RAIRO. RECHERCHE OPÉRATIONNELLE

J. L. SOURROUILLE

# Identification d'un phénomène bactériologique

*RAIRO. Recherche opérationnelle*, tome 17, n° 1 (1983), p. 43-61

<http://www.numdam.org/item?id=RO_1983__17_1_43_0>

R.A.I.R.O. Recherche opérationnelle/Operations Research
(vol. 16, n° 1, février 1982, p. 43 à 62.

# IDENTIFICATION D'UN PHÉNOMÈNE BACTÉRIOLOGIQUE (*)

par J. L. SOURROUILLE (1)

Résumé. — *Pour mettre en évidence la présence des systèmes de restriction/modification, la voie biochimique techniquement possible demanderait un investissement en temps et en travail très important. Une autre voie choisie ici, moins précise mais plus rapide, consiste à identifier le phénomène concerné. Ce texte présente une étude du modèle débouchant sur les conditions d'existence d'une solution et propose des méthodes d'identification adaptées aux cas réels rencontrés.*

Mots clés : Système d'équations non linéaires à variables mixtes ; équations pseudo-booléennes ; intervalles ; systèmes de restriction/modification.

Abstract. — *In order to exhibit the presence of the modification/restriction systems, the biochemic way, technically possible, would require a very large time and work investment. An other method chosen here, less precise but quicker, is to identify directly the concerned phenomenon. This paper presents a study of the model giving the existence conditions of a solution and proposes identification methods for observed concrete cases.*

Keywords: Mixed integer nonlinear system of equations; pseudo-booleans equations; intervals; modification/restriction systems.

## INTRODUCTION

Dans l'industrie laitière, les bactéries sont sélectionnées en fonction du produit fini à obtenir, mais aussi de leur résistance aux phages (virus), leurs parasites naturels. Les bactéries possèdent plusieurs moyens de défense, et une première étude a conclu que le principal moyen exploitable serait axé sur les systèmes de restriction/modification ($R/M$). Sur la base de faits connus et d'hypothèses de travail, un modèle mathématique prenant en compte les aspects quantitatifs les plus discriminants du phénomène a été retenu.

(*) Reçu octobre 1980.
(1) Département informatique appliquée, I.N.S.A., 20, avenue Albert-Einstein, 69621 Villeurbanne.

R.A.I.R.O. Recherche opérationnelle/Operations Research, 0399-0559/1983/43/$ 5.00

Une brève description du phénomène bactériologique précède la première partie de ce texte consacrée à l'étude du modèle mathématique. La détermination des paramètres de ce modèle conduit à résoudre un système d'équations non linéaires à variables discrètes et continues. L'examen des propriétés des solutions permet, par l'adjonction de contraintes, de remplacer le problème général par un problème équivalent admettant un nombre fini de solutions. Nous présentons ensuite les conditions d'existence d'une solution avant d'aborder la résolution proprement dite.

La qualité et la quantité des données recueillies rendent tellement différents les contextes de l'identification qu'une seule méthode de résolution ne peut être uniformément efficace. Nous proposons dans une deuxième partie diverses techniques permettant de faire face aux situations rencontrées pratiquement, accompagnées d'une brève description de l'application réalisée.

## LE PHÉNOMÈNE

D'une manière très schématique, et en l'absence de tout autre moyen de défense, les hypothèses faites sur l'action des systèmes $R/M$ peuvent se résumer ainsi :

Soit une population de bactéries $A_x$ ne possédant qu'un seul système $R/M$ $x$;

— en présence d'une population de phages non protégés contre ce système $R/M$ $x$, une proportion $p_x$ très petite de phages se développent; ils deviennent alors protégés contre ce système $R/M$ $x$;

— en présence d'une population de phages protégés contre ce système $R/M$ $x$, les phages se développent et détruisent, dans certaines conditions et limites, les bactéries qu'ils atteignent.

Chaque bactérie peut posséder plusieurs systèmes $R/M$ et chaque phage peut être protégé contre plusieurs systèmes $R/M$; l'action d'un système $R/M$ est considérée indépendante des populations de bactéries et de phages mises en présence.

Supposons par exemple qu'un procédé de fabrication utilise des bactéries $A_{xy}$ ne possédant que les ystèmes $R/M$ $x$ et $y$. Si un phage attaque victorieusement l'une de ces bactéries, il deviendra protégé contre les systèmes $x$ et $y$ et pourra détruire les autres bactéries. En utilisant à la place de $A_{xy}$ un mélange de bactéries $A_{xy}$ et $B_{tuv}$, les phages victorieux de $A_{xy}$ seront en très grande majorité détruits par $B_{tuv}$. Divers mécanismes peuvent ainsi être imaginés pour augmenter la sécurité de fabrication, mais il faut pour cela choisir convenablement les bactéries en fonction des systèmes $R/M$ possédés.

Sur les bases de données expérimentales obtenues en laboratoire ([2]) et d'un modèle qualitatif, deux sociétés ([3]) interessées par ces questions nous ont proposé d'identifier ce procédé de défense des bactéries. Nous ne présentons pas la justification du modèle, celle-ci relevant plutôt de la bactériologie.

L'étude du modèle mathématique n'étant pas particulière à l'application, nous appellerons par la suite « individus » les bactéries, « caractères » les systèmes $R/M$ tandis que les « coefficients » caractériseront la proportion $p_x$ de phages survivant à la réaction phages/bactéries décrite ci-dessus.

## I. DÉFINITION DU PROBLÈME

### I.1. Formulation générale

L'identification des paramètres du modèle mathématique retenu pose le problème suivant :

PROBLÈME 1 : Partant de données $d_{ij} \in \mathbb{R}^+$ connues pour certains couples $(i, j)$ d'individus $(i \neq j)$ et assimilées à une somme pondérée de caractères possédés par $i$ sans l'être par $j$, il faut trouver pour un échantillon de $m$ individus :

- les données $d_{ij} \in \mathbb{R}^+$ inconnues;
- le nombre $n$ de caractères explicatifs à retenir;
- les coefficients attachés à chaque caractère :

$$C = \{ c_p \in [\mathbb{R}^+ - 0];\ p = 1 \text{ à } n \};$$

- les caractères possédés par chacun des individus.

En posant :

$$\begin{cases} x_{ip} = 1 & \text{si l'individu } i \text{ possède le caractère } p \\ x_{ip} = 0 & \text{sinon,} \end{cases}$$

ce dernier point revient à déterminer :

$$X = \{ x_{ip} \in \{ 0, 1 \};\ p = 1 \text{ à } n,\ i = 1 \text{ à } m \}$$

tels que :

$$\sum_{p=1}^{n} c_p x_{ip} (1 - x_{jp}) = d_{ij}. \qquad (1)$$

([2]) Air Liquide.
([3]) Akzo et Vitex.

## I.2. Interprétation sur un exemple

Supposons connus pour un échantillon de trois individus le nombre $n$ de caractères, le vecteur $C$ des coefficients et la matrice $X$ de description des caractères possédés par les individus ( $\{ n, C, X \}$ ). Il apparaît sur la figure 1 que :

$$\sum_{p=1}^{n} c_p x_{ip}(1-x_{jp}) = d_{ij}$$

représente la somme des coefficients des caractères possédés par l'individu $i$ que l'individu $j$ ne possède pas. Il est visible sur cet exemple qu'un jeu de paramètres $\{ n, C, X \}$ ne conduit qu'à un seul tableau $D$. Par contre le problème inverse — partir de $D$ afin de déterminer $\{ n, C, X \}$ — conduit à une infinité de solutions.

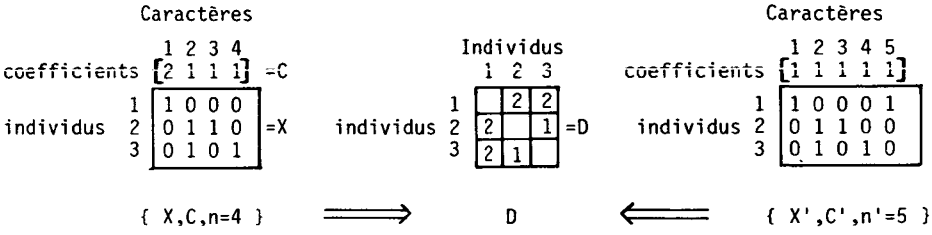


**Figure 1. Exemple de deux solutions conduisant au même tableau de données D.**

## I.3. Propriétés des solutions

De fait, toutes ces solutions sont d'un intérêt inégal. Afin de n'obtenir que des solutions distinctes, nous allons ajouter au problème initial des contraintes (ou conditions) qui en limiteront le nombre.

CONDITION 1 : *Aucune solution ne doit comporter de caractère possédé par tous les individus ou par aucun, car il est indétectable à travers les mesures.*

S'il existe $q$ tel que :

$$x_{iq} = 1 \text{ (ou } x_{iq} = 0), \qquad \forall i = 1 \text{ à } m,$$

alors :

$$x_{iq}(1-x_{jq}) = 0, \qquad \forall i, j = 1 \text{ à } m,$$

d'où :

$$\sum_{\substack{p=1 \\ p \neq q}}^{n} c_p x_{ip}(1-x_{jp}) = \sum_{p=1}^{n} c_p x_{ip}(1-x_{jp}), \qquad \forall i, j = 1 \text{ à } m.$$

Pour éliminer ces caractères indésirables il suffit d'introduire des contraintes de la forme :

$$\left\{ \begin{array}{l} \exists i \in \{ 1, \ldots, m \} : x_{ip} = 0 \\ \exists i \in \{ 1, \ldots, m \} : x_{ip} = 1 \end{array} \right\}, \qquad \forall p = 1 \text{ à } m. \tag{2}$$

CONDITION 2 : *Aucune solution ne doit comporter de caractères ayant la même répartition dans l'échantillon d'individus.*

Si dans une solution le caractère $u$ est remplacé par deux caractères $v$ et $r$ tels que :

$$\left\{ \begin{array}{l} c_v + c_r = c_u, \\ x_{iv} = x_{ir} = x_{iu}, \end{array} \right. \qquad \forall i = 1 \text{ à } m,$$

une nouvelle solution est obtenue car $\forall i, j = 1$ à $m$ :

$$\sum_{\substack{p=1 \\ p \neq u}}^{n} c_p x_{ip} (1 - x_{jp}) + c_v x_{iv} (1 - x_{jv}) + c_r x_{ir} (1 - x_{jr}) = \sum_{p=1}^{n} c_p x_{ip} (1 - x_{jp}).$$

Les données ne permettant pas de différencier ces deux solutions, il ne faut chercher que les solutions dont le nombre de caractères est minimal, d'où la contrainte :

$$\exists i \in \{ 1, \ldots, m \}; \qquad x_{ip} \neq x_{iq}, \forall \underset{p \neq q}{p, q} = 1 \text{ à } m. \tag{3}$$

CONDITION 3 : *Les différentes solutions obtenues ne doivent pas se déduire l'une de l'autre par permutation des indices des caractères, le choix des indices des caractères étant arbitraire.*

### I.4. Problème modifié

L'ajout de ces trois contraintes au problème initial nous définit un nouveau problème que nous appellerons problème 2. Ces restrictions sur les solutions du problème 1 ne peuvent être considérées comme limitatives car il est clair que la connaissance du tableau $D$ ne suffit pas à la détermination d'un jeu unique de paramètres $\{ n, C, X \}$ d'une part, et d'autre part que le noyau de solutions trouvé, en y adjoignant les conditions précédentes, permet de bâtir toute solution du problème 1.

## II. CONDITIONS D'EXISTENCE D'UNE SOLUTION

### II.1. Représentation générale des solutions : notations

A chaque caractère $q$ correspond un vecteur binaire :

$$x_{.q} = \{ x_{iq}; \ i = 1 \text{ à } m \}.$$

Pour un échantillon de $m$ individus, le nombre de vecteurs binaires possible est $2^m-2$ puisque la condition 3 implique que deux vecteurs ne peuvent être identiques et que la condition 2 élimine les deux vecteurs formés uniquement de 0 et de 1. La figure 2 présente par exemple une solution portant sur trois individus.

| | | Caractères | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| Coefficients | | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ |
| Individus | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| | 2 | 0 | 1 | 1 | 0 | 0 | 1 |
| | 3 | 1 | 0 | 1 | 0 | 1 | 0 |

**Figure 2. Représentation générale d'une solution pour trois individus.**

Pour plus de commodité, toute expression telle que :

$$\sum_{p=1}^{n} c_p x_{ip}(1-x_{jp}) x_{kp}(1-x_{hp})$$

sera écrite sous la forme : $x_i \bar{x}_j x_k \bar{x}_h$, les variables étant surlignées lorsqu'elles apparaissent en complément à 1 dans l'expression.

Remarquons que $x_1 \bar{x}_2 = x_1 \bar{x}_2 x_3 + x_1 \bar{x}_2 \bar{x}_3$.

La connaissance des $2^m-2$ produits de $m$ variables pour un échantillon de $m$ individus permet de déterminer complètement le jeu de paramètres $\{ n, C, X \}$. Si par exemple (*fig.* 2) $\bar{x}_1 \bar{x}_2 x_3$ prend la valeur $v \in [\mathbb{R}^+ - 0]$, cela signifie qu'il existe un caractère d'indice arbitraire 1 possédé par le seul individu 3 et de coefficient $c_1 = v$.

## II.2. Résolution analytique d'un système d'équations particulier

Pour déterminer les conditions d'existence d'une solution, il est nécessaire de résoudre un système d'équations indéterminé dont les variables sont soumises à des contraintes. Nous verrons que les particularités de ce système permettent d'obtenir l'ensemble des solutions, généralement définies par des intervalles.

### II.2.1. *Résolution dans l'ensemble des réels positifs*

Soit à résoudre :

$$\begin{cases} a_i + a_j = b_{ij}, & a_i,\ a_j,\ b_{ij} \in \mathbb{R}^+ \\ \text{où les inconnues sont } a_i \text{ et } a_j & i, j = 1 \text{ à } k \\ & i \neq j \end{cases} \tag{4}$$

$b_{ij}$ n'étant pas obligatoirement connu pour tous les couples $(i, j)$, le nombre d'équations est quelconque et au maximum de $k(k-1)/2$.

Représentons ce système d'équations par un graphe symétrique dont les sommets $i$ prennent les valeurs des variables $a_i$ ; une arête marquée $b_{ij}$ joint les sommets $i$ et $j$ lorsqu'il existe une équation $a_i + a_j = b_{ij}$.

PROPOSITION 1 : *Si le graphe possède un cycle de longueur impaire (i. e. dont le nombre d'arêtes est impair), toutes les variables de la composante connexe du graphe à laquelle appartient ce cycle peuvent être déterminées.*

Soit 1, 2, . . ., $2p+1$ les indices des sommets d'un cycle de longueur impaire :

$$a_1 + a_2 = b_{1,2}$$

$$a_2 + a_3 = b_{2,3}$$

. . . . . .

$$a_{2p}{}_{f0013+} a_{2p+1} = b_{2p,2p+1}$$

$$a_{2p+1} + a_1 = b_{2p+1,1}$$

En effectuant la « somme alternée » des équations on obtient :

$$a_1 = \frac{1}{2}(b_{1,2} - b_{2,3} + \ldots - b_{2p,2p+1} + b_{2p+1,1}).$$

De proche en proche, toutes les variables du cycle peuvent être déterminées.

PROPOSITION 2 : *Si le graphe possède un cycle de longueur paire, la somme alternée des arêtes calculée le long de ce cycle doit être nulle. Si cette relation est vérifiée, l'une quelconque des équations de ce cycle peut être éliminée.*

Soit 1, 2, . . ., $2p$ les indices des sommets d'un cycle de longueur paire. En effectuant comme ci-dessus la somme alternée des équations de la 1re à la $2p^{\text{ieme}}$ on obtient :

$$b_{1,2} - b_{2,3} + \ldots - b_{2p,1} = 0.$$

Il y a donc une équation redondante et n'importe quelle équation parmi les $2p$ du cycle peut être supprimée.

PROPOSITION 3 : *Dans un graphe sans cycle, l'intervalle* $[m_i, M_i]$ *où la variable* $a_i$ *prend ses valeurs se calcule comme suit :*

$m_i$ *(resp.* $M_i$*) est la plus grande (resp. petite) valeur obtenue le long des chaînes simples d'origine i et de longueur paire (resp. impaire). Pour chaque chaîne de* $2p$ *(resp.* $2p+1$*) arêtes :*

— *on calcule la somme relative aux arêtes de rang impair de la* 1re *à la* $2p$*-1-ième (resp.* $2p+1$*-ième) ;*

— *de cette somme on soustrait la somme relative aux arêtes de rang pair de la 2-ième à la 2p-ième.*

*La chaîne de longueur nulle est considérée de longueur paire et de somme nulle. Si $m_i > M_i$, il n'y a pas de solution.*

Soit $G$ l'ensemble des sommets d'un graphe connexe sans cycle et $P_j$ (resp. $I_j$) l'ensemble des sommets reliés au sommet $j$ du graphe par une chaîne simple de longueur paire ou nulle (resp. impaire).

Appelons $S_j^p$ la somme alternée des arêtes calculée le long d'une chaîne simple commençant au sommet $j$ et se terminant au sommet $p$ :

$$S_j^p = b_{jk} - b_{kt} + \ldots - b_{sp} \qquad \text{pour} \quad p \in P_j$$

et posons $S_j^j = 0$. Avec ces notations, l'intervalle défini par la proposition 3 s'écrit :

$$[m_i = \max_{p \in P_i} (S_i^p),\ M_i = \min_{p \in I_i} (S_i^p)].$$

Le système (4) admettra une solution si :

$$\begin{cases} m_i \geqq 0, \quad \forall i \in G, \\ \forall a_j \in [m_j, M_j] : \exists a_i \in [m_i, M_i] : \qquad a_i + a_j = b_{ij} \end{cases}$$

pour toutes les arêtes $(i, j)$ du graphe.

Ceci peut encore s'écrire :

$$\begin{cases} M_i \geqq m_i \geqq 0, \qquad \forall i \in G, \\ m_i + M_j = m_j + M_i = b_{ij} \text{ pour toutes les arêtes (i, j) du graphe.} \end{cases}$$

(*a*) Si $a_i < m_i$ (ou $a_i > M_i$), il n'y a pas de solution car on sait que :

$$S_i^p = a_i - a_p \qquad \text{pour} \quad p \in P_i$$

d'où :

$$a_i < \max_{p \in P_i} (S_i^p) = \max_{p \in P_i} (a_i - a_p)$$

ou encore :

$$\min_{p \in P_i} (a_p) < 0,$$

ce qui est contraire aux hypothèses.

ce qui est contraire aux hypothèses.

(*b*) Si $a_i \in [m_i, M_i]$ alors :

$$\rightarrow M_i \geqq m_i \geqq 0, \qquad \forall i \in G,$$

car :

$$m_i = \max_{p \in P_i} (S_i^p) \quad \text{donc} \quad m_i \geqq S_i^i = 0 :$$

$$\rightarrow m_i + M_j = b_{ij} (= M_i + m_j),$$

car :

$$m_i + M_j = \max_{p \in P_i} (S_i^p) + \min_{p \in I_j} (S_j^p)$$

pour $i$ et $j$ adjacents, $S_j^p = b_{ij} - S_i^p$ et $I_j \equiv P_i$ d'où :

$$m_i + M_j = \max_{p \in P_i} (S_i^p) + \min_{p \in P_i} (b_{ij} - S_i^p) = b_{ij}.$$

REMARQUE : Lorsque dans une composante connexe d'un graphe, l'intervalle $[m_i, M_i]$ est déterminé pour un sommet $i$, ceux de tous les autres sommets de cette composante existent et peuvent être déterminés de proche en proche par la formule ci-dessus.

II . 2 . 2. *Extension aux intervalles*

Partant de $b_{ij}$ réels positifs, les solutions sont décrites par des intervalles. Lorsque les $b_{ij}$ sont eux mêmes des intervalles, les solutions doivent être décrites par les intervalles les plus serrés possible. Alors les propositions 1, 2 et 3 ne présentent plus d'intérêt, des résultats plus fins pouvant être obtenus en utilisant la proposition suivante où $b_{\text{inf}}$ et $b_{\text{sup}}$ sont respectivement la borne inférieure et supérieure de l'intervalle considéré :

$$\textit{Les bornes } m_j = \max_{p \in P_j} (b_{\text{inf}} S_j^p)$$

$$\textit{et } M_j = \min_{p \in I_j} (b_{\text{sup}} S_j^p)$$

*de l'intervalle où la variable $a_j$ prend ses valeurs doivent être calculées le long de toutes les chaînes simples menant du sommet $j$ au sommet $p$ et vérifier $m_j < M_j$. Les cycles doivent être conservés et les calculs effectués pour tous les sommets.*

De cette manière, les intervalles $[m_j, M_j]$ définissent l'ensemble des valeurs $a_j$ potentiellement acceptables (permuter $b_{\text{inf}}$ et $b_{\text{sup}}$ dans les deux expressions

aurait pu conduire à la conclusion erronée qu'il n'existait pas de solution). Mais étant donné leur mode de calcul, les $S_j^p$ ne sont pas indépendants donc même si cet intervalle existe, on ne peut garantir l'existence d'une solution. Ces remarques peuvent être illustrées par un exemple très simplifié :

Soit $a$, $b$, $c$ trois variables et $[m, M]$ un intervalle à calculer tel que :

$$m = a - b$$

et :

$$M = a - c.$$

En adoptant la méthode de calcul ci-dessus, on trouve pour $a \in [3, 5]$, $b \in [0, 1]$, $c \in [2, 3]$ :

$$\begin{cases} m = b_{\text{inf}}(a-b) = 2, \\ M = b_{\text{sup}}(a-c) = 3. \end{cases}$$

L'intervalle existe donc et pourtant il ne peut manifestement exister de solution car $b < c$.

Pour $a \in [3, 5]$, $b \in [2, 3]$ et $c \in [0, 1]$, on aurait trouvé en permutant $b_{\text{inf}}$ et $b_{\text{sup}}$ :

$$\begin{cases} m = b_{\text{sup}}(a-b) = 3, \\ M = b_{\text{inf}}(a-c) = 2. \end{cases}$$

L'intervalle n'existe pas et pourtant il existe toujours une solution quels que soient $a$, $b$ et $c$ choisis dans leur intervalle respectif.

## II. 3. Existence d'une solution en fonction de la taille de l'échantillon

Les conditions d'existence varient suivant la taille de l'échantillon étudié. Pour deux individus, la solution est immédiate et existe toujours. Examinons le problème pour plus de deux individus.

### II. 3. 1. *Problème à trois dimensions*

Avec les notations choisies, le système d'équations s'écrit :

$$\left.\begin{array}{l} d_{12} = x_1 \bar{x}_2 = x_1 \bar{x}_2 x_3 + x_1 \bar{x}_2 \bar{x}_3 \\ d_{13} = x_1 \bar{x}_3 = x_1 x_2 \bar{x}_3 + x_1 \bar{x}_2 \bar{x}_3 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ d_{32} = \bar{x}_2 x_3 = x_1 \bar{x}_2 x_3 + \bar{x}_1 \bar{x}_2 x_3 \end{array}\right\} \begin{array}{c} 6 \text{ équations} \\ \text{non indépendantes} \\ 6 \text{ inconnues} \end{array}$$

A ce système correspond le graphe isomorphe à un cube de la figure 3.

— Le long du cycle de longueur paire, les arêtes doivent vérifier la relation :

$$d_{21} - d_{31} + d_{32} - d_{12} + d_{13} - d_{23} = 0. \tag{5}$$

Si cette équation est vérifiée le système est redondant et une quelconque des équations correspondant à une arête du cycle peut être supprimée.

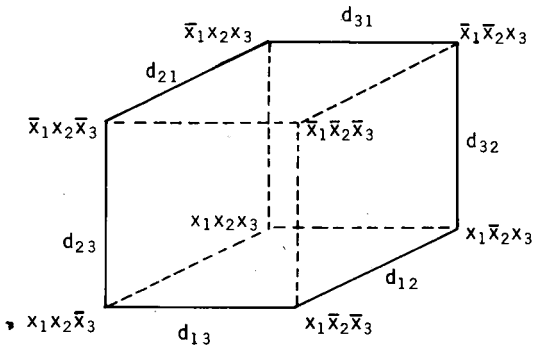


**Figure 3. Graphe du système d'équations à trois dimensions.**

— Le calcul de l'intervalle correspondant à une variable sommet suffit pour montrer l'existence d'une solution. Examinons le cas de $\bar{x}_1 x_2 \bar{x}_3$ après avoir enlevé par exemple l'arête marquée $d_{12}$ du cycle.

$$\begin{cases} m = \max(0, d_{21} - d_{31},\ d_{23} - d_{13}), \\ M = \min(d_{21}, d_{23}, d_{21} - d_{31} + d_{32}). \end{cases}$$

Il faut avoir $m \leqq M$ d'où :

$$\begin{cases} d_{21} \leqq d_{23} + d_{31} & d_{23} \leqq d_{21} + d_{13} \\ d_{31} \leqq d_{32} + d_{21} \end{cases} \tag{6}$$

Pour qu'il existe une solution à trois dimensions, il faut que les contraintes (5) et (6) soient vérifiées.

### II.3.3. Problème à quatre dimensions

Chaque variable est à nouveau décomposée en deux nouvelles variables :

$$\left.\begin{array}{r} \bar{x}_1 x_2 x_3 = \bar{x}_1 x_2 x_3 x_4 + \bar{x}_1 x_2 x_3 \bar{x}_4 \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ \bar{x}_2 \bar{x}_3 x_4 = x_1 \bar{x}_2 \bar{x}_3 x_4 + \bar{x}_1 \bar{x}_2 \bar{x}_3 x_4 \end{array}\right\} \begin{array}{c} 24 \text{ équations} \\ \text{non indépendantes} \\ 14 \text{ inconnues} \end{array}$$

A ce système d'équations correspond un graphe isomorphe à un hypercube (*fig.* 4). Pour qu'il existe une solution à 4 dimensions il faut que les conditions d'existence des solutions à trois dimensions soient respectées. Les variables du

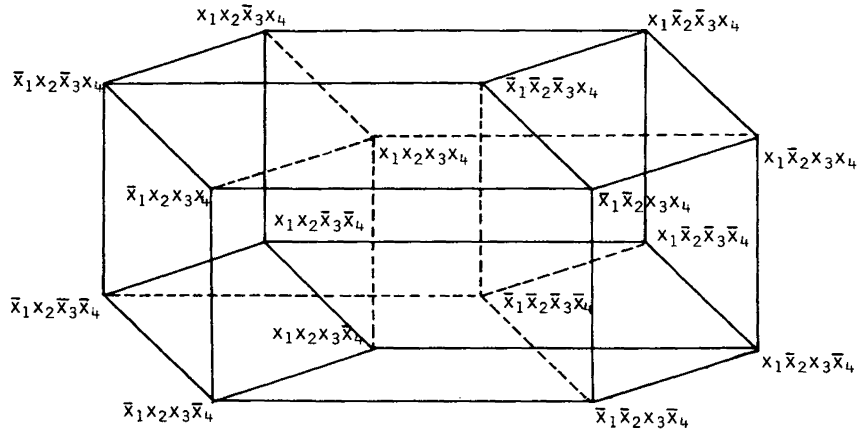


**Figure 4. Graphe du système d'équations à quatre dimensions.**

type $\overline{x}_1 x_2 x_3$ sont des intervalles et ce sont les résultats du paragraphe II.2.2 qui s'appliquent. Les calculs deviennent trop longs pour être effectués manuellement, et il a seulement été vérifié que d'autres conditions existaient. Il faut par exemple :

$$d_{13} + d_{42} \leqq d_{14} + d_{43} + d_{32} + d_{12}. \tag{7}$$

## II.4. Conclusion

Suivant les valeurs $d_{ij}$, il pourra ne pas exister de solution. Les contraintes à satisfaire, dénombrables facilement pour un échantillon de trois individus, doivent être trouvées par un calcul formel pour des tailles supérieures. Même si ces conditions étaient connues elles seraient de toute façon trop longues à vérifier pour présenter un intérêt pratique réel : du point de vue temps de calcul, il paraît plus rentable de chercher directement les solutions numériques.

## III. RECHERCHE DES SOLUTIONS

Deux questions liées à l'incomplète spécification du problème 2 se posent avant d'aborder sa résolution : quelles sont la qualité et la quantité des données ?

Dans le cas de notre application, les données (ordre de grandeur 0 à 10) sont obtenues par comptage des phages survivant après mise en présence d'un très grand nombre de bactéries et phages ($10^{11}$ phages/ml). Ce comptage peut être effectué de façon plus ou moins précise (et coûteuse), mais dans le cas le plus favorable considéré l'incertitude sur les données est encore élevée ($\sim 0{,}7$). Du fait du procédé de mesure, il est beaucoup plus économique d'obtenir des données sous forme d'intervalle que de valeur fixe.

En ce qui concerne la quantité des données, il faut noter que l'action des systèmes $R/M$ peut être efficace au point qu'il ne reste plus de phage après réaction, d'où une donnée inconnue. Mais l'action des systèmes $R/M$ ne peut être observée qu'à travers le comportement global des moyens défensifs des bactéries : ce bruit se traduit aussi par une absence de phage qui ne peut donc être interprétée uniquement en terme d'action des systèmes $R/M$. Certaines de ces données masquées peuvent être déterminées au prix de mesures supplémentaires, ce qui élimine partiellement le bruit.

Pratiquement, il s'est avéré efficace de prévoir plusieurs traitements très différents. Le premier, réservé aux échantillons de bactéries de taille importante, accepte des données peu nombreuses de type intervalle ou valeur fixe. Il est essentiellement destiné à sélectionner les bactéries qui vont subir le deuxième traitement réservé à des échantillons moins importants où les données sont proportionnellement nombreuses et de type valeur fixe. Dans ce cas les mesures sont effectuées avec une technique plus précise que dans le premier cas. Le troisième traitement est destiné à des échantillons de n'importe quelle taille, à condition que les données soient toutes connues (mesurées ou calculées) et très précises. Ce traitement suppose que la proportion $p_x$ de phages non protégés survivant à l'action d'un seul système $R/M$ $x$ est contante.

### III. 1. Faible proportion de données $d_{ij}$ connues

Ce cas correspond, pour notre expérimentation, à une connaissance d'environ 30% des $d_{ij}$, la plupart sous forme d'intervalle, la taille de l'échantillon étant d'environ 30 individus. Devant l'imprécision du problème il est préférable d'utiliser le pouvoir prédictif du modèle à la détermination des $d_{ij}$ inconnus plutôt que rechercher des solutions ( $\{ n, C, X \}$ ) peu informatives. La seule contrainte égalitaire (5) permet d'évaluer l'ensemble des $d_{ij}$ à partir d'une proportion de $1/2+1/m$ données « bien placées » (*fig.* 5).

L'algorithme proposé traite indifféremment les valeurs fixes et les intervalles :

Après avoir fixé les $d_{ij}$ inconnus à l'intervalle $[0, \infty[$, on recalcule tous les $d_{ij}$ de façon à vérifier les contraintes (5) et (6) ou leur extension aux intervalles.

Cet algorithme a deux effets :

— il diminue la largeur des intervalles $d_{ij}$ connus lorsque le contexte permet une meilleure évaluation,

— il fournit une évaluation des intervalles $d_{ij}$ inconnus.

Comme toutes les conditions d'existence ne sont pas prises en compte, les résultats sont des estimations par excès des intervalles réels. Mais il est apparu inutile de chercher une estimation très fine d'intervalles que l'on sait de toute façon peu précise parce que les informations de départ sont insuffisantes.

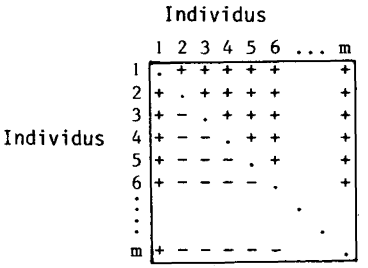


**Figure 5. A partir des $d_{ij}$ marqués « + » tous les $d_{ij}$ marqués « − » peuvent être déterminés (il existe de nombreuses autres dispositions de ces marques).**

Relativement peu de données étant connues, les conflits avec les contraintes sont rares. Pour les élminer, en dehors de la voie consistant à refaire la mesure des $d_{ij}$ incriminés, on peut élargir les intervalles concernés soit jusqu'à supprimer le conflit, soit en éliminant la donnée. Dans ces deux derniers cas, la correction d'une erreur revient à diminuer l'information prise en compte et à élargir les intervalles résultat.

L'intérêt de ce procédé est d'extraire des informations de la confrontation modèle-données, sans qu'il soit pour cela nécessaire d'effectuer ni une campagne de mesures complète ni des mesures très précises. Cet intérêt est renforcé en exploitant les données au fur et à mesure de leur obtention : ainsi, compte tenu des relations (5) et (6), la ou les nouvelles mesures à effectuer pour obtenir les informations désirées peuvent être déterminées à l'avance (*fig.* 5).

### III. 2. Forte proportion de données $d_{ij}$ connues

Nous sommes dans le cas où environ 80 % des données $d_{ij}$ sont connues sous la forme de valeurs fixes, pour un échantillon sélectionné de 10 à 20 individus. La connaissance de la quasi-totalité des données implique un très grand nombre de conflits avec les contraintes. Contrairement au cas précédent, il est

plus adapté ici de chercher directement des solutions approchées. Pour cela nous définissons un critère de performance $Q$ à minimiser :

$$Q = \sum_{\substack{i=1 \\ d_{ij} \text{ défini}}}^{m} \sum_{j=1}^{m} \left\{ \sum_{p=1}^{n} [c_p x_{ip}(1-x_{jp})] - d_{ij} \right\}^2.$$

Ce critère permet la recherche des paramètres $\{ n, C, X \}$ donnant un tableau $D'$ le moins éloigné possible du tableau observé. Pour la résolution de ce problème presque classique — hormis la détermination de $n$ — de minimisation, il existe plusieurs types de méthodes : approche totalement discrète [3, 6], méthode d'énumération utilisant le problème continu associé [1], résolution du problème continu associé soumis à des contraintes supplémentaires [4]. Une étude comparative détaillée [5] a conclu que la méthode par pénalisation était la plus adaptée, surtout au niveau du temps de calcul.

Cette méthode fournit tout ou partie du domaine sur lequel la fonction garde une valeur proche de l'optimum et considérée acceptable. Si l'on désire plus d'informations que la liste des solutions simultanément acceptables, il faut recommencer la mesure des $d_{ij}$ jusqu'à obtenir un critère nul ou inférieur à l'écart normal dû à l'imprécision des mesures.

### III. 3. Recherche des solutions dans un cas dégénéré

Lorsque tous les coefficients $c_p$ sont égaux à une constante $c$ connue, l'équation générale du problème devient :

$$\sum_{p=1}^{n} x_{ip}(1-x_{jp}) = d_{ij}/c = k_{ij}. \tag{8}$$

Plaçons-nous dans le cas où les $k_{ij}$ sont des entiers supposés connus, et où les conditions d'existence sont satisfaites.

En appelant $X_i$ l'ensemble des caractères possédés par l'individu $i$, l'équation (8) s'écrit :

$$|X_i - X_j| = k_{ij}.$$

Le problème revient donc à chercher la composition relative d'ensembles connaissant le cardinal de leurs différences.

III. 3. 1. *Énoncé du problème*

PROBLÈME 3 : Connaissant la valeur de $k_{ij} \in \mathbb{N}^+$, $\forall i,\ j=1$ à $m$ ($k_{ii}=0$, $\forall i=1$ à $m$), trouver :

- $X=\{ x_{ip} \in \{ 0, 1 \} / \forall i=1 \text{ à } m, \forall p=1 \text{ à } n \}$ ;
- le nombre de caractères $n$

tels que :

$$\sum_{p=1}^{n} x_{ip}(1-x_{jp})=k_{ij} \tag{9}$$

sous les conditions :

$$\left\{ \begin{array}{l} \exists i : {}_{ip}=0 \\ \exists i : x_{ip}=1 \end{array} \right\} \forall p=1 \text{ à } n,$$

les solutions ne pouvant se déduire l'une de l'autre par permutation des indices.

REMARQUE : Lorsque les coefficients $c_p$ sont constants, la condition 2 n'est plus applicable.

Ce problème 3 est sensiblement différent des problèmes précédent, et il est apparu opportun d'utiliser une approche différente. Il s'agit en fait de résoudre un système d'équations pseudo-booléennes non linéaires. Mais en utilisant les propriétés des solutions, on peut se ramener à une méthode itérative ne demandant que la résolution de plusieurs systèmes d'équations linéaires pseudo-booléennes.

III. 3. 2. *Propriétés des solutions*

Soit $S$ une solution pour les individus d'indice 1 à $m_1$ ($m_1 < m$), avec $n_1$ caractères. Les équations (9) sont vérifiées pour tout couple ($i$ $j$) tel que $i, j=1$ à $m$.

*Pour ajouter un nouveau caractère à une solution, il faut l'affecter aux $m_1$ premiers individus ou à aucun.*

Soit $n_1+1$ l'indice du caractère ajouté. Les équations (9) étant toujours vérifiées :

$$\sum_{p=1}^{n_1} x_{ip}(1-x_{jp})+x_{i,\,n_1+1}(1-x_{j,\,n_1+1})=k_{ij}, \qquad \forall i, j,$$

ce qui implique que $x_{i,n_1+1}(1-x_{j,n_1+1})=0$. De même on trouve

$$x_{j,n_1+1}(1-x_{i,n_1+1})=0$$

d'où :

$$x_{j,n_1+1}=x_{i,n_1+1}, \qquad \forall i, j=1 \text{ à } m_1.$$

Cherchons une solution pour $m_1+1$ individus. Soit $v$ le nouvel individu et $n_3$ le nombre de caractères de cette solution. Il faut vérifier les nouvelles équations :

$$\begin{aligned} &\sum_{p=1}^{n_3} x_{ip}(1-x_{vp})=k_{iv} \qquad \forall i=1 \text{ à } m, \\ &\sum_{p=1}^{n_3} x_{vp}(1-x_{ip})=k_{vi} \qquad \forall i=1 \text{ à } m. \end{aligned} \tag{10}$$

Compte tenu de la propriété ci-dessus, la solution ne pourra avoir que la forme donnée sur la figure 6. A partir des équations (10), les bornes de $n_2$ et $n_3$ peuvent être déterminées :

$$b_2 \leqq n_2 \leqq B_2 \qquad \text{et} \qquad b_3 \leqq n_3 \leqq B_3$$

avec :

$$b_2=n_1+\max_i\left(0,\ k_{iv}-\sum_{p=1}^{n_1} x_{ip}\right), \qquad B_2=n_1+\min_i(k_{iv}),$$

$$b_3=n_2+\max_i\left(0,\ k_{vi}-\sum_{p=1}^{n_1} x_{vp}\right), \qquad B_3=n_2+\min_i(k_{vi})$$

Dans (10), on peut isoler :

$$\text{card}(v)=\sum_{p=1}^{n_3} x_{vp}=\sum_{p=1}^{n_2} x_{ip}-k_{iv}+k_{vi}$$

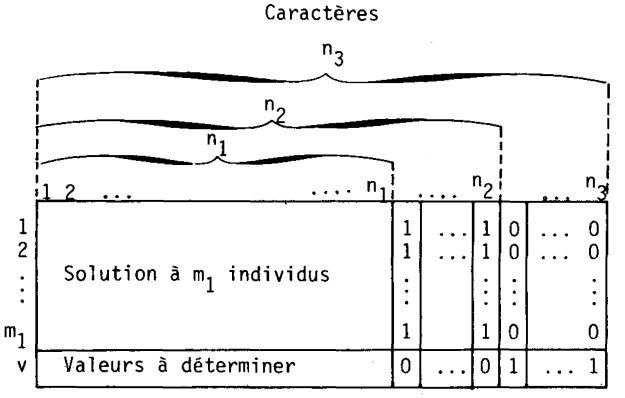


**Figure 6. Forme générale d'une solution portant sur $m_1+1$ individus.**

III.3.3. *Procédure de recherche des solutions*

Lorsque $n_2$ et $n_3$ sont connus, les équations (10) définissent en fait un système de $m_1+1$ équations pseudo-booléennes à $n_1$ inconnues. Comme tous les coefficients des variables sont booléens, la méthode [3] qui ordonne les coefficients perd de son intérêt. Si nous supposons $x_{vp}=0$, $\forall p=1$ à $n_3$, trouver une solution revient à placer card $(v)$ fois la valeur 1 dans le vecteur ligne $x_v$, tout en vérifiant les équations (10). Le procédé de construction (*fig.* 6) implique la satisfaction de la condition (1). L'algorithme doit être construit de manière à éliminer les solutions identiques à une permutation d'indice près.

Principales étapes de l'algorithme :

* Créer une solution fictive avec 0 individu.

* Pour $m_1=0$ jusqu'à $m-1$ faire :

. Pour toutes les solutions de dimension $m_1$ trouvées à l'itération précédente faire.

. . Pour $n_2=b_2$ jusqu'à $B_2$ faire.

. . . . Construire une représentation de $n_2$ caractères avec $x_{vp}=0$, $\forall p=1$ à $n_2$.

. . . . Envisager toutes les possibilités de placer card $(v)$ fois la valeur 1 dans le vecteur $x_v$ de taille $n_3$ variable, $n_3=b_3$ à $B_3$.

. . . . Placer les solutions trouvées dans la file des solutions à $m_1+1$ dimensions.

III.3.4. *Application réalisée*

Lorsque les $k_{ij}$ sont obtenus directement, la résolution est immédiate. Dans notre cas le problème s'est posé à la suite d'une transformation du problème 2 dans lequel les coefficients $c_p$ étaient constants. L'équation générale de ce problème est devenue :

$$\sum_{p=1}^{n} x_{ip}(1-x_{jp})=d_{ij}/c$$

Les valeurs $k_{ij}$ sont obtenues par arrondi à une valeur entière du rapport $d_{ij}/c$. Cet arrondi pouvant introduire des conflits n'existant pas préalablement ou confirmer ceux existants, il est effectué en parallèle avec une procédure de vérification et de correction, de manière à satisfaire les conditions d'existence.

## CONCLUSION

Le choix d'une méthode d'identification ne peut se faire qu'en fonction de la qualité et de la quantité des données. Nous avons proposé des méthodes adaptées à l'application considérée, mais l'étude du modèle reste vraie quel que

soit le contexte et fournit un point de départ pour la recherche de solutions dans des cas différents. Ces méthodes permettent de construire une typologie de bactéries à partir de données incertaines mais redondantes, la cohérence interne du modèle palliant leur manque de qualité. En outre, nous avons montré comment diminuer le coût des campagnes de mesure en utilisant le pouvoir prédictif du modèle. Enfin, sur le plan de la connaissance, cette étude a permis de vérifier la cohérence des hypothèses de travail [2], ce qui par la voie biochimique prendra plusieurs années.

## BIBLIOGRAPHIE

1. J. ABADIE, *Une méthode de résolution des programmes non linéaires partiellement discrets sans hypothèse de connexité*, R.I.R.O., 5e année, vol. 1, 1971, p. 23-38.
2. J. P. BOUSSEMAER, P. SCHRAUWEN, J. L. SOURROUILLE et P. GUY, *Multiple Restriction/Modification Systems in Lactic Streptococci and their Significance for the Definition of a Phage Typing System*, Journal of Dairy Research, vol. 47, 1980, p. 401-409.
3. P. L. HAMMER et S. RUDEANU, *Boolean Methods in Operations Research and Related Areas*, Springer, New York, 1968.
4. D. M. HIMMELBLAU, *Applied Nonlinear Programming*, McGraw-Hill, 1972.
5. J. L. SOURROUILLE, *Identification des systèmes de restriction/modification dans un échantillon de bactéries*, Thèse Docteur-ingénieur, Lyon, 1978.
6. H. A. TAHA, *Integer Programming. Theory, Applications and Computations*, Academic Press, New York, 1975.