UNIVERSITÉ CLAUDE BERNARD - LYON I

L'IDENTIFICATION BIOLOGIQUE ASSISTÉE PAR ORDINATEUR

Mémoire présenté en vue de l'obtention du

DIPLÔME D'HABILITATION À DIRIGER DES RECHERCHES

(application de l'arrêté du 5 juillet 1984)

Renaud FORTUNER

Juin 1992

L'IDENTIFICATION BIOLOGIQUE ASSISTÉE PAR ORDINATEUR

INTRODUCTION

Toute ma carrière s'est déroulée dans des organismes de recherche sans composante enseignement. J'ai travaillé d'abord à l'ORSTOM, au Sénégal puis en Côte d'Ivoire, à une époque où il y avait très peu d'élèves en nématologie. De plus je n'avais pas encore accumulé l'ancienneté et l'expertise qui auraient justifié que l'on me confie une direction de recherches. Je travaille depuis 1980 aux Etats-Unis, au *California Department of Food and Agriculture* (CDFA), dans un laboratoire qui n'a aucune responsabilité d'enseignement. Je n'ai donc jamais encadré d'étudiants mais on peut prendre l'expression "direction de recherches" dans un sens plus large que le seul encadrement de travaux de thèse et la comprendre comme la direction d'une équipe de recherche. J'ai essayé de montrer dans le présent mémoire quelles sont mes aptitudes dans ce domaine en présentant le projet NEMISYS, *Nematode Identification System*, au sein duquel je coordonne les activités d'une équipe pluridisciplinaire composée de nématologistes et d'informaticiens.

Toute recherche originale débute par une idée qu'il faut analyser et traduire en une succession d'étapes réalisables pratiquement. Il faut ensuite trouver les moyens en hommes et en argent pour mener cette recherche à bien, puis il faut faire le travail et, une fois les résultats obtenus, les disséminer par des publications, des conférences, des ateliers, etc. Dans le cas d'une thèse, les responsabilités sont partagées entre le directeur de thèse et l'étudiant. Le directeur souffle l'idée à l'étudiant et trouve les moyens financiers. L'étudiant, lui, s'occupe des détails pratiques et fait la plus grosse part du travail. Dans le cas du projet NEMISYS, j'ai fait à la fois le travail du directeur de thèse et celui de l'étudiant, surtout en ce qui concerne la partie biologique. De plus, j'ai dirigé les travaux de mes collègues informaticiens et j'ai coordonné les efforts des taxonomistes qui participent au projet. Le succès de ces activités de direction devraient permettre de juger mon aptitude à diriger la thèse d'un étudiant, si j'ai un jour à le faire.

§

Le sujet du projet NEMISYS est l'identification, une activité souvent traitée en parent pauvre de la systématique bien que, sous le nom trompeur de "classification", elle soit l'objet de nombreux travaux de pointe en informatique, systèmes experts, réseaux neuronaux et bien d'autres. En biologie, l'identification devrait être tenue, non pour une activité mineure mais pour le fondement de toute étude, des plus théorique aux plus appliquées. Comment étudier la physiologie d'un animal, ou, s'il est phytophage, ses effets sur les cultures, si l'on n'a pas d'abord établi son identité? L'étude de la diversité biologique, cette expression à la mode, commence par la détermination des espèces présentes dans divers biotopes. Il faudrait pouvoir développer les études faunistiques, surtout dans les régions menacées par le développement. Les forêts vierges en particulier devraient être prospectées avant leur disparition complète car elles sont souvent le refuge d'espèces intéressantes pour la systématique. Par exemple, j'ai identifié dans la forêt de Taï, en Côte d'Ivoire, les nématodes *Rotylenchoides intermedius* et *R. affinis,* depuis transférées dans le genre *Helicotylenchus*, qui dérivent des espèces typiques d'*Helicotylenchus* par la perte de la branche génitale postérieure chez les femelles. Ces espèces sont précieuses car elles peuvent permettre de tester l'hypothèse selon laquelle la perte de la branche postérieure est accidentelle et ne donne aucune indication phylogénétique (Sternberg & Horvitz, 1982). D'autres espèces rares existent dans cette forêt vierge (*Criconemella yapoense, Xiphinema yapoense, X. douceti, Hylonema ivorense*) et elles disparaissent toutes lors du défrichage précédant la mise en culture (Fortuner & Couturier, 1983).

L'identification est importante également pour la protection de l'agriculture. Par exemple, l'espèce *Hirschmanniella oryzae* cause des dégâts graves au riz irrigué dans de nombreuses régions mais elle n'existe pas en Californie où elle est classée en catégorie A, c'est-à-dire qu'elle est considérée comme étant "*an organism*

of known economic importance subject to state enforced action involving: eradication, quarantine regulation, containment, rejection, or other holding action[1]" *(CDFA, Plant Industry Policy Letter 89-2)*. Par contre, l'espèce proche *H. belli* est classée en catégorie D (*no action, organisms of little or no economic importance*[2]) parce que cette espèce indigène ne semble pas faire de dégâts. Ces deux espèces sont morphologiquement si proches l'une de l'autre qu'il est difficile de les identifier. Ceci complique l'application des lois de quarantaine et pose d'autre part la question de l'identité de ces deux espèces, car il se pourrait que *H. belli* ne soit qu'une variante locale de l'espèce cosmopolite *H. oryzae* et qu'elle ne puisse attaquer le riz que par suite de conditions adverses de sol, de climat, ou d'autres facteurs existant en Californie. J'ai pu montrer (Fortuner & Maggenti, 1991) que ces deux espèces pouvaient en fait être différenciées à l'aide d'analyses discriminantes, ce qui résout le premier problème et qui pourrait indiquer une réponse négative à la deuxième question (voir § 3.6).

L'identification permet aussi d'aborder l'étude de l'écologie et de définir la structure des peuplements sous diverses conditions. Par exemple, les rizières du nord du Sénégal (Région du Fleuve) supportent des peuplements composés presque uniquement de *H. oryzae,* tandis que l'espèce *H. spinicaudata* est rarement identifiée (Fortuner, 1975). Au contraire, les peuplements des rizières de Casamance, au sud du pays, sont composés surtout de *H. spinicaudata* (Fortuner & Merny, 1974). Ceci est en rapport avec le régime de l'eau dans les deux régions. Les terrains de Casamance restent humides toute l'année, avec une végétation adventice en inter campagne, quand même il ne s'y fait pas une double culture annuelle de riz. Dans la région nord au contraire, une sécheresse totale de huit mois succède à quatre mois de culture rizicole souvent conduite sous plus d'un mètre d'eau. *H. oryzae* est l'une des rares espèces capable à la fois de survivre de longues semaines en l'absence d'oxygène et de supporter le dessèchement du sol en se mettant en régime d'anhydrobiose (Fortuner, 1976). Ces comparaisons n'auraient pas pu être faites si les espèces n'avaient pas été clairement identifiées.

Il faut identifier, mais qui s'occupe de faire les identifications? Traditionnellement ce sont les experts identificateurs, systématiciens, techniciens, spécialistes, et leurs étudiants. A l'heure actuelle, la systématique est déconsidérée par les pouvoirs administratifs. De nombreux postes libérés par les départs en retraite ne sont pas remplacés et le nombre des personnes expertes en identification diminue sans cesse. Les méthodes moléculaires souvent présentées comme la panacée pour toutes les questions dont s'occupe la systématique traditionnelle, y compris l'identification. Il est certain qu'elles offriront bientôt un moyen sûr et rapide pour identifier les espèces économiquement importantes. En nématologie, plusieurs équipes s'occupent de développer des sondes moléculaires ou des dessins d'électrophorèse spécifiques, en particulier Triantaphyllou aux Etats-Unis, l'équipe de l'INRA à Antibes, et d'autres chercheurs. Au vu des résultats déjà obtenus il est certain que ces méthodes (aidées par la nouvelle technique PCR) offriront bientôt une méthode sure pour l'identification de parasites tels les *Meloidogyne,* les hétérodérides, peut-être certains *Ditylenchus* et *Aphelenchoides*. La gravité des dégâts causés aux cultures par les espèces appartenant à ces groupes justifie l'engagement de moyens puissants pour leur identification. Par contre, je ne crois pas qu'il soit pratiquement possible d'engager les fonds, les hommes et le matériel nécessaires pour développer des sondes moléculaires pour chacune des millions d'espèces animales qui ont déjà été décrites. Trois à quatre mille espèces de nématodes phytoparasites sont déjà connues et bien d'autres restent à découvrir. Les méthodes moléculaires ne peuvent à elles seules assurer tous les besoins en identification biologique. Elles doivent être réservées aux espèces économiquement importantes et venir en complément des autres aides à l'identification.

Les méthodes d'identification traditionnelles sont conçues pour être utilisées par des experts. Les experts disparus, les biologistes généralistes ne savent pas s'en servir, l'identification ne se fait plus ou elle se fait mal. La seule alternative est l'identification assistée par un ordinateur ayant accès à une base de connaissances où serait engrangée l'expertise des meilleurs identificateurs actuels. Il est urgent de recueillir le savoir des chercheurs qui partent à la retraite et de le préserver pour le mettre à la disposition des futures générations. Ceci n'est pas chose aisée. L'identification est souvent décrite comme un art et se refuse à

---

[1] un organisme connu pour son importance économique soumis à une action de l'état comprenant: éradication, mise sous quarantaine, endiguement, rejet, ou autre action d'arrêt.

[2] aucune action, organisme ayant une importance économique faible ou nulle.

l'analyse. Avant d'entreprendre le sauvetage des connaissances pour l'identification biologique, il faut d'abord définir les différentes stratégies d'identification puis décomposer les connaissances mises en jeu en les classant en diverses catégories: données, métadonnées, règles heuristiques, etc. Ce n'est qu'à la suite de cette mise en ordre qu'il devient possible de formuler les questions à poser aux experts pour obtenir d'eux des réponses sans ambiguïté. Il faut ensuite transformer ces réponses en données qui puissent être entrées dans une base de connaissances adéquate, définir la structure de cette base de connaissances, et l'organiser de façon que les utilisateurs futurs puisse l'interroger facilement pour récupérer les connaissances qui lui ont été confiées.

De nombreuses méthodes d'identification utilisant les ordinateurs ont été proposées, mais la plupart s'adressent à un petit groupe d'espèces, généralement au niveau du genre. Une méthode globale, résolvant d'un coup le problème de l'identification de milliers d'espèces, impose d'autres contraintes. Par exemple, il n'est plus possible de redécrire toutes les espèces considérées et il faut bien se résoudre à utiliser les données de la littérature en dépit de tous leurs défauts, en particulier données manquantes et absence d'un format commun.

§

J'ai été confronté à ces problèmes depuis le début de ma carrière de nématologiste. A mon arrivée au laboratoire ORSTOM de Dakar en 1971, j'ai été chargé du relevé faunistique des nématodes des plantes cultivées au Sénégal. J'ai fait le même genre d'études en Côte d'Ivoire à partir de 1976. Passant en Californie en 1980, j'ai trouvé un emploi dans le laboratoire de nématologie du CDFA où je dois identifier les espèces présentes dans les plantes sous quarantaine. J'ai utilisé plusieurs moyens traditionnels d'aide à l'identification, clé dichotomique (Fortuner, 1970) et clé tabulaire (Fortuner, 1974), puis j'ai proposé NEMAID, un logiciel pour le calcul de coefficients de similarité entre les spécimens à identifier et toutes les espèces d'un genre (Fortuner, 1983). Ce logiciel est tout à fait satisfaisant et je continue de le développer (Fortuner & Wong, 1983; Fortuner & Ahmadi, 1986; Fortuner, 1986). Il souffre cependant de certaines limitations qui interdisent son emploi par des non-spécialistes.

Après l'achèvement de ma thèse de doctorat (Fortuner, 1986) consacrée à l'étude de la variabilité intra-spécifique et de ses conséquences, j'ai pensé qu'il était nécessaire de repenser le problème, définir le domaine des nématodes phytoparasites, définir les contraintes qui se posent lors de leur identification, et arriver ainsi à une solution globale satisfaisante. En 1987, j'ai débuté une collaboration qui s'est avérée être des plus fructueuse avec deux informaticiens de l'université de Californie à Davis, Jim Diederich et Jack Milton. Le but de nos efforts est la création de NEMISYS (*NEMatode Identification SYStem*), une station de travail experte pour l'identification des nématodes.

Avec l'aide de mes collègues et poussé par leurs questions, j'ai abordé des points fondamentaux pour la définition du processus d'identification. Définissant les diverses stratégies suivies par les experts, je me suis rendu compte que la plus commune, la reconnaissance immédiate, était décrite comme un acte un peu mystérieux. Pankhurst (sous presse) écrit que "*in order to find the name of an unknown specimen, the methods are (...) to know what it is already*[3]". J'ai essayé de mieux cerner cette approche en proposant le concept de *promorphe* qui désigne les formes reconnaissables par tout nématologiste avant l'examen détaillé de leur morphologie. J'ai précisé mes vues sur la nature probabiliste des données morphologiques, vues qui apparaissent déjà dans ma thèse consacrée à l'étude de la variabilité. Pour contourner les problèmes crées par cette variabilité et par les difficultés qu'il y a à décrire certains caractères, j'ai défini les concepts de *caractères primaires d'identification* (caractères faciles à observer) et de *nids d'espèces* (groupes réunissant toutes les espèces ayant les mêmes caractères primaires d'identification).

A mon tour, j'ai aidé mes collègues informaticiens à mettre au point leurs concepts théoriques: station de travail experte, outils, etc. En particulier, j'ai collaboré avec Jim Diederich pour définir la nature d'un "*bio-DBMS*" (système de management de bases de données biologiques) capable de prendre en compte la complexité des divers types de caractères morphologiques et de leurs relations. Il a très tôt été décidé de créer une base de données morphologiques (NEMbase) à partir des descriptions d'espèces et de populations publiées

[3] Pour trouver le nom d'un spécimen inconnu, les méthodes sont (...) de savoir d'emblée qui il est.

depuis une centaine d'années dans la littérature spécialisée. Un logiciel, le Terminator, a été créé pour l'extraction semi-automatique de ce type de données et leur stockage dans une base de données ad-hoc.

Toutes ces activités ont joué dans le cadre du *NEMISYS International Project* (NIP). Ce projet a démarré en 1988 avec un *Advanced Research Workshop* (atelier de recherches avancées) financé par l'OTAN. Cet atelier a réuni une trentaine de spécialistes mondiaux pendant une semaine à l'université de Raleigh (Caroline du Nord) où nous avons défini les besoins en identification pour la nématologie (Fortuner, 1989). NEMISYS étant supposé recueillir l'expertise des spécialistes mondiaux de l'identification pour la mettre au service de tous, le recrutement de participants a continué et à l'heure actuelle, plus de 75 personnes reçoivent le bulletin bimensuel que je publie (*NEMISYS International Project Update*). Des prototypes de NEMISYS et du Terminator ont été établis pour démontrer l'intérêt de nos méthodes.

Le financement du projet a été assuré par plusieurs sources, la *National Science Foundation* (NSF), Sun Microsystems, Hewlett Packard, et le CDFA. Je viens d'obtenir une somme de 20.000 dollars pour tester le Terminator et commencer de peupler la base de données.

Le projet NEMISYS continue de se développer selon plusieurs directions. De nouveaux outils sont en cours d'élaboration et seront ajoutés au prototype de NEMISYS pour le rendre plus proche du système final. La base de données en cours de création permettra de tester le système en vraie grandeur. La consultation des experts du NIP se poursuivra pour la création d'une base de connaissances qui permettra une meilleure utilisation de NEMISYS. Finalement, nous comptons étudier la possibilité de transférer les outils et méthodes créées pour les nématodes phytoparasites à d'autres groupes biologiques.

§

Le présent mémoire reprend quelques unes des questions qu'il a fallu résoudre pour la définition de NEMISYS, une approche globale au problème de l'identification des nématodes.

DESCRIPTION DU DOMAINE

Avant de chercher des solutions à un problème il est nécessaire de bien le définir et cette première partie va s'attacher à décrire le domaine dont je m'occupe, celui des nématodes phytoparasites.

Ce sont des animaux microscopiques (Fig. 1) qui font environ un millimètre de long et 20 à 30 μm de diamètre. Après un examen rapide à la loupe binoculaire, les observations se font au plus fort grossissement du microscope optique ou même au microscope électronique à balayage. Elles portent soit sur des spécimens vivants, soit sur des spécimens fraîchement tués, soit sur des spécimens conservés dans les collections d'un musée depuis parfois des vingtaines d'années. On comprend qu'il est en général très difficile de voir les caractères d'identification chez les nématodes. Il faudrait sélectionner les caractères les plus faciles à observer pour réduire les risques d'erreurs d'identification. Cependant, si l'on demande à un expert de donner une liste de caractères faciles pour l'identification des espèces d'un genre, il est à craindre qu'il se rabatte sur les caractères traditionnellement utilisés dans les clés d'identification, dont certains ne sont pas faciles du tout!

Figure 1: *Helicotylenchus dihystera,* un exemple typique de nématode phytoparasite.

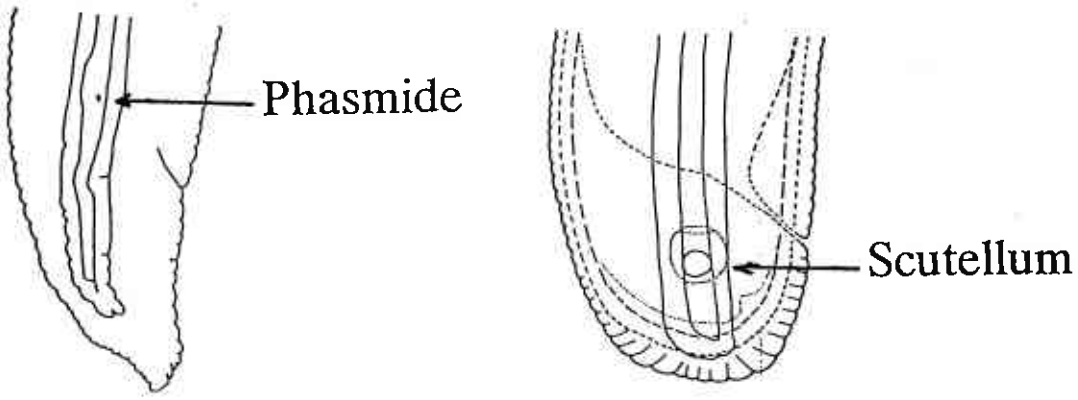

Figure 2: Phasmide: organe chimio-sensible s'ouvrant à l'extérieur par un pore peu visible. Scutellum: phasmide modifiée à large ouverture.

Il a été décidé de ne pas aborder cette question de front mais de décomposer le concept de caractère facile en ses constituants de base. J'ai défini un caractère facile comme étant un caractère non ambigu et non variable qui décrit un aspect d'un organe bien visible dans le groupe considéré (genre, espèce ou nid d'espèce). Ces notions doivent à leur tour être précisées avant que l'on puisse demander à un expert de les définir pour chaque caractère dans un groupe donné.

1.1. Les métadonnées

1.1.1. Visibilité des organes

Si le nématode est mal fixé, et souvent même s'il est bien fixé, certains organes sont pratiquement impossibles à voir et on ne peut au mieux qu'en deviner de vagues traces à l'endroit où ils sont supposés se trouver. L'entrée des données tient alors plus de l'imagination que de l'observation.

Certains organes sont plus visibles que d'autres et le même organe peut être bien visible chez une espèce et pratiquement impossible à voir chez une autre. Un cas extrême est la *phasmide,* une glande chimio-sensible (Fig. 2) qui, dans la majorité des cas, s'ouvre à l'extérieur par un simple pore souvent à la limite de la visibilité. Chez quelques genres (*Scutellonema, Hoplolaimus*) la phasmide est transformée en un *scutellum* à large ouverture, pourvu d'une ampoule sous-cuticulaire visible par transparence.

La visibilité des organes n'est jamais prise en compte par les systèmes d'identification traditionnels qui se fient à l'utilisateur pour savoir observer les caractères nécessaires. Il faudrait pouvoir communiquer au système d'identification ce type de donnée pour qu'il en tienne compte. A vrai dire, la visibilité d'un organe n'est pas une donnée au sens habituel du terme mais une donnée qui va influencer la façon dont les données descriptives vont être utilisées. C'est une donnée au sujet des données, c'est à dire une métadonnée.

1.1.2. Ambiguïté des caractères

L'ambiguïté des caractères est une autre métadonnée. En principe, un organe ayant une faible variabilité sera décrit par des caractères ambigus, mais l'inverse n'est pas toujours vrai. On peut avoir des organes très visibles dont certains caractères sont ambigus. Par exemple, les *spicules*, organes sexuels secondaires des mâles, sont parfois recourbés en épine de rose (Fig. 3). La longueur du spicule peut être mesurée le long du bord dorsal, le long du bord ventral, le long de l'axe, ou bien en ligne droite entre les points extrêmes de l'organe. Chaque façon de mesurer donne bien sûr un résultat différent.

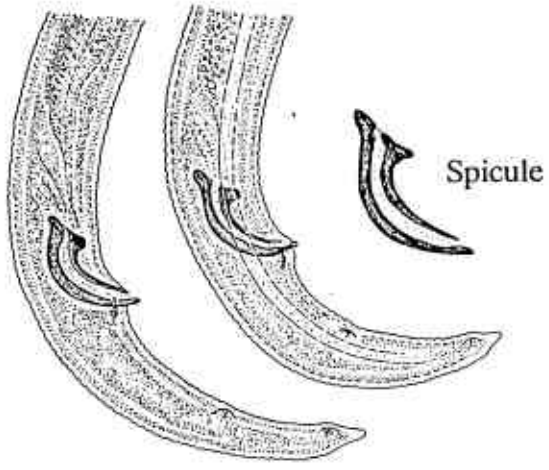

Figure 3: Les spicules, organes copulateurs mâles, chez *Aphelenchoides fragariae.*

Un autre exemple d'ambiguïté est donné par des bourrelets cuticulaires longitudinaux appelés champs latéraux. Le nombre de lignes qui composent ces champs latéraux est un caractère souvent employé pour la définition des genres. Le genre *Trilineellus*, comme son nom l'indique, a été défini par la présence de trois lignes (Lewis & Golden, 1981). En fait, des photos au microscope électronique à balayage (Fig. 4) montrent que son champ latéral est constitué de deux bourrelets et que l'on ne voit que trois lignes parce que les deux lignes centrales sont pressées l'une contre l'autre. Le résultat, trois ou quatre lignes, dépend de l'interprétation choisie par l'observateur.

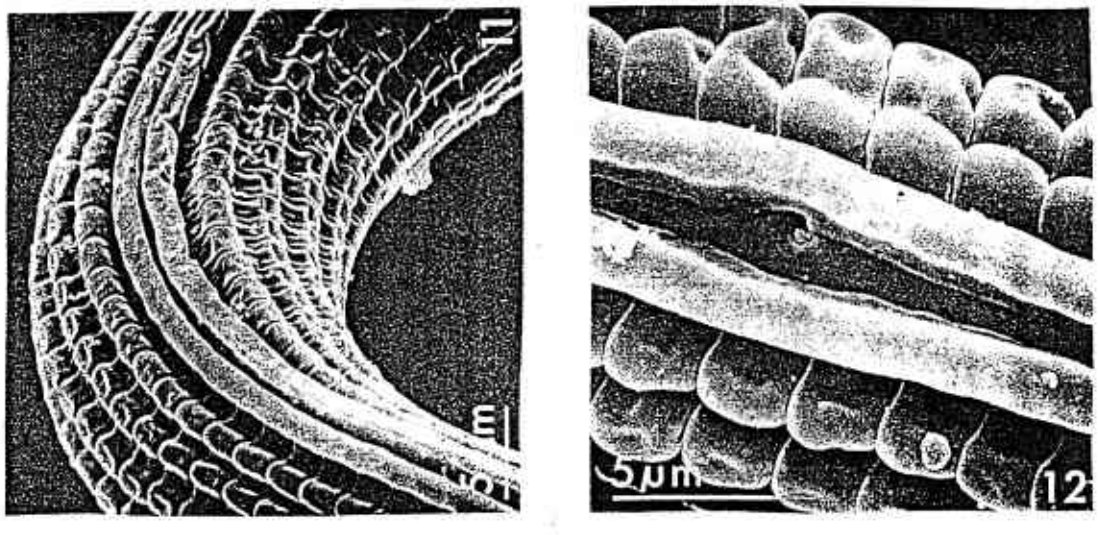

Figure 4: Les champs latéraux de *Trilineellus clathrocutis* vus au microscope électronique à balayage.

