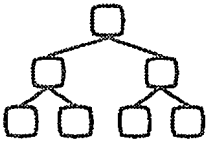
UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS
DIVISION DE ESTUDIOS DE POSGRADO

METODOS EN LA IDENTIFICACION BIOLOGICA AUTOMATIZADA



TESIS QUE PARA OBTENER EL GRADO ACADEMICO DE

MAESTRO EN CIENCIAS (BIOLOGIA) PRESENTA

MIGUEL MURGUIA ROMERO

MEXICO, D.F. 1992

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

Contenido

Capítulo I

Introducción

La diversidad del mundo vivo es grande, en la actualidad se conocen más de tres millones de especies y quizá existan cerca de diez millones (Margulis y Schwartz, 1981).

Cuando el hombre desea conocer algo de la realidad, después de la observación intenta clasificar los fenómenos u objetos observados. La clasificación es un acto que establece orden a los hechos observados.

Es tarea del taxónomo proponer sistemas de clasificación para los organismos, que permitan establecer cierto orden en el conocimiento de la diversidad del mundo vivo, con criterios cada vez más universales. A medida que se obtiene más información de los organismos, los sistemas de clasificación se modifican, intentando incorporar mas conocimientos dentro de las clasificaciones.

Al construir clasificaciones más complejas y completas, los procesos de identificación de los organismos, también se modifican.

Es importante indicar que un sistema de clasificación en el que los objetos clasificados sean difíciles de ubicar, es poco útil. Al crear una clasificación se debe estar consciente de que el hecho implícito posterior de diagnosis esté intimamente relacionado con la estructura de esa clasificación.

Por eso, si se generan estrategias y metodologías de identificación o diagnóstico más eficientes, el proceso de clasificación será más sencillo, pues al crear las clasificaciones, no se tendrá que estar observando con detalle cómo repercutirá el nuevo esquema de clasificación en el proceso de identificación de los objetos.

I. 1) Alcances

El presente trabajo tiene como antecedentes principales los trabajos sobre la creación de una policlave para familias de plantas con flores de México, presentados en Murguía (1987) y en Villaseñor y Murguía (1992). Uno de los productos finales de estos trabajos es la policlave FAMEX, además de generar una metodología para la automatización de la identificación biológica que contempla la recopilación de datos, su verificación y la implementación de algoritmos operables en una computadora.

En este trabajo se analizan y ordenan los diferentes enfoques utilizados para abordar el problema de la automatización de la identificación biológica. Algunos de los enfoques presentados no se han llevado a la práctica en el campo de la taxonomía biológica, pero han demostrado su eficiencia en procesos de diagnóstico análogos a los de la taxonomía biológica.

Aunque existen ya textos en donde se intenta recopilar las tecnologías para la identificación biológica automatizada, como el de Pankhurst (1975), es importante recapitular las nuevas metodologías surgidas de las Ciencias de la Computación.

Este trabajo no pretende agotar y exponer todas las metodologías de la identificación automatizada que en la actualidad se han generado, principalmente alrededor de la robótica, sino más bien extraer las que se observen más prometedoras o adecuadas para la taxonomía biológica. En este trabajo sólo se consideran características cualitativas, pues el tratamiento de las cuantitativas es diferente. Algunas de las características cuantitativas pueden tratarse como cualitativas, haciendo intervalos, aunque esta solución puede no ser la más adecuada para algunos casos.

Aunque muchos de los métodos de identificación presentados también son aplicables al proceso de la clasificación taxonómica y con enfoques que seguramente enriquecen a los métodos actuales de la Taxonomía Numérica incluyendo a la Cladística-, no se hace explícita su utilidad. La modificación de los esquemas de clasificación a partir de información de la identificación es un tema interesante no tratado en este trabajo, al que puede introducirse mediante la lectura de Booker et al. (1990).

I.2) Objetivos

* Presentar un esquema de clasificación de las metodologías de identificación biológica automatizada.

* Explicar los principales procedimientos que cada clase de metodología aplica con la idea de facilitar posteriores implementaciones en computadora.

* Exponer un nuevo modelo formal que permita su análisis, crítica y modificaciones o adiciones.

* Presentar una implementación en computadora del modelo y un caso de aplicación para un grupo taxonómico.

Capítulo II

Objetivos de la Identificación Biológica

La identificación biológica tiene como principal objetivo hacer operativo al sistema de clasificación elegido para ordenar a los seres vivos. La mayoría de las clasificaciones biológicas, en la actualidad, desean reflejar a las relaciones evolutivas entre las especies, razón por lo que los atributos que definen a las especies no necesariamente están presentes en los individuos como unidad a identificar. Esta situación lleva a la necesidad de utilizar, durante la identificación, características no consideradas, o con una ponderación menos útil, durante el proceso de clasificación.

El problema principal surge del hecho de que casi siempre las unidades estudiadas directamente por los taxónomos son los individuos, y pocas veces son organismos, poblaciones o especies. El taxónomo colecta individuos, no organismos, ni poblaciones y menos especies.

Desde este punto de vista, la identificación biológica desempeña un papel fundamental en el desarrollo de las ideas en Biología, pues constituye una confrontación de las teorías con la realidad observada.

II. 1) Definición de Identificación Biológica

Debido a la diversidad de significados que para los científicos tiene la clasificación en Biología, la identificación biológica tiene diferentes enfoques. A continuación se brindan diferentes definiciones de identificación biológica encontradas en la bibliografía:

[Crisci, 1983]
Ubicación de un objeto no identificado en la clase o grupo al que corresponde, conforme a una clasificación construida previamente.

[Font-Quer, 1985]
Acto de reconocer la familia, género, especie, etc. en que se clasificó a un organismo.

[Murguía, 1992]
Proceso en el que los objetos (ejemplares) se hacen corresponder con una clasificación preestablecida.

[Pankhurst, 1978]
Encontrar el nombre para un espécimen de animal o planta.

[Radford, et al., 1974]
Acto de reconocer o establecer el taxon al que pertenece un ejemplar.

[Walter, 1975]
La práctica de asignar a los especímenes, nombres de taxa.

Muchas de las definiciones tienen inmersa la palabra "clasificacion" y todas presuponen, de manera implícita o explícita, la existencia de una clasificación. A continuación se analizan algunas de las características de las definiciones mencionadas:

- Crisci (1983) es tautológico al incorporar la palabra identificación en su definición.

- Las definiciones de Pankhurst (1978) y Walter (1975) hablan de "encontrar el nombre", acto que parece implicar que la nomenclatura tiene que ver mas que la clasificación con la identificación.

- Pankhurst presupone una clasificacion del mundo vivo en animales y plantas.

- Font-Quer (1985) asume la jerarquía Linneana como la única manera de clasificar.

- Radford (1973) asume al taxon como la unidad de clasificación.

En esta tesis, el significado de identificación que se asume es:

> **"Proceso en el que los objetos se hacen corresponder con una clasificación preestablecida."**

Esta definición es práctica y tiene las siguientes características:

- No contiene la palabra "nomenclatura". La identificación está relacionada sólo indirectamente con la nomenclatura mediante una clasificación.
- No presupone alguna estructura para la clasificación.
- Se habla de una "correspondencia" con una clasificación, lo que no limita a la identificación a un proceso de pertenencia.

La identificación en Biología debe concebirse como un proceso dinamico, en el que los objetos y las relaciones con los que opera cambian no sólo en el tiempo, sino en mas dimensiones, como por ejemplo el tipo de clasificación considerada, el nivel taxonómico de identificación deseado, el grupo de organismos o hasta el mismo propósito que se persiga de la identificación. Las mismas clasificaciones son dinámicas, pues aunque uno de los objetivos de la sistematica biológica es crear clasificaciones robustas; es casi imposible encontrar clasificaciones que cambien poco con el tiempo, ya sea debido al descubrimiento de nuevos taxa o a la incorporación de nuevas evidencias que obligan a realizar ajustes a la clasificación existente.

La identificacion biológica es un paso posterior a la clasificación. Sin una clasificación taxonómica la identificación biológica no es posible. Sin embargo, por ser un proceso estrechamente relacionado y posterior a la clasificacion, la identificación biologica es el lugar donde se pueden encontrar elementos para retroalimentar al proceso de clasificacion: sus criterios, sus bases filosóficas, sus métodos y su aportación practica-metodológica en el quehacer de los biologos u otro tipo de agrupaciones humanas.

II. 2) Marco Histórico

A continuación se enumeran algunos de los hechos históricos importantes en el desarrollo de los métodos de identificación. Aunque la búsqueda continúe y se intenten encontrar fechas más precisas y trabajos aún no conocidos, estos hechos dan una visión global de los acontecimientos ocurridos en la historia de la identificación biológica. La mayoría han sido extraídos de Pankhurst (1978).

