 3D chromatogram separation and its solution by multi-areas genetic algorithm

Lizhi Cui[1,4†], Josiah Poon[1*], Simon K Poon[1†], Hao Chen[1†], Junbin Gao[2†], Paul Kwan[3†], Kei Fan[1†], Zhihao Ling[4†]

From IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2013)
Shanghai, China. 18-21 December 2013

Abstract

Background: The 3D chromatogram generated by High Performance Liquid Chromatography-Diode Array Detector (HPLC-DAD) has been researched widely in the field of herbal medicine, grape wine, agriculture, petroleum and so on. Currently, most of the methods used for separating a 3D chromatogram need to know the compounds' number in advance, which could be impossible especially when the compounds are complex or white noise exist. New method which extracts compounds from 3D chromatogram directly is needed.

Methods: In this paper, a new separation model named parallel Independent Component Analysis constrained by Reference Curve (pICARC) was proposed to transform the separation problem to a multi-parameter optimization issue. It was not necessary to know the number of compounds in the optimization. In order to find all the solutions, an algorithm named multi-areas Genetic Algorithm (mGA) was proposed, where multiple areas of candidate solutions were constructed according to the fitness and distances among the chromosomes.

Results: Simulations and experiments on a real life HPLC-DAD data set were used to demonstrate our method and its effectiveness. Through simulations, it can be seen that our method can separate 3D chromatogram to chromatogram peaks and spectra successfully even when they severely overlapped. It is also shown by the experiments that our method is effective to solve real HPLC-DAD data set.

Conclusions: Our method can separate 3D chromatogram successfully without knowing the compounds' number in advance, which is fast and effective.

Background

For thousands of years, plants have played a dominant role in the development of sophisticated traditional herbal medicine (HM) systems [1,2]. And nowadays, HM has also attracted much interest of both patients and scientists [3]. However, herbal medicines are extracted with boiling water during the decoction process, which makes it very difficult to realize quality control [4]. In 1991 [5], the World Health Organization (WHO) accepted chromatography fingerprint, which reflects the complex chemical composition of the analyzed sample based on spectroscopic, chromatographic or electrophoretic techniques [6], as a methodology for the assessment of natural products. But, two disadvantages exist for chromatography fingerprint: it relies on retention time which is not stable; it is chosen from only one specific wavelength which misses much information from other wavelength. So, 3D chromatogram generated by High Performance Liquid Chromatography-Diode Array Detector (HPLC-DAD) was researched widely [7-10]. The construction of HPLC-DAD dataset is illustrated in Figure 1, in which there are three compounds contained in the solution for example.

A drop of sample is injected at the top of a column with absorbent. A cup of solvent, same as that in the sample, carries the sample through the column. Different

* Correspondence: Josiah.poon@sydney.edu.au
† Contributed equally
[1]School of information technologies, University of Sydney, Sydney, NSW 2006, Australia
Full list of author information is available at the end of the article

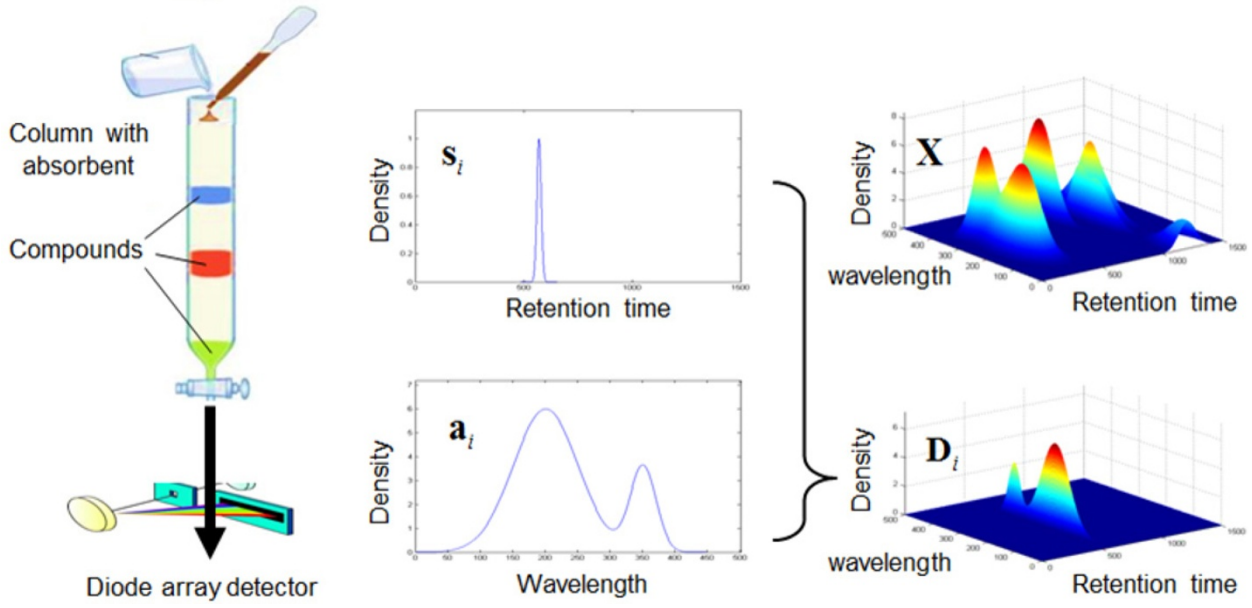

Figure 1 The principle of HPLC-DAD dataset. The vector of $\mathbf{s}_i$ represents chromatogram peaks and the vector of $\mathbf{a}_i$ represents spectra. The matrix of $\mathbf{D}_i$ represents 3D dataset containing only one compound, which is formed by $\mathbf{a}_i \times \mathbf{s}_i^T$. The matrix of X represents the dataset containing more than one compound.

compounds will receive different resistance when they go through the column. Given an ultraviolet detector at the bottom of the column, a chromatogram peak represented by $\mathbf{s}_i (i = 1, 2, 3)$ is formed to reflect the concentration for corresponding compound. If using a DAD for detection, which has more than one thousand channels to detect multi-wavelength simultaneously, besides chromatogram peaks of the outflowing compounds, spectra represented by $\mathbf{a}_i (i = 1, 2, 3)$ will also be recorded. The matrices of $\mathbf{D}_i$ and $\mathbf{X}$ represent i^{th} compound and the mixture of all the compounds respectively. Their relationship is given as

$$\mathbf{X} = \sum_{i=1}^{n} \mathbf{D}_i = \sum_{i=1}^{n} \mathbf{a}_i \times \mathbf{s}_i^T = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n] \times \begin{bmatrix} \mathbf{s}_1^T \\ \mathbf{s}_2^T \\ \vdots \\ \mathbf{s}_n^T \end{bmatrix} = \mathbf{AS} \quad (1)$$

where, the variable of n is the number of compounds contained in the solution, which equals to 3 for the example in Figure 1.