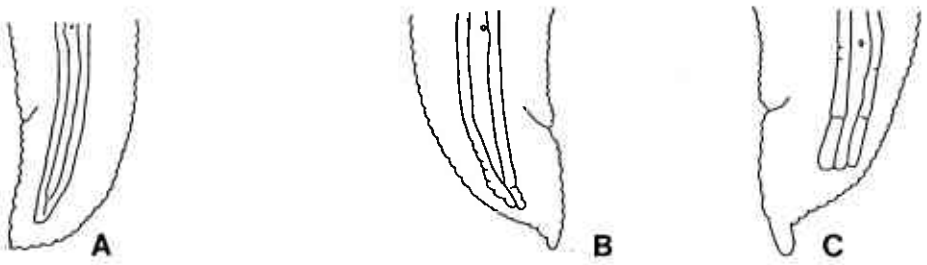

Figure 5: Fusion des lignes internes du champ latéral sur la queue. A: *Helicotylenchus dihystera;* B-C: *H. pseudorobustus.*

1.1.3. Variabilité

Tout varie dans les nématodes, les mesures varient bien sûr, mais aussi les caractères de forme et, pire encore, la variabilité elle même est variable. Par exemple, chez *Helicotylenchus dihystera* les lignes internes du champ latéral fusionnent sur la queue en formant un Y (Fig. 5). Ce type de fusion est présent chez tous les spécimens de l'espèce. Par contre, la variabilité du même caractère est énorme chez l'espèce très proche *H. pseudorobustus*. Dans la population type, 80% des spécimens ont une fusion en U et 20% ont une fusion en Y. D'autres populations ressemblent à *H. dihystera* en ce sens que tous les spécimens on une fusion en Y, et chez d'autres populations enfin, tous les spécimens ont une fusion en U (Fortuner et al., 1981; 1984)

Cette variabilité pose bien sûr un problème pour l'identification des espèces. On serait tenté de rejeter tous les caractères variables, mais l'ennui c'est qu'ils le sont presque tous et il n'en resterait aucun pour différencier les espèces (Fortuner, 1984). En fait, si on connaît bien la variabilité et ses limites, il est possible de se servir d'un caractère aussi variable que la fusion caudale des lignes du champ latéral. Si un spécimen a une fusion en U et si d'autres caractères ont auparavant réduit le choix aux deux espèces en cause, il ne peut s'agir que de *H. pseudorobustus*. La description d'une métadonnée "*variabilité*" permet d'informer le système de cette observation. Le système doit alors être capable d'en tenir compte, par exemple en appliquant une règle ad-hoc.

A noter que la variabilité des caractères morphologiques s'exerce à trois niveaux. Il y a variation d'une population à l'autre dans une espèce donnée, il y a aussi variation d'un individu à l'autre dans une population donnée, et, pour certains caractères, il y a variation d'un endroit du corps à l'autre dans le même spécimen.

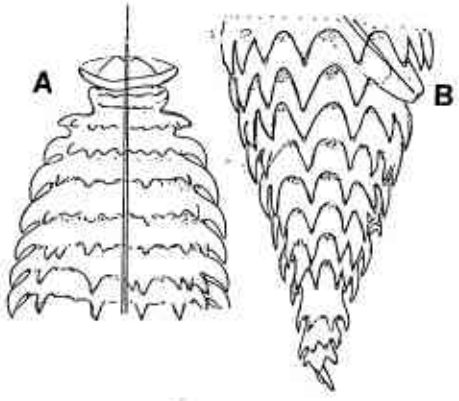

Figure 6: Ornementation cuticulaire chez *Ogma cobbi:* A: extrémité antérieure; B: extrémité postérieure.

Par exemple le diamètre du corps peut prendre différentes valeurs en fonction de l'endroit où on le mesure. Chez certains criconématides, la forme de l'ornementation cuticulaire, un caractère qualitatif, est différente d'une extrémité du corps à l'autre (Fig. 6).

1.1.4. Nature probabiliste des données

De nombreuses méthodes d'identification présument que la nature est déterministe, en ce sens que l'observation d'un caractère suffit pour rejeter une espèce. L'alternative est une vue probabiliste de la nature, selon laquelle l'observation d'un caractère ne donne qu'une certaine probabilité pour le rejet d'une telle espèce. Pour juger de la validité des différentes méthodes d'identification, je crois qu'il est important de décider laquelle de ces deux façons de voir correspond le mieux à la réalité.

Données quantitatives. Si un caractère quantitatif a une moyenne M et un écart-type s chez un taxon T, on peut dire que 95% des spécimens de T tombent dans l'intervalle $M \pm 2s$, ce qui est un énoncé éminemment probabiliste. De nombreuses méthodes de systématique ou d'identification divisent en un petit nombre d'intervalles les caractères quantitatifs, tant nombres réels que nombres entiers. A supposer que l'un de ces intervalles soit pris exactement égal à $M \pm 2s$, 5% des spécimens se trouveront en dehors de l'intervalle défini pour l'espèce à laquelle ils appartiennent. Même un intervalle plus large, égal par exemple à $M \pm 3s$, rejetterait encore 1% des spécimens de l'espèce. A noter qu'un tel intervalle n'aurait plus grand pouvoir discriminant. Par exemple, l'espèce *Helicotylenchus dihystera* mesure 685 μm de long, avec un écart-type de 50 μm. L'intervalle $M \pm 3s$ vaudrait alors 535-835 μm et recouvrirait au moins en partie les intervalles de longueur de la plupart des espèces du genre. Les méthodes qui prennent en compte les mesures en tant que telles (voir § 3.4) ne souffrent pas des mêmes difficultés et sont à préférer.

Données qualitatives. Dans la plupart des descriptions, les caractères qualitatifs sont donnés d'une façon déterministe: *"le taxon T possède l'état S pour le caractère C".* Ce déterminisme apparent n'est parfois que le reflet de l'insuffisance de nos connaissances. Par exemple, le genre *Pratylenchus* a toujours été décrit avec, entre autres caractères, une région antérieure basse, aplatie et des glandes oesophagiennes recouvrant l'intestin sur une courte distance. L'espèce *P. morettoi* a récemment été proposée dans ce genre (Luc et al., 1986) parce qu'elle présente la plupart des autres caractères diagnostiques pour *Pratylenchus* mais *P. morettoi* a une région antérieure en dôme et un long recouvrement intestinal par les glandes oesophagiennes. Depuis la découverte de cette espèce, il n'est plus possible de parler de ces caractères en termes déterministes. Il est préférable de dire est que 99.99% des spécimens de *Pratylenchus* ont les caractères considérés jusqu'alors comme typiques du genre.

§

Visibilité, ambiguïté, variabilité et autres métadonnées, s'appliquent à des caractères qui appartiennent à de nombreux types différents et qui sont liés entre eux par des relations d'une grande complexité, types et relations qu'il faut maintenant décrire.

1.2. La nature des données

Si l'on demande à un systématicien quels sont les types de caractères morphologiques, il répondra sans doute qu'il y en a deux, caractères quantitatifs et caractères qualitatifs. De même, il définira les relations entre caractères sur un plan purement statistique et il parlera d'indépendance et de corrélations. En fait, il est nécessaire de définir bien d'autres types et bien d'autres relations si l'on veut créer un système informatique élaboré, car ces définitions dirigent à la fois la façon dont les caractères sont engrangés dans la base de données et la façon dont ils sont utilisés par le système.

1.2.1. Les types de caractères

Anderberg (cité par Rypka, 1972) a proposé une classification croisée des caractères en fonction d'une part de l'étendue, qui distingue entre caractères continus, discontinus, et binaires, et d'autre part de l'échelle, qui fait la différence entre nombres réels, intervalles, caractères ordinaux et caractères nominaux.

Les types de caractères

	Etendue		
Echelle	Continu	Discontinu	Binaire
Nombre	Longueur du corps	Nombre d'anneaux	?
Intervalle	??	Lignes champ latéral	colonnes de l'utérus (3 ou 4)
Ordinal	Posture du corps	Forme queue (effilée à courte)	stylet court/long
Nominal	Absurde	Forme queue (conoïde ou cylindr.)	présence/absence

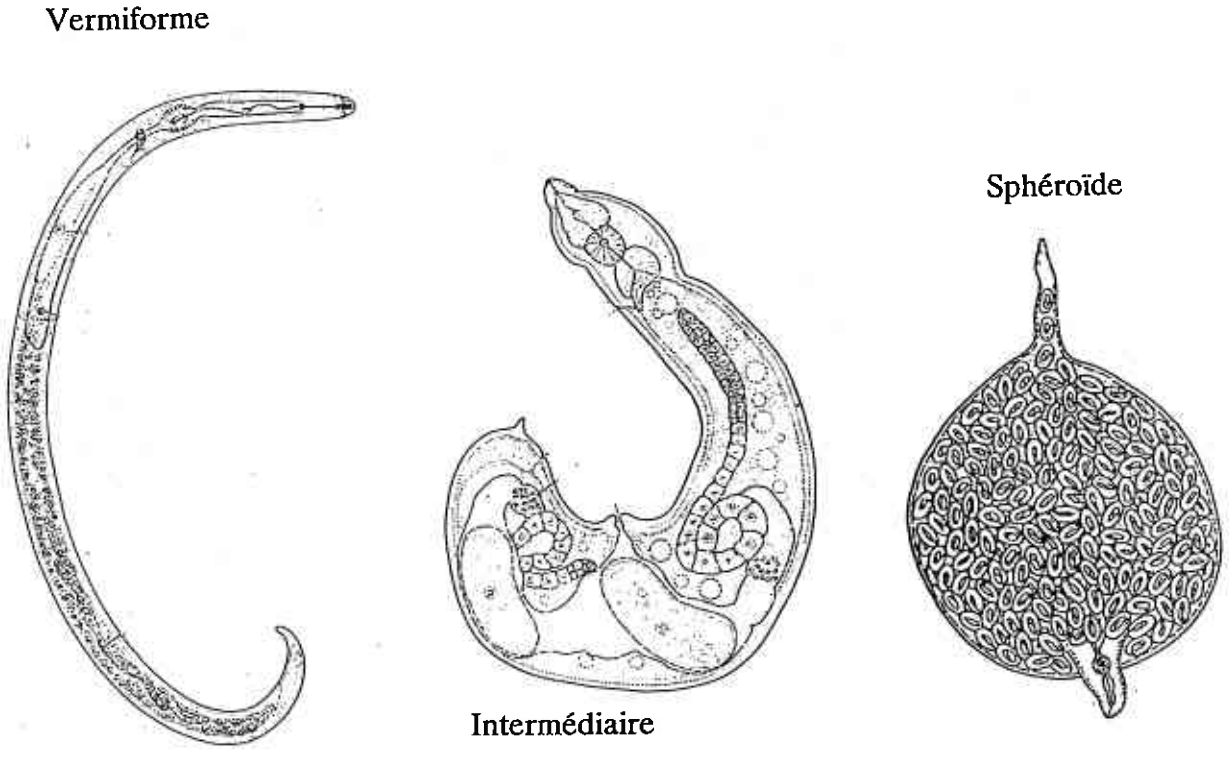

Figure 7: Les trois types de corps chez les nématodes phytoparasites.

L'échelle permet de savoir quelles sont les opérations mathématiques qui peuvent être effectuée sur le caractère. Par exemple si un caractère nominal prend les valeurs *A* et *B* dans deux formes différentes, on peut dire seulement que *A* est différent de *B*. Si le caractère est ordinal, on peut dire de plus que *A* est plus grand que *B*. On peut mesurer cette différence ($A - B = x$ unités) dans le cas d'un intervalle et enfin on peut calculer le rapport *A/B* si l'on a affaire à un nombre. Quand à l'étendue, elle donne une idée de la façon dont le caractère est enregistré.

Certains caractères sont difficiles à classer. Par exemple, le type de corps est soit vermiforme soit sphéroïde, et ce caractère est à première vue nominal/binaire, mais en fait on trouve des formes intermédiaires (Fig. 7). De plus on peut rattacher ce caractère à une expression mathématique (indice α = longueur divisée par diamètre) et dire que le corps est vermiforme tant que ce rapport est supérieur à 3, et qu'il est sphéroïde quand il devient inférieur à 2. Le type de corps est donc en dernière analyse un caractère ordinal/discontinu.

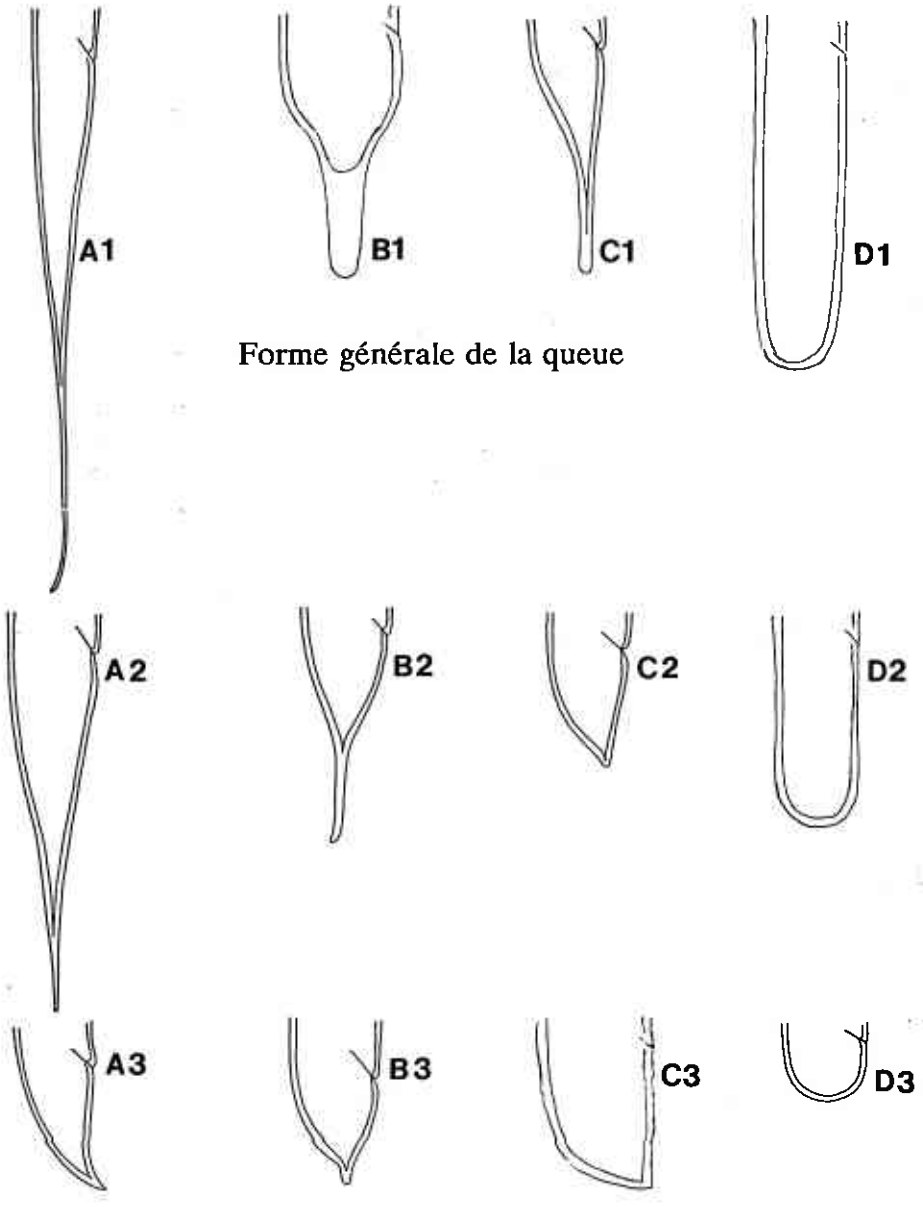

Figure 8: Formes typiques de queues. A-D: Etats nominaux du caractère; 1-3: Etats ordinaux pour chaque état nominal.

Un autre exemple, encore plus délicat à résoudre, concerne certains caractères qui comportent un mélange d'états ordinaux et d'états nominaux. La forme de la queue (Fig. 8) apparait à deux reprises dans le tableau ci-dessus. Il existe quelques types de queues nettement différent les uns des autres (conoïde, cylindroïde, à prolongement axial, à prolongement ventral) qui semblent avoir évolué indépendamment les uns des autres. Ces types fondamentaux représentent donc des états nominaux pour ce caractère. A l'intérieur de chacun de ces types, on observe des variations que l'on peut ordonner à l'aide de l'indice *c'* (longueur de la queue divisée par diamètre) et qui représentent donc une série ordinale.

1.2.2. Les relations entre caractères

Quelque soit le type auquel ils appartiennent, les caractères morphométriques sont souvent liés entre eux par des relations qui peuvent être très complexes. Ces relations sont soit artificielles, créées par la façon dont les auteurs rédigent les descriptions, soit naturelles et liées à la nature des données biologiques. Tout ceci fait que le nombre de caractères qui ont été utilisés par les auteurs est bien supérieur au nombre minimum de caractères qui seraient nécessaires pour complètement décrire une espèce. La description de ces relations permettra de "nettoyer" les données trouvées dans les descriptions publiées et les rendre plus aptes à servir à l'identification spécifique.

Dans la littérature, on trouve un certain nombre de caractères faisant double emploi (caractères redondants) lorsque, par exemple, tel auteur décrit la position du pore excréteur en fonction de celle de l'hémizonide tandis que tel autre décrit la position de l'hémizonide en fonction de celle du pore excréteur. Il est bien évident qu'il sera nécessaire de changer l'un de ces caractères en l'autre pour pouvoir comparer de telles descriptions. La liste des caractères faisant double emploi a été établie pour les nématodes phytoparasites, et les règles permettant de passer de l'un à l'autre sont en cours d'élaboration.

La longueur d'un organe est souvent donnés par sa mesure en micromètres, par exemple *"longueur du stylet = 20,5 μm"*, mais certains auteurs emploient une expression floue telle *"stylet court"*. Pour comparer mesure exacte et caractère flou, il faut utiliser des techniques faisant correspondre les divers états possibles du caractère flou à des valeurs numériques grâce à une fonction mathématique ad-hoc. La définition de telles fonctions est en cours et s'appuie d'une part sur des enquêtes auprès des experts participants au projet et d'autre part sur l'étude des quelques descriptions où les deux approches sont utilisées pour la mesure d'un même organe. Une difficulté supplémentaire vient du fait que la correspondance entre caractère flou et valeur numérique peut varier d'un groupe à l'autre. Un stylet de 20,5 μm sera dit court dans le groupe des hoplolaimides, et long chez les tylenchides.

Les difficultés évoquées ci-dessus sont dues à l'absence de règles à suivre lors de la publication de descriptions d'espèces. D'autres problèmes sont posés par la complexité inhérente aux données morphologiques.

Les caractères sont assumés être indépendants pour l'application de nombreuses méthodes, tant en systématique qu'en identification. Ceci n'est pas toujours le cas. Par exemple, les contraintes mécaniques qui résultent d'une forte élongation du stylet (Fig. 9) exigent que le procorpus soit épais avec un lumen oesophagien replié sur lui-même quand le long stylet est rétracté, et que le bulbe musculaire médian soit fort avec une large valve. Il serait donc faux de considérer ces caractères comme indépendants. Ce genre

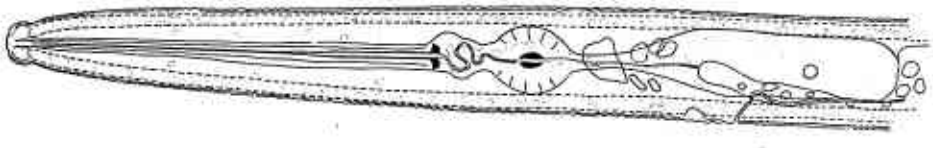

Figure 9: Extrémité antérieure de *Morulaimus arenicolus*.

d'information ne se trouve nulle part consignée par écrit, mais fait partie de l'expertise des systématiciens. Le projet NEMISYS consiste en partie à recueillir ce type de connaissances, les formaliser sous forme de règles et les stocker dans une base de connaissances pour utilisation ultérieure par le système.

Les caractères résumant, eux, sont des caractères dont chaque état possible implique qu'un certain nombre d'autres caractères, qu'on peut appeler les caractères résumés, soient présents sous une forme déterminée. Par exemple, si la phasmide, qui est un organe chimio-sensible, est dite être un *scutellum* (Fig. 2), cela implique qu'elle a une large ouverture et qu'elle est pourvue d'une ampoule sous-cuticulaire, tandis que la phasmide simple s'ouvre à l'extérieur par un simple pore et n'a pas d'ampoule sous-cuticulaire. Il faudra que le système puisse passer du caractère résumant aux caractères résumés et vice versa.

Les caractères diagnostiques pour le genre auquel appartient l'espèce décrite représentent un cas particulier de caractère résumant. Ces caractères sont supposés être présents dans chaque espèce du genre et en général ils ne sont pas inclus dans les descriptions spécifiques. Il faudrait les ajouter aux données de la description mais ceci est compliqué par le fait que la diagnose des genres change continuellement au gré des réviseurs successifs. Il faudra donc stocker les diagnoses génériques qui se sont succédées au cours des ans et il faudra que le système sache quelle est celle qui s'applique à chaque description.

D'autres problèmes concernent non les caractères eux-mêmes mais les états que peuvent prendre ces caractères. Par exemple, certains caractères ne sont utilisables que si un autre caractère prend une certaine valeur. Un exemple simple est donné par tous les caractères décrivant un organe, utilisables seulement si le caractère "*présence de l'organe*" prend la valeur "*organe présent*". Cette dépendance s'observe parfois au niveau des valeurs du caractère. Le corps des nématodes peut être de type vermiforme, intermédiaire ou sphéroïde (Fig. 7). Chaque type présente plusieurs formes possibles et le caractère "*forme du corps = filiforme*" ne peut s'observer que si le corps est de type vermiforme.

1.2.3. Qualificatifs

On trouve souvent dans les descriptions des adjectifs et adverbes attachés aux caractères ou aux états des caractères. Par exemple la posture du corps est *souvent* spirale, *rarement* en forme de C, ou bien la posture est *presque* droite.

Dans le premier cas (*souvent* spirale, *rarement* en forme de C), on trouve une indication assez peu précise de la variabilité du caractère et de la probabilité qu'il y a d'observer chaque état. Pour attacher une probabilité à de tels énoncés j'ai fait une enquête parmi les systématiciens qui participent au projet NEMISYS. Par exemple, d'après les 22 réponses reçues (sur les 58 personnes interrogées), un pourcentage moyen de 76,9% (écart-type = 8,5) peut-être attribué au terme *souvent*. Cela veut dire que l'on peut remplacer l'énoncé "*caractère C souvent présent*" par "*caractère C présent dans 76,9% des spécimens*". On peut aussi aborder cette question comme un problème de donnée floue et dire qu'un caractère décrit comme *souvent présent* est vu dans au moins 59,9% et au plus 93,9% des spécimens ($76,9 \pm 2x8,5$).

Dans le deuxième cas (*presque* droite), c'est l'état observé qui est qualifié: la posture n'est pas tout à fait droite, mais presque. Il peut s'agir, soit d'un nouvel état possible qui vient s'ajouter aux états définis à priori, soit d'une expression synonyme d'un état déjà existant. Les expressions "*corps presque droit*", "*corps faiblement courbé*" et "*corps en forme de C très largement ouvert*" décrivent toutes la même posture. Là encore, l'expertise des participants du projet sera mise à contribution pour décider des synonymies.

