INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Les Cahiers d'EDORA II

Alain PAVE - Jean-Luc GOUZE

N° 2530
Avril 1995

PROGRAMME 5

INRIA
SOPHIA ANTIPOLIS

Les Cahiers d'EDORA II

Alain Pavé*, Jean-Luc Gouzé** éditeurs

Programme 5 — Traitement du signal, automatique et productique
Projet Miaou

Rapport de recherche n° 2530 — Avril 1995 — 88 pages

Résumé : Cet ouvrage constitue les actes d'une réunion du club EDORA, qui rassemble des biologistes, informaticiens et mathématiciens autour de la modélisation en biologie.

Mots-clé : modélisation en biologie, représentation des connaissances.

(Abstract: pto)

*. Laboratoire de Biométrie, Université Claude Bernard, 43, Boulevard du 11 novembre 1918, F 69622 Villeurbanne Cedex
**. INRIA Sophia-Antipolis

Unité de recherche INRIA Sophia-Antipolis
2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex (France)
Téléphone : (33) 93 65 77 77 – Télécopie : (33) 93 65 77 65

EDORA Letters II

Abstract: These are the proceedings of a workshop of the Edora scientific club. This club gathers biologists, computer scientists and mathematicians concerned with the modelling in the biological sciences.

Key-words: biological modelling, knowledge representation.

SOMMAIRE

PRÉFACE

EDORA et ses enfants
- la modélisation et quelques questions d'actualité -

Alain Pavé

Ce deuxième ouvrage du club EDORA rend compte d'une activité de modélisation suivie dans les secteurs biologiques et écologiques, modélisation mathématique mais aussi informatique. On retrouve ici les idées avancées dans le numéro 1 des cahiers d'EDORA, publié en 1988, d'une conception large de la modélisation. Ce club aussi est l'un des lieux où se retrouvent des mathématiciens, automaticiens, informaticiens, biologistes et écologistes français.

Cependant les textes qui suivent ne témoignent que d'une partie de ce qu'on pourrait appeler le rayonnement d'EDORA. Si nous faisons aujourd'hui un bref bilan, on peut avancer sans grand risque de se tromper que sans EDORA :

- une part importante de l'activité « informatique et génome » ne se serait pas établie sur les bases actuelles. Elles n'auraient même peut être pas été possibles.

- L'initiative lancée par le Programme Environnement du CNRS de création d'un sous-programme thématique « Méthodes, Modèles et Théories pour l'Environnement » n'aurait pas vu le jour.

- Plus généralement, la cristallisation d'une communauté de biomathématiciens et de bio-infomaticiens n'aurait pas été aussi rapide.

D'autres exemples plus spécifiques et plus précis pourraient être cités à travers des résultats scientifiques et des réalisations concrètes qui illustrent l'activité du Club.

Il ne s'agit pas ici de faire un rapport d'activité. Mais depuis une dizaine d'année, on peut considérer que dans la mouvance d'EDORA, une quinzaine de thèses et quatre HDR ont été soutenues, une centaine d'articles ont été publiés. De même, des logiciels d'aide à la modélisation et à l'analyse de données ont été conçus, sans insister sur les contributions théoriques. Ce qui caractérise enfin EDORA, c'est que les progrès réalisés l'ont été certes en modélisation de situations biologiques et écologiques, mais aussi en informatique et en mathématiques. C'est là, à mon avis, l'une des conditions essentielles d'un travail multidisciplinaire : les différents acteurs, dans leurs disciplines respectives, doivent trouver matière à développements et résultats nouveaux.

On peut être gré à l'INRIA d'avoir été à l'initiative d'EDORA, puis d'avoir accueilli et soutenu ce club. Ceci a permis le développement d'une véritable activité méthodologique originale entre des chercheurs appartenant à des organismes différents (INRIA, bien sûr, CNRS, INRA, ORSTOM et Universités). Le croisement des compétences est toujours source de progrès. Au delà des frontières institutionnelles et disciplinaires, une communauté scientifique nationale a émergé pour le développement d'outils d'analyse des données et de modélisation pour l'écologie et la biologie, par exemple : des environnements (informatiques !) de résolution de problèmes, des systèmes à bases de connaissances, sans insister sur la modélisation mathématique qui fut le point de départ d'EDORA.

EDORA a-t-il un avenir ? En fait sa structure informelle peut permettre de compléter les nouvelles initiatives citées ci-dessus, voire d'en proposer des nouvelles, à risques. EDORA peut se le

permettre. Il peut aussi s'agir d'assurer une zone refuge en cas d'incident institutionnel majeur (quand même peu probable ou peu raisonnable étant donné les enjeux et la qualité de la communauté scientifique réunie).

Quelques résultats récents me semblent devoir être considéré, source de réflexions et de recherches méthodologiques nouvelles. J'en retiendrai deux :

1- En virologie, l'avancée faite dernièrement sur la compréhension des mécanismes de l'infestation d'un organisme humain par le VIH.

Ces résultats ont changé le point de vue qui semblait bien établi sur les premières phases du SIDA : celui du comportement « furtif » du virus, alors qu'on assiste à une lutte sans merci que les lymphocytes mènent dès le début de l'infestation. Cette vision nouvelle vient d'expériences qui analysent en termes quantitatifs la dynamique du « système » virus-lymphocytes et sa réponse à des antiviraux qui jouent le rôle d'une perturbation de ce système. La modélisation mathématique a aussi joué un rôle essentiel dans ce processus de recherche.

En fait, pendant plus de dix ans on a privilégié les approches réductionnistes, statiques et structurelles de la biologie moléculaire et de la pharmacologie, pour découvrir qu'une approche simultanée de la dynamique des populations virales et de celle des lymphocytes faisait voir d'un autre jour les mécanismes de l'infection. Gageons que des concepts de l'écologie, utilisant le modèle mathématique comme médiateur, permettront d'aller plus loin dans la compréhension et la résolution de ce problème majeur. D'autant plus que cela se joue à deux niveaux : celui du malade, avec ses populations virales et de lymphocytes, puis celui des populations humaines, comme phénomène épidemique.

A ce sujet trois remarques peuvent être faites :

- Depuis plus de 10 ans la communauté EDORA, et plus largement celle des biomathématiciens « écologues » étaient compétents pour aborder l'analyse des données et la modélisation de tels phénomènes. Mais le contact avec ceux qui les étudiaient expérimentalement a été en grande partie perdu depuis le début : les voies choisies et démarches étaient intellectuellement trop différentes.

- Ce n'est pas parce que l'écologie et ses diverses spécialités s'adressent à des populations et des peuplements de niveaux d'organisation supérieurs et surtout d'échelles spatiales plus grandes, que ses concepts ne peuvent pas être utiles ailleurs. Si un écologue a tout à gagner à être au fait des techniques et avancées de la biologie moléculaire, une biologiste moléculaire, un pharmacologue, une virologiste ou un microbiologiste tirera bénéfice d'une culture écologiste. Pourrions-nous assister au même type de progrès que ceux enregistrés sur la lutte contre les maladies parasitaires par l'introduction, plus naturelle il est vrai, d'une démarche écologiste ?

- Le rôle essentiel joué par la modélisation pour la compréhension du phénomène doit se poursuivre. Mais cette méthode peut aussi être précieuse pour imaginer les moyens de lutte contre l'infection, notamment ceux utilisant les combinaisons d'antiviraux. Par ailleurs et outre les aspects dynamique des populations, les aspects génétiques doivent être examinés et modélisés. La stratégie du virus est d'engendrer rapidement une grande diversité phénotypique, gagnant ainsi de vitesse le système immunitaire. Cette stratégie est fatale à l'hôte et au virus, car il tue son hôte. Peut-on imaginer un « modèle » stable de virus vivant en harmonie avec son hôte tout en stimulant son système immunitaire ?

Dans la mouvance d'EDORA au moins une partie de ces problèmes de modélisation peuvent être pris en compte. Des avancées dans ce domaine pourraient être utiles pour d'autres maladies infectieuses.

2- Dans les recherches sur l'environnement et le développement, thèmes aussi d'actualité, la modélisation des dynamiques couplées de systèmes naturels et de systèmes sociaux et plus

généralement de modèles de systèmes « intégrés » à diverses composantes (bio-écologiques, socio-économiques, géo-physico-chimiques) est à l'ordre du jour.

Des approches nouvelles, notamment celles issues de l'Intelligence Artificielle distribuée, comme les systèmes multi-agents, ont permis d'aborder ce type de problème. L'intérêt, parmi d'autres, est de pouvoir représenter des comportements des acteurs naturels ou sociaux. En revanche, les possibilités de représentations réalistes et détaillées risquent de conduire à des résultats peu faciles à utiliser sinon en analysant de façon traditionnelle (par exemple statistique) les simulations produites, véritables expériences artificielles. On peut tomber dans le paradoxe d'Umberto Eco de la carte à l'échelle 1:1 ! Malgré ces risques et ces limites, il me semble intéressant de travailler sur ces mondes virtuels, connectés (quand même !) à la réalité, d'y effectuer de véritables expériences, elles aussi virtuelles, et de modéliser de façon plus synthétique certains résultats, par exemple à l'aide d'objets mathématiques simples. Cela peut être une voie permettant d'éclairer, dans des situations complexes, le constat fait depuis longtemps et démontré dans quelques cas élémentaires : malgré le nombre et la complexité des processus sous-jacents, l'évolution de certaines variables d'état, souvent macroscopiques, peuvent être représentées par des modèles simples.

L'avenir n'est-il pas de considérer les relations entre les trois pôles : monde réel et son observation - monde virtuel, représentation informatique de ce monde réel - représentations formelles très simplifiées du modèle mathématique « utilisable ».

EDORA et ses enfants me semblent bien placés pour aborder ce type de problème et continuer à participer à l'effort collectif sur la modélisation de la gestion des ressources renouvelables, de la dynamique et du contrôle des écosystèmes, sur la conception d'outils d'intérêt général comme les systèmes à bases de connaissances pour la systématique.

Alain Pavé
le 4 mars 1995

Ce volume des cahiers d'EDORA fait suite au tome 1, publié sous la forme d'un rapport de recherche INRIA numéro 866 en juillet 1988. Le présent ouvrage constitue les actes d'une réunion du club EDORA (dont le président est Alain Pavé) organisée par A. Pavé et J.L. Gouzé à Lyon (Laboratoire de Biométrie) les 30 et 31 mai 1991. Divers exposés ont eu lieu, dont certains n'ont pas donné lieu à article dans ce volume. Il est peut-être intéressant de rappeler leurs titres, afin de donner une meilleure idée des sujets abordés :

- P. Nival (Villefranche-sur-Mer): Choix du niveau de complexité biologique et validation d'un modèle.
- I. Till (Orsay): Théorie des jeux en biologie évolutive.
- S. Maurice (Orsay): Evolution des systèmes de reproduction avec un déterminisme nucléocytoplasmique du sexe: modèles de simulation.
- R. Ferrière (ENS Paris): Explorer la densité dépendance: modèles en temps discret structurés en âge.
- M. Farza (LAG Grenoble): Système d'aide à la modélisation et à l'estimation de bioprocédés.

Je tiens enfin à remercier A. Guiteau pour son aide lors de l'édition de ces actes.

Jean-Luc Gouzé, INRIA

Modélisation des connaissances méthodologiques en analyse des données

François CHEVENET
URA CNRS 243
Université Claude Bernard - Lyon1
F-69622 VILLEURBANNE Cedex
chevenet@biomserv.univ-lyon1.fr

Jutta WILLAMOWSKI
IMAG-ARTEMIS
BP 53 X
F-38041 GRENOBLE Cedex
jutta@everest.imag.fr

Résumé: Dans cet article les problèmes liés au triplet modèles mathématiques/programmes/utilisateurs sont abordés dans le contexte de l'analyse des données. DANAIDE, une application à l'ordination linéaire simple des concepts développés autour de la notion d'environnement de résolution de problèmes est présentée. Le formalisme de représentation est le schéma, qui permet de décrire sous forme d'objets non seulement les données manipulées au cours d'une analyse, mais aussi les tâches qui manipulent ces dernières.
Mots-clefs: environnement de résolution de problèmes, représentation par objet, représentation de tâches, analyse des données, ordination linéaire simple.

Abstract: This article discusses the problems concerning the relations between mathematical models, programs and users in the domain of data analysis. It presents DANAIDE, an application of concepts developped within the context of problem solving environments. The representation formalism is the schema ; it allows to describe the data manipulated during an analysis as well as the tasks manipulating the data.
Key-words: problem solving environment, object centered representation, task representation, data analysis, simple linear ordination.

Sommaire:

1- Introduction

L'objectif de notre étude consiste en la recherche d'un environnement de résolution de problèmes appliqué aux méthodes d'ordination linéaire simple (statistique descriptive). Il s'agit d'une part de préciser, valider une approche informatique (1) et d'autre part de modéliser les connaissances méthodologiques (2) nécessaires à la mise en oeuvre de ces méthodes. L'effort de formalisation permettant d'acquérir de la connaissance sur la connaissance, nous attendons de ces travaux un modèle d'organisation, d'abord au niveau des méthodes, à plus long terme au niveau de l'utilisateur.

Pourquoi, quand et comment doit-on faire appel au modèle euclidien d'analyse des données ? Cette question, quoique naturelle, n'a de réponse que sous le couvert de connaissances mathématiques (modèles), informatiques (logiciels, procédures) et demande expérimentale (finalité). Les trois entités: modèles, programmes et utilisateurs sont au centre des réflexions menées sur la recherche de la définition d'un environnement de résolution de problèmes (par exemple [ROUS88]).

Le modèle euclidien d'analyse des données est très riche, il génère de nombreuses méthodes décomposables en modules paramétrables. En ordination linéaire simple, les données sont sous la forme d'un seul tableau de n lignes et p colonnes difficile à appréhender en raison de sa nature multivariée, et les méthodes associées (analyse des correspondances simple, multiples, inter ou intra classes, non symétriques, partielles ou globales, analyse en composantes principales générale, centrée par colonnes ou par lignes, doublement centrée...) ont toutes pour objectif la recherche d'une représentation dans un espace de faible dimension (plans factoriels).

Les logiciels qui mettent en oeuvre ces méthodes sont aussi très nombreux et largement diffusés, mais ils ne diminuent pas le déséquilibre entre d'une part un flot intense et un niveau élevé de problèmes actuels de l'analyse des données dans différents domaines et d'autre part la productivité créative d'un corps limité d'expert en statistiques [AIVA91].
Ces logiciels sont parfois trop spécialisés, trop rigides, trop "boîte noire" et les solutions qu'ils proposent sont en fait celles du concepteur du programme et ne répondent pas toujours aux objectifs particuliers de l'utilisateur.

La mise au point du système intégré A.D.E. (Analyse des Données Ecologiques [CHES&91][THIO90]) a permis des développements récents en matière d'analyse des données (approches linéaires et graphiques) et d'environnement matériel et logiciel. Le système pilote un ensemble de modules paramétrables documentés par des piles HyperCard.

Il est nécessaire de compléter ces travaux par la modélisation d'une expertise méthodologique qui puisse rendre compte de la variabilité des utilisateurs potentiels des méthodes d'analyse des données. Cette variabilité est grande, tant du point de vue du niveau de connaissances de l'utilisateur en analyse des données que de celui du champ scientifique concerné (biologie moléculaire, écologie,..., économie).

Pour le non statisticien, l'environnement de résolution de problèmes doit l'*aider à résoudre des tâches qu'il juge complexes*, en limitant la sous-, ou la sur-exploitation, de ses données. L'environnement doit être capable de prendre des décisions concernant le processus de résolution, il doit permettre de *guider* l'utilisateur dans le choix et la mise en oeuvre d'une démarche méthodologique menant à son objectif, de l'*assister* dans le choix entre les différentes méthodes disponibles et leur enchaînement, d'*automatiser* les

---

(1) *Projet SHERPA, "Dynamique des bases de connaissances", F. Rechenmann, IMAG - ARTEMIS, BP 53 X, F-38041 Grenoble Cedex.*

(2) *Projet PIREN "Méthodes, Modèles et Théories", les problèmes de multiplicité d'échelles d'espace et de temps dans les recherches sur l'environnement, méthodes et logiciels pour l'analyse des données écologiques, D. Chessel, A.Pavé: URA 1451, URA 243, Université Claude Bernard Lyon 1, 43 bd du 11 novembre 1918 F-69622 Villeurbanne Cedex.*

séquences les plus techniques et *contrôler* leur exécution. S'il y a échec d'une de ces méthodes, le système doit pouvoir revenir en arrière sur ses choix et tester d'autres stratégies ou d'autres méthodes.

Pour l'expert en statistiques, il faut pouvoir laisser un champ d'investigation dans la recherche de nouvelles méthodes, permettre un travail "pour voir" par exemple. L'environnement informatique doit alors faire preuve d'*ouverture*. Il doit être doué d'un potentiel de réorganisation de méthodes élémentaires prédéfinies pour la construction de nouvelles méthodes, faciliter la manipulation d'une programmathèque par exemple.

Pour le concepteur de la base il faut donc fournir au système un potentiel dynamique, une ouverture "en amont", dirigée vers l'utilisateur. Nous pouvons aussi distinguer une ouverture "en aval": le système doit pouvoir accueillir de nouvelles méthodes sans remise en cause de sa structure. Il s'agit par exemple de l'enrichissement d'une bibliothèque de programmes. L'intégration de méthodes concerne d'un côté des modules déjà existants, de l'autre les nouveaux programmes développés régulièrement dont la connaissance nécessaire à leur application est de plus en plus évoluée et spécialisée.

Une autre caractéristique essentielle de l'environnement de résolution de problèmes est sa capacité d'*interaction* avec l'utilisateur via une *interface graphique*, mais l'interaction passe aussi par le choix d'un *formalisme de représentation* adéquat. Le modèle des "frames" **[MINS75][FIKE&85]** est un formalisme proche du modèle conceptuel de l'utilisateur. Ce type de représentation fournit une souplesse dans la conception d'un système manipulant des objets complexes. Sous les formes classiques, procédurales, connaissance et contrôle sont fondus ensemble au sein de l'algorithme et du langage utilisés. Sous une forme déclarative, la connaissance (explicite, discursive) est accessible indépendamment des mécanismes d'exploitation de ces connaissances.

Enfin, il doit y avoir *couplage* entre le système et les procédures externes pour permettre *l'inférence des propriétés des objets*, pour la *mémorisation* des propriétés des données calculées et *l'interprétation des résultats* : "les données font partie intégrante du problème" **[ROUS88]**.

Au sein du projet SHERPA SCAI (Scientific Computing with Artificial Intelligence), un prototype permettant de construire un environnement de résolution de problèmes a été développé sur la base du système de gestion de bases de connaissances à objets Shirka, dont nous rappelons les principales caractéristiques. Le concept de méthode, puis celui de tâches se rapportant au système SCAI est développé, nous décrivons ensuite son interface graphique.

Nous terminons par la présentation de DANAIDE (Data ANalysis with Artificial Intelligence and Data Expert), une application des concepts développés par SCAI à l'ordination linéaire simple. DANAIDE constitue notre introduction dans la modélisation d'algorithmes en analyse des données. Le principe selon lequel la méthode encapsule le programme, et non l'inverse, est à la base de ce travail.

2- Shirka

2.1- Représentation des connaissances

Shirka **[RECH&90][RECH&91]** est un modèle de connaissance centré-objet inspiré du modèle des "frames" **[FIKE&85]**, mais qui s'appuie sur la distinction classe - instance. Une classe et ses instances sont définies dans un schéma (**Figure-1**) par un ensemble d'attributs ; un attribut est lui-même défini par une liste de facettes et une facette par une liste de valeurs ; une valeur est soit un schéma soit une référence à un schéma.

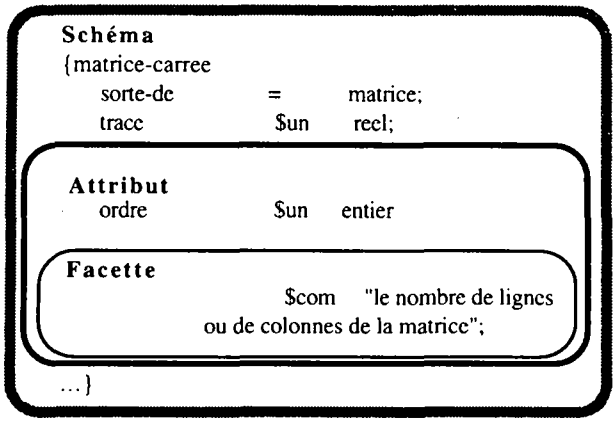

Figure 1. Schéma simplifié de classe décrivant une matrice carrée dans la base DANAIDE. Tous les attributs et facettes ne sont pas représentés et détaillés ici.

Les classes sont organisées en un graphe acyclique ; une classe donnée domine des classes d'objets plus spécifiques auxquelles elle transmet, par un mécanisme d'héritage, la connaissance sur ses attributs (**Figure-2**). Cette connaissance peut être précisée et augmentée dans les sous-classes, mais pas remise en cause. La description d'un attribut peut être précisée par des restrictions de domaine de valeurs ou par de nouveaux moyens de détermination de ses valeurs, ou encore en fixant ces valeurs.

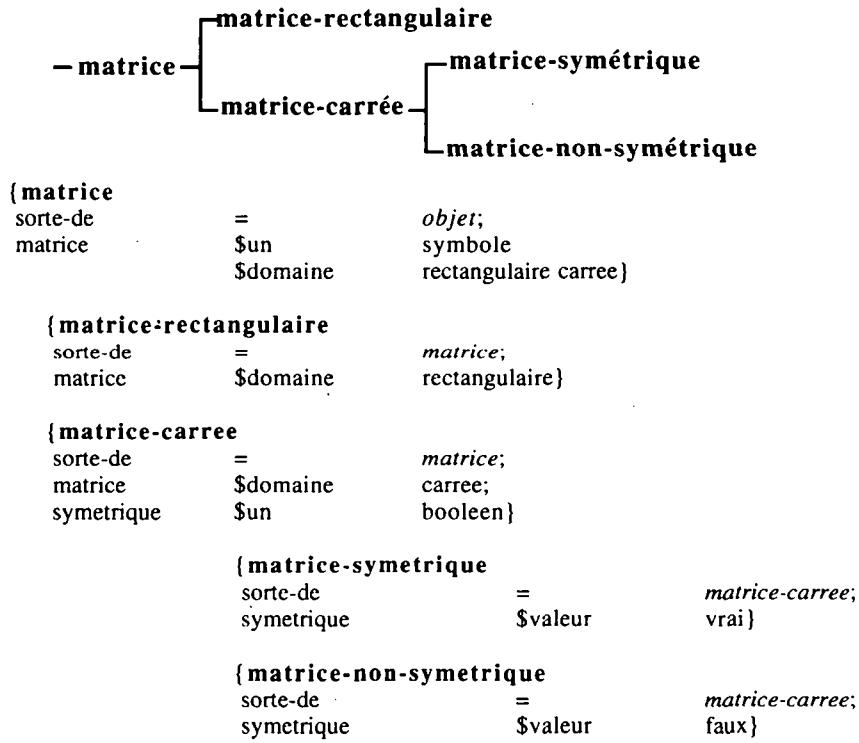

Figure-2. Exemple d'organisation en classes. La classe `matrice-symetrique` est définie comme une sous-classe de `matrice-carree`, qui est elle-même une sous-classe de `matrice`. Le type de la matrice (carrée ou rectangulaire) dans `matrice-symetrique` est donc hérité de sa sur-classe `matrice-carree` ; l'attribut `matrice` est défini par restriction du domaine de ce même attribut dans sa sur-classe : la seule valeur possible est carrée (extrait simplifié de DANAIDE).

2.2- Exploitation des connaissances

Utiliser Shirka consiste à définir des classes, puis à créer des instances de ces classes. Ensuite, deux types d'actions sont possibles : identifier la ou les sous-classes d'appartenance de ces instances – à l'aide d'un mécanisme de classification qui exploite la hiérarchie des classes et sous-classes – ou chercher les valeurs indéterminées d'attributs. Shirka tente alors de répondre à ces requêtes en utilisant la connaissance introduite dans les classes lors de leur définition. Pour ce faire, il dispose de mécanismes d'inférence, comme l'attachement procédural et le filtrage qui peuvent être combinés.

2.2.1- L'attachement procédural

L'attachement procédural, existant dans tous les systèmes faisant appel à la notion de "frame" présente cependant diverses originalités dans Shirka. La procédure attachée à l'attribut doit elle-même être définie, d'un point de vue externe, par un schéma de classe, dont les attributs sont ses paramètres. Appeler cette procédure revient donc à instancier ce schéma de classe et transmettre l'instance ainsi créée à une fonction. Cette dernière accède alors aux paramètres d'entrée et effectue les calculs requis. L'attachement procédural peut être utilisé dans plusieurs types de facettes. Il permet notamment dans la facette `sib-exec` d'inférer des valeurs d'attributs indéterminées; la fonction complète l'instance par les valeurs des attributs associés aux paramètres de sortie. Ces valeurs sont alors propagées à d'autres attributs du schéma faisant appel à cet attachement procédural (**Figure-3**).

```
{matrice-rectangulaire
...
nombre-lignes       $un              entier
                    $sib-exec...
 nombre-colonnes    $un              entier
                    $sib-exec...
nombre-elements     $un              entier
                    $sib-exec
                    {methode-nb-elements
                    objet-fichier    $sib-filtre
                                     {objet-modele
                                     objet-mathematique    $var<- lui;
                                     objet-fichier         $var-> objet-fichier};
                    nb-elements $var-> nombre-elements}
                    $a-verifier
                    {predicat-nbc*nbl=nbe
                    nbl     $var<- nombre-lignes;
                    nbc     $var<- nombre-colonnes;
                    nbe     $var<- nombre-elements}
                    $si-succes
                    {stocker ...nombre-lignes, nombre-colonnes...}}
```

Figure-3. Exemple d'inférence des propriétés d'un objet. La classe `matrice-rectangulaire` dispose de la possibilité d'inférer la valeur des attributs `nombre-lignes`, `nombre-colonnes` et `nombre-éléments` d'une matrice rectangulaire. C'est un exemple de combinaison de l'attachement procédural (facette `sib-exec`) et du filtrage (facette `sib-filtre`) . Le calcul du nombre d'éléments du fichier par une méthode spécifique permet de vérifier les valeurs déterminées indépendamment pour le nombre de lignes et de colonnes (la facette `a-verifier` est un autre type de facette permettant de faire appel à l'attachement procédural). En cas de succès les résultats sont stockés (facette `si-succes`)(extrait simplifié de DANAIDE).

2.2.2- Le filtrage

Le filtrage est un mécanisme d'inférence propre à Shirka. Il consiste à rechercher des instances satisfaisant une description donnée sous la forme d'un schéma de classe et à extraire des instances trouvées des valeurs qui deviendront, par le biais de variables, les valeurs des attributs indéterminés. Il est mis en oeuvre par la facette `sib-filtre`, qui peut contenir plusieurs filtres essayés séquentiellement. Un filtre se présente sous la

forme d'un schéma de classe, spécialisation d'un schéma existant. Il représente un ensemble de conditions que doivent satisfaire des instances du schéma filtré.

Généralement, le filtre porte sur un sous-ensemble des attributs du schéma filtré. Un filtre étant un schéma de classe, le processus de filtrage consiste donc à instancier le filtre, puis à comparer les instances du schéma filtré avec l'instance de filtre créée. Quand une instance trouvée satisfait les conditions, la valeur d'un de ses attributs est considérée comme valeur possible pour l'attribut auquel ce filtre est attaché par la facette `sib-filtre` (**Figure-3**).

3- Représentation des connaissances dans SCAI

Sur la base de Shirka une couche supplémentaire a ensuite été ajoutée, permettant de représenter sous forme d'objets les méthodes élémentaires et la démarche de résolution de problèmes complexes : SCAI, un prototype permettant la construction d'environnements de résolution de problèmes. La représentation unique et cohérente par objets dans SCAI peut intégrer à la fois des données en grand nombre et des programmes d'analyse très diversifiés.

Au niveau le plus abstrait la démarche de résolution de problèmes est décrite en termes de tâches ; plusieurs types de tâches ont été identifiés. Une tâche complexe est décomposée en sous-tâches. Elle peut modéliser une séquence de plusieurs sous-tâches ou un choix parmi différentes sous-tâches possibles. La tâche élémentaire fait directement appel à une ou plusieurs méthodes (programmes) de résolution possibles. La représentation choisie offre l'avantage de pouvoir intégrer une forme de mode d'emploi des méthodes comme on le verra par la suite.

3.1- Les méthodes

Chaque méthode est modélisée sous forme d'une classe. Celle-ci précise quelles sont les caractéristiques de ses données d'entrée et de sortie et quel est le module exécutable associé. Cette représentation en classes permet de raisonner sur les méthodes, en particulier pour la choisir dans un contexte donné (**Figure-4**).

```
{procedure-mmm
sorte-de                =           procedure
                        $com        "MULTIPLICATION DE 2 MATRICES";
nom-fct                 $valeur     procedure-mmm;
&entree e1              $un         matrice
                        $com        "premiere matrice";
&entree e2              $un         matrice
                        $com        "deuxieme matrice"
                        $a-verifier
                        {predicat-nbc1=nbl2
                         matrice-1  $var<-      e1
                         matrice-2  $var<-      e2};
&sortie s1              $un         matrice}
```

Figure-4. Exemple de méthode, le produit de deux matrices. Les contraintes de typage correspondent aux contraintes sur les types attendus par les entrées (utilisation de différentes facettes faisant référence à différentes classes de schémas). La valeur `matrice` est le nom de la classe à laquelle doit appartenir l' instance pour les entrées e1 et e2. Un second type de contrainte statique est associé à l'entrée e2 en utilisant la facette `a-verifier` qui consiste à tester si le nombre de colonnes de la première matrice égale le nombre de lignes de la deuxième. L'attribut `nom-fct` a pour valeur le nom de la fonction Lisp responsable de la liaison avec la procédure externe. Le résultat du produit matriciel est une instance de la classe matrice, valeur de l'attribut s1 (extrait simplifié de DANAIDE).

Ces pré- et postconditions permettent donc de déclarer les connaissances nécessaires à l'emploi et à l'invocation des méthodes. L'invocation d'une procédure revient à instancier le schéma de méthode associé.

La définition de classes pour les méthodes permet en outre de les structurer. Elles sont organisées en hiérarchies de classes. Les méthodes à l'intérieur d'une hiérarchie ont un même rôle, mais elles sont plus ou moins adaptées à un certain contexte de données. Une méthode plus générale se trouve à un niveau plus haut dans la hiérarchie de classes qu'une méthode plus spécifique. En outre il y a héritage des caractéristiques d'une classe de méthode plus générale vers une classe de méthode plus spécifique. Ceci concerne plus particulièrement des contraintes définies sur ses entrées. Une classe de méthode plus spécifique définit par exemple des contraintes supplémentaires concernant les entrées qu'elle est capable de traiter ou elle a tout simplement besoin de données supplémentaires.

La structuration des méthodes en hiérarchie selon leur spécificité induit une autre possibilité, elle permet de supporter un mécanisme de choix de la méthode la plus adaptée à un contexte précis. Ceci s'effectue à l'aide d'un mécanisme de classement, qui à partir d'une classe racine, va chercher parmi toutes ses sous-classes lesquelles sont adaptées au contexte donné. Ce classement se fait principalement sur les contraintes définies sur les entrées des méthodes, d'autres caractéristiques peuvent également être prises en compte (par exemple il existe de nombreuses variantes de réalisations de l'inversion matricielle, des choix implicites dus aux valeurs réelles ou complexes, mais aussi des choix explicites quant à la précision ou du temps des calcul) .

L'exécution d'une méthode est contrôlée par le système. Elle consiste d'abord à instancier son schéma de classe. Lors de cette instanciation les contraintes sur les entrées, définies dans son schéma de classe, sont vérifiées. Il y a donc ici un premier contrôle du système sur l'exécution. Ensuite celui-ci se charge de l'appel du module couplé. Il récupère les données de sorties et vérifie également les contraintes définies sur celles-ci. Un échec d'une méthode peut donc être dû soit à des problèmes concernant les contraintes sur les entrées ou sur les sorties, soit à l'échec de l'exécution du module associé.

La représentation des méthodes sous forme de classes permet par ailleurs d'en intégrer facilement de nouvelles. Les classes constituent une interface déclarative entre le système ou l'utilisateur et les modules proprement dits. La structuration des classes de méthodes en hiérarchie facilite non seulement le choix de méthodes mais aussi leur intégration, car il suffit de connaître la fonctionnalité d'une méthode et les contraintes requises pour ses entrées et ses sorties pour introduire la classe correspondante à l'endroit approprié dans la hiérarchie.

3.2- Le concept de tâche

La représentation explicite des méthodes permet ainsi de lancer des traitements élémentaires sur des entités définies. Le choix de la méthode la plus adaptée au contexte peut être pris en charge par le système. L'étape suivante consiste à modéliser des stratégies de résolution de problèmes plus complexes qui ont besoin d'enchaîner des méthodes. C'est ici que le concept de tâche est introduit. Une tâche est d'un côté caractérisée par ses fonctionnalités, ses entrées et ses sorties comme une méthode, de l'autre par la "stratégie de résolution" qu'elle décrit .

Comme les méthodes, les tâches sont définies par des classes d'objet et structurées en hiérarchies. Les tâches à l'intérieur d'une hiérarchie résolvent le même problème global mais elles sont plus ou moins adaptées aux différents contextes de l'information à traiter. Elles constituent , ainsi que les classes de méthodes, une interface entre l'utilisateur et les stratégies disponibles. Le même mécanisme de classement peut être utilisé pour choisir la tâche (ou stratégie) la plus adaptée à une situation précise (voir exemples chapitre 5).

Dans SCAI différents types de tâches ont été définis selon la forme de la stratégie qu'elles modélisent. Il y a d'abord les tâches complexes qui se décomposent en sous-tâches : les tâches séquentielles et les tâches de choix. Les tâches séquentielles permettent de définir des enchaînements prototypiques de sous-tâches. Les tâches de choix

réunissent les différentes possibilités pour résoudre un même problème. Les sous-tâches d'une tâche de choix sont a priori équivalentes, c'est-à-dire qu'elles sont toutes aussi adaptées au problème à résoudre, mais la définition d'une classe de tâche et de sous-classes plus spécialisées permet au système de décider a priori (en connaissance du contexte) laquelle de ces sous-classes est la plus appropriée. En combinant les tâches séquentielles et les tâches de choix il est possible de définir des tâches itératives et récursives.