There are many methods to separate $\mathbf{X}$ in (1), such as evolving factor analysis (EFA)[11], heuristic evolving latent projections (HELP)[12], window factor analysis (WFA)[13], orthogonal projection resolution (OPR)[14], evolving window orthogonal projections (EWOP)[15], iterative target transformation factor analysis (ITTFA)[16], alternating regression (AR)[17], parallel factor analysis (PARAFAC1/2)[18], multivariate curve resolution-alternating least squares (MCR-ALS)[19] and interactive self-modelling mixture analysis (SIMPLISMA)[20], alternating trilinear decomposition (ATLD)[21] and immune algorithm (IA)[22]. However, all these method need the number of the compounds to be known in advance. And the method to obtain the compounds' number is based on Eigenvalue, which will miss small peaks especially when noise is severe. Recently, Independent Component Analysis (ICA)[23] was introduced in this field, which considered compounds and noises as independent components. But two disadvantages existed: 1) noises was considered as independent components, which gave unexpected and useless information in the results; 2) identifying compounds from noises after separating all the independent components was still needed.

In order to extract compounds directly from the data set, this paper proposed a parallel model of Independent Component Analysis constrained by Reference Curves (pICARC) and its solution by multi-areas Genetic Algorithm (mGA). In section 2, the principle of pICARC and mGA were proposed. In section 3, simulations and experiments were provided to show the performance of our method. Finally, conclusions and future works were summarized in section 4.

Methods

The principle of our method is illustrated in Figure 2. The left thick box is the mathematical part, and the right thick box are corresponding data sets in the reality. Firstly, we construct a kind of Reference Curve (RC)

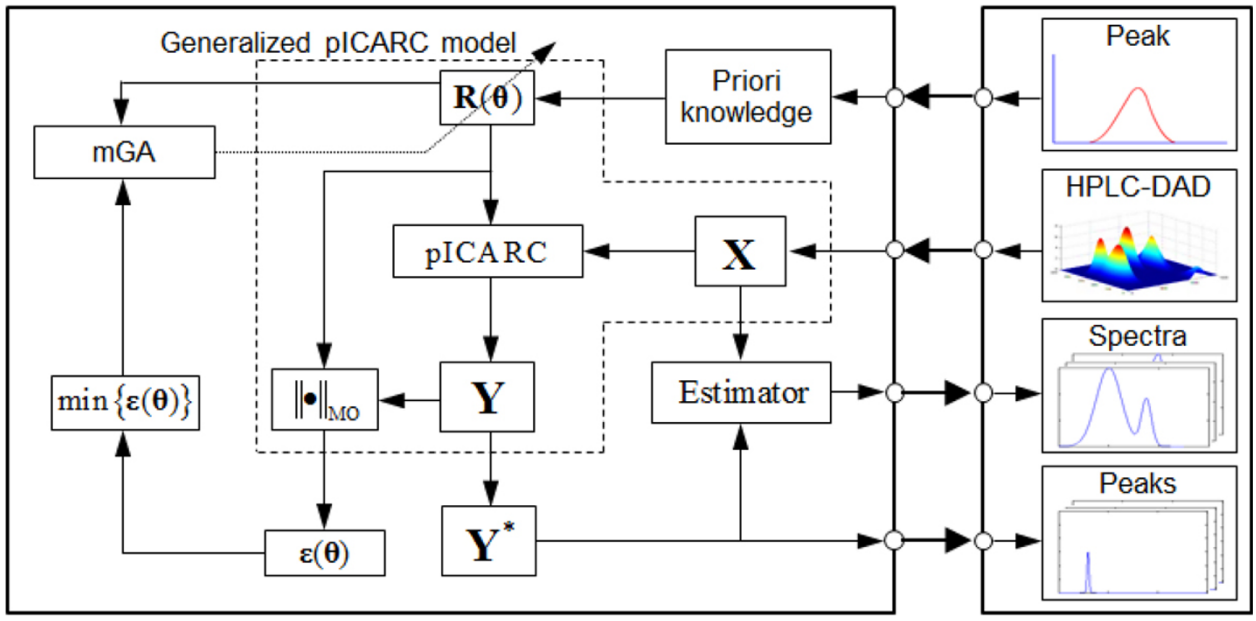

Figure 2 The structure of our method. $\mathbf{R}(\boldsymbol{\theta})$ is Reference Curve (RC), $\mathbf{Y}$ is Calculated Curve (CC), $\|\bullet\|_{\mathrm{MO}}$ is Measurement Operator (MO) to give the distance $\boldsymbol{\varepsilon}(\boldsymbol{\theta})$ between $\mathbf{R}(\boldsymbol{\theta})$ and $\mathbf{Y}$

$\mathbf{a}_i \times \mathbf{s}_i^T$ with different parameter of $\boldsymbol{\theta}$ based on the priori knowledge. Then inputting $\mathbf{R}(\boldsymbol{\theta})$ and $\mathbf{X}$ into the pICARC model, Calculated Curves (CCs) $\mathbf{Y}$ will be obtained. The distances, $\boldsymbol{\varepsilon}(\boldsymbol{\theta})$, between $\mathbf{R}(\boldsymbol{\theta})$ and $\mathbf{Y}$ are calculated by the Measurement Operator (MO) $\|\bullet\|_{\mathrm{MO}}$. Combining all the elements contained in the dash polygon together, it is called generalized pICARC model, which has converted the separation problem to a multi-parameter optimization issue. Next, what is needed is to find the parameters of $\boldsymbol{\theta}^*$, which minimizes the value of $\boldsymbol{\varepsilon}(\boldsymbol{\theta})$.

In this paper, the algorithm of mGA, which is proposed with reference to the features of $\mathbf{R}(\boldsymbol{\theta})$, is used to search $\boldsymbol{\theta}^*$. After all the $\boldsymbol{\theta}^*$ have been found, $\mathbf{Y}^*$ will be obtained, whose row vectors are the chromatogram peaks for individual compounds. By using an estimator, the spectra of the compounds will be obtained. The priori knowledge has been introduced within our method. What the user needed to do is just to input your data set$\mathbf{X}$.

Following, the priori knowledge about chromatogram peaks, the pICARC model, MO, mGA and the estimator will be introduced one by one.

The priori knowledge

According to the physical principle of chromatography, each peak of chromatogram looks like certain curves such as Gaussian curve, Log-normal curve, Gamma curve and Weibull curve, etc. [24], among which, the Gaussian function was used prevalently for simulating chromatogram peaks in many relevant researches. [25,26] So, we will use Gaussian curve, which is illustrated in equation (2) and Figure 3, in our research. There are two parameters: μ and

$$\gamma = \exp[-\frac{(x-\mu)^2}{2*\sigma^2}].$$

$$\gamma = \exp[-\frac{(x-\mu)^2}{2*\sigma^2}] \tag{2}$$

where, the factor of $1/\sqrt{2\pi}\sigma$ is removed to set the maximum value to 1. The amplitude information will be found in the spectra.

The range of the parameter μ is decided by the data set, whose minimum value is $\mu_1 = 1$ and maximum

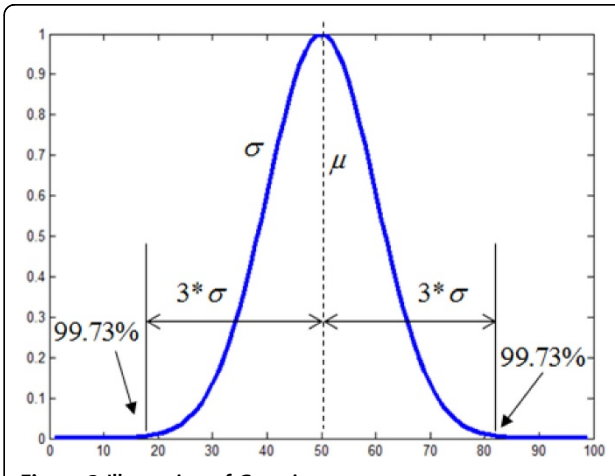

Figure 3 Illustration of Gaussian curve.

value is $\mu_2 = column(\mathbf{X})$. According to the quantile of 99.73%, we have the following inequality

$$\sigma < \frac{\mu_2 - \mu_1 + 1}{6} \tag{3}$$

pICARC model

The model of ICA is represented as (4)

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_t] = \mathbf{A} \times \mathbf{S} = \mathbf{A} \times [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_t] = \mathbf{A} \times \begin{bmatrix} \mathbf{s}_1^T \\ \mathbf{s}_2^T \\ \vdots \\ \mathbf{s}_n^T \end{bmatrix} \tag{4}$$

where $\mathbf{X}$ are the observation vectors; $\mathbf{A}$ is the mixed matrix; $\mathbf{S}$ are the source vectors; the subscript of $\boldsymbol{t}$ is the number of the samples in $\mathbf{A}$ and $\mathbf{S}$. The purpose of this model is to obtain $\mathbf{A}$ and $\mathbf{S}$ only based on $\mathbf{X}$ under four assumptions[27]. According to the separability theorem of ICA[28], we could trust ICA to separate HPLC-DAD data set as long as the number of the wavelength is greater than the number of the compounds. In 1999, the algorithm of fastICA was proposed to solve (4) [29,30], of which the parallel form is shown as

$$\begin{cases} \max\ E\left\{G(\mathbf{B}^T\tilde{\mathbf{X}})\right\} = E\left\{G\left(\begin{bmatrix} \mathbf{b}_1^T \\ \mathbf{b}_2^T \\ \vdots \\ \mathbf{b}_n^T \end{bmatrix}[\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \cdots, \tilde{P}\mathbf{x}_n]\right)\right\} \\ \text{subject to } E\left\{\left(\mathbf{b}_i^T\tilde{\mathbf{x}}\right)^2\right\} = \|\mathbf{b}_i\|^2 = 1,\ (i = 1, 2, \cdots, n) \end{cases} \tag{5}$$

where, $\mathbf{b}_i$ is a m-dimensional vector such that$E\left\{\left(\mathbf{b}^T\tilde{\mathbf{x}}\right)^2\right\} = 1$, $\tilde{\mathbf{X}}$ is a matrix preprocessed from$\mathbf{X}$, $G(x) = 1/4 \bullet (x^4)$ is a nonlinear function. Applying KKT condition [31] and Newton method [32], the iterative equation of $\mathbf{B}$ is shown as

$$\begin{cases} \mathbf{B}^+ = E\left\{\tilde{\mathbf{X}}g(\mathbf{B}^T\tilde{\mathbf{X}})\right\} - E\left\{g'(\mathbf{B}^T\tilde{\mathbf{X}})\right\}\mathbf{B} \\ \mathbf{B}^+ = (\mathbf{B}\mathbf{B}^T)^{-1/2}\mathbf{B} \end{cases} \tag{6}$$

where, $g(\bullet)$ is the first derivative of $G(\bullet)$.

As mentioned in the introduction, two major problems were found when applying the ICA model to HPLC-DAD data set. Therefore, suitable modification about ICA based on the priori knowledge of chromatogram peaks is needed to constrain the shape of the source signals. What should be noted is that the variable used for calculation in ICA model is the $\tilde{\mathbf{X}}$, which generates a signal of $\mathbf{y}'_i$ shown in equation (7). However, the curve that should be constrained is $\mathbf{y}_i$, which is a calculated signal to approximate $\mathbf{s}_i^T$. The $\mathbf{D}_w$ is the whitening matrix. There is a difference between $\mathbf{y}'_i$ and $\mathbf{y}_i$, which is caused by pre-processing.

$$\begin{cases} \mathbf{y}'_i = [\gamma'_{i1}, \gamma'_{i2}, \cdots, \gamma'_{it}] = \mathbf{b}_i^T\tilde{\mathbf{X}} = \mathbf{b}_i^T[\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \cdots, \tilde{\mathbf{x}}_t] \\ \mathbf{y}_i = [\gamma_{i1}, \gamma_{i2}, \cdots, \gamma_{it}] = \mathbf{b}_i^T\mathbf{D}_w\mathbf{X} = \mathbf{b}_i^T\mathbf{D}_w[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_t] \end{cases} \tag{7}$$

In order to avoid introducing a new variable, which should represent the difference between $\mathbf{y}'_i$ and $\mathbf{y}_i$, the model is constructed as

$$\begin{cases} \max[E\{G(\bar{\mathbf{B}}^T\tilde{\mathbf{X}}')\} - \left\|\bar{\mathbf{B}}^T\tilde{\mathbf{X}}'' - \mathbf{R}(\boldsymbol{\theta})\right\|_{\mathrm{MO}}] \\ \mathbf{R}(\boldsymbol{\theta}) = [\mathbf{r}_1(\boldsymbol{\theta}), \mathbf{r}_2(\boldsymbol{\theta}), \cdots, \mathbf{r}_n(\boldsymbol{\theta})]^T \\ \bar{\mathbf{B}}^T = [\mathbf{B}^T, \mathbf{d}] = [[\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_n]^T, \mathbf{d}_{1\times n}] \\ \tilde{\mathbf{X}}' = [\tilde{\mathbf{X}}^T, \mathbf{0}]^T \\ \tilde{\mathbf{X}}'' = [\tilde{\mathbf{X}}^T, \mathbf{1}]^T \\ E\{(\mathbf{b}_i^T\tilde{\mathbf{x}}')^2\} = E\{(\bar{\mathbf{b}}_i^T\tilde{\mathbf{x}}')^2\} = \|\mathbf{b}\|^2 = 1 \end{cases} \tag{8}$$

where $\mathbf{d}$ represents the differences existing in the equation (7). The equation (8) is called pICARC.

Measurement operator

The purpose of MO is to measure the distance between $\mathbf{y}_i$ and $\mathbf{r}_i(\boldsymbol{\theta})$. Here, the vector of $\boldsymbol{\varepsilon}(\boldsymbol{\theta}) = [\varepsilon_1(\boldsymbol{\theta}), \varepsilon_2(\boldsymbol{\theta}), \cdots, \varepsilon_n(\boldsymbol{\theta})]$ contains all these distances. We gave the definition of $\varepsilon_i(\boldsymbol{\theta})$as

$$\varepsilon_i(\boldsymbol{\theta}) = \left\|\mathbf{y}_i - \mathbf{r}_i(\boldsymbol{\theta})\right\|_2^2 = \sum_j [\mathbf{y}_i(j) - \mathbf{r}_i(j;\boldsymbol{\theta})]^2 \tag{9}$$

where, j represents every element in the vectors of $\mathbf{y}_i$ and $\mathbf{r}_i(\boldsymbol{\theta})$

Multi-area genetic algorithm

GAs are a family of computational models inspired by evolution. These algorithms encode a potential solution to a specific problem on a simple chromosome-like data structure, and apply recombination operators to these structures in such a way as to preserve critical information. Usually, GAs are used to find one global optimal point in the searching plane. But in our problem, several points in the $\mu - \sigma$ plane should be found as solutions simultaneously. In order to search multi optimal points, we propose an algorithm named multi-areas Genetic Algorithm (mGA).

Areas are circles which are composed of chromosomes that are closed to one another, where the candidate solution will be found. If the fitness of one solution is much better than the others, the area around it will have better fitness as well. This will lead many elites assemble into this area. In order to balance the number of elites among all the areas, emigrant and immigrant policy are adopted. Except limited chromosomes left in every area, the others will be selected as emigrants,

which will keep balance among population in all areas. Immigrant policy under certain criterion introduces new chromosomes into all areas.

The flow chart of mGA is illustrated in Figure 4. Firstly, parameters, which will be depicted in step 1), are initialized. Secondly, initial population is generated. Then, multi-areas have been formed among the population. Each of these areas has a center and a radius. Only in the areas, whose radius are big enough, the elite chromosomes have the right to match with each other and generate children. Children with high fitness values will replace the inferior individuals. Only a certain number of those top ranking chromosomes in an area will be kept for the next generation, the others will be selected as emigrants to avoid too many chromosomes converging in one dominant area. Later, the emigrants will be allocated into different areas under certain law, which will be descripted in step 6).

After the migration process, new generation has been formed. New multi-areas will be separated among the new generation again. The iteration continues until there is no area with the size big enough to generate children. Finally, only several areas left, which contain the candidates for solution. Solutions are found according to the decision as described in step 7).

1) Parameters initialization: The major parameters include the number of the population, the number of elites, length of the chromosome and the fitness function. The first three are decided according to the size of the search plane. The fitness of every chromosome is given by equation (9). Here, smaller the fitness is, more superior the chromosome is.
2) Population initialization: The population is equal to the row number of $\mathbf{R}(\boldsymbol{\theta})$. In order to get initial population with better fitness, two steps were adopted. Firstly, the search space is separated into several sub spaces equally, and maximum number of chromosomes is randomly generated in every sub spaces. Then, top ones according to their fitness value are selected as initial population.
3) Multi-areas separation: Multi-areas are formed among the population according to following steps: (a) select the chromosome with best fitness as the center of an area; (b) give an user-defined radius and draw a circle; (c) calculate the distance between center and the other chromosomes contained in this area, update the radius of this area with the largest distance; (d) for chromosomes not belonging to a specific area, repeat Step (a) to (d) until no more chromosome left.
Some of these circles (areas) may overlap with each other. For those circles with severe overlapping, they should be merged as a big area for immigration; this will be further described in step 6).
4) Mating and reproduction: Every elite finds a mate which has the highest hamming distance, i.e. the number of different bits between them, by the sequence ordered of fitness. New chromosomes, the children, are generated by crossing every different bit between the mated pairs. If a child has better fitness value than any one of its parent, it will be classified as excellent child, which will be reserved; otherwise, it will be dropped.
5) Select migrants: Only limited chromosomes, whose number is user-defined depending on application, according to their fitness will stay in one area, the others are selected as emigrants. In this step, only the number of the migrants is decided. The generation of immigrants will be descripted in step 6).
6) Migrant quota among areas allocation: For every circle (area), more elites in the previous generation, less immigrants will be allowed, to avoid too many elites in one area. If two areas are severely overlapped, they will be merged as one big area to receive immigrants. Otherwise, the quota will be unfair because of

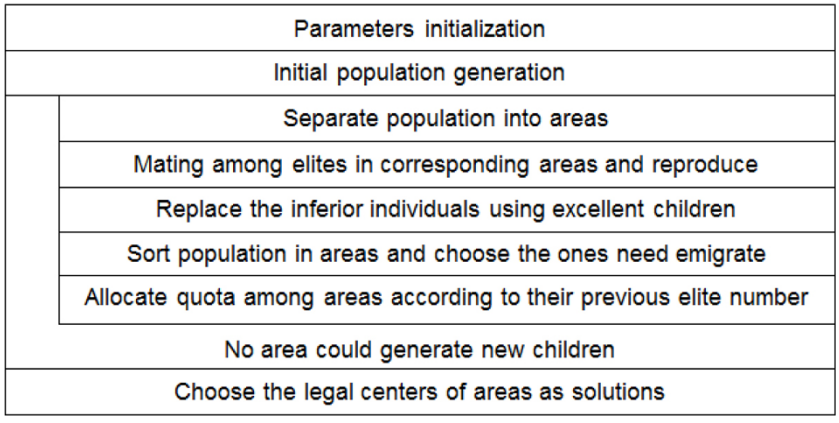

Figure 4 Flow chart of mGA.

the overlap. Take the experiments in this paper as an example. There is a severe overlap between area 1 and area 3. And there are 23 elites in area 1 and 1 elite in area 3. If the area 1 and area 3 are not combined as one, there will be a small number of migrants for area 1 but many for area 3. The immigrant in area 3 can also be considered as immigrant for area 1 because of the overlap. There is also an overlap between area 1 and area 5, but the center of area 5 is at the boundary of the $\mu - \sigma$ plane, which is ignored for further evolution.

The chromosomes with the same number as the initial population will be generated randomly for every circle (area), but only top ones decided by the quota are selected as immigrants.

7) Find solution: As the algorithm proceeds, the radii of the areas will become smaller. When all the radii become small enough, the program ends. The center of each area is the candidate solution. If the center is a local minimum, it will be selected as a solution.

Estimator

After obtaining all the chromatogram peaks, $\mathbf{Y}^* \approx \mathbf{S}$, by solving equation (8), the spectra can be estimated with $\mathbf{Y}^*$ and $\mathbf{X}$ with considering the noise contained in $\mathbf{X}$. For simplicity, we ignore the noise in this paper to calculate spectra by

$$\mathbf{A} = \mathbf{X} * \text{pinv}(\mathbf{S}) \tag{10}$$

Where the pinv(•) is the pseudo inverse function. Equation (10) is derived from equation (1) directly.

Results and discussion

In this section, a group of simulations were given to explain the principle of our method. Then several experiments on a real HPLC-DAD data set were given to demonstrate the practicability of our method. Two criteria were used to evaluate our method: 1) to see whether the compounds' number found by our method was right; 2) to see the errors between the true spectra and calculated spectra.

Simulations and discussion

As illustrated in Figure 5, five compounds' chromatogram peaks, which are represented by different parameters of (μ, σ) respectively, are constructed in the simulation dataset. The parameters for the compounds' chromatogram peaks from 1^{st} to 5^{th} are: (50, 21), (75, 12), (90, 10), (155, 17) and (175, 9) respectively. These initial parameters' distribution on the $\mu - \sigma$ plane is shown in Figure 6. There are many areas for the initial parameters, but few areas for the final parameters. The program is available from the corresponding author.

From the simulation, we can see:

(1) The method proposed in this paper could separate 3D chromatogram into chromatogram peaks and spectra effectively without know the compounds' number in advance even severe overlap exist. The pICARC model transformed the separation problem to a multi-parameter optimization issue, which could be solved by swarm intelligent algorithm. The algorithm of mGA could find all the solutions simultaneously.

(2) Sometimes, the result given by this method was incorrect. This is because that the chromosome is initialized randomly, which will cause undetermined situation. This problem can be solved by running program multiple times and compare the candidate results to obtain the final results. Ten times of simulation have been done in this simulation and seven times have the correct result, which is list in Table 1.

(3) The implementation of the mGA is fast. Among ten times of simulation, the slowest one took 11 steps and 4.3652 seconds.

Experiments and discussion

The program of experiment is available for free at from the corresponding author. The data set of "adataset.mat", which is used in the experiment, can be downloaded

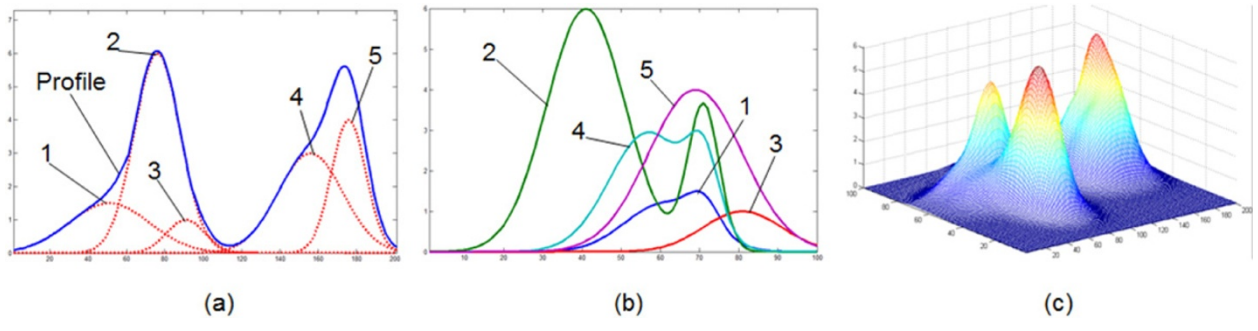

Figure 5 Dataset for simulation. (a) Chromatogram peaks for compounds. (b) Spectra for compounds. (c) 3D chromatogram.

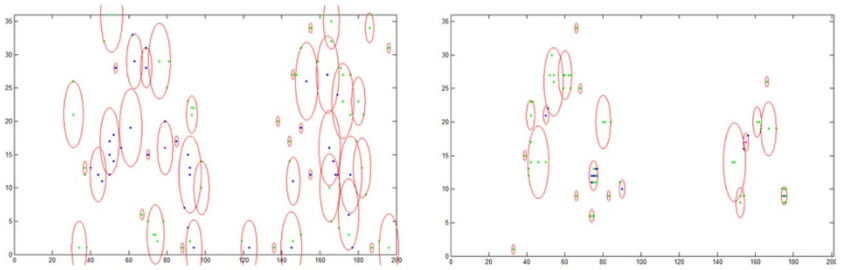

Figure 6 Distribution of the parameters. The abscissa is μ, the ordinates is σ.(a) Distribution of the initial parameters. (b) Distribution of the final parameters.

from http://www.mcrals.info for free [33]. The data set is illustrated in Figure 7. The data set is a three-compound system with two pesticides identified and one unknown interferent. The three-way data set is formed by one matrix with the three compounds and two matrices of standards with one known compound.

Ten experiments were run totally. As the population initialization was randomly, we just list the initial population and initial multi-areas for the first experiment in Table 2. Among the ten experiments, eight gave the same results: (19, 7), (31, 13) and (60, 10). The initial population distribution and the final population distribution of one experiment are illustrated in Figure 8. The calculated results are illustrated in Figure 9.

From the results, we can see:

(1) The method proposed in this paper can separate true HPLC-DAD data set into chromatogram peaks and spectra without know the compounds' number in advance. In the ten experiments, eight gave correct results. And the maximum step and time cost are 16 and 2.6854 seconds.

(2) In (a) of Figure 9, the Profile is the projection of the data set along the axis of wavelength. And the chromatogram peaks are products of the calculated curves by the maximum value of the corresponding calculated spectra. The reason that the second peak is higher than the profile is that there is some values in the spectra are negative. The reason for the negative will be discussed in next item.

(3) In (b) and (c) of Figure 9, there are still errors between the calculated spectra and true spectra. This could be caused by following two reasons: noise existed in the data set; difference between the reference curve and true chromatogram peaks. For the first reason, the estimator, shown in equation (10) should be improved. For the second reason, more detailed reference curve should be proposed to replace equation (2).

Table 1 Correct results for the simulation.

Areas	Center	Radius	NumE	NumP	ErrorC	ErrorS
1	(155,17)	1	6	6	6.8102e-9	1.8503e-7
2	(75,12)	1.4142	2	5	1.0044e-8	1.1303e-8
3	(90,10)	1	21	26	1.0864e-8	1.372e-7
4	(175,19)	3.1623	4	6	1.6714e-8	5.1367e-8
5	(50,21)	1	1	3	2.4052e-8	1.4096e-7

Column of NumE is the elites' number contained in this area. Column of NumP is the population's number contained in this area. Column of ErrorC is the $\boldsymbol{\varepsilon}(\boldsymbol{\theta})$ between the reference curves with the center as parameters and the chromatogram peaks. Column of ErrorS is the error between the calculated spectra and the true spectra.

Conclusions

In this paper, the pICARC model and its solution by mGA were proposed. A priori knowledge of chromatogram peaks is introduced into ICA model. And the shape of the chromatogram peaks, which are the source signals in the ICA model, is constrained with certain kind of function with parameters. Because the Gaussian curve is used widely in relevant field, we use Gaussian curve to constrain the shape of the chromatogram peaks. In order to solve this model, which contained several objective points in the $\mu - \sigma$ plane, we modified the algorithm of GA to propose the algorithm of mGA. This algorithm separated all the population into multi-areas according to their fitness and distances from each other. Immigrant policy was adopted to keep the variety of the population. In order to avoid gathering too many elites into one dominant area, the immigrants were allocated according to the former elite number of the destination area. Finally, simulations and experiments were done to prove the performance of our model and its solution algorithm. In this section, conclusions were summarized firstly, and then future works were prospected.

Conclusions

(1) The pICARC model transformed the separation problem of HPLC-DAD data set to a multi

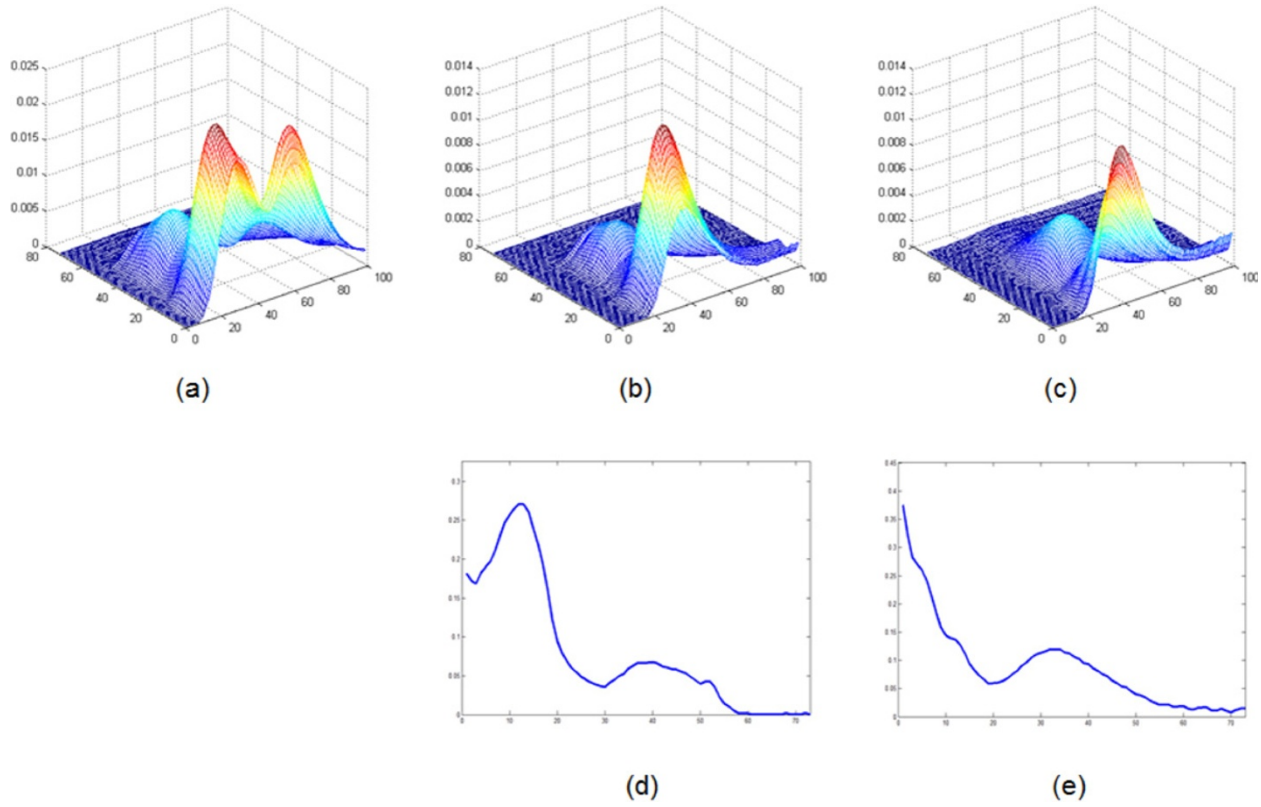

Figure 7 HPLC-DAD data set used in the experiments. (a) 3D chromatogram contains all the three compounds. (b) 3D chromatogram contains only the first compound. (c). 3D chromatogram contains only the second compound. (d) The spectrum of the first compound. (e) The spectrum of the second compound.

parameters optimization issue, which can be solved by abundant optimization algorithms.

(2) By introducing a priori knowledge of chromatogram into the ICA model, only the useful signals will be picked out from the mixed data set. It is not necessary for us to know the compounds' number in advance or to discriminate the signal from noises after finding out all the independent components.

Table 2 Initial population and initial multi-area for the first experiment.

Areas	Center	Radius	Error	NumE	NumP	Big area
1	(31,13)	5.6569	0.0045	23	25	1
2	(21,7)	5.6569	0.0062	8	13	2
3	(25,12)	5	0.0099	1	2	1
4	(60,10)	5.8310	0.0100	3	11	3
5	(37,16)	5.0990	0.0108	1	8	0
6	(53,17)	5	0.0173	0	9	0
7	(44,17)	1	0.0180	0	2	0
8	(98,17)	5	0.0205	0	3	0

Column of error is the $\varepsilon(\theta)$ between the reference curves with the center as parameters and the chromatogram peaks. Column of NumE is the elites' number contained in this area. Column of NumP is the population's number contained in this area.

This means that this model improves the accuracy as well as saves the time for calculation.

(3)The algorithm of mGA is a useful method to search multiple objective points in the $\mu - \sigma$ plane simultaneously. The information of chromosome' fitness and their distance from each other were used to cluster them into multi-areas. Immigrant policy and quota allocation law were made to keep the variety for every areas. Genetic operators were used to keep the evolution among areas.

Future works

According to the discussions in section 3, three major works are needed to be done in the future:

(1) Improved estimator should be developed to eliminate the effect caused by noise existing in the data set. The estimator used in this paper, shown by equation (10), ignores the noise. So there are errors existing in the results.

(2) More accurate reference curves should be introduced into the pICARC model. The other functions referred in section II should be used to see whether

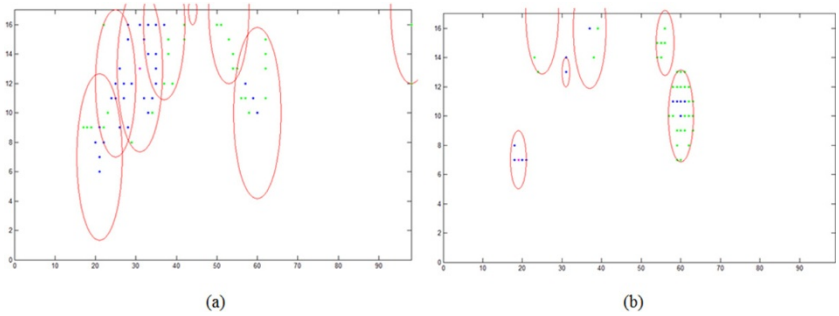

Figure 8 Distribution of population. (a) The initial population. (b) The final population.

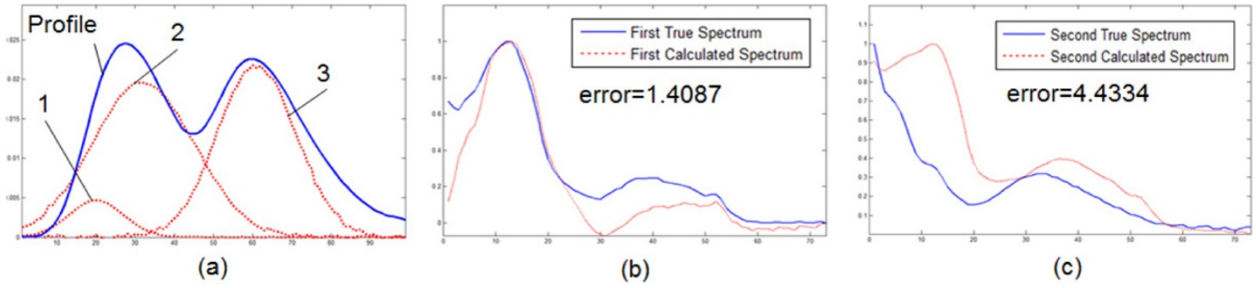

Figure 9 Results of the experiment. (a) Three calculated chromatograms. (b) Compare between first true spectrum and first calculated spectrum. (c) Compare between second true spectrum and second calculated spectrum.

better results could be obtained. In reality, the chromatogram curve maybe more complex than the functions referred in this paper, new reference curves with more parameters could be proposed based on the experiments for out model.

(3) Facing new reference curves, which could have more parameters than that used in this paper, new optimization algorithm should be developed to fit the new parameters space.

List of abbreviations

HPLC-DAD: High Performance Liquid Chromatography-Diode Array Detector
ICA: Independent Component Analysis
pICARC: parallel Independent Component Analysis constrained by Reference Curve
GA: Genetic Algorithm.
mGA: multi-area Genetic Algorithm.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LC designed the pICARC model, designed the mGA algorithm and drafted the manuscript. JP and SKP supervised the manuscript, conceived of the whole method and revised the manuscript carefully. HC participated in the revision of this manuscript. JG and PK participated in the design of pICARC model and helped to draft the manuscript. KF supervised the manuscript from pharmacy view. ZL participated in the design of the simulations and experiments. All authors read and approved the final manuscript.

Acknowledgements

Lizhi Cui thanks school of information technologies, the University of Sydney for providing him with a Ph.D. fellowship. Lizhi Cui thanks China Scholarship Council for providing living expenses during the study in Sydney, and the students No. is 201206740061.

Declarations

Publication of this paper has been funded by the corresponding author. This article has been published as part of *BMC Bioinformatics* Volume 15 Supplement 12, 2014: Selected articles from the IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2013): Bioinformatics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/15/S12.

Authors' details

[1]School of information technologies, University of Sydney, Sydney, NSW 2006, Australia. [2]School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW 2795, Australia. [3]School of Science and Technology, Univeristy of New England, Armidale, NSW 2350, Australia. [4]Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai, 200237, China.

Published: 6 November 2014

References

1. Cragg GM, Grouthaus PG, Newman DJ: **Impact of natural products on developing new anti-cancer agents.** *Chem Rev* 2009, **B**:3012-3043.
2. Tistaert C, Dejaegher B, Heyden YV: **Chromatographic separation techniques and data handling methods for herbal fingerprints: a review.** *Anal Chim Acta* 2011, **690**:148-161.
3. WHO: *Traditional Medicine strategy 2002-2005* Geneva: Switzerland; 2002.
4. Liang YZ, Xie P, Chan K: **Quality control of herbal medicines.** *J Chromatogr B* 2004, **812**:53-70.

5. WHO: *Guidelines for the assessment of herbal medicines* Geneva: Switzerland; 1991.
6. Royal Society of Chemistry: *IUPAC Compendium of Chemical Terminology* Cambridge: UK; 1997.
7. Marini F, Aloise AD, Bucci R, Buiarelli F, Magri AL, Magri AD: **Fast analysis of 4 phenolic acids in olive oil by HPLC-DAD and chemometrics.** *Chemom Intell Lab Syst* 2011, **106**:142-149.
8. Vosough M, Mojdehi NR, Salemi A: **Chemometrics assisted dispersive liquid-liquid microextraction for quantification of seven UV filters in urine samples by HPLC-DAD.** *J Sep Sci* 2012, , **35:** 3575-3585.
9. Liu X, Wu Z, Yang K, Ding H, Wu Y: **Quantitative analysis combined with chromatographic fingerprint for comprehensive evaluation of Danhong injection using HPLC-DAD.** *J Pharm Biomed Anal* 2013, **76**:70-74.
10. Li G, Bu H, Yang MQ, Zeng X, Yang JY: **Selecting subsets of newly extracted features from PCA and PLS in microarray data analysis.** *BMC Genomics* 2008, **9(Suppl2)**:S24.
11. Maeder M: **Evolving factor analysis for the resolution of overlapping chromatographic peaks.** *Anal Chem* 1987, **59**:527-530.
12. Kvalheim OM, Liang Y: **Heuristic evolving latent projections: resolving two-way multicomponent data. 1. Selectivity, latent-projective graph, datascope, local rank, and unique resolution.** *Anal Chem* 1992, **64**:936-946.
13. Malinowski ER: **Window factor analysis: theoretical derivation and application to flow injection analysis data.** *J Chemom* 1992, **6**:29-40.
14. Liang Y, Kvalheim OM: **Diagnosis and resolution of multiwavelength chromatograms by rank map, orthogonal projections and sequential rank analysis.** *Anal Chim Acta* 1994, **292**:5-15.
15. Xu C, Jiang J, Liang Y: **Evolving window orthogonal projections method for two-way data resolution.** *Analyst* 1999, **124**:1471-1476.
16. Vandeginste B, Essers R, Bosman T, Reijnen J, Kateman G: **Three-component curve resolution in liquid chromatography with multiwavelength diode array detection.** *Anal Chem* 1985, **57**:971-985.
17. Karjalainen EJ: **The spectrum reconstruction problem use of alternating regression for unexpected spectral components in two-dimensional spectroscopies.** *Chemom Intell Lab Syst* 1989, **7**:31-38.
18. Vosough Maryam, Amir Salemi: **Exploiting second-order advantage using PARAFAC2 for fast HPLC-DAD quantification of mixture of aflatoxins in pistachio nuts.** *Food Chemistry* 2011, **127**:827-833.
19. Tauler R, Barcelo D: **Multivariate curve resolution applied to liquid chromatography-diode array detection.** *Trends Anal Chem* 1993, **12**:319-327.
20. Cao L, Harrington PB, Liu J: **SIMPLISMA and ALS applied to two-way nonlinear wavelet compressed ion mobility spectra of chemical warfare agent simulants.** *Anal Chem* 2005, **77**:2575-2586.
21. Zhao J, Wu H, Niu J, Yu Y, Yu L, Kang C, Li Q, Zhang X, Yu R: **Chemometric resolution of coeluting peaks of eleven antihypertensives from multiple classes in high performance liquid chromatography: A comprehensive research in human serum, health product and Chinese patent medicine samples.** *J chromatogr B* 2012, **902**:96-107.
22. Shao X, Liu Z, Cai W: **Resolving multi-component overlapping GC-MS signals by immune algorithms.** *Trends Anal Chem* 2009, **28**:1312-1321.
23. Debrus B, Lebrun P, Ceccato A, Caliaro G, Govaerts B, Olsen BA, Rozet E, Boulanger B, Hubert P: **A new statistical method for the automated detection of peaks in UV-DAD chromatograms of a sample mixture.** *Talanta* 2009, **79**:77-85.
24. Grimalt J, Iturriaga H: **The resolution of chromatograms with overlapping peaks by means of different statistical functions.** *Analytica Chimica Acta* 1982, **139**:155-166.
25. Lin D, Lu X: **Resolution of noisy HPLC-DAD data by two-dimensional wavelet transform and subwindow factor analysis.** *Chem J Chin Univ* 2005, **26**:1039-1042.
26. Zhang z, Chen S, Liang Y: **Peak alignment using wavelet pattern matching and differential evolution.** *Talanta* 2011, **83**:1108-1117.
27. Lathauwer LD, Moor BD, Vandewalle J: **An introduction to independent component analysis.** *J Chemom* 2000, **14**:123-149.
28. Eriksson J, Koivunen V: **Identifiability, separability, and uniqueness of linear ICA models.** *IEEE Signal Process, Lett* 2004, **11**:601-604.
29. Hyvarinen A: **Fast and Robust Fixed-Point Algorithms for Independent Component Analysis.** *IEEE Trans On Neural Networks* 1999, **10**:626-634.
30. Hyvarinen A, Oja E: **Independent Component Analysis: Algorithms and Applications.** *Neural Networks* 2000, **13**:411-430.
31. Belegundu AD, Chandrupatla TR: **Constrained minimization.** *Optimization Concepts and applications in Engineering* 2nd edition. Edited by Cambridge University Press; 2011, 190-214.
32. Luenberger DG: **Iterative methods of optimization.** *Optimization by vector space methods* Edited by John Wiley & Sons, Inc; 1969, 271-311.
33. Tauler R, Lacorte S, Barceló D: **Application of multivariate curve self-modeling curve resolution for the quantitation of trace levels of organophosphorous pesticides in natural waters from interlaboratory studies.** *J of Chromatogr A* 1996, **730**:177-1.

doi:10.1186/1471-2105-15-S12-S8
Cite this article as: Cui *et al.*: An improved independent component analysis model for 3D chromatogram separation and its solution by multi-areas genetic algorithm. *BMC Bioinformatics* 2014 **15**(Suppl 12):S8.