La même imprécision peut s'observer pour les données chiffrées et un caractère tel "*nombre de lignes du champ latéral*" peut être une donnée exacte (6 lignes), une étendue de valeurs (5 à 7 lignes), une étendue ouverte (5 lignes ou plus), ou une approximation (une demi-douzaine de lignes).

"*Une demi-douzaine*" est une donnée imprécise, qu'il ne faut pas confondre avec une donnée incertaine dont la valeur ne peut être donnée avec certitude, par exemple "*phasmides apparemment absentes*". L'identification doit souvent s'appuyer sur des données incertaines ou imprécises, et doit s'affranchir des problèmes posés par les données manquantes, dont l'imprécision ou l'incertitude est trop grande pour pouvoir être données par les auteurs. Ceci est vrai à la fois pour les caractères décrivant les spécimens à identifier que ceux qui décrivent les espèces auxquelles sont comparés ces spécimens.

§

Ce court survol de quelques uns des problèmes posés par les données morphologiques montre bien pourquoi l'identification est souvent considérée comme un art. Dans notre effort pour la transformer en une science nous nous sommes heurtés de plus à d'autres contraintes, liées aux circonstances de l'identification: but des études, nature des spécimens, et expertise des identificateurs.

1.3. Circonstances de l'identification

1.3.1. But des études

L'identification des nématodes est faites pour des raisons bien diverses, chacune ayant ses besoins et ses contraintes. Par exemple, les études faunistiques visent à reconnaître quelles sont les espèces présentes dans une région donnée. Ces espèces sont à priori inconnues et il faudra étudier soigneusement les individus prélevés avant de pouvoir établir leur identité. En contrepartie, les nématologistes intéressés accepteront de passer un temps assez long à faire les identifications, puisque la détermination des espèces est justement l'un des buts de ce genre d'étude.

A l'inverse, il y a des cas où l'on sait d'avance ce qu'on va trouver. Par exemple, des prélèvements successifs sont faits lors d'un essai au champ pour suivre l'évolution des populations en fonction des divers traitements. Après le premier prélèvement, les espèces présentes sont parfaitement connues. Il faut alors une méthode de vérification très rapide car l'utilisateur est surtout intéressé par son essai et il refusera de passer plus de quelques secondes à identifier des espèces qu'il connaît déjà.

Dans l'organisme où je travaille, je dois vérifier l'absence de certaines espèces dangereuses dans des échantillons prélevés sur les plantes introduites en Californie. Dans un tel cas il est plus important de s'assurer que certaines formes ne se trouvent pas dans l'échantillon plutôt que de perdre son temps à identifier toutes les espèces présentes. De plus, il est nécessaire de travailler très vite car les importateurs attendent impatiemment le résultat, tout en se gardant de tout risque d'erreur qui pourrait avoir des conséquences économiquement graves.

1.3.2. Origine de l'échantillon

La nature des espèces que l'on risque de trouver dans un échantillon dépend des circonstances dans lesquelles a été fait ce prélèvement, nature de l'hôte, origine géographique mais aussi partie de la plante échantillonnée. Par exemple si on extrait les nématodes de racines nues on ne trouvera en principe que des formes endoparasites. Il faudrait pouvoir communiquer cette information au système et qu'il sache s'en servir pour diriger l'identification vers la réponse la plus probable en fonction des circonstances données. Dans un premier temps, NEMISYS interrogera une base de données où ce genre d'information sera engrangée pour en tirer une simple liste d'espèces cible. Plus tard, le système offrira la possibilité de se servir de ces informations d'une façon plus élaborée, par exemple pour le calcul de probabilités Bayesiennes (voir § 3.5).

1.3.3. Les identificateurs

Situations très diverses là-aussi, depuis les experts complets jusqu'aux complets débutants. Dreyfus & Dreyfus (1986) ont défini cinq degrés d'expertise (novice, débutants avancés, personnes compétentes, connaisseurs et experts). Ils ont montré que les personnes dans chaque catégorie attaquent un problème de façon différente. Par exemple, les débutants avancés et les personnes compétentes font grandement confiance à un plan d'action prédéfini (par eux-mêmes ou par d'autres), tandis que les experts peuvent se libérer des contraintes d'un plan et se fier à leur intuition, fruit de leur connaissance intime du problème. Il est évident que chaque utilisateur suivra une stratégie différente lors d'une session d'identification.

Rares sont les gens qui prétendent être un expert complet. Le cas le plus fréquent est l'identificateur qui connaît un certain nombre de groupes, quelques familles, quelques genres, dans lesquels il est très à l'aise,

mais dont l'expertise est assez limitée ou nulle dans d'autres groupes. Les étudiants et autres novices n'ont aucune expertise du tout. Non seulement ils ne connaissent pas les genres, familles ou espèces ni la morphologie en général, mais aussi ils ne savent même pas comment identifier. Ils sont incapables de choisir un plan d'action. L'aide à l'identification devra leur fournir non seulement les données mais aussi la stratégie à suivre et devra les conduire pas à pas jusqu'à la réponse finale. Le système idéal devra donc être capable de s'adapter aux circonstances de l'identification et au niveau d'expertise de l'utilisateur, pour pouvoir toujours être un système à la fois sûr et rapide. Je montrerai plus bas (voir § 4.7) comment l'agencement des outils d'identification de NEMISYS permet de résoudre ce problème.

§

Après avoir défini le domaine correspondant à l'identification des nématodes phytoparasites, il est maintenant possible de traduire les besoins qui viennent d'être évoqués en un certain nombre d'exigences et de principes généraux qu'il faudra satisfaire pour créer un système d'identification global.

PRINCIPES GÉNÉRAUX ET EXIGENCES

Il est nécessaire, par exemple pour l'étude de la diversité biologique, de disposer d'un système qui permette l'identification de toutes les espèces d'un groupe donné et ne pas se limiter à quelques espèces économiquement importantes. Pour les nématodes phytoparasites, cela veut dire que le système doit permettre l'identification de plusieurs milliers d'espèces. Le système doit aussi être utilisable n'importe où, même dans des régions où la faune est encore inconnue. Enfin le système doit être utilisable par tous, et non seulement par quelques spécialistes. Tout ceci crée des contraintes qui ne se posent pas quand on se limite à un système utilisable seulement par un expert pour identifier quelques espèces importantes dans une région limitée. De plus, nous l'avons vu, le système doit être rapide, simple et fiable.

2.1. Rapidité

La rapidité s'obtient de plusieurs manières. Il faut bien sûr un système installée dans un ordinateur de haut niveau et utilisant un code bien écrit. Il faut de plus que l'approche biologique soit bien conçue. Par exemple, il ne faudrait pas forcer les utilisateurs à répéter ce qu'ils savent déjà. Si un simple coup d'oeil leur donne une bonne idée de l'espèce, du genre ou de la famille auquel appartient le spécimen, il ne faudrait pas qu'ils soient obligés de partir du plus haut niveau, par exemple de l'ordre, et de descendre pas à pas les degrés de la classification pour faire découvrir au système la réponse qu'ils connaissent déjà.

2.1.1. Un raccourci: la notion de promorphe

Les systèmes traditionnels lorsqu'ils ne forcent pas l'utilisateur à ce long parcours, le font en assumant que le genre auquel appartient le spécimen a déjà été reconnu, et proposent une clé pour les espèces du genre en question. La question de l'identification correcte du genre est évoquée ci-dessous (voir § 2.3.1) et une solution proposée (voir § 2.3.2).

Si le genre n'est pas immédiatement et sûrement identifié, les biologistes, qu'ils soient ou non des systématiciens, peuvent souvent reconnaître d'un coup d'oeil un certain nombre de groupes qui ne correspondent pas exactement aux divisions de la classification. Par exemple, le genre *Hoplorhynchus* a été décrit par Andrássy (1985) dans la famille des Hoplolaimidae avec des caractéristiques intermédiaires entre celles de cette famille et ceux du genre *Tylenchorhynchus* (Belonolaimidae). Luc (1986) démontra qu'il s'agissait en fait d'un synonyme du genre *Pratylenchoides* dans la famille Pratylenchidae. Si un observateur tente d'identifier l'espèce utilisée par Andrássy et, faisant la même erreur, utilise une clé des Hoplolaimidae il aura éliminé d'emblée la bonne réponse qui ne peut se trouver que dans une clé des Pratylenchidae. La solution traditionnelle est de partir du niveau de l'ordre des Tylenchida et suivre le long chemin décrit plus haut.

J'ai proposé la notion de promorphe pour résoudre cette difficulté (Fortuner, 1989). Un promorphe est une forme reconnaissable d'un coup d'oeil, avant l'étude détaillée de sa morphologie. Parmi les groupes animaux mieux connus que les nématodes, un promorphe poisson regrouperait par exemple poissons et cétacés. Chez les insectes, le papillon *Sesia apiformis* serait placé dans le promorphe "guêpe" ou le scarabée *Eccoptoptera cupricollis* dans le promorphe "fourmi". Un promorphe n'est pas un taxon, famille ou genre, au sens phylogénétique du terme, mais l'idée que l'on se fait d'un tel groupe. La liste des promorphes que l'on peut définir dans un groupe biologique donné est excessivement floue et dépend surtout de la culture scientifique de chacun. Les non-spécialistes ne reconnaîtront qu'une douzaine de promorphes, représentant les formes les plus typiques des familles les plus courantes. Par contre, le spécialiste d'un groupe pourra reconnaître des genres à l'intérieur de ces familles, et même des groupes d'espèces à l'intérieur de ces genres.

Les promorphes ne constituent pas une classification en ce sens qu'ils ne sont pas arrangés hiérarchiquement et qu'ils sont égaux entre eux, en dépit du fait que certains sont définis de façon plus large que d'autres. Le nématologiste généraliste reconnaîtra les formes juvéniles ($2^{ème}$ stade) de toutes les espèces du genre *Meloidogyne* comme appartenant au promorphe "meloidogynide", tandis qu'un expert du groupe pourra reconnaître les groupes "exigua" et "graminis" (Eisenback, 1989). Une autre différence importante entre le concept de promorphe et la classification Linnéenne est qu'une espèce donnée peut très bien appartenir à deux ou plusieurs promorphes différents (arrangement non exclusif) ou au contraire se trouver en dehors de tous les promorphes reconnus par un observateur donné (arrangement non exhaustif). Un exemple du premier cas est offert par l'espèce utilisée par Andrássy pour décrire le genre *Hoplorhynchus* qui appartient aux promorphes pratylenchide, telotylenchide et hoplolaimide. Le deuxième cas est représenté par toutes les formes qui sont si rarement observées que la plupart des nématologistes sont incapables de les reconnaître au premier coup d'oeil.

La reconnaissance d'un promorphe par un observateur permet de limiter la liste des espèces possibles et cette information est communiquée au système de façon simple par l'entrée d'un seul mot: le nom du promorphe.

§

La rapidité du processus d'identification peut aussi être assurée par le choix d'une méthode convenable pour l'entrée de la description des spécimens à identifier.

2.1.2. Entrée des données

Est à rejeter d'emblée tout système qui force les utilisateurs à entrer les données dans un certain ordre, ou qui exige que toutes les données soient entrée d'un coup. Il faut laisser à l'utilisateur le choix de la nature des données ainsi que de l'ordre et de l'opportunité de leur entrée. Il pourra ainsi entrer un petit nombre de données, les analyser et ne décider d'entrer de nouvelles données que dans la mesure où cela s'avère nécessaire. Cette approche permet à un utilisateur averti d'arriver à une réponse satisfaisante avec le minimum de données. Par contre le débutant risque de ne pas savoir choisir les données les plus efficaces et le système devra dans ce cas être capable de l'aider. NEMISYS calcule l'efficacité des caractères discriminants pour les candidats restant en course et montre à l'utilisateur quel est le pourcentage minimum de candidats qui seraient rejetés par chaque caractère.

Il y a de nombreuses façons d'entrer les données et le choix d'une méthode plutôt qu'une autre dépend des préférences personnelles et aussi de la situation.

Langage écrit. L'entrée écrite des données suit souvent une décomposition hiérarchique des caractères soit par système (système génital, système nerveux, etc.), soit par fonction (par exemple, fonction digestive qui réunit des éléments appartenant aux systèmes digestif, bien sûr, mais aussi musculaire et glandulaire), soit enfin par région du corps (par exemple, tous les organes se trouvant à l'extrémité du corps, quelque soit le système auquel ils appartiennent). Une telle hiérarchie est arrangée selon un format assez rigide, que doit bien connaître l'utilisateur. Ce qui est possible pour une douzaine ou une vingtaine de caractères devient une organisation très lourde pour les 465 caractères déjà reconnus chez les nématodes phytoparasites. Il n'est pas simple de se souvenir de tous ces caractères et de leur place dans la hiérarchie, et il est assez lent de naviguer dans cette hiérarchie pour aller trouver le caractère cherché.

L'idéal serait d'abandonner toute exigence de format et de langage et d'accepter l'entrée des données en langage naturel. Ceci pose immédiatement le problème des différences de terminologie d'un utilisateur à l'autre. Telle forme de queue (Fig. 10) sera décrite par l'un comme courbée dorsalement par l'autre comme convexe-conoïde. Il faut que le système soit capable de d'interpréter ces différences. Le Terminator, un outil créé pour l'extraction semi-automatique des données publiées dans la littérature, a pu être utilisé pour permettre à Nemisys de comprendre les données entrées par l'utilisateur à condition que celui-ci se serve du langage restreint que l'on trouve dans toutes les descriptions morphologiques. Le fonctionnement du Terminator est décrit plus bas (voir § 4.3.2.)

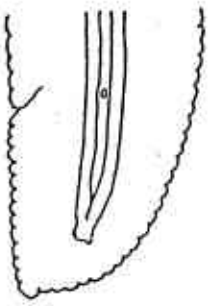

Figure 10: Une forme de queue souvent observée chez *Helicotylenchus dihystera.*

Langage parlé. Les systèmes de reconnaissance de la voix commencent à être disponibles sur le marché. Ils permettront bientôt d'entrer les caractères qualitatifs en les décrivant oralement au cours de l'observation des spécimens sous le microscope. L'utilisation conjointe des méthodes de compréhension du langage naturel et de reconnaissance de la voix permettra d'entrer une description qualitative complète en quelques secondes.

Entrée graphique. Cependant, il est souvent difficile à des gens qui ne sont pas des experts de décrire par des mots les formes qu'ils observent. Il leur serait beaucoup plus facile de comparer leurs spécimens à un tableau graphique des formes possibles et de choisir celle ou celles qui s'en approchent le plus.

Une base de données graphiques pourrait aussi permettre à l'usager de dessiner à petites touches un "portrait-robot" du spécimen comparable à celui employé par la police pour aider les témoins à décrire un criminel. Les caractères attachés aux divers éléments du portrait seraient ensuite transformés en données qui serviraient à l'identification.

Entrée assistée. Les méthodes décrites ci-dessus sont utilisables seulement pour les caractères qualitatifs ou pour les caractères quantitatifs déjà mesurés. Cette prise de mesures est une opération longue et fastidieuse et il sera nécessaire d'adjoindre à NEMISYS un système d'aide à la prise de mesures. Il suffira d'aménager un passage entre NEMISYS et un des nombreux logiciels commerciaux déjà existant.

Reconnaissance automatique des formes. Dans l'avenir il sera possible à l'ordinateur de reconnaître les limites des organes et de procéder aux mesures et évaluations qualitatives sans la moindre intervention de l'usager. La mise au point d'une telle méthode, idéale pour l'entrée des données, demandera encore bien des études, tant du côté informatique que du côté biologie. J'ai accompli le premier pas en faisant la liste de tous les organes (ou parties d'organes) utilisables, et des caractères qui leur sont attachés. Cette liste de caractères dits biologiques devra ensuite être transformée en caractères utilisables par l'ordinateur (caractères calculables). Cette transformation est surtout nécessaire pour les caractères qualitatifs et pourra s'effectuer de plusieurs manières.

Par exemple, l'extrémité antérieure est souvent décrite comme étant basse ou haute, et ce caractère biologique peut être relié à un indice (calculable par l'ordinateur) égal au quotient de la hauteur par la largeur maximum de cette région. L'utilisation de tels quotients (ou rapports) a récemment fait l'objet d'une controverse entre les nématologistes qui rejettent totalement leur usage (Roggen & Asselberg, 1971; Roggen et al., 1987) et ceux qui l'admettent dans certaines circonstances (Fortuner, 1990). En nématologie, il est d'usage de donner des rapports tels le rapport α égal à longueur L divisée par le diamètre D du corps. A première vue, il semble que ces rapports fassent partie d'une fonction mathématique simple, $L = \alpha D$. En fait, la croissance étant allométrique, L et D sont reliée par une expression de la forme $L = \alpha_1 D + b$ où α_1 est différent de α. Comme il a été montré par Roggen et al. (1990) l'expression $\alpha = L/D$ n'est pas valide, quand aux valeurs α_1 et b, elles ne sont jamais calculées par les nématologistes.

Tout en acceptant ces conclusions, j'ai montré qu'il est toujours possible de diviser l'une par l'autre

deux longueurs mesurées avec la même unité chez le même individu. Le résultat de cette division, que j'appelle l'indice α pour le différencier nettement des rapports traditionnels discutés par Roggen, est l'expression chiffrée d'un caractère qualitatif. Par exemple, l'indice α donne une idée de l'aspect général du corps, mince ou épais. Il n'est plus ici question de relier un indice aux longueurs qui lui ont donné naissance mais de décrire un nouveau caractère. En particulier, la valeur moyenne de l'indice dans un échantillon ne permet pas de calculer l'une des longueurs à partir de l'autre. De nombreuses descriptions donnent les valeurs moyennes pour la longueur du corps et l'indice α mais pas pour le diamètre. Il serait faux de dire que le diamètre moyen du corps soit égal à α/L. Il n'en reste pas moins que les indices sont la façon la plus simple de transformer un caractère biologique qualitatif en une caractère calculable.

Un autre moyen serait de transformer les courbes représentant la forme des organes en une expression mathématique obtenue par transformation de Fourier, courbe de Bézier, ou toute autre transformation. L'expression mathématique pourrait ensuite être analysée par une analyse discriminante. Chaque forme caractéristique d'un organe serait représentée par un nuage de points dans un système d'axes. Il serait alors possible d'analyser de la même façon la forme de l'organe chez le spécimen à identifier et de s'adresser aux fonctions discriminantes pour déterminer quelle est la forme caractéristique dont elle est la plus proche.

Dans un premier temps, la reconnaissance automatique des formes utilisera des contours d'organes tracés à la main. Plus tard, il sera peut-être possible de faire reconnaître par l'ordinateur les limites des organes et d'obtenir ainsi ce tracé d'une façon automatique. Ceci demandera des études poussées car l'absence de contraste entre les différents organes rendra difficile la découverte de leurs limites.

2.2. Simplicité

Un système destiné à des usagers qui ne sont pas nécessairement des spécialistes doit être très simple à utiliser. Cette simplicité doit s'exercer au niveau biologique comme au niveau informatique.

Au niveau biologique d'abord, le système doit pouvoir être utilisé par des gens n'ayant que d'assez vagues connaissances du domaine. Il faudra leur apporter de l'aide, non seulement en leur offrant accès aux définitions, par textes ou graphiques, des organes et de leurs caractéristiques, mais aussi en les guidant dans le processus d'identification. Par contre, puisque le système doit aussi pouvoir servir aux spécialistes, cette aide ne doit pas gêner le déroulement des opérations, limiter la liberté des utilisateurs, ou ralentir le système.

Au niveau informatique, le système doit être accessible à des gens n'ayant pas grande habitude des ordinateurs et son mode d'emploi doit être clair, même pour des gens qui ne s'en serviront que rarement. De nombreuses méthodes d'identification par ordinateur se contentent d'une interface malhabile déroulant lignes par ligne un dialogue au formalisme rigide. Un tel système rebuterait vite l'usager qui n'aurait bientôt plus l'envie de s'en servir. Nous avons mis l'interface au premier plan des considérations qui nous guident dans l'élaboration de NEMISYS. Nous verrons plus bas comment le système répond aux exigences parfois contradictoires exposées ici (voir § 3.7 *seq.*)

2.3. Sécurité

2.3.1. Caractères d'identification et caractères phylogénétiques

De nombreux systèmes d'identification traditionnels sont basés sur les groupes définis par les systématiciens quand ils créent leurs classifications. Un système pour l'identification des genres débute par exemple au niveau de l'ordre, et consiste en une série de clés pour l'identification successivement des sous-ordre, super-famille, famille, sous-famille et finalement du genre auquel appartient le spécimen. Une classification est définie (ou devrait être définie) à partir de caractères phylogénétiques, c'est-à-dire de caractères qui trahissent l'existence d'un ancêtre commun à tous les membres de chaque taxon. Ces caractères ne sont pas forcément faciles à voir. Par exemple, une clé proposée par Siddiqi (1986) pour la détermination des sous-ordres de l'ordre des Tylenchida utilise soit des caractères écologiques (cycle vital, mode de parasitisme) qui ne sont pas immédiatement visibles, soit des caractères morphologiques variables. D'une façon générale, il y a peu de chances qu'un caractère phylogénétique soit aussi un caractère facile à observer. Une méthode fiable doit n'utiliser que des caractères faciles et doit donc s'écarter, au moins temporairement, des taxons traditionnels.

2.3.2. Nids d'espèces

Abandonnant temporairement le système Linnéen, j'ai proposé le concept de *"nid d'espèces"* pour désigner un groupement de toutes les espèces qui partagent un même ensemble de caractères primaires d'identification. J'appelle *"caractère primaire d'identification"* un caractère qui s'attache à un organe facile à voir et qui décrit une caractéristique non ambiguë et à la variabilité intra-spécifique soit nulle, soit nettement délimitée dans chacune des espèces du groupe.

Comme les promorphes définis plus haut, les nids d'espèces ne représentent pas une classification parallèle à la classification traditionnelle. Ils ne sont pas arrangés de façon hiérarchique, ils ne sont ni exclusifs ni exhaustifs. Le concept de nid permet de reconnaître un groupe d'espèces sans grand risque de se tromper puisque les nids sont définis à partir de caractères faciles à voir. Bien entendu, l'identification ne s'arrête pas au nid mais continue jusqu'à l'espèce. Une fois l'espèce identifiée, il est possible d'aller voir quelle est sa position dans la classification, par exemple en consultant une base de données de nomenclature et taxonomiques. Ceci permet de donner la réponse selon la nomenclature binomiale habituelle.

2.3.3. Dégradation ménagée du système

Quand on continue l'identification jusqu'au niveau de l'espèce, on est bien obligé d'utiliser des organes peu visibles, des caractères ambigus ou des caractères ayant une forte variabilité intra-spécifique, ce qui rend très réelle la possibilité de faire des erreurs. Il n'est pas possible d'éliminer totalement cette possibilité mais un bon système doit en minimiser les conséquences. Il faut que les performances du système ne se dégradent que lentement quand on fait des erreurs, et que cette dégradation soit proportionnelle au nombre d'erreur commises.

La dégradation lente est une qualité inhérente à certaines méthodes (par exemple coefficient de similarité) et absente chez d'autres (par exemple clé dichotomique). Elle existe chez NEMISYS car le système offre à l'utilisateur la possibilité d'utiliser une méthode à dégradation lente. Même si la méthode choisie est à dégradation brutale, NEMISYS offre à l'utilisateur la possibilité de faire une ou deux erreurs avant que la bonne réponse soit rejetée. Bien que n'étant pas vraiment une dégradation lente, ce procédé retarde le moment où cette propriété deviendrait nécessaire.