Siglo IV A.C. Aristóteles escribe su "Teoría de las Plantas", realiza investigaciones botánicas, clasifica y describe minerales, animales y vegetales al igual que Teofrasto.

c.a. 1500. Se da a conocer al "Viejo Mundo" algo sobre los sistemas de clasificación en América (v.g. el Códice Badiano).

1672. Morison publica una monografía sobre Umbelliferae, en donde se incluyen dibujos de frutos y diagramas para auxiliar en la identificación.

1686. Grew publica, en su "Historia Piscicum", un diagrama dicotómico para mostrar la clasificación de peces cartilaginosos.

1736. Linneo introduce el termino "clave".

1778. Lamarck introduce el uso explícito de las claves dicotómicas para identificar.

1943. Aparece la idea de "clave tabular" o "clave sinóptica".

1968. Aparece el primer programa en computadora para identificación.

1969. Duke introduce el término "policlave"

Respecto al trabajo que Aristóteles realizó, se conoce solo una fracción, pues de los cuatrocientos libros de que constaba su obra, según Diogenes, se conocen sólo unos cuantos. En términos generales se puede decir que el trabajo de Aristóteles en la descripción y clasificación de los seres vivos, constituyó "el ejemplo" durante casi dos milenios.

En cuanto a los criterios de clasificación e identificación de seres vivos en América precolombina, se puede decir que existía una nomenclatura pictográfica, como así lo indica Francisco del Paso y Troncoso (López, 1975). En sus códices, los antiguos mexicanos utilizaban glifos específicos para representar a cada clase de organismo. Quizá fueron las características representadas en esos glifos uno de los atributos útiles para su identificación. Por la naturaleza de sus clasificaciones, algo más ecológicas que las actuales, es muy probable que otras características útiles en su identificación hayan sido atributos del ambiente en que se colectaba el organismo, también olores, sabores y diversas reacciones químicas y fisiológicas pudieron servir como atributos útiles en la identificación de plantas, animales y minerales.

Las claves, como herramienta para identificación en Biología, pasaron por etapas primitivas en las que su principal objetivo era mostrar la clasificación de los grupos, atendiendo poco a su uso como método de identificación.

A más de dos siglos de su creación, las claves dicotómicas constituyen la principal herramienta para identificar. Es a mediados de este siglo cuando se reconoce la necesidad de construir y utilizar métodos de identificación que permitan manipular más relaciones que las que permiten las claves dicotómicas: tal es el motivo del término "policlave", como indicando que el "poli" se contrapone al "di".

Aunque en general se reconocen las ventajas de las policlaves contra las claves dicotómicas, las primeras son utilizadas en menos del 1% de los casos. Sin duda, las claves dicotómicas impresas en papel, serán reemplazadas por métodos más eficientes y quedarán relegadas a un uso menor, pero eso sucederá dentro de varios años o décadas.

En México, la práctica de la identificación biológica automatizada empieza a desarrollarse, pues la incorporación de la computadora en el trabajo diario de los taxónomos es reciente. Algunas instituciones de tamaño mediano y grande cuentan con acceso a computadoras desde hace más de una década, pero la mayoría de los investigadores han tenido contacto directo con ellas, apenas hace unos cinco años.

II. 3) Clasificación, Nomenclatura e Identificación

Los criterios para ordenar a los seres vivos se establecen principalmente mediante un sistema de clasificación, en donde el concepto de especie desempeña un papel fundamental. Los

criterios para establecer unidades naturales y las mismas unidades naturales así consideradas, constituyen conceptos de orden.

Gonzalez (1991) habla del concepto Individuo-Organismo-Población-Especie (IOPE) como "una aproximación a la unidad teórica de la Biología" y destaca la importancia de transformar a las unidades continuas, ontológicamente, en unidades discretas de conocimiento. Dentro del contexto de su Teoría de los Procesos Alterados, observa a la especie como un "proceso alterado". Dentro de la teoría de los procesos alterados, el adjetivo "alterado" hace énfasis en que la observación de un proceso constituye, por sí misma, una alteración a dicho proceso.

En la actualidad, las clasificaciones están acotadas por los límites de la identificación biológica, entre otros factores. El concepto de especie que pueda prevalecer al construir una clasificación, es elegido tomando en cuenta la facilidad de identificar a las unidades clasificadas. Por ejemplo, al proponer como unidad de clasificación a procesos alterados (IOPE), es obvio que los medios de identificacion deben adaptarse para brindar el servicio adecuado a las clasificaciones propuestas, que tomen en cuenta los tipos de atributos que definen a las unidades a clasificar; además, muchos deben ser atributos que varían con el tiempo, es decir dinámicos y no estáticos. Aunque se reconoce que muchos de los atributos que definen a las unidades de clasificación las especies- son dinámicos por naturaleza, esto no se refleja ampliamente en los métodos de identificación.

El nombre de las clases es un nexo entre los procesos de identificación y las clasificaciones. Si el nombre de las clases contiene información de las clases, la identificación tiene otro punto de contacto a su favor con las clasificaciones.

Es importante establecer la metodología para nombrar a los grupos de objetos en un sistema de clasificación. En la nomenclatura se debe expresar la clasificación. Si los nombres asignados a las clases expresan atributos de ellas, entonces la clasificación y la nomenclatura tendrán estructuras más estables y útiles a la ciencia que si operan de manera aislada.

En un diálogo de Platón, Sócrates habla del significado y la relación del nombre de las cosas. El nombre de las cosas contiene un significado, en ocasiones parte de ese significado es la clase a la que pertenece en determinado sistema de clasificación.

Francisco del Paso y Troncoso, al encontrar que los pueblos precolombinos tenían un sistema de clasificación, que se reflejaba en la nomenclatura de los vegetales y animales,

intenta a toda costa hacerlo corresponder con la jerarquía Linneana y su nomenclatura. Así, ejemplifica como el vocablo "etl" era equivalente al nivel genérico de la jerarquía linneana, en particular para designar a las plantas del género Phaseolus. Además hace notar que si bien no todos los vocablos nahuas que nombran clases de plantas contienen sólo dos raíces, si se empeñaban en poner nombres con un significado natural acerca de la planta, no como en la nomenclatura actual que permite nombres totalmente ajenos a la naturaleza de las clases. Otro ejemplo son los amates o plantas del género Ficus: el **amacuáhuitl** (de **amatl** papel y **cuáhuitl** arbol). Distinguian el tlilámatl (de **tlilic** negro), hoy amate prieto (Ficus cotinifolia); **texcalámatl**, (de **texcalli**, piedra volcánica, lava) o sea amate de los pedregales de lava (Ficus petiolaris).

II. 4) Metodologías tradicionales

Las metodologías para identificación durante los años 50s a 80s y aún en la actualidad, se auxilian mucho de manuales de identificación. Entre los más populares están los de la serie "How to know..." de WM Company (v.g. Ehrlich y Ehrlich, 1961).

Las claves de Casas y McCoy (1979) para reptiles y anfibios de México, presentan esquemas intercalados dentro de las claves, lo que facilita la identificacion, al poder observar gráficamente lo que la clave indica.

Otra representacion muy recurrida son las tablas sinópticas, en donde se grafican los atributos contra las clases. En CIBA-GEIGY (1979) se presentan tablas para elección de herbicidas y para enfermedades del maíz, auxiliándose de fotografías en color que ilustran los síntomas más conspicuos. El auxilio de fotografías es muy recurrido en obras de tipo divulgación-artísticas, como los trabajos de Gilbert (1988) y Graf (1986), en donde las ilustraciones "dan vida" a las obras. Phillips (1989) muestra excelentes fotografías de las hojas de varios árboles que auxilian considerablemente en la identificación. Graf (1986) utiliza los colores y tonalidades de partes vegetativas y reproductivas que también auxilian mucho en la identificacion de "plantas exóticas".

Son las claves dicotómicas la herramienta más recurrida en la actualidad para identificar organismos. Las claves de Smith y Taylor (1966) para anfibios y reptiles de México, siguen siendo una herramienta diaria de trabajo para los herpetólogos. Las claves de Smith (1950) para algas de agua dulce, también son aún insustituibles. En Limnología, el texto de Edmondson (1963), que contiene claves y esquemas de organismos invertebrados de aguas dulces, es aún una herramienta muy útil. En Entomología, el

texto de Ross (1978) es útil para una identificacion a nivel de familia. Las claves de Davis y Cullen (1979) para familias de plantas con flores, también son frecuentemente utilizadas. Para hongos de importancia económica, las claves de Guzmán (1977) brindan una ayuda considerable.