Un autre type de tâche défini, la tâche-à-spécialiser, sert à structurer les connaissances. Elle n'a pas de stratégie de résolution directement associée. Pour sa réalisation il faut exécuter une des tâches définies par ses sous-classes.

Enfin, pour faire le lien entre tâches et méthodes les tâches élémentaires sont introduites. Elles constituent les feuilles de la décomposition de tâches complexes. Une tâche élémentaire ne se décompose plus en sous-tâches mais fait directement appel à une classe de méthode qui permet de la résoudre. Les différents types de tâches sont résumés dans **figure-5.**

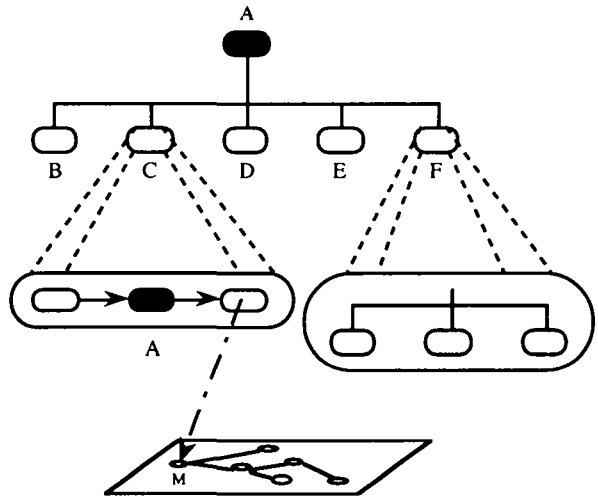

Figure-5. Structures de décomposition séquentielle, alternative et récursive. *C* est une tâche séquentielle, *F* est une tâche choix et *A* est une tâche récursive. La dernière tâche qui compose *C* est une tâche terminale à laquelle est attachée une hiérarchie de méthodes **[UVIE&91]**.

Les tâches sont donc structurées selon deux aspects : d'un côté elles sont organisées en hiérarchie de classes selon les objectifs qu'elles modélisent et qui sont plus ou moins spécialisés, de l'autre elles ont chacune une décomposition différente. Ces deux aspects sont indépendants. La position d'une tâche dans la hiérarchie de spécialisation par rapport à une autre n'influence pas sa décomposition. Il est par exemple possible de définir une classe de tâche séquentielle A avec comme sous-classe une tâche B qui est une tâche élémentaire. C'est l'exploitation simultanée de ces deux aspects qui permet une résolution efficace des problèmes. Le contrôle de l'exécution des tâches dépend donc de ces deux structurations et est décrit dans le chapitre suivant.

4- Exploitation des connaissances dans SCAI

La représentation des connaissances stratégiques et opérationnelles par le moyen de tâches et de méthodes constitue la base pour une résolution automatique de problèmes. Pour gérer l'exécution des tâches et des méthodes un mécanisme approprié a été défini : c'est le moteur de tâches. Ce chapitre expose son fonctionnement. Les différentes phases de l'exécution qui sont distinguées par le moteur sont détaillées. Par la suite sont décrits les problèmes susceptibles de se présenter pendant la résolution d'une tâche et la réaction du moteur face à de tels problèmes.

4.1- Le moteur de tâches

Le moteur distingue quatre phases essentielles dans l'exécution d'une tâche : la première phase consiste à adapter la tâche au contexte donné; la deuxième à l'instancier; la troisième à la décomposer et la quatrième à déterminer les sorties.

L'adaptation d'une tâche ou d'une méthode au contexte s'effectue donc par un mécanisme de classement qui détermine quelles spécialisations (sous-classes) de la tâche sont adaptées à la situation actuelle, généralement aux données à traiter. Dans le système actuel c'est l'utilisateur qui pilote cette adaptation, c'est-à-dire que c'est lui qui choisit la branche à explorer dans l'arbre des spécialisations de la tâche ou de la méthode. Les données manquantes peuvent pendant le classement être inférées ou demandées à l'utilisateur. Il est envisageable d'automatiser complètement ce processus d'adaptation au contexte de sorte que ce soit le système qui détermine de façon autonome la ou les tâches les plus adaptées. L'instanciation d'une tâche ou d'une méthode consiste d'abord à renseigner toutes ses entrées. Les contraintes définies sur celles-ci sont ainsi vérifiées. Si ces contraintes ne sont pas satisfaites, l'instanciation, et par là l'exécution échoue. Pour la décomposition d'une tâche deux cas peuvent se présenter : (a) la tâche retenue est complexe, elle est alors décomposée en ses sous-tâches qui à leur tour sont instanciées et exécutées en appliquant récursivement le même traitement, (b) la tâche retenue est élémentaire, la méthode associée est alors spécialisée, donc adaptée au contexte, instanciée et exécutée.

Avec ce fonctionnement du moteur chaque (sous-)tâche ou chaque méthode n'est à son tour adaptée au contexte qu'au début de sa propre exécution. Ceci permet d'adapter en cours du raisonnement la stratégie de résolution prototypique et prédéfinie au contexte actuel du problème. La décomposition de tâches en sous-tâches conduit ainsi finalement à un enchaînement spécifique de méthodes, qui mène à la solution d'un problème particulier.

A la fin d'une exécution "normale" les sorties de la tâche doivent être déterminées. L'instance est alors complétée. Les contraintes concernant les sorties sont vérifiées. Au cas où celles-ci sont violées il y a échec de la tâche (ou de la méthode) concernée.
L'échec d'une tâche ou d'une méthode peut se produire pendant chacune de ces quatre phases, soit parce que la spécialisation choisie n'était pas la bonne, soit parce que les contraintes sur les entrées ou sur les sorties ne sont pas satisfaites, ou encore parce qu'une des sous-tâches échoue. En cas d'échec la tâche ou la méthode est marquée et la raison de l'échec est mémorisée. Le moteur revient dans ce cas en arrière : il remet en cause le dernier choix effectué, lors de l'exécution d'une tâche de choix ou lors d'une spécialisation.

Le moteur de tâches intègre ainsi des connaissances stratégiques, tactiques et opérationnelles: connaissances stratégiques grâce au pouvoir de description de tâches complexes et de leur décomposition, connaissances tactiques exprimées dans la structure hiérarchique des tâches et des méthodes exploitées lors de la spécialisation pour adapter le processus de résolution au contexte spécifique et connaissances opérationnelles permettant de contrôler l'exécution de procédures externes au système.

4.2- L'interface avec l'usager

L'environnement SCAI dispose d'une interface graphique qui permet de visualiser le processus de résolution, l'enchaînement des méthodes et les données utilisées, modifiées ou créées. Son rôle consiste à présenter à l'utilisateur non seulement la solution d'un problème mais aussi le processus de résolution suivi. L'utilisateur doit avoir la possibilité de demander des renseignements supplémentaires concernant la solution obtenue et le raisonnement qui l'a produite. L'interface graphique de SCAI est conçue de la façon suivante. Elle permet tout d'abord à l'utilisateur de définir son problème. Il choisit parmi toutes les tâches définies dans la base celle qui le concerne, il l'instancie et renseigne ses entrées. Puis c'est le moteur de tâche qui prend en charge le traitement et exécution de la

tâche. Il spécialise et décompose la tâche en question pour arriver à l'enchaînement nécessaire pour la résolution du problème.

Pendant la phase de spécialisation le système montre à l'utilisateur quelles spécialisations de la tâche ou méthode concernée sont possibles et adaptées, et lesquelles sont inappropriées. Les informations (entrées) manquantes sont déterminées soit en les demandant à l'utilisateur, soit en les calculant (tâches, attachements procéduraux...). Sur la base des propositions du système la spécialisation à exécuter est choisie à l'écran par l'utilisateur.

Le processus de spécialisation et de décomposition des tâches est visualisé en continu par un écran spécifique, l'écran de planification. Cet écran est actualisé dynamiquement au fur et à mesure du raisonnement. L'utilisateur peut ainsi voir à tout moment où en est le système dans la résolution du problème. L'écran contient l'arbre de spécialisation et de décomposition créé progressivement à partir de la tâche lancée initialement (**Figure-6**).

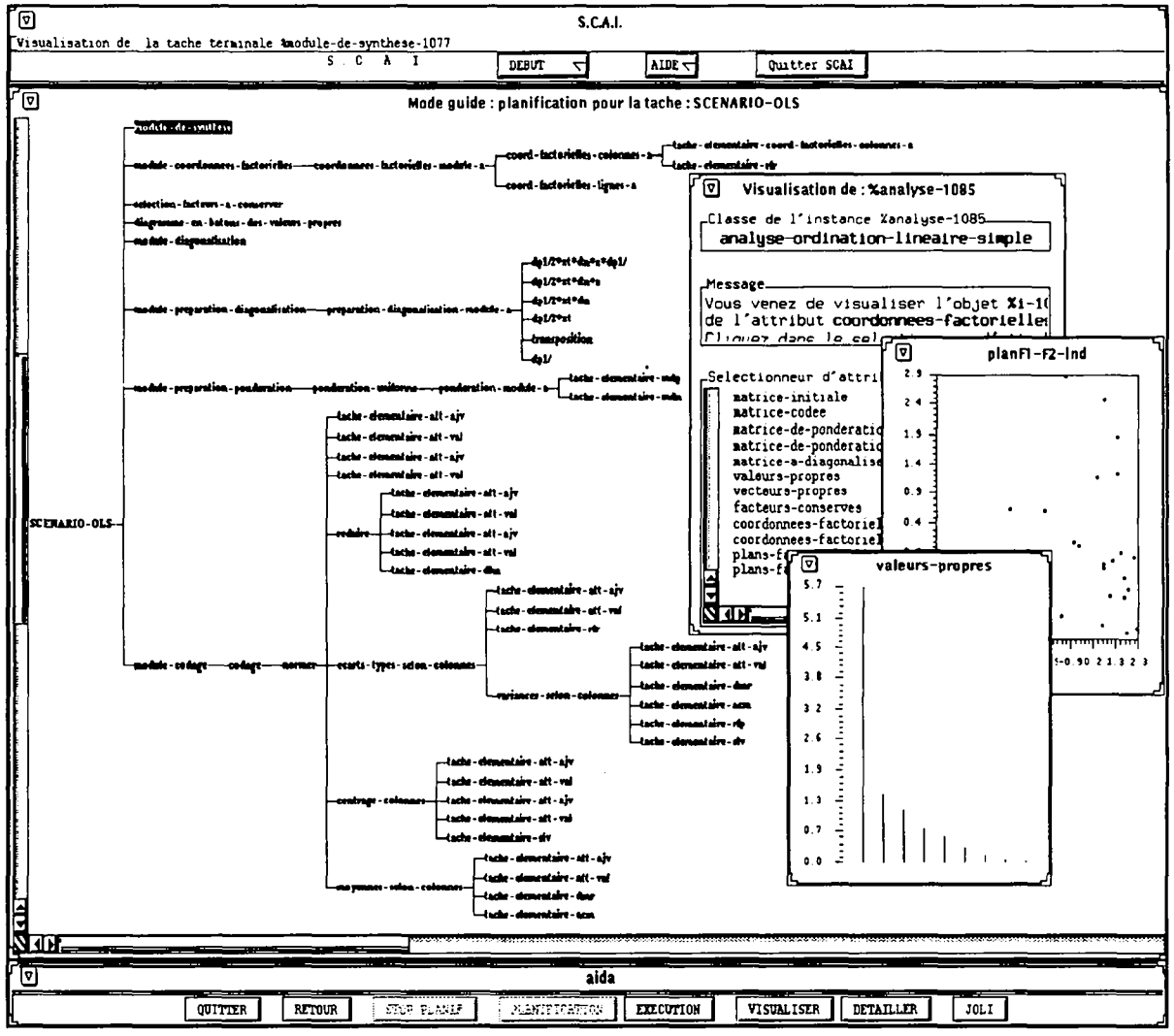

Figure 6. Ecran de planification : Cet écran visualise l'arbre de décomposition et de spécialisation qui a conduit à la résolution du problème SCENARIO-OLS, un problème d'ordination linéaire simple. Pendant la résolution différentes tâches graphiques sont activées : le diagramme en bâtons des valeurs propres après diagonalisation d'une matrice par exemple. Après la résolution l'utilisateur peut demander plus de renseignements sur les tâches exécutées ce qui lui permet en particulier d'accéder aux valeurs des attributs d'entrée et de sortie, ici par exemple l'instance %analyse-1085.

Chaque spécialisation ou décomposition se manifeste en une instanciation de la classe de la tâche ou de la méthode correspondante. Chaque noeud de l'arbre représente donc une de ces instances créées et traitées par le moteur. A la fin du raisonnement l'utilisateur peut consulter ces instances. Il les choisit sur l'écran et les visualise ensuite avec un éditeur approprié. Il a ainsi accès à toutes les informations concernant la tâche, en particulier ses entrées et ses sorties, c'est-à-dire les instances utilisées, modifiées ou créées par la tâche. Pour des tâches élémentaires l'utilisateur peut de même invoquer un écran graphique associé qui lui donne des informations sur la méthode exécutée.

5- DANAIDE

Le formalisme de représentation et les possibilités du système Shirka/SCAI appliqués à la construction d'un environnement de résolution de problèmes en analyse des données constituent un cadre d'étude particulièrement intéressant. D'une part, la diversité des méthodes exprimées par le modèle mathématique euclidien d'analyse des données offre de vastes horizons (numériques, graphiques et conceptuels) et d'autre part les techniques d'analyse multivariée sont sujettes à une forte demande dans de nombreux domaines d'applications.

Nous présentons dans ce chapitre la maquette DANAIDE en précisant son organisation générale, la formalisation des connaissances portant sur les objets manipulés en ordination linéaire simple ainsi que l'organisation en structures hiérarchiques des tâches qui les manipulent.

5.1- Domaine d'application

La première étape dans la mise au point d'une maquette consiste à identifier, spécifier le sous-problème à traiter, qui doit être suffisamment complexe pour tester les performances du système et significatif, représentatif du problème général. Nous avons porté notre réflexion sur les méthodes d'ordination linéaire simple (à un tableau): l'Analyse en Composantes Principales (ACP), l'Analyse Factorielle des Correspondances (AFC) et l'Analyse des Correspondances Multiples (ACM). Elles permettent une réflexion de fond suffisamment complexe pour tester le comportement du système tout en étant représentatives des méthodes multitableaux associées.

Les données à analyser sont sous la forme d'un tableau de n lignes et p colonnes difficile à appréhender en raison de sa nature multivariée: deux espaces vectoriels euclidiens à p et n dimensions contenant n et p vecteurs respectivement. Les méthodes d'ordination linéaire simple ont toutes pour objectif la recherche d'une représentation dans un espace de faible dimension (plans factoriels) généré par de nouveaux caractères synthétiques obtenus par combinaisons linéaires des caractères initiaux (*i.e.* méthodes factorielles et linéaires).

Les **Figures 7 et 8** [DOLE&91] montrent les différents modules de référence qui composent une ACP, une AFC ou une ACM. La première étape consiste en un codage des données (**Figure-7**) : définition d'un triplet (X, Dp, Dn). Il s'agit du passage des observations brutes (matrice Z: un tableau individus*variables, une table de contingence etc.) en observations tranformées (matrice X: centrage par colonnes, dénombrements->fréquences etc.) associées aux matrices diagonales des poids des lignes (Dn) et des colonnes (Dp). C'est cette étape de codage qui différencie l'ACP, l'AFC et l'ACM et qui est fonction de la nature de la matrice initiale Z et des objectifs poursuivis par l'utilisateur.

Tous les étapes suivantes sont communes (**Figure-8**): déterminer la matrice à diagonaliser, diagonalisation (génération d'une base orthonormée), calcul des coordonnées factorielles lignes et colonnes et enfin interprétations (graphiques essentiellement).

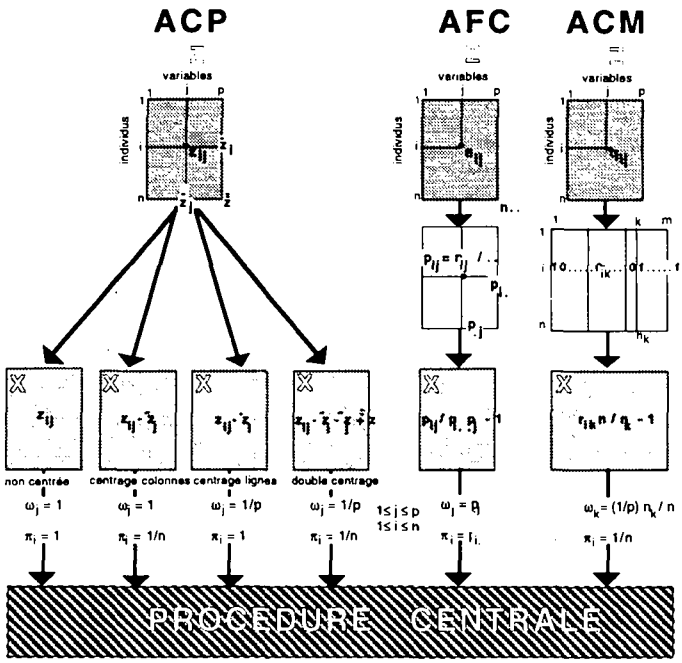

Figure-7 Codage, transformation des données initiales et matrices de pondération, [DOLE91].

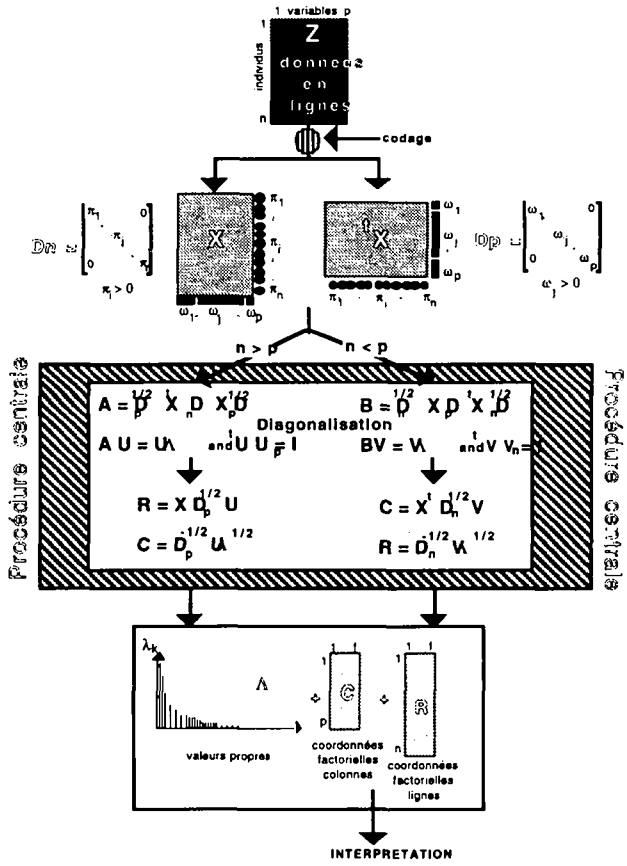

Figure-8 Procédure centrale de calcul d'une analyse (ACP, AFC, ACM): calcul de la matrice à diagonaliser, diagonalisation et calcul des coordonnées factorielles. Notations: Z: matrice des données brutes; X: matrice Z après codage; Dn, Dp: matrices diagonales de pondération lignes et colonnes; In, Ip: matrices identités; V: matrice des vecteurs propres, Λ: matrice des valeurs propres; R, C: matrices des coordonnées factorielles lignes et colonnes. Le choix de l'option n> p (nombre de lignes supérieur au nombre de colonnes) ou l'inverse n'a pas d'effet sur les résultats mais il permet de sélectionner la plus petite matrice à diagonaliser (n*n ou p*p), [DOLE91].

5.2- Organisation générale de la base

La seconde étape dans la mise au point de notre prototype consiste en une formalisation des connaissances en ordination linéaire simple: séparation des connaissances nécessaires à la résolution du problème de celles utiles pour justifier les raisonnements, des connaissances portant sur les objets manipulés et des connaissances portant sur les objets qui manipulent.

Le système se compose d'une base d'objets et d'une base de tâches et de méthodes.

La base d'objets contient la description, sous forme de schémas Shirka, des différents objets manipulés au cours d'une analyse (données brutes, données codées...). Toutes ces entités sont décrites par un schéma de classe: l'*objet-modèle*. C'est le principal objet dont les instances sont les fonctionnalités des tâches. Ces instances peuvent être créées soit par l'utilisateur, soit générées par les tâches elles-mêmes (enrichissement de la base).

Par ailleurs, un autre type d'objet décrit les différents types d'analyses (ACP...) ainsi que leurs paramètrages (ACP non centrée, ACP centrée par colonnes, ACP normée...), il s'agit du schéma de classe *objet-analyse* .

Nous décrivons plus en détail l'*objet-modèle* et l' *objet-analyse* dans le chapitre 5.3.

La base de tâches et de méthodes contient la description des algorithmes de calcul des différentes analyses sous forme de schémas SCAI. Le fonctionnement des tâches se réalise en combinant *tâches-séquentielles*, *tâches-de-choix*, *tâche-a-spécialiser* et finalement *tâches-élémentaires*.

Tous les schémas (*objet-modèle*, *objet-analyse*, *tâches et méthodes SCAI*) sont organisés en structures hiérarchiques: structuration de la connaissance au sein des objets manipulés, structuration fonctionnelle (spécialisations d'objectifs) au sein des tâches et structuration opérationnelle (spécialisations de réalisations) au sein des schémas méthodes (procédures).

5.3- La base d'objets

Nous discutons dans ce chapitre de la structuration des connaissances se rapportant aux différents types d'objets gérés par DANAIDE: l'*objet-modèle* et l'*objet-analyse*.

5.3.1- L'objet-modèle

Un *objet-modèle* est un objet composite intégrant cinq aspects différents sur les propriétés des entités manipulées par les tâches: *objet-conceptuel, objet-informatique*, *objet-graphique*, *objet-historique* et *objet-mathématique*. Chaque aspect est décrit au sein de l'*objet-modèle* par un attribut dont la valeur fait référence à un schéma, instance d'une classe appartenant éventuellement à une hiérarchie. Par exemple (**Figure-9**), l'instance %i-0 décrit les propriétés d'un jeu de données que l'on souhaite analyser (matrice Z). La valeur de l'attribut *objet-conceptuel* est une instance (%i-1) qui décrit le jeu de données selon le modèle conceptuel de l'utilisateur, ici une table individus-variables brute, de même la valeur de l'attribut *objet-mathématique* est une instance (%i-5) qui formalise ici les propriétés mathématiques de ce même jeu de données, une matrice rectangulaire. Les instances de l'*objet-modele* décrivent de la même façon toutes les entités manipulées au cours d'une analyse: le jeu de données après codage, ..., des coordonnées factorielles etc.

L'*objet-informatique* est l'aspect décrivant les propriétés informatiques des données : le nom du fichier les contenant, le type de ce fichier.... L'*objet-graphique* décrit les caractéristiques graphiques d'une ou plusieurs application(s) visualisée(s) par l'utilisateur: diagramme en bâtons, carte par caractères... Enfin, l'*objet-historique* mémorise l'origine de l'objet et la liste des tâches auxquelles il a été soumis.

Un aspect fondamental de la base consiste donc en la subdivision des connaissances en aspects indépendants (*objet-conceptuel, objet-mathématique*...), qui sont éventuellement décomposables en sous-points de vue, selon une organisation hiérarchique de la connaissance. Un extrait de la hiérarchie se rapportant à l'objet-conceptuel est présenté **Figure-10**.

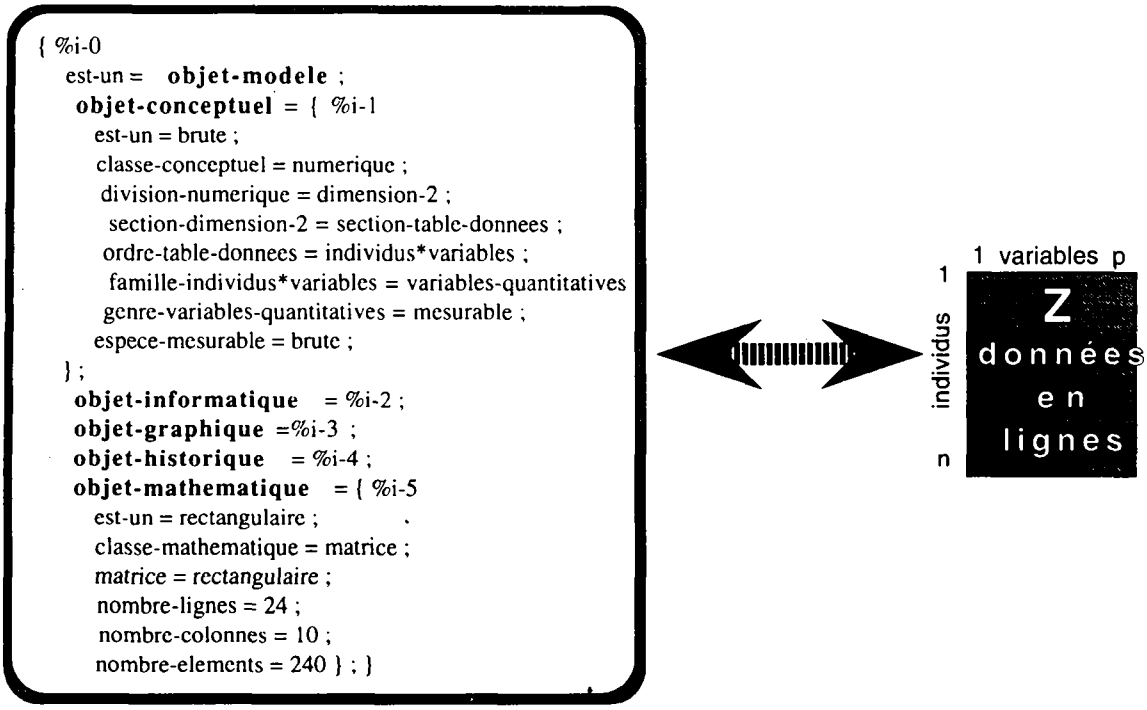

Figure-9. Dualité objet-modele/objet du monde réel en analyse des données : L'objet-modèle agrége différentes façons de considérer une même entité, ici un même jeu de données : ce sont les aspects conceptuel, - informatique, - graphique, - historique, et mathématique.

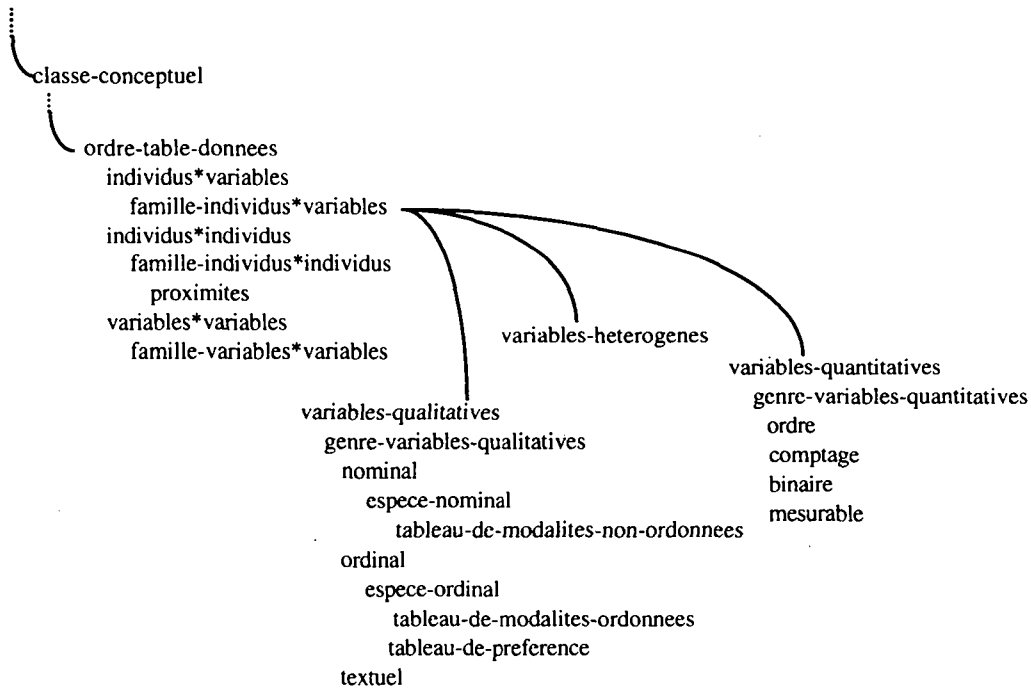

Figure-10. Extrait de la hiérarchie correspondant à l'aspect *objet-conceptuel* . La classification tente de réaliser une partition complète: c'est à dire un découpage exhaustif de l'ensemble en classes exclusives.

5.3.2- L'objet-analyse

Un autre type d'objet que l'*objet-modèle* a pour rôle de synthétiser, résumer tous les résultats essentiels concernant un type d'analyse répertorié. Il s'agit de *l'objet-analyse* qui se spécialise selon les méthodes. La **figure-11** montre par exemple l'objet ACP.

```
{acp
 sorte-de                          =          objet-analyse;
 matrice-initiale                  $un        objet-modele;
 matrice-codee                     $un        objet-modele;
 matrice-a-diagonaliser            $un        objet-modele;
 matrice-de-ponderation-lignes     $un        objet-modele;
 matrice-de-ponderation-lignes     $un        objet-modele;
 valeurs-propres                   $un        objet-modele;
 vecteurs-propres                  $un        objet-modele;
 facteurs-conserves                $un        entier;
 coordonnees-factorielles-lignes   $un        objet-modele;
 coordonnees-factorielles-colonnes $un        objet-modele;
 plans-factoriels-lignes           $liste-de  objet-modele;
 plans-factoriels-colonnes         $liste-de  objet-modele}
```

Figure-11. L'objet ACP, spécialisation de l'objet-analyse : Cet objet résume tous les résultats obtenus en cours d'une analyse en composantes principales.

Les attributs de cette classe référencent les caractéristiques essentielles de l'analyse en composantes principales. Au sein de leurs instances, les valeurs prises par les attributs sont des instances d'*objet-modèle* générés par les tâches.

5.4- La base de tâches et de méthodes

Si les tâches élémentaires et les méthodes associées permettent de coupler au système des programmes de tous types, la version actuelle de la base ne pilote que des procédures au contenu sémantique élémentaire (addition de deux matrices, division des éléments d'une matrice par un réel...). C'est cette granularité suffisamment importante qui nous permet de définir un mode de fonctionnement en "stand-alone" particulier, un aspect "programmathèque intelligente" qui concerne aussi bien des tâches élémentaires de calcul numérique que graphique. La définition de tâches élémentaires au sens de DANAIDE demande une refonte importante des programmes actuels, une extraction des connaissances vis-à-vis des calculs numériques n'ayant pour responsabilité que la rapidité et la fiabilité. Cette recherche d'une base de tâches élémentaires permet au système de disposer d'un potentiel de création agréable à gérer. Par exemple la définition de quelques tâches élémentaires graphiques donne la possibilité de générer des applications complexes par constructions pyramidales et en ne manipulant que des schémas. Ainsi, la génération d'un graphe x-y avec boutons de changement d'échelle, d'affichage de graduation etc., n'est pas une tâche élémentaire mais une tâche séquentielle, modulable par l'utilisateur selon les tâches de choix intégrées.

La description d'un objet complexe via une classe intégrant différents aspects importants permet de véhiculer simplement des objets porteurs de suffisamment d'information pour les inférences et les procédures externes associées aux tâches. Les instances d'*objet-modèle* sont le principal type de données traité par les tâches de DANAIDE. Un rôle essentiel de ces tâches, autre que celui de modélisation d'algorithmes, consiste à générer des instances d'*objet-modèle* (une ou plusieurs selon les tâches). Cet aspect d'enrichissement de la base concerne chaque tâche, certaines d'entre elles étant d'ailleurs spécialisées dans cette fonction (génération d'objets particuliers de type relationnel, ou d'objets correspondant à des tâches complexes, exemple de l'*objet-analyse*).

Sans rentrer dans trop de détails techniques nous expliquons le principe de génération de ces objets en l'illustrant sur une tâche élémentaire. Nous développons ensuite l'organisation en structures hiérarchiques des tâches de la base.

Pour générer une instance d'*objet-modèle*, il est nécessaire de générer des instances pour chaque classe (*objet-conceptuel*, *objet-mathématique*...), et en présence de hiérarchie, chercher la sous-classe de rattachement qui décrit le mieux l'objet sous cet

aspect. Les noms de ces sous-classes sont des valeurs qui sont traitées comme des entrées par les tâches, elles sont inférées par défaut selon une échelle d'attribution fonction du niveau d'instanciation de la tâche ou calculées par attachement procédural.

```
{tache-elementaire-mmm
sorte-de            =           operations
                    $com        "MULTIPLICATION DE 2 MATRICES";
est-un              =           tache-term;
lui-meme            $var-nom    lui;
&entree me1         $un         objet-modele
                    $com        "premiere matrice";
&entree me2         $un         objet-modele
                    $com        "deuxieme matrice";
classe-conceptuel   $un         objet-conceptuel
                    $defaut     dimension-2;
classe-mathematique $un         objet-mathematique;
                    $a-verifier {predicat-nb11=nbc2
                                objet-modele-1  $var<-  me1;
                                objet-modele-2  $var<-  me2}
                    $si-succes  {stocker
                                instance        $var<-  lui;
                                attribut        $valeur cm-1;
                                valeur          $var<-  carrée}
                    $si-echec   {stocker
                                instance        $var<-  lui;
                                attribut        $valeur cm-1;
                                valeur          $var<-  rectangulaire}
&sortie ms1         $un         objet-modele
                    $var-nom    s1
proc                =           {sorte-de       =       procedure-mmm;
                                e1              $var<-  me1;
                                e2              $var<-  me2;
                                e3              $var<-  classe-conceptuel;
                                e4              $var<-  classe-mathematique;
                                s1              $var->  ms1}}
```

Figure-12. Exemple de tâche élémentaire : Le produit de deux matrices. Les entrées me1 et me2 ont pour valeur les deux instances d'objet-modele décrivant les matrices à multiplier. Le résultat, également une instance d'objet-modele, est stocké dans l'attribut ms1. Les valeurs des attributs classe-conceptuel et classe-mathémathique décrivent les propriétés de l'instance générée.

Dans l'exemple de la **Figure-12** (Le produit matriciel), la valeur attribuée par défaut à l'entrée *classe-conceptuel* est le nom de la sous-classe pour laquelle il faut créer une instance, valeur de l'attribut *objet-conceptuel* de l'*objet-modèle* généré. Une instanciation à un niveau supérieur permet de disposer d'une information plus précise, et donc d'attribuer une sous-classe de rattachement plus spécialisée dans la hiérarchie correspondante. Par exemple le produit matriciel utilisé dans le contexte du calcul d'une matrice de variances-covariances. Un autre mode de détermination du nom de la sous-classe de rattachement est attribué à l'entrée *classe-mathematique* . Dans cet exemple, cette valeur est fonction du résultat déterminé par attachement procédural et qui consiste à comparer le nombre de lignes de la première matrice avec le nombre de colonnes de la seconde. En cas d'égalité, la sous-classe de rattachement décrit une matrice carrée, dans le cas contraire la matrice résultant du produit est considérée comme rectangulaire. La distinction entre les attributs *sorte-de* et *est-un* permet de comprendre le fait qu'une tâche peut être une spécialisation d'une autre tout en ayant une décomposition spécifique et indépendante.

DANAIDE distingue trois types d'applications: les applications de gestion, les applications de calcul graphique et les applications de calcul numérique. **Figure-13** montre un exemple d'organisation hiérarchique de tâches.

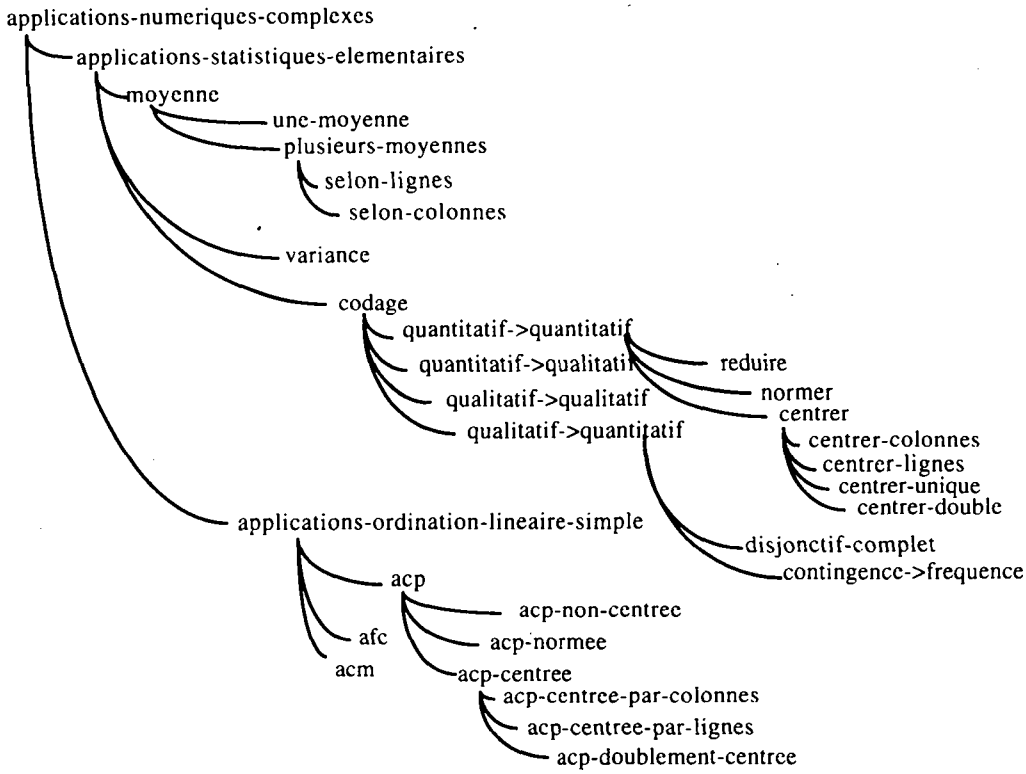

Figure-13. hiérarchie de tâches décrites dans DANAIDE : Cette hiérarchie décrit les différentes applications numériques complexes de la base.

La tâche *applications-numeriques-complexes* est une tâche à spécialiser. La tâche *applications-ordination-lineaire-simple* dispose elle d'une décomposition permettant différents types d'analyses alors que ses sous-classes correspondent à des objectifs plus précis. Par exemple si l'utilisateur expérimenté souhaite contrôler toutes les étapes d'une ACP, il instancie la tâche *ordination-linéaire-simple* qui séquence quelques tâches mais qui sont complexes . A l'inverse, le choix d'une tâche spécialisée comme l'*ACP-normee* séquence directement de nombreuses tâches qui sont élémentaires, à ce niveau l'utilisateur n'intervient que très peu.

6- Conclusion

DANAIDE est une application à l'ordination linéaire simple du formalisme de représentation de connaissances par objet et des mécanismes de raisonnement par tâches développés dans les systèmes Shirka et SCAI. La construction d'environnements de résolution de problèmes en analyse des données est une solution aux problèmes générés par le triplet modèles mathématiques, programmes , utilisateurs.

Le potentiel dynamique de la base se formalise de plusieurs façons. D'une part il est toujours possible pour l'utilisateur de chaîner manuellement et via l'interface graphique les tâches qu'il souhaite afin de répondre à son objectif particulier. D'autre part l'organisation hiérarchique de tâches complexes permet de proposer à l'utilisateur d'interagir dans les choix méthodologiques à différents niveaux de complexité (voir exemple de la tâche *applications-ordination-lineaire-simple*).

Un autre intérêt de la définition de tâches vient de la description rapide de synonymes. Par exemple les tâches élémentaires de centrage unique et soustraction des éléments d'une matrice par un réel font appel à la même procédure. Cette information n'est pas redondante du point de vue de l'utilisateur.

Nous souhaitons introduire un premier niveau d'expertise en assistant l'utilisateur confronté au choix d'une tâche parmi plusieurs. Un mécanisme d'identification de tâches pour un *objet-modèle* donné est à l'étude. Il est fondé sur l'association de l'inférence par classification sur classes d'attribut et de l'attachement procédural. Le processus se concrétise par une tâche SCAI, qui peut donc être instanciée soit par l'utilisateur, soit par une autre tâche. Les modalités peuvent être les suivantes: identification selon la sémantique de la tâche, selon le domaine d'application de la tâche, selon l'indice de complexité de la tâche ou encore selon l'historique de l'instance *objet-modèle*. Ainsi, un exemple d'application peut consister à dire qu'il est inutile d'identifier une tâche si celle-ci a déjà été instanciée pour l'objet concerné, ou qu'il est possible d'identifier une tâche si et seulement si une séquence précise de tâches a déjà été effectuée. Ce mécanisme peut ainsi permettre l'introduction d'un menu dynamique et issue d'une expertise.

L'application symétrique qui consiste à identifier un ou plusieurs objet(s) pour une tâche donnée peut aussi être développée selon le même principe. Il peut s'agir par exemple de l'identification d'objets prototypiques avec lesquels l'utilisateur peut faire des comparaisons.

C'est aussi une de nos préoccupations que de définir dans un premier temps un noyau de calcul (numérique et graphique) capable dans un deuxième temps de se spécialiser dans un domaine d'application particulier. Au laboratoire de Biométrie, un exemple de d'application peut se réaliser dans le domaine de la biologie moléculaire par le couplage de ColiGène [PERR91] et DANAIDE. L'application génère de nouveaux points de vue, par exemple un objet-biologique, qui s'intègrent dans une nouvelle spécialisation d'*objet-modèle*.

BIBLIOGRAPHIE

[ADER91] H.J. ADER.
Formalizing Statistical Expert Knowledge.
In, Diday E. and Lechevallier Y. (Eds). *Symbolic-Numeric Data Analysis and Learning.*
Nova Science Publishers, New York, 1991.

[AIVA91] S. AIVAZIAN.
Instruments mathématiques et logiciels pour la construction de systèmes experts dans une discipline.
In, Diday E. and Lechevallier Y. (Eds). *Symbolic-Numeric Data Analysis and Learning.*
Nova Science Publishers, New York, 1991.

[AUDA83] Y. AUDA.
Rôle des méthodes graphiques en analyse des données: application au dépouillement des enquètes écologiques.
Thèse 3° cycle.
Université Claude Bernard, Lyon 1, France, 127 p., 1983

[CHAI86] J. CHAILLOUX, M. DEVIN, F. DUPONT, J.M. HULLOT, B. SERPETTE, J.VUILLEMIN.
Le_Lisp Version 15.2,*Manuel de référence.*
INRIA, Rocquencourt, France.

[CHES&91] D. CHESSEL, S. DOLEDEC.
ADE Software. Multivariate Analyses and Graphical Display for Environmental Data. Version 3.1.
User's Manual, 1991

[DOLE&91] S. DOLEDEC, D. CHESSEL.
Recent developments in linear ordination methods for environmental sciences.
In *Trends in Ecology, Council of Scientific Research.Integration Ed.*
Research Trends (Publishers), India. 1991.

[ESCO87] Y. ESCOUFIER.
The duality diagram: a means for better practical applications.
In, Legendre P. and Legendre L. (Eds) *Developments in Numerical Ecology.*
NATO ASI Series G (Ecological Sciences) Springer Verlag, Berlin, pp. 139-156, 1987.

[FIKE&85] R. FIKES, T. KEHLER.
The role of frame-based representation in reasoning.
Comm. ACM, 28, n°9, pp.904-920, 1985.

[GRIV92] S. GRIVAUD.
IVAN : Interface de Visualisation et d'Aide à la Navigation dans une base de connaissances à objets.
Mémoire d'ingénieur CNAM, Grenoble, France, à paraître début 1992

[ILOG91] ILOG.
AIDA: un environnement de développement d'interfaces graphiques.
Manuel d'utilisation.
Gentilly, France, 1991.

[MASI&90] G. MASINI, A. NAPOLI, D. COLNET, D. LEONARD, K. TOMBRE.
Les langages à objets: langages de classes, langages de frames, langage d'acteurs.
Collection iia, InterEditions, 584p., 1990.

[MINS75] M. MINSKY.
A Framework for Representing Knowledge.
In *The psychology of Computer Vision*, Winston P.H. Ed., McGrawHill, 1975.

[PERR91] G. PERRIERE.
Utilisation d'une base de conaissances centrée-objet pour l'étude de l'expressivité des gènes de *E.Coli*
Actes du colloque IMABIO. Paris 2-3 avril, 1991.

[RECH&90] F. RECHENMANN, P. UVIETTA.
Shirka: système de gestion de bases de connaissances centrées-objet.
Manuel d'utilisation.
Laboratoire Artémis, INRIA, Grenoble, France, 1990.

[RECH&91] F. RECHENMANN, P. UVIETTA.
Shirka: an object-centered knowledge based management system.
In *Artificial Intelligence in Numerical and Symbolic Simulation*. Pavé A. and Vansteenkiste eds.
Aléas, Lyon, France, pp. 9-23, 1991.

[ROUS88] B. ROUSSEAU.
Vers un environnement de résolution de problèmes en biométrie, Apport des techniques de l'intelligence artificielle et de l'intéraction graphique.
Thèse de doctorat.
Université Claude Bernard, Lyon 1, France, 252p., 1988.

[THIO&90] J. THIOULOUSE, J. DEVILLERS, D. CHESSEL, Y. AUDA.
Graphical techniques for multidimensional data analysis.
In Devillers J. and Karcher W. (eds), *Applied Multivariate Analysis in SAR and Environmental Studies*, pp. 153-205, 1990.

[THIO90] J. THIOULOUSE.
MacMul and GrapMu: two Macintosh programms for the display and analysis of multivariate data.
Computers and Geosciences, 8, pp. 1235-1240, 1990.

[UVIE&91] P. UVIETTA, J. WILLAMOWSKI.
Modélisation des connaissances en Biologie Moléculaire
Congrès RFIA 27-29 novembre 1991

Positivité en biologie et convergence vers l'équilibre.

Jean-Luc Gouzé
INRIA Sophia-Antipolis
BP 109
06561 Valbonne Cédex

Résumé:
Les modèles mathématiques en biologie utilisent très souvent des variables positives. Nous montrons, pour le modèle différentiel linéaire et pour le modèle de Lotka-Volterra en dimension n, qu' imposer a priori la positivité "stricte" des variables implique un comportement régulier du modèle; de plus, supposer l'existence d'un équilibre positif, implique alors la convergence globale des solutions vers cet équilibre dans tout l'espace.

Abstract:
Mathematical models of biological phenomena have often non-negative variables (numbers, concentrations,...). We show, for the differential linear case and the Lotka-Volterra generalized model in dimension n, that, if the model is built with these a priori "strict" positivity constraints, then the solutions have a regular behaviour (neither periodic stable solutions, nor chaos ...); if one suppose moreover that a positive equilibrium exists, then all the solutions converge towards this equilibrium in the whole space.

1. Introduction:

Quand on construit un modèle mathématique d'un phénomène biologique, on est amené à définir et à manipuler des variables mathématiques qui représentent, dans une certaine unité de mesure, une quantité biologique. Le plus souvent, cette quantité sera a priori positive; par exemple, en dynamique des populations, en biochimie et en chimie, on manipule des nombres ou des concentrations.

Le biologiste sait donc a priori que les variables mathématiques qu'il manipule n'auront de sens (par rapport à la réalité) que si elles restent positives, ou éventuellement nulles. Pour construire le modèle, il a alors le choix entre deux attitudes possibles:

- soit il construit un modèle mathématique quelconque (c'est à dire ne respectant pas les contraintes de positivité sur les variables) ; il doit alors définir un domaine de validité du modèle, dans lequel les variables resteront positives. Si une solution sort de ce domaine, elle n'a plus de sens biologique.
- soit il construit un modèle où, mathématiquement, les solutions restent toujours positives.

Nous ne voulons pas entrer ici dans la discussion (difficile) du choix entre ces deux attitudes. Dans cet article, nous choisirons la seconde, qui est évidemment la plus contraignante mathématiquement.

Ce choix de positivité pose donc des contraintes au modèle; nous traduirons ces contraintes dans le cas de deux exemples: le modèle différentiel linéaire et celui de Lotka-Volterra généralisé. Il est intéressant de constater que ces contraintes ont des implications fortes sur le comportement des solutions, en excluant par exemple tout comportement de type périodique, récurrence ou chaos. Pour être honnête, il faut cependant dire que , dans le cas de Lotka-Volterra, nous imposons des contraintes de positivité "robustes" plus fortes que les contraintes usuelles.

Une autre hypothèse a priori fréquemment faite par le biologiste est que le modèle admet un point d'équilibre, car cela correspond à l'observation expérimentale d'un état stationnaire. Cette hypothèse a aussi des conséquences mathématiques. Toujours dans l'exemple linéaire et de Lotka-Volterra, nous montrons alors que l'existence mathématique d'un équilibre (positif bien sûr), ajoutée à l'hypothèse de positivité, implique alors que toutes les solutions convergent dans tout l'espace vers le point d'équilibre, qui est donc globalement stable.

Dans la suite, nous ne considérerons que des modèles définis par des équations différentielles ordinaires. Nous mettrons plutôt l'accent sur les idées intuitives que sur les démonstrations mathématiques. Nous travaillerons dans un espace de dimension n (il y a n variables): nous appelerons positif un vecteur n-dimensionnel dont toutes les composantes sont positives. Nous appellerons orthant positif la région de l'espace formée par les vecteurs positifs (en dimension deux, c'est le "quadrant" positif); rappelons qu'en français, "positif" veut dire "positif ou nul".

Pour fixer les idées, nous nous placerons dans le cadre de la dynamique des populations, mais cela n'a pas d'importance. Pour des exposés sur les modèles en biologie, on peut consulter (Lebreton et Millier 1982) et (Hofbauer and Sigmund 1988); pour la théorie des matrices positives, voir (Berman and Plemmons 1979); pour une revue des résultats sur les systèmes coopératifs, voir (Smith 1988).

1. Le modèle linéaire:

Ce modèle est un peu un cas d'école, car il semble qu'en biologie les modèles les plus intéressants soient non-linéaires. Il permet cependant d'exposer les outils utilisés et les idées de base.

Ce modèle s'écrit donc:

$$x' = Ax + b$$

où X et b sont deux vecteurs n-dimensionnels et A une matrice carrée à n lignes et n colonnes.

L'hypothèse de positivité est donc que, si on part d'une condition initiale x(0) positive, alors la solution x(t) de l'équation différentielle restera positive pour tout temps t positif. Il est facile de voir que cette condition est équivalente au fait que, sur chacune des faces $x_i = 0$ de l'orthant positif, le champ défini par l' équation différentielle est répulsif, et donc que la solution ne peut pas franchir cette face.

On traduit cela mathématiquement par:

$x_i = 0 \Rightarrow x'_i \geq 0$, les autres x_j étant quelconques positifs

Voyons ce que cela implique pour l'équation linéaire; d'abord, si on est au point origine 0, la condition ci-dessus donne que $b \geq 0$. La condition s'écrit alors:

$$\sum_{i \neq j} A_{ij} x_j + b_i \geq 0$$

Comme cette condition doit être vraie pour tout les x_j, cela implique que $A_{ij} \geq 0$ pour tout i différent de j. Supposons en effet qu'il existe un tel A_{ij} strictement négatif, alors en prenant un x_j assez grand, on pourrait toujours rendre l'expression négative.On a donc:

Proposition 1. Pour le modèle linéaire, la positivité implique que le vecteur b est positif et que la matrice A est positive, sauf éventuellement sur la diagonale.

Dans ce cas linéaire, on sait de plus que la solution ira soit vers l'équilibre soit partira à l'infini. Le comportement est très restreint à la base de par la formulation linéaire. Voyons maintenant le rôle de l'équilibre. L'hypothèse a priori est qu'il existe un unique équilibre

positif; nous avons ajouté le mot unique pour simplifier les choses; si l'équilibre n'est pas unique, il y a une infinité d'équilibre, ce qui est une situation dégénérée (voir cependant la remarque à la fin de cette section).

La traduction est donc que l' équation linéaire

$$Ax = -b$$

admet une solution x positive; de plus A est bijective pour avoir l'unicité. Mais d'après ci-dessus, b est positif et A positive sauf sur la diagonale. Un théorème de matrices positives (Berman et Plemmons 1979) nous dit alors que l'équation a une unique solution positive si et seulement si -A est une M-matrice, et que cela implique que A est stable. Donc:

Proposition 2. Sous l'hypothèse de positivité, l'existence d'un équilibre positif implique la stabilité de cet équilibre .

Géométriquement, on peut mettre en évidence des régions invariantes et des rectangles (n-dimensionnels) contractants invariants: si on part d'un tel rectangle, on ne peut plus en sortir; le rectangle de plus se contracte au cours du temps jusqu'à se réduire au point d'équilibre. Pour trouver un tel rectangle autour du point d'équilibre, on trace la droite passant par l'origine et l'équilibre; alors deux points sur cette droite de part et d'autre du point d'équilibre définissent un rectangle invariant contractant (figure 1).

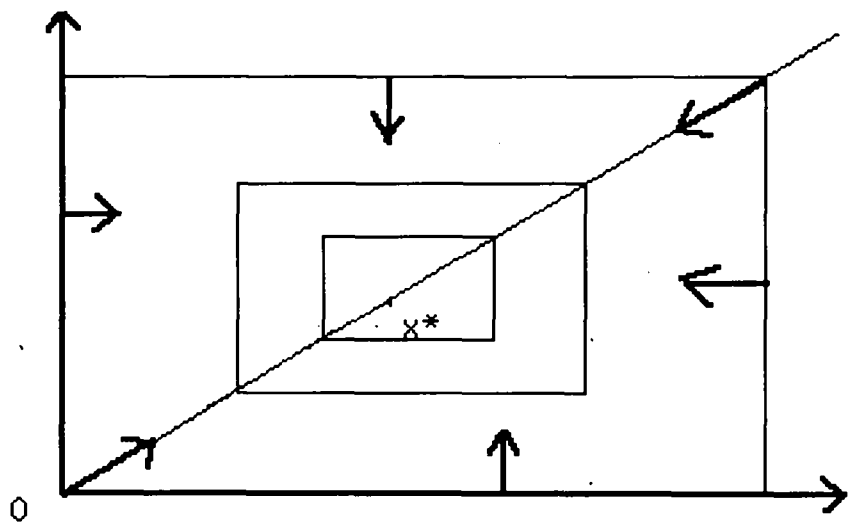

Figure 1

Remarques:

- l'existence d'un équilibre implique en particulier que la matrice A est négative sur la diagonale.
- s'il n'y a pas d'équilibre positif, alors toutes les solutions partent à l'infini, puisqu'elles ne vont pas vers l'équilibre.
- comme il est dit plus haut, on a supposé A bijective pour avoir unicité de l'équilibre; cependant il y a un cas intéressant où A n'est pas bijective, c'est celui d'un système conservatif. En analyse compartimentale par exemple, on a des modèles avec conservation de la masse, ou du nombre total de substance, et donc l'équation s'écrit:

$$x' = Ax$$

où la matrice A a la somme de ses colonnes nulles (la quantité Σx_i se conserve au cours du temps). On a alors une droite d'équilibre, mais la solution reste sur la surface $\Sigma x_i = C$, et donc finalement, on a un équilibre unique sur cette surface. L'hypothèse de positivité et l'existence d'un équilibre conduisent exactement au même résultat.

2. Le modèle généralisé de Lotka-Volterra:

Le plus souvent, on entend sous ce nom le modèle en dimension 2 qui s'écrit:

$$x' = ax - bxy \qquad (1)$$
$$y' = cxy - dy$$

et dont les solutions sont toutes périodiques dans le quadrant positif; il représente une interaction de type prédateur-proie (figure 2).

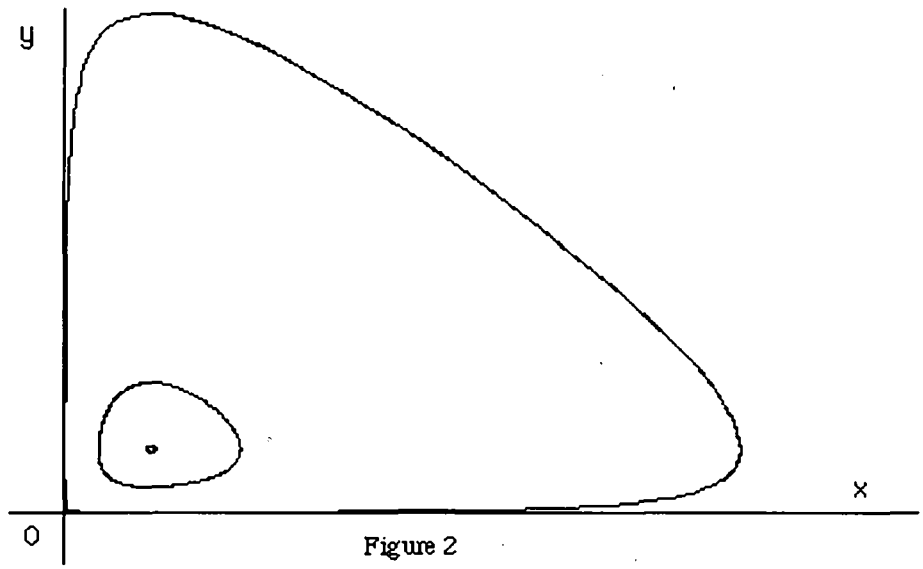

Figure 2

Nous considérerons ici un modèle beaucoup plus général en dimension n qui s'écrit:

$$x'_i = x_i \left(\sum_{j=1}^{n} A_{ij}\, x_j + b_i \right) \qquad i = 1...n$$

La matrice A représente les interactions entre les espèces (par exemple $A_{ij} \leq 0$ veut dire que l'espèce j fait diminuer l'espèce i; $A_{ji} \geq 0$ veut dire que l'espèce i fait augmenter l'espèce j; c'est donc un phénomène de prédation). Le signe des coefficients de la matrice A est lié à la nature biologique des interactions (prédation, mutualisme, compétition). Ces interactions sont de nature quadratique ($x_i x_j$). Le terme b_i représente une croissance ou une décroissance exponentielle de l'espèce isolée.

On sait beaucoup de choses sur le plan mathématique en ce qui concerne ce système (Hofbauer and Sigmund 1988); on sait en particulier qu'il peut avoir un comportement très compliqué, avec des cycles limites ou du chaos.

En ce qui concerne la positivité, il semble que les choses soient simples: en effet, d'après les équations, sur une face $x_i = 0$, on a $x'_i = 0$; les faces sont donc invariantes, on ne pourra donc pas les franchir d'après la propiété d'unicité de la solution, et l'intérieur de l'orthant positif l'est donc aussi. D'un point de vue formel, il n'y a plus rien à dire: de par leur structure, les modèles de Lotka- Volterra vérifient l' hypothèse de positivité.

Cependant, on peut exiger plus de la positivité pour rester réaliste biologiquement: en effet, sans hypothèses supplémentaires, une solution de Lotka-Volterra peut aller aussi près que l'on veut d'une face où une variable s'annule, puis en repartir; le système proie-prédateur (1), par exemple, se comporte comme cela (figure 2). Biologiquement, cela n'a pas beaucoup de sens, car une quantité biologique ne peut être arbitrairement petite: en effet, les quantités mesurent toujours des populations d'individus, et sont forcément, à petite échelle, discrètes. La modélisation par équation différentielle n'a alors plus beaucoup de sens. Nous ne poursuivrons pas ici cette intéressante discussion; nous exigerons donc une positivité plus "robuste", et imposerons que toutes les variables restent à une distance finie (éventuellement petite) des faces de l'orthant. C'est ici une énorme hypothèse que fait le biologiste: il suppose a priori qu'aucune des espèces ne deviendra trop petite; il suppose donc a priori le système complet viable, et qu'aucune espèce n'ira vers l'extinction.

Une traduction mathématique simple possible est la suivante: on se fixe un seuil S au dessous duquel il est interdit de tomber; on écrit alors que sur la surface $x_i = S$, le champ est répulsif, c'est à dire qu'il fait augmenter x_i. On obtient l'équation:

$$S\left(A_{ii}\, S + \sum_{j \neq i} A_{ij} x_j + b_i \right) \geq 0 \qquad \text{pour tout i et tous les } x_j \geq S$$

ce qui implique que $A_{ij} \geq 0$ pour $i \neq j$, par le même raisonnement que dans le cas linéaire. Donc on retrouve le fait que la matrice A est positive sauf éventuellement sur la diagonale. Cette propriété de A est importante mathématiquement, car elle signifie que le champ est coopératif (voir Smith 1988), c'est à dire que les éléménts en dehors de la diagonale de la matrice jacobienne sont positifs. Des résultats mathématiques disent alors que les solutions ont un comportement régulier, au sens où, en gros, soit elles partent à l'infini, soient elles convergent vers le point d'équilibre; donc cela exclut des comportements du type solutions périodiques stables, chaos,... Donc:

Proposition 3. L'hypothèse de positivité robuste dans le modèle de Lotka-Volterra implique un comportement régulier du système.

Remarquons cependant qu'il n' y a pour l'instant pas de contraintes sur les b_i. Traduisons mainteant le fait qu'il y ait un point d'équilibre positif x^* (on fera encore ici l'hypothèse usuelle de l'unicité de l'équilibre et donc de la bijectivité de la matrice A). On a :

$$A\, x^* + b = 0$$

Mais par hypothèse, le champ est positif au point (S,S,...,S) noté s ; c'est le point le plus "petit" du domaine admissible (figure 3); donc

$$A\, s + b \geq 0$$

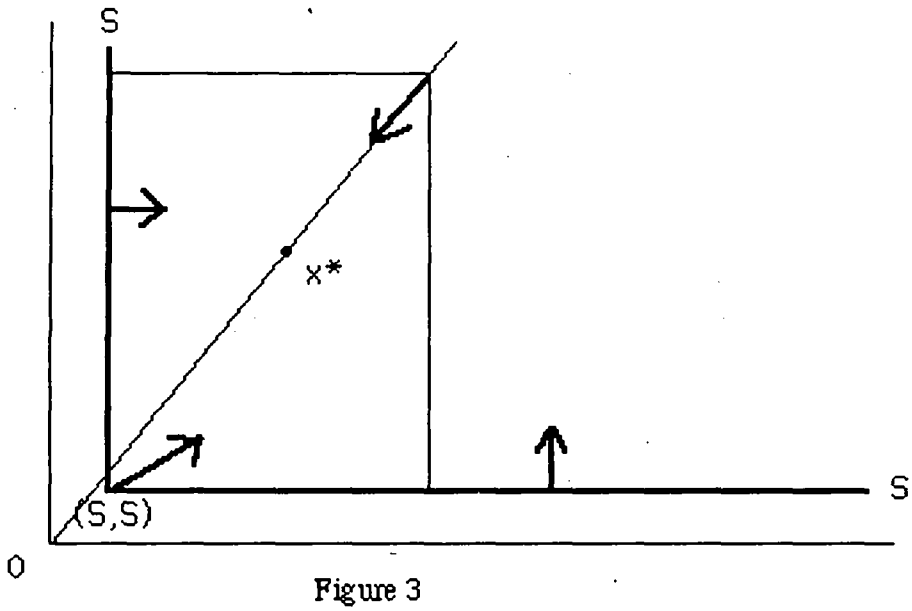

Figure 3

Si on soustrait les deux équations, on obtient:

$$A (s - x^*) \geq 0$$

et le vecteur $(s - x^*)$ est évidemment négatif, car x^* est à l'intérieur du domaine admissible. On retombe donc, comme dans le cas linéaire, sur une équation du type

$$A u = v$$

où u est négatif et v positif, et A positive en dehors de la diagonale. Donc -A est encore une M-matrice et elle est donc stable, ce qui ici ne nous donne que la stabilité locale de l'équilibre. Mais on a beaucoup plus, car on peut encore trouver des rectangles contractants invariants; le point inférieur sera le point s, et le point supérieur pourra être pris égal à αx^*, pour un α positif plus grand que 1 et aussi grand que l'on veut; le champ en ce point est alors négatif. La théorie des champs coopératifs dit alors que ce rectangle est contractant invariant; on a donc la stabilité globale dans tout le domaine admissible. Donc:

Proposition 4. L'hypothèse d'existence d'un équilibre entraîne, sous les conditions de positivité robuste, la stabilité globale de cet équilibre dans tout le domaine.

Remarques:

- les rectangles invariants contractants impliquent aussi la bornitude de la solution.

- les conditions de positivité trouvées dans les deux cas ne dépendent que du signe des éléments et pas de leur valeur numérique.

- dans le cas de Lotka-Volterra, on a fait une hypothèse très forte de survie de toutes les espèces, et on obtient alors que toutes ces populations doivent n'avoir que des interactions de type mutualisme: il ne doit y avoir aucun prédateur.

- dans le cas de Lotka-Volterra, si l'hypothèse de positivité robuste n'est pas vérifiée (comme dans le modèle classique (1)), alors on est sûr que, suivant les conditions

initiales, il peut y avoir des trajectoires qui s'approchent aussi près que l'on veut des faces de l'orthant. Il faut alors se poser la question de la réalité biologique de ces trajectoires.

3.Conclusion:

Dans ces deux exemples, on a montré que les hypothèses (parfois augmentées) de positivité vers l'équilibre impliquaient des contraintes fortes sur la structure et le comportement des modèles. Bien sûr, il s'agissait plutôt de soulever des problèmes liés à la modélisation, et nous n'avons apporté que peu de réponses: nous n'avons rien dit sur les autres classes de modèles; on n'aura pas alors de résultats aussi forts.

De plus, dans le cas de Lotka-Volterra, on a dû renforcer l'hypothèse habituelle de positivité; il nous paraît néanmoins intéressant de réfléchir plus avant sur le rôle des hypothèses a priori dans la modélisation. Nous pensons personnellement qu'il est parfois plus convaincant de se donner des hypothèses biologiques a priori, puis de traduire ces hypothèses biologiques en propriétés mathématiques, puis de construire des modèles mathématiques vérifiant ces propriétés.

Références

<1> A. Berman and R.J. Plemmons, "Nonnegative matrices in the mathematical sciences" Academic Press (1979)
<2> J. Hofbauer and K. Sigmund, "The theory of evolution and dynamical systems", Cambridge University Press (1988)
<3> J.D. Lebreton et C. Millier, "Modèles dynamiques déterministes en biologie", Masson (1982)
<4> H.L. Smith, Systems of ordinary differential equations which generates an order preserving flow, SIAM Review, 30, pp. 87-113 (1988)

COLIGENE — EXEMPLE D'UNE BASE DE CONNAISSANCES CENTREE-OBJET POUR L'ETUDE DE L'EXPRESSIVITE DES GENES DE *E. COLI*

G. Perrière et C. Gautier

Laboratoire de Biométrie, URA CNRS 243, Université Claude Bernard Lyon I, 43 Boulevard du 11 Novembre 1918, F-69622 Villeurbanne Cedex, France.

E-mail :
perriere@biomol.univ-lyon1.fr

MOTS-CLES :
Intelligence Artificielle, Base de Connaissances, Représentation Centrée-Objet, Expressivité des Gènes, Analyse des Séquences.

RESUME

ColiGene est une base de connaissance centrée-objet modélisant certaines des relations connues entre séquence génomique et expressivité des gènes chez *Escherichia coli*. ColiGene a été développée à l'aide du système de gestion de bases de connaissances SHIRKA. Les objets représentés dans la base sont des structures biologiques comme des gènes ou des signaux de régulation. Ces objets sont organisés en une structure hiérarchique de classes, sous-classes et instances. La base intègre également des méthodes d'analyse des séquences, en particulier des méthodes permettant de localiser des éléments génétiques comme des séquences codantes ou des signaux de régulation. Ces méthodes vont posséder des modules de création d'instances qui vont permettre d'intégrer immédiatement dans la base les structures prédites. La navigation dans la base et son interrogation ainsi que l'utilisation des méthodes sont réalisées par l'intermédiaire d'interfaces graphiques. Enfin un exemple d'utilisation de ColiGene est présenté ainsi qu'une discussion sur les limites actuelles du modèle de SHIRKA.

ABSTRACT

ColiGene is an object-centered knowledge base designed for the modelling of some of the known relationships between genomic sequence and expressivity in *Escherichia coli*. ColiGene was developed with the knowledge base management system SHIRKA. Objects represented in the base are biological structures such as genes or regulatory signals. These objects are organized in hierarchical structure of classes, sub-classes and instances. The base also includes sequence analysis methods, particularly those that localize genetic elements such as coding sequences or regulatory signals. These methods possess modules for instance creation that allow to integrate immediately the predicted structures in the base. Navigation and interrogation of the base and the use of the methods are made through graphical interfaces. At last, an example of study using ColiGene is given as also a discussion on the present limitations of the SHIRKA model.

1. INTRODUCTION

La quantité de séquences biologiques disponibles ayant augmenté de façon exponentielle au cours de ces dernières années, le besoin de disposer de systèmes permettant de manipuler aisément ces séquences ainsi que les informations qui leur étaient associées s'est révélé nécessaire. Depuis 1981, deux organismes sont en charge de collecter de façon concertée les séquences nucléotidiques : le laboratoire européen de biologie moléculaire à Heidelberg, pour l'EMBL data library (Cameron, 1988 ; Kahn et Cameron, 1990) et le laboratoire national de Los Alamos, pour la GenBank genetic sequence data bank (Burks *et al.*, 1985 ; 1990). Plus récemment, la DNA data bank of Japan (Miyazawa, 1990) a rejoint ce travail de collaboration. Aucun système d'interrogation n'étant fourni avec ces collections, plusieurs logiciels d'extraction et d'interrogation ont donc été développés (Bishop *et al.*, 1987). De tels systèmes offrent généralement des possibilités de sélection multicritères sur les séquences, l'accès pouvant se faire à n'importe quel niveau de l'arbre phylogénétique, ceci sur des données bibliographiques, des mots-clés, ou sur des informations quant au découpage des séquences en régions fonctionnelles. Dans ce contexte notre groupe a développé le système d'interrogation ACNUC (Gouy *et al.*, 1984 ; 1985) qui est, à ce jour, le système le plus largement utilisé en France. Cependant, du fait du volume actuel de séquences disponibles et, du fait des relations de plus en plus complexes mises en évidence entre séquence et fonctionnement cellulaire, le développement de bases de données dédiées, complémentaires des banques généralistes, apparaît indispensable. Les techniques de l'Intelligence Artificielle constituent dans ce domaine un champ d'expérimentation privilégié, de part les possibilités qu'elles offrent de représenter de manière explicite les connaissances puis de les exploiter. En particulier les modèles de représentation par objets apparaissent particulièrement adaptés à la construction de bases de *connaissances* en biologie moléculaire.

Dans ce contexte notre objectif a donc été de construire une base — ColiGene — afin d'examiner certaines relations existant entre séquence et expressivité. En effet, il est bien connu que des relations de ce genre existent dans les gènes de *E. coli*, que ce soit au niveau de l'initiation de la transcription (Mulligan *et al.*, 1984 ; O'Neill, 1989), de l'initiation de la traduction (Stormo, 1986 ; Thanaraj et Pandit, 1989 ; Sprengart *et al.*, 1990) et de l'élongation peptidique (Ikemura, 1981 ; Grantham *et al.*, 1981 ; Gouy et Gautier, 1982). Qui plus est, des interactions plus complexes ont été suggérées dans le cas où les gènes se trouvent organisés en opérons multigéniques (Gouy, 1987). Pour construire cette base nous avons donc choisi de prendre en compte :

- Les grands traits de l'organisation génétique de *E. coli*, incluant la structuration en opérons et la modulation de l'expression des gènes à l'aide de certains signaux comme les promoteurs ou les sites d'initiation de la traduction.

- Le polymorphisme génétique, potentiellement important dans le cas où l'on souhaite saisir l'impact sur l'expressivité d'un gène, de faibles variations dans sa composition en bases.

- Des méthodes mathématiques susceptibles soit de prédire l'organisation génétique d'une séquence lorsque celle-ci est inconnue, soit de déterminer les caractéristiques d'un signal, soit enfin d'estimer l'expressivité d'un gène.

- Des informations générales sur le produit du gène, comme par exemple sa fonction dans la cellule.

Ces choix nécessitent l'utilisation de structures de données complexes qui n'existent pas dans les bases de données classiques. Prenons l'exemple des opérons bactériens ; ceux-ci sont constitués de différents objets biologiques : un ou plusieurs promoteurs de transcription, des gènes, des signaux de régulation intercistroniques, un ou des terminateurs de transcription. Ces objets biologiques peuvent eux-mêmes être des objets complexes, constitués par l'association de différentes structures. Dans le cas des promoteurs de trancription, ceux-ci sont caractérisés par deux régions : la région -10 ou TATA box et la région -35 ou site de fixation de la RNA polymérase (Hawley et McClure, 1983). Représenter une telle organisation implique l'utilisation de modèles capables à la fois de modéliser des objets intégrés dans une structure hiérarchique et entretenant des relations d'« appartenance » avec les autres objets de la hiérarchie. Ces conditions, ainsi que la nécessité de pouvoir incorporer dans la structure de la base des méthodes mathématiques complexes nous ont conduits à nous intéresser aux modèles de représentation par objets. C'est pourquoi ColiGene a été développée à l'aide du système de représentation des connaissances centré-objet SHIRKA (Rechenmann et Uvietta, 1991), ceci en collaboration avec ses concepteurs. SHIRKA est écrit en langage LE_LISP (Chailloux *et al.*, 1986) disponible sur un grand nombre de machines. SHIRKA a déjà été utilisé en biologie, en particulier pour construire des bases à objectifs taxonomiques (Gautier et Pavé, 1990 ; Pavé *et al.*, 1991) ou des bases de modèles mathématiques (Pavé et Rechenmann, 1986). Enfin les différentes intérfaces graphiques développées pour ColiGene ont été réalisées à l'aide du logiciel AIDA développé par la société ILOG (1991).

2. LA BASE DE CONNAISSANCE

2.1. ORGANISATION DE LA CONNAISSANCE

Le modèle de représentation des connaissances sous SHIRKA est basé sur une notion similaire aux « frames » décrits par Fikes et Keller (1985). Sous SHIRKA, ces objets prennent le nom de *schémas*, un schéma pouvant représenter soit une *classe*, soit un élément de cette classe, on parle alors d'*instance*. Une classe et ses instances sont définis par des attributs, un *attribut* étant défini par une liste de facettes, et une *facette* par une liste de valeurs. Fort classiquement la base est constituée d'un réseau acyclique de classes et de sous-classes organisé en une structure hiérarchique. Une sous-classe décrivant des objets plus spécifiques que ses super-classes.

2.1.1. Structure hiérarchique

La construction de la base a donc tout d'abord consisté en l'établissement de cette hiérarchie des différents objets que l'on souhaitait représenter. Ces objets se répartissent sous ColiGene en deux grandes catégories :

- Des objets correspondant à des structures *biologiques*, eux-mêmes subdivisés en deux sous-ensembles :
  - Des objets simples, constitués d'une part de gènes codant soit pour des protéines soit pour des RNA structuraux (rRNA et tRNA), et d'autre part de signaux de régulation associés à ces gènes (promoteurs et terminateurs de transcription, sites d'initiation de la traduction).
  - Des objets complexes regroupant soit un ensemble d'objets simples, ce sont les opérons, soit un ensemble d'objets eux-mêmes complexes, ce sont les régulons.

- Des objets correspondant à des *descripteurs* de structures biologiques, il existera ainsi des descripteurs de sites d'initiation de la traduction ou de sites d'initiation de la transcription.

Concernant les objets biologiques cette organisation hiérarchique va en fait comprendre deux couches bien distinctes (figure 1). La premiére couche comprend des classes très générales qui ne vont servir qu'à décrire les propriétés communes à certains objets. Par exemple on y trouvera la classe décrivant la structure générale d'un gène protéique. La deuxième couche va comprendre elle, des classes correspondant à des structures *individuelles*. Ainsi chaque gène ou chaque signal connu de *E. coli* sera représenté par un schéma de classe qui lui sera propre. Il est ainsi possible de trouver plusieurs instances dérivant d'une même classe correspondant à une structure biologique individuelle, chaque instance correspondant en fait à un allèle particulier. A ce jour des informations sur près de 1500 gènes ainsi que sur leurs signaux de régulation associés ont été récoltées et intégrées dans la base. Le total des instances figurant dans la base se situant aux environs de 3000. Du fait que notre objectif premier est d'étudier les relations entre séquence et expressivité, des structures comme les éléments IS, les sites de fixation des phages ou les origines de réplication ne sont pas représentées dans la base.

2.1.2. Ecriture des schémas

La deuxième étape dans notre modélisation a été de formaliser la connaissance sur les objets de la hiérarchie, ceci en décrivant chaque type de structure à l'aide d'une liste d'attributs. Un certain nombre de ces attributs va être commun à plusieurs classes. Ainsi l'attribut **statut** va servir a préciser si une structure à une existence qui est *prouvée*, *potentielle* ou *putative*. Une structure sera ainsi déclarée comme prouvée si son existence a été démontrée par l'expérimentation biologique, potentielle si elle a été décrite comme telle dans la documentation des séquences des banques, et putative si elle a été mise en évidence en utilisant des méthodes mathématiques de prédiction. L'attribut **ref-num** quant à lui correspond au nom de la séquence qui sera associé à une structure biologique, que cette séquence se trouve dans la banque ou dans un fichier de données séparé.

Sur la figure 2 est représentée la structure générique du schéma de classe correspondant aux gènes protéiques. Dans ce schéma vont figurer des attributs qui vont nous donner la position du gène sur la carte chromosomique, sa fonction, son code EC dans le cas ou il code pour une protéine enzymatique, le trait phénotypique qui lui est rapporté, ainsi qu'un mot-clé. Nous avons également fait figurer deux attributs qui vont nous servir à estimer son expressivité : **NMD** et **BC**. Ces deux attributs correspondent à des indices d'usage du code définis par Gouy et Gautier (1982), NMD signifiant nombre moyen de discriminations de tRNA par cycle d'élongation et BC, coefficient de bon choix en troisième position des codons. Suivant la valeur numérique prise par ces deux indices il est possible d'inférer la valeur de l'attribut **expressivite**, qui pourra être *faible*, *moyenne* ou *forte*. Nous avons également considéré que le site d'initiation de la traduction ou Ribosome Binding Site (RBS) devait être intégré dans la structure d'un gène protéique. Comme il s'agit là d'une structure complexe, l'attribut **RBS** va renvoyer a un schéma particulier qui possède ses propres attributs.

2.1.3. Liens avec les banques de séquences

La base doit être alimentée à partir de données possédant déjà une certaine structuration, en effet il n'était pas concevable de gérer directement les collections de séquences à partir de SHIRKA. Dans ce but la base a été connectée au système ACNUC (Gouy *et al.*, 1985). Ce système comprend un logiciel d'interrogation : QUERY, qui permet d'accéder aux collections EMBL et GenBank structurées en base de données de type entité-association. Sous ACNUC il est possible d'accéder directement aux découpages des séquences en régions fonctionnelles homogènes (parties codantes et RNA structuraux) et de les manipuler ; ces régions étant considérées comme des sous-séquences. L'interface entre ColiGene et ACNUC a été primitivement développée en FORTRAN, plus récemment cette interface a été entièrement

réécrite en C. Il est à noter que, pour des raisons de rapidité d'accès, n'est utilisé qu'un sous-ensemble d'ACNUC ne comprenant que les séquences de *E. coli*.

Un problème réside dans le fait que les limites de structures comme les signaux de régulation ne ne sont pas définies en tant que sous-séquences dans ACNUC. Une telle formalisation n'a en effet jamais été tentée dans les différentes collections de séquences. Ainsi il est difficile d'assigner des limites précises quant au début et à la fin d'un promoteur de transcription, par exemple. Pour tourner ce problème nous avons utilisé des limites arbitraires, basées sur des analyses statistiques de biais dans la composition en nucléotides, biais observés dans les régions correspondant à ces signaux (Schneider *et al.*, 1986 ; Berg et von Hippel, 1987 ; O'Neill, 1989). Dans le cas des promoteurs, ceux-ci sont représentés par des séquences de 58 nucléotides dans lesquelles, par convention, la base numéro vingt est la dernière base de la région -35. Nous avons ensuite étendu le schéma conceptuel d'ACNUC avec un fichier supplémentaire de nom SIGNAL. Les données figurant dans ce fichier sont relatives à la position du signal dans la séquence mère, elles ont été recueillies manuellement, soit dans la documentation des séquences de GenBank, soit dans des compilations de la littérature (Hawley et McClure, 1983).

2.1.4. Intégration des méthodes

En biologie moléculaire, dans le domaine de l'analyse des séquences, une bonne partie de la connaissance est d'ordre méthodologique. Nous avons donc introduit un certain nombre de méthodes d'aide à l'analyse des séquences dans la structure de la base. Ces programmes sont déclenchés soit sur intervention de l'utilisateur, soit automatiquement dans le cas de l'attachement procédural (*cf* §2.2.3.). Du fait du très grand nombre de méthodes disponibles (voir par exemple la revue de Waterman, 1990), seuls quelques outils de base ont été intégrés. On trouvera ainsi des programmes permettant :

- De visualiser la structure primaire ou secondaire des séquences, ceci avec la possibilité de mettre en évidence les différentes régions fonctionnelles d'une séquence.

- D'extraire des séquences ou des informations associées à des instances dans des fichiers de données.

- De déterminer « l'efficacité » d'un signal de régulation ceci soit en fonction de sa conformité au consensus dans le cas des promoteurs de transcription (Schneider *et al.*, 1986 ; O'Neill, 1989), soit en fonction de la qualité de ses régions caractéristiques dans le cas des sites d'initiation de la traduction (Thanaraj et Pandit, 1989 ; Sprengart *et al.*, 1990).

- De prédire la présence de régions fonctionnelles à l'intérieur d'une séquence, ceci qu'il s'agisse de régions correspondant à des gènes ou à des signaux de régulation. La détection de ces zones est facilitée par l'emploi d'outils graphiques.

Toutes les précédentes fonctionnalités sont également représentées dans la base par des objets, au sens donné par SHIRKA à ce terme, et il est donc aisé de les manipuler. Il est ainsi possible de rajouter facilement des méthodes en fonction des besoins de l'utilisateur. Dans le cas des programmes de prédiction, il est possible d'intégrer immédiatement dans la base les structures prédites comme étant des structures biologiques. Tous les programmes de prédiction couplés à ColiGene possèdent en effet un module de création d'instances. Il est donc possible une fois une analyse effectuée, d'intégrer dans la base les résultats de cette analyse.

2.2. EXPLOITATION DES CONNAISSANCES

2.2.1. Les commandes SHIRKA

Il est possible d'interroger la base de connaissance en utilisant simplement le mode de commande en ligne de SHIRKA. Ces commandes vont permettre d'effectuer des interrogations de base et sont décrites en détail dans le manuel de SHIRKA (Rechenmann et Uvietta, 1990). Citons simplement la commande **`l-inst`** qui permet de récupérer la liste des instances appartenant à une classe donnée ; la commande **`l-spec`** qui permet de déterminer quelles sont toutes les spécialisations d'une classe ; et enfin la commande **`vi`** qui permet de visualiser, avec ou sans déclenchement des mécanismes d'inférence, les valeurs des différents attributs d'une instance.

2.2.2. Le filtrage

Le filtrage est un mécanisme d'inférence spécifique à SHIRKA (Rechenmann et Uvietta, 1991). Il consiste à rechercher des instances satisfaisant une description donnée sous la forme d'un schéma de classe. Il est mis en œuvre par la facette **`$sib-filtre`**, qui peut contenir plusieurs filtres essayés séquentiellement. Un filtre est donc un ensemble de conditions que doivent satisfaire des instances.

Exemple : récupération des instances de la classe `gene` correspondant à des gènes se situant entre la minute 0 et la minute 5 du chromosome de *E. coli*.

```
{ gene
   sorte-de        =            objet-biologique;
   map             $un          reel
                   $intervalle  [0.0 510.0] }

{ genes-0-5
   sorte-de        =            objet;
   genes           $liste-de    gene
                   $sib-filtre  { gene
                                    lui-meme   $var-> genes;
                                        map    $intervalle
                                                        [0.0 5.0] } }
```

L'interrogation se fera en créant une instance vide de la classe `gene-0-5` et en demandant à visualiser cette instance avec inférence :

```
{ liste-de-genes-0-5
   est-un          =      genes-0-5;
   genes           =      aceE aceF araA araB araC araD arl
                          ... }
```

Une telle procédure est évidemment assez lourde, et il paraît peu envisageable pour un non spécialiste de construire ses propres filtres, c'est pourquoi une bibliothèque des requêtes les plus communes a été écrite. Dans cette bibliothèque figurent des filtres permettant de récupérer des gènes en fonction de leur localisation sur le chromosome ou bien en fonction de leur expressivité. Il existe aussi des filtres qui vont permettre de sélectionner des sites d'initiation de la traduction en fonction de leur efficacité supposée.

2.2.3. L'attachement procédural

L'attachement procédural est un mécanisme assez répandu dans les systèmes de représentation par objets. Sous SHIRKA c'est la facette **$sib-exec** qui permet d'associer à un attribut une ou plusieurs méthodes de calcul qui vont permettre de déterminer la valeur prise par cet attribut. Chaque méthode est décrite par un schéma de classe, spécialisation du schéma **methode**, dans lequel l'attribut prédéfini **nom-fct** a pour valeur le nom de la fonction LISP qui effectuera le calcul. Les autres attributs décrivent les paramètres d'entrée-sortie de la méthode. Le plus souvent sous ColiGene cette fonction LISP va faire appel elle-même à une fonction C, liée dynamiquement au core LE_LISP.

Exemple : détermination la valeur de l'indice NMD dans les schémas représentant les gènes protéiques.

```
{ gene-proteique
      sorte-de        =          gene;
      provenance      $un        symbole
                      $defaut    banque;
      ref-num         $un        symbole;
      NMD             $un        reel
                      $sib-exec  { calcul-flc
                                     source     $var<- provenance;
                                     mnemo      $var<- ref-num;
                                     resultat   $var-> NMD } }
```

Le schéma **methode** qui correspond est le suivant :

```
{ calcul-flc
      sorte-de        =          methode;
      nom-fct         $valeur    flc
      source          $un        symbole;
      mnemo           $un        symbole;
      resultat        $un        reel }
```

Le code de la fonction LISP appelée étant :

```
(de flc (inst)
    (affect 'resultat (_flc (string (val? 'mnemo))
                            (string (val? 'source)))))
```

Dans ce cas **_flc** est le nom de la fonction C accédant à la séquence et retournant la valeur numérique de l'indice NMD.

Parmi les autres méthodes déclenchées par attachement procédural on trouve en particulier des méthodes permettant de prédire l'efficacité des sites d'initiation de la traduction ou des promoteurs de transcription. Il est alors intéressant de mettre en parallèle les résultats obtenus par ces méthodes avec ceux concernant l'expressivité prédite par l'intermédiaire de l'usage du code.

2.2.4. Le module de navigation et d'interrogation IVAN

Il semble évident que la navigation dans la base de connaissances et son interrogation, à l'aide des seules commandes en ligne ne peuvent se concevoir si l'on envisage une diffusion plus large de la base. Dans cette optique, une interface graphique a récemment été développée au laboratoire ARTEMIS (Grivaud, 1991). Cette interface, de nom IVAN (Interface Visuelle d'Aide à la Navigation), va comporter une fenêtre dans laquelle une partie ou la totalité de la hiérarchie de la base peut être représentée (figure 3). Il est alors possible de sélectionner dans cette fenêtre une classe donnée, puis d'accéder à la structure de cette classe ainsi qu'aux instances qui en dérivent. Une particularité étant qu'au niveau des instances, il est possible

d'accéder aux structures imbriquées. Ainsi à partir d'un schéma d'opéron, on pourra successivement accéder à un gène faisant partie de cet opéron, puis au RBS de ce gène, et enfin au descripteur de ce RBS.

Le système IVAN comporte également un module d'interrogation qui va permettre de construire des requêtes bien plus facilement qu'à l'aide du filtrage (figure 4). Là encore les requêtes porteront sur les valeurs prises par les attributs. A noter qu'il est possible de sauvegarder un requête afin de la réutiliser plus tard, de même qu'il est possible de conserver le résultat d'une requête — c'est à dire la liste des instances sélectionnées — sur lequel il sera possible d'effectuer de nouvelles sélections.

2.3. UN EXEMPLE DE SESSION SOUS COLIGENE

Cet exemple est basé sur l'étude d'une séquence *E. coli* figurant dans la collection GenBank (release 68) sous le mnémonique ECOACE. Elle est connue comme contenant quatre gènes protéiques : *A*, *aceE*, *aceF* et *lpd*. Deux de ces gènes — *aceE* et *aceF* — étant incorporés dans une structure en opéron.

2.3.1. Recherche des parties codantes

La première chose à faire va être de déterminer s'il existe dans la séquence des parties codantes (CDS). La commande à utiliser porte le nom de **`frame`**. L'algorithme employé va procéder au découpage de la séquence en une suite de blocs d'égale longueur, ensuite la valeur moyenne du NMD pour chaque bloc est portée en ordonnée (figure 5). La présence de terminateurs de traduction à l'intérieur d'un bloc est également signalée. Dans les régions où aucun codon de terminaison n'a été signalé et où la valeur moyenne pour l'indice NMD se situe en dessous de 45, on considère que la probabilité de trouver un cadre ouvert de lecture correspondant à une protéine exprimée est bonne. Au cas où dans une même région, plusieurs sites d'initiation de la traduction seraient en phase, il est possible de regarder la « qualité » prédite de chacun de ces sites, afin de choisir celui qui serait potentiellement le meilleur.

Une fois la recherche terminée, il est possible de procéder à la création de une ou plusieurs instances, dérivant de la classe **`gene-proteique`**, et correspondant aux structures révélées par le programme. Dans notre exemple quatre régions apparaissent comme étant susceptibles de contenir un gène protéique, nous avons donc créé quatre nouvelles instances qui vont correspondre à ces structures putatives. Par la suite nous avons interrogé la base afin de déterminer le niveau d'expressivité de ces parties codantes. Il apparaît alors que le premier CDS correspondrait à un gène faiblement exprimé alors que les trois autres correspondraient à des gènes fortement exprimés. Ces CDS seront désignés sous le nom de CDS I, II, III et IV. Les instances correspondantes créées par le programme ont toutes un statut putatif et leur attribut **`type`** est valué avec le symbole **`ORF`**.

2.3.2. Recherche des promoteurs

La deuxième étape dans notre étude va être d'essayer de localiser des promoteurs susceptibles d'agir sur ces CDS. La fonctionnalité générale de recherche des signaux a pour nom **`signal`** et, dans le cas des promoteurs, cette recherche est basée sur l'algorithme de O'Neill (O'Neill, 1989). Cet algorithme utilise une matrice de fréquences de bases pour chacune des trois principales "classes" de promoteurs recensées (O'Neill et Chiafari, 1989). Avec cette méthode, trois promoteurs ont été identifiés : un dans la région 5' en amont du CDS IV et deux dans la région 3' en aval de ce même CDS.

2.3.3. Recherche des terminateurs

Pour parfaire notre étude nous allons essayer de localiser les terminateurs de transcription de la séquence. Ceux-ci sont recherchés à l'aide de l'algorithme de Brendel et Trifonov (1984), une option de visualisation de la structure secondaire étant en outre proposée pour chaque fragment ayant été prédit comme étant un terminateur (figure 6). La structure secondaire de la séquence est prédite à l'aide du programme CRUSOE de Papanicolaou *et al.* (1984) et la visualisation de cette structure est effectuée à l'aide du programme LoopTool, du package GDE (1991). Dans notre exemple nous ayons choisi de limiter la recherche aux régions ne contenant pas de CDS et deux structures de terminateurs potentiels ont été prédites entre les CDS III et IV, et en aval du CDS IV.

2.3.4. Conclusions de l'étude

Nous n'avons pas trouvé de promoteurs pour les CDS I, II, et III alors que les auteurs de la séquence en avaient détecté un pour *aceEF* (Stephens *et al.*, 1983a ; 1983b). Il est possible que ce promoteur soit sous la dépendance d'un ou de plusieurs facteurs de régulation de la transcription et que, par conséquent, il ne suive pas le consensus édicté par O'Neill (1989). Par contre le promoteur détecté en 5' du CDS IV ainsi que les deux terminateurs correspondent bien à ceux décrits par les auteurs. On voit que le CDS IV est bordé en 5' d'un promoteur et en 3' d'un terminateur, il est donc probable que ce CDS soit transcrit de façon indépendante des autres gènes.

3. DISCUSSION

3.1. VALIDATION DE LA BASE

3.1.1. Du point de vue biologique

La base a d'ores et déjà été employée dans des études biologiques. En particulier elle a été utilisée pour aider à localiser le promoteur du gène *iclR*, ainsi que pour analyser la composition en codons et l'expressivité d'un ORF se situant entre les locus *iclR* et *aceB* de *E. coli* (Cortay *et al.*, 1991 ; Galinier *et al.*, 1991). Il s'avère que cet ORF possède une composition en codons non adaptée aux fréquences de tRNA de *E. coli*. Il est donc possible que le gène correspondant soit exprimé dans la cellule à un taux extrêmement bas, voire qu'il ne soit pas du tout exprimé. Ceci est confirmé par l'étude de son site d'initiation de la traduction qui peut être considéré comme mauvais, suivant les critères énoncés par Thanaraj et Pandit (1989).

Certains résultats sur la qualité prédite des signaux de traduction et de leur relation avec l'expressivité ont également été obtenus. En particulier une étude réalisée sur tous les gènes protéiques de la base ainsi que sur leur sites d'initiation associés semble montrer que certains critères qui avaient été proposés pour estimer l'efficacité des initiateurs (Sprengart *et al.*, 1990) ne sont pas en fait vérifiés. Des séquences censées ne se trouver que dans les gènes fortement exprimés sont ainsi présentes dans la quasi-totalité des sites d'initiation des gènes étudiés.

3.1.2. Du point de vue méthodologique

Sous ColiGene il est possible d'effectuer des sélections sur des critères qu'il serait très difficile d'intégrer dans des bases de données classiques. Il est ainsi par exemple possible de récupérer des gènes protéiques en fonction de leur expressivité prédite. De même il est possible de récupérer des signaux de régulation classés suivant leur efficacité supposée ou encore des

opérons bactériens contenant des gènes liés par couplage traductionnel. Des requêtes croisées sont également possibles, par exemple pour récupérer des gènes fortement exprimés, situés dans une région précise du chromosome et précédés par un site d'initiation de la traduction « fort ».

ColiGene peut être également utilisée pour tester des méthodes en analyse de séquences. Des méthodes de prédiction d'efficacité de promoteurs de transcription (Berg et von Hippel, 1987 ; O'Neill, 1989) ont ainsi été testées sur des cas où l'efficacité *relative* des dits promoteurs était connue expérimentalement (Bardonnet, 1991). Après classification de ces promoteurs en fonction de leur efficacité prédite on a essayé de voir si cette prédiction cadrait effectivement avec l'efficacité réelle. Dans ce cas il s'est avéré que ces méthodes présentaient un certain nombre de limitations. Ainsi leur emploi ne peut être effectué que sur des promoteurs constitutifs de *E. coli* c'est-à-dire des promoteurs dont l'efficacité est non modulée par l'intermédiaire d'activateurs ou d'inhibiteurs de transcription.

3.2. LE DEVELOPPEMENT DE NOUVEAUX OUTILS DE MODELISATION

3.2.1. Introduction d'une notion de « point de vue »

La complexité des objets biologiques représentés fait que l'utilisation d'une seule hiérarchie est très limitative. Jusqu'à présent les structures biologiques présentes dans ColiGene ne sont en effet représentées que sous une perspective *fonctionnelle*, toute la hiérarchie est basée sur la fonction des objets, avec une division en classes de gènes, de signaux, etc. Il serait également possible de considérer une perspective évolutive, dans laquelle serait prise en compte par exemple le caractère de répétitivité d'une séquence, c'est-à-dire si une séquence est présente en un ou plusieurs exemplaires dans le génome et, dans le cas où la séquence est répétée, quel est son type et son degré de répétition. En effet il existe une catégorie de signaux, dans les génomes bactériens, dont le rôle sera en grande partie lié au fait que ces structures se trouvent présentes en de nombreux exemplaires sur le chromosome. Ces signaux sont connus sous le nom d'unités palindromiques ou REP (Gilson *et al.*, 1984).

Deux solutions sont disponibles pour éviter cette perte d'information, la première consiste à définir plusieurs hiérarchies *indépendantes* pour le même type d'objets de la base. La seconde implique la modification du modèle de SHIRKA en y incluant un nouveau concept : le concept de *point de vue*. Ce concept autorise la coexistence de plusieurs niveaux de connaissances, chaque niveau décrivant les mêmes objets vus selon des points de vue différents. Dans ce cas l'ensemble de niveaux donnerait une description complète des objets modélisés (figure 7).

3.2.2. Vers un système de résolution coopérative de problèmes

A l'heure actuelle la façon dont sont gérées les méthodes sous ColiGene est rudimentaire. Ceci ne présentant que peu d'inconvénients du fait de leur nombre peu élevé. Cependant si on envisage d'augmenter le nombre de ces méthodes, la nécessité de les gérer d'une façon évoluée va se poser. En effet, l'acquisition des compétences nécessaires à l'utilisation optimale de méthodes en analyse de séquences constitue un lourd investissement en temps pour le néophyte, et celui-ci préfèrera le plus souvent n'utiliser qu'un nombre restreint de méthodes bien connues. Dans le but de pouvoir disposer de bases modélisant au mieux les connaissances méthodologiques, le gestionnaire de méthodes SCAI intégré à SHIRKA, a été développé au laboratoire ARTEMIS (Poncabaré et Rechenmann, 1991). Dans ce système, la connaissance méthodologique est modélisée sous la même forme que l'a été précédemment la connaissance biologique : sous la forme de schémas de classes. Dans ce cas les attributs de ces schémas vont constituer les entrées/sorties d'une méthode.

Le concepteur de la base devra donc modéliser les raisonnements suivis en analyse de séquences, ceci à l'intérieur du formalisme SHIRKA. Nous donnerons ici un exemple élémentaire de cette décomposition du raisonnement : il s'agit d'une stratégie possible de recherche de parties codantes dans un organisme non mammifère. Dans cet exemple, le problème peut être résolu de deux façons : soit on connaît les fréquence des tRNA de l'organisme étudié, soit on ne les connaît pas. Dans le premier cas la méthode à employer utilisera ces fréquences afin de déterminer les régions dans lesquelles un biais dans l'utilisation des codons est constaté. Dans le deuxième cas, il sera nécessaire d'employer une méthode utilisant le test de χ^2, capable de mettre en évidence le rythme de trois existant dans les régions codantes. Un raisonnement simple comme celui-là peut-être modélisé sous SCAI par le schéma représenté sur la figure 8.

Au moment de la résolution la méthode la plus adaptée au contexte sera déterminée grâce au mécanisme de classification intégré dans SHIRKA. Dans le modèle présent, deux formes de choix de sous-tâche sont possibles : un « choix automatique » et un « choix utilisateur ». Ils sont à exploiter différemment : dans le premier cas le choix parmi les sous-tâches possibles est effectué par le système, qui exploite cette représentation, tandis que dans le deuxième cas c'est l'utilisateur, qui, bénéficiant éventuellement de connaissances supplémentaires, prend une décision.

3.3. LES AUTRES PROBLEMES RENCONTRÉS

3.3.1. La maintenance de la base

Un des problèmes rencontré est un des problèmes classique des bases de données, il s'agit de la maintenance et de la mise à jour de la base. Tous les trois mois, de nouvelles versions des banques de séquences sont disponibles, à ce moment là il est nécessaire de procéder à la mise à jour. Une grande part de ce travail peut être effectuée par des programmes automatisant la procédure, mais la gestion de nombreuses exceptions ne peut être effectuée que manuellement. En particulier le séquençage de nouvelles structures fait qu'il est nécessaire d'écrire au moins les schémas de classe correspondant à ces structures, la génération des instances pouvant être automatisée. En fait le problème le plus aigu est celui des données concernant les localisations des différents signaux génomiques de *E. coli*. En effet ces données figurent dans un fichier qu'il sera nécessaire d'éditer manuellement, cette édition passant par trois étapes :

- Tout d'abord il faut explorer la liste des séquences apparues dans la nouvelle release, ceci afin de voir si dans ces séquences des signaux de régulation ont été identifiés.
- Ensuite, il faut enlever du fichier toutes les références aux séquences ayant disparu.
- Enfin il faut vérifier les positions concernant les séquences ayant été modifiées (le plus généralement ces séquences ayant été concaténées avec d'autres). Dans ce cas il peut être nécessaire de modifier les valeurs indiquant les positions de début et de fin du signal.

Une fois cette édition effectuée, il faut modifier manuellement les instances qui faisaient appel à des structures qui ont disparu dans la nouvelle release. La conséquence principale d'une telle organisation est que la maintenance de la base n'est réalisable que par son concepteur.

3.3.2. La taille de la base

Un autre problème existe, lui aussi lié à l'utilisation d'une base de connaissances basée sur des objets. Dans des systèmes de ce type, toute la connaissance doit résider en mémoire

centrale. Ceci implique l'utilisation de grosses configurations possédant au minimum 16 Mo de RAM. Une étape future dans les développements à envisager pour le modèle SHIRKA serait donc la création d'un gestionnaire d'objets pouvant exploiter de la connaissance se trouvant sur disque, tout en évitant une pénalisation trop lourde sur la vitesse d'accès.

4. CONCLUSIONS — PERSPECTIVES

ColiGene est un système aujourd'hui opérationnel mais qui doit cependant encore être amélioré. Les améliorations que nous comptons lui apporter se situeront au niveau du modèle, bien sûr, mais aussi au niveau de l'interface utilisateur. Les améliorations au niveau de la modélisation sont conduites conjointement avec l'équipe du laboratoire ARTEMIS ; elles porteront, comme nous l'avons dit précedemment, sur l'introduction de structures multi-hiérarchiques et de gestionnaires de méthodes, intégrés dans la base. Les améliorations au niveau de l'interface porteront sur la construction de nouveaux outils graphiques, pour les différentes méthodes associées à la base.

ColiGene est actuellement utilisée dans l'analyse des signaux de certains opérons multigéniques et l'expression différentielle des gènes de ces opérons. La question posée est : cette expression différentielle constatée peut-elle s'expliquer par la présence de signaux intercistroniques augmentant l'expression de certains gènes ou empêchant leur dégradation trop rapide ? Sont également étudiés les liens existant entre efficacité des signaux de traduction et expressivité résultante des gènes. Ainsi est-il possible de prédire, à partir de la simple séquence nucléotidique, si un site d'initiation de la traduction est efficace ou non ? Et si oui, cela se traduit-il par une différence mesurable au niveau de l'expression.

Enfin il faut souligner que si ColiGene a été développée avec un objectif de recherche précis qui était l'étude de l'expressivité des gènes en relation avec la structure des séquences génomiques, les résultats méthodologiques obtenus permettent d'envisager d'autres applications. Ainsi deux nouvelles structures de bases de connaissances utilisant les résultats précedemment acquis sont en cours d'élaboration :

- Une base de méthodes en analyse des séquences, de nom AnalSeq. Ce système devant pouvoir être utilisé dans l'aide à l'expertise de grands fragments génomiques provenant d'organismes variés. Il incluerait un ensemble de tâches modélisant les raisonnements suivis en analyse de séquence et serait donc susceptible de guider l'utilisateur dans les choix et les enchaînements de méthodes à suivre pour effectuer une étude complète sur une séquence ou sur un groupe de séquences.

- Une base dédiée à l'étude de l'organisation spatiale des génomes de mammifères. L'objectif majeur de cette base étant d'étudier la composante spatiale de l'évolution des génomes, ce qui implique la maîtrise des relations existant entre les différents niveaux de cartographie.

BIBLIOGRAPHIE

Bardonnet, N. (1991) Elaboration d'un système de fusion d'opérons utilisant le gène *uidA* d'*Escherichia coli* K-12. Contribution à l'étude des signaux de transcription de la bactérie industrielle *Corynebacterium glutamicum*. *Thèse de Doctorat*. Université Claude Bernard Lyon I.

Berg, O.G. & von Hippel, P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723-750.

Bishop, M.J., Ginsburg, M., Rawlings, C.J. & Wakeford, R. (1987) Molecular sequence databases. In *Nucleic acid and protein sequence analysis, a practical approach*. M.J. Bishop and C.J. Rawlings eds., IRL Press, Oxford, Washington D.C., pp. 83-113.

Brendel, V. & Trifonov, E.N. (1984) A computer algorithm for testing potential prokaryotic terminators. *Nucleic Acids Res.*, **12**, 4411-4427.

Burks, C., Cinkosky, M.J., Gilna, P., Hayden, J.E.-D., Abe, Y., Atencio, E.J., Barnhouse, S., Benton, D., Buenafe, C.A., Cumella, K.E., Davison, D.B., Emmert, D.B., Faulkner, M.J., Fickett, J.W., Fischer, W.M., Good, M., Horne, D.A., Houghton, F.K., Kelkar, P.M., Kelley, T.A., Kelly, M., King, M.A., Langan, B.J., Lauer, J.T., Lopez, N., Lynch, C., Lynch, J., Marchi, J.B., Marr, T.G., Martinez, F.A., McLeod, M.J., Medvick, P.A., Mishra, S.K., Moore, J., Munk, C.A., Mondragon, S.M., Nasseri, K.K., Nelson, D., Nelson, W., N'Guyen, T., Reiss, G., Rice, J., Ryals, J., Salazar, M.D., Stelts, S.R., Trujillo, B.R., Tomlinson, L.J., Weiner, M.G., Welch, F.,J., Wiig, S.E., Yudin, K. & Zins, L.B. (1990) GenBank : current status and future directions. *Methods Enzym.*, **183**, 3-22.

Burks, C., Fickett, J.W., Goad, W.B., Kanehisha, M., Lewitter, F.I., Rindone, W.P., Swindell, C.D. & Bilofsky, H.S. (1985) The GenBank nucleic acid sequence database. *Comput. Applic. Biosci.*, **1**, 225-233.

Cameron, G.N. (1988) The EMBL data library. *Nucleic Acids Res.*, **16**, 1865-1867.

Cortay, J.-C., Nègre, D., Galinier, A., Duclos, B., Perrière, G., & Cozzone, A.J. (1991) Regulation of the acetate operon in *Escherichia coli* : purification and functional characterization of the IclR repressor. *EMBO J.*, **10**, 675-679.

Chailloux, J., Devin, M., Dupont, F., Hullot, J.-M., Serpette, B. & Vuillemin, J. (1986) Le_Lisp Version 15.2, Manuel de Référence, second edition, INRIA, Rocquencourt, France.

Fikes, R. & Kehler, T. (1985) The role of frame-based representation in reasonning. *Comm. ACM*, **28**, 904-920.

Galinier, A., Bleicher, F., Nègre, D., Perrière, G., Duclos, B., Cozzone, A.J. & Cortay, J.-C. (1991) Primary structure of the intergenic region between *aceK* and *iclR* in the *Escherichia coli* chromosome. *Gene*, **97**, 149-150.

Gautier, N. & Pavé, A. (1990) Object-centered representation for species systematics and identification in living systems in nature. *Comput. Applic. Biosci.*, **6**, 383-386.

GDE (1991) Genetic Data Environment software. University of Illinois, Urbana, Illinois.

Gilson, E., Clement, J.-M., Brutlag, D. & Hofnung, M. (1984) A family of repetitive extragenic palindromic DNA sequences in *E. coli*. *EMBO J.*, **3**, 1417-1421.

Gouy, M. (1987) Origine et fonction de l'utilisation de la dégénerescence du code génétique chez *Escherichia coli* : structuration en banque de données et analyse statistique des séquences nucléotidiques. *Thèse de Doctorat d'Etat*. Université Claude Bernard Lyon I.

Gouy, M. & Gautier, C. (1982) Codon usage in bacteria : correlation with gene expressivity. *Nucleic Acids Res.*, **10**, 7055-7073.

Gouy, M., Gautier, C., Attimonelli, M., Lanave, C. & di Paola, G. (1985) ACNUC — a portable retrieval system for nucleic acid sequence databases : logical and physical designs and usage. *Comput. Applic. Biosci.*, **1**, 167-172.

Gouy, M., Milleret, F., Mugnier, C., Jacobzone, M. & Gautier, C. (1984) ACNUC : a nucleic acid sequence data base and analysis system. *Nucleic Acids Res.*, **12**, 121-127.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.*, **9**, r43-r74.

Grivaud, S. (1991) IVAN : Interface de Visualisation et d'Aide à la Navigation dans une base de connaissances centrée-objet. Mémoire d'ingénieur CNAM, Grenoble, France.

Hawley, D.K. & McClure, W.R. (1983) Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucleic Acids Res.*, **11**, 2237-2255.

Ikemura, T. (1981) Correlation between the abundance of *Escherichia coli* transfers RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.*, **146**, 1-21.

ILOG (1991) AIDA : un environnement de développement d'interfaces graphiques, Manuel d'utilisation. ILOG, Gentilly, France.

Kahn, P. & Cameron, G. (1990) EMBL data library. *Methods Enzym.*, **183**, 23-31.

Miyazawa, S. (1990) DNA data bank of Japan : present status and future plans. In *Computers and DNA, SFI Studies in the Sciences of Complexity*, Vol. VII, G. Bell and T. Marr eds., Addison-Wesley, pp. 47-62.

Mulligan, M.E., Hawley, D.K., Entriken, R. & McClure, W.R. (1984) *Escherichia coli* promoter sequences predict *in vitro* RNA polymerase selectivity. *Nucleic Acids Res.*, **12**, 789-800.

O'Neill, M.C. (1989) Consensus methods for finding and ranking DNA binding sites application to *Escherichia coli* promoters. *J. Mol. Biol.*, **207**, 301-310.

O'Neill, M.C. & Chiafari, F. (1989) *Escherichia coli* promoters. II. A spacing class-dependant promoter search protocol. *J. Biol. Chem.*, **264**, 5531-5534.

Papanicolaou, C., Gouy, M. & Ninio, J. (1984) An energy model that predicts the correct folding of tRNA and the 5S RNA molecules. *Nucleic Acids Res.*, **12**, 31-44.

Pavé, A., Gautier, N. & Bernstein, C. (1991) Object centered representation and problems related to living systems in nature : systematics, biogeography and population dynamics. In *Artificial Intelligence in Numerical and Symbolic Simulation*. A. Pavé and G.C. Vansteenkiste eds., Aléas, Lyon, France, pp. 51-74.

Pavé, A. & Rechenmann, F. (1986) Computer aided modelling in biology : an Artificial Intelligence approach. *S.C.S. Simul. Serie*, **18**, 52-66.

Poncabaré, T. & Rechenmann, F. (1991) SCAI : un environnement de développement de systèmes à base de connaissances en calcul scientifique et technique.

Rechenmann, F. & Uvietta, P. (1990) Shirka : système de gestion de bases de connaissances centrées-objet, Manuel d'utilisation. INRIA et laboratoire ARTEMIS, Grenoble, France.

Rechenmann, F. & Uvietta, P. (1991) Shirka : an object-centered knowledge based management system. In *Artificial Intelligence in Numerical and Symbolic Simulation*. A. Pavé and G.C. Vansteenkiste eds., Aléas, Lyon, France, pp. 9-23.

Schneider, T.D., Stormo, G.D., Gold, L. & Ehrenfeucht, A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415-431.

Sprengart, M.L., Fatscher, H.P. & Fuchs, E. (1990) The initiation of translation in *E. coli* : apparent base pairing between the 16S rRNA and downstream sequences of the mRNA. *Nucleic Acids Res.*, **18**, 1719-1723.

Stephens, P.E., Darlison, M.G., Lewis, H.M. & Guest, J.R. (1983a) The pyruvate dehydrogenase complex of *Escherichia coli* K12. Nucleotide sequence encoding the dihydrolipoamide acetyltransferase component. *Eur. J. Biochem.*, **133**, 481-489.

Stephens, P.E., Lewis, H.M., Darlison, M.G. & Guest, J.R. (1983b) Nucleotide sequence of the lipoamide dehydrogenase gene of *Escherichia coli* K12. *Eur. J. Biochem.*, **135**, 519-527.

Stormo, G. (1986) Translational initiation. In *Maximizing gene expression*. Reznikoff, W. and Gold, L. eds., Butterworth Publishers, Stoneham, Massachusetts, pp. 195-224.

Thanaraj, T.A. & Pandit, M.W. (1989) An additional ribosome-binding site on mRNA of highly expressed genes and a bifunctionnal site on the colicine fragment of 16S rRNA from *Escherichia coli* : important determinants of the efficiency of translation initiation. *Nucleic Acids Res.*, **17**, 2973-2985.

Waterman, M.S. (1990) *Mathematical methods for DNA sequences*. CRC Press, Inc., Boca Raton, Florida.

Figure 1 – Oganisation hiérarchique de la base de connaissance ColiGene. Sur ce schéma sont représentées les classes générales décrivant la structure des objets biologiques étudiés. Dans la partie terminale de la hiérarchie on a simplement indiqué le nombre de sous-classes correspondant à des structures individuelles, sachant qu'à chaque classe sera rattachée au moins une instance.

Figure 2 – Structure du schéma de classe décrivant la structure d'un gène protéique. Les attributs `trait-phénotypique`, `code-EC`, `keyword` et `fonctions` sont valués dans les schémas de classe individuels de la hiérarchie de la base. Les valeurs des attributs `RBS`, `acides-aminés`, `NMD`, `BC` et `expressivité` sont déterminés par attachement procédural. Les attributs `provenance` et `statut` sont valués par défaut. Enfin l'attribut `ref-num` sera lui valué au niveau de l'instance.

Figure 3 – Organisation de la fenêtre de navigation du système IVAN. Le panneau `Arbre des classes` montre l'organisation hiérarchique complète d'un concept sélectionné par l'utilisateur, dans le cas présent, la classe `objet-biologique`. Un concept sous IVAN sera en effet constitué par une classe dérivant immédiatement de la classe la plus générale de SHIRKA, la classe `objet`. Les autres panneaux vont permettre de visualiser les sous-classes et les sur-classes d'une classe donnée ainsi que les attributs qui leur sont associés. Dans le cas ou des classes d'attributs auraient été créées, il est possible de ne visualiser que les attributs appartenant à une ou plusieurs classes données (panneau `Classes d'attribut`). Au niveau du panneau `Instances` est affichée la liste de toutes les instances rattachées à la classe sélectionnée, c'est à ce niveau là qu'il est possible de sélectionner une instance en particulier et de visualiser sa structure.

Figure 4 – Organisation de la fenêtre d'interrogation du système IVAN. Dans la partie gauche on trouve, regroupés sous différents en-têtes, les boutons de commande permettant de se déplacer dans la hiérarchie des repertoires (`Systeme`), d'initialiser la classe sur laquelle on veut effectuer une interrogation (`Constructeur`), de lancer, de sauvegarder ou de vider une requête (`Requete`), et enfin de conserver ou de sauvegarder le résultat d'une requête (`Selection`). Dans la partie droite se trouve le constructeur de requêtes et le panneau de visualisation des résultats. Toutes les instances suivant les critères définis dans le constructeur de requête vont s'afficher à ce niveau et là encore il sera possible de sélectionner une instance afin de visualiser sa structure.

Figure 5 – Exemple d'utilisation de la fonctionnalité de détection des parties codantes dans ColiGene. Dans chacune des trois phases, l'usage du code est porté en ordonnée. Sur la partie inférieure de la figure, la présence de codons de terminaison est signalée (zones noires sur la figure). La détection des parties codantes se fait en mettant en regard les régions dépourvues de codons de terminaison avec les régions pour lesquelles un biais dans la composition en codons a été détecté. Dans ce cas là, les valeurs prises par l'indice NMD seront basses. Deux modes de visualisation sont disponibles : le mode "click" permet de préciser les coordonnées du bloc cliqué tandis que le mode "zoom" permet de ne visualiser qu'une partie du diagramme.

Figure 6 – Sur cette figure est représenté le menu principal de la fonctionnalité de recherche de signaux ainsi que l'output de la fonctionnalité de visualisation des structures secondaires de RNA. Les deux structures secondaires visualisées sont celles des deux terminateurs de transcription détectés dans la séquence.

Figure 7 – Une possibilité d'organisation en multi-perspective sous ColiGene. Seuls les niveaux les plus élevés des deux hiérarchies sont représentés ici. Dans ce schéma, nous avons considéré qu'un fragment génomique pouvait être vu sous deux points de vue ; le premier point de vue prenant en compte l'aspect fonctionnel des objets biologiques en question et le deuxième, l'aspect évolutif. Les classes qui portent le même nom dans les deux hiérarchies correspondent aux mêmes objets. Dans cet exemple, la classe `REP`, correspondant aux unités palindromiques, est considérée sous les deux perspectives. Une instance de cette classe, `lacZY1r`, sera donc décrite sous ces deux perspectives.

Figure 8 – Décomposition du raisonnement pour une stratégie de recherche de parties codantes dans une séquence provenant d'un organisme non-mammifère. Ce problème se modélise sous SCAI sous la forme d'une tâche de choix (`T-rech-ORF-non-mam`) qui peut être résolue soit par la sous-tâche `T-coeff+flc+diag`, dans le cas ou les fréquences de tRNA de l'organisme sont connues, soit par la sous-tâche `T-chi2+diag` dans le cas ou ces fréquences sont inconnues. Chacune de ces sous-tâches est elle-même une tâche séquentielle : `T-chi2+diag` est la suite de `T-chi2` et de `T-diag`, `T-coeff+flc+diag` la suite de `T-choix-coeff`, de `T-flc` et de `T-diag`. Ces tâches, dites tâches terminales, vont être résolues directement par le lancement de méthodes, ainsi `T-chi2` va-t-elle déclencher une méthode calculant la valeur d'un test de χ^2 tandis que `T-diag` va lancer une méthode effectuant le tracé du graphe des valeurs. C'est à partir de ce graphe qu'il sera possible de repérer les éventuelles parties codantes présentes dans la séquence.

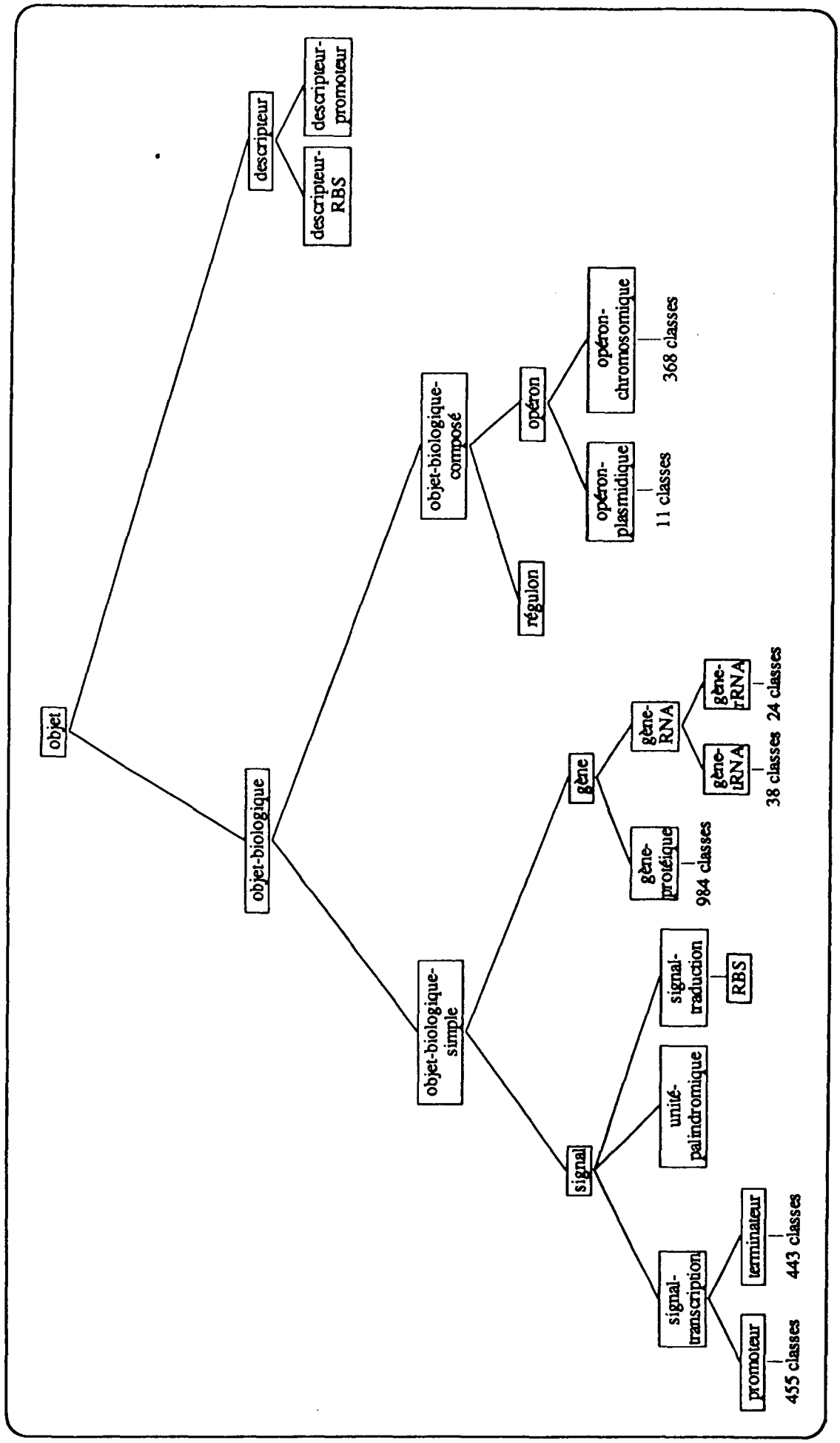
objet
descripteur
descripteur-promoteur
descripteur-RBS
objet-biologique
objet-biologique-composé
opéron
opéron-chromosomique
368 classes
opéron-plasmidique
11 classes
régulon
objet-biologique-simple
gène
gène-RNA
gène-rRNA
gène-tRNA
38 classes
24 classes
gène-protéique
984 classes
signal
signal-traduction
RBS
unité-palindromique
signal-transcription
terminateur
443 classes
promoteur
455 classes

Fig. 1

```
( gene-proteique
     sorte-de              =          gene;
     ref-num               $un        symbole;
     provenance            $un        symbole
                           $defaut    banque;
     statut                $un        symbole
                           $domaine   prouve potentiel
                                      putatif;
     map                   $un        reel
                           $intervalle [0.0 510.0];
     keyword               $un        chaine;
     RBS                   $un        RBS;
     trait-phenotypique    $un        chaine;
     code-EC               $un        chaine;
     fonctions             $liste-de  symbole;
     acides-amines         $un entier
                           $sib-exec  ( nb-aa
                                        ... );
     NMD                   $un reel
                           $sib-exec  ( FLC
                                        ... );
     BC                    $un reel
                           $sib-exec  ( RLC
                                        ... );
     expressivite          $un        symbole
                           $domaine   haute moyenne faible
                           $sib-exec  ( expr
                                        ... )
)
```

Fig. 2

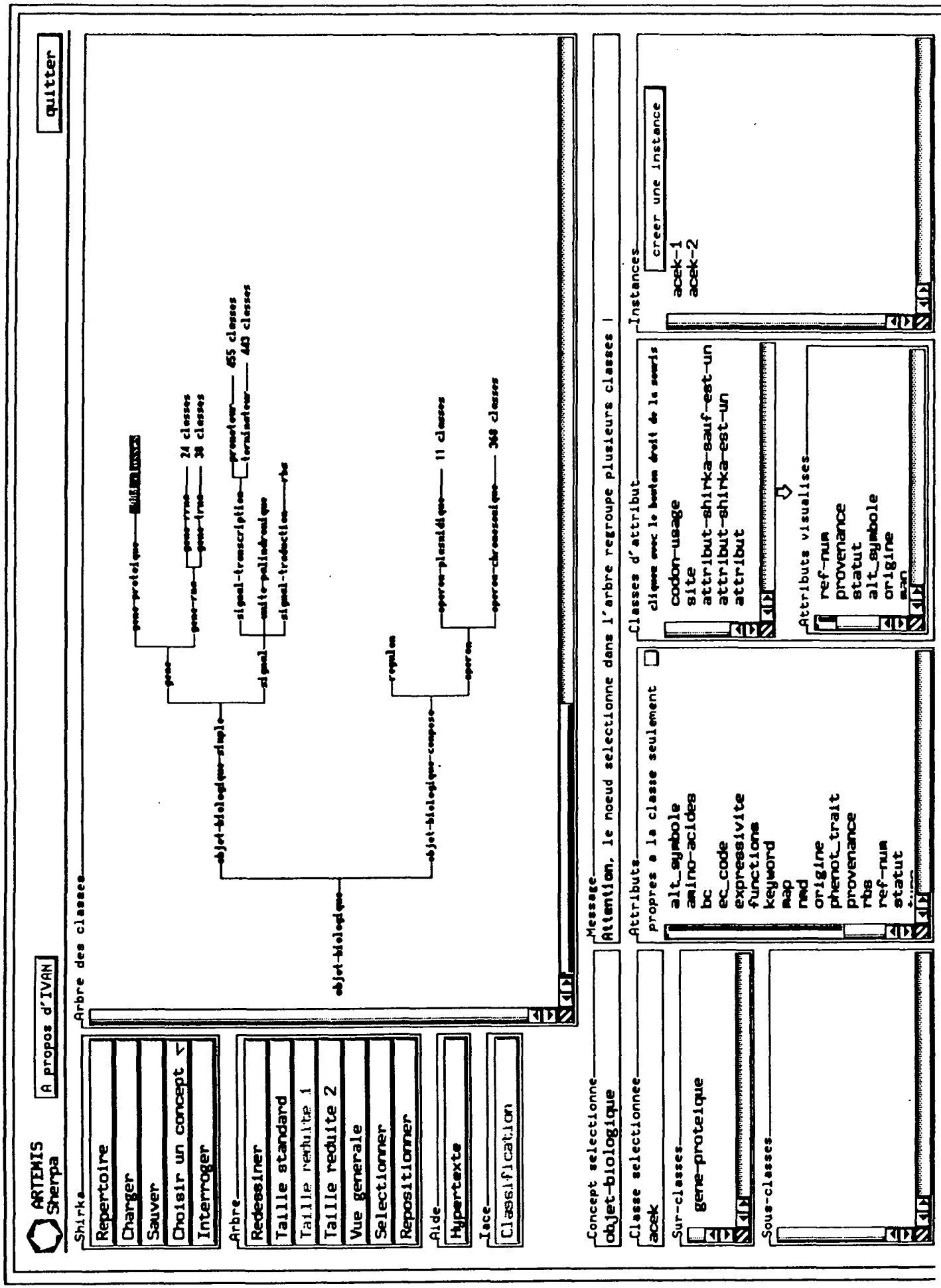
ARTEMIS
Sherpa
A propos d'IVAN
quitter
Shirka
Repertoire
Charger
Sauver
Choisir un concept
Interroger
Arbre
Redessiner
Taille standard
Taille reduite 1
Taille reduite 2
Vue generale
Selectionner
Repositionner
Aide
Hypertexte
Face
Classification
Arbre des classes
objet-biologique
objet-biologique-simple
objet-biologique-composé
gene
signal
region
operon
gene-proteique
gene-rna
gene-trna
signal-transcription
signal-polyadenylique
signal-traduction
operon-plasmidique
operon-chromosique
24 classes
38 classes
455 classes
443 classes
11 classes
368 classes
Concept selectionne
objet-biologique
Message
Attention, le noeud selectionne dans l'arbre regroupe plusieurs classes !
Classe selectionnee
acek
Attributs
propres a la classe seulement
alt_symbole
amino-acides
bc
ec_code
expressivite
functions
keyword
map
nad
origine
phenot_trait
provenance
rbs
ref-num
statut
Classes d'attribut
clique avec le bouton droit de la souris
codon-usage
site
attribut-shirka-sauf-est-un
attribut-shirka-est-un
attribut
Attributs visualises
ref-num
provenance
statut
alt_symbole
origine
Instances
creer une instance
acek-1
acek-2
Sur-classes
gene-proteique
Sous-classes

Fig. 3

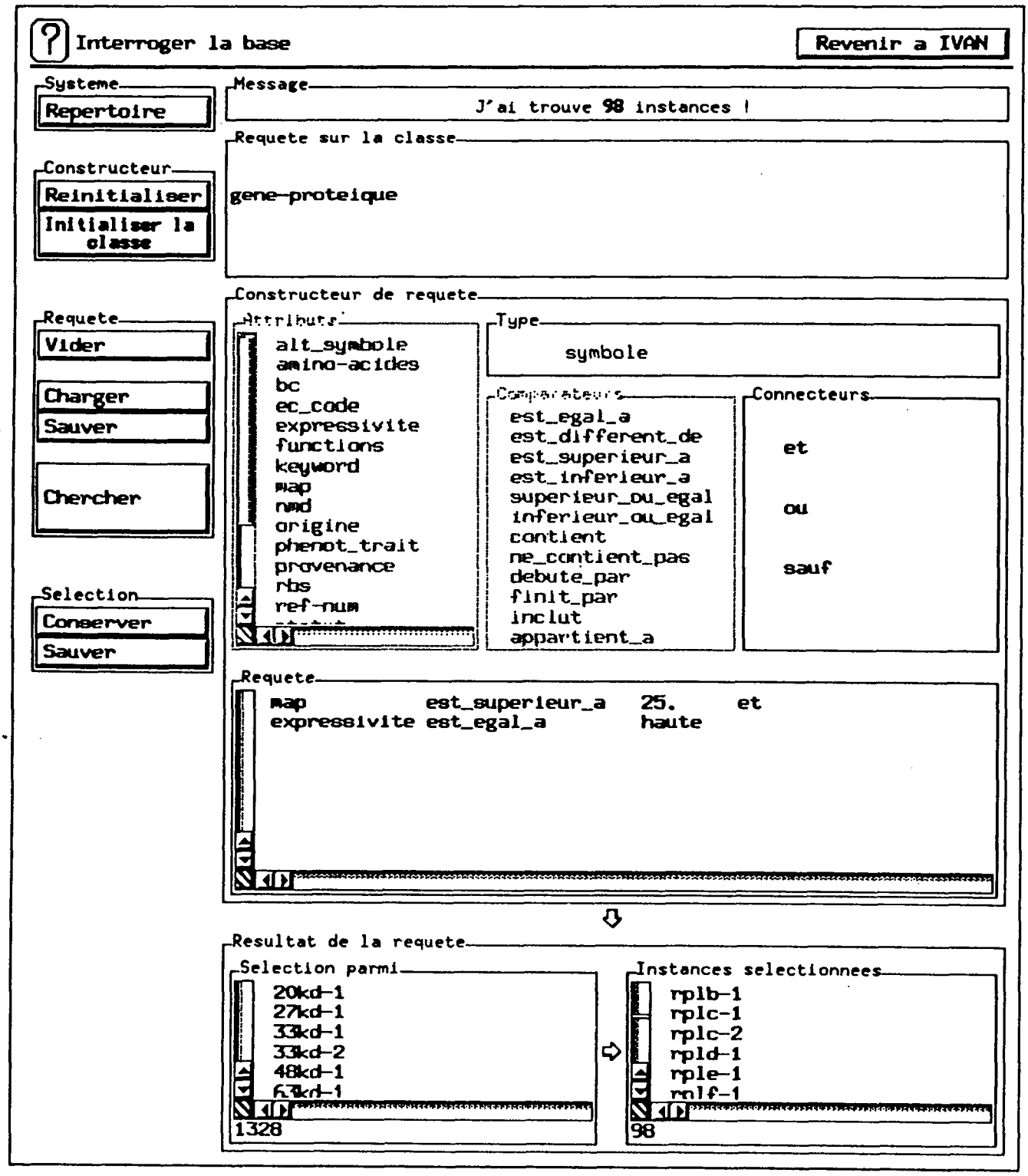
Interroger la base
Revenir a IVAN
Systeme
Repertoire
Message
J'ai trouve 98 instances !
Requete sur la classe
gene-proteique
Constructeur
Reinitialiser
Initialiser la classe
Constructeur de requete
Requete
Vider
Charger
Sauver
Chercher
Attributs
alt_symbole
amino-acides
bc
ec_code
expressivite
functions
keyword
map
nmd
origine
phenot_trait
provenance
rbs
ref-num
Type
symbole
Comparateurs
est_egal_a
est_different_de
est_superieur_a
est_inferieur_a
superieur_ou_egal
inferieur_ou_egal
contient
ne_contient_pas
debute_par
finit_par
inclut
appartient_a
Connecteurs
et
ou
sauf
Selection
Conserver
Sauver
Requete
map est_superieur_a 25. et
expressivite est_egal_a haute
Resultat de la requete
Selection parmi
20kd-1
27kd-1
33kd-1
33kd-2
48kd-1
1328
Instances selectionnees
rplb-1
rplc-1
rplc-2
rpld-1
rple-1
98

Fig. 4

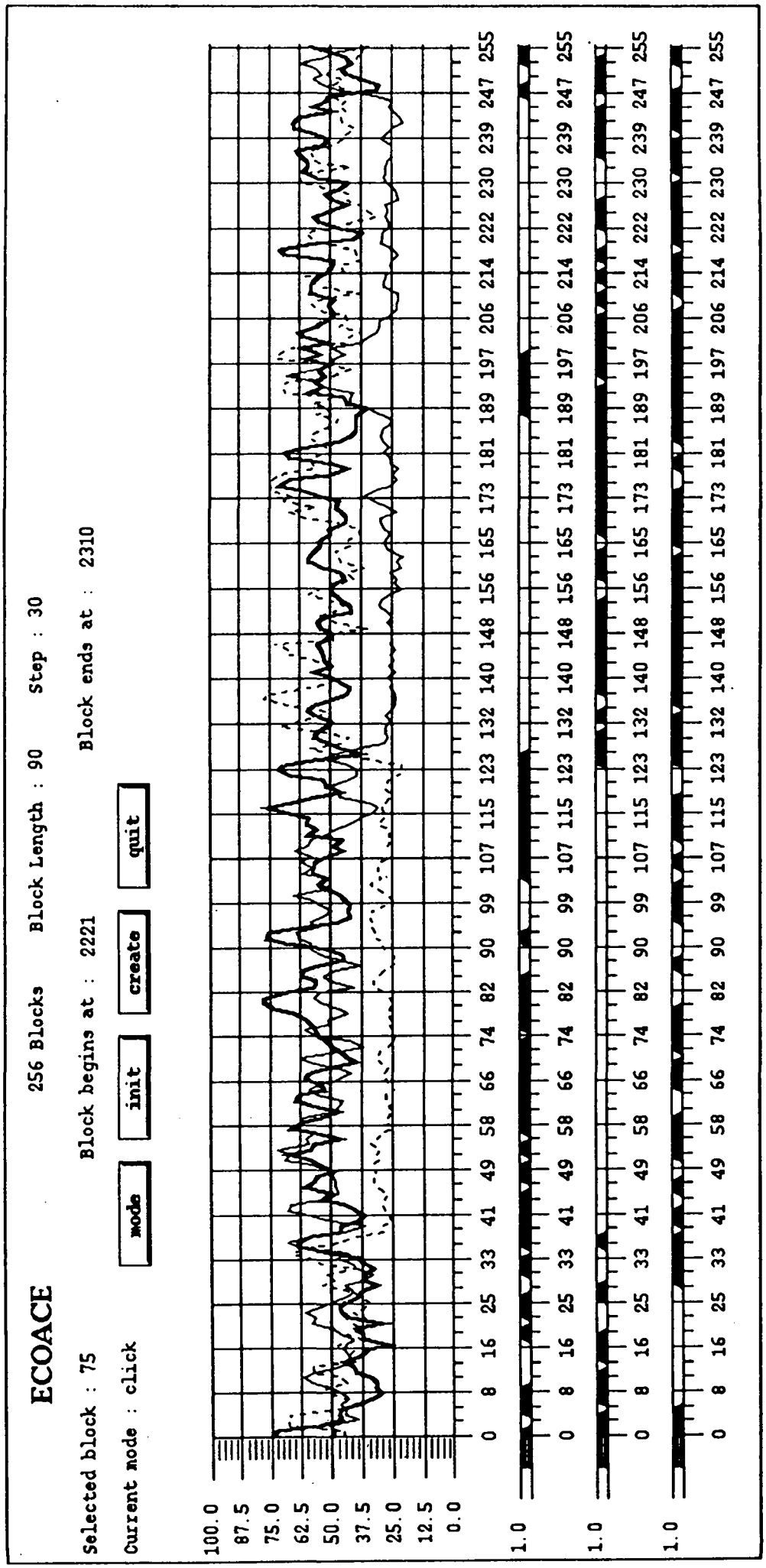
ECOACE
256 Blocks
Block Length : 90
Step : 30
Selected block : 75
Block begins at : 2221
Block ends at : 2310
Current mode : click
mode
init
create
quit
100.0
87.5
75.0
62.5
50.0
37.5
25.0
12.5
0.0
1.0
0 8 16 25 33 41 49 58 66 74 82 90 99 107 115 123 132 140 148 156 165 173 181 189 197 206 214 222 230 239 247 255

Fig. 5

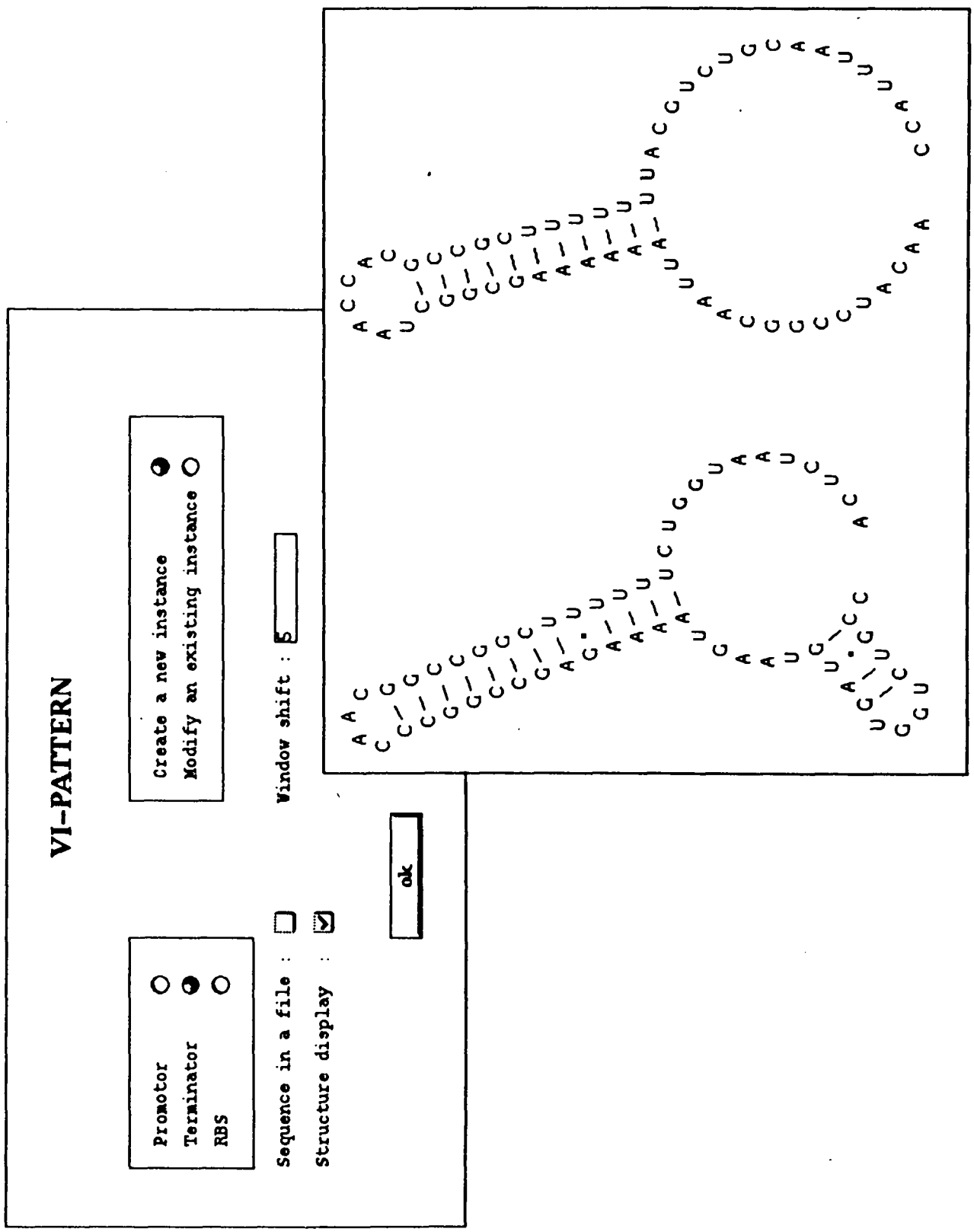
VI-PATTERN
Create a new instance
Modify an existing instance
Window shift : 5
Promotor
Terminator
RBS
Sequence in a file :
Structure display :
ok

Fig. 6

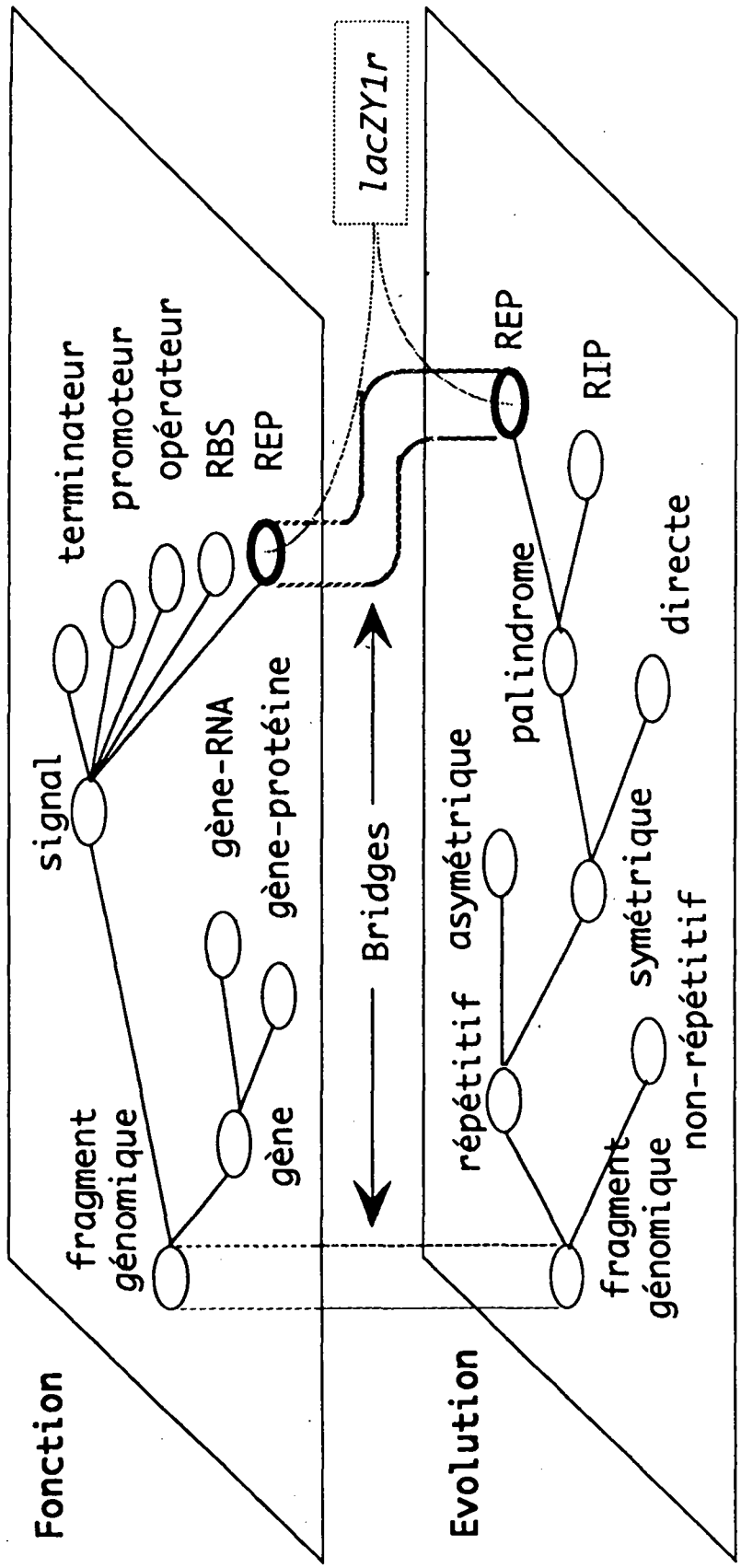
Fonction
Evolution
Bridges
lacZY1r
fragment génomique
signal
gène
gène-RNA
gène-protéine
terminateur
promoteur
opérateur
RBS
REP
répétitif
asymétrique
palindrome
REP
RIP
symétrique
directe
fragment génomique
non-répétitif

Fig. 7

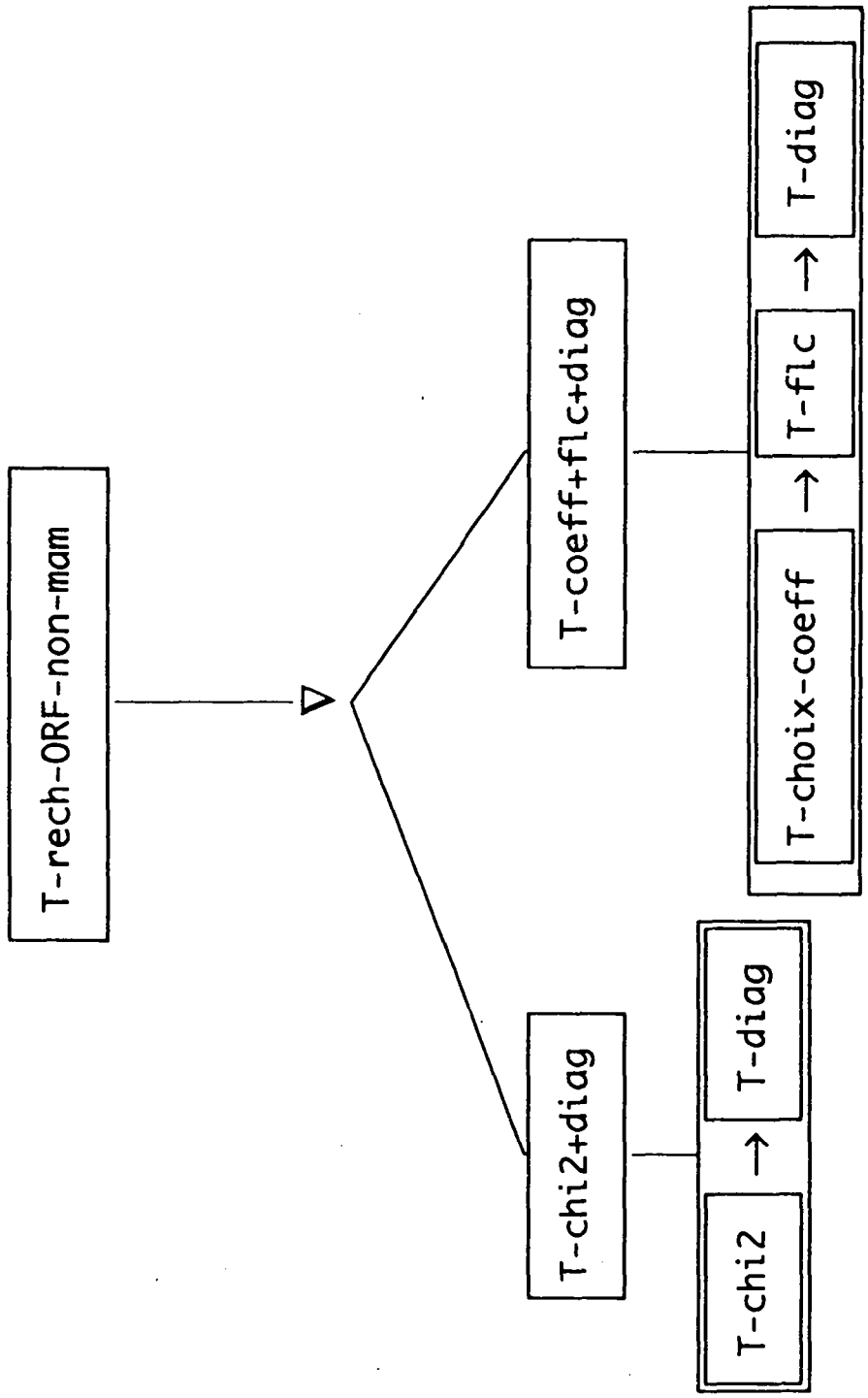
T-rech-ORF-non-mam
T-coeff+flc+diag
T-choix-coeff
T-flc
T-diag
T-chi2+diag
T-chi2
T-diag

Fig. 8

Un Estimateur du Maximum de Vraisemblance pour la Constante de Milieu du Modèle Log-linéaire d'Abondance.

R. Pupier

Laboratoire de Biologie Animale et Appliquée

Université Jean Monnet

23, rue du Dr. Paul Michelon

42023 SAINT-ETIENNE Cedex 2

Résumé.- On montre que l'indice de concentration de Gini permet de construire un estimateur de maximum de vraisemblance pour la constante de milieu **a** dans le modèle log-linéaire de MOTOMURA. Les propriétés statistiques de cet estimateur sont comparées avec celle de l'estimateur classique par moindres carrés, après linéarisation, sur des simulations de relevés faites pour différentes valeurs de **a**; les meilleures performances de cet estimateur devraientt entrainer son utilisation systématique, d'autant que les calculs numériques sont aussi simples que les calculs de régression.

Summary.- Gini's concentration coefficient (1912) is used for building the maximum likelihood estimator of the parameters in a log-linear species abundance distribution. We compare the statistical properties of this estimator with those of the classical mean square estimator, by means of a computer simulation. We recommend the use of this estimator for the simplicity of its computation and its best statistical performances.

Depuis quelques années l'indice de concentration de GINI (1912), ou la courbe de concentration de LORENZ (1905), très prisés par les économistes, font leur apparition en biologie ou en écologie numérique. En voici quelques exemples : déjà en 1976, dans un ouvrage classique, DAGET suggère (pp. 30-31) l'utilisation de la courbe de LORENZ pour étudier une distribution d'abondance; TAILLIE (1979) propose la définition d'indices cohérents d'équitabilité au moyen de cette même courbe; KNOX *et al.* (1989), à la suite des travaux de WERNER and SOLBRIG (1984) se servent de l'indice de GINI pour étudier les distributions de taille des arbres en fonction de la densité initiale de plantation et montrent que la concentration augmente avec la densité; LYONS and HUTCHINSON (1989) comparent, grâce à cet indice, la dispersion spatiale de populations d'insectes (on relève dans cet article quelques erreurs dans les formules théoriques qui ne nuisent heureusement pas à ce travail). On pourra se reporter à la bibliographie de ces auteurs pour obtenir d'autres exemples.

Malgrè quelques mises en garde (DIAZ, 1979, GOUZE & SCIANDRA, 1988) sur le risque d'introduction de biais dans les estimations, on continue à utiliser assez systématiquement, en écologie numérique, un critère de moindres carrés pour

estimer des coefficients sans trop se soucier de sa signification. Ces considérations incitent à inventorier d'autres méthodes adaptées au modèle qu'on a en vue et, si possible, dont la mise en oeuvre soit aussi aisée que celle des moindres carrés.

Dans ce travail, après un bref rappel des propriétés de l'indice de concentration de GINI et de ses liens avec les modèles d'abondance, on définit la problématique des modèles discrets, en précisant leur rôle inférentiel; on met ensuite en évidence, dans le cas du modèle log-linéaire de MOTOMURA (1947), un estimateur du maximum de vraisemblance pour les paramètres du modèle, qui est étroitement lié à la connaissance de l'indice de GINI d'un relevé. Une simulation de tels relevés dans un peuplement ayant les caractéristiques d'un modèle de MOTOMURA permet de comparer les performances de cet estimateur avec celles des estimateurs de moindres carrés; les résultats obtenus et la simplicité de la mise en oeuvre numérique devraient conduire à abandonner pour ces ajustements la méthode de régression linéaire.

1. L'indice de GINI et les distributions d'abondance.

Défini dans un copieux article d'économétrie (GINI, 1912), le coefficient de concentration se calcule pour une variable aléatoire positive discrète ou non. Dans le cas d'une variable X possédant une fonction de densité f, d'espérance E(X) = m et de fonction de répartition F, on définit pour tout $x \geq 0$ la fonction

$$q(x) = \frac{1}{m}\int_0^x t.f(t)\,dt$$

qui conduit à la courbe de concentration de la figure 1, paramétrée par q = q(x), p = F(x).

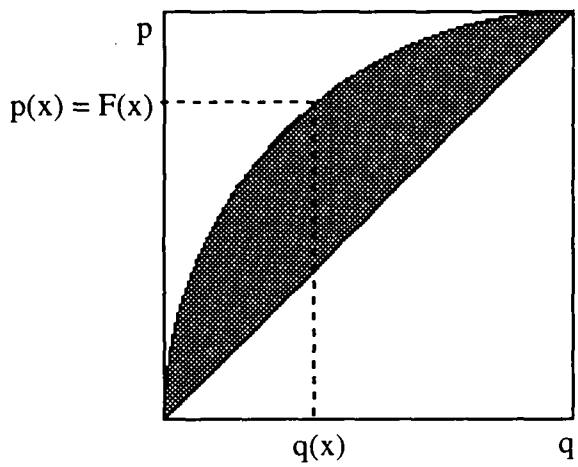

Fig. 1.- Courbe de concentration.

Le double de l'aire hachurée dans la figure précédente est appelé indice de

concentration de GINI qui a ainsi pour expression :

$$IG = 2\int_0^1 F\,dq \; - \; 1. \qquad (1.1)$$

Considérons alors un échantillon de taille n de la variable X et la statistique d'ordre correspondante :

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}. \qquad (1.2)$$

Si l'on utilise la définition précédente avec la fonction de répartition empirique déduite de l'échantillon, on obtient une fonction continue affine par morceaux dont les sommets de la courbe représentative ont pour abscisses $q_k = \sum_{i=1}^{k} \frac{x_{(i)}}{S}$ où $S = \sum_{i=1}^{n} x_i$ et pour ordonnées $p_k = \frac{k}{n}, k = 0, \ldots, n$. C'est la courbe de LORENZ de l'échantillon.

On obtient alors une estimation de l'indice IG de GINI de X par

$$\mathbf{ig} = \frac{1}{(n-1)S}\sum_{i=1}^{n} (2i-n-1)x_{(i)}, \qquad (1.3)$$

et l'on notera que le double de l'aire limitée par la courbe de LORENZ et la première bissectrice est $\frac{n-1}{n}\,\mathbf{ig}$.

De nombreux travaux ont porté sur cet indice de GINI; en particulier DOWNTON (1966) propose comme estimateur sans biais de l'écart-type d'une variable normale

$$\sigma^* = \sqrt{\pi}.\bar{x}.\mathbf{ig},$$

estimateur très efficace étudié par BARNETT *et al* . 1967).

Conformément aux habitudes concernant les modèles d'abondance, on utilisera la statistique d'ordre opposée :

$$x_{(1)} \geq x_{(2)} \geq \ldots \geq x_{(n)} \qquad (1.4)$$

pour laquelle la formule donnant l'indice de GINI devient :

$$(1.5) \qquad \mathbf{ig} = \frac{1}{(n-1)S}\sum_{i=1}^{n}(n+1-2i)x_{(i)}.$$

Les modèles les plus connus de distributions d'abondance pour les espèces d'un peuplement sont ceux de MACARTHUR (1957), MOTOMURA (1947) et PRESTON (1948, 1962). Il ne sera pas étudié ici celui préconisé par FRONTIER (1977), à la suite de MARGALEF (1957), et défini dans un tout autre cadre par MANDELBROT (1953). On peut donner de ces modèles une formulation aléatoire discrète générale qui permettra de décrire correctement les paramètres qui entrent en jeu dans les estimations (ce n'est d'ailleurs pas la seule formulation aléatoire possible, cf. e.g. DENNIS & PATIL, 1979). Etant donné un peuplement composé de n espèces, un modèle d'abondance, aléatoire, discret, consiste en une distribution ordonnée de probabilités p_r, r = 1, ..., n qui donne la probabilité, p_r, pour qu'un individu tiré au hasard dans le peuplement appartienne à l'espèce de rang r (par ordre d'abondance décroissante). Dans le meilleur des cas on ne connait qu'une liste L d'espèces susceptibles d'appartenir au peuplement; en théorie on peut **restreindre** cette liste aux n espèces appartenant ré&ellement à celui-ci. Par abus, on notera encore L = $\{E_1,...,E_i,...,E_n\}$ cette liste restreinte; le modèle définit donc une permutation (ou plusieurs) ρ de $\{1,...,n\}$ où $\rho(r) = i$ désigne le numéro i dans la liste L de l'espèce dont le rang dans le modèle est r. On verra que cette complication apparente est essentielle pour le problème inférentiel de la description du peuplement à partir d'un relevé (échantillon). D'autre part le nombre n est un paramètre important du modèle, ainsi que, s'il y en a, les paramètres propres de la distribution $(p_r)_{r=1,...,n}$.

A chaque modèle on peut donc associer un indice de concentration

$$(1.6) \qquad \mathbf{ig} = \frac{1}{(n-1)S}\sum_{i=1}^{n}(n+1-2i)x_{(i)}.$$

Par exemple pour le modèle de MAC ARTHUR

$$(1.7) \qquad p_r = \frac{1}{n}\sum_{k=1}^{n+1-r}\frac{1}{n+1-k}$$

on vérifie que **ig** = 0.5, ce qui ôte toute pertinence à ce modèle en ce qui concerne l'étude de la diversité. Dans le modèle log-normal de PRESTON, les p_r sont définis par une variable aléatoire log-normale LN(m,σ) :

$$(1.8) \qquad p_r = \frac{e^{\sigma u_r}}{S} \quad \text{o} \quad S = \sum_{r=1}^{n} e^{\sigma u_r}$$

formule dans laquelle u_r est le $\frac{n+1-r}{n+1}$ -fractile de la loi normale centrée réduite. On montre que pour une variable log-normale LN(m,σ) l'indice théorique de concentration est IG = $2\Phi(\frac{\sigma}{\sqrt{2}}) - 1$, Φ étant la fonction de répartition d'une variable de Gauss N(0,1), et comme l'indice **ig** calculé au moyen de (1.6) et (1.8) est une estimation de IG, on aurait là un moyen d'estimer σ.

Enfin le modèle log-linéaire de MOTOMURA, qui fait l'objet essentiel de ce travail, est défini par

$$(1.9) \qquad p_r = \frac{1-\mathbf{a}}{1-\mathbf{a}^n}\,\mathbf{a}^{r-1} \qquad 0 < \mathbf{a} < 1$$

où **a** est appelé **constante de milieu**. L'indice de GINI correspondant vaut :

$$(1.10) \qquad \mathbf{ig} = \frac{n+1}{n-1} - \frac{2}{(n-1)(1-\mathbf{a})} + \frac{2n}{n-1}\cdot\frac{\mathbf{a}^n}{1-\mathbf{a}^n}.$$

La relation (1.10) va être le fondement de l'estimateur que nous allons maintenant étudier. La représentation graphique de **ig** en fonction de **a** fournit les courbes de la figure 2 ci-dessous :

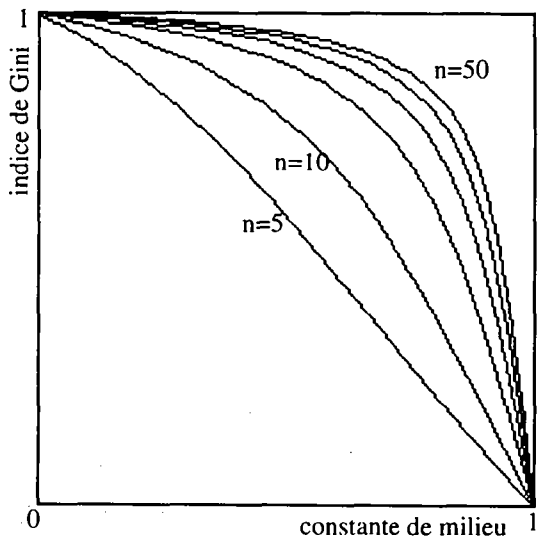

Fig. 2.- Indice de Gini en fonction de la constante de milieu

2. Un estimateur du maximum de vraisemblance.

On considère un peuplement P satisfaisant au modèle de MOTOMURA; L

désignant la liste des espèces, indexée a priori, le problème de l'écologiste de terrain consiste à déterminer au mieux la constante de milieu **a**, le nombre n d'espèces et l'ordre d'abondance des espèces, c'est à dire, la permutation ρ définie ci-dessus. Il a donc à estimer le paramètre $\theta = (\mathbf{a},n,\rho)$ où $\mathbf{a} \in]0,1[$, $n \in \mathbf{N}^*$ et ρ est une permutation de $\{1,...,n\}$. Pour ce faire il possède un relevé effectué dans la biocénose, pratiqué par un tirage dont les conditions d'échantillonnages précisées au préalable essaient de le rapprocher au mieux d'un tirage au hasard. Ce relevé se présente donc sous la forme d'un échantillon $\omega = (k_1, ..., k_i, ...,k_n)$, les indices étant ceux de la liste L, et où les k_i sont les effectifs observés de chaque espèce, certains d'entre eux pouvant être nuls.

On note $N = \sum_{i=1}^{n} k_i$ l'effectif total de l'échantillon. La statistique d'ordre décroissant définit une (ou plusieurs) permutation(s) δ, équivalentes :

$$k_{\delta(1)} \geq k_{\delta(2)} \geq ... \geq k_{\delta(n)}.$$

L'ordre, non nécessairement unique, défini sur L par δ est en général différent de celui défini par ρ. Enfin l'échantillon réordonné par ρ, $(k_{\rho(1)}, ..., k_{\rho(n)})$, est une réalisation d'une variable multinomiale $M(N,p_1,...,p_n)$. Cette description n'est d'ailleurs pas spécifique du modèle de MOTOMURA.

On désignera par $n' = n(\omega)$ le nombre d'effectifs non nuls dans l'échantillon ω et par $a_G = A_G(\omega)$ l'estimation de **a** obtenue par la résolution dans l'intervalle [0,1] de l'équation

$$(2.2) \qquad \frac{n'+1}{n'-1} - \frac{2}{(n'-1)(1-x)} + \frac{2n'}{n'-1}\cdot\frac{x^{n'}}{1-x^{n'}} = ig'$$

où

$$(2.3) \qquad ig' = \frac{1}{(n'-1)N}\sum_{i=1}^{n'} (n'+1-2i)k_{\delta(i)}.$$

Il est clair que (a_G,n',δ) est une estimation de $(\mathbf{a},n,\rho)$.

Théorème.- *L'estimateur du maximum de vraisemblance de* $\theta = (\mathbf{a},n,\rho)$ est :

$$\Theta : \omega \rightarrow (a_G,n',\delta).$$

Le logarithme de la fonction de vraisemblance associée à l'échantillon ω :

$$(2.4) \qquad \varphi_\omega(\mathbf{a},n,\rho) = \ln\left(\frac{1-\mathbf{a}}{1-\mathbf{a}^n}\right) + \frac{\ln \mathbf{a}}{N}\sum_{r=1}^{n} (r-1)k_{\rho(r)}$$

dépend de façon essentielle de l'ordre défini par la permutation ρ inconnue. Posons pour simplifier $c(n,\tau) = \frac{1}{N}\sum_{r=1}^{n}(r-1)k_{\tau(r)}$ pour toute permutation τ. En particulier, en notant encore δ la restriction de δ aux indices 1,...,n' (ceux pour lesquels l'effectif $k_{\delta(i)}$ est non nul), on a :

$$c(n,\delta) = c(n',\delta) = \frac{1}{N}\sum_{r=1}^{n'}(r-1)k_{\delta(r)}$$
$$= \frac{1}{N}\sum_{r=1}^{n'} rk_{\delta(r)} - 1.$$

Lemme.- *Soit* $(x_i)_{i=1,...,n}$ *une famille décroissante de nombres réels. Pour toute permutation* τ *on a* :

$$\sum_{i=1}^{n} ix_i \leq \sum_{i=1}^{n} ix_{\tau(i)}.$$

Le détail de la preuve est laissé aux bons soins du lecteur.

Ainsi $c(n,\rho) \geq c(n,\delta) = c(n',\delta)$.

La dérivée

$$\frac{d\varphi}{da} = -\frac{1}{1-a} + \frac{na^{n-1}}{1-a^n} + \frac{1}{a}c(n,\rho)$$

s'annulle une fois et une seule dans l'intervalle]0,1] car l'équation équivalente

$$\frac{x}{1-x} - \frac{nx^n}{1-x^n} = c(n,\rho) \qquad (2.5)$$

a pour premier membre une fonction strictement croissante de [0,1] sur $[0,\frac{n-1}{2}]$ et les constantes $c(n,\rho)$ prennent justement leurs valeurs dans $[0,\frac{n-1}{2}]$.

Désignons par $M(n,\rho)$ le maximum de $\varphi_\omega(a,n,\rho)$; comme $c(n,\rho) \geq c(n,\delta)$ on a

$$\varphi_\omega(a,n,\rho) \leq \varphi_\omega(a,n,\delta)$$

et donc $M(n,\rho) \leq M(n,\delta)$ ce qui montre déjà que le maximum de la fonction de vraisemblance est obtenu pour $\rho = \delta$. L'équation (2.5) s'écrit alors :

$$\frac{x}{1-x} - \frac{nx^n}{1-x^n} = c(n',\delta)\,. \qquad (2.6)$$

Pour tenir compte du paramètre n on considère l'indice de GINI :

$$ig(n) = \frac{1}{(n-1)N}\sum_{r=1}^{n'}(n+1-2r)k_{\delta(r)}$$

$$= \frac{1}{(n-1)N}\sum_{r=1}^{n'}(n-1-2(r-1))k_{\delta(r)}$$

$$= 1 - \frac{2}{(n-1)N}\sum_{r=1}^{n'}(r-1)k_{\delta(r)}$$

et en utilisant (2.6) :

$$ig(n) = 1 - \frac{2x}{(n-1)(1-x)} + \frac{2n}{n-1}\cdot\frac{x^n}{1-x^n}$$

$$= \frac{n+1}{n-1} - \frac{2}{(n-1)(1-x)} + \frac{2n}{n-1}\cdot\frac{x^n}{1-x^n}.$$

Ainsi l'équation (2.6) est-elle équivalente à celle donnant la constante de milieu **a** en fonction de l'indice de GINI de l'échantillon complet ω (dans lequel on tient compte de la présence de n espèces). Si enfin on remarque que $\varphi_\omega(\mathbf{a},n,\delta) \leq \varphi_\omega(\mathbf{a},n'\delta)$ pour tout $\mathbf{a} \in]0,1[$ car $\frac{1}{1-\mathbf{a}^n} \leq \frac{1}{1-\mathbf{a}^{n'}}$ et, comme on l'a déjà vu, $c(n,\delta) = c(n',\delta)$, on obtient en définitive $M(n,\rho) \leq M(n',\delta)$ pour tout $n \geq n'$ et toute permutation ρ.

Il convient maintenant d'examiner les qualités de cet estimateur du maximum de vraisemblance. En ce qui concerne les estimation $\hat{n} = n'$ et $\hat{\rho} = \delta$, elles ne sont ni très surprenantes ni très enthousiasmantes. Par contre l'estimation $\hat{\mathbf{a}} = a_G$ est à confronter avec les estimations communèment utilisées. La plus courante est celle obtenue par moindres carrés après linéarisation du modèle :

$$(2.7) \qquad \ln\left(\frac{k_{\delta(i)}}{N}\right) = \ln C + i.\ln \mathbf{a}.$$

On note A_{MC} cet estimateur. On peut également concevoir, avec les progrès de l'analyse numérique dans les laboratoires de biologie, que l'on cherche à affiner cette estimation par une méthode de moindres carrés non linéaire, du type GAUSS-NEWTON, et on notera A_{GN} un tel estimateur (on a utilisé la version préconisée par DIAZ (op. cit.)). Cette comparaison a été réalisée au moyen d'une simulation de relevés.

3. Les performances des estimateurs de a dans une simulation.

On a vu que dans un modèle aléatoire discret d'abondance, un relevé dans un peuplement s'assimile à la réalisation d'une variable aléatoire multinomiale $M(N,p_1,...,p_n)$ où les probabilités p_r sont données par le modèle et N désigne le nombre

d'individus récoltés dans l'échantillon. Pour simuler une telle réalisation il faut utiliser un générateur de nombres pseudo-aléatoires répartis uniformément dans [0,1], dont la principale qualité est l'indépendance des tirages. ANTONIADIS *et al.* (1985) ont fait une revue critique de certains des générateurs disponibles, et nous remercions ces auteurs d'avoir bien voulu nous fournir celui qu'ils considèrent comme le mieux adapté aux micro-ordinateurs de notre laboratoire.

En faisant l'hypothèse qu'un peuplement suivait exactement un modèle de MOTOMURA, plusieurs séries de Q relevés de N individus pour n espèces ont été réalisées, et ce, pour différentes valeurs de **a** : nous donnons les résultats concernant Q = 100, N = 500 et n = 12 (cf. annexe).

Pour chaque relevé les estimations a_G, a_{MC} et a_{GN} sont calculées, la distribution théorique d'effectifs correspondant à ces estimations est déterminée, et les ajustements testés par un test de χ^2, tant avec les échantillons qu'avec la distribution originelle. On remarquera que dans une telle simulation, contrairement au cas réel, l'indexation de l'échantillon se fait par rapport au modèle théorique, ici connu, et non par rapport à une liste d'espèces établie a priori. Ceci n'ôte cependant rien aux conclusions car c'est la comparaison des permutations ρ et δ qui importe. On trouvera en annexe quelques relevés entièrement traités.

* Quelques statistiques élémentaires :

	A_{MC}			A_{GN}			A_G		
a	m	var	t	m	var	t	m	var	t
0,30	0,3233	0,001451	6,116	0,3007	0,000568	0,305	0,3055	0,000265	3,371
0,40	0,4177	0,001210	5,101	0,3973	0,000670	-1,061	0,4024	0,000293	1,402
0,40	0,4192	0,001290	5,344	0,3972	0,000620	-1,139	0,4032	0,000334	1,742
0,50	0,5076	0,000905	2,523	0,5020	0,000618	0,824	0,5046	0,000302	2,675
0,50	0,5052	0,001053	1,616	0,4946	0,000771	-1,956	0,4987	0,000306	-0,736
0,56	0,5638	0,000521	1,658	0,5598	0,000590	-0,077	0,5614	0,000213	0,929
0,60	0,6008	0,000496	0,352	0,6007	0,000595	0,298	0,6012	0,000221	0,774
0,70	0,6887	0,000349	-6,050	0,7024	0,000237	1,315	0,6990	0,000135	-0,902
0,80	0,7996	0,000236	-2,209	0,8012	0,000176	0,910	0,8002	0,000117	0,177

Tableau I.- Statistiques élémentaires des estimateurs pour les 9 groupes.

On constate d'une part que l'estimateur A_{MC} donne les résultats les plus éloignés des valeurs attendues; dans la plupart des groupes l'intervalle de confiance à 5% de la moyenne observée ne contient pas la vraie valeur **a**. De plus cet estimateur semble posséder une dérive liée à la valeur de **a** : il sous-estime les fortes valeurs de **a** et sur-estime les faibles valeurs. Par contre les deux autres estimateurs ont des résultats assez comparables, même si certains groupes donnent une valeur de T largement significative; on ne peut dégager une tendance comme pour A_{MC}, et on remarque en particulier que

l'un des groupes à **a** = 0,50 est mal ajusté par A_{GN} et l'autre par A_G. En ce qui concerne la variance c'est A_G qui réalise la meilleure performance avec des valeurs pratiquement indépendantes de **a**, alors que les variances pour A_{MC} et A_{GN} sont des fonctions décroissantes de **a**, A_{GN} ayant néanmoins la plus faible variance des deux. En résumé c'est l'estimateur traditionnelement utilisé, A_{MC}, qui a les moins bonnes performances.

* Les ajustements sur les échantillons :

Un test d'ajustement du χ^2 a été pratiqué entre chaque échantillon simulé et les distributions "théoriques" calculées avec les valeurs estimées a_{MC}, a_{GN} et a_G. Le niveau descriptif du χ^2 observé a été calculé et nous a permis, selon la pratique commune, de rejeter l'hypothèse que l'échantillon suivait un modèle de MOTOMURA lorsque ce niveau descriptif était inférieur à 5% (resp. 1%). Le tableau II nous donne le nombre de rejets pour chaque groupe de 100 échantillons avec les trois estimateurs.

	A_{MC}		A_{GN}		A_G	
%	1	5	1	5	1	5
0,30	17	35	9	16	0	5
0,40	16	35	6	12	0	4
0,40	16	29	8	11	0	4
0,50	13	20	7	16	0	4
0,50	4	16	4	9	0	2
0,56	3	9	4	5	0	1
0,60	2	12	4	6	0	2
0,70	3	5	0	3	0	1
0,80	0	1	0	0	0	0

Tableau II.- Nombre d'ajustements refusés par un test de χ^2 aux risques de 1% et 5% pour les trois estimateurs.

On constate la très mauvaise qualité de l'estimateur A_{MC} qui conduit, au risque 5%, à rejeter 35 échantillons pour **a** = 0,30 ou **a** = 0,40. L'estimateur A_{GN} n'est guère meilleur pour les faibles valeurs de **a**. Seul A_G conserve un pourcentage de rejet compatible avec le risque. Ceci montre combien la qualité de l'estimation d'un paramètre peut influencer la réponse à une question très qualitative. La valeur précise de **a** n'a aucun intérêt pratique pour l'écologiste, mais une mauvaise estimation conduira dans des limites souvent très étroites au rejet d'un modèle parfaitement valable (on trouvera en annexe quelques cas traités intégralement).

On dispose donc de trois méthodes pour, à partir d'un relevé, calculer un "relevé" théorique satisfaisant au modèle de MOTOMURA, la constante de milieu étant estimée par l'un des estimateurs précédents. Par abus de langage on notera encore A_{MC}, A_{GN} et A_G ces trois méthodes. Certains des échantillons simulés sont vraisemblablement

très éloignés du modèle originel (voir annexe pour un exemple); on a donc recensé le nombre d'échantillons pour lequel le modèle est refusé par 1, 2 ou 3 méthodes. Le tableau III nous donne cette information, où l'on constate qu'aucun échantillon est rejeté par la seule méthode A_G, que bien sûr, A_{MC} est celle qui rejette, seule, le plus d'échantillons, mais que A_{GN} pour sa part en rejette un certain nombre qui sont, par ailleurs, acceptés par A_{MC}. Ceci montre que, en dehors de leurs mauvaises qualités statistiques, les estimateurs A_{MC} et A_{GN} sont peu fiables. A trois exceptions près les échantillons rejetés par A_G le sont aussi par les deux autres.

	1	2	3	4	5	6
0,30	24	5	6	0	0	5
0,40	25	3	6	0	1	3
0,40	20	1	6	1	0	3
0,50	13	9	3	0	0	4
0,50	11	5	3	0	1	1
0,56	7	3	1	0	0	1
0,60	6	0	4	0	0	2
0,70	3	1	1	0	0	1
0,80	1	0	0	0	0	0

Tableau III.- Nombre d'ajustements refusés par A_{MC} seul (1), A_{GN} seul (2), A_{MC} et A_{GN} (3), A_{GN} et A_G (4), A_G et A_{MC} (5), tous les estimateurs (6).

* Ajustements avec le modèle originel :

On a également comparé chacune des distributions théoriques, calculées précédemment, avec les effectifs du modèle originel; dans le calcul du χ^2 c'est celui-ci qui fournit bien entendu les effectifs dits théoriques.

	A_{MC}		A_{GN}		A_G					
	1%	5%	1%	5%	1%	5%	χ_2	niv.des	A_{MC}	A_{GN}
0,30	13	29	0	1	0	1	9,02	2,9%	22	0
0,40	14	21	0	3	0	0	7,23	20,4%	26	4
0,40	11	19	0	1	0	0	7,95	15,9%	25	2
0,50	3	6	0	2	0	0	7,92	24,4%	17	8
0,50	5	10	0	1	0	0	7,55	27,3%	17	8
0,56	1	1	1	1	0	0	7,93	33,9%	6	8
0,60	0	1	1	1	0	0	7,89	44,4%	6	9
0,70	0	0	0	0	0	0	5,61	84,7%	13	9
0,80	0	0	0	0	0	0	5,49	90,5%	8	1

Tableau IV.- Comparaison des ajustements avec le modèle originel.

Le tableau IV des rejets observés est très comparable au tableau II, même si les écarts sont moindres, car le remplacement d'un échantillon par un ajustement revient à effectuer un

lissage des données et fournit dans tous les cas une valeur de χ^2 inférieure.

On a par ailleurs, pour affiner la comparaison, indiqué le χ^2 maximum observé avec l'estimateur A_G ainsi que son niveau descriptif, et le nombre de χ^2 observés pour A_{MC} et A_{GN} qui sont supérieurs à cette valeur. On constate là encore la moins grande précision de ces deux derniers estimateurs (à une exception près)

Conclusion.

L'estimation de la constante de milieu **a** dans le modèle log-linéaire d'abondance joue un rôle fondamental dans l'acceptation ou le rejet du modèle. De façon traditionnelle cette estimation s'effectue par moindres carrés après linéarisation du modèle. Les biais introduits par cette méthode sont importants. On peut également utiliser directement une méthode de type GAUSS-NEWTON sans linéarisation préalable. En analysant l'ensemble des paramètres qui interviennent dans le modèle – outre la constante de milieu, il y a le nombre n d'espèces, inconnu, et l'ordre d'abondance de ces espèces symbolisé par une permutation ρ de l'ensemble $\{1,...,n\}$ – on a pu montré que l'estimateur du maximum de vraisemblance du paramètre $\theta = (\mathbf{a}, n, \rho)$ s'obtient en utilisant l'indice de concentration de Gini d'un relevé d'espèces observé dans le peuplement, les estimations de n et de ρ étant données respectivement par le nombre n' d'espèces et l'ordre d'abondance du relevé. Le calcul de l'estimation de **a** se fait par résolution d'une équation polynomiale de degré n' par une méthode standard (Newton). Le calcul numérique de cette estimation n'est pas plus compliqué que le calcul de régression linéaire de la méthode classique. Par contre les résultats statistiques sont nettement meilleurs : l'estimateur de **a** ainsi obtenu a une variance très faible et ne présente pas de biais. Des ajustements effectués après des simulations de relevés issus de modèles log-linéaires montrent dans tous les cas que les performances de cet estimateur sont supérieures à celles des deux autres, puisque le nombre de refus du modèle est de l'ordre de grandeur du risque pris dans le test de χ^2, ce qui est loin d'être le cas avec la méthode traditionnelle ou la méthode de GAUSS-NEWTON.

Annexe.

Voici quelques relevés simulés et leurs traitements par les diverses méthodes. Les paramètres de la simulation sont **a** = 0,56 et N = 500.

Effectifs du modèle théorique :

220,21 123,32 69,06 38,67 21,66 12,13 6,79 3,80 2,13 1,19 0,67 0,37

Relevé n°4.

Il s'agit d'un relevé atypique.

258	121	49	31	17	12	5	2	1	1	3	0

Ordre δ :

258	121	49	31	17	12	5	3	2	1	1	0

Estimations de **a** et valeurs des χ^2 :

$\hat{a}_{MC} = 0{,}566871$ $\hat{a}_{GN} = 0{,}469006$ $\hat{a}_{G} = 0{,}514409$

A_{MC}	$\chi^2 = 18{,}05$	DL = 6	Niveau descriptif : 0,6%	Refus
A_{GN}	$\chi^2 = 18{,}45$	DL = 5	Niveau descriptif : 0,2%	Refus
A_{G}	$\chi^2 = 7{,}08$	DL = 5	Niveau descriptif : 21,4%	Accept.

On remarquera que le relevé simulé est très éloigné du modèle originel (χ^2 = 15,66, DL = 7). La "géométrie" du modèle est cependant conservée avec une constante de milieu différente (â = 0,515); les deux estimations par moindres carrés s'écartent en sens opposé de cette valeur.

Relevé n°55.

Ce relevé ne permet pas d'ajuster un modèle log-linéaire.

247	99	76	26	24	14	5	4	1	2	1	1

Ordre δ :

247	99	76	26	24	14	5	4	2	1	1	1

Estimations de **a** et valeurs des χ^2 :

$\hat{a}_{MC} = 0{,}588761$ $\hat{a}_{GN} = 0{,}497114$ $\hat{a}_{G} = 0{,}546810$

A_{MC}	$\chi^2 = 20{,}72$	DL = 7	Niveau descriptif : 0,4%	Refus
A_{GN}	$\chi^2 = 25{,}33$	DL = 6	Niveau descriptif : 0,01%	Refus
A_{G}	$\chi^2 = 13{,}58$	DL = 6	Niveau descriptif : 3,7%	Refus

Relevé n°82.

Exemple de biais important pour l'estimateur A_{MC}.

222	114	77	42	20	16	7	1	1	0	0	0

Estimations de **a** et valeurs des χ^2 :

$\hat{a}_{MC} = 0{,}500072$ $\hat{a}_{GN} = 0{,}564592$ $\hat{a}_{G} = 0{,}559606$

A_{MC}	$\chi^2 = 21{,}57$	DL = 5	Niveau descriptif : 0,06%	Refus

A_{GN}	$\chi^2 = 5,78$	DL = 6	Niveau descriptif : 44,8%	Accept.
A_G	$\chi^2 = 5,87$	DL = 6	Niveau descriptif : 43,8%	Accept.

Alors que les effectifs du relevé sont tout à fait dans l'ordre de grandeur du modèle, l'estimateur A_{MC} sous-estime fortement la constante de milieu et l'ajustement par cette méthode est en défaut; les estimations par les deux autres méthodes sont très proches l'une de l'autre.

Relevés réels.

A titre de comparaison nous avons appliqué notre méthode à deux relevés tirés de la littérature (DAGET, 1976). L'un concerne une pèche dans le ruisseau Etea (Zaïre), et le modèle de Motomura ne semble pas convenable; l'autre un peuplement de mollusques benthiques pour lequel le modèle log-linéaire semble s'appliquer.

Pour le relvé de poissons on trouve respectivement :

$a_{MC} = 0,7308$ $a_{GN} = 0,5191$!!! $a_G = 0,7060$.

Les χ^2 sont très grands pour les deux valeurs proches : 0,7308 et 0,7060 et conduisent au rejet du modèle. On peut par ailleurs expliquer la grande déviation observée pour l'estimateur a_{GN} par le fait que le principe géométrique de descente de la méthode de GAUSS-NEWTON est particulièrement sensible à l'adéquation du modèle avec les données numériques, ce qui met l'accent sur la démarche circulaire qui consiste à estimer **a** pour juger de cette adéquation. Ce autre défaut de cet estimateur mériterait d'être étudié en détail (voir DIAZ, *op. cit.*.).

Pour le relevé de mollusques, on obtient :

$a_{MC} = 0,5133$ $a_{GN} = 0,4615$ $a_G = 0,4627$

et les χ^2 respectifs :

33,13 21,53 21,48.

Références.

ANTONIADIS A., BERRUYER J. et FILHOL A. (1985), Multidétecteurs Bidimensionnels. I. Simulation d'un spectre avec un ou plusieurs pics de Bragg, Rap-port 85AN19T, Institut Max von Laue-Paul Langevin, Grenoble.

BARNETT F.C., MULLEN K. and SAW J.G. (1967). Linear estimates of a population scale parameter, *Biometrika*, 54: 551-554.

DAGET J. (1976). Les Modèles mathématiques en Ecologie, Masson Ed., Paris, 172 pp.

DENNIS B., PATIL G.P., ROSSI O., STEHMAN S. and TAILLIE C. (1979). A Bibliography of Literature on Ecological Diversity and related Methodology, in *Ecological Diversity in Theory and Practice*, J.F. GRASSLE, G.P. PATIL, W. SMITH & C. TAILLE Eds, International Co-operative Publishing House, Fairland, Maryland, USA : 320-353

DIAZ G. (1979). Ajustement de données expérimentales par des sommes d'exponentielles au sens des moindres carrés, *Pré-publications du Département de Mathématiques*, Université de Saint-Etienne, n°3.

DOWNTON F. (1966). Linear estimates with polynomial coefficients, *Biometrika*, 53 : 129-141.

FRONTIER J. (1977). Réflexions pour une Théorie des Ecosystèmes, *Bull. Ecol.*, 8, (4) : 445-464.

GINI C. (1912). Variabilita e Mutabilita, contributo allo studio delle distribuzioni e relazioni statistice., *Studi Economico-Giudirici della R. Universita di Cagliari*, 3, 2.

GOUZE J.L. et SCANDRIA A. (1988), La loi exponentielle et ses vérifications expérimentales en Biologie, *Les Cahiers d'EDORA*, Rapport Recherche INRIA n°866 : 109-116.

KNOX R.G., PEET R.K. and CHRISTENSEN N.L. (1989), Population Dynamics in loblolly Pine stands : changes in skewness and size inequality, *Ecology*, 70 (4) : 1153-1166.

LORENZ M.O. (1905). Methods of measuring concentration of wealth, *J. Amer. statist. Assoc*, 9 : 209-219.

LYONS N.I. and HUTCHESON, K. (1989). Measures of the Dispersion of a Population Based on Ranks, in *Estimation and Analysis of Insect Populations*, J. BERGER, S. FIENBERG, J. GANI, K. KRICKEBERG and B. SINGER, Eds, Lecture Note in Statistics, n° 55, Springer-Verlag, Berlin : 370-377.

MAC ARTHUR R.H. (1957). On the relative abundance of bird species, *Proc. Nat. Acad. Sci.*, 43 : 293-295.

MANDELBROT B (1953). Contribution à la théorie mathématique des communications, Thèse Univ. Paris, Publ. Inst. Stat. Univ. Paris, 2, (1/2), 121 pp.

MARGALEF R. (1957), La teoria de la informacion en Ecologia, *Mem. real. Acad. Ciencias Artes Barcelona*, 37 : 373-449.

MOTOMURA I. (1947). Further notes on the law of geometrical progression of the population density in animal association (en japonnais, résumé en anglais), *Seiri Seitai*, 1, 55-60.

PRESTON F.W. (1948). The commonness and rarity of species, *Ecology*, 29, 254-283.

PRESTON F.W. (1962). The canonical distribution of commonness and rarity, *Ecology*, 43 : 185-215 et 410-432.

TAILLIE C. (1979). Species Equitability : A Comparative approach, in *Ecological Diversity in Theory and Practice*, J.F. GRASSLE, G.P. PATIL, W. SMITH & C. TAILLIE Eds, International Co-operative Publishing House, Fairland, Maryland, USA : 51-61.

WEINER J. and SOLBRIG O.T. (1984), The meaning and measurement of size hierarchies in plant populations, *Oecologia* (Berlin), 61 : 334-336.

ETUDE DE LA CROISSANCE ET DE L'ABSORPTION DE NITRATE CHEZ UNE ALGUE PHYTOPLANCTONIQUE (*PROROCENTRUM MINIMUM* : DINOPHYCEAE) SOUMISE À DES APPORTS IMPULSIONELS ET PÉRIODIQUES DE NITRATE

ANTOINE SCIANDRA

Station Zoologique - UA 716

Villefranche-sur-mer

B.P. 28 - 06230

ABSTRACT

The growth and nitrate uptake rates of the red-tide dinoflagellate *Prorocentrum minimum* were measured in a chemostat culture system in which nitrate was added in the same proportions every 1, 2 or 3 days. In comparison with continuous nitrate supply, the rate of cell division was not affected by the 1 or 2 day pulse treatments, whereas it fell drastically when a nitrogen source was added only every 3 days. Delayed steady uptake rates were reached during the 1 or 2 day pulse phases, which reflected a mid-term adaptation of the cell uptake process under discontinuous nutrient supply. This adaptation permitted *P. minimum* to maintain a steady growth rate under these regimes. During the one pulse per 3 day treatment, the maximal uptake rate measured during each pulse experiment increased importantly, which reflected a long term adaptation, but it was not sufficient to maintain the initial growth rate. For low frequencies of nitrate supply, uptake and growth rate became largely uncoupled. It is concluded that *P. minimum* is a species able to form a large internal pool of nitrogen which constitutes a competitive advantage discussed in the light of *in situ* observations.

RESUME

Les processus de croissance et d'absorption de nitrate ont été mesurés chez une algue phytoplanctonique *Prorocentrum minimum* (Dinophycée) au sein d'un système de culture en continu. Après une période d'alimentation continue en nitrates, ces derniers sont injectés par impulsions successives avec des périodicités de 1, 2 et 3 jours, et des concentrations telles que la culture reçoive en moyenne au cours du temps la même quantité de nitrate. En comparaison avec un mode d'alimentation continu, le taux de division cellulaire n'est pas affecté pour les périodes impulsionelles de 1 et 2 jours. Par contre, il chute brutalement pour des impulsions effectuées une fois tous les trois jours. Pour les périodicités de 1 et 2 jours, un taux d'absorption stable n'est observé à la suite de l'impulsion qu'après un temps de latence qui reflète une adaptation à moyen terme permettant aux cellules alimentées de façon discontinue de maintenir un taux de croissance stable. Pour des apports de nitrate effectués seulement une fois tous les trois jours, les cellules atteignent un taux maximum d'absorption beaucoup plus important qui traduit une adaptation à long terme qui reste néanmoins insuffisante pour maintenir un taux de croissance constant. Pour les basses fréquences impulsionelles, croissance et absorption deviennent des processus très découplés. En définitive, *P. minimum* est une espèce capable de former des réserves intracellulaires de nitrate importantes qui lui confèrent par rapport à d'autres espèces un avantage compétitif qui est discuté au vu d'observations faites *in situ*.

INTRODUCTION

Cette étude traite des rapports entre le processus d'absorption de sels nutritifs d'une espèce du phytoplancton, *Prorocentrum minimum* , et sa croissance, celle-ci étant représentée par le taux de division cellulaire de la population. L'absorption désigne le mécanisme par lequel des ions nutritifs inorganiques en solution dans l'eau de mer, en l'occurrence les nitrates dans notre étude, traversent la membrane de la cellule. L'assimilation désigne la chaîne de réactions enzymatiques que subissent les ions devenus intracellulaires avant d'être incorporés sous forme de protéines associées soit à l'activité enzymatique de la cellule, soit à la synthèse d'ADN, soit au stockage d'énergie. Cette étude s'inscrit dans la problématique de l'adaptation des espèces phytoplanctoniques au sein du milieu océanique où le facteur qui limite la croissance (dans notre cas, les nitrates) peut avoir une fluctuation temporelle complexe. La prise en considération que la couche superficielle de mélange de l'océan (environ les 100 premiers mètres) n'est pas un environnement stable, et la démonstration expérimentale que les processus physiologiques du phytoplancton sont adaptés à un environnement fluctuant (Harris, 1986), ont suscité durant la dernière décade un engouement important pour l'étude des processus de croissance et d'absorption dans des conditions instationnaires de sels nutritifs.

Les premières tentatives menées pour quantifier l'influence de la concentration de sels nutritifs sur l'activité des cellules phytoplanctoniques ont été faites dans des conditions dites stables, où les algues sont acclimatées au niveau nutritif externe avant que les mesures de croissance ou d'absorption ne soient faites. Les cinétiques ainsi obtenues mettent en évidence des relations hyperboliques entre d'une part la concentration externe de sels nutritifs (x, μg-atN.l^{-1}) et d'autre part la croissance $\mu(x)$ en jour^{-1} (modèle de Monod), et la vitesse d'absorption $\rho(x)$ en μg-at N.jour^{-1}. cellule^{-1} (relation de Michaëlis-Menten).

$$\mu(x) = \mu m \frac{x}{k\mu + x}$$

$$\rho(x) = \rho m \frac{x}{k\rho + x}$$

A l'état stable, tel qu'il peut être obtenu dans un chémostat, ces deux expressions représentent les mêmes flux de matière: tout l'azote absorbé est investi sous forme de biomasse azotée, donc de croissance algale (si les processus d'excrétion sont négligeables).

L'expression de Monod, qui a été utilisée pour décrire la croissance du phytoplancton en milieu océanique par Dugdale (1967), met en relation la croissance μ qui est un indice endogène, avec une variable exogène, la concentration de sel nutritif externe. Droop (1968) a établi une relation également de type

hyperbolique, mais plus réaliste puisqu'elle prend en considération la concentration interne de l'élément qui limite la croissance (q, en µg-atN.cellule^{-1}), que l'on appelle également le quota interne:

$$\mu(q) = \mu'_m \ (1-\frac{k_q}{q})$$

où μ'_m est le taux de croissance maximal théorique obtenu pour une valeur infinie de q, et k_q le quota interne minimum en dessous duquel la croissance n'est plus possible. Morel (1987) a montré l'équivalence de ces trois relations à l'équilibre lorsque les flux sont stationnaires. Cela signifie que dans des conditions de stabilité, la croissance algale peut être connue d'après la simple détermination du taux d'absorption de sels nutritifs.

Quand la croissance est loin de la stationnarité, les modèles classiques de Monod, de Droop, et de Michaëlis-Menten ne sont plus aptes à représenter en détail ni la croissance, ni l'absorption de sels nutritifs. Nombreux sont les cas cités dans la littérature où ces modèles ne peuvent être adaptés aux cinétiques observées (Murphy, 1980; Goldman et McCarthy, 1978; McCarthy, 1981; Cunningham et Maas, 1978; Burmaster, 1979; Cunningham, 1984; DeManche *et al.*, 1979). Cette inaptitude a initialement été démontrée en soumettant des cellules carencées à une impulsion unique du facteur limitant Conway *et al.*, 1976). Ce type de perturbation quasi instantanée constitue bien entendu un cas extrême de variation. Même si de telles impulsions sont improbables dans le milieu naturel, en raison des mécanismes de diffusion spatiale qui lissent toutes variations importantes, ces expérimentations se sont néanmoins révélées utiles pour révéler l'existence des processus d'adaptation potentiels chez différentes espèces du phytoplancton. En régime instationnaire, les cinétiques d'absorption et de stockage de la plupart des nutriments deviennent autrement complexes, en raison du nombre important d'inter-relations physiologiques au sein de la cellule, et des priorités données à la synthèse des différents constituants cellulaires (Harris, 1986). Les constantes (coefficients de demi-saturation k, taux maximum, ρ_m, μ_m) deviennent en réalité des variables qui peuvent évoluer dans le temps en fonction des conditions intra et extracellulaires. La diversité des types d'adaptation enregistrés reflète également la complexité des processus d'adaptation qui diffèrent d'une espèce à l'autre (Collos, 1986).

L'étude des processus d'adaptation dans l'environnement marin est chose ardue, non seulement parce que ces processus, étant non linéaires, sont difficiles à analyser, mais aussi parce que les modes de variation lagrangienne des facteurs limitants *in situ* sont complexes et mal connus (Marra, 1990; McCarthy et Altabet, 1984). Quelques modèles mathématiques ont néanmoins permis de prédire l'existence durable de certaines fréquences caractéristiques, comme les bouffées de

nitrate à travers la nitracline (Klein, 1984; Woods et Wiley, 1972). En dépit du fait que des modes plus complexes existent probablement en raison de la simultanéité et de la diversité des processus physiques qui entretiennent l'instabilité dynamique de l'environnement marin (Denmann et Gargett, 1983), on peut néanmoins utiliser comme descripteur de la variabilité d'un facteur limitant sa période de variation (Harris, 1984).

Etant donné que la gamme des fréquences de variabilité des nutriments est assez large, il apparaît judicieux de mesurer les taux de croissance et d'absorption de cellules soumises à des impulsions de différentes fréquences et amplitudes. Les travaux de Suttle *et al.*, (1987), Olsen *et al.* (1989), Sommer (1985), ainsi que d'autres, ont clairement démontré qu'il n'est plus possible d'aborder la problématique de la compétition entre espèces sans considérer les modes de variation temporelle des facteurs limitants. Quarmby *et al.* (1982) ont par exemple montré que des nitrates ajoutés périodiquement à des cultures de diatomées modifient leur cinétique de croissance par rapport à une alimentation continue. Malheureusement, toutes ces études sont généralement limitées à l'influence d'une seule fréquence impulsionelle, et par conséquent sont inaptes à illustrer la diversité des comportements cellulaires face à différentes fréquences de variation.

En outre, les études des effets d'impulsions sur la croissance et sur le taux d'absorption sont le plus souvent accomplies à la suite d'un pré-conditionnement des cellules réalisé dans un système de culture en continu, de sorte que les pools internes sont stabilisés à un niveau fixé par le taux de renouvellement de la culture. L'apport impulsionel est donc effectué sur des cellules qui sont supposées être à l'état stable, situation improbable en mer. En procédant à l'étude de trains d'impulsions de différentes fréquences, il devient par contre possible d'analyser les processus de croissance et d'absorption de cellules plus où moins éloignées d'un état d'équilibre, condition nécessaire pour optimiser l'étude de processus non linéaires. Il est clair que l'intervalle de temps ente deux impulsions consécutives ainsi que leur amplitude, sont des facteurs cruciaux, puisqu'ils déterminent l'état physiologique des cellules au moment de l'impulsion (Collos, 1980).

Une telle approche est ardue, parce qu'une importante quantité de données est nécessaire pour analyser la non-linéarité des cinétiques observées (Collos, 1983; Goldman *et al.*,1981; Harrison *et al.*, 1989). Des modèles mathématiques comme ceux de Turpin *et al.* (1981) ont tenté de fournir des résultats théoriques sur l'effet d'apports intermittents sur la croissance phytoplanctonique. Cependant, étant donné la complexité des modes d'absorption et d'utilisation des nutriments par les cellules (Dortch, 1982), de tels modèles sont peu crédibles, surtout s'ils formalisent la croissance et l'absorption à partir de relations établies pour l'état stable.

Sont par conséquent présentés dans ce paragraphe les résultats d'une étude expérimentale dont le but a été d'évaluer l'adaptation de *Prorocentrum minimum* à divers régimes de fluctuation d'azote inorganique, en l'occurrence des nitrates. *P. minimum* est une espèce dont la distribution et la dynamique de croissance ont été très largement étudiées en relation avec les facteurs physiques de l'environnement (Tyler et Seliger, 1978; 1981). Les implications écologiques de nos résultats sur le développement d'eaux colorées dues à cette espèce sont discutées dans la conclusion.

MATERIELS ET METHODES

Le détail de la mise en culture de *Prorocentrum minimum* est exposé dans Sciandra, 1991). Rappelons simplement que sa prolifération s'effectue dans un chémostat de 3,3 litres. A l'aide de ce dispositif, le taux de croissance de l'algue, sa concentration ainsi que celle des nitrates se stabilisent après un certain temps à des niveaux fixés par le taux de renouvellement de la culture (0,16 jour^{-1}) par un milieu neuf, ainsi que par la concentration de nitrate (le facteur limitant) dans ce dernier. Lorsque la stabilité est atteinte, la culture est toujours alimentée avec un milieu neuf et dépourvu de nitrate. Ceux-ci sont injectés séparément de façon discontinue. Suivant la fréquence des impulsions, la concentration de la solution injectée est ajustée de sorte que l'apport moyen de nitrate dans la culture reste identique à l'apport effectué en mode d'alimentation continu. Sur toute la durée de l'expérience, la culture reçoit en moyenne la même quantité de nitrate. Seule la fréquence d'approvisionnement change. L'ensemble du système (acquisition sels nutritifs, température, pH, basculement du mode continu vers le mode impulsionel) est piloté par un automate élaboré à la Station Zoologique de Villefranche-sur-mer (Malara et Sciandra, 1991). Durant cette expérience, l'azote inorganique total a été dosé sans distinction entre nitrate et nitrite. Le taux de dilution a été maintenu constant durant toute l'expérience, de sorte que le taux de variation instantanée des nitrates dans le chémostat s'écrit:

$$\frac{dx}{dt} = -\rho(x)\ y + D\ (C-x)$$

x (µg-at N .l^{-1}) et y (cell.l^{-1}) étant respectivement la concentration d'azote inorganique et la densité cellulaire dans le chémostat, $\rho(x)$ (µg-at N .cell^{-1}.h^{-1}) le taux d'absorption d'azote inorganique, D (h^{-1}) le taux de dilution, et C (µg-at N .l^{-1}) la concentration de nitrate dans le milieu de renouvellement. Pour les impulsions où un taux d'absorption stable a été observé durant un certain temps avant la limitation due à l'épuisement du sel, on a supposé que la cinétique de variation suivait la loi de Michaëlis-Menten:

$$\rho(x) = \rho m \frac{x}{k_N + x}$$

ρm (μg-at N .cell^{-1}.h^{-1}) étant le taux d'absorption maximum, et k_N (μg-at N.l^{-1}) la constante de demi-saturation. Les estimations de k_N et ρm sont obtenues par régression non-linéaire grâce à l'algorithme de Gauss-Marquard. La fonction estimée est la taux de variation de la concentration en azote inorganique:

$$\frac{dx}{dt} = -\rho m \frac{\bar{x}}{k_N+\bar{x}} y + D (c-\bar{x})$$

qui est ajustée aux valeurs calculées à partir de la variation entre deux mesures consécutives de x:

$$\frac{\Delta x}{\Delta t} = \frac{x_{t+\Delta t}-x_t}{\Delta t}$$

La concentration moyenne de x au cours de cette variation est calculée d'après:

$$\bar{x} = \frac{x_t+x_{t+\Delta t}}{2}$$

Le taux de croissance μ est estimé à partir de la pente d'une fonction polynomiale ajustée aux concentrations cellulaires, qui représente l'évolution moyenne de la culture au cours du temps (Fig. 1).

$$\mu = \frac{1}{\Delta t} \mathrm{Ln}\left(\frac{y_{t+\Delta t}}{y_t}\right) + D$$

RESULTATS

La Fig. 1 illustre la concentration en azote inorganique et la population de *Prorocentrum minimum* durant les 70 jours d'expérimentation. Les nitrates ne sont plus décelables (moins de 0,02 μg-at N.l^{-1}) à partir du 8ème jour, bien que la concentration cellulaire continue d'augmenter jusqu'au 12ème jour. A ce moment, la culture se stabilise suite à une limitation de la croissance par les nitrates qui sont très faibles, et est considérée à l'état stable. Le tableau I donne les différents modes d'introduction de nitrate dans le chémostat.

Périodes (jours)	Nitrate présent dans le milieu de renouvellement	Impulsion de NO_3		
		Nombre (par période)	Quantité (µgat N par litre de culture)	Fréquence ($Jour^{-1}$)
0 - 4	-	0	-	-
4 - 17	+	0	-	-
14	+	1	6.4	1.00
17 - 24	-	7	6.4	1.00
24 - 32	-	5	12.5	0.50
32 - 38	-	2	19	0.33
39	-	1	19	1.00
40 - 70	+	0	-	-

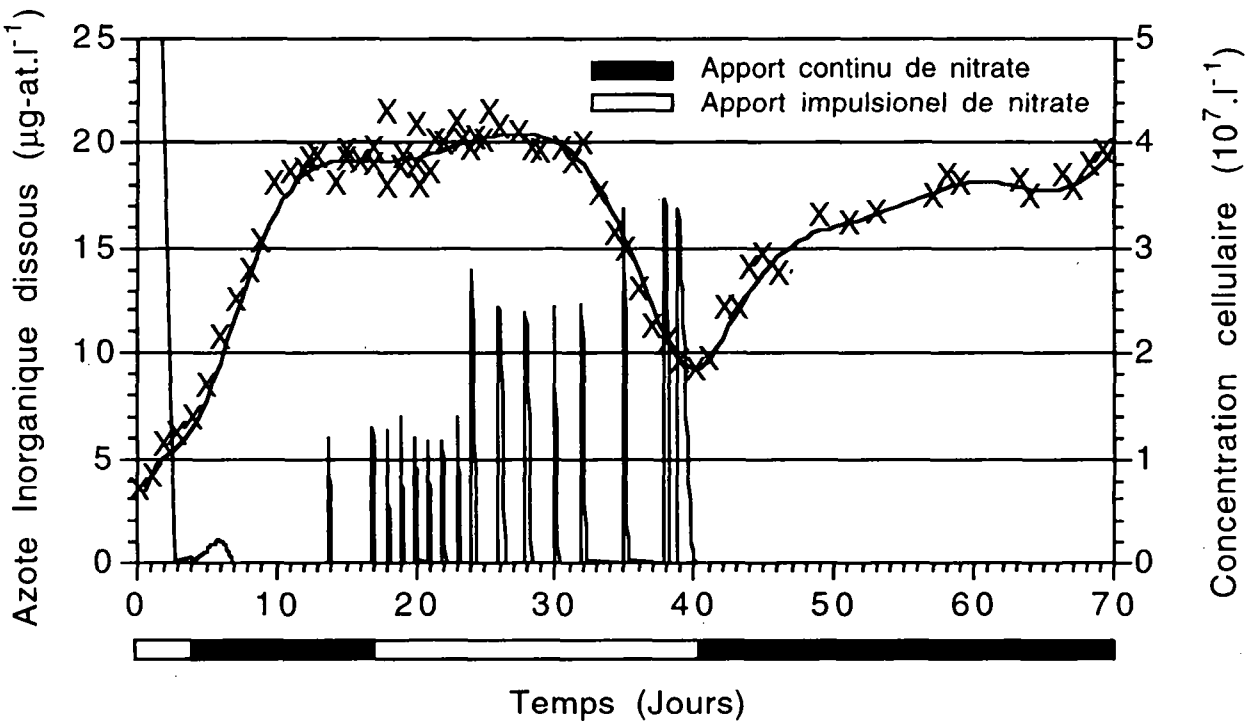

Figure 1. Evolution de Prorocentrum minimum (points et courbe ajustée) et de l'azote inorganique (ligne) dans le chémostat. Pendant la période d'apport impulsionel de nitrate, le milieu de renouvellement continu est dépourvu de nitrate (sauf pour la première impulsion)

L'apport impulsionel de nitrate n'affecte pas significativement le taux de croissance de *Prorocentrum minimum* jusqu'au 30ème jour par rapport au taux de croissance stationnaire obtenu en régime d'apport continu. (Fig. 2). Pendant les 12 premiers jours du régime impulsionel, la concentration cellulaire augmente légèrement, probablement parce que l'état stable n'était pas complètement atteint le

17ème jour, ou suite à une légère baisse de débit de la pompe qui assure le renouvellement du milieu de culture.

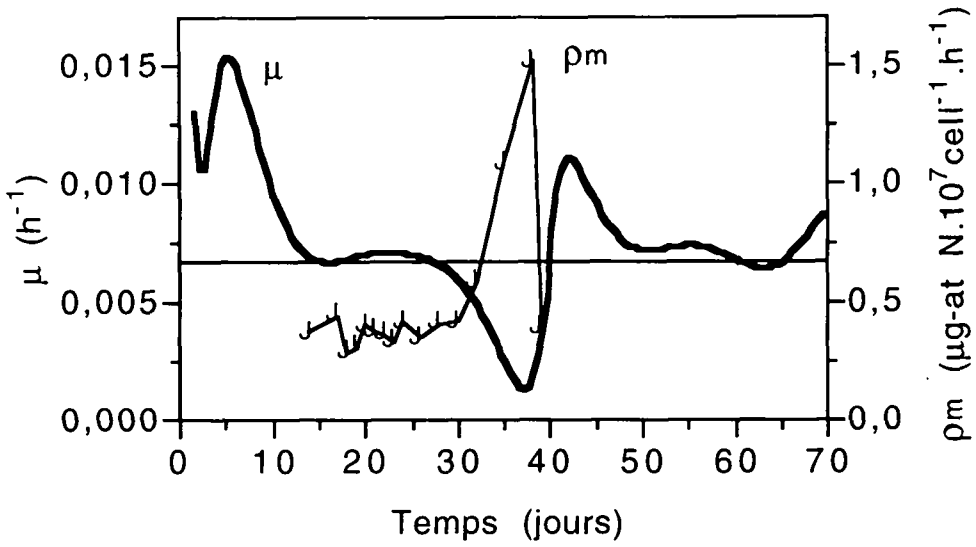

Figure 2. Taux de croissance (ligne continue) et taux maximum d'absorption (points) de *Prorocentrum minimum* mesurés au cours de chaque impulsion.

Inversement, le taux de croissance décroît rapidement jusqu'à de très faibles valeurs (0,03 jour^{-1}) quand l'addition de nitrate est faite une fois tous les trois jours (du 23ème au 38ème jour). On remarque également que le taux de croissance retrouve rapidement des valeurs supérieures au taux de dilution lorsque la dernière impulsion avant le retour à un régime d'apport continu est effectuée seulement 1 jour après la précédente, au lieu de trois. On voit également sur la Fig. 1 que, lorsque le régime d'apport en continu est rétabli, la culture tend à retrouver son niveau initial, mais plus lentement.

On peut calculer que, durant la phase impulsionelle, les cellules de *Prorocentrum minimum* ont absorbé la même masse d'azote qu'en régime continu, moins 4 %. Une différence si faible ne permet vraisemblablement pas d'expliquer en soit la chute du taux de croissance. Au cours de cette expérience, la réactivité de *P. minimum* à la limitation par l'azote dépend étroitement de la fréquence des impulsions. Dans les conditions expérimentales présentes, des périodes d'apport de 1 et 2 jours n'affectent pas la croissance de *P. minimum* , alors que les périodes de 3 jours sont limitantes.

La Fig. 3 illustre les variations temporelles de la concentration en azote inorganique et du taux d'absorption pour quelques impulsions. Il apparaît clairement sur l'ensemble des séries recueillies une augmentation de la non linéarité du processus d'absorption durant la période impulsionelle. Le taux mesuré chez les

cellules à l'état stable le 17ème jour atteint son maximum (ρ=4.0 10^{-8} µgatN.cell^{-1}.h^{-1}) immédiatement après l'impulsion.

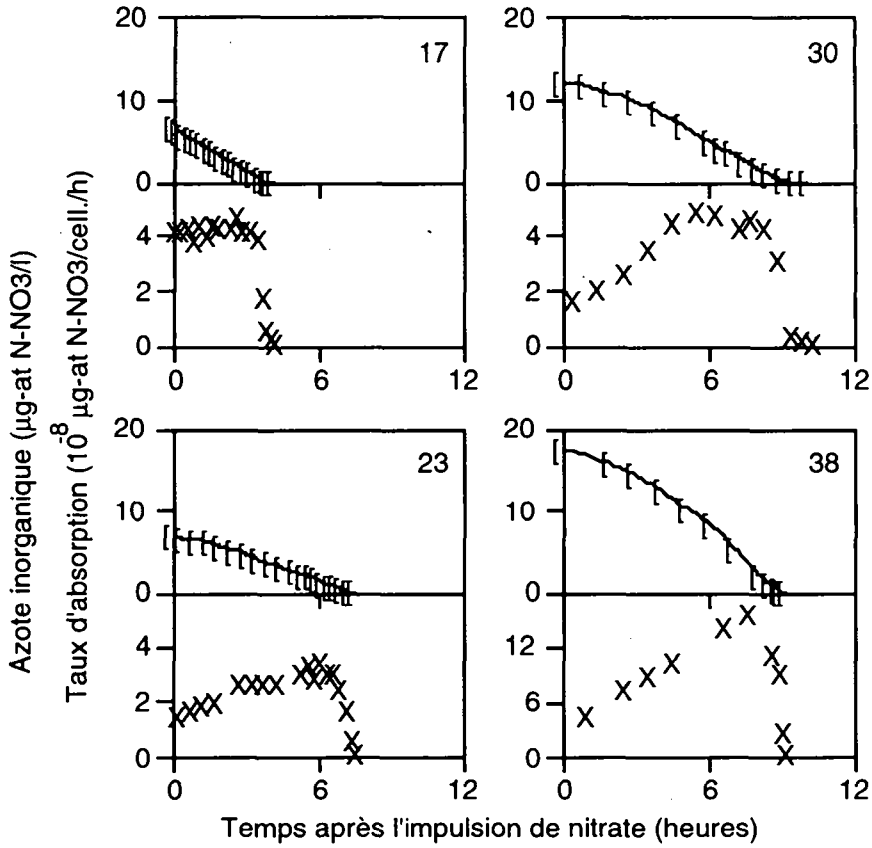

Figure 3. Evolution de l'azote inorganique dissous (B: nitrate et nitrite confondus), et de son taux d'absorption par les cellules (J) suite aux apports impulsionels effectués les jours 17, 23, 30 et 28. Noter que les échelles diffèrent.

Les taux initiaux mesurés pour les impulsions suivantes sont plus faibles et décroissent au fil des impulsions. Quelques heures sont nécessaires pour qu'un taux relativement stabilisé soit atteint avant le contrôle externe du à l'épuisement des nitrates. Ce mode est caractéristique des cellules limitées en azote (Eppley *et al.*, 1969; Caperon et Meyer, 1972; Collos, 1980; Romeo et Fisher, 1982). Les taux maximum atteints au cours de chaque impulsion sont représentés sur la Fig. 2. Par opposition avec la période précédant le 32ème jour durant laquelle on n'observe pas de variation significative du taux d'absorption maximum, une augmentation marquée se produit ensuite, qui est concomitante avec la réduction du taux de croissance. On remarquera également que la première valeur du taux d'absorption qui excède significativement 5.0 10^{-8} µgat N.cell^{-1}.h^{-1} est obtenue à la fin de la période où les impulsions sont distribuées 1 fois tous les deux jours.

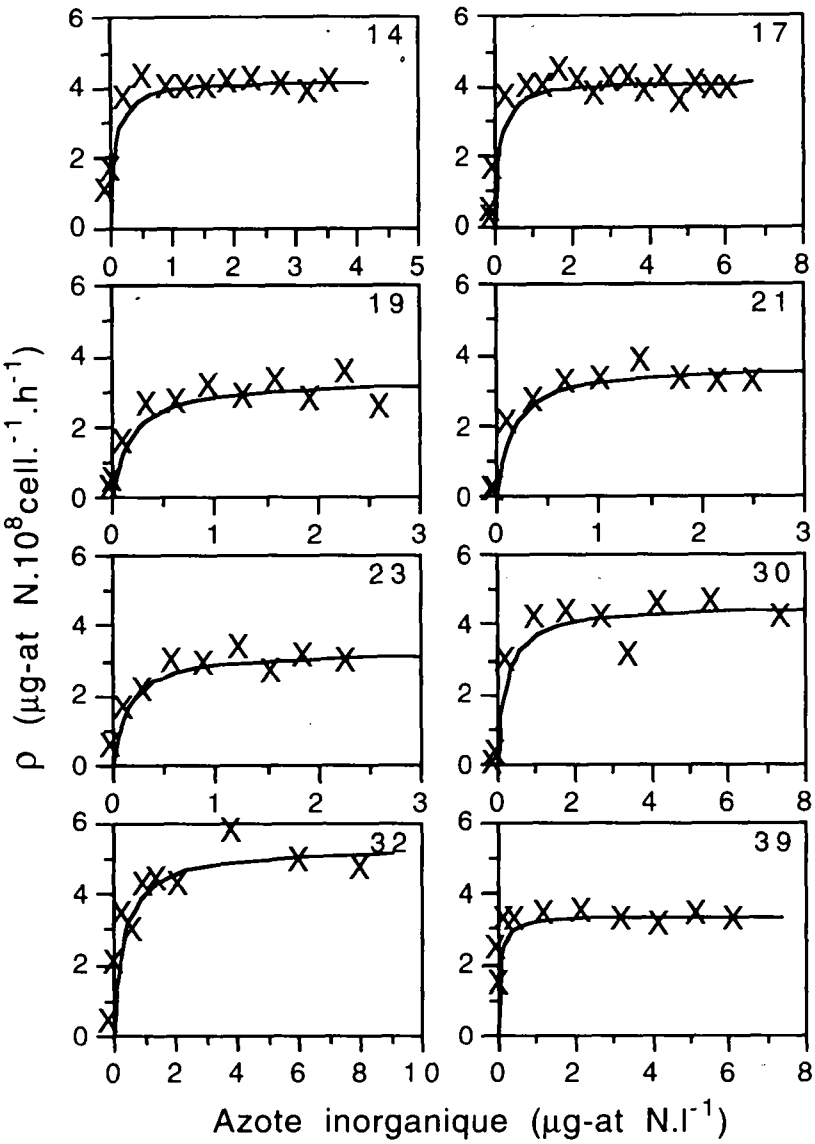

Figure 4. Cinétiques du taux d'absorption obtenues pour les impulsions où un taux maximum stabilisé a été mesuré avant le contrôle externe du à l'épuisement des nitrates par les cellules. Les nombres se réfèrent aux jours des impulsions.

La Fig. 4 montre les cinétiques d'absorption estimées qui ont été obtenues au cours des impulsions pour lesquelles un taux stationnaire a été atteint avant le contrôle externe du à l'épuisement des nitrates. En comparaison avec les valeurs de k_N communément obtenues pour les dinoflagellés, celles que nous mesurons pour *P. minimum* sont étonnamment basses. Ceci est expliqué par le fait que nos mesures d'azote inorganique concernent les nitrates et les nitrites sans distinction. La Fig. 5 montre l'excrétion de nitrite consécutive à une addition de 8 µg-at $N\text{-}NO_3.l^{-1}$ dans une culture de *P. minimum* limitée par les nitrates dans un chémostat identique et avec le même taux de dilution. Plus de 5 % de l'azote apporté est excrété à un taux moyen de 1.7 10^{-9} µg-at $N.cell^{-1}.h^{-1}$. Pour les concentrations de nitrate

inférieures à 2 µg-at $N.l^{-1}$, les nitrites sont à leur tour consommés. Il s'en suit que, si les nitrates ne sont pas mesurés séparément des nitrites, et si ces derniers sont présents en quantités non négligeables, le taux d'absorption mesuré pour les nitrates peut être sous-estimé quand les cellules excrètent des nitrites, et sur-estimé quand les cellules réabsorbent les nitrites en fin d'incubation. Cet artefact conduit à une sous-estimation du k_N (Collos, 1982).

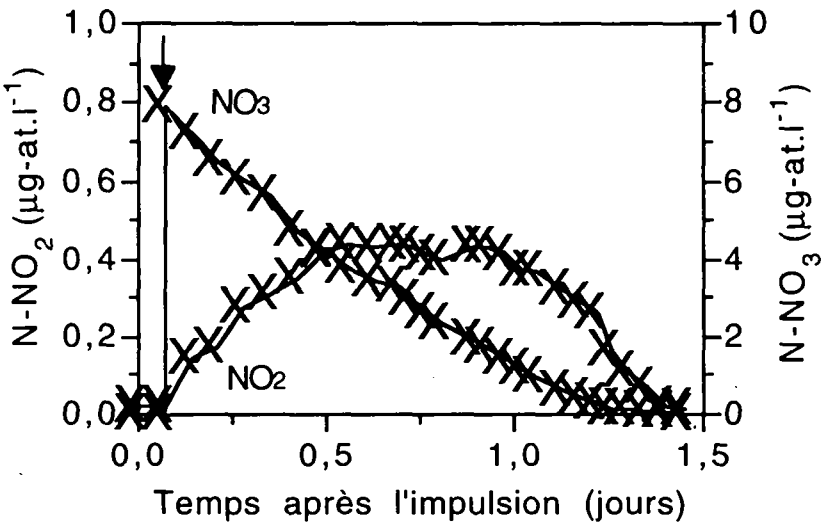

Figure 5. Variation des teneurs en nitrate et nitrite dans une culture en continu de *Prorocentrum minimum* , consécutive à une addition impulsionelle de nitrate (↓)

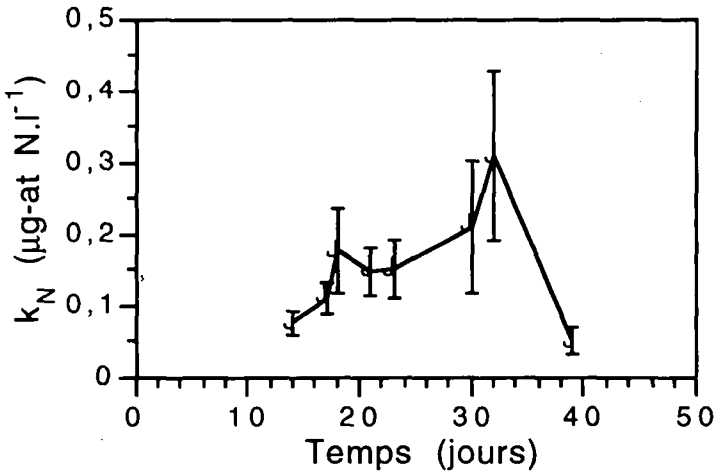

Figure 6. Variation du coefficient de demi-saturation des cinétiques présentées Figure 4. L'intervalle de confiance représente un écart-type.

Nous avons néanmoins reporté les valeurs de k_N sur la Fig. 6, ce qui permet d'avoir un indice de l'évolution de l'"affinité globale" des cellules pour l'azote

inorganique. Une augmentation notable du coefficient de demi-saturation se produit à partir du 32ème jour, comme c'est également le cas pour le ρ_m. Cependant, il est impossible de stipuler si cette variation reflète une réelle modification de l'affinité des cellules, où une diminution de l'excrétion de nitrite avec le temps. Quoiqu'il en soit, cette variation pourrait traduire une modification du métabolisme azoté.

DISCUSSION

La discordance des résultats trouvés dans la littérature qui concernent les effets d'impulsions de nutriments sur la croissance et la consommation (Collos, 1986) trouve probablement son origine dans la diversité des états physiologiques des cellules au moment de l'impulsion (Raimbault et Mingazzini,1987; Elrifi et Turpin, 1987). La plupart du temps, un stimulus unique est porté sur des cellules dont le degré de limitation n'est pas forcément bien défini (Collos, 1983; Dortch *et al.*, 1984). La distinction entre "limité" et "carencé" est un peu arbitraire, si l'on doit comprendre que, dans le premier cas, le quota interne n'est pas à son maximum théorique, et que dans le second, il est proche de son minimum. Mais le "quota interne" (au sens de Droop) est en réalité un index global qui représente la totalité de l'azote cellulaire, sans précision sur sa composition. En fait, suivant les conditions de limitation, de carence ou de réplétion, les divers éléments de structure, de synthèse et génétiques sont différemment affectés (Dortch, 1982), suivant des voies complexes et hiérarchiques (Falkowski *et al.*, 1989; Wheeler, 1983).

Au cours de cette étude, différentes adaptations physiologiques apparaissent chez *Prorocentrum minimum* suivant la fréquence des impulsions. Pour toutes les perturbations, exceptée la première, il apparaît une non-linéarité qui reflète une adaptation à moyen terme des cellules, qui suit l'impulsion. Durant la phase d'une addition par jour, et une partie de la période d'une addition tous les deux jours, un taux d'absorption maximum relativement stabilisé a été obtenu au cours de chaque impulsion, et qui était sensiblement le même que celui obtenu à l'état stable en régime continu. On peut donc supposer qu'à chacune de ces impulsions, les cellules ont eu le temps de retrouver leur équilibre initial.

Par opposition à cette période où il n'apparaît aucune variation significative ni du taux de croissance, ni du taux d'absorption, une adaptation à plus long terme apparaît lorsque les cellules recoivent des nitrates seulement 1 fois tous les trois jours. Cette adaptation est caractérisée par une augmentation du taux maximum d'absorption au fil des impulsions. Durant le régime de périodicité impulsionelle de 2 jours, le taux d'absorption maximum ne commence à augmenter significativement qu'au bout de la 5ème impulsion, ce qui suggère que l'effet de ce régime sur la consommation de sel est retardé. Par la suite, le taux maximum d'absorption mesuré

dépasse le taux obtenu à l'état stationnaire d'un facteur 3 à 4. Il apparaît d'après la Fig. 2 que la croissance commence à être sérieusement ralentie lorsque l'adaptation à long-terme devient plus visible.

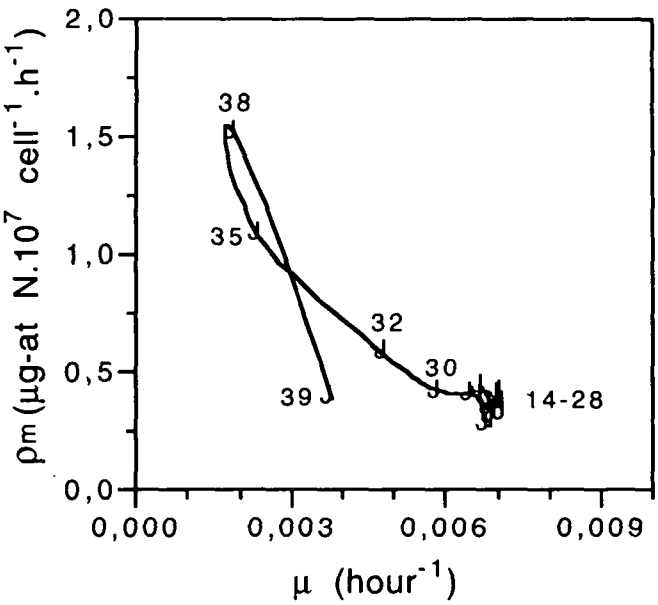

Figure 7. Plan de phase de la croissance et du taux d'absorption maximum. Les deux processus deviennent largement découplés pour la période où les nitrates sont ajoutés seulement 1 fois tous les trois jours. Les nombres indiquent les jours d'impulsions.

Croissance et absorption sont très fortement découplées chez *Prorocentrum minimum* (Fig. 7). Bien que les cellules aient quasiment consommé la même quantité d'azote durant toute l'expérience, sa conversion en croissance dépend de la fréquence des apports de nitrate. La fréquence de 1 jour^{-1} induit une non-linéarité dans le processus d'absorption qui traduit une adaptation à moyen terme suffisante pour permettre aux cellules de garder un taux de croissance stable. Pour des fréquences plus faibles et des amplitudes plus importantes, la non linéarité, qui est accrue, reflète une adaptation plus importante, mais cependant insuffisante pour préserver le taux de croissance. En se référant aux travaux de Dortch (1982), on peut suspecter que, durant la période d'apport journalier, seules les formes les plus labiles de l'azote intracellulaire comme les nitrates et les acides aminés sont mobilisées, alors que pour des intervalles d'addition plus importants, des éléments plus structuraux ou fonctionnels comme les protéines sont utilisés pour satisfaire la demande énergétique des cellules. Pour compenser cette perte de matériel, les cellules augmentent leur consommation de nitrate quand celui-ci est à nouveau disponible, suivant une cinétique qui dépend du degré de limitation. Comme l'assimilation de l'azote nouvellement incorporé nécessite une dépense énergétique et enzymatique, un temps d'induction est nécessaire avant que la consommation atteigne

un régime maximum.

La division cellulaire commence à s'accélérer environ un jour après la fin de la période impulsionelle (39ème jour) qui s'arrête lorsque l'apport de nitrate en continu est réinstallé. Un tel laps de temps et le fait que des périodicités d'apports impulsionels de 1 et 2 jours n'affectent pas le taux de croissance de *Prorocentrum minimum* , suggèrent que cette espèce est apte à accumuler d'importantes réserves nutritives. Ce type de réponse présente un avantage écologique dans les situations où la fréquence de disponibilité en nitrate est inférieure au taux de division cellulaire (Collos, 1986).

Parallèlement aux multiples modes d'adaptation que *Prorocentrum minimum* peut développer face à différentes formes et intensités du spectre de la lumière (Vogel et Sager, 1985; Harding, 1988; Coats, 1988), Paasche *et al.* (1984) ont démontré l'aptitude chez cette espèce de maintenir des taux d'absorption quasiment inchangés d'ammoniaque ou de nitrate pendant la phase nocturne d'un cycle lumineux diurne de 12h:12h. Cette faculté peut permettre de découpler dans le temps la photosynthèse (proche de la surface durant le jour) et la consommation de nitrate (en profondeur, durant la nuit), et pourrait rendre cette espèce compétitive par rapport à d'autres espèces comme les diatomées (Eppley et Harrison, 1975; Harrison, 1976). Tyler et Seliger (1981) rapportent que, au moment où les nutrilites deviennent rares dans les eaux stratifiées de la baie de Chesapeake à la fin de l'été, des efflorescences de *P. minimum* persistent néanmoins avec une consommation de carbone élevée. Des observations préliminaires dans la même zone ont par ailleurs démontré que cette espèce est capable d'effectuer une migration nocturne vers les couches plus riches en nutrilites. Cette aptitude peut effectivement constituer un avantage, seulement si *P. minimum* est à même de supporter plusieurs heures de privation par jour dans les couches supérieures. Notre déduction expérimentale selon laquelle *P. minimum* peut maintenir un taux de croissance constant dans un milieu où l'élément qui limite sa croissance fluctue avec une périodicité de 1 jour, conforte l'hypothèse selon laquelle cette espèce peut effectivement développer un comportement migratoire qui contribue significativement à réduire les effets de la imitation par les nutrilites.

CONCLUSION

Cette étude expérimentale illustre comment certains processus considérés *a priori* comme pouvant être représentés par des cinétiques simples s'avèrent en réalité complexes lorsqu'ils sont étudiés au sein d'un environnement variable. La diversité des conditions expérimentales a permis de révéler la diversité des cinétiques d'absorption chez *Prorocentrum minimum* , traduisant divers modes

d'adaptation face aux variations externes de nitrate. Au delà de l'aspect purement expérimental, se pose le problème de la formalisation des lois observées. Il apparaît clairement que la non-linéarité des processus, ainsi que le découplage entre absorption, assimilation et croissance ne pourraient être correctement restitués par un modèle qu'à la condition que ce dernier intègre les processus de stockage, les phénomènes d'induction enzymatiques, autrement-dit les mécanismes internes de la cellule qui sont régis par des réactions biochimiques. Il va de soi que la réalisation d'un tel modèle est chose ardue à cause de l'effort expérimental nécessaire pour la calibration, et qu'elle n'est donc pas souhaitable pour qui veut se contenter d'une formalisation qui restituerait globalement les phénomènes sans les représenter explicitement. On peut en effet oublier un instant la physiologie de l'espèce que l'on considère alors comme une boîte noire dont les mécanismes et la structure interne se manifestent uniquement par les relations que l'on peut expérimentalement établir entre une fonction d'entrée (fréquence et amplitude des impulsions de nitrate par exemple) et une ou plusieurs sorties (taux de croissance et d'absorption des cellules). En multipliant les expériences avec différents mode d'entrée, on peut obtenir différentes sorties cohérentes les unes par rapport aux autres parce que résultant de lois physiologiques qui restent néanmoins transparentes pour l'expérimentateur. L'idée est donc d'exercer un contrôle sur les cellules avec des forçages qui prennent des valeurs discontinues dans la gamme de ce que peut tolérer l'espèce. En connaissance des entrées et des sorties, on doit chercher à établir une représentation du processus étudié pour une fonction d'entrée continue, à l'aide d'un formalisme qu'il reste à définir en cherchant à le simplifier au maximum.

REFERENCES BIBLIOGRAPHIQUES

Burmaster, D.E. (1979). The continuous culture of phytoplankton: mathematical equivalence among three steady-state models. Am. Nat. 113: 123-134.

Caperon, J., Meyer, J. (1972). Nitrogen-limited growth of marine phytoplankton. I. Changes in population characteristics with steady-state growth rate. Deep-Sea Res. 19: 601-618.

Coats, D.W., Harding, L.W. (1988). Effect of light history on the ultrastructure and physiology of Prorocentrum mariae-lebouriae (Dynophyceae). J. Phycol. 24: 67-77.

Collos, Y. (1980). Transient situations in nitrate assimilation by marine diatoms. I. Changes in uptake parameters during nitrogen starvation. Limnol. Oceanogr. 25: 1075-1081.

Collos, Y. (1982). Transient situations in nitrate assimilation by marine diatoms. II. Changes in nitrate and nitrite following a nitrate perturbation. Limnol. Oceanogr. 27: 528-535.

Collos, Y. (1983). Transient situations in nitrate assimilation by marine diatoms. 4. Non-linear phenomena and the estimation of the maximum uptake rate. J. plankton Res. 5: 677-691.

Collos, Y. (1986). Time-lag algal growth dynamics: biological constraints on primary production in aquatic environments. Mar. Ecol. Prog. Ser. 33: 193-206.

Conway, H.L., Harrison, P.J., Davis, O. (1976). Marine diatoms grown in chemostats under silicate or ammonium limitation. II. Transient response of Skeletonema costatum to a single addition of the limiting nutrient. Mar. Biol. 35: 187-199.

Cunningham, A. (1984). The impulse response of Chlamydomonas reinhardii in nitrite-

limited chemostat culture. Biotechnol. Bioengen. 26: 1430-1435.
Cunningham, A., Maas, P. (1978). Time lag and nutrient storage effects in the transient growth response of Chlamydomonas reinhardii in nitrogen-limited batch and continuous culture. J. General Microbiol. 104: 227-231.
DeManche, J.M., Curl, H.C., Lundy, D.W., Donaghay, P.L. (1979). The rapid response of the marine diatom Skeletonema costatum to changes in external and internal nutrient concentration. Mar. Biol. 53: 323-333.
Denmann, K.L., Gargett, A.E. (1983). Time and space scales of vertical mixing and advection of phytoplankton in the upper ocean. Limnol. Oceanogr. 28: 801-815.
Dortch, Q. (1982). Effect of growth conditions on accumulation of internal nitrate, ammonium, amino acids, and protein in three marine diatoms. J. Exp. Mar. Biol. Ecol. 243-264.
Dortch, Q., Clayton, J.R., Thoresen, S.S., Ahmed, S.I. (1984). Species differences in accumulation of nitrogen pools in phytoplankton. Mar. Biol. 81: 237-250.
Droop, M.R. (1968). Vitamin B12 and marine ecology. IV. The kinetics of uptake growth and inhibition in Monochrysis lutheri. J. Mar. Biol. Assoc. U.K. 48: 689-733.
Dugdale, R.C. (1967). Nutrient limitation in the sea: dynamics, identification and significance. Limnol. Oceanogr. 12: 685-695.
Elrifi, I.R., Turpin, D.H. (1987). Short-term physiological indicators of N deficiency in phytoplankton: a unifying model. Mar. Biol. 96: 425-432.
Eppley, R.W., Harrison, W.G. (1975). Physiological ecology of Gonyaulax polyedra , a red water dinoflagellate of southern california. In: LoCicero, V.R. (ed.) Toxic dinoflagellate blooms, Vol. Proc. Intl. Conf. (Ist) Mass. Sci. Technol. Fndn., Wakefield, 12-22
Eppley, R.W., Rogers, J.N., McCarthy, J.J. (1969). Half-saturation 'constants' for uptake of nitrate and ammonium by marine phytoplankton. Limnol. Oceanogr. 14: 912-920.
Falkowski, P.G., Sukenik, A., Herzig, R. (1989). Nitrogen limitation in Isochrysis galbana (Haptophyceae). II. Relative abundance of chloroplast proteins. J. Phycol. 25: 471-478.
Goldman, J.C., McCarthy, J.J. (1978). Steady state growth and ammonium uptake of a fast-growing marine diatom. Limnol. Oceanogr. 23: 695-703.
Goldman, J.C., Taylor, C.D., Glibert, P. (1981). Nonlinear time-course uptake of carbon and ammonium by marine phytoplankton. Mar. Ecol. Prog. Ser. 6: 137-148.
Harding, W.H. (1988). The time course of photoadaptation to low-light in Prorocentrum mariae-lebouriae (Dinophyceae). J. Phycol. 24: 274-281.
Harris, G.P. (1984). Phytoplankton productivity and growth measurements: past, present and future. J. Plankton Res. 6: 219-237.
Harris, G.P. (1986). Phytoplankton ecology. Structure, function and fluctuation. London, New York,
Harrison, P.J., Parslow, J.S., Conway, H.L. (1989). Determination of nutrient uptake kinetic parameters: a comparison of methods. Mar. Ecol. Progr. Ser. 52: 301-312.
Harrison, W.G. (1976). Nitrate metabolism of the red tide dinoflagellate Gonyaulax polyedra Stein. J. exp. mar. Biol. Ecol. 21: 199-209.
Klein, P., Coste, B. (1984). Effects of wind-stress variability on nutrient transport into the mixed layer. Deep-Sea Res. 31: 21-37.
Malara, G., Sciandra, A. (1991). A multiparameter phytoplanktonic culture system driven by microcomputer. J. Applied Phycol. in press:
Marra, J., Bidigare, R.R., Dickey, T.D. (1990). Nutrients and mixing, chlorophyll and phytoplankton growth. Deep-Sea Res. 1: 127-143.
McCarthy, J.J. (1981). The kinetics of nutrient utilization. In: Platt, T. (ed.) Can. Bull. Fish. Aquat. Sci., Vol. 210. 211-233
McCarthy, J.J., Altabet, M.A. (1984). Patchiness in nutrient supply: implication for phytoplankton ecology. In: Meyers, D.G., Strickler, J.R. (ed.) Trophic interactions within aquatic ecosystems, Vol. 85. AAAS Selected Symposium, 29-45
Morel, F.M.M. (1987). Kinetics of nutrient uptake and growth in phytoplankton. J. Phycol. 23: 137-150.
Murphy, T.P. (1980). Ammonia and nitrate uptake in the lower Great Lakes. Can. J. Fish. Aq. Sci. 37: 1365-1372.

Olsen, Y., Vadstein, O., Andersen, T., Arne, J. (1989). Competition between Staurastrum luetkemuellerii (Chlorophyceae) and Microcystis aeruginosa (Cyanophyceae) under varying modes of phosphate supply. J. Phycol. 25: 499-508.

Paasche, E., Bryceson, I., Tangen, K. (1984). Interspecific variation in dark nitrogen uptake by Dinoflagellates. J. Phycol. 20: 394-401.

Quarmby, L.M., Turpin, D.H., Harrison, P.J. (1982). Physiological responses of two marine diatoms to pulsed additions of ammonium. J. Exp. Mar. Biol. Ecol. 63: 173-181.

Raimbault, P., Mingazzini, M. (1987). Diurnal variations of intracellular nitrate storage by marine diatoms: effects of nutritional state. J. Exp. Mar. Biol. Ecol. 112: 217-232.

Romeo, A.J., Fisher, N.S. (1982). Intraspecific comparisons of nitrate uptake in three marine diatoms. J. Phycol. 18: 220-225.

Sciandra, A. (1991). Coupling and uncoupling between nitrate uptake and growth rate in Prorocentrum minimum (Dinophyceae) under different frequencies of pulsed nitrate supply. Mar. Ecol. Prog. Ser. 72: 261-269.

Sommer, U. (1985). Comparison between steady and non-steady state competition: experiments with natural phytoplankton. Limnol. Oceanogr. 30: 335-346.

Suttle, C.A., Stockner, J.G., Harrison, P.J. (1987). Effects of nutrient pulses on community structure and cell size of a freshwater phytoplankton assemblage in culture. Can. J. Fish. Aquat. Sci. 44: 1768-1774.

Turpin, D.H., Parslow, J.S., Harrison, P.J. (1981). On the limiting nutrient patchiness and phytoplankton growth: a conceptual approach. J. plankton Res. 3: 421-431.

Tyler, M.A., Seliger, H.H. (1978). Annual subsurface transport of a red tide dinoflagellate to its bloom area: water circulation patterns and organism distributions in the Chesapeake Bay. Limnol. Oceanogr. 23: 227-237.

Tyler, M.A., Seliger, H.H. (1981). Selection for a red tide organism: physiological responses to the physical environment. Limnol. Oceanogr. 26: 310-324.

Vogel, H., Sager, J.C. (1985). Photosynthetic response of Prorocentrum mariae-lebouriae (Dinophyceae) to different spectral qualities, irradiances, and temperatures. Hydrobiologia. 128: 143-153.

Wheeler, P.A. (1983). Phytoplankton nitrogen metabolism. In: Carpenter, E.J., Capone, D.G. (ed.) Nitrogen in the marine environment, Vol. 9. Academic Press, Inc., New York, 309-346

Woods, J.D., Wiley, R.L. (1972). Billow turbulence and ocean microstructure. Deep-Sea Res. 19: 87-121.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - B.P. 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Lorraine - Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 Villers lès Nancy Cedex (France)
Unité de recherche INRIA Rennes - IRISA, Campus Universitaire de Beaulieu 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes - 46, avenue Félix Viallet - 38031 Grenoble Cedex 1 (France)
Unité de recherche INRIA Rocquencourt - Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 Le Chesnay Cedex (France)

ISSN 0249 - 6399