Il faut aussi que le système connaisse ses limites et soit capable de se rendre compte qu'il n'a pas assez de données pour conclure. Il y a deux aspects à ce dernier point. D'abord, si l'utilisateur n'a pas fourni suffisamment de données le système attribue à la réponse un coefficient représentant la probabilité qu'elle soit exacte et offre des suggestions pour améliorer ce coefficient. D'autre part, si les données entrées ne correspondent à aucune des espèces de la base de données du système, le système est capable de reconnaître qu'il a affaire à une espèce nouvelle et aide l'utilisateur à la décrire. Les règles qui président à de telles décisions sont en cours d'étude pour NEMISYS.

2.3.4. Evaluation des données et endossement

Alors que la plupart des systèmes d'aide à l'identification font confiance à l'utilisateur et prennent pour argent comptant les informations qu'il leur fournit, NEMISYS juge le risque d'erreur attaché à chaque donnée en fonction d'une douzaine de facteurs liés au matériel optique utilisé (type de matériel, grossissement), aux spécimens eux-même (nombre, état de conservation), au caractère en question (caractère aisé ou difficile), et enfin à l'utilisateur lui-même, son degré d'expertise et ses *Personal Intuitive Feelings* (PIF) c'est à dire la mesure de son degré de confiance dans la donnée qu'il vient d'entrer. Un simple algorithme tire de ces facteurs un coefficient d'endossement qui sera utilisé à plusieurs reprises dans le système, par exemple comme poids w dans le coefficient de similarité NEMAID (voir § 3.4) ou pour attribuer un coefficient d'endossement de la validité de la réponse finale.

LES MÉTHODES DE L'IDENTIFICATION

Les principes généraux et les exigences de l'identification biologique une fois définis, il devient possible de juger les performances et les limitations des principales méthodes d'identification en usage à l'heure actuelle.

3.1. Reconnaissance immédiate

Dans la majorité des actes d'identification, l'expert reconnait une espèce du premier coup d'oeil et se contente ensuite de vérifier rapidement que sa première impression était correcte. Aucune aide à l'identification ne tient compte de cette pratique pourtant si courante et lui offre le support nécessaire. En particulier, si l'expert ne connaît pas par coeur les caractères à vérifier, il doit aller les chercher dans un dossier dans lequel il a engrangé les copies des descriptions de toutes les espèces dont il s'occupe. Chaque laboratoire devrait maintenir une collection complète de tels dossiers, d'où duplication d'efforts et perte de temps et d'argent. Souvent, les dossiers ne sont pas tenus à jour, en particulier dans les pays en voie de développement qui manquent des ressources pour acheter les journaux où sont publiés les descriptions d'espèces. Il est prévu d'associer à NEMISYS une base de données textuelles et graphiques, avec toutes les descriptions publiées dans la littérature. Cette base de données permettra à l'expert de vérifier immédiatement la valeur de sa première impression.

Lorsque l'identificateur ne devine pas tout de suite l'espèce en cause, il a quand même souvent une bonne idée du groupe auquel appartiennent les spécimens, c'est à dire qu'il reconnait un promorphe tel qu'il a été défini plus haut (voir § 2.1.1). Tout nématologiste peut faire la différence entre un tylenchide et un hoplolaimide (Fig. 11), deux promorphes qui correspondent en gros à deux familles différentes. Je connais bien les hoplolaimides et, à l'intérieur de ce promorphe, je peux distinguer d'un coup d'oeil un "scutello" d'un "helico". Un spécimen reconnu d'abord comme scutello peut ensuite se révéler appartenir au genre *Helicotylenchus*. Même si je me trompe, mon erreur elle-même aide à l'identification. Là apparait le problème posé par les méthodes d'identification traditionnelles qui assument que le groupe, famille ou genre, a été correctement identifié et qui ne s'intéressent qu'à la différenciation des espèces à l'intérieur du groupe. Si je prends un *Helicotylenchus* pour un scutello et que j'essaye de l'identifier avec une clé du genre *Scutellonema,* je n'arriverai bien sûr à rien. Le concept de promorphe permet à un système intelligent de savoir que dans un tel cas il ne peut s'agir que de l'une des quelques espèces intermédiaires entre les genres *Helicotylenchus* et *Scutellonema.* Si par exemple le spécimen a été prélevé aux USA, il appartient sans doute à l'espèce *H. vulgaris.*

3.2. Clé dichotomique

C'est la méthode traditionnellement la plus utilisée en nématologie. Ses inconvénients sont bien connus et se placent surtout à deux niveaux. D'une part l'approche dichotomique rend impossible la prise en compte de la variabilité intra-spécifique, d'autre part l'arrangement de la clé, avec des renvois à des lignes qui peuvent être éloignées l'une de l'autre, fait que toute erreur a des conséquences "mortelles". De plus, les clés dichotomiques sont difficiles à mettre à jour.

Le succès des clés dichotomiques auprès des systématiciens tient au fait que les clés sont le moyen le plus rapide d'identifier une espèce parce qu'elles éliminent à chaque ligne de larges fractions du groupe. A condition de ne pas faire d'erreur, une clé bien faite apporte la réponse souhaitée en utilisant le nombre minimum de caractères. Le risque d'erreur est minimisé par le fait que les clés sont le plus souvent utilisées par des identificateurs expérimentés, qui connaissent bien à la fois les clés et les groupes auxquels elles se

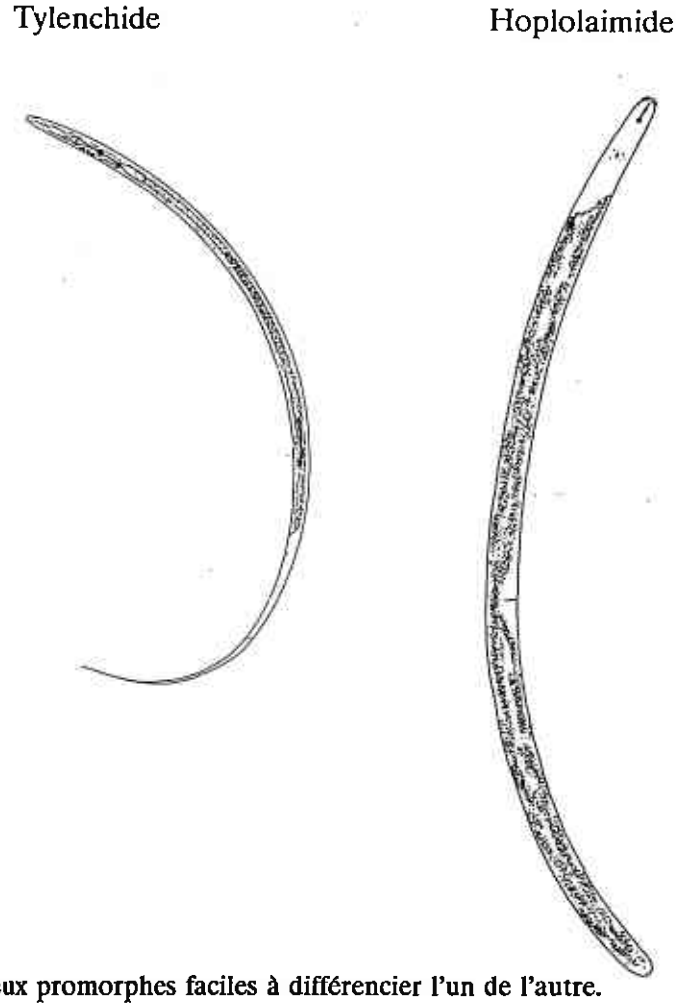

Figure 11: Deux promorphes faciles à différencier l'un de l'autre.

rapportent. Ceci leur permet de suppléer par leur expertise aux déficiences de la méthode. Ile savent quelles sont les lignes dangereuses, celles qui s'appuient sur des caractères difficiles, et ils savent comment surmonter ces difficultés. Par exemple, la première ligne de toutes les clés dichotomiques pour les espèces du genre *Pratylenchus* utilise le caractère *nombre d'anneaux labiaux*, soit deux, soit trois à quatre anneaux. Les spécialistes du genre savent que plus il y a d'anneaux, plus ils sont minces et difficiles à voir. Si donc les anneaux sont bien visibles, cela veut dire qu'il y en a deux, si au contraire on a du mal à les voir, il y en a trois ou quatre (Loof *in* Fortuner, 1989). Il est évident que les débutants, ou même les systématiciens experts pour d'autres groupes que les *Pratylenchus,* ignorent cette astuce et ne peuvent dépasser la première ligne de la clé.

La méthode dichotomique est déterministe et il a été montré plus haut (voir § 1.1.4) que les données morphologiques sont probabilistes. Pourtant, il est possible d'utiliser la rapidité offerte par ce type de clé en prenant un certain nombre de précautions. Dans NEMISYS, la méthode d'élimination déterministe est employée surtout pour sélectionner un nid d'espèce. Selon la définition de ce concept donnée plus haut (voir § 2.3.2), ces groupes d'espèces sont fondés sur des caractères qui sont déterministes, du moins en pratique. Des procédés ad-hoc (endossement des données, élimination d'un nid seulement s'il présente plus de une ou deux différences avec la description du spécimen à identifier) permettent de réduire le risque d'une dégradation brutale des performances du système.

3.3. Clé tabulaire

Les limitations des clés dichotomiques sont en partie surmontées par la clé tabulaire où tous les caractères sont présentés simultanément pour chaque espèce (la clé est polytomique). Une erreur faite sur un caractère peut être rattrapée grâce aux autres caractères. De plus l'utilisateur a une certaine liberté dans le choix des caractères qu'il va utiliser. Alors que dans la clé dichotomique le caractère employé à chaque ligne est choisi une fois pour toute par l'auteur de la clé, l'utilisateur d'une clé tabulaire est libre d'écarter certains des caractères, soit parce qu'il ne leur fait pas confiance, soit parce que qu'il ne peut pas les observer dans le spécimen à identifier. L'arrangement en tableau permet de plus à l'auteur de la clé de noter une certaine variabilité spécifique. La clé tabulaire est facile à mettre à jour lorsque de nouvelles espèces sont décrites dans le genre. Le gros inconvénient de ce type de clé est qu'elles cessent d'être d'un emploi pratique quand elles comportent plus d'une ou deux douzaines d'espèces.

3.4. Coefficient de similarité: NEMAID

Pour les grands genres, l'ordinateur peut aider à faire les comparaisons que l'on ne peut plus faire à vue et pour ce faire, la méthode la plus directe est le calcul d'un coefficient de similarité entre le spécimen à identifier et toutes les espèces du genre auquel il appartient. D'une façon générale, un coefficient de similarité compare chaque couple de caractères chez le spécimen à identifier et chez une espèce du genre et leur attribue un score de 0 (totalement différents) à 1 (parfaitement semblables). Les données manquantes sont neutralisées. L'ordinateur fait la moyenne de toutes les notes obtenues par les caractères décrivant chaque espèce pour obtenir un coefficient entre 0 et 1 qui indique dans quelle mesure l'espèce ressemble au spécimen. On aboutit à une liste des espèces du genre classées d'après leur similarité avec le spécimen.

Gower (1971) a proposé un coefficient de similarité générale S_G utilisable pour trois types de caractères (binaires, à états multiples et quantitatifs). Ce coefficient compare les individus j et k et donne à chaque caractère i un score $0 \leq S_{ijk} \leq 1$. Chaque caractère reçoit aussi un poids w_{ijk}. Le coefficient final est égal à:

$$S_G = \frac{\sum_{i=1}^{n} w_{ijk} s_{ijk}}{\sum_{i=1}^{n} w_{ijk}}$$

Pour les caractères binaires, s_{ijk} est égal à 1 quand les caractères ont le même état chez i et j et à 0 dans le cas contraire. Le même système est utilisé pour chaque état d'un caractère à état multiple, et le nombre d'états possibles n'est pas pris en compte, ce qui introduit un certain biais. Pour les caractères quantitatifs, Gower prend $s_{ijk} = 1 - (|X_{ij} - X_{ik}|)/R_i$ où R_i est l'étendue du caractère i dans le genre considéré.

Il est assez facile de corriger l'importance excessive donnée aux caractères à états multiples et prenant pour score du caractère la moyenne des scores obtenus par chacun de ses états possibles. Plus délicate est la modification de ce coefficient pour la prise en compte de la variabilité intra-spécifique.

§

Depuis 1983 j'ai mis au point un programme, NEMAID, qui modifie en ce sens le coefficient proposé par Gower (1971). Le programme laisse la plus grande liberté possible à l'utilisateur dans la définition des limites de cette variabilité, laquelle est prise en compte de façon différente pour les caractères quantitatifs et qualitatifs.

Caractères quantitatifs. Avec les caractères quantitatifs, il est tentant d'utiliser l'écart-type comme mesure de la variabilité. Si une mesure a une moyenne M et un écart-type e chez une certaine espèce, un individu appartenant à l'espèce a 95% de chances de se trouver dans l'intervalle $X \pm e$. On pourrait penser modifier

l'algorithme de Gower pour accepter comme semblable à l'espèce en question tous les spécimens tombant dans cet intervalle. Cependant, de nombreuses espèces de nématodes ont été décrites de façon succincte, sans que les paramètres statistiques soient donnés pour les mesures.

Plus grave est le fait que les facteurs extérieurs (plante hôte par exemple) font varier les mesures de façon statistiquement significative. Pour le montrer, la progéniture d'une femelle de l'espèce parthénogénétique *Helicotylenchus dihystera* a été élevée sur des hôtes différents. La longueur du stylet est de 24,2 μm (e = o,34) sur riz Moroberekan et de 26,05 μm (e = 0,916) sur poivron Early California Wonder. L'intervalle $X \pm e$ serait de 23,5 - 24,9 μm pour le riz, ce qui éliminerait la plupart des spécimens élevés sur poivron. Les moyennes des échantillons de *H. dihystera* prélevés au champ sont encore plus variables et vont de 23,9 à 27,7 μm (Fortuner, 1987).

Le programme NEMAID prend en compte la variabilité des mesures observée dans la nature en demandant aux experts d'un groupe de proposer un pourcentage de correction p basé sur leur expérience. Les experts forment cette expérience en étudiant en détail quelques une des espèces du groupe, et cette expérience est ensuite extrapolée pour le genre entier. Le pourcentage p sert à calculer une valeur P de correction pour le couple inconnu/espèce prenant en compte la valeur moyenne C_u du caractère dans la population inconnue et la valeur moyenne C_s du même caractère dans une espèce connue. Cette valeur corrective est égale à:

$$P=\frac{C_u+C_s}{2}*p$$

Si R est la différence entre la plus grande et la plus petite moyenne spécifique pour ce caractère dans le genre, la similarité S pour les caractères quantitatifs est égale à:

$$S=1-\frac{|C_u-C_s|-P}{R-P}$$

La différence C_u - C_s n'est prise en compte que lorsqu'elle dépasse la valeur seuil P, qui représente la limite de la variabilité intra-spécifique admissible. La similarité S est prise comme étant égale à 1 quand la différence $|C_u - C_s|$ est plus petite que P. Elle commence à diminuer dans la mesure où la différence dépasse P. Elle serait égale à 0 si l'on comparait la plus petite et la plus grande des espèces du genre ($|C_u - C_s| = R$).

Caractères qualitatifs. Les caractères qualitatifs possèdent deux ou plusieurs états. Les pourcentages de spécimens d'une espèce donnée qui possèdent chaque état sont enregistrés dans la base de données. Par exemple, la fusion caudale des lignes du champ latéral chez la population type de *H. pseudorobustus* est enregistrée comme : fusion en Y: 20%; fusion en U: 80%. Lorsque plusieurs populations ont été décrites pour une espèce, il est possible de calculer deux valeurs, M et A à partir de K_1 et K_2, le pourcentage le plus petit et le pourcentage le plus grand observés pour chaque état dans les populations décrites:

$$M=\frac{K_2+K_1}{2}$$

$$A=\frac{K_2-K_1}{2}$$

M et A sont calculés une fois pour toutes pour chaque espèce et enregistrés dans la base de données. Lors d'une identification, le pourcentage U de spécimens ayant l'état en question est calculé dans la population à identifier et le coefficient de similarité s pour l'état est donné par:

$$s=1-(|U-M|-A)$$

Le même calcul est effectué pour tous les états du caractère. S'il y a n états, la similarité S pour le caractère entier est donnée par:

$$S = \frac{\sum_{k=1}^{k=n} s}{n}$$

La variabilité intra-spécifique du caractère est prise en compte dans la mesure où elle a été enregistrée pour chaque espèce. Quand une espèce est connue par sa seule population type, M est égal à A et les calculs ne prennent pas en compte la variabilité qui existe certainement dans la nature mais qui nous est malheureusement inconnue. Il faudrait qu'un plus grand nombre de populations soit décrit pour chaque espèce connue.

Les similarités S calculées pour chaque caractère reçoivent un poids w proposé par l'expert pour chaque genre et que l'usager du programme peut accepter ou modifier à sa guise. NEMISYS, qui offre l'accès aux algorithmes de NEMAID, calcule automatiquement le poids w en utilisant l'algorithme d'endossement évoqué plus haut (voir § 3.4.4). La moyenne des valeurs Sw pour tous les caractères est calculée pour trouver le coefficient de similarité générale S_G décrit plus haut. Le résultat final est donné sous la forme d'une liste des espèces classées par ordre décroissant de leur similarité avec la population à identifier.

Les performances de NEMAID ne se dégradent que lentement en cas d'erreur. Si par exemple l'utilisateur fait une erreur sur dix caractères utilisés, l'espèce correcte sera encore donnée comme étant 90% semblable au spécimen à identifier. Un autre avantage est la grande liberté donnée à l'utilisateur dans les choix des caractères qu'il veut utiliser, et dans la définition des limites de la variabilité. Un inconvénient majeur est la nécessité d'entrer un nombre relativement élevé de caractères pour bénéficier de ces avantages. Si par exemple on n'utilise que deux caractères, une seule erreur fera que l'espèce correcte ne sera donnée que comme 50% semblable aux spécimens à identifier. D'autre part, la réponse correcte n'est pas nécessairement en tête de liste. Une fois la liste obtenue, il est indispensable que l'utilisateur aille étudier dans la littérature la description complète des espèces les plus semblables et décide quelle est la bonne. Ceci implique qu'il ait l'expertise nécessaire pour faire ce choix en toute connaissance de cause et pour cette raison, NEMAID ne peut pas être utilisé tel quel par les non-experts.

3.5. Méthodes probabilistes: Règle de Bayes

La règle de Bayes est à la base des méthodes d'identification qui toutes en dérivent. D'après cette règle, la probabilité d'avoir le taxon T si le caractère C est présent ($P(T|C)$ est égale à :

$$P(T|C) = P(C|T) \times P(T) / P(C)$$

c'est à dire la probabilité d'observer ce caractère si on a le taxon, $P(C|T)$, multipliée par $P(T)$, la probabilité d'observer ce taxon donnée à priori, c'est-à-dire en l'absence de tout autre renseignement, et divisée par $P(C)$, la probabilité d'observer le caractère en question.

Il est difficile d'estimer la probabilité $P(T)$, qui représente dans l'opinion d'un expert quelle est la probabilité d'observer l'espèce en question dans un prélèvement pris au hasard. Cette opinion se fonde sur les observations, publiées ou non, des peuplements de nématodes observées au champ. Les observations publiées consistent souvent en une simple liste d'espèces trouvées dans certaines circonstances (plante hôte, origine géographique du prélèvement, nature du sol, etc.). Elles comportent parfois les indices suivants:

(1) La *densité absolue* (ou abondance) d'une espèce dans un échantillon, c'est à dire le nombre de spécimens de cette espèce par unité de poids ou de volume de sol ou de racines de cet échantillon;

(2) la *fréquence absolue* (on constance) d'une espèce dans les circonstances étudiées, c'est à dire le pourcentage d'échantillons où cette espèce à été observée dans les circonstances en question;

(3) la *proéminence* d'une espèce qui est sa densité absolue multipliée par la racine carrée de sa fréquence absolue.

La proéminence est une valeur absolue, mais elle peut être rendue relative en la divisant par la somme des proéminences de toutes les espèces présentes dans les circonstances données. Si aucune publication n'est disponible pour des conditions locales données, les nématologistes travaillant dans la région peuvent proposer des probabilités d'observer les différentes espèces présentes.

Dans la formule de Bayes, la probabilité d'observer une espèce, $P(T)$, est supposée être valide "pour l'univers entier". En fait les probabilités sont toujours conditionnées par certaines informations (origine géographique de l'échantillon, plante hôte, et partie de la plante échantillonnée) et ce sont en fait des probabilités conditionnelles. Par exemple, il y a 98,8% de chances d'observer l'espèce *Hirschmanniella oryzae* dans un échantillon de racine prélevé sur riz inondé au Nord Sénégal, mais seulement 1,2% de chances d'observer *H. spinicaudata* et pratiquement aucune chance d'observer les autres 3700 espèces de nématodes phytoparasites (Fortuner & Merny, 1974). En Casamance, au Sud Sénégal, ces probabilités deviennent 20,5% pour *H. oryzae* et 79,5% pour *H. spinicaudata*. Dans la même région, ces deux espèces n'ont pratiquement aucune chance d'apparaître dans un prélèvement fait sur riz de plateau où il y a au contraire 64,8% de chances de trouver *Pratylenchus brachyurus* et 32,2% de chances de voir *P. sefaensis* (Fortuner, 1975). Si l'on me demandait de donner la probabilité d'observer *H. oryzae* valable "pour l'univers entier" je serais bien en peine de décider entre 98,8%, 20,5% ou 0%.

Pour l'utilisation pratique de la Règle de Bayes dans un algorithme d'identification, les probabilités d'observer à priori chaque espèce peuvent être données des trois façons suivantes: 1) plusieurs probabilités pour chaque espèce, chacune étant valable dans des circonstances bien définies; 2) une seule probabilité pour chaque espèce, valable dans toutes les circonstances; et 3) une seule probabilité valable pour toutes les espèces, ce qui revient à admettre notre ignorance de leur répartition.

La première solution est possible dans bien des cas, surtout pour les grandes cultures dans les régions où des laboratoires de nématologie sont installés. On pourra l'utiliser pour les identifications de routine, par exemple le suivi d'un essai au champ où les nématodes présents sont bien connus. Par contre, cette approche est interdite dans les régions ou les plantes hôtes qui n'ont jamais été étudiées, et surtout pour les études faunistiques puisque le but de ces études est justement d'établir la liste et la proéminence des espèces présentes.

Proposer une seule probabilité par espèce, valide sinon pour tout l'univers du moins pour toute la Terre, est un exercice délicat et aux résultats suspects. Il serait difficile de proposer une telle probabilité pour chacune des trois à quatre mille espèces de nématodes phytoparasites décrites. Il faudrait pour cela pouvoir estimer la probabilité d'observer chaque espèce pour chaque situation possible (c'est à dire sur chaque plante hôte possible, dans chaque région possible et pour chaque partie de la plante), et aussi avoir une idée de la probabilité de se trouver dans chacune de ces situations. Par exemple, *H. oryzae* est fréquent dans les racines du riz inondé dans le monde entier mais est presque inconnu sur les autres plantes. En supposant que la probabilité d'observer cette espèce sur riz inondé est de 90% et que les 140.000 hectares cultivés en riz inondé représentent 1% des terres cultivées mondiales, ou 0,00028% de la surface de la planète, la probabilité d'observer *H. oryzae* est de 0,9 fois 0,01 = 0,009 si on se limite aux terres cultivées, 0,00025 pour la planète entière. J'ai pu proposer ces chiffres pour cette espèce bien connue mais il me serait difficile d'en faire autant pour la grosse majorité des espèces décrites.

Prendre toutes les probabilités égales entre elles serait une façon de dire que nous ne savons pas grand chose de leur répartition. Avec 3700 espèces nominales décrites à ce jour, la probabilité d'observer chacune serait égale à 1/3700 = 0,00027. Ce nombre est presque égal à la probabilité d'observer *H. oryzae* dans un prélèvement pris n'importe où sur Terre (0,00025), mais il est bien plus petit que la probabilité trouvée dans le cas où l'on se limite aux terres cultivées (0,009). Avec une probabilité aussi faible, les performances de l'algorithme tiré de la règle de Bayes ne sont pas meilleures que celles de méthodes traditionnelles, telle la clé dichotomique. En supposant que dans la formule donnée ci-dessus $P(C|T)$ soit égal à 1 (chaque caractère utilisé a 100% de chances d'être présent dans l'espèce considérée), que $P(C)$ soit égal à 0.5 (il y a 50% de chances d'observer le caractère en question, en d'autres termes, chaque caractère est présent dans 50% des espèces), et que la probabilité à priori $P(T)$ soit égale à 0,00027, il faudrait une douzaine de caractères pour parvenir à une réponse (probabilité à posteriori égale à 100%). Ce nombre est identique au nombre de caractères qu'il faudrait entrer dans une clé dichotomique pour 3,700 espèces dans les mêmes conditions. Ce résultat est normal puisque la clé dichotomique est simplement un cas particulier de la règle de Bayes dans

lequel toutes les espèces sont considérées avoir la même chance d'être présentes. Si on utilise une probabilité définie dans des circonstances particulières (par exemple rizières du Nord Sénégal, $P(T) = 0.98$), un seul caractère suffit pour arriver au même résultat.

En résumant cette discussion, un algorithme tiré de la règle de Bayes donne de bons résultats lorsque l'on se limite à des conditions d'observations bien définies (l'"univers" consiste en une certaine partie d'une plante connue dans une région connue), mais ses performances sont similaires à celles d'autres méthodes lorsque l'on ne peut pas (ou que l'on ne veut pas) prendre avantage de telles limitations. Il est impossible de proposer des probabilités pour tous les cas de figure possibles car on ne connaît pas toutes les faunes possibles. En particulier les faunes des régions non cultivées sont très peu connues, et un système basé sur l'application directe de la règle Bayesienne serait de peu d'intérêt pour l'étude de la diversité biologique. La même remarque est vrai pour de nombreux pays en développement dont la faune nématologique n'a jamais été prospectée.

3.6. Méthodes statistiques

Depuis Fisher (1936), nous savons que l'analyse discriminante offre une solution analytique à la recherche de la séparation optimale entre deux (ou plusieurs) groupes d'objets préalablement définis. Il est possible d'analyser par cette technique toutes les espèces d'un genre ou d'un nid d'espèces pour définir un espace multidimensionnel. On peut ensuite placer le spécimen à identifier dans cet espace et calculer sa distance aux différentes espèces. J'ai utilisé les analyses discriminantes avec succès pour séparer deux espèces très proches l'une de l'autre, *Hirschmanniella oryzae* et *H. belli* (Fortuner & Maggenti, 1991). J'ai d'abord mesuré la dérive des mesures dans les spécimens types conservés sur lame depuis vingt ans, puis des analyses discriminantes m'ont permis de sélectionner des critères raisonnablement à l'abri des influences extérieures, y compris des déformations subies au cours du stockage. Cette sélection s'est appuyée sur la comparaison des coefficients de structure calculés les uns avec les populations types des deux espèces, les autres avec plusieurs populations appartenant à *H. belli*. Sept caractères ont une corrélation élevée pour la première analyse (ils participent à la séparation des deux espèces), mais une faible corrélation dans la seconde (ils n'ont que peu d'influence sur la séparation des diverses populations de la même espèce) et ils permettent la différenciation de ces espèces.

L'analyse discriminante requiert que toutes les variables soient normalement distribuées dans tous les groupes et que ces groupes aient des matrices de covariance identiques. Il existe des variantes non paramétriques qui peuvent être utilisées quand ces assomptions ne sont pas vérifiées, mais elles sont moins performantes que les méthodes paramétriques classiques. Dans tous les cas, il faut que chaque espèce soit représentée par un échantillon de taille convenable, et que tous les caractères inclus dans l'analyse soient connus pour tous les spécimens de chaque échantillon et pour le spécimen à identifier. Ceci interdit l'utilisation de ce genre de méthodes dans le cas général où nous nous sommes placés, pour lequel les espèces ne sont représentées que par leur description publiée dans la littérature. Les descriptions publiées ne donnent au mieux que les valeurs moyennes des mesures et on ne connaît pas les valeurs individuelles qui ont servi à les calculer. De plus, de nombreux caractères manquent dans chaque description.

A noter que nous pensons proposer NEMbase, la base de données associée à NEMISYS, comme une sorte de journal électronique dans lesquels les auteurs pourront déposer de nouvelles descriptions d'espèces déjà connues. Il sera possible d'entrer séparément la description de chaque spécimen pour chaque population. De plus, les auteurs auront à leur disposition la liste de tous les caractères à enregistrer pour offrir une description complète de leurs spécimens, et on peut espérer voir se réduire le nombre de données manquantes. Tout ceci permettra de constituer peu à peu une matrice de données utilisable pour les méthodes statistiques.

3.7. Réseaux neuronaux

Le dernier cri en matière d'identification est l'utilisation des réseaux neuronaux. Contrairement à l'approche traditionnelle, un réseau neuronal n'a besoin ni de règles, ni d'algorithmes de calcul. Il consiste en un certain nombre d'éléments interconnectés. L'activité de chaque élément est contrôlée par l'activité des éléments auxquels il est relié et par la force ou poids de ces connections. Le réseau "apprend" de façon soit autonome soit dirigée à fournir la réponse qui est attendue de lui (Zerwekh, sous presse).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
TALO-FIBULAR FACET														
1 sloping	8	8	8	8	8	8	0	0	0	0	0	0	0	0
2 steep-sided	0	0	0	0	0	0	8	8	8	8	8	8	8	8
TALO-TIBIAL MEDIAL FACET														
3 broad & extending plantar surface	8	8	8	8	8	8	8	0	0	0	0	0	0	0
4 narrow & elevated above plantar surface	0	0	0	0	0	0	0	8	8	8	8	8	8	8
FLEXOR GROOVE POSITION (posterior talus)														
5 lateral position	8	8	8	8	8	8	0	0	0	0	0	0	0	0
6 midline position	0	0	0	0	0	0	8	8	8	8	8	8	8	8
SUBTALAR JOINT														
7 continuous distal facets w/ long proximal facet	7	7	7	4	7	7	7	7	7	7	0	0	7	7
8 separate distal facets w/ a short proximal facet	0	0	0	0	0	0	0	0	0	0	8	8	0	0
CALCANEO-CUBOID JOINT														
9 moderately deep pivot	7	0	0	7	7	7	0	7	7	7	0	0	0	0
10 deep pivot	0	8	8	0	0	0	0	0	0	0	0	0	8	8
11 shallow pivot	0	0	0	0	0	0	0	0	0	0	8	8	0	0
12 flat surface	0	0	0	0	0	0	8	0	0	0	0	0	0	0
---- etc. -----														
TALONAVICULAR JOINT														
35 laterally rotated & oblong talar head shape	0	0	0	8	8	0	4	7	7	7	0	0	8	8
36 medially rotated & oblong talar head shape	8	0	8	0	0	0	0	0	0	0	0	0	0	0
37 laterally rotated & very round talar head shape	0	0	0	0	0	8	0	0	0	0	0	0	0	0
38 laterally rotated & asymmetrical talar head shape	0	0	0	0	0	0	0	0	0	0	8	8	0	0
39 medially rotated & flattened talar head shape	0	8	0	0	0	0	0	0	0	0	0	0	0	0

A-Cheirogaleidae	B-Lorisidae	C-Galagidae	D-Daubentoniidae
E-Lemuridae	F-Indriidae	G-Tarsiidae	H-Callitrichidae
I-Cebidae	J-Atelidae	K-Cercopithecinae	
L-Colobinae	M-Hylobatidae	N-Pongidae	

Figure 12. Identification de familles de primates par des caractères de l'anatomie du pied. Base de connaissances d'un système neuronal développé par Zerwekh (sous presse).

La figure 12 représente une table pour l'identification de familles de primates basée sur les os du pied (Zerwekh, sous presse). Les données ont trait à 39 caractères pour 14 familles. Une échelle de 0 à 8 mesure l'indication donnée par chaque caractéristique pour chacune des familles considérées. Par exemple, une facette talo-fibulaire en pente est fortement indicative des prosimiens (note 8) et n'est pas indicative du tout des anthropoïdes (note 0). La création d'une telle table pour les 465 caractères décrivant 4000 espèces de nématodes phytoparasites représenterait un travail énorme.

Une fois de plus, les méthodes de pointe donnent des résultats très impressionnants, mais elles exigent des bases de données nouvelles qu'il faut bien recueillir. Ceci est possible pour des petits groupes tels celui de l'exemple ci-dessus, mais c'est humainement impossible pour les larges groupes tel celui qui m'intéresse.

3.8. Les stratégies

Le processus d'identification comprend un certain nombre de fonctions de base, prise puis entrée des données, leur analyse soit pour élimination, soit pour comparaison, vérification d'un résultat préliminaire, flânerie au hasard parmi les descriptions des taxons d'un groupe, focalisation des recherches sur les taxons les plus probables, etc. (Diederich & Milton, sous presse, b). La stratégie de base pour l'identification consiste à entrer des données, les analyser, puis à vérifier la réponse obtenue, mais les stratégies possibles sont innombrables. Par exemple, un expert peut reconnaître d'un coup d'oeil l'espèce en cause et aller d'emblée la vérifier. Si les circonstances sont familières, on peut limiter le nombre des espèces à considérer en focalisant les recherches. Un utilisateur qui, sans être un débutant, n'est pas très à l'aise dans le groupe qu'il essaye d'étudier, va peut-être commencer par flâner un peu au hasard parmi les descriptions des espèces du groupe dans l'espoir de reconnaître en l'une d'elles le spécimen qu'il étudie; il peut ensuite focaliser les recherches sur un petit groupe, entrer quelques données, les analyser, soit pour éliminer les espèces qui ne conviennent pas, soit pour comparer son spécimen aux espèces possibles. Les possibilités sont très nombreuses et on peut parcourir de nombreux chemins au cours d'une session d'identification.

Ceci est vrai surtout pour un expert. Il a été dit plus haut que la différence fondamentale entre experts et débutants est l'utilisation d'un plan d'action. L'expert se fie à son intuition, fruit d'une longue expérience, pour décider de la meilleure marche à suivre dans chaque circonstance. Les novices au contraire doivent suivre un plan d'action bien défini car ils manquent de l'expérience nécessaire pour fabriquer une stratégie sous l'inspiration du moment. Ceci est un défi pour un système comme NEMISYS qui se veut universel car il lui faut aider le débutant sans pour autant gêner l'expert.

Nous proposons de résoudre cette difficulté en offrant un "Plan Visuel" dans la fenêtre des outils de NEMISYS. Les différents panneaux de la fenêtre de l'outil et les boutons d'action sont judicieusement placés pour suggérer une marche à suivre aux utilisateurs occasionnels ou débutants. L'expert, lui, se sert de l'outil comme il l'entend, soit en suivant le plan suggéré, soit au gré de son inspiration. Je montrerai plus bas (voir § 4.2.2) une réalisation pratique de ce principe pour l'outil BASIC ID.

§

Pour conclure, chaque méthode d'identification a ses avantages mais aussi présente des inconvénients et des limitations qui la rendent impropre à être utilisée dans telle ou telle circonstance. NEMISYS transcende ces méthodes en les incluant toutes dans un système qui sera la première pierre d'une station de travail experte pour le nématologiste, système que je vais maintenant présenter un peu plus en détail.

NEMISYS

Depuis 1987, je travaille avec Jim Diederich et Jack Milton à la mise au point de NEMISYS (NEMatode Identification SYStem), une station de travail experte pour l'identification des nématodes. Le projet NEMISYS est à la charnière de deux sciences, informatique et biologie, il est dirigé par trois chercheurs, il dépend en partie de la collaboration de plus de soixante-dix participants et il vise à créer un système très nouveau par bien des aspects. Le coût de sa réalisation se chiffre déjà en dizaines de milliers de dollars. Il est évident que tous ces facteurs se sont traduits par un certain nombre de problèmes dont il a fallu tenir compte pour l'organisation du projet. Avant de décrire les réalisations pratiques de NEMISYS, je veux évoquer quelques unes de ces difficultés et la façon dont nous les avons résolues.

4.1. L'organisation du Projet NEMISYS

4.1.1 Le but du projet

Lors du démarrage du projet, les buts de mes collègues informaticiens étaient bien différents des miens. Je voulais un système utilisable en vraie grandeur dans les conditions définies dans les deux premières parties de ce mémoire: un système global, utilisable par tous dans toutes les circonstances possibles, et qui soit rapide, simple et sûr. J'avais alors des idées assez vagues sur les implications pratiques et théoriques de ce que je demandais et sur les possibilités offertes par l'informatique moderne. Jim et Jack, eux, avaient travaillé sur un système utilisant le langage Smalltalk pour la création de bases de données orientées-objet et ils cherchaient un domaine d'application pour tester leurs idées. Pour eux, la réalisation pratique des outils devait céder le pas aux recherches de base.

Il nous a fallu trouver un terrain d'entente. Plutôt que de mettre au point rapidement un outil simple qui me serve dans mes futures recherches en systématique et en écologie, j'ai fait du développement de l'outil le but de mes activités de recherches et j'ai étudié des questions fondamentales telles la définition de nouveaux algorithmes d'identification, la formulation de nouveaux concepts (promorphes, nids d'espèces), la définition des types de caractères et de leur relations ainsi que leur traduction sous forme de métadonnées ou de règles.

De leur côté, mes collègues ont accepté de consacrer un pourcentage assez important de leur temps à l'écriture du code pour la réalisation pratique des outils. En contrepartie, ces outils leurs ont permis de tester leurs idées en vraie grandeur.

4.1.2. L'attaque des problèmes

Aussi bien pour la définition du projet dans son ensemble que pour la réalisation de ses différentes parties, le principe directeur a été de commencer par définir le but à atteindre de façon aussi complète que possible. J'ai vu bien des gens s'engager dans des recherches sans avoir la moindre idée de ce qu'ils en attendaient. Il me semble évident qu'il faut au contraire bien savoir ce que l'on cherche. Le but peut changer au cours du voyage, des erreurs de parcours peuvent apparaître, mais Christophe Colomb n'aurait jamais découvert l'Amérique s'il n'avait pas un jour décidé d'aller aux Indes!

Ma façon de voir s'accordait très bien avec les préoccupations de mes collègues qui n'étaient pas satisfaits de la méthode traditionnelle de construction d'un système expert selon laquelle le moteur d'inférence et la base de connaissances sont d'abord crées et l'interface est ensuite plaqué tant bien que mal sur le système déjà existant. Eux voulaient inverser le processus et développer l'interface en premier. Ces deux types de considérations se sont mariées très heureusement pour nous faire définir le développement des outils de NEMISYS selon le schéma suivant:

(1) *Définition du but à atteindre.* La création d'un outil commence lorsque je définis une activité que je considère comme importante pour l'identification, telle l'entrée des données, la sélection d'une réponse possible, le calcul d'un coefficient de similarité, etc. J'explique quelles sont les fonctions primaires à inclure et la façon dont elles seront utilisées. Par exemple, l'outil Terminator est destiné à l'extraction des données morphologiques d'un texte publié, et il comprend un certain nombre de fonctions primaires: ouvrir un dossier avec une description à traiter, accepter ou rejeter le choix proposé par l'outil, proposer un choix différent, chercher un caractère, ajouter un organe, un caractère ou un état de caractère aux listes de termes possibles, inclure les qualificatifs rencontrés dans le texte, récapituler les caractères déjà traités, sortir de l'outil, etc. Les conséquences de chaque action sont examinées en détail. Par exemple, doit-on donner à l'opérateur la possibilité d'ajouter de nouveaux termes aux listes utilisées par l'outil à mesure qu'il les découvre dans les descriptions? L'avantage est une mise à jour immédiate de la liste de termes qui est au coeur de l'outil, les inconvénients sont le risque de se tromper et d'ajouter des caractères déjà inclus, ou d'utiliser un mauvais format pour la définition des nouveaux caractères. Ici, nous avons atteint un compromis selon lequel l'opérateur entre ces termes dans une liste provisoire qui est ensuite vérifiée par le responsable de la base de données.

(2) *Dessin de l'outil.* Une fois but et fonctions de l'outil clairement définis, nous essayons d'arranger un ensemble de panneaux, boutons et menus qui apparaîtront sur l'écran de l'ordinateur pour remplir les fonctions ainsi définies. Les textes à traiter, les données entrées, le résultat d'analyses ou de recherches faites dans la base de données, etc., apparaissent dans les panneaux. Boutons et menus permettent à l'utilisateur de piloter l'outil. Il faut atteindre un équilibre entre diverses exigences contradictoires. Par exemple, il faut avoir des panneaux assez nombreux pour pouvoir remplir toutes les fonctions de l'outil tout en étant assez grands pour être utilisables; il faut avoir tous les boutons nécessaires au pilotage, sans risquer de voir l'utilisateur occasionnel se perdre dans les commandes, bref, il est évident que le dessin final ne peut être réussi du premier coup.

(3) *Prototype.* C'est là qu'intervient la flexibilité du langage Smalltalk choisi par mes collègues car, entre autres avantages, il permet la création rapide de prototypes. Nous avons donc le luxe de créer une fenêtre provisoire, l'essayer, voir ce qui ne convient pas et la modifier rapidement.

(4) *Outil fonctionnel.* Ce n'est que lorsque nous sommes tous d'accord sur les fonctions de l'outil et l'aspect de son interface que les mécanismes de fonctionnement sont mis en place. Il reste à faire les essais en vraie grandeur qui entraînent souvent de nouvelles modifications de détail mais les grandes lignes ne changent plus. De plus, la façon dont l'outil a été dessiné, qui met en avant les considérations liées à son utilisation, assure que l'interface sera bien adaptée à sa fonction et facile à utiliser.

Comme on le voit, les considérations biologiques dirigent le processus, ce qui est normal en raison du but du système que nous sommes en train de créer, mais l'informatique facilite la tâche, par exemple en permettant la mise au point rapide de prototypes qui nous permettent de tester nos idées. Ces bénéfices de l'interdisciplinarité se retrouvent dans l'étude des problèmes plus fondamentaux qui apparaissent au cours de nos travaux, comme nous allons maintenant le voir.

4.1.3. Collaboration et interdisciplinarité

Le projet NEMISYS est dirigé par un monstre à deux têtes (informatique et biologie) et trois jambes (Jim, Jack et moi). En plus des conflits qui apparaissent dans toute collaboration en raison de différences de caractères, compliquées dans notre cas par des cultures différentes, française et américaine, notre ignorance réciproque des méthodes et traditions de l'autre science a causé de nombreux problèmes. Chaque science a son propre vocabulaire technique et certains mots ont une signification différente, voir opposée en informatique et en biologie. Par exemple, les informaticiens appellent "classification" ce que les biologistes appellent "identification".

Plus graves auraient pu être les problèmes causés par les différences de méthodes mais nous avons joué le jeu et accepté de considérer soigneusement les points soulevés par les autres. Très vite nous nous sommes rendus compte que ces difficultés avaient un aspect positif, en nous offrant un point de vue inhabituel

sur des questions que nous ne nous posions même plus tant elles semblaient aller de soi. Chacun a dû expliquer sa science, préciser des concepts parfois vagues et les creuser. Ce faisant nous avons obtenu une meilleure idée de ce que nous voulions dire et les nouvelles contraintes soulevées par une discipline inconnue jusque là nous ont forcé à plus de logique, plus de clarté.

Finalement chacun approche les problèmes des autres avec l'enthousiasme d'un néophyte et les résultats sont parfois étonnant. Confronté au problème de l'intelligence artificielle des textes, une des difficultés majeures qui se posent à l'informatique moderne actuelle, j'ai demandé quelle était la raison qui nous interdisait d'utiliser les fonctions d'un logiciel de traitement de textes, et ma question naïve a donné à Jim l'idée de créer le Terminator (voir § 4.3.2).

Les problèmes soulevés par la définition de l'"utilité" (*usefulness*) d'un caractère sont un bon exemple de la façon dont nous organisons les recherches et des avantages de l'interdisciplinarité. Au début du projet, j'avais défini les nids d'espèces comme les groupements des espèces qui partagent le même ensemble de caractères primaires d'identification, et ces caractères primaires comme étant à la fois constant et assez facile à observer dans un groupe d'espèces donné. Certains caractère variables ou difficiles étaient donnés comme "caractères secondaires" s'ils pouvaient aider à l'identification (Fortuner *in* Fortuner, 1989). L'utilité d'un caractère était définie comme une métadonnée proposée par les experts: les caractères utiles étaient ceux qu'un expert regarde en priorité lorsqu'il cherche à identifier un nid (Diederich et al., 1989). Comme on le voit, il y avait une nette circularité dans la définition des nids et des caractères primaires, et le concept d'utilité se basait sur l'opinion des experts.

En janvier 1990, j'ai tenté de redéfinir le concept de l'utilité des caractères dans l'espoir de découvrir des règles ou algorithmes permettant au système de proposer lui-même des nids d'espèces de façon plus objective que les experts. Je me suis très vite heurté à ces problèmes de circularité et de subjectivité. De plus, lorsque j'ai transmis mes observations à Jim et Jack, ceux-ci s'aperçurent que les définitions des caractères primaires et secondaires aboutissaient à une liste non-minimale de caractères primaires, ce qui heurtait leurs habitudes d'informaticiens. Ils m'expliquèrent qu'un système à hautes performances doit utiliser le nombre de données le plus petit possible pour accomplir une tâche et que, puisque les nids pouvaient être définis avec un petit nombre de caractères, ils ne voyaient pas l'utilité d'autres caractères primaires, encore moins des caractères secondaires.

Nos discussion, souvent très vives, durèrent jusqu'à la mi-février. Nous mîmes en question des définitions que nous avions cru acquises et nous mîmes parfois le projet lui-même en question. Finalement, un armistice fut conclu et des définitions provisoires furent proposées pour nous permettre de continuer. L'utilité n'était plus considérée comme une métadonnée et les caractères primaires et secondaires étaient simplement définis comme étant ceux proposés par les experts.

J'ai récemment décidé de rouvrir la question avec un plan d'études rigoureux. Il me faudra d'abord bien expliquer à Jim et Jack la différence entre discrimination d'un objet existant et création d'un tel l'objet par agglomération. L'agglomération procède de bas en haut, d'un niveau inférieur (par exemple des espèces) vers un niveau supérieur (par exemple un nid). La discrimination au contraire va en sens inverse, d'un groupe de nids possible vers le nid qui seul convient. Si la discrimination s'accommode d'un nombre minimal de caractères et cesse dès qu'un seul nid reste en jeu, l'agglomération doit tenir compte de tous les caractères communs aux espèces réunies dans le même nid. Le processus aboutit à une liste de caractères qui est plus que minimale et qui comprend à la fois des caractères primaires et des caractères secondaires. Ceci est un avantage pour la vérification de la réponse fournie par discrimination. Ayant trouvé le seul nid possible parmi les nids existant à l'aide de la liste minimale de caractères primaires, les autres caractères primaires et les caractères secondaires permettent de s'assurer que ce nid est le bon et que le spécimen inconnu n'appartient pas en fait à un nouveau nid non encore décrit.

Une fois ces points acquis, il deviendra possible de proposer une méthode pour la sélection des caractères primaires et des nids. Une agglomération directe est impossible car bien des caractères sont manquants dans la description de chaque espèce (ce qui interdit d'utiliser la procédure *"cluster analysis"* des logiciels SAS par exemple). Par contre, NEMAID (voir 3.4) fonctionne même en l'absence d'une partie des caractères et il peut donc être utilisé pour créer une matrice de similarité pour toutes les espèces considérées entre elles. L'agglomération peut alors être faite à partir de cette matrice.

Cet exemple montre que je suis à l'origine de la définition des problèmes qui se posent sur le plan biologique, parfois à la suite de questions posées par mes collègues. S'ils n'avaient pas mis l'accent sur la nécessité d'avoir une liste minimale de caractères, je n'aurais sans doute pas cherché à mieux définir les concepts employés. Eux agissent en examinant les solutions que je propose à la lumière des théories et pratiques de l'informatique, et nous travaillons tous à trouver un terrain d'entente.

4.1.4. Communications

Bien entendu notre collaboration n'est possible que grâce à des communications très serrées. Nous nous rencontrons chaque semaine à Davis et nous échangeons des messages électroniques plusieurs fois par jour. Ceci pose des problèmes techniques (la transmission de dossiers entre mon IBM AT et leurs Sun et Macintosh en passant par le système Unix de l'université de Davis est un vrai casse-tête). Les communications à distance exacerbent les conflits qui apparaissent parfois entre nous. Le message écrit est un message mort, figé, qu'on ne peut pas ajuster au vu des réactions immédiates du partenaire. Nous avons failli plusieurs fois rompre notre collaboration jusqu'au jour où nous nous sommes rendu compte qu'il suffisait de se rencontrer en personne pour que, comme par magie, les échanges acrimonieux fassent place à la recherche d'un terrain d'entente. Il est possible de collaborer à distance mais pour débattre de questions fondamentales, le contact personnel est indispensable, soit réel, soit par téléconférences télévisées qui en donnent l'illusion.

4.1.5. Le *"NEMISYS International Project"* (NIP)

J'ai décidé dès le début du projet NEMISYS de faire appel aux systématiciens nématologistes pour deux raisons. L'expertise dans ce domaine est très distribuée et bien peu peuvent se dire experts complets pour tous les groupes de nématodes. Ce n'est certainement pas mon cas, et j'ai donc besoin de l'aide de mes collègues pour les groupes qui ne me sont pas familiers. En 1988, le projet a pu démarré parce j'ai invité une trentaine de personnes à un atelier de travail aux USA grâce à une subvention de l'OTAN. Cet atelier m'a donné l'occasion d'attirer les meilleurs experts mondiaux et de réunir ainsi un bon groupe de soutien. Sur la lancée de l'atelier j'ai pu recruter une quarantaine d'autres collaborateurs.

A cette époque, l'un des participants s'est targué de contacts personnels avec le directeur d'une organisation internationale pour faire miroiter l'espoir d'une grosse subvention. En contrepartie il demandait un titre qui lui fasse honneur et lui donne de l'importance vis-à-vis de ses collègues. Alléché par son offre, j'ai décidé de couper le globe en régions, chaque région ayant un directeur (voici notre homme casé) et chaque directeur s'occupant de recruter collaborateurs et futurs utilisateurs de NEMISYS. Je dois avouer que j'ai essuyé un échec total. Non seulement mon vantard s'est montré incapable de remplir ses promesses, non seulement lui et les autres directeurs n'ont pas trouvé un seul nouveau participant, mais de plus cette organisation régionale, mise en place au cours de l'année 1989 si riche en événements historiques, m'a causé toutes sortes d'ennuis. Les chercheurs que j'ai contacté dans les pays de l'est se sont indignés d'avoir été mis dans la même région que l'URSS! Je ne regrette pas d'avoir essayé car le jeu en valait la chandelle. La réussite nous aurait assuré un financement très confortable. De plus la rédaction de la demande de subvention nous a forcé de coucher sur papier bien des aspects du projet qui resserviront pour d'autres requêtes. Quoi qu'il en soit, l'organisation en régions a été mise en sommeil et le NIP continue seulement avec des groupes d'experts, un par promorphe.

Là-aussi, les résultats ont été décevants, un peu par la faute des circonstances. Il était nécessaire de créer rapidement un prototype de NEMISYS avec une ou deux douzaines de nids, pour faire démonstrations et tests. J'ai demandé aux collaborateurs de décrire autant de nids qu'ils pouvaient. La réponse a été assez bonne et sept collaborateurs ont décrit quatorze nids. L'ennui est que ces nids appartiennent à divers promorphes et ne permettent donc pas des comparaisons fines. Pour faciliter les essais, j'ai dû décrire moi-même une vingtaine de nids dans le seul promorphe hoplolaimide et j'ai demandé aux participants du NIP de cesser de décrire de nouveaux nids. A l'heure actuelle, un seul groupe est actif et m'a fourni les descriptions des nids d'espèces du promorphe dolibel (dolichodorides/bélonolaimides). Le même groupe a maintenant commencé l'étude du promorphe pratylenchide. Je compte relancer bientôt les autres participants pour terminer rapidement cet aspect du projet.

Le NIP survit grâce au *NEMISYS International Project Update*, un bulletin que j'envoie aux participants pour les informer de l'état d'avancement des recherches et pour recueillir leur opinion sur certains points. Environ la moitié des personnes interrogées répondent à mes questionnaires. J'ai mis en place un réseau électronique qui devrait permettre d'accélérer les communications.

Il sera difficile de maintenir l'intérêt de collaborateurs dont les préoccupations sont parfois bien éloignées de l'identification. Pour recueillir leur expertise il faudrait pouvoir les inviter à un nouvel atelier et les faire travailler sur une liste de questions soigneusement définies à l'avance. J'ai maintenant une bonne expérience de l'organisation de tels ateliers car en plus de l'atelier financé par l'OTAN en 1988 j'ai organisé en 1990 l'atelier ARTISYST, sur fonds de la *National Science Foundation* (NSF), pour le survol des méthodes informatiques modernes utilisables pour la systématique en général.

4.1.6. Financement du projet

J'ai parlé à plusieurs reprises des considérations monétaires et il est vrai qu'elles ont souvent décidé de l'orientation à donner à nos activités.

Aux USA, les dépenses de fonctionnement assurées par l'organisation où l'on travaille sont ridiculement faible, deux ou trois milliers de dollars par an dans les meilleurs cas, pratiquement rien dans les circonstances où nous nous trouvons, mes deux collègues et moi-même. Notre seul espoir est d'obtenir un financement extérieur, par exemple auprès de la NSF. La rédaction d'une demande de subvention demande deux à trois mois de travail et, avant d'entreprendre un tel exercice, il faut s'assurer que les chances de succès sont bonnes, ou du moins raisonnables, ou bien que le texte pourra être réemployé pour d'autres demandes ou pour un article.

Le projet doit bien cadrer avec les préoccupations de l'organisme auquel on demande une subvention. Par exemple, en nématologie, le *U.S. Department of Agriculture* (USDA) subventionne uniquement les études portant sur l'étude du stress causé aux plantes par l'attaque des parasites. J'ai bien essayé d'expliquer que NEMISYS permettait de déterminer quel était le nématode causant le stress et que donc nos projets méritaient d'être subventionnés en dépit des restriction de l'USDA, mais l'argument était un peu trop spécieux et il n'a pas été accepté.

La section systématique de la NSF a elle-aussi rejeté nos premières demandes car le jury de sélection a jugé l'identification comme une activité de service indigne d'un financement NSF, organisme dévoué à la science "pure". Cette attitude a porté un rude coup à nos projets car la NSF est la principale source de financement de la recherche aux USA. En désespoir de cause, nous avons décidé d'essayer de changer la politique de la NSF! Nous avons obtenu une subvention de cet organisme pour organiser un atelier de travail destiné à passer en revue les possibilités offertes à la systématique par les méthodes informatiques modernes. Nous avons réussi à faire inclure dans la liste des questions à débattre les thèmes qui nous intéressent (identification, bases de données et stations de travail expertes). Utilisant au mieux notre position d'organisateurs de l'atelier, nous avons fait admettre aux participants que l'identification est une activité scientifique à part entière et que la NSF devrait financer les études qui s'y rapportent, ainsi que la création de bases de données sur des sujets pratiques (Fortuner, 1992). Armé de ces conclusions, je suis plus optimiste des chances de succès de ma prochaine demande.

Il est parfois nécessaire de s'écarter un moment du but à atteindre pour mieux obtenir les fonds qui permettront d'y parvenir. Par exemple, j'ai accepté une subvention du CDFA pour l'étude d'un point d'un intérêt certain mais très limité (différentiation de *H. oryzae* et de *H. belli*, voir § 3.6). Cette subvention m'a permis d'acheter un système de prise informatisée des données. Une stratégie semblable vient de me rapporter une subvention de 20.000 dollars, grâce à laquelle je vais pourvoir commencer de peupler NEMbase, la base de données attachée à NEMISYS.

Tout ceci m'a donné une bonne expérience de directeur de projet et m'a appris comment se diriger dans le maquis des règles administratives. Mes efforts ont parfois été couronnés de succès et mes collègues informaticiens eux-aussi ont pu trouver des fonds pour faire avancer le projet et construire un prototype de NEMISYS. Ce prototype est loin de représenter le système complet, bien sûr, mais il nous a permis de tester nos conceptions et de les démontrer au cours de nombreuses conférences scientifiques. Je vais maintenant présenter quelques aspects du système, surtout pour donner une idée de ses capacités présentes et potentielles.

4.2. Le système NEMISYS

Je ne veux donner dans ce mémoire donne qu'une description très brève de la structure du système NEMISYS. Seuls sont décrits l'interface et les principes et concepts qui gouvernent sa conception sur le plan biologique. La partie informatique du projet est sous la direction de mes collègues Jim Diederich et Jack Milton. Dans un mémoire qui veut démontrer mes aptitudes à diriger des recherches, il m'a semblé qu'il ne m'appartenait pas de parler de recherches dirigées par autrui!

Ma collaboration avec mes deux collègues est bien structurée. Je décris les principes et concepts de l'identification biologique, et je les traduis en un ensemble de fonctions et d'outils informatisables. Nous travaillons tous les trois à la définition de l'interface qui doit satisfaire à la fois aux exigences des informaticiens qui le construisent et à celles des biologistes qui l'utiliseront. Mes collègues s'occupent seuls de la partie informatique. Je me contente de recevoir les résultats de leurs efforts et de vérifier qu'ils correspondent bien à mon attente.

A la différence des approches traditionnelles, basées sur une méthodes d'identification bien définie, NEMISYS est un ensemble d'outils qui veut intégrer toutes les méthodes. Chaque outil accomplit une tâche parmi toutes les tâches possibles liées au processus d'identification. Je peux décrire le fonctionnement de certains outils, je ne peux pas décrire celui du système entier car il dépend de chaque utilisateur, et de chaque utilisation. On peut décrire chacun des outils d'un charpentier, le marteau, la scie, le ciseau etc., mais l'utilisation de ces outils dépend du charpentier et de ce qu'il veut faire. L'artisan qui réalise une salle à manger en chêne sculpté se servira des mêmes outils que le bricoleur qui installe une étagère pour les jouets des enfants, mais les utilisera de manière bien différente. NEMISYS est un ensemble d'outils dont le fonctionnement dépend de l'utilisateur qui lui apporte son expérience.

4.2.1. Concept d'un ensemble d'outils

Il est peu probable que l'on arrive un jour à définir une méthode d'identification unique qui soit la meilleure approche possible dans toutes les circonstances possibles. En fonction de l'utilisateur, du matériel à identifier, du but de l'identification, et surtout des différentes phases d'une même session d'identification, l'une ou l'autre des méthodes existantes sera optimale. Les paramètres présidant au choix d'une méthode sont si changeants qu'il serait illusoire de vouloir les définir à priori. Il est préférable d'offrir une série d'outils à l'utilisateur, chaque outil ayant une fonction différente et utilisant des méthodes différentes, et lui laisser l'entière liberté de choisir l'outil qui convient le mieux à chaque instant. C'est le concept de *station de travail experte* qui dit à l'utilisateur: "*Fais ce que tu veux, quand tu le veux*".

NEMISYS utilise le langage Smalltalk-80. Un des avantages de ce langage est qu'il est sans "modes". Le programmeur peut corriger une erreur, tester sa correction puis l'insérer dans le programme principal sans avoir jamais à passer d'un mode à l'autre. Dans le même ordre d'idées, l'utilisateur de NEMISYS pourra passer d'un outil à l'autre en conservant les données et les résultats partiels déjà acquis et sans avoir à les copier.

Dans une station de travail experte, un outil permet d'accomplir une fonction importante pour l'application en question, ici l'identification. Un outil doit pouvoir conserver son utilité dans les versions successives du système. Un outil peut être utilisé conjointement avec d'autres outils, et son utilisation peut varier au gré de l'imagination de l'utilisateur. Finalement, un outil n'est pas complet en soi. Il n'accomplit qu'une partie du travail nécessaire et il ne fonctionne qu'en présence d'un utilisateur et conjointement avec d'autres outils. Un outil ne remplace pas un expert mais il lui permet d'accomplir sa tache plus rapidement, surement et aisément (Diederich & Milton, sous presse, a). Quelques exemples d'outils NEMISYS permettront de mieux faire voir comment fonctionne ces outils.

4.2.2. Exemples d'outils de NEMISYS

"*DATA ENTRY*" Cet outil, qui n'a pas encore été construit, se composera d'une fenêtre avec quatre panneaux principaux. Un panneau permettra l'entrée des observations en langage naturel, sans aucune contrainte. Un bouton activera un microphone et un système de reconnaissance de la voix qui évitera à l'observateur d'avoir

à taper ses observations au clavier de l'ordinateur. Il lui suffira de décrire à haute voix ce qu'il voit dans son microscope. Un autre bouton permettra de demander au système de comparer les données entrées à la liste des caractères et de leurs valeurs constituant la base de données NEMbase. L'utilisateur pourra accepter ou rejeter chaque caractère. Si le système ne reconnait pas l'une des données entrée, l'observateur pourra souligner un mot ou une expression du texte et les rechercher dans la décomposition hiérarchique par organe, partie d'organe et caractère qui apparaîtra dans un autre panneau. Cette décomposition pourra être vue par système (système musculaire, système digestif, etc.), par fonction (par exemple, la fonction digestive comprend aussi les muscles attachés aux organes digestifs) ou par région du corps. Un troisième panneau permettra de voir le dessin d'un organe ou caractère souligné dans l'un des deux panneaux supérieurs. Il sera aussi possible de voir apparaître une définition textuelle du même caractère ou organe, ainsi qu'un résumé de toutes les données entrées.

D'autres méthodes d'entrée des données seront proposées dans de futures versions de NEMISYS, y compris la prise de données assistée par ordinateur. Ces méthodes apparaîtront soit dans l'outil DATA ENTRY, soit dans d'autres outils spécialisés.

"*BASIC ID*" L'un des outils de base de NEMISYS, le BASIC ID, est destiné aux identification de routine par un utilisateur de compétence moyenne, pas assez expert pour pouvoir deviner directement quel est la bonne réponse, mais ayant pourtant suffisamment d'expérience pour ne pas avoir à être guidé pas à pas par le système.

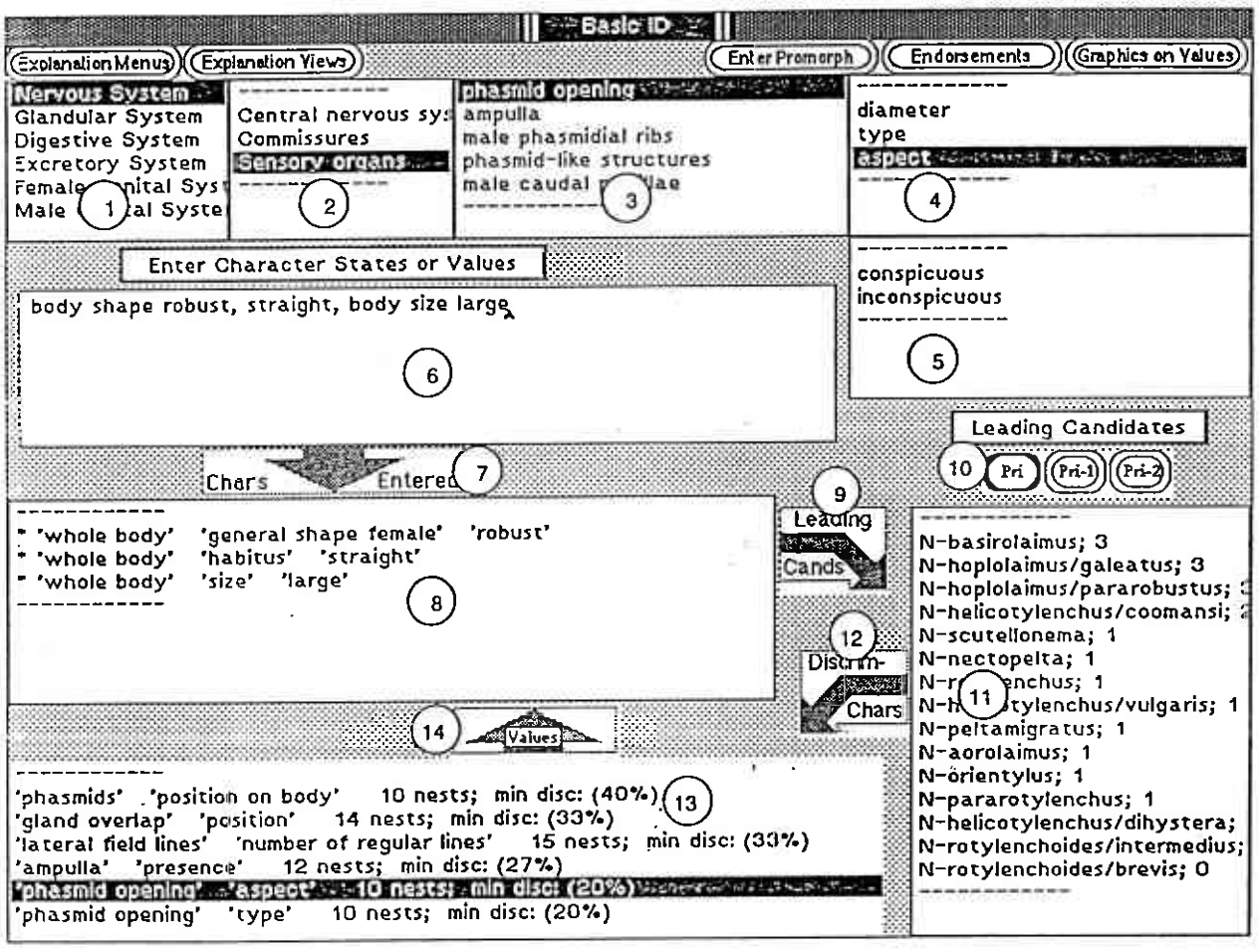

Figure 13: Un outil de NEMISYS: le BASIC ID.

La fenêtre de cet outil (Fig. 13) comprend plusieurs panneaux et des boutons. Les panneaux 1 à 6 permettent l'entrée rapide des données sans passer par l'outil spécialisé dans cette fonction (DATA ENTRY).

Quand l'utilisateur presse le bouton 7 (par l'intermédiaire de la souris), la liste des données entrées apparait dans le panneau 8. Les panneaux 8, 11 et 13, et les boutons 9, 12 et 14 forment la *boucle d'identification* qui est le coeur de l'outil. Cette boucle est un exemple de Plan Visuel qui illustre la stratégie la plus couramment suivie pour une identification de routine: l'utilisateur accepte les données du panneau 8, presse le bouton 9 pour voir une liste de candidats possibles en fonction de ces données apparaître dans le panneau 11, presse le bouton 12 pour obtenir dans le panneau 13 la liste des caractères qui permettraient de séparer ces formes, et il retourne enfin aux premiers panneaux pour entrer de nouvelles données avant de recommencer la boucle. La boucle d'identification aide l'utilisateur en lui suggérant la marche à suivre, mais elle le laisse libre de sauter d'un panneau à l'autre, ou bien d'aller utiliser les fonctions offertes par d'autres outils.

La méthode d'identification utilisée dans le Basic ID est déterministe dans son principe. Les données entrées par l'utilisateur éliminent les groupes d'espèces (nids) qui ne leur correspondent pas. Pour éviter les erreurs, ces nids sont définis à partir de *caractères primaires d'identification*, c'est à dire des caractères qui sont si facile à observer dans un groupe donné que même un débutant a peu de risques de s'y tromper (voir § 2.3.2). L'utilisateur prudent peut aussi s'octroyer la possibilité de faire une ou deux erreurs avant que le bon nid d'espèce soit rejeté.

"SHOW ME" Il arrive souvent qu'un expert croit avoir reconnu les spécimens à identifier et qu'il veuille comparer leur description à celle d'une espèce ou de tout autre taxon. L'outil SHOW ME va lui montrer la diagnose de la forme sélectionnée. L'outil se compose de trois panneaux principaux. Le premier panneau permet de sélectionner un promorphe, un nid, une espèce ou tout autre taxon. Le deuxième panneau montre un dessin de la forme sélectionnée tandis que le dernier panneau donne une description textuelle de ses caractères diagnostiques. Il sera possible de noter quels sont les caractères qui correspondent bien à ce que l'on observe dans les spécimens à identifier.

Si l'on craint d'être influencé en voyant sur l'écran la "bonne" réponse, on peut faire disparaître les valeurs des caractères dans la forme sélectionnée. Il faut alors entrer les valeurs de chaque caractère observés chez le spécimen inconnu. Un coefficient de similarité est alors calculé pour chaque caractère et pour le nid ou l'espèce sélectionnés.

A noter que les données entrées par n'importe quel outil deviennent partie intégrale de la description de spécimens, au même titre que celles entrées par l'outil spécialisé DATA ENTRY.

4.3. La base de données NEMbase

4.3.1. Etude des données publiées

Les aides à l'identification de tel ou tel taxon sont généralement proposées à partir de données fraîches, réunies par l'auteur de la révision. Cela permet de partir d'une matrice de données complète (pas de données manquantes), cohérente (toutes les données sont entrées selon le même format) et de qualité uniforme (pas de variabilité introduite par l'observateur). En contrepartie de ces avantages, chaque système ne peut raisonnablement porter que sur un petit nombre de taxons. Il y a environ 3700 espèces dans le seul groupe des nématodes phytoparasites et leur redescription à partir de données fraîches est impossible en pratique.

L'alternative est de se servir de toutes les descriptions d'espèces qui ont été publiées dans la littérature spécialisée. Cette solution a deux inconvénients majeurs. D'abord la qualité des données publiées est très variable d'une description à l'autre en fonction de l'expertise de l'auteur. Il faut cependant remarquer que le même problème se pose pour les données entrées par les futurs usagers du système et que les solutions retenues pour se protéger des erreurs des utilisateurs s'appliquent également aux erreurs commises par les auteurs de descriptions.

Plus grave est le fait que les données publiées ne suivent aucun format commun. Le vocabulaire et le choix des caractères varient d'une description à l'autre, ce qui multiplie le nombre de caractères utilisés dans l'ensemble des descriptions publiées. J'ai déjà trouvé 465 caractères dans les descriptions des espèces de

l'ordre des Tylenchida et la liste ne cesse de s'allonger. Il devient presque impossible de se rappeler tous les caractères de la liste et, même si l'on s'en souvient, il est difficile de retrouver le caractère cherché dans une base de données comportant 465 colonnes. La création de la base de données devient une tâche au dessus des possibilités humaines et il faut chercher de l'aide électronique.

4.3.2. Le Terminator

La compréhension par les systèmes d'intelligence artificielle des textes écrits en langage naturelle est l'un des points chauds des recherches actuelles en informatique mais les résultats n'ont pas encore atteint la maturité nécessaire pour devenir utilisables en pratique. Avec Jim Diederich et Jack Milton, j'ai tourné la difficulté en créant le *Terminator,* un système qui cherche dans les phrases successives d'une description les termes et expressions réunis dans la liste des noms d'organes et caractères en usage pour un groupe biologique donné. Quelques règles permettent au système de deviner dans la liste le caractère auquel se réfère chaque phrase de la description.

L'opérateur peut accepter le caractère proposé par le système ou en sélectionner un autre au terme d'une recherche aidée par le système. Dans la figure 14 le Terminator est en train d'étudier la phrase *the body is ventrally curved in the shape of a widely open C* et l'opérateur a souligné l'expression *widely open C.*

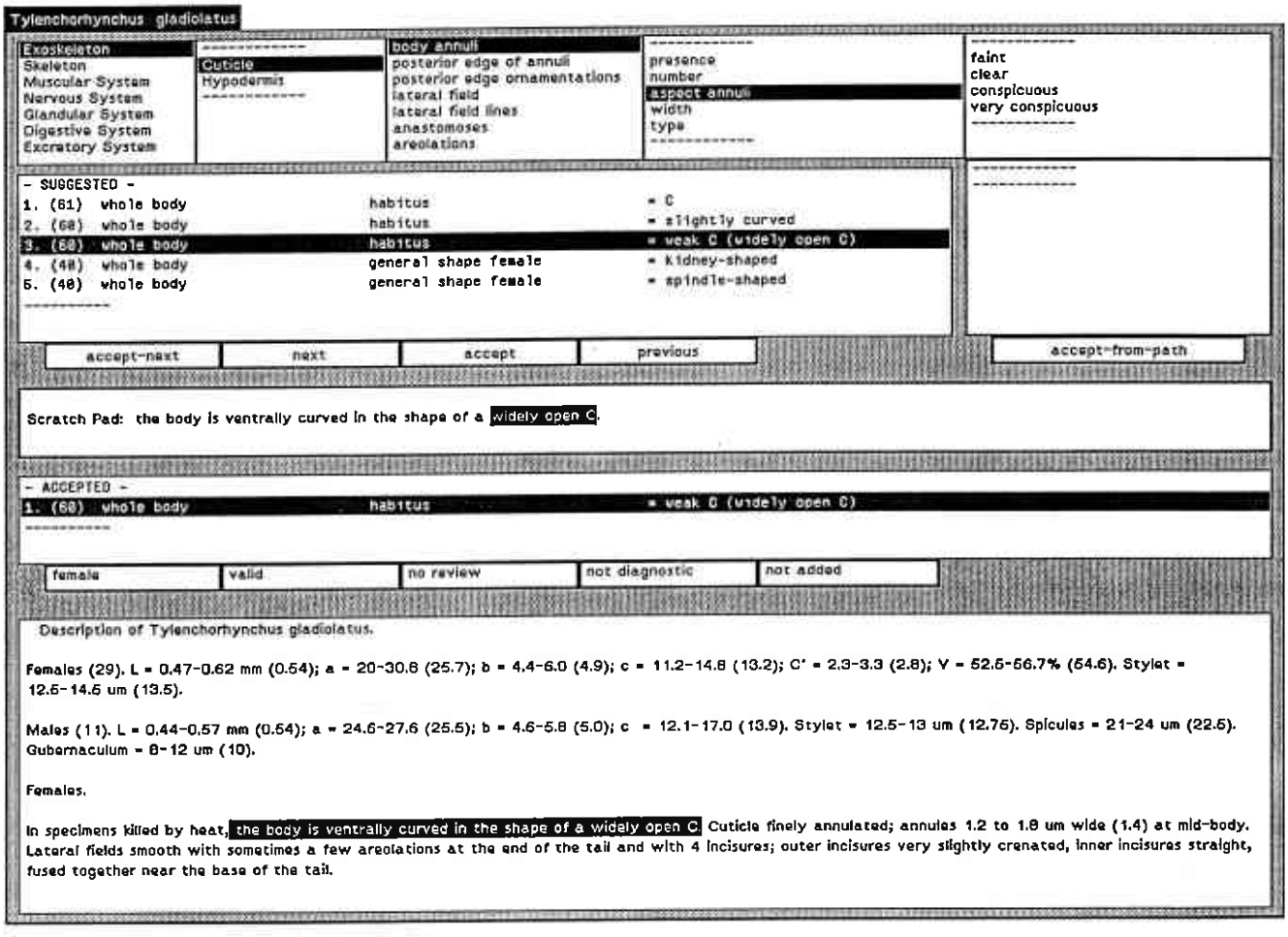

Figure 14: Le Terminator, un outil pour l'extraction semi-automatique des données morphologiques de la littérature.

Dans le panneau marqué SUGGESTED, le Terminator a proposé plusieurs possibilités pour le placement de cette information dans la base de données. L'opérateur voit que *whole body, habitus* est le caractère qui convient et que l'état *widely open C* correspond à l'état *wide C*. L'opérateur sélectionne cet état comme correct et il entre l'expression *widely open C* trouvée dans le texte comme un synonyme de l'état *wide C*. Il peut aussi directement accepter l'expression trouvée dans le texte d'un simple clic de la souris et le Terminator la place dans le panneau marqué ACCEPTED.

Les premiers essais montrent qu'il faut environ un quart d'heure pour extraire les données d'une description. Il sera possible à un opérateur de traiter en un an la dizaine de milliers de descriptions de populations de nématodes phytoparasites qui ont été publiées jusqu'à présent. A noter que les articles publiées en langues autres que l'anglais pourront être analysés en traduisant simplement la liste de termes et d'expressions utilisés par le Terminator.

4.3.3. Schéma d'une base de données morphologiques

Le stockage en ordinateur des données extraites par le Terminator pose un certain nombre de problèmes car les systèmes de gestion de bases de données (*DBMS* en anglais) ont été conçus pour les milieux d'affaire et d'industrie. Le format qu'ils imposent convient mal à la richesse, pour ne pas dire l'exubérance, des données biologiques. En collaboration avec Jim et Jack, j'ai défini les caractéristiques des données morphométriques auxquelles nous avons affaire. Le but de cet exercice est la définition d'un *bio-DBMS*, un système de gestion de bases de données capable de prendre en compte tous les aspects des données biologiques tels qu'il ont été définis plus haut (voir § 1.2).

Les études en cours permettront de mettre un peu d'ordre dans le fouillis extrait de la littérature. Des règles et des relations mathématiques permettront de compresser les données dans une base de données finale (NEMbase) la plus petite possible, c'est à dire comprenant le nombre minimum de caractères nécessaires à la description complète d'un spécimen ou d'une espèce. Si l'usage de cette base de données se répand parmi les nématologistes, il est à espérer que les auteurs prendront peu à peu l'habitude d'utiliser les caractères et les termes retenus, ce qui se traduira à long terme par une amélioration de la qualité des descriptions.

Il est prévu de mettre NEMbase à la disposition des auteurs pour qu'ils y déposent de nouvelles descriptions d'espèces connues ce qui leur évitera d'avoir à les publier sur papier. Les journaux scientifiques n'auront plus qu'à publier les descriptions des populations types des espèces nouvelles. Le stockage dans une base de données unique de nombreuses populations de la même espèce permettra de proposer une définition statistique des espèces.

4.4. La base de connaissances

Une base de connaissances est une collection d'information sur l'expertise (plus éventuellement des données brutes) se rapportant à un certain domaine et représentée d'une manière (par exemple règles, réseaux sémantiques, etc.) qui rend cette expertise explicitement disponible pour un logiciel d'application. La base de connaissance de NEMISYS est indépendante de la base de données NEMbase, ce qui permettra d'utiliser NEMbase conjointement à d'autres bases de connaissances pour des applications autres que l'identification.

L'expertise nécessaire pour l'identification comprend diverses catégories de connaissances que je vais brièvement passer en revue.

4.4.1. Les métadonnées

Dès le début du projet NEMISYS (Diederich et al., 1989), nous avons défini les métadonnées suivantes:

(1) *Métadonnées indépendantes des instances.* Ce genre de métadonnées comprend le type de caractère avec échelle et étendue telles qu'elles ont été définies plus haut (voir § 1.2.2), l'unité pour les mesures, et certaines relations entre caractères, par exemple les caractères dont la value peut être calculée à partir de celle d'autres caractères (par exemple, épaisseur d'un anneau = longueur du corps divisée par nombre d'anneaux).

(2) *Métadonnées relatives aux caractères pour chaque instance.* Ce sont les métadonnées dont j'ai parlé

plus haut (voir § 1.1), visibilité des organes, ambiguïté, variabilité et utilité du caractère. La priorité sépare les caractères d'un nid ou d'une espèce en caractères primaires, secondaires ou tertiaires. Elle est en principe calculée à partir d'autres métadonnées. Un caractère primaire n'est ni ambigu ni variable, et il est lié à un organe bien visible dans le nid ou l'espèce en question. La métadonnée "priorité" permet aux experts de passer outre au classement proposé automatiquement par le système.

(3) *Métadonnées relatives aux valeurs des caractères*. Ce dernier type comprend par exemple l'intervalle des valeurs prises par un caractère dans une espèce, la valeur la plus typique du caractère pour un nid d'espèces donné, la fréquence d'apparition de chaque état ou classe de valeur dans un nid, etc.

Ces métadonnées sont, soit recueillies auprès des experts du NIP, soit calculées à partir des données de NEMbase. Leur définition a demandé de longues discussions avec mes collègues car les exigences de l'informatique vont souvent à l'encontre des pratiques courantes en nématologie. Par exemple, une certaine expertise est cachée dans les noms de caractères tels "phasmide absente, peu visible, présente, ou très visible". Ce genre d'information, que l'on trouve très souvent dans les descriptions morphologiques, est un mélange de deux types d'information bien différents. La présence ou l'absence de l'organe est une donnée qui a sa place dans la base de donnée NEMbase. Lorsque la phasmide est présente, sa plus ou moins grande visibilité est une métadonnée qui doit être engrangée dans la base de connaissance. A noter que la visibilité de la phasmide est peut-être liée au diamètre de son ouverture, ce qui est une autre donnée!

4.4.2. Relations entre caractères

De nombreux caractères sont liés entre eux par des relations de redondance, de dépendance, de dérivation, de récapitulation, etc., telles celles qui ont été définies plus haut (voir § 1.2.2). Dans chaque cas, il est nécessaire de définir les règles à appliquer pour élucider ces relations. Par exemple, les positions respectives de l'hémizonide et du pore excréteur sont des caractères redondants pour lesquels il faut définir les règles permettant de passer de l'un à l'autre en fonction des valeurs rencontrées dans les descriptions. Ces règles sont du type SI condition ALORS action, par exemple SI hémizonide, position par rapport au pore excréteur = postérieure, ALORS pore excréteur, position par rapport à l'hémizonide = antérieur. Il existe des centaines de relations possibles qu'il faudra définir et stocker dans la base de connaissance.

Les participants au Projet NEMISYS seront mis à contribution pour cette tâche. Pour cela il leur sera demandé de répondre à des enquêtes publiées dans le *NEMISYS International Project Update*. Ce bulletin sert aussi pour les enquêtes destinées à résoudre le problème de la correspondance entre caractères flous et données chiffrées et il est à craindre que les experts se lassent bien vite et cessent de répondre aux questionnaires. Pour les aider, un réseau électronique a été mis en place. Questions et réponses passeront par courrier électronique.

4.5. Etat d'avancement du projet

Il faut bien comprendre que NEMISYS ne sera jamais fini! C'est un système ouvert, un ensemble d'outils, et il est prévu de continuer de créer outil après outil, à mesure que leur besoin s'en fera sentir et en fonction des possibilités en temps et en argent. Il est donc impossible de décrire l'état d'avancement du projet entier, mais seulement celui de composantes, base de données, base de connaissances, et d'outils particuliers.

La base de données comporte les descriptions de promorphes, de nids d'espèces et d'espèces. Les données se rapportant aux promorphes et aux nids d'espèces sont obtenus auprès des experts associés au projet. A l'heure actuelle deux des onze promorphes définis pour les nématodes phytoparasites (hoplolaimides et dolicho/bélonolaimides) sont complètement décrits ainsi que tous leurs nids d'espèces. Une douzaine de nids isolés ont été décrits dans les autres promorphes. La description complète d'un troisième promorphe (pratylenchides) a démarré avec un protocole plus léger, ce qui devrait accélérer les choses. Les données existantes sont suffisantes pour tester les outils qui s'arrêtent au niveau nid d'espèces, tels le BASIC ID.

Il a été très difficile d'obtenir le financement de la création d'une base de données pour les descriptions d'espèces, activité traitée avec mépris par les rapporteurs de la *National Science Foundation* qui n'y ont vu que manipulation d'une information existante. Ma persévérance a cependant été récompensée et j'ai obtenu 20.000 dollars (112.000 francs au cours du jour) pour acheter l'ordinateur (un Mac IIfx poussé a

20 méga octets de mémoire vive) qui me permet d'utiliser le Terminator. J'ai embauché des étudiants qui ont commencé la capture des descriptions d'espèces publiées dans la littérature. Je disposerai bientôt des données pour une cinquantaine d'espèces dans deux genres (*Radopholus* et *Globodera*). Cela peut paraître faible en regard de la tâche à accomplir (environ 4.000 espèces) mais c'est un début et surtout cela permettra de tester les outils pour l'identification spécifique tels NEMAID. De plus, je pense que je pourrai faire faire bien plus que les deux genres exigés par le contrat de financement. Il faut environ une demi heure de travail par description. Comme mes étudiants sont payés 7,10 dollars de l'heure, une description me coûte moins de quatre dollars. Il me reste environ 6.000 dollars après l'achat du matériel, soit l'équivalent de 1.500 descriptions. Même en tenant compte des impondérables ce travail représentera une partie significative de la tâche à accomplir (environ 10.000 descriptions pour les 4.000 espèces du groupe).

La base de connaissance recueillera l'expertise des spécialistes comme décrit ci-dessus (paragraphe 4.4.). Les métadonnées attachées aux nids d'espèces one été réunies pour les nids déjà décrits. Les métadonnées décrivant la nature des caractères ont été définies sur le plan conceptuel et deux articles sont en cours de rédaction, l'un dans lequel je pose le problème sur le plan biologique, l'autre dans lequel Jim Diederich explique les solutions informatiques que nous nous proposons d'appliquer. Certaines de ces solutions font déjà partie du système, d'autres lui seront ajoutées à mesure que leur besoin se fera sentir. Notre situation financière précaire nous interdit d'embaucher l'aide technique massive qui serait nécessaire à la mise en oeuvre rapide du système, et il est impératif de bien ménager nos faibles ressources. Certains aspects du projet sont délibérément mis en sommeil, s'ils ne sont pas absolument indispensables à la version de base du système.

Cette version de base est bientôt prête. Nous avons déjà un prototype qui a été démontré en juin dernier au cours d'une tournée dans plusieurs laboratoires en France. Ce prototype comprenait la liste minimum d'outils nécessaires pour illustrer le concept de NEMISYS: BASIC ID, SHOW ME (décrits ci-dessus, paragraphe 4.2.2.), ASK ME (proche du SHOW ME), PROMORPH (pour la sélection d'un promorphe). Le prototype incluait aussi des sous-outils, fragments de futurs outils destinés à illustrer nos plans d'avenir. Il s'agissait surtout de fonctions graphiques (comparaison des illustrations de deux nématodes, tableau de toutes les formes possibles de queues, etc.) et de ce que Jack et Jim ont baptisé *pixel menus,* menus qui changent de contenu en fonction du point (pixel) de l'image sur lequel se trouve le curseur lorsqu'on les appelle. Ceci permettra de sélectionner un détail morphologique d'une image pour obtenir le nom de l'organe en question, sa définition et les caractères qui lui sont attachés. Tous ces outils (ou certains de leurs panneaux) peuvent être doublés et chaque copie traitée indépendamment de l'autre. Enfin les données entrées dans un outil apparaissent automatiquement dans tous les autres.

La version 1 de NEMISYS permettra à l'utilisateur d'identifier jusqu'au niveau de l'espèce les formes incluses dans la base de données. NEMISYS 1 comprendra les mêmes outils que le prototype plus les outils pour l'identification des espèces, en particulier l'outil NEMAID. Certains de ces outils (BASIC ID) sont terminés, d'autres en sont aux derniers ajustements. L'outil NEMAID est en cours de réalisation. Son concept, basé sur le calcul d'un coefficient de similarité selon les principes exposés plus haut (paragraphe 3.4.) est bien au point. NEMISYS 1 devrait être prêt à la fin de l'année. Le Terminator est en cours de testage. Il est déjà question de l'adapter à d'autres disciplines (je suis en pourparlers avec des entomologistes américains et avec un ichtyologiste du Muséum National d'Histoire Naturelle).

4.6. Test des performances

NEMISYS est un ensemble d'outils dont l'utilisation dépend des circonstances et de l'utilisateur. Chaque outil et chaque fonction peuvent être testés séparément mais les performances du système entier dépendront de la façon dont il est utilisé.

Les outils sont testés en permanence à mesure qu'ils sont créés et mis au point. Par exemple, l'interface et les régles du Terminator ont été analysées et changées pour rendre cet outil plus performant et plus facile à utiliser. Lorsque le logiciel ne parvient pas à trouver le caractère contenu dans une phrase d'une description d'espèce, il est possible de placer dans un dossier la phrase en question, les caractères suggérés par erreur, et le bon caractère sélectionné par l'opérateur. Le contenu de ce dossier permet ensuite à mes collègues informaticiens de chercher la raison de l'échec initial du système. Il faut en moyenne un quart

d'heure pour extraire les données d'une description d'espèce en utilisant le prototype du logiciel. Le système final sera certainement encore plus rapide. Le BASIC ID, lui, fonctionne parfaitement avec les nids du promorphe utilisé pour les tests (hoplolaimides). Il est peu probable que l'addition des dix autres promorphes diminuera beaucoup les performances car, en régime normal, l'utilisateur commence son identification en sélectionnant un promorphe et il part donc d'un nombre de nids du même ordre de grandeur que celui utilisé pour les tests (une vingtaine).

Les données sont prêtes pour tester l'outil NEMAID au cours de sa réalisation. Les descriptions des espèces du genre *Radopholus* seront disponibles quand l'outil sera prêt à être testé. L'outil NEMAID est destiné à l'identification spécifique après qu'un ou deux nids aient été retenus par l'outil BASIC ID. Considérant qu'un nid moyen contient le même nombre d'espèces qu'un genre tel *Radopholus,* on peut dire que les tests de NEMAID seront conduits en vraie grandeur.

En conclusion, les grand traits du système de base sont en place ou seront en place dans quelques mois, au moment où les données seront disponible pour l'expérimentation en vraie grandeur. Cette expérimentation n'a pas encore été conduite d'une façon systématique, mais les tests limités déjà effectués permettent d'être optimiste sur les chances de succès.

4.7. Accueil de NEMISYS par les futurs utilisateurs

Il est bien sûr trop tôt pour dire comment la première version opérationnelle de NEMISYS sera accueillie par les utilisateurs. Je peux cependant rapporter l'accueil très favorable qui a été donné au prototype quand il a été démontré au cours d'une tournée organisée en France l'année dernière sous l'égide du CNRS et lors de divers séminaires et ateliers de travail (Diederich *et al.,* 1990; Fortuner, 1991a; 1991b).

Les nématologistes sont très intéressés par nos idées puisque le *NEMISYS International Project* groupe plus de soixante-dix taxonomistes, un chiffre énorme pour notre petite discipline. Une démonstration de NEMISYS a été filmée et une bande vidéo circule en ce moment parmi les participants du projet. Quand au Terminator, les biologistes à qui je parle du concept d'une extraction semi-automatique des données publiées dans la littérature ont tous la même réaction que Dallwitz, le père du langage DELTA, qui m'a dit: "J'en rêve depuis vingt ans mais c'est impossible à réaliser". La démonstration du Terminator l'a laissé sans voix!

L'accueil des informaticiens pour les concepts développés par mes collègues dans le cadre de ce projet n'est pas moins encourageant. Par exemple, Dr. Hafner a écrit en ces termes à Jim Diederich:

> *"I am definitely going to use Chapter 3 as part of the readings for my seminar at Harvard in the spring. [...]. I was very intrigued by your ideas about the expert workstation. I think this approach is extremely valuable in applying AI concepts to research tasks, in which traditional advisory expert systems do not seem particularly useful. I'm delighted that you have some good publications on this project, which I can use to bring these issues on my class. The course is entitled "Computational Analysis of Legal Research," and its purpose is to develop a model of the legal research process, which we can then translate into the design of an expert legal researcher's workstation along the same lines as your expert biology researcher's workstation."*[4]

Ces commentaires montrent clairement que notre concept d'un ensemble d'outils est applicable en dehors de notre domaine et même, comme le suggère Carol Hafner, à la recherche de points de Droit. Il faudra attendre la sortie de NEMISYS 1 et ses premiers tests à l'extérieur de notre petit groupe pour être sûr de son succès, mais les premières réactions me laissent très optimiste sur ses chances.

---

[4]: J'utiliserai certainement votre chapitre 3 dans la liste des travaux à lire pour mon séminaire à Harvard au printemps. (...) J'ai été très intéressée par vos idées sur la station de travail experte. Je pense que cette approche a une grande valeur pour l'application des concepts d'intelligence artificielle aux travaux de recherche pour lesquels les systèmes experts consultatifs traditionnels ne semblent pas très utiles. Je suis très heureuse que vous ayez quelques bons articles sur ce projet, que je peux utiliser pour introduire ces points dans mon cours. Le cours est intitulé "Analyse de la recherche légale par ordinateur" et il vise à développer un modèle du processus de recherche légale, que nous pourrons ensuite traduire dans la conception d'une station de travail de recherche pour l'expert en Droit dans la ligne de votre station de travail de recherche pour l'expert biologiste.

CONCLUSION

NEMISYS, comme son nom l'indique, c'est d'abord l'identification des nématodes. Le système grandira pour offrir de plus en plus d'outils et de plus en plus de fonctions. Par exemple, des systèmes experts, des réseaux neuronaux ou des méthodes statistiques pourraient être employés si les données (règles, métadonnées ou données statistiques) existent pour leur application dans un petit groupe d'espèces. J'ai pu séparer les espèces du groupe *Hirschmanniella oryzae / H. belli* à l'aide d'analyses discriminantes (Fortuner & Maggenti, 1991). lorsqu'une session d'identification sera bloquée au niveau de ces deux espèces, une future version de NEMISYS donnera accès à une fonction qui utilisera la matrice de données établie au cours de mon étude pour classer le spécimen à identifier par rapport aux axes que j'ai définis à l'époque. Dans un autre ordre d'idées, un petit système expert sera disponible pour guider l'utilisateur débutant dans son utilisation des divers outils.

Le but lointain du développement de NEMISYS est l'identification automatique sans intervention humaine. Cela demandera un certain nombre de travaux dont certains sont déjà en cours. Il faudra d'abord établir une liste minimale de caractères ainsi que les règles qui permettent, à partir de ces caractères, de retrouver tous les autres (par exemple, passage d'une donnée chiffrée au caractère flou correspondant). Cette liste a déjà été établie pour les membres de l'ordre des Tylenchida à l'aide d'un outil spécialisé (outil SCHEMA DEVELOPMENT) qui permet de définir les organes et leurs caractères, ainsi que les relations qui existent entre ces caractères.

Une fois établie la liste minimale de caractères, il faudra rédiger des instructions précises pour mesurer ces données à partir de points caractéristiques (*landmarks*). Ces règles peuvent être très simples (par exemple la mesure de la distance qui sépare deux points), ou plus compliquées, comme dans le cas de la reconnaissance des formes. Des systèmes de reconnaissance automatique des formes existent déjà. Par exemple, Chris Meacham à Berkeley a mis au point un système qui applique une transformée de Fourier au tracé (manuel ou automatique) du contour d'un organe pour obtenir à une expression mathématique de la forme de cet organe. Ces expressions mathématiques peuvent être analysées par une analyse discriminante qui permet d'associer à chaque forme un point dans un système d'axes. L'identification de la forme de l'organe dans un spécimen inconnu se fait en calculant sa position dans le système d'axes ainsi défini.

§

Le prototype de NEMISYS comporte des bases de données et de connaissances et des outils qui serviront d'abord à l'identification des nématodes phytoparasites. NEMISYS a été conçu comme un système ouvert, et il est prévu de le développer selon deux directions orthogonales l'une à l'autre.

Tout d'abord, il sera possible de créer de nouveaux outils utilisant de nouvelles bases de connaissances pour des applications différentes dans le même groupe biologique. Nous avons commencé de réfléchir à ce que pourrait être un outil pour l'alpha-taxonomie (description d'espèces nouvelles de nématodes phytoparasites), un processus calqué sur celui de l'identification. En présence d'une espèce supposée nouvelle, l'outil aidera l'usager à vérifier qu'elle appartient bien à un genre donné. Il établira la liste des espèces du genre auxquelles la nouvelle forme est reliée, et il proposera les caractères diagnostiques de l'espèce nouvelle. *Tout ceci sera lié à la question de la confiance que l'on peut accorder aux données et donc à la notion* d'endossement dont je parle plus haut (paragraphe 2.3.4.).

Plus intéressants seront les outils qui s'adresseront à la question de l'évolution des caractères et des formes. Nous avons dans nos cartons ELECTRONEMA, le nématode électronique, qui utilisera une série de photos pour montrer, soit l'aspect de tous les organes dans une espèce donnée (y compris leur développement au cours des stades larvaires), soit l'aspect d'un organe donné chez des espèces différentes.

L'ensemble devrait aider à la formulation d'hypothèses sur les plésiomorphies et les apomorphies pour certains caractères (l'inclusion de vues de pseudodiplogastérides est envisagée pour aider à ce genre d'études). Chaque application se servira des bases de données et de connaissances créés pour d'autres usages, complétées par des bases s'adressant aux problèmes spécifiques à l'application nouvelle. Par exemple, un outil pour l'aide à la systématique cladistique se servira sans doute de la base de données morphologiques, mais il lui faudra une base de connaissances avec les apomorphies pour chaque taxon.

Je pense aussi à un outil pour la modélisation de l'évolution des formes qui utiliserait les mêmes transformées que l'outil d'identification automatique dont je parlais plus haut. Si deux formes, l'une présumée ancestrale, l'autre dérivée, sont placées dans le système d'axes discriminants, on peut utiliser le système à l'envers et obtenir l'aspect des formes intermédiaires à partir des coordonnées d'un point situé à mi-chemin des deux points donnés. Il est bien sûr impossible d'en déduire immédiatement que ces formes intermédiaires représentent bien le chemin suivi au cours de l'évolution pour passer de la forme ancestrale à la forme dérivée. Les expressions mathématiques qui permettent de passer d'une forme à l'autre ne sont pas nécessairement calquées sur l'évolution biologique. Il n'empêche que ces formes intermédiaires fournissent un modèle qu'il devrait être possible de tester. Une indication intéressante serait la découverte dans la nature d'une forme identique à celle produite par les fonctions mathématiques. Cet outil s'appuie sur plusieurs méthodes existantes qu'il suffirait de réunir dans un outil unique pour pouvoir réaliser des expériences qui devraient être plein d'intérêt.

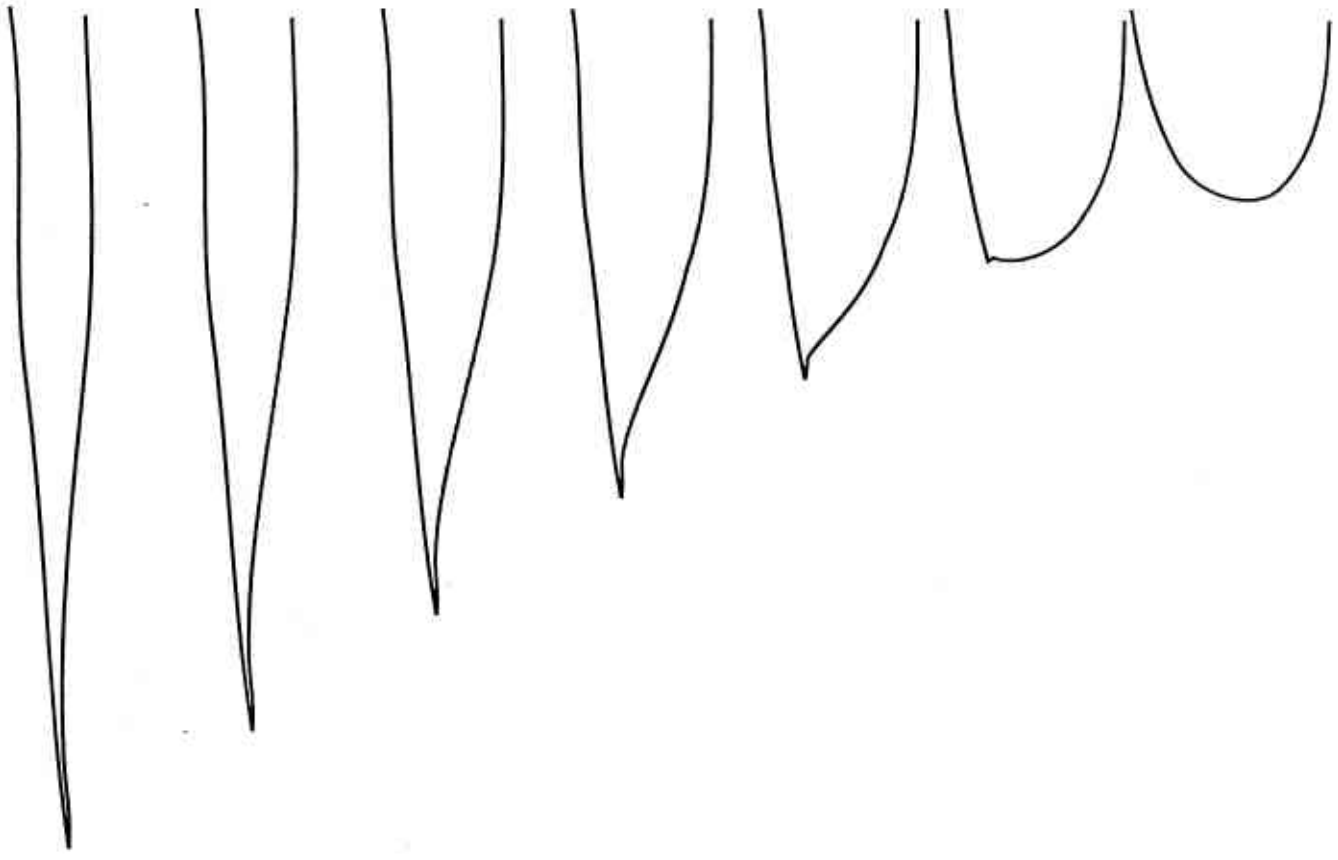

Figure 15. Passage d'une queue effilée à une queue ronde. Fonction "Blending" du logiciel Adobe Illustrator.

La figure 15 montre le passage d'une queue pointue à une queue ronde. Elle a été obtenue par la fonction Blending de *Adobe Illustrator,* un logiciel de design. Elle n'a aucune valeur mathématique mais elle donne une idée des résultats auxquels on peut s'attendre. Il est amusant de noter que l'évolution de la forme passe par un stade "projection ventrale" qui est visible chez certains nématodes.

L'écologie est une discipline très liée à la taxonomie et elle fournira sans doute la matière aux premières applications de NEMISYS en dehors du cadre étroit qui lui a donné naissance. NEMISYS permettra aux écologistes d'identifier plus rapidement les espèces et on peut espérer voir se multiplier le nombre d'études de peuplements nématologiques rencontrés dans des circonstances particulières. Les écologistes pourront se servir de NEMbase pour déposer les descriptions des spécimens témoins correspondant à ces études faunistiques. Les performances du système d'identification ne peuvent que bénéficier d'un accroissement du nombre de données qu'il utilise et on voit que les rétroactions multiples entre identification et écologie serviront les deux sciences.

Les relations entre écologie et identification sont également apparente dans certaines fonctions du processus d'identification (règle de Bayes, focalisation des recherches), qui ont besoin d'informations d'ordre écologique. Ces données, liste de plantes hôtes, répartition des espèces en fonction de l'origine géographique des prélèvements, de la nature du sol, du régime hydrique, de la partie de la plante étudiée, etc., seront réunies dans des bases de données ad-hoc. Ces bases de données seront peuplées à partir des études publiées par les écologistes, et ceux-ci seront sans doute très heureux d'avoir accès à un index de leurs travaux.

Un outil simple permettra de décrire les peuplements observés. L'entrée des données comportera la liste des espèces (couplée peut-être avec les outils d'identification), le nombre de spécimens, la description du site et son emplacement, ce dernier point se faisant avec l'assistance d'un système d'information géographique. La sortie serait des cartes de répartition géographique, des graphiques décrivant l'influence des facteurs de milieu, ou les paramètres descriptifs habituels, densité et fréquence de chaque espèce, soit relative, soit absolue, indice de dominance, biomasse, etc. Tous ces concepts sont bien connus et il suffira de disposer de temps et d'argent pour les mettre à la disposition des chercheurs par l'intermédiaire d'un outil NEMISYS.

Plus difficile sera l'élaboration d'outils permettant les études plus poussées et en particulier la modélisation. Un outil NEMISYS pourrait être utilisé par les écologistes pour proposer une série d'algorithmes pour la modélisation d'un biotope donné.

Les outils de NEMISYS sont conçus pour être indépendants du groupe biologique utilisé pour leur création. Un deuxième axe de développement du système consistera à étendre l'utilisation des outils créés pour les nématodes à d'autres groupes biologiques (par exemple, les poissons des eaux continentales françaises) en changeant simplement les bases de données et de connaissances. La nouvelle base de données sera créée à l'aide du Terminator qui doit être modifié pour permettre la création du schéma simultanément à l'entrée des données, chaque nouveau caractère ou organe découvert dans les descriptions d'espèces étant placé à l'endroit convenable par l'utilisateur avec l'assistance du système. La base de connaissance sera à développer en collaboration avec les experts de chaque domaine considéré, mais il est à espérer que les solutions aux problèmes fondamentaux étudiés pour les nématodes seront applicables aux problèmes similaires rencontrés dans les autres groupes. Les remarques de Carol Hafner (paragraphe 4.7.) montrent que le principe de la station de travail experte est applicable en dehors de la biologie.

§

Le projet NEMISYS dépasse les possibilités d'un homme seul, et même de trois hommes en comptant mes collègues informaticiens. Ceci est vrai pour le projet tel qu'il existe, c'est à dire pour l'identification des nématodes phytoparasites, ce l'est encore plus lorsque l'on envisage l'extension à d'autres groupes et à d'autres disciplines. J'ai essayé de montrer que je me sentais capable de diriger un tel projet dans ses développements les plus ésotériques, mais je n'aurai jamais le temps de réaliser seul le détail des applications. Il faut donc prévoir que dans un avenir proche le projet NEMISYS s'efface pour laisser la place à une réalisation de toute autre envergure, celle d'un centre de bioinformatique. La définition de la mission et des moyens d'un tel centre en est encore à ses premiers pas et il ne m'est pas possible de le décrire en détail. Je me contente ici d'en donner ici une vue d'ensemble.

Le Centre de Bioinformatique

Je propose avec Jim Diederich et Jack Milton la création d'un centre de bioinformatique qui, sans vouloir remplacer les équipes existantes ou limiter en quoi que ce soit leur autonomie, devrait accomplir les fonctions suivantes:

1. coordonner les recherches en cours pour les applications biologiques de l'informatique;
2. démarrer des recherches originales en informatique pure et assurer le transfert de technologie vers la biologie;
3. intégrer les résultats obtenus au centre et ailleurs dans un ensemble d'outils type NEMISYS;
4. former les biologistes à l'utilisation pratique de ces outils; et
5. aider les structures d'enseignement existantes (université, grandes écoles) à développer des cours spécialisés dans le domaine de la bioinformatique.

Le centre ne viendrait pas en concurrence des équipes déjà en place, mais leur offrirait une ressource nouvelle. Les chercheurs en bioinformatique qui ont développé des systèmes originaux seraient invités à venir les intégrer à un système général. Par exemple, NEMISYS pourrait très facilement avoir un outil XPER donnant accès au système expert développé par Lebbe (1984). Cet outil serait accessible pour l'identification dans les groupes biologiques couverts par XPER. L'outil NEMAID en cours de développement est en fait l'intégration au système d'un logiciel existant (Fortuner, 1983; 1986; Fortuner & Wong, 1983; 1985).

Coordination des recherches

Le centre devrait développer une base de données pour informer les chercheurs intéressés des activités des autres bioinformaticiens en France ou dans le monde. Les chercheurs d'un labo ignorent souvent que des chercheurs d'un autre centre travaillent sur un sujet parallèle au leur dans un autre domaine. La base de données servirait de trait d'union entre tous.

Le centre pourrait aussi organiser des "opération de sauvetage des connaissances" au cours desquelles seraient capturés l'expérience, le savoir et le savoir-faire de biologistes de renom. Le centre essayerait de mettre cette expertise sous une forme qui puisse être engrangée dans une base de connaissances.

Chaque année, le centre organiserait ateliers, rencontres et conférences pour faire le tour d'un sujet particulier. Des experts nationaux et internationaux spécialistes du domaine seraient invités à venir travailler au centre pendant plusieurs mois ou même un an.

Activités de recherche

Les informaticiens du centre travailleraient sur les problèmes liés aux applications biologiques, comme cela a été le cas au cours du projet NEMISYS qui a fortement modifié l'orientation des recherches auxquelles se livraient mes collègues Diederich et Milton. Le centre s'intéresserait aux problèmes qui couvrent des domaines multiples plutôt qu'à ceux qui n'ont qu'un champ d'action limité. Ce genre de recherche est difficile à justifier dans un labo à la mission spécifique bien délimitée, mais il s'appliquerait parfaitement au concept d'un centre de bioinformatique. Un exemple est la mise au point d'un système de gestion des bases de données biologiques par l'adaptation au domaine biologique des logiciels mis au point pour le commerce et l'industrie.

Les biologistes du centre, eux, se tiendraient au courant des recherches en cours en biologie pour repérer les problèmes susceptibles de recevoir une solution informatique. Ils devraient identifier les spécialistes de ces problèmes et servir d'intermédiaires entre ces experts et les informaticiens du centre.

Le centre aurait l'oeil sur les recherches de pointe en informatique (ordinateurs parallèles, réalité virtuelle, réseaux neuronaux, etc.) et être prêt à les utiliser.

Formation pratique

La formation informatique des biologistes ferait partie de la mission du centre dans le cadre de la dissémination de ses résultats. Il faudrait non seulement les former à l'utilisation des outils mis au point par le centre et par d'autres chercheurs, mais aussi les préparer à collaborer avec les informaticiens pour

développer leurs propres recherches en bioinformatique. Le projet NEMISYS m'a donné une bonne expérience des difficultés de communication entre spécialistes des deux sciences.

Le centre devrait aussi participer au transfert de technologie dans le cadre du dialogue Nord/Sud. Les pays en développement manquent des outils modernes, des données et des connaissances nécessaires pour utiliser les méthodes modernes. Le centre pourrait chercher les fonds nécessaires pour inviter des chercheurs de ces pays à venir se former sur place à l'utilisation des outils et méthodes mis au point au centre. Ils repartiraient chez eux munis du matériel nécessaire à l'utilisation de leur nouveau savoir. Les communications (si possible par voie électronique) avec les anciens élèves du centre seraient cruciales pour briser leur isolement, les mettre au courant des développements les plus récents et leur donner les compléments de formation qui leur seraient nécessaires.

Enseignement

Il serait bon que les universités associées au centre proposent des classes spécialisées en informatique appliquée à la biologie. Au début ces classes pourraient être rattachées à l'enseignement traditionnel en biologie. Plus tard, elles pourraient donner naissance à un certificat "Bioinformatique" spécialisé. Dans tous les cas, les élèves devraient recevoir un haut niveau de connaissances à la fois en biologie et en informatique. Certains des étudiants pourraient d'ailleurs avoir déjà un diplôme dans l'une des deux sciences et vouloir se spécialiser dans l'autre. Les jury de thèse devraient inclure des spécialistes de toutes les disciplines concernées.

§

NEMISYS devra se transformer peu à peu en une station de travail experte pour le biologiste. Le système ne sera plus alors lié à un certain type d'activités (identification, systématique ou écologie), ni à un groupe biologique donné (nématodes phytoparasites). Il sera le modèle d'un nouveau type de système alliant toutes sortes de méthodes mathématiques et statistiques à une interface permettant un dialogue naturel entre l'utilisateur et la machine. Ce modèle pourra être appliqué à toutes les disciplines biologiques et même, si l'on en croit le Dr. Hafner, à des domaines bien éloignés de la biologie tels le Droit.

RÉFÉRENCES

Références des travaux personnels (voir ci-dessous la liste complète de mes publications)

Fortuner, R., 1970. On the morphology of *Aphelenchoides besseyi* Christie, 1942 and *A. siddiqii* n.sp., (Nematoda, Aphelenchoidea). *J. Helminth.*, 44(2): 141-152.

Fortuner, R., 1974. Description de *Pratylenchus sefaensis* n.sp. et de *Hoplolaimus clarissimus* n.sp. (Nematoda: Tylenchida). *Cah. ORSTOM, sér. Biol.*, No. 21 (1973): 25-34.

Fortuner, R., 1975. Les nématodes parasites des racines associés au riz au Sénégal (Haute-Casamance et régions Centre et Nord) et en Mauritanie. *Cah. ORSTOM, sér. Biol.*, 10(3): 147-159.

Fortuner, R. 1976. Etude écologique des nématodes des rizières du Sénégal. *Cah. ORSTOM, sér. Biol.*, 11(3): 179-191.

Fortuner, R., 1983. Computer assisted semi-automatic identification of *Helicotylenchus* species. *Calif. Pl. Dis. Rept*, 2: 45-48.

Fortuner, R., 1984. Morphometrical variability in *Helicotylenchus* Steiner, 1945. 6: Value of the characters used for specific identification. *Revue Nématol.*, 7(3): 245-264.

Fortuner, R., 1986. A better assessment of variability of qualitative characters for the computer identification program Nemaid. *Revue Nématol.*, 9(3): 277-279.

Fortuner, R., 1987. *Variabilité et identification des espèces chez les nématodes du genre* Helicotylenchus. Collection Etudes et Thèses, ORSTOM, v + 232 pages.

Fortuner, R., 1989 (Editor). *Nematode identification and expert-system technology*. New York, Plenum Publishing Corp., ix + 386 pp.

Fortuner, R., 1991a. L'identification biologique, problèmes et besoins; évaluation des méthodes traditionnelles. *Exposé, Muséum National d'Histoire Naturelle, Paris; ORSTOM, Bondy; ENGREF, Nancy; INRIA, Grenoble; URA CNRS 243 Université de Lyon, Villeurbanne; ORSTOM, Montpellier, INRA, Toulouse, 24 juin-5 juillet 1991.*

Fortuner, R. 1991b. Constitution d'une base de données à partir des descriptions publiées dans la littérature. *Exposé, Muséum National d'Histoire Naturelle, Paris; ORSTOM, Bondy; ENGREF, Nancy; INRIA, Grenoble; URA CNRS 243 Université de Lyon, Villeurbanne; ORSTOM, Montpellier, INRA, Toulouse, 24 juin-5 juillet 1991.*

Fortuner, R. & Ahmadi, A., 1986. *NEMAID 2.0. Computer program for identification of Nematodes. User's manual.* California Department of Food and Agriculture. 49 pages.

Fortuner, R. & Couturier, G., 1983. Les nématodes parasites de plantes de la forêt de Taï (Côte d'Ivoire). *Revue Nématol.*, 6(1): 3-10.

Fortuner, R. & Maggenti, A.R., 1991. A statistical approach to the objective differentiation of *Hirschmanniella oryzae* from H. belli (Nemata: Pratylenchidae). *Revue Nématol.* 14(1): 165-180.

Fortuner, R., Maggenti, A.R., & Whittaker, L.M., 1984. Morphometrical variability in *Helicotylenchus* Steiner, 1945. 4. Study of field populations of *H. pseudorobustus* and related species. *Revue Nématol.*, 7(2): 121-135.

Fortuner, R. & Merny, G., 1974. Les nématodes parasites des racines associés au riz en Basse-Casamance (Sénégal) et en Gambie. *Cah. ORSTOM, sér. Biol.*, No. 21 (1973): 3-20.

Fortuner, R., Merny, G. & Roux, C., 1981. Morphometrical variability in *Helicotylenchus* Steiner, 1945. 3. Observations on African populations of *Helicotylenchus dihystera* and considerations on related species. *Revue Nématol.*, 4(2): 235-260.

Fortuner, R. & Wong, Y., 1983. *NEMAID. Computer program for identification of nematodes. User's manual.* Publ. No. 640, California Department of Food and Agriculture. 44 pages.

Fortuner, R. & Wong, Y., 1985. Review of the genus *Helicotylenchus* Steiner, 1945. 1. A computer program for identification of the species. *Revue Nématol.*, 7(4): 385-392.

Autres références

Andrássy, I., 1985. A dozen new nematode species from Hungary. *Opusc. zool. Bdpest* 19/20: 3-39.

Diederich, J., Fortuner, R. & Milton, J., 1989. Building a knowledge base for plant-parasitic nematodes: description and specification of metadata. In: Fortuner, R. (Ed), *Nematode identification and expert-system technology.* New York, Plenum Publishing Corp.: 65-76.

Diederich, J., Fortuner, R. & Milton, J, 1990. NEMISYS, an expert workstation for nematode identification. *Communication, Second International Nematology Congress, 11-17 August 1990, Veldhoven, The Netherlands.*

Diederich, J. & Milton, J., (sous presse, a). Expert workstations: a tool-based approach. In: Fortuner, R. (Ed.), *Advances in computer methods for systematic biology: artificial intelligence, databases, computer vision.* Baltimore, Maryland, The Johns Hopkins University Press.

Diederich, J. & Milton, J., (sous presse, b). NEMISYS, a computer perspective. In: Fortuner, R. (Ed.), *Advances in computer methods for systematic biology: artificial intelligence, databases, computer vision.* Baltimore, Maryland, The Johns Hopkins University Press.

Dreyfus, H. L. & Dreyfus, S. E., 1986. *Mind over machine.* New York, The Free Press, MacMillan, Inc.

Eisenback, J. D., 1989. Identification of meloidogynids. In: Fortuner, R. (Ed), *Nematode identification and expert-system technology.* New York, Plenum Publishing Corp.: 123-137.

Fisher, R. A., 1936. The use of multiple measurements in taxonomic problems. *Annls Eugenics* 7: 179-188.

Gower. J. C., 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27: 857-871.

Lebbe, J., 1984. *Manuel d'utilisation du logiciel XPER.* Paris, Micro Applications.

Lewis, S. A. & Golden, A. M., 1981. Description of *Trilineellus clathrocutis* n.g., n.sp. (Tylenchorhynchinae: Tylenchida Thorne, 1949) with a key to species and observations on *Tylenchorhynchus* sensu stricto. *J. Nematol.* 13: 135-141.

Luc, M., 1986. *Hoplorhynchus* Andrássy, 1985, a junior synonym of *Pratylenchoides* Winslow, 1958 (Nemata: Pratylenchidae). *Revue Nématol.* 9: 198.

Luc, M., Baldwin, J. G. & Bell, A. H., 1986. *Pratylenchus morettoi* n.sp. (Nemata: Pratylenchidae). *Revue Nématol.* 9: 119-123.

Pankhurst, R. J. (sous presse). Principles and Problems of identification. In: Fortuner, R. (Ed.), *Advances in computer methods for systematic biology: artificial intelligence, databases, computer vision.* Baltimore, Maryland, The Johns Hopkins University Press.

Roggen, D. R. & Asselberg, R., 1971. The use of ratios in nematology. *Nematologica* 17: 187-189.

Roggen, D. R., Revets, S & Van den Berghe, W., 1987. Using ratios. *Nematologica* 32: 398-407.

Rypka, E. W., 1975. Pattern recognition and microbial identification. In: Pankhurst, R. J. (Ed.) *Biological identification with computers.* London, New York, San Francisco: Academic Press: 153-180.

Sternberg, P. W. & Horvitz, R. H., 1982. Gonadal cell lineage of the nematode *Panagrellus redivivus* and implications for the evolution by modification of the cell lineage. *Developm. Biol.* 88: 147-168.

Zerwekh, R. (sous presse). Information processing with neural networks. In: *Advances in computer methods for systematic biology: artificial intelligence, databases, computer vision.* The Johns Hopkins University Press.