II. 5) Elementos de un Sistema de Identificación Biológica

La identificación biológica es un sistema en el que intervienen diversos objetos o procedimientos (fig. 1):

a) **Ejemplar bajo identificación.** Es el objeto al que se le tiene que asignar el taxon al que pertenece. Puede estar completo o incompleto.

b) **Sistema de Clasificación.** Es el grupo de taxa en que se divide el universo considerado por el sistema. Puede estar estructurado por uno o más niveles. En este elemento estan el conjunto de características de los taxa.

c) **Nivel y Grupo taxonómico de partida.** Es el nivel taxonómico más alto en que se divide al sistema, v.g. Clase, familia o género.

d) **Nivel taxonómico de identificación.** Es la precisión con que el sistema diagnosticara. Como límite se tiene al nivel taxonómico más inferior considerado.

e) **Taxa candidatos.** Son los taxa a los que puede pertenecer el ejemplar bajo identificación, considerando un subconjunto del Sistema de Clasificación.

f) **Métodos y herramientas de Identificación.** Son las estrategias para realizar el diagnóstico. Tienen que ver con la estructura de datos en que se representa a la información del sistema de clasificación.

g) **Identificador.** Es la persona que identifica, es decir, el usuario.

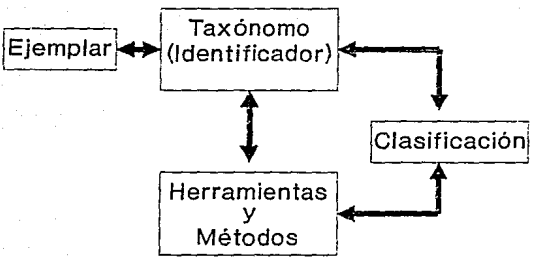


Fig. 1: Elementos de un sistema de identificación biológica

Desde el punto de vista de sistemas computacionales, Booker et al. (1990) proponen los siguientes componentes para los sistemas de clasificación (i.e. identificación):

1) **Interfase de entrada:** Incorpora mensajes al sistema

2) **Clasificador:** Contiene los criterios para identificar.

3) **Lista de Mensajes:** Contiene todos los mensajes introducidos mediante la interfase de entrada, más aquellos deducidos por el propio sistema (en el caso de que cuente con una maquina de inferencias).

4) **Interfase de salida:** Traduce algunos de los mensajes en acciones que modifican el estado del medio.

En este esquema, los mensajes desempeñan el papel de las caracteristicas del ejemplar. La interfase de salida, para el caso de identificadores taxonomicos, modifica los criterios o puntos de vista del usuario respecto al ejemplar bajo identificación. En otros tipos de identificadores, el sistema tiene la capacidad de modificar al objeto en identificación.

Capítulo VIII

Formalización de la Identificación Biológica

Toda formalización que se intente hacer de la identificación biológica tendrá que revisar los criterios de la clasificación biológica. La naturaleza de la clasificación biológica influye en los métodos de la identificación biológica. La identificación con un sistema de clasificación construido con criterios estadísticos es diferente a la identificación con sistemas construidos con criterios deductivo-lógicos.

Los límites de la identificación provienen o están influenciados por los criterios de clasificación; estos criterios son de diferente naturaleza, por ejemplo criterios de orden, observabilidad, unidad y criterios de relación entre esos criterios. Las bases en que se fundamenta un proceso de clasificación es importante para establecer formal y detalladamente el proceso de identificación. Sin un modelo estructurado del conocimiento, la identificación biológica corre peligro de caer en terreno de lo puro afectivo y emocional.

Contestar ¿por qué un individuo o pedazo de él pertenece a una determinada clase? es tarea que requiere fundamentos lógico-metodológicos bien establecidos. ¿ Un individuo tetrápodo, blanco, con rayas es una cebra porque es claro que no es un caballo, pero se parece mucho a él? ¿La pertenencia a una clase se justifica por el hecho de la no pertenencia a otra clase?. Decir cuales son las razones por las que un individuo pertenece a una clase requiere de una estructuración del conocimiento, que debe estar expresada en la clasificación biológica.

III. 1) Sistemas Formales

La ciencia es una de las más valiosas herencias de la humanidad. Para muchos, el estado de la ciencia representa el estado cultural de los pueblos.

La formalización del conocimiento es un camino que los científicos han buscado seriamente a través de la historia. El científico se apoya, cada vez con mayor frecuencia, en ecuaciones matemáticas, modelos y otro tipo de abstracciones para representar los fenómenos que ocurren en la naturaleza. En la medida en que nuestro sistema de comunicación sea más riguroso y congruente, los conocimientos así representados serán más fáciles de analizar, de encontrar errores e incluso, descubrir nuevos fenómenos. Aunque la ciencia se auxilia muchas veces de la intuición total y de técnicas a menudo artesanales, es en el camino de la formalización en donde se pueden encontrar terrenos más sólidos que pisar. Cabe aclarar que la formalización no es garantía de la acumulación de conocimientos verdaderos o correctos.

III. 2) Lógica Matemática y Demostración Científica

La ciencia se fundamenta en un sistema lógico, que le brinda estrategias y criterios para validar la incorporación de nuevos conocimientos.

Aristóteles estableció y estudió la Lógica como formalismo, para poder tener mayor seguridad de lo que se cree que son las cosas o fenómenos. La Lógica es un instrumento orgánico para apreciar o evaluar la certeza del razonamiento (Copi, 1979). Aristóteles también estudió la demostración como proceso lógico para dar validez a los descubrimientos. Así, en la demostración se pueden deducir cosas, fundamentando los antecedentes y la conexión que entre ellos existe. En el descubrimiento científico, para pasar de premisas a conclusiones, debe existir siempre algo que lo justifique. Como ejemplo, en Lógica se conoce el "Modus Ponens". A continuación se muestra un ejemplo donde se utiliza esta regla:

```
1)   p -> q     Proposición
2)   p          Hecho
     ------
3)   q          Deducción, de 1) y 2),
                por "Modus Ponens"
```

es decir, si existe una proposición p -> q que lee "si p entonces q" y además se conoce como hecho a p, entonces, aplicando "Modus Ponens", q también es un hecho. Esta regla general, es un formalismo que permite la acumulación de conocimientos. A veces la ciencia no tiene aún elementos formales para poder descubrir; en esos casos se justifica el descubrimiento intuitivo, aunque por desgracia, no comprobado. Pero en el caso de existir modelos formales para la representación de los conocimientos y su descubrimiento, no es

válido intuir, pues se considera más apropiado utilizar aquella metodología que permita su comprobación. Al no utilizar mecanismos formales, se corre mayor riesgo de desviar los resultados. En la fig. 2 se intenta representar el "hilo" de los métodos formales para el descubrimiento y cómo el separarse de ellos, al utilizar intuición, puede generar conclusiones erróneas. Al respecto, Szilasi (1945) menciona que "cuando no formamos de antemano un sistema, sino que tenemos que ir avanzando, poco a poco, con nuestro sistema de comprensión, entre las sombras de lo no creado por nosotros, las cosas no son sencillas". La reflexión anterior se refiere a que existe una continuidad en los "proyectos de comprensión", los cuales se van sucediendo unos a otros; los proyectos de comprensión surgen o se fundamentan en proyectos de comprensión precedentes. La fig. 2 es sólo una "caricatura", pues aún utilizando formalización puede llegarse a conclusiones erróneas, como anteriormente ya se indicó en **III.1.**

Actualmente, en Informática, se siguen generando metodologías para el diseño, desarrollo, implementación y hasta uso de los sistemas computacionales. Estas metodologías se crean mediante formalismos matemáticos y constituyen verdaderas teorías. El hecho de que existan formalismos creados o descubiertos mediante deducciones demostradas, justifica el que, para construir y utilizar sistemas computacionales, sea conveniente contar con ciertos elementos teóricos y no sólo con conocimientos totalmente prácticos y poco teóricos. Además como Bachelard (1982) señala, lo unitario y pragmático puede representar un obstáculo para el conocimiento científico.

Es importante recalcar que cualquier método que se implemente para la automatización de la identificación biológica, debe o se desearía que estuviera formalmente planteado, pues esto permitirá conocer más acerca de sus propiedades, sus debilidades y los casos en que es adecuado. Es decir, se debe tender a una "Identificación Científicamente Demostrable", a una Identificación realizada con el formalismo y rigor más extremosos.

III. 3) El Biólogo en la generación de Software

El biólogo debe intervenir en la construcción de los sistemas que él utilizará. Por supuesto que lo anterior se refiere a programas de computadora que se tienen que construir, porque no existen aplicaciones que se parezcan lo suficiente a las que el biólogo necesita, y por lo tanto, es difícil encontrar que los sistemas preconstruidos respondan a sus necesidades. Por otra parte, el grado de especialización tiene límites: por ejemplo, pensar en construir un procesador de

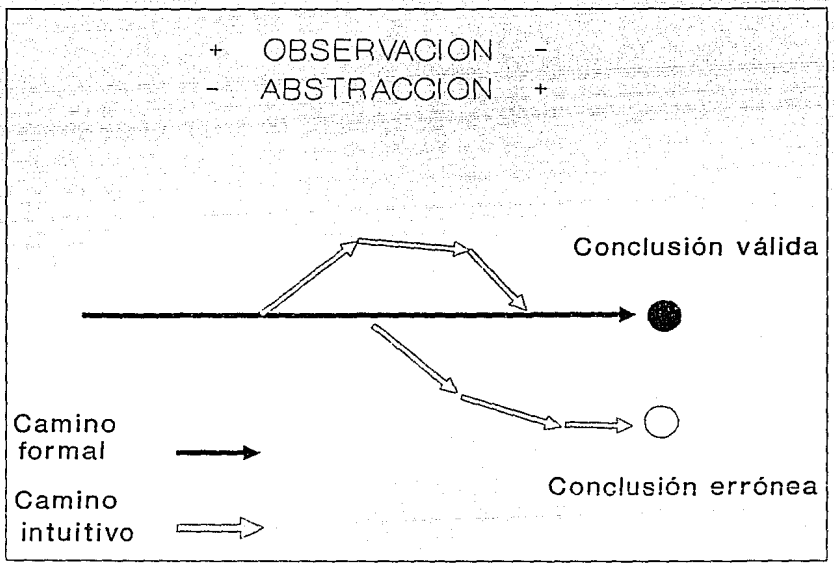


Fig. 2: Hilo de los Métodos Formales y de los Intuitivos en el descubrimiento

textos para biólogos, es menos frecuente que pensar en programas para identificación taxonomica. El biólogo es el que conoce sus necesidades, más que el informatico; por eso es necesario una interacción entre ambos. De hecho, esta comunicación biólogo-informático es tan necesaria como la computadora misma, pues las automatizaciones que se construyen por un equipo de biólogos quizá tengan muchos errores, pero las que se realizan entre biólogos e informáticos con conocimientos nulos de Biología, difícilmente llegan a concluirse.

En la figura 3 se muestra la relación que guarda el biólogo (usuario) con la construcción de los sistemas que utiliza. En este proceso se observan tres elementos básicos (Usuario, Diseño y Desarrollo) que originan otro tipo de relaciones u objetos. Los usuarios especifican sus necesidades al informático, quien se encarga de diseñar el sistema; el diseñador especifica a los desarrolladores lo que tienen que programar; los programadores (desarrolladores) lo brindan al usuario. Al centro de estos tres objetos se encuentra la generación de un cuarto objeto: el producto (sistema o programa de computadora). Entre el usuario y el diseñador se puede generar otro concepto: Requerimientos de Software, que se refiere a las metodologías utilizadas para establecer la comunicación entre estos dos objetos. Aquí también se sitúa el mundo de las actualizaciones o "up-date", si es que se están realizando modificaciones a un sistema o programa preexistente. Entre Diseño y Desarrollo, se encuentra la Ingeniería de Software, con sus metodologías de programación, lenguajes de programación y diversos tipos de herramientas. Entre el Desarrollo y el Usuario se sitúa el mundo de las especificaciones técnicas (manuales, requerimientos de equipo y memoria e instalación, entre otros elementos).

Esta visión simplificada del proceso de construcción de "software de ayuda al biólogo" muestra como, ademas de considerar la dirección Software -> Biólogo, también se debe tener en cuenta la dirección Biólogo -> Software, que es menos común encontrarla entre los desarrollos de software que utiliza el biólogo. Es decir, es mas comun encontrar sólo adaptaciones, pero existen muchos procesos que el biólogo puede automatizar y que por desgracia, no existen análogos en otras actividades del hombre. El software para inferencia evolutiva es una excepción, pues ahora empieza a desarrollarse en cantidades más grandes y con más formalidad.

Para situar la cantidad de conocimientos que deseablemente deben tener los biólogos, obsérvese la tabla I, en donde se cataloga, en forma discreta, a los niveles de cultura informática. Respecto a esa clasificación, el nivel adecuado para el biologo es el IV. En este nivel se tienen elementos para ser independiente moderadamente, pues sin auxilio de otras personas se pueden resolver alrededor del 80% de las tareas automatizables. Ademas, permite una interaccion con informaticos

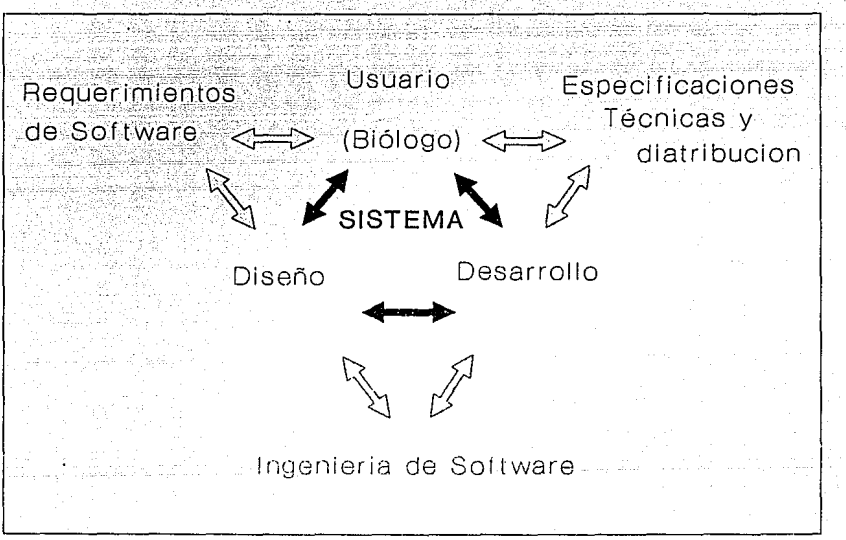


Fig. 3: Relación entre el Biólogo y el Desarrollo de Sistema a su servicio.

que se sitúan en niveles superiores de cultura informática y que son capaces de concretizar los deseos del biólogo.

VII	**Análisis:** Capacidad para conceptualizar soluciones y concretarlas mediante una estrategia de solución e implementarlas en un sistema de cómputo. Conocimiento y dominio de la Teoría de Cómputo y de Algoritmos y su relación con la solucion de problemas.
VI	**Programación:** Conocimiento y dominio de varios lenguajes de programación y capacidad para decidir cuál es el adecuado para cada tipo de problema que se presenta. Conocimiento de la interrelación práctica y teórica entre las diversas herramientas de programación, así como la Ingeniería de Software.
V	**Pequeña programación:** Conocimiento y dominio de uno o dos lenguajes de programacion. Formación en teorías matemáticas de la Informática y teoría de algoritmos, pero no en Ingeniería de Software o de Computación. Se entienden las necesidades de los usuarios y hay capacidad para desarrollar pequeños sistemas que resuelvan problemas específicos.
IV	**Nivel Usuario:** Captura de datos. Capacidad para manipular uno o varios paquetes, como un procesador de textos, una hoja de cálculo, un manejador de bases de datos o un programa de taxonomía o análisis evolutivo. Programación de algunas macros o aplicaciones dentro de los paquetes.
III	**Visual-Recepción:** Desenvolvimiento dentro de un ambiente en que se utilizan las computadoras para diversos procesos. Uso de listados u otro tipo de producto directo de un sistema de cómputo, como gráficas, fichas o reportes.
II	**Audición:** Se desarrolla en un ambiente de cultura informática medio o alto; lectura y comprensión de libros y revistas de computación. Quizá se tienen manuales al alcance pero no a la computadora, en la que no se practica.
I	**Analfabetismo:** Conocimientos nulos de Computación. Ideas equivocas y a veces ficticias de lo que son las computadoras en la actualidad. En ocasiones es extremista: o se piensa que las computadoras no sirven para nada o que lo pueden todo.

Tabla I. Clasificación de la cultura informática.

III. 4) La Identificación en el Contexto Taxonómico

Como se definió en la sección II.2 la determinación o identificación taxonómica es un proceso en el que los objetos se hacen corresponder con una clasificación preestablecida. Para entender el proceso de identificación, debe comprenderse y definirse con exactitud el proceso de clasificación. Los sistemas de clasificación son fundamentales para la comunicación entre investigadores.

Las descripciones taxonómicas de los taxa son un medio para enlazarlos con los sistemas de clasificación. Pero el manejo automático de las descripciones, que generalmente se expresan en lenguaje natural, es escaso en la actualidad. Por esta razón, se crea el concepto de "clave de identificación", que auxilia metódicamente a establecer la relación ejemplares-clasificación. Las claves de identificación son una expresión estructurada de las descripciones taxonómicas y dicha estructura es susceptible de automatización.

En la siguiente sección se presenta un modelo para representar la relación entre estos elementos: ejemplares, descripciones taxonómicas, clasificación e identificación. La representación es sólo un punto de vista para auxiliar en la metodología de la identificación taxonómica, aunque es adecuada para la automatización de otros procesos relacionados, como la generación por computadora de descripciones taxonómicas. Aunque aquí no se hablara de la generación automática de descripciones taxonómicas, los fundamentos para entender estos dos procesos son similares. Un ejemplo de la generación automática de descripciones taxonómicas se presenta en Murguía (1990c).

III. 5) Teoría de Gráficas para ubicar a la Identificación Biológica en el Contexto Taxonómico

Para poder automatizar procesos, es necesario contar con un sistema de representación de los objetos y las relaciones entre ellos; se debe tener un sistema de símbolos con los que se pueda operar y definir funciones.

A continuación se presenta un enfoque Popperiano del proceso de construcción de teorías en el descubrimiento científico en Biología. El enfoque desde el que se abordará el estudio es la Teoría de Gráficas y algunos conceptos se presentan de manera no totalmente formal. Para una mayor información acerca de la representación del conocimiento mediante gráficas, consúltese a Oliver y González (1979) o a González (1984). En la representación intervienen objetos y

relaciones. Las relaciones pueden ser directas e indirectas. Los objetos (o nodos) se encadenan en una secuencia lógica y congruente mediante las relaciones. Las relaciones pueden formar áreas básicas, que normalmente representan la síntesis de nuevos conceptos que surgen a partir de las relaciones definidas entre los objetos.

Entendido el sistema de representación mediante gráficas, a continuación se propone una representación, mediante este formalismo, para la metodología en los estudios Taxonómicos y Evolutivos.

En la figura 4 se muestra una gráfica (adaptada de Salazar, 1979), que explica, en términos generales, el proceso de acumulación del conocimiento en la ciencia. En esta gráfica intervienen seis nodos:

I Ejemplares
II Descripciones Taxonómicas
III Clasificación
IV Relaciones Evol. Directas
V Arbol Evolutivo o Relaciones Evolutivas Indirectas
VI Teorías y Tendencias Evolutivas del Grupo

A continuación se describe el proceso que se representa en la figura 4. El investigador colecta ejemplares y los describe. Después tiene que ubicarlos en un sistema de clasificación, identificandolos o clasificándolos si constituyen una entidad taxonómica nueva para la ciencia. Establecida una clasificación, se pueden encontrar relaciones directas entre las clases que la conforman, relaciones como semejanzas o diferencias. Detectadas las relaciones directas entre los grupos, continúa una concatenación de las relaciones directas, en donde se entrelazan estas relaciones para formar una red de relaciones más estructurada. Cuando se establece esta red de relaciones, se pueden formular hipótesis o teorías. Las teorías son "comprobadas" por la observación de la realidad objetiva, proceso en el que pueden existir reajustes a las teorías, no sin antes recorrer el camino descrito.

En la tabla II se muestra cómo cada nodo de la figura 4 corresponde a un nivel de abstracción en el descubrimiento científico en general.

El nodo CLASIFICACION (III) tiene un valor especial, pues de él parten dos relaciones indirectas. Obsérvese como el nodo EJEMPLAR (I) es una "fuente", pues de él solo parten relaciones y no llega ninguna a el. Mientras que el nodo TEORIAS Y TENDENCIAS EVOLUTIVAS DEL GRUPO (VI) es un sumidero, pues a él solo llegan relaciones y no salen.

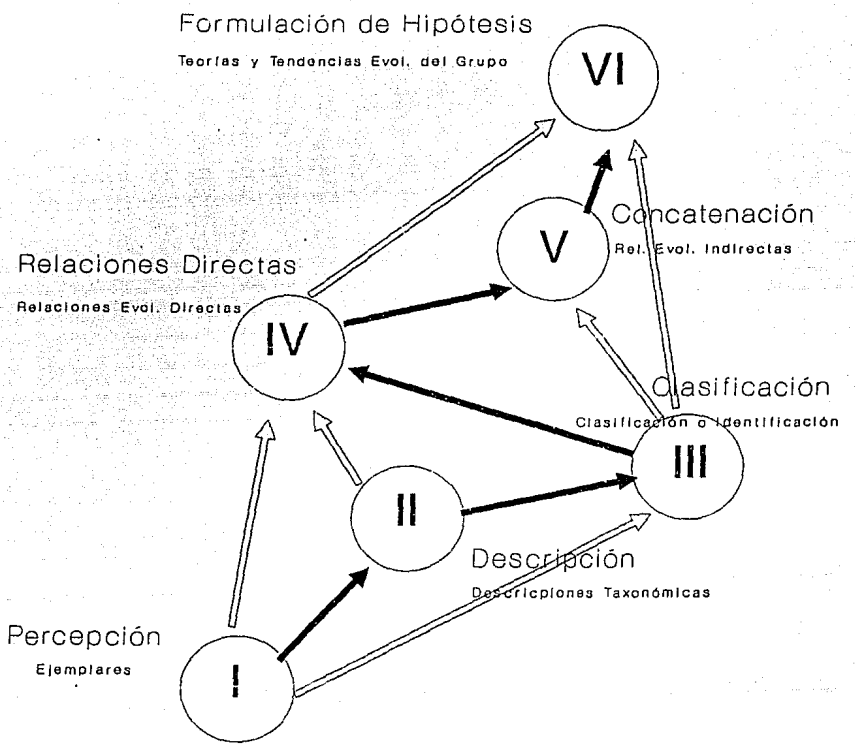


Fig. 4: Representación Gráfica para las Metodologías de los estudios Taxonómico-Evolutivos.

→ Relaciones Directas

⇒ Relaciones Indirectas

Nivel	Nodo	Nivel de Abstracción
I	Ejemplares	Percepción
II	Descripciones Taxonómicas	Descripción
III	Clasificación o Identificación	Clasificación
IV	Relaciones Evolutivas Directas	Relaciones Directas
V	Relaciones Evolutivas Indirectas	Concatenación
VI	Teorías y Tendencias Evolutivas del Grupo	Formulación de Hipótesis

Tabla II. Equivalencia en Taxonomía de los Niveles de Abstracción.

Desde esta perspectiva, se observa que los ejemplares son el punto de partida de la Taxonomía, mientras que la meta es la formulación de hipótesis, es decir la generación de Teorías Evolutivas globalizadoras.

La relación que va de CLASIFICACION a FORMULACION es un paso indirecto que se infiere después de haber recorrido la trayectoria CLASIFICACION (III) -> RELACIONES EVOLUTIVAS DIRECTAS (IV) -> RELACIONES EVOLUTIVAS INDIRECTAS (V) -> TEORIAS Y TENDENCIAS EVOLUTIVAS DEL GRUPO (VI). Este paso puede observarse como una consecuencia lógica de un análisis previo, o bien, como una propuesta de carácter intuitivo, válido para investigadores con suficiente experiencia que justifique el brinco de tres relaciones. En ese caso, el investigador debe estar consciente de que para realizar dicha inferencia se habrá recorrido antes el camino de las relaciones directas, pero su capacidad de abstracción, análisis y experiencia en el área, justifica este brinco, que se realiza de forma explícita. De otro modo, ese recorrido sería un error metodológico y por lo tanto, se estarían realizando inferencias no válidas, muy probablemente incorrectas.

La subgráfica conformada por los nodos: EJEMPLARES, DESCRIPCIONES y CLASIFICACION (fig. 5) contiene dos relaciones directas y una indirecta, que va de EJEMPLARES a CLASIFICACION. Es esta relación indirecta la que se refiere a la identificación taxonómica.

Entre los tres nodos se genera un área de síntesis que se identifica con el proceso de clasificación y diagnóstico taxonómicos. La relación que va de EJEMPLARES a DESCRIPCIONES se

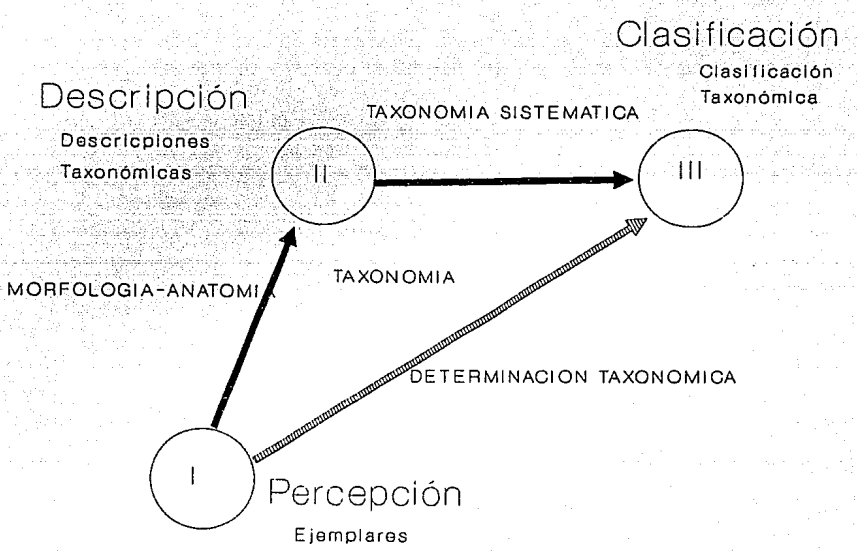


Fig. 5: Gráfica en donde se genera la relación de determinación.

identifica con la Morfologia y la Anatomia, en la mayoria de los casos. La relación que va de DESCRIPCIONES a CLASIFICACION se identifica como Taxonomia-Sistemática.

Además se observa una relación indirecta de EJEMPLARES a CLASIFICACION, que encuentra su expresión como el proceso de identificación taxonómica. Así, el proceso de identificación es sólo una consecuencia del recorrido EJEMPLARES -> DESCRIPCIONES -> CLASIFICACION. Nuevamente se puede encontrar el caso de realizar la inferencia indirecta, sin antes pasar por las relaciones directas, proceso justificado por la experiencia que el investigador tenga en el grupo taxonómico.

El área de TAXONOMIA, definida en la subgráfica de la figura 5, sólo puede concebirse si se tiene en cuenta una interacción entre las tres relaciones:

Morfologia-Anatomia,
Taxonomia-Sistematica
Identificación y Clasificación Taxonómicas.

Así, se observan los elementos que intervienen en el proceso de la identificación taxonómica; a veces intervendrán en igual medida, en otras ocasiones prevalecera alguno de ellos.

Capítulo IV

Automatización de la Identificación Biológica

En este capítulo se presentan diversos trabajos representativos de los intentos por automatizar a la identificación taxonómica. Para realizar automatización, se debe contar con modelación de los procesos a automatizar. Aunque para la mayoria de las propuestas encontradas en la bibliografía no se realiza una modelación explicita y previa a la automatización, se vislumbra que en cada una subyace un modelo conceptual.

En **III. 5** se presentó un modelo de la identificación dentro de un contexto taxonómico, y en **VII. 2** se particulariza con el objeto de realizar su automatización.

IV. 1) Los inicios

La era de la computación moderna comienza en 1947, con la computadora ENIAC. Pero las aplicaciones en identificación biológica comienzan hasta los inicios de los años 60s. Así, Pankhurst (1978) cita que, en 1968, Boughey construye el primer programa para identificación y en la bibliografía el artículo de Goodal (1968) describe un programa para identificacion taxonómica programado en lenguaje PL-1 y que corria bajo una IBM 360. Recuérdese que la palabra policlave aparece hasta 1969. Sin embargo, Fichefet et al. (1984) citan un articulo de Dybowski de 1968, titulado "Conditional probability and the identification of bacteria: a pilot study" publicado en la revista Journal of Geneneral Microbiology.

La generación de policlaves por computadora o programas para identificación automatica, esta muy ligada a la generación de programas para construcción de claves dicotómicas y descripciones taxonómicas. Los conceptos surgidos de estas tres tareas y de los de la taxonomia numérica, en realidad son el antecedente para la futura creación de métodos mas avanzados en la automatización de la taxonomía en general.

En la siguiente sección se explican algunas implementaciones de policlaves en computadora, para familiarizar al lector con la práctica de la identificación biológica automatizada y mostrar mediante esos ejemplos su estado actual de desarrollo.

IV. 2) Policlaves por Enunciados

Se denomina aquí "policlaves por enunciados" a las claves en computadora basadas en matrices de datos de taxa vs. características, en donde las características constituyen en la clave enunciados que adquieren valores independientes de los demás enunciados.

IV. 2. a) La policlave FAMEX

La policlave FAMEX es una clave por computadora para las más de 200 familias de plantas con flores de México (Murguía, 1988; Villaseñor y Murguía, en prensa). Esta policlave está programada para que la utilicen personas con conocimientos mínimos de computación; el ambiente es a base de menús. El menú principal se presenta en la figura 6.

El usuario indica a la computadora las características que presenta el ejemplar bajo determinación. La forma de introducirlas es indicando el número de la característica que presenta el ejemplar e inmediatamente el programa indica las posibles familias. El usuario puede seguir introduciendo más características para reducir el número de familias posibles para la diagnosis del ejemplar.

La policlave FAMEX se está utilizando en diversas instituciones nacionales y extranjeras, brindando facilidades en la determinación de ejemplares a nivel familia y como material didáctico en cursos de Botánica. La policlave FAMEX está a disposición de cualquier persona que la solicite.

En la opción I) INFORMACION DE FAMILIAS, el programa presenta las características que los ejemplares de una determinada familia presentan, en otras palabras, su descripción.

```
POLICLAVE COMPUTARIZADA "FAMEX" PARA DETERMINACION DE
FAMILIAS DE PLANTAS CON FLORES PRESENTES EN MEXICO

                    B) BUSCAR FAMILIA

                    I) INFORMACION DE FAMILIAS

                    D) DETERMINACION

                    N) NUEVO EJEMPLAR

                    S) SALIDA

SU OPCION:
```

Fig. 6. Menú principal de la clave FAMEX.

IV. 2. b) La policlave TROPIFAM

Esta es una policlave para familias de angiospermas del oeste de Estados Unidos elaborada por Duncan y Meacham (1986a, b), que utiliza la información de Simpson y Jones (1972).

La interacción con el programa es mediante comandos. Cada comando tiene un nombre y parámetros, así, el comando para adicionar caracteres, que indican que están presentes en el ejemplar bajo determinación, es AC:

```
AC 2 4/6 10- 13-/17-
```

lo que indica que el ejemplar tiene los caracteres 2, 4, 5 y 6 y no tiene los caracteres 10, 13, 14, 15 16 y 17.

IV. 3) Policlaves por Carácter - Estado de Carácter

Las policlaves por estado de carácter difieren de las policlaves por enunciados en que los enunciados no son totalmente independientes, sino que guardan cuando menos una relación de exclusión con otros. Los enunciados de estados de carácter del mismo carácter son excluyentes entre sí.

IV. 3. a) El programa SATAX

El programa SATAX (Sistema de Ayuda al TAXónomo) (Murguía y Téllez, 1988; Murguía, 1990c) utiliza la representación de Carácter- Estado de Carácter. La matriz de datos está constituida por los caracteres, los estados de carácter y los taxa. Los datos específicos de cada taxon están asociados mediante una lista a los estados de caracteres.

Cada taxon tiene un número entero asociado, que es el que aparece en las listas de los estados de carácter. Así, la presencia de un estado de carácter en un taxa se representa mediante la inclusión del correspondiente número entero en la lista del estado de carácter.

El ambiente de esta clave contiene sólo dos áreas: el menú y la información al usuario. En la pantalla de información al usuario es donde se establece un "diálogo" entre la clave y el usuario. El usuario selecciona la característica y la clave responde con una lista de los posibles estados de carácter, de los que el usuario selecciona uno. Seleccionado el estado de carácter, la clave brinda una lista de los taxa a los que puede pertenecer el ejemplar bajo determinación. Acto seguido el usuario puede seleccionar otro carácter (fig. 7).

El programa cuenta además con un módulo de generación de descripciones taxonomicas. Este módulo toma los datos de la misma matriz de datos que se utiliza para la diagnosis, de forma que al modificarse la informacion de la policlave, automáticamente se modifican las descripciones taxonómicas.

SATAX Tambien contiene otro módulo para consultar, en linea, un diccionario de términos taxonómicos del grupo.

```
                                    CACTUS
Archivo Caracteres Descripción Estados Modificar Taxa Policlave Glosario Fin

3) tallo (tipo en globosas)                    4) tallo (desarrollo morfismo)
5) tallo (tipo de ramificacion)                6) tallo (superficie)
7) tallo (tricomas)                            8) raiz (tipo)
9) hojas (presencia)                           10) hojas (desarrollo)
11) hojas (tipo)                               12) aréolas (tipo en el tallo)
13) aréolas (del tallo: surco)                 14) surco areolar (tamaño)
15) aréolas del tallo (glándulas)              16) tallo (espinas)
17) espinas (tipo)                             18) aréolas (glóquidas)

¿ Cuál No. de caracter (0 para terminar)? 12
aréolas (tipo en el tallo):

1) monomorfas
2) dimorfas

¿ Cuál No. de estado ? 2

6 posibilidades.
Posibles taxa:
8) WILCOXIA              46) NEOLLOYDIA          52) ARIOCARPUS
53) EPITHELANTHA         54) PELECYPHORA         58) MAMMILLARIA

Presione cualquier tecla para continuar ...
```

Figura 7: Policlave CACTUS en el ambiente de SATAX (Sistema de Ayuda al TAXónomo).

IV. 3. b) La Policlave CACTUS

La policlave CACTUS (Gama et al., 1990) es para géneros de Cactáceas de México. Está implementada en el programa SATAX y por lo tanto basada en la representación de carácter - estado de carácter y no en la de enunciados.

La matriz incorporada cuenta con más de 60 caracteres, cerca de 170 estados de carácter y 59 géneros de la familia.

Contiene un diccionario de términos taxonómicos que el programa es capaz de manipular para auxiliar al usuario poco familiarizado con la terminología del grupo.

También se pueden obtener las descripciones taxonómicas en lenguaje natural de los géneros, pues es una de las facilidades que el programa SATAX brinda.

IV. 3. c) El programa ONLINE

El programa ONLINE (Pankhurst y Atchison, 1975) permite una identificación utilizando una matriz de datos en formato DELTA (Dallwitz, 1980).

El formato DELTA está basado en la representación por estados de carácter. Los datos se asocian a los taxa, por lo que se tienen que especificar mediante una pareja de números, el carácter y el estado de carácter.

La interacción con el usuario se logra mediante comandos que indican al programa que realice determinadas acciones.

Watson y Dallwitz (1983) han construido un banco de datos en formato DELTA para los géneros de la subfamilia Caesalpinioideae (Leguminosae). La matriz de datos consta de 134 caracteres, con dos a nueve estados de carácter cada una. Los caracteres se agrupan en morfológicos (1-80), química de la semilla y germinación (81-86), anatomía de la hoja (87-118), anatomía de la madera (119-123), polen (124-130), citología y distribución geográfica (131-134). Con este banco de datos, han generado de forma automática descripciones taxonómicas y claves dicotómicas.

El usuario debe editar la matriz de datos directamente en un procesador de palabras, en un formato estricto (DELTA) lo que dificulta su uso. Peláez y Vargas (1990) han construido un programa (CAPDELTA) que facilita la captura de datos en formato DELTA.

IV. 4) Comparación de los programas descritos

Varias características se pueden analizar de las diferentes implementaciones existentes. Los ejemplos presentados en el inciso anterior, muestran algunos de los atributos que en general presentan las policlaves por computadora. Para resaltar los beneficios que tiene la identificación por computadora, frente a los métodos tradicionales (claves dicotómicas), se comparan los programas presentados. En la tabla III se observan algunas características de estos programas.

	FAMEX	TROPIFAM	SATAX	ONLINE
Interacción con el usuario	Menús	Comandos	Menús	Comandos
Ayuda en linea	No	No	No	No
Manejo de diccionario	No	No	Sí	No
Modificación de matriz de datos por el usuario	No	No	Sí	Sí
Módulo para generación de descripciones	No	No	Sí	No
Año de creación	1988	1986	1988	1976
Lenguaje de programación	Pascal	Fortran	Prolog	Fortran

Tabla III. Características de las policlaves en computadora descritas.

La interfase con el usuario puede ser a base de menús o comandos; un ambiente de menús es más agradable. La incorporación del concepto de "botón" de la computación actual, hará el uso de la computadora más atractivo para el taxónomo. Mientras un programa sea más fácil de usar, más se usa, por lo que el éxito de la automatización de la identificación taxonómica depende también de la interfase con el usuario. A este respecto hay que señalar que la incorporación de ayudas interactivas y manejo de diccionario es un punto fundamental.

La posibilidad de modificar la matriz o de incorporar nuevas matrices, hace una diferencia en la intención de los programas, pues mientras unos funcionan con información no actualizable en el programa, otros pueden utilizarse como

"cascarón" de policlaves por especialistas que no desean programar.

Como la identificación taxonómica está intimamente relacionada con otros quehaceres del taxónomo, es atractiva la idea de incorporar ayuda automática a esos quehaceres, como por ejemplo diccionarios, generación de descripciones taxonómicas y claves dicotómicas o análisis filogenéticos, entre otros.

Por último, mientras no existan librerías especificas para la programación de policlaves en computadora, cualquier lenguaje de programación es buen candidato para el desarrollo.

IV. 5) Objeciones a la Identificación Automatizada

Esta sección brinda un panorama de los diferentes tipos de objeciones a la automatización de la identificación biológica.

Para muchos, la automatización de procesos es un beneficio en términos de tiempo, dinero o nuevos enfoques. Sin embargo, existen objeciones a la automatización de la identificación biológica. Principalmente surge la pregunta: ¿ podrá una computadora desempeñar el papel del taxónomo en la identificación biológica sin diferencias substanciales ?, es decir, ¿ existen situaciones en las que un taxónomo será insustituible por una computadora?.

Las objeciones a la situación de que las computadoras nunca podrán substituir a los taxónomos, en la tarea de diagnosticar ejemplares, son una particularidad de las objeciones analizadas por Turing (1950) a la pregunta ¿ pueden pensar las máquinas ?.

Asi, las personas suelen reaccionar en contra de la "totipotencialidad" de las computadoras, con objeciones que Turing clasifica en nueve categorías:

1) **Objeción Teológica:** El pensamiento es una función inmortal del hombre. Dios ha dado un alma inmortal a todos los hombres y mujeres, pero a ningún otro animal o maquinas. Por lo tanto, ningún animal o máquina puede pensar.

2) **Objeción "Cabeza en la Arena":** Las consecuencias del hecho de que las máquinas pensaran serian demasiado horribles. Esperemos y creamos que no puedan hacerlo.

3) **Objeción Matemática:** Existen muchos resultados de la Logica Matemática que pueden utilizarse para demostrar

que existen serias limitaciones en las computadoras actuales (máquinas de estado finito). Por ejemplo, el teorema de Gödel demuestra que, en cualquier sistema lógico suficientemente poderoso, se pueden formular proposiciones que no pueden demostrarse ni refutarse dentro del sistema, a menos que el sistema mismo sea contradictorio.

4) **El argumento de la Conciencia:** Este argumento está expresado en la "Oración de Lister" (del año 1949 que Jefferson citó) : "Hasta que una máquina no sepa escribir un soneto o componer un concierto, con base en los pensamientos y las emociones que siente, y no a consecuencia de la caida venturosa de símbolos, no podremos estar de acuerdo en que la máquina pueda ser igual que un cerebro, es decir, que no solamente sepa escribirlos, sino también saber que los ha escrito..."

5) **Argumentos desde el punto de vista de diferentes incapacidades:** Estos argumentos tienen la forma siguiente:

"Admito que usted puede compeler a las máquinas a hacer todas las cosas que acaba de mencionar, pero nunca podrá inducir a una máquina a hacer X."

6) **La objeción de Lady Lovelace:** Un informe sobre la Máquina analítica de Babbage, elaborado por Lady Lovelace declara: "La máquina analítica no pretende crear nada. Puede hacer cualquier cosa que sepamos ordenarle que haga". Es decir, a menudo se declara que las máquinas no pueden dar sorpresas (opinión que Turing rechaza).

7) **Argumento de la continuidad en el sistema nervioso:** El sistema nervioso no es una máquina de estado discreto, un error pequeño, acerca de la dimensión de un impulso nervioso que tropieza con una neurona, puede representar una diferencia grande para el volumen del impulso saliente.

8) **El argumento de la informalidad del comportamiento:** No es posible elaborar un conjunto de reglas que describa lo que una persona debería hacer en cualquier serie concebible de circunstancias.

9) **El argumento de la percepcion extrasensorial:** Existen fenómenos que parecen negar muchas de las ideas científicas habituales, como la telepatía, la clarividencia, la precognición y la psicocinesis.

Estos argumentos, citados por "el padre de la Inteligencia Artificial", también son aplicables cuando se intenta poner límites al papel que pueden desempeñar las computadoras en el futuro de la identificación biológica y, en general, en muchas de las tareas del científico contemporáneo.

Estas actitudes, que se observan a nivel individuo, producen efectos y actitudes emergentes en grupos de individuos. Todavía faltan muchos estudios acerca del efecto de la automatización electrónica en las ciencias y en la sociedad, pero se vislumbran problemas serios; además no se está seguro que las computadoras puedan resolverlos adecuadamente.

IV. 6) Bases de Datos, Descripciones Taxonómicas e Identificación.

Es la intención de esta sección recalcar la importancia que tiene la información en la construcción de las policlaves. Sin información no es posible construir una herramienta para identificar.

Como las claves son información estructurada de cierta manera, las bases de datos taxonómicas desempeñan un papel importante en la construcción de policlaves. De hecho, a lo que se tiende es a construir bases de datos taxonómicas que brinden información para la generación de descripciones taxonómicas y para el proceso de identificación, además de tener relación con bases de datos florísticas. La cantidad de información que se debe compartir entre las Bases de Datos Taxonómicas y las Bases de Datos Florísticas justifica el hecho de intentar construir sistemas integrados florístico-taxonómicos.

Watson y Milne (1972) y Raynal (1974) explican de manera superficial la dinámica de los programas generadores de claves dicotómicas y los beneficios de utilizarlos.

Moreno y Alikin (1988) hablan de la importancia que tienen las bases de datos en la generación automática de descripciones taxonómicas y de claves dicotómicas y policlaves.

En la medida que se construyan sistemas integrados que el taxónomo pueda acceder de diferentes maneras, la rapidez en la acumulación de información y conocimientos aumentará. El poder tener bases de datos florísticas, taxonómicas, claves de identificación, y descripciones botánicas interrelacionadas en un sistema, brinda al usuario un beneficio enorme. Además, el que todos estos componentes provengan de una misma información, hace que la congruencia del sistema sea más fácil de mantener. Por ejemplo, si la información para generar las descripciones y

para generar las claves es la misma que la que está contenida en la base de datos, entonces, al modificar la base de datos, se modifican automáticamente, las descripciones y las claves.

También es importante observar cómo se empiezan a establecer comunicaciones entre diferentes bases o bancos de datos en diferente lugar académico y geográfico (telecomunicaciones).

Capítulo V

Perspectivas de la Identificación Biológica Automatizada

En este capítulo se explican las diferentes maneras de abordar la automatización del proceso de identificación biológica. El proceso de identificación no es exclusivo de la Biología, por lo que las metodologías para automatizarlo surgen de varias disciplinas técnicas y científicas.

Los procesos de identificación, comúnmente llamados en otras áreas "de clasificación", requieren de información y conocimientos especializados. Se debe saber qué información es importante o relevante para poder identificar.

Para automatizar el proceso de identificación se debe elegir un método acorde con el tipo de información disponible.

Desde esta perspectiva, los métodos automáticos de identificación deben estructurarse con base en modelos del conocimiento que consideren, implícita o explícitamente, una clasificación; por esa razón, en la medida en que la clasificación contenga una estructura lógica más sólida, la automatización de procesos que involucren a la clasificación serán más fáciles de realizar y en mayor amplitud.

Así, los métodos de identificación contienen conocimiento, conocimiento sobre como manipular el conocimiento, es decir, metaconocimiento. Es interesante observar la retroalimentación que existe entre la identificación y la clasificación. Como los métodos de identificacion exigen o buscan una estructura lógica en la clasificación, para poder automatizar procesos, la clasificación se verá obligada a revisar sus propios criterios, además de atender a la tarea práctica de la identificación.

V. 1) Enfoque Matricial

En este enfoque el conocimiento de la clasificación taxonómica se representa mediante matrices de caracteristicas

contra clases. Así, por ejemplo, se pueden representar los datos de un grupo taxonomico en una matriz en donde los renglones sean las clases y las columnas las caracteristicas.

V. 1. a) Algoritmos numéricos

Es muy sencillo representar matrices con valores binarios, es decir, matrices de presencia - ausencia. Pero cuando se desea aceptar más de dos valores posibles, las operaciones se tornan más complejas. Generalmente se suele representar la ausencia mediante el cero y la presencia mediante el número uno. Por ejemplo, la siguiente matriz M contiene datos sobre la clasificación de un grupo hipotético:

	CLASES				
CARACTERISTICAS	A	B	C	D	E
a	0	1	0	1	1
b	1	0	1	1	0
c	0	0	0	0	1
d	1	0	0	1	0
e	1	1	1	1	0
f	0	1	0	1	1

MATRIZ EJEMPLO 1

En esta matriz de 5x6 se representa la presencia o ausencia de seis características (a a f) y su distribución en cinco clases (A a E).

En general, en el esquema de carácter - estado de carácter, cada renglón representa la presencia de un estado de carácter. Por ejemplo, los estados de carácter a y b pueden pertenecer al carácter C1. los estados c, d, e y f al C2. Por lo que una matriz en que cada columna representa a un carácter y cada valor a un estado de caracter podria ser:

	CLASES				
CARACTERISTICAS	A	B	C	D	E
C1	b	a	b	a,b	a
C2	d,e	e,f	e	d,e,f	c,f

Para poder hacer operaciones con matrices fácilmente, se prefiere el primer esquema presentado, el de representación de los estados de carácter con estados binarios.

Las características del ejemplar a determinar se representan mediante un vector de número de elementos, igual al número de renglones de la matriz de datos del grupo. Por ejemplo, el vector de datos V del ejemplar bajo determinación podría ser:

```
        a     b     c     d     e     f
     [ 0     1     0     1     0     0 ]
     ---------     -----------------
        C1              C2
```

Lo que representa que el ejemplar tiene los estados de carácter b (para la característica C1) y d (para la característica C2).

Ahora se puede realizar alguna operación entre el vector y la matriz. En particular, el resultado de multiplicar el vector por la matriz es un producto interesante. Ese producto es precisamente un vector R, resultado de la diagnosis, en donde cada elemento del vector indica si el ejemplar puede pertenecer o no a cada clase. Si el elemento es igual al número de unos en el vector del ejemplar, en este ejemplo 2, el ejemplar puede pertenecer a esa clase, y si el elemento es menor, entonces el ejemplar no puede pertenecer a esa clase:

$$V \times M = R$$

$$R = [\ 2 \quad 0 \quad 1 \quad 2 \quad 0\ ]$$

$$A \quad B \quad C \quad D \quad E$$

Lo que indica que el ejemplar sólo puede pertenecer a las clase A y D, y no a B, C y E.

El significado de la multiplicación del vector V (características presentes en el ejemplar en determinación) por la matriz M (matriz de características) es el siguiente: como cada elemento r(i) del vector producto es la suma de los productos del i-ésimo elemento de V por el elemento m(i,j) de M (j representa a cada renglón de la matriz M), entonces la suma máxima es el número de renglones que contenga la matriz, pues cada producto

$$v(i) \times m(i,j)$$

tiene sólo dos posibles valores: 1 ó 0, ya que cada v(i) y m(i,j), también es 1 o 0. De hecho la cota de cada r(i) es el número de unos que contenga V, es decir, $\Sigma v(i)$

Así, cuando r(i) = $\Sigma v(i)$, es decir, r(i) adquiere el valor máximo, significa que la columna i de la matriz M (que representa las características de un taxon) contiene unos donde V también tiene unos, es decir, ese taxon presenta todas las características que presenta el ejemplar en determinación. Así, cada r(i) representa el número de características que presenta el ejemplar en determinación y que también presenta el taxon i.

La implementación de éste método es muy simple, pues la multiplicación de matrices es sencilla. En muchos paquetes o lenguajes de programación ya está dada como una utilería o procedimiento interno.

En este método, el vector que representa a las características del ejemplar bajo determinación, puede tener cualquier combinación de valores. Si el vector está compuesto sólo de ceros, entonces ninguna clase resulta la posible, y si el vector contiene sólo unos, entonces, todas las clases son posibles. Es importante hacer notar que el vector debe contener unos en las características que se dude, así puede introducir en el vector la presencia de dos estados de carácter para un mismo carácter, lo que significa una disyunción lógica.

Es interesante el hecho de poder establecer un margen de tolerancia para que el ejemplar pertenezca o no a las clases, por ejemplo, si en el ejemplo anterior se establece un margen de error de uno, lo que significa que también son candidatas las clases que difieran en un valor del vector, entonces no sólