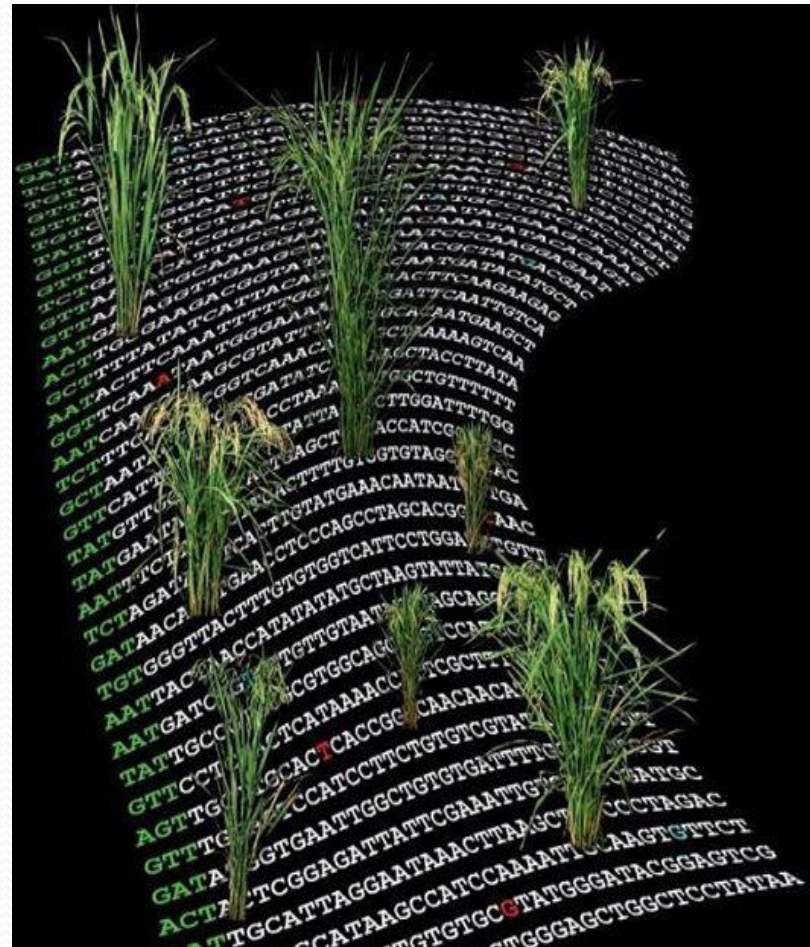


Một hướng tiếp cận tập thô để phân tích biểu hiện gene

Nguyễn Xuân Thảo, Ngọc Minh Châu

Bm toán tin ud – Khoa CNTT



Đặt vấn đề

TABLE 3.1 A Simplified DT Illustrating the Relationship of the Gene-Expression Profiles of Three Genes to the Presence or Absence of Phospholipidosis Induced by Eight Hypothetical Molecules^a

<i>Mol</i>	<i>gene-1</i>	<i>gene-2</i>	<i>gene-3</i>	<i>p</i>
1	↑	■	■	+
2	■	↓	↑	○
3	■	■	↑	+
4	↑	↑	↑	+
5	■	↓	↑	○
6	↑	■	■	○
7	■	↑	↑	+
8	■	↓	↑	○

Phân tích biểu hiện của *p* dựa trên các biểu hiện của {*gene-1*, *gene-2*, *gene-3*} → Luật (trị thức)

Đặt vấn đề_(tiếp)

- Loại bỏ gen không ảnh hưởng đến biểu hiện của p (nếu có)
- Rút trích các luật
- Ứng dụng (ra quyết định)

Nội dung

- 1. Giới thiệu
- 2. Bảng quyết định
- 3. Lý thuyết tập thô
- 4. Ví dụ
- 5. Rút trích luật
- 6. Biểu hiện gene do tác dụng của thuốc và biểu hiện Phospholipit trên tế bào ung thư gan HEPG2

1. Giới thiệu

- Một khía cạnh quan trọng của nghiên cứu gen chức năng liên quan đến việc xác định các mối quan hệ giữa mức độ biểu hiện của các gen và phản ứng sinh học cụ thể.
 - Về vấn đề này, một số nghiên cứu đã được tiến hành để đánh giá cơ sở di truyền cơ bản của phản ứng tế bào gây ra bởi các hóa chất khác nhau.
-
- Các dữ liệu từ nghiên cứu này có thể được biểu diễn dưới dạng một bảng quyết định (DT)
 - Vì số lượng gen thường được xét đến có xu hướng khá lớn, nó gây khó khăn cho việc xác định các tập con tối thiểu của các gen chịu trách nhiệm cho một phản ứng tế bào (?) (cellular response) hoặc chức năng sinh học.

1. Giới thiệu (tiếp)

- Lý thuyết tập thô (RST): là công cụ hữu hiệu xử lý các dữ liệu rời rạc dạng bảng quyết định(DT)
- RST cung cấp các thủ tục tính toán để xác định các tập con nhỏ nhất của các thuộc tính, duy trì các mối quan hệ quan trọng để hỗ trợ quá trình sinh các luật đơn giản.

2. Bảng quyết định

TABLE 3.1 A Simplified DT Illustrating the Relationship of the Gene-Expression Profiles of Three Genes to the Presence or Absence of Phospholipidosis Induced by Eight Hypothetical Molecules^a

<i>Mol</i>	<i>gene-1</i>	<i>gene-2</i>	<i>gene-3</i>	<i>p</i>
1	↑	■	■	+
2	■	↓	↑	○
3	■	■	↑	+
4	↑	↑	↑	+
5	■	↓	↑	○
6	↑	■	■	○
7	■	↑	↑	+
8	■	↓	↑	○

^aUp regulated genes are indicated by a upwards arrow (↑), down regulated genes by a downward arrow (↓), and normally expressed genes by a filled square (■). Phospholipidosis *p* is indicated by a plus (+) if it occurs and an open circle (○) if it does not occur.

3. Lý thuyết tập thô

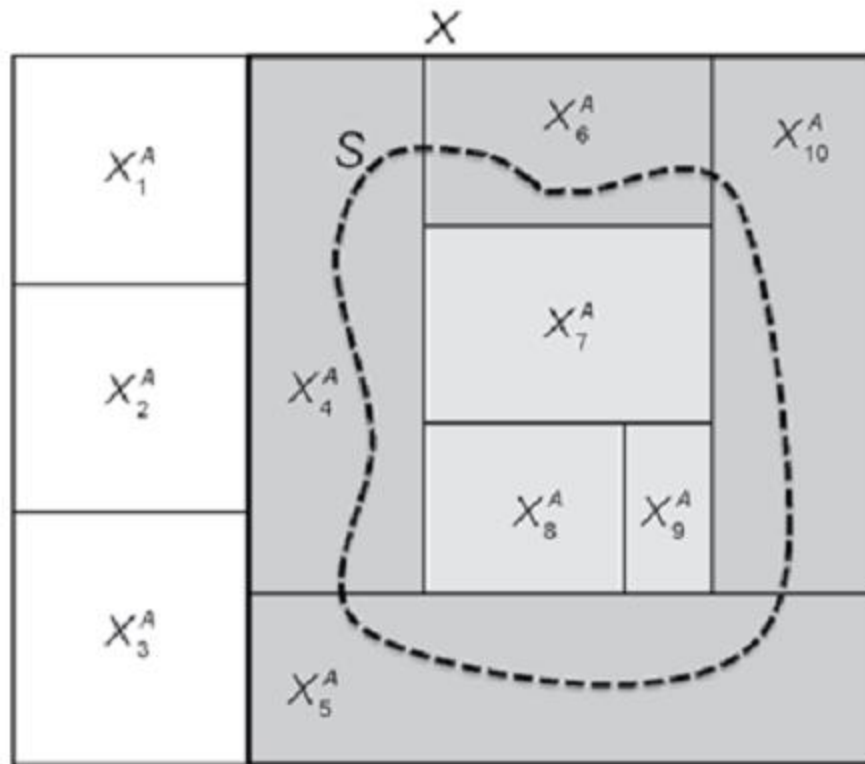
- Các hàng (rows): tập các đối tượng $X = \{x_1, x_2, \dots, x_N\}$
- Các cột (columns) : tập các thuộc tính $A = \{a_1, a_2, \dots, a_M\}$
- Trong bảng quyết định $A = C \cup D.$
- Dựa vào tập con A_k của A ta có sự phân lớp của X theo A_k

$$X \xrightarrow{A_k} X(A_k) = \{X_1^{A_k}, X_2^{A_k}, \dots, X_Q^{A_k}\} \quad (3.2.7)$$
- Tập xấp xỉ dưới của tập con S của X theo A_k

$$\underline{A}_k S = \{x \in X | X_\ell^{A_k} \subseteq S, \ell = 1, 2, \dots, Q\} \quad (3.2.10)$$
- Tập xấp xỉ trên của tập con S của X theo A_k

$$\bar{A}_k S = \{x \in X | X_\ell^{A_k} \cap S \neq \emptyset, \ell = 1, 2, \dots, Q\} \quad (3.2.11)$$
- Biên của S theo A_k là $B_{A_k}(S) = \bar{A}_k S - \underline{A}_k S.$

3. Lý thuyết tập thô (tiếp)



Đo độ chính xác và chất lượng của phân lớp

- Theo công thức (3.2.7) ta có

$$\underline{A}_k X(A_k) = \{ \underline{A}_k X_1^{A_k}, \underline{A}_k X_2^{A_k}, \dots, \underline{A}_k X_Q^{A_k} \} \quad (3.2.15)$$

và
$$\bar{A}_k X(A_k) = \{ \bar{A}_k X_1^{A_k}, \bar{A}_k X_2^{A_k}, \dots, \bar{A}_k X_Q^{A_k} \}. \quad (3.2.16)$$

là họ các lớp xấp xỉ trên và dưới của X dựa theo phân hoạch $X(A_k)$

- Độ chính xác

$$\text{Acc}[X(A_k)] = \frac{\sum_{L=1}^Q \text{Card}(\underline{A}_k X_L^{A_k})}{\sum_{L=1}^Q \text{Card}(\bar{A}_k X_L^{A_k})} \quad (3.2.17)$$

- Chất lượng

$$\text{Qual}[X(A_k)] = \frac{\sum_{L=1}^Q \text{Card}(\underline{A}_k X_L^{A_k})}{\text{Card}(X)}. \quad (3.2.18)$$

4. Ví dụ

TABLE 3.1 A Simplified DT Illustrating the Relationship of the Gene-Expression Profiles of Three Genes to the Presence or Absence of Phospholipidosis Induced by Eight Hypothetical Molecules^a

<i>Mol</i>	<i>gene-1</i>	<i>gene-2</i>	<i>gene-3</i>	<i>p</i>
1	↑	■	■	+
2	■	↓	↑	○
3	■	■	↑	+
4	↑	↑	↑	+
5	■	↓	↑	○
6	↑	■	■	○
7	■	↑	↑	+
8	■	↓	↑	○

^aUp regulated genes are indicated by a upwards arrow (↑), down regulated genes by a downward arrow (↓), and normally expressed genes by a filled square (■). Phospholipidosis *p* is indicated by a plus (+) if it occurs and an open circle (○) if it does not occur.

4.1 Phân tích bảng 3.1

- Đối tượng (rows) $Mol = \{1, 2, 3, 4, 5, 6, 7, 8\}$
- Thuộc tính đk (column) $G = \{gene-1, gene-2, gene-3\}$
- Phân lớp theo các thuộc tính điều kiện

$$X(G) = \{X_1^G, X_2^G, X_3^G, X_4^G, X_5^G\}$$

ta có bảng 3.2

- Phân lớp theo thuộc tính quyết định (biểu hiện của p)

$$X(D) = \{X_1^D, X_2^D\}$$

ta có bảng 3.3

4.2 Phân lớp bảng 3.1 theo thuộc tính đk

TABLE 3.2 The DT Derived from the Data Given in Table 3.1 with the Five Indiscernibility Classes X_1^G , X_2^G , X_3^G , X_4^G , and X_5^G Induced by the Set of Condition Attributes $G = \{gene - 1, gene - 2, gene - 3\}$.

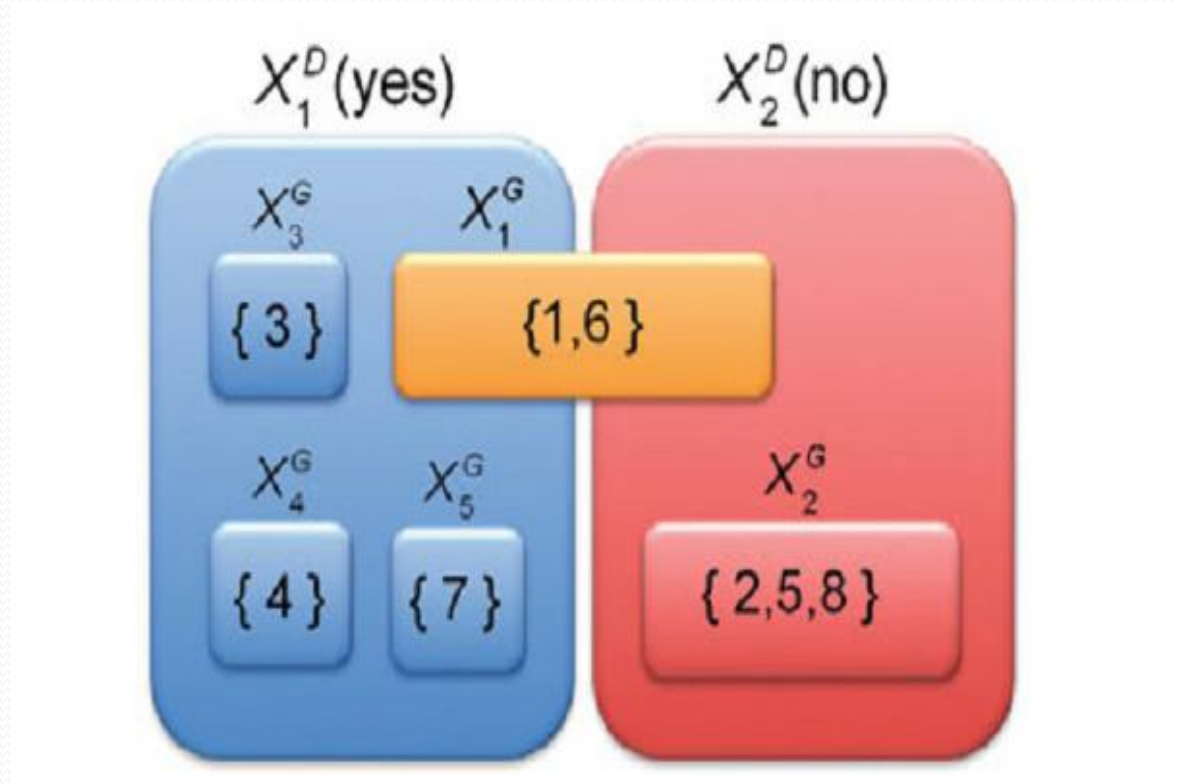
Class	<i>Mol</i>	<i>gene-1</i>	<i>gene-2</i>	<i>gene-3</i>	<i>p</i>
X_1^G	1	↑	■	■	+
	6	↑	■	■	○
X_2^G	2	■	↓	↑	○
	5	■	↓	↑	○
	8	■	↓	↑	○
X_3^G	3	■	■	↑	+
X_4^G	4	↑	↑	↑	+
X_5^G	7	■	↑	↑	+

4.3 Phân lớp bảng 3.1 theo thuộc tính qd

TABLE 3.3 The DT Obtained from the Data Given in Table 3.1 with the Two Indiscernibility Classes X_1^D and X_2^D Induced by the Decision Attribute Phospholipidosis p

Class	<i>Mol</i>	<i>gene-1</i>	<i>gene-2</i>	<i>gene-3</i>	<i>p</i>
X_1^D	1	↑	■	■	+
	3	■	■	↑	+
	4	↑	↑	↑	+
	7	■	↑	↑	+
X_2^D	2	■	↓	↑	○
	5	■	↓	↑	○
	6	↑	■	■	○
	8	■	↓	↑	○

4.4 Biểu hiện gene



- Độ chính xác phân lớp $\text{Acc}[X(D)] = 0.60$
- Chất lượng phân lớp $\text{Qual}[X(D)] = 0.75$

4.5 Thuộc tính dư thừa

- Thuộc tính dư thừa là thuộc tính không ảnh hưởng đến quá trình phân lớp của bảng quyết định
- Trong bảng qd 3.1 thì *gen-3* là dư thừa. Tập $\{gen-1, gen-2\}$ tạo thành các C-rút gọn và D-rút gọn.
- Các tập $\{gen-1, gen-3\}$, $\{gen-2, gen-3\}$ không tạo thành các C-rút gọn và D-rút gọn.

5. Rút trích luật

- Luật (R) có dạng **If A then B**
- Do {gen-1, gen-2} là tập rút gọn duy nhất nên rút trích luật ta dựa vào tập thuộc tính này.

TABLE 3.5 Rules Generated from the Example Presented in Sections 3.2.3–3.2.5

	Rules	Support
<i>Rule-1</i>	IF (<i>gene-2</i> , ↓) THEN (<i>phospholipidosis</i> , ●)	3
<i>Rule-2</i>	IF (<i>gene-2</i> , ↑) THEN (<i>phospholipidosis</i> , +)	2
<i>Rule-3</i>	IF (<i>gene-1</i> , ↑) AND (<i>gene-2</i> , ■) THEN (<i>phospholipidosis</i> , +) OR (<i>phospholipidosis</i> , ●)	1,1
<i>Rule-4</i>	IF (<i>gene-1</i> , ■) AND (<i>gene-2</i> , ■) THEN (<i>phospholipidosis</i> , +)	1

5. Rút trích luật (tiếp)

Class	Mol	gene-1	gene-2	gene-3	p
X_1^G	1	↑	■	■	+
	6	↑	■	■	0
X_2^G	2	■*	↓	↑	0
	5	■*	↓	↑	0
	8	■*	↓	↑	0
X_3^G	3	■	■	↑	+
X_4^G	4	↑*	↑	↑	+
X_5^G	7	■*	↑	↑	+

	Rules	Support
Rule-1	IF (gene-2, ↓) THEN (phospholipidosis, 0)	3
Rule-2	IF (gene-2, ↑) THEN (phospholipidosis, +)	2
Rule-3	IF (gene-1, ↑) AND (gene-2, ■) THEN (phospholipidosis, +) OR (phospholipidosis, 0)	1,1
Rule-4	IF (gene-1, ■) AND (gene-2, ■) THEN (phospholipidosis, +)	1

6. Biểu hiện gen do tác dụng của thuốc và p trên tế bào ung thư gan HEPG2

- Ứng dụng của RST vào một tập dữ liệu thực tế trên mối quan hệ giữa biểu hiện gen do thuốc tới sự biểu hiện của p trên tế bào ung thư gan [[xem 20](#)]
- Phản ứng của tế bào HEPG2 dựa trên 30 loại thuốc (bảng 3.6) với sự thay đổi mức độ biểu hiện trên 17 gen (bảng 3.8)
- Kết quả: 17 loại thuốc chia vào 8 lớp có biểu hiện P
13 loại thuốc chia vào 5 lớp không có biểu hiện p
- Bảng 3.7 liệt kê các tập hợp các gen đánh dấu 13 ứng cử viên ban đầu được đề xuất bởi Sawada et al. [20] cùng với kí hiệu của gen, sản phẩm gen, và các chức năng sinh học.

TABLE 3.6 Thirty Drugs That Induce Changes in Gene Expression Levels That May or May Not Be Associated with the Onset of Phospholipidosis [20]^a

Mol No.	Drug	Drug Class ^b	Phospholipidosis ^c	Mol No.	Drug	Drug Class ^b	Phospholipidosis ^c
1	Amitriptyline	AD	+++	16	Loratidine	AI	+
2	Chlorcyclizine	AI	+++	17	Pentamidine	AF	+
3	Fluoxetine	AD	+++	18	Acetaminophen	NSAID	–
4	Amiodarone	TH	++	19	Clarithromycin	AB	–
5	AY-9944	AC	++	20	Disopyramide	AA	–
6	Chlorpromazine	AD	++	21	Erythromycin	AB	–
7	Imipramine	AD	++	22	Flecainide	AA	–
8	Perhexiline	AG	++	23	Haloperidol	AP	–
9	Tamoxifen	AN	++	24	Levofloxacin	AB	–
10	Clozapine	AP	++	25	Ofloxacin	AB	–
11	Sertraline	AD	++	26	Procainamide	AA	–
12	Clomipramine	AD	+	27	Quinidine	AA	–
13	Thioridazine	AP	+	28	Sotalol	AA	–
14	Zimelidine	AD	+	29	Sulfamethazole	AB	–
15	Ketoconazole	AF	+	30	Sumatriptan	MH	–

^aNote that the degree of phospholipidosis indicated by the number of plus signs (“+”) is not considered in this work, only whether phospholipidosis is or is not induced by a given drug molecule. The minus sign (“–”) indicates that phospholipidosis did not occur.

^bAA, antiarrhythmic; AB, antibiotic; AC, anticholesteric; AD, antidepressant; AF, antifungal; AG, antianginal; AI, antiinflammatory; AN, antineoplastic; AP, antipsychotic; MH, migraine headaches; TH, thyroid.

^cThe degree of phospholipidosis induced by a given drug is indicated by the number of pluses (“+”) in the fourth and eighth columns.

Bảng 3.7 liệt kê các tập hợp các gen đánh dấu 13 ứng cử viên ban đầu được đề xuất bởi Sawada et al. [20] cùng với biểu tượng của họ gen, sản phẩm gen, và các chức năng sinh học.

TABLE 3.7 Candidate Marker Genes Proposed by Sawada et al. [20] and Their Gene Products.

Gene No.	Gene Symbol	Gene Regulation ^a	Gene Product	Biological Function
1	ASAH1 ★	↑	<i>N</i> -acylsphingosine amidohydrolase (acid ceramidase) 1	Lipid metabolism/phospholipid degradation
2	MGC4171 ★	↑	Hypothetic protein MGC4171	Lipid metabolism/phospholipid degradation
3	LSS ★	↑	Lanosterol synthase (2,3-oxidosqualene-lanosterol cyclase)	Lipid metabolism/cholesterol biosynthesis
4	NR0B2	↑	Nuclear receptor subfamily 0, group B, member 2	Lipid metabolism/regulation of cholesterol biosynthesis
5	PHYH	■	Phytanoyl-CoA hydroxylase (Refsum disease)	Lipid metabolism/fatty acid alpha-oxidation
6	FABP1 ★	↑	Fatty acid binding protein 1 (Liver)	Lipid metabolism/fatty acid transport
7	INHBE	↑	Activin beta E	Cell cycle, proliferation, death
8	P8 ★	↑	P8 protein (candidate of metastasis 1)	Cell cycle, proliferation, death
9	HPN ★	↑	Hepsin (transmembrane protease, serine 1)	Proteolysis and peptidolysis
10	SERPINA3 ★	↑	Serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member-3	Endopeptidase inhibition
11	ASNS	■	Asparagine synthase	Miscellaneous
12	C10orf10	↑	Chromosome 10 open reading frame 10	Unknown/EST
13	FLJ10055	↑	Hypothetical protein FLJ10055	Unknown/EST
14	FRCP1 ★	↑	Likely ortholog of mouse fibronectin type III repeat containing protein 1	Unknown/EST
15	AP1S1	■	Adaptor-related protein complex 1, sigma 1 subunit	Golgi vesicle
16	SLC2A3	↓	Solute carrier family 2 (facilitated glucose transporter), member 3	Transport
17	TAGLN ★	↓	Transgelin	Smooth muscle, cell specific cytoskeletal protein

Genes highlighted in gray were considered by Sawada et al. to be only weakly correlated with the presence of phospholipidosis. Genes marked by a star (★) were determined in this work to be associated with the onset of phospholipidosis. See Table 3.8 for detailed expression data on all 17 genes.

^aThe upward arrows (↑) or downward arrows (↓) correspond to genes that are primarily, but not exclusively, up regulated or down regulated, respectively. The filled square (■) denotes a gene that is generally normally expressed.

6.1 Tập dữ liệu

TABLE 3.8 Fold Change Expression Values (“mRNA Scores”) of Candidate Phospholipidosis Marker Genes in Human Hematoma HepG2 Cells Determined By Real Time PCR and Pathology Analysis [20]^{a,b,c}

Mol. No.	Gene Number																	Phospholipidosis
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
1	3.10	2.22	1.79	1.64	1.53	2.39	4.19	2.82	2.20	3.01	1.84	5.08	2.11	1.80	0.88	0.22	0.23	+++
2	3.23	2.43	2.98	2.46	0.67	2.62	1.68	1.85	2.97	2.23	1.41	3.24	2.47	2.96	0.74	0.20	0.30	+++
3	6.55	1.34	3.51	2.70	1.37	2.80	3.61	2.40	2.39	1.73	1.68	3.08	1.93	2.15	0.39	0.09	0.20	+++
4	1.69	1.32	1.82	1.35	0.93	2.01	1.92	1.88	2.15	1.17	1.16	1.12	1.56	1.07	0.64	0.35	0.56	++
5	4.71	1.63	4.78	2.71	0.73	3.16	0.89	1.22	2.68	1.97	1.05	3.32	1.68	1.45	0.63	0.22	0.21	++
6	1.29	1.74	1.99	1.20	1.69	2.56	2.72	2.51	1.70	1.94	1.39	2.95	1.90	1.70	1.19	0.43	0.65	++
7	4.52	1.81	2.69	3.41	1.37	2.45	1.87	2.00	2.32	2.12	1.57	3.36	2.11	1.80	0.56	0.15	0.15	++
8	3.12	1.93	3.10	2.09	0.60	2.49	1.17	1.42	2.22	2.50	1.46	2.55	1.98	3.42	0.65	0.22	0.26	++
9	4.50	1.60	3.01	2.56	0.52	1.90	0.81	1.21	2.19	2.12	1.36	2.02	1.85	1.53	0.66	0.24	0.27	++
10	2.75	1.88	2.69	1.97	0.87	3.89	5.25	3.39	2.35	1.96	2.45	5.15	3.34	1.92	0.71	0.24	0.24	++
11	3.08	1.94	3.91	1.79	2.62	4.14	5.19	3.88	3.07	2.69	2.56	4.44	2.38	3.49	1.01	0.14	0.37	++
12	0.82	1.41	1.52	0.98	1.63	2.36	2.75	2.63	2.39	2.03	1.25	2.08	3.83	1.46	0.61	0.63	0.78	+
13	1.36	1.32	2.03	0.77	1.64	5.07	3.48	3.14	2.85	2.46	0.91	2.02	5.79	1.57	0.63	0.38	0.46	+
14	2.91	1.12	1.05	1.33	1.14	1.42	1.77	1.52	1.74	1.95	1.58	2.67	2.24	1.71	1.05	0.70	0.59	+
15	1.47	1.53	4.33	2.36	1.45	5.30	3.09	2.59	2.71	2.37	3.40	2.29	2.24	1.95	1.04	0.48	0.71	+
16	1.23	1.57	1.54	1.63	1.20	1.58	2.21	1.81	1.52	1.83	2.23	0.79	1.42	1.12	1.11	0.46	0.44	+
17	0.93	0.73	0.52	1.83	0.94	0.15	10.7	4.89	2.50	2.12	7.72	2.88	1.82	2.49	0.93	0.50	0.68	+
18	1.18	0.75	0.91	0.69	0.55	1.11	1.19	0.91	1.10	0.99	1.14	0.94	1.13	1.02	0.90	1.50	0.93	○
19	1.55	1.18	0.65	0.78	0.59	0.91	1.51	1.10	0.96	1.39	1.02	0.93	1.48	1.00	0.79	0.86	0.86	○
20	1.82	1.21	1.14	0.63	1.04	1.02	1.18	1.39	1.12	0.99	0.83	0.90	0.94	0.98	1.08	1.03	0.99	○

(continued)

(↑) mRNA Scores greater than 1.50

(■) mRNA Scores that lie between 0.70 and 1.50

(↓) mRNA Scores less than 0.70

Tập dữ liệu_(tiếp)

TABLE 3.8 (Continued)

Mol. No.	Gene Number																	Phospholipidosis
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
21	0.90	0.86	0.87	0.90	0.77	0.73	0.86	0.84	0.87	0.85	1.10	0.71	0.78	0.92	1.03	0.73	0.85	●
22	2.42	1.57	1.22	1.02	1.04	1.50	2.25	1.82	1.26	1.60	2.54	1.73	2.10	1.65	1.13	0.46	0.86	●
23	0.77	1.24	1.52	1.49	0.92	1.50	1.01	1.50	1.43	1.31	1.25	1.38	1.32	0.93	0.95	0.44	0.52	●
24	1.10	0.79	0.75	0.79	0.46	0.88	0.99	0.80	0.83	0.92	0.80	0.83	0.70	0.83	0.80	1.07	0.90	●
25	1.17	0.77	0.78	0.81	0.75	0.75	0.79	0.74	0.72	0.82	0.67	0.79	0.86	0.68	0.79	0.94	0.74	●
26	0.86	1.08	0.90	0.74	0.67	1.00	0.89	0.91	1.08	0.96	0.89	0.98	1.78	0.73	0.87	0.88	0.85	●
27	1.43	1.20	1.30	1.03	0.77	1.38	1.73	1.20	1.21	1.22	2.01	1.06	1.27	1.26	0.96	0.37	0.63	●
28	1.39	0.72	0.88	0.63	0.54	0.98	1.12	0.84	0.79	0.78	0.67	0.80	0.84	0.85	0.78	1.53	0.80	●
29	1.26	0.78	0.82	0.82	0.52	0.85	0.95	0.87	0.89	0.98	0.95	0.70	0.84	0.87	0.92	1.11	0.99	●
30	1.35	0.99	1.23	0.63	0.84	1.01	1.12	1.05	1.16	0.95	0.97	1.05	1.34	1.26	1.22	0.99	1.06	●

^aCells shaded in dark gray represent over-expression of the gene (↑), uncolored cells represent normal expression of the gene (■), and cells colored light gray represent under-expression of the gene (↓).

^bSee Table 3.6 and Table 3.7 for the names of the molecules and genes, respectively, listed in this table and text for further details.

^cNote that the degree of phospholipidosis indicated by the number of plus signs (“+”) is not considered in this work, only whether phospholipidosis is or is not induced by a given drug molecule. The open circle (“○”) indicates that phospholipidosis did not occur.

6.2 Xác định mức độ biểu hiện gen

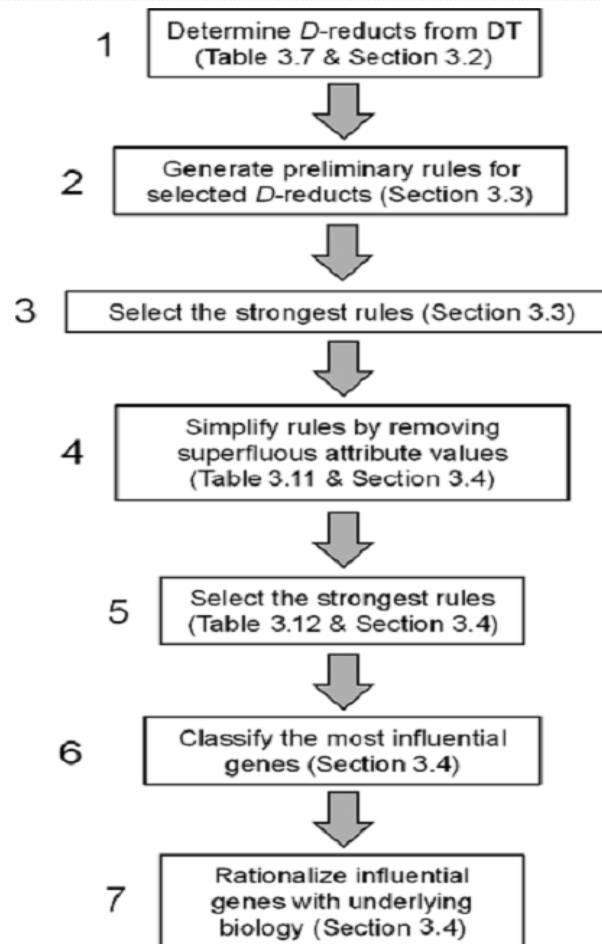


FIGURE 3.4 The process used in this work to identify gene expression levels associated with drug-induced phospholipidosis in HepG2 cells.

Bước1. Xác định các D-rút gọn

TABLE 3.9 The Set of 14 D-Reducts of Lengths $L=1,3$ Generated from the Data in Table 3.8

p	$D\text{-reducts, } D_p$	Length, L
1	{ gene-9 }	1
2	{ gene-6, gene-8, gene-17 }	3
3	{ gene-2, gene-6, gene-12 }	3
4	{ gene-2, gene-3, gene-8 }	3
5	{ gene-6, gene-14, gene-17 }	3
6	{ gene-1, gene-6, gene-17 }	3
7	{ gene-1, gene-10, gene-17 }	3
8	{ gene-1, gene-8, gene-17 }	3
9	{ gene-2, gene-6, gene-10 }	3
10	{ gene-6, gene-10, gene-17 }	3
11	{ gene-2, gene-6, gene-8 }	3
12	{ gene-2, gene-6, gene-14 }	3
13	{ gene-6, gene-13, gene-17 }	3
14	{ gene-6, gene-12, gene-17 }	3

$$X_1^{D_1} = X_1^{\{p\}} = \{1,2,\dots,17\}$$

$$X_2^{D_1} = X_2^{\{p\}} = \{18,19,\dots,30\}$$

Bước 2. Các luật ban đầu

TABLE 3.10 Eleven Top Rules Describing the Relationship of Gene-Expression Levels to Drug-Induced Phospholipidosis in Hepatoma HepG2 Cells^a

Rule No.	Rule	LHS Supp	RHS Supp	RHS Acc	LHS Cov	RHS Cov
1	IF (<i>gene-9</i> , $\uparrow$) THEN (<i>phospholipidosis</i> , $\oplus$)	17	17	1.00	0.57	1.00
2	IF (<i>gene-6</i> , $\uparrow$) AND (<i>gene-10</i> , $\uparrow$) AND (<i>gene-17</i> , $\downarrow$) THEN (<i>phospholipidosis</i> , $\oplus$)	12	12	1.00	0.40	0.71
3	IF (<i>gene-2</i> , $\uparrow$) AND (<i>gene-6</i> , $\uparrow$) AND (<i>gene-10</i> , $\uparrow$) THEN (<i>phospholipidosis</i> , $\oplus$)	11	11	1.00	0.37	0.64
4	IF (<i>gene-1</i> , $\uparrow$) AND (<i>gene-10</i> , $\uparrow$) AND (<i>gene-17</i> , $\downarrow$) THEN (<i>phospholipidosis</i> , $\oplus$)	10	10	1.00	0.33	0.59
5	IF (<i>gene-1</i> , $\uparrow$) AND (<i>gene-6</i> , $\uparrow$) AND (<i>gene-17</i> , $\downarrow$) THEN (<i>phospholipidosis</i> , $\oplus$)	10	10	1.00	0.33	0.59
6	IF (<i>gene-6</i> , $\uparrow$) AND (<i>gene-14</i> , $\downarrow$) AND (<i>gene-17</i> , $\downarrow$) THEN (<i>phospholipidosis</i> , $\oplus$)	10	10	1.00	0.33	0.59
7	IF (<i>gene-2</i> , $\uparrow$) AND (<i>gene-6</i> , $\uparrow$) AND (<i>gene-12</i> , $\uparrow$) THEN (<i>phospholipidosis</i> , $\oplus$)	10	10	1.00	0.33	0.59
8	IF (<i>gene-6</i> , $\uparrow$) AND (<i>gene-8</i> , $\uparrow$) AND (<i>gene-17</i> , $\downarrow$) THEN (<i>phospholipidosis</i> , $\oplus$)	10	10	1.00	0.33	0.59
9	IF (<i>gene-1</i> , $\uparrow$) AND (<i>gene-8</i> , $\uparrow$) AND (<i>gene-17</i> , $\downarrow$) THEN (<i>phospholipidosis</i> , $\oplus$)	8	8	1.00	0.27	0.47
10	IF (<i>gene-2</i> , $\uparrow$) AND (<i>gene-3</i> , $\uparrow$) AND (<i>gene-8</i> , $\uparrow$) THEN (<i>phospholipidosis</i> , $\oplus$)	8	8	1.00	0.27	0.47
11	IF (<i>gene-2</i> , $\uparrow$) AND (<i>gene-6</i> , $\uparrow$) AND (<i>gene-8</i> , $\uparrow$) THEN (<i>phospholipidosis</i> , $\oplus$)	8	8	1.00	0.27	0.47

$LHSSup(R) = |A|$, $RHSSup(R) = |B|$, $RHSAcc(R) = |B|/|A|$
 $LHSCov = |A|/30$; $RHSCov = |B|/17$

- Luật (R): If A then B

Bước 3. Chọn các luật mạnh nhất

- RHS Cov: đo mức độ mạnh của luật nên ta quan tâm đến hệ số này.
- Chọn các luật có RHS Cov ≥ 0.5
- Kết hợp với các D-reduct để tiên đoán các biểu hiện của p do tác dụng của thuốc
- Rule-2 và D-reduct D₁₀

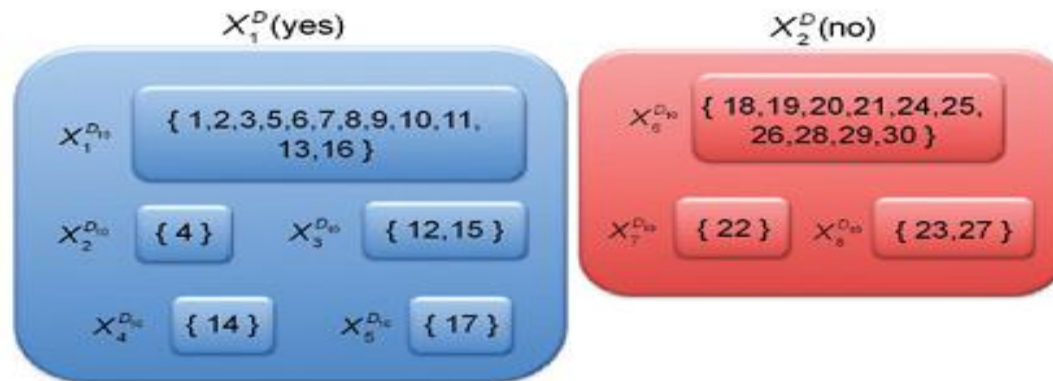


FIGURE 3.5 Schematic depiction of the indiscernibility classes induced by the *D*-reduct {*gene-6*, *gene-10*, *gene-17*} obtained from the hepatoma HepG2 cell gene expression data set presented in Section 3.3.1. The objects in the rectangles shaded in blue correspond to molecules that definitely induce phospholipidosis, while those rectangles shaded in red correspond to molecules that definitely do not induce phospholipidosis. In this case the boundary set is null, so all of the *C*-indiscernibility classes are subsets of the indiscernibility classes induced by the decision attribute that indicates the presence or absence of phospholipidosis in the hepatoma HepG2 cells. Hence, all of the rules generated from this partitioning are deterministic. For color details, please see color plate section.

Bước 4. Chọn luật đơn giản, loại các thuộc tính dư thừa

TABLE 3.11 Rules Derived from the Tenth D -Reduct, D_{10} , in Table 3.10 after Reduction of the Attribute Values^a

Rule No.	Rule	LHS Supp	RHS Supp	RHS Acc	LHS Cov	RHS Cov
1	IF (<i>gene-6</i> , ■) AND (<i>gene-10</i> , ■) THEN (<i>phospholipidosis</i> , ●)	12	12	1.00	0.400	0.923
2	IF (<i>gene-6</i> , ↑) THEN (<i>phospholipidosis</i> , +)	15	15	1.00	0.500	0.882
3	IF (<i>gene-6</i> , ■) AND (<i>gene-17</i> , ■) THEN (<i>phospholipidosis</i> , ●)	11	11	1.00	0.367	0.846
4	IF (<i>gene-10</i> , ↑) AND (<i>gene-17</i> , ↓) THEN (<i>phospholipidosis</i> , +)	14	14	1.00	0.467	0.824
5	IF (<i>gene-10</i> , ■) AND (<i>gene-17</i> , ■) THEN (<i>phospholipidosis</i> , ●)	10	10	1.00	0.333	0.769
6	IF (<i>gene-6</i> , ↓) THEN (<i>phospholipidosis</i> , +)	1	1	1.00	0.333	0.059

^aUp regulated genes are indicated by an upward arrow (↑), down regulated genes by a down arrow (↓), and normally expressed genes by a filled square (■). A plus sign (+) indicates that phospholipidosis has occurred; while an open circle (●) indicates that it has not occurred.

which give rise to the following values for the rule characteristics

$$\begin{aligned}
 \text{LHS Supp} &= \text{Card}(X_1^{D_{10}}) = 12 \\
 \text{RHS Supp} &= 12 \\
 \text{RHS Acc} &= \text{LHSSupp} / \text{RHSSupp} = 12 / 12 = 1.00 \\
 \text{LHS Cov} &= \text{LHSSupp} / \text{Total Drugs} = 12 / 30 = 0.40 \\
 \text{RHS Cov} &= \text{Card}(X_1^{D_{10}}) / \text{Card}(X_1^{[p]}) = 12 / 17 = 0.71
 \end{aligned}
 \tag{3.3.7}$$

Bước 5. chọn các luật mạnh nhất_(tiếp)

TABLE 3.12 Most Important Rules for the Induction of Phospholipidosis*

Rule No.	Rules	RHS Cov
1	IF (<i>gene-9</i> , ↑) THEN (<i>phospholipidosis</i> , +)	1.00
2	IF (<i>gene-6</i> , ↑) THEN (<i>phospholipidosis</i> , +)	0.88
3	IF (<i>gene-10</i> , ↑) AND (<i>gene-17</i> , ↓) THEN (<i>phospholipidosis</i> , +)	0.82
4	IF (<i>gene-14</i> , ↑) AND (<i>gene-17</i> , ↓) THEN (<i>phospholipidosis</i> , +)	0.71
5	IF (<i>gene-8</i> , ↑) AND (<i>gene-17</i> , ↓) THEN (<i>phospholipidosis</i> , +)	0.71
6	IF (<i>gene-3</i> , ↑) AND (<i>gene-8</i> , ↑) THEN (<i>phospholipidosis</i> , +)	0.71
7	IF (<i>gene-1</i> , ↑) AND (<i>gene-17</i> , ↓) THEN (<i>phospholipidosis</i> , +)	0.65
8	IF (<i>gene-2</i> , ↑) AND (<i>gene-3</i> , ↑) THEN (<i>phospholipidosis</i> , +)	0.65

*Up regulated genes are indicated by an upward arrow (↑) and down regulated genes by a downward arrow (↓). A plus sign (+) indicates that phospholipidosis has occurred; while an open circle (○) indicates that it has not occurred.

Bước 6+7. Phân lớp các gen quan trọng nhất

TABLE 3.13 Approximate Ranking of Marker Genes Determined in This Work (See Table 3.7 for Additional Details)

Rank	Gene Number	Regulation	Gene Symbol
1	9	↑	HPN
2	6	↑	FABP1
3	17	↓	TAGLN
4	10	↑	SERPINA3
5	8	↑	P8
6	14	↑	FRCP1
7	3	↑	LSS
8	1,2	↑,↑	ASAH1, MGC4171



Trân trọng cảm ơn!

Tài liệu tham khảo

- [1]. JOACHIM PETIT, NATHALIE MEURICE, JOSÉ LUIS MEDINA-FRANCO, and GERALD M. MAGGIORA, A rough set theory approach to analysis of genes expression profiles, Chemoinformatics for Drug Discovery, First Edition. Edited by Jürgen Bajorath. © 2014 John Wiley & Sons, Inc. Published 2014 by John Wiley & Sons, Inc.
 - [2]. Sawada H, Takami K, Asahi S. A toxicogenomic approach to drug-induced phospholipidosis: Analysis of its induction mechanism and establishment of a novel in vitro screening system. Toxicol Sci 2005;83:282–292.
 - [3]. Ripphausen P, Nisius B, Bajorath J. State-of-the-art in ligand-based virtual screening, Drug Discov Today 2011; 16:372–376.
 - [4]. Komorowski J, Polkowski L, Skowron A. Rough sets: a tutorial. Synthesis 1999;46:2–8.
-
- 1. Heijne WHM, Lamers R-JAN, van Bladeren PJ, et al. Profiles of metabolites and gene expression in rats with chemically-induced hepatic necrosis. Toxicol Pathol 2005;33: 425–433.
 - 2. Leavitt J, Goldman D, Merrill C, et al. Changes in gene expression accompanying chemically-induced malignant transformation of human fibroblasts. Carcinogenesis 1982;3:61–70.
 - 3. Lu J, Pei H, Kaeck M, et al. Gene expression changes associated with chemically induced rat mammary carcinogenesis. Mol Carcinog 1997;20:204–215.
 - 4. Lambert CB, Spire C, Renaud M-P, et al. Reproducible chemical-induced changes in gene expression profiles in human hepatoma HepaRG cells under various experimental conditions. Toxicol In Vitro 2009;23:466–475.
 - 5. Moggs JG, Tinwell H, Spurway T, et al. Phenotypic anchoring of gene expression changes during estrogen-induced uterine growth. Environ Health Perspect 2004;112:1589–1606.
 - 6. Pawlak Z. *Rough Sets – Theoretical Aspects of Reasoning About Data*. Dordrecht: Kluwer Academic Publishers; 1991.
 - 7. Qian Y, Liang J, Dang C. Converse approximation and rule extraction from decision tables in rough set theory. Comput Mathematics 2008;55:1754–1765.

Tài liệu tham khảo (tiếp)

8. Nasiri JH, Mashinchi M. Rough set and data analysis in decision tables. *J Uncertain Syst* 2009;3:232–240.
9. Xiao J-M. New rough set approach to knowledge reduction in decision table[s]. *Proc Int Conf Mach Learn Cybern* 2004;4:2208–2211.
10. Ziarko W. On learnability of decision tables. *Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing; Uppsala*. New York: Springer Verlag; 2004 p 394–401. *Lectures Notes in AI* 3066.
11. Cios K, Pedrycz W, Swimiarski R. Rough sets. *Data Mining—Methods for Knowledge Discovery*. Dordrecht: Kluwer Academic Publishers; 1998.
12. Midelfart H, Kormorowski J, Norsett K, et al. Learning rough set classifiers from gene expressions and clinical data. *Fund Inform* 2002;53:155–183.
13. Hvidsten TR, Lægreid A, Komorowski J. Learning rule-based models of biological process from gene expression time profiles using Gene Ontology. *Bioinformatics* 2003;19: 1116–1123.
14. Lee JA, Verleysen M. *Nonlinear Dimensionality Reduction*. New York: Springer; 2007.
15. Quinlan JR. Induction of decision tress. *Mach Learn* 1986;1:81–106.
16. Breiman L, Friedman JH, Olshen RA, et al. *Classification and Regression Trees*. Monterey: Wadsworth & Brooks/Cole Advanced Books & Software; 1984.
17. Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press; 2000.
18. Saaty TL. *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process*. Pittsburg: RWS Publications; 2000.
19. Bender EA. *Mathematical Methods in Artificial Intelligence*. Los Alamitos: IEEE Computer Society Press; 1996.

Tài liệu tham khảo (tiếp)

20. Sawada H, Takami K, Asahi S. A toxicogenomic approach to drug-induced phospholipidosis: Analysis of its induction mechanism and establishment of a novel *in vitro* screening system. *Toxicol Sci* 2005;83:282–292.
21. Reasor MJ, Hastings KL, Ulrich RG. Drug-induced phospholipidosis: issues and future directions. *Expert Opin Drug Safety* 2006;5:567–583.
22. Nonoyama T, Fukuda R. Drug induced phospholipidosis pathological aspects and its prediction. *J Toxicol Pathol* 2008;21:9–24.
23. Tengstrand-Baronas E, Lee JW, Alden C, et al. Biomarkers to monitor drug-induced phospholipidosis. *Toxicol Appl Pharmacol* 2007;218:72–78.
24. Occam's razor, Wikipedia: http://en.wikipedia.org/wiki/Occam's_razor. Accessed 2012 May 10.
25. Fakih SJ, Das TK. LEADF: a methodology for learning efficient approaches to medical diagnosis. *IEEE Trans Inf Technol Biomed* 2006;10:220–228.
26. Collette TW, Szladow AJ. Use of rough sets and spectral data for building predictive models of reaction rate constants. *Appl Spectrosc* 1994;48:1379–1386.
27. Hashemi RR, Young JF. The prediction of methylmercury elimination half-life in humans using animal data: a neural network/rough sets analysis. *J Toxicol Environ Health, Part A* 2003;66:2227–2252.
28. Chèvre N, Gagné F, Gagnon P, et al. Application of rough sets analysis to identify polluted aquatic sites based on a battery of biomarkers: a comparison with classical methods. *Chemosphere* 2003;51:13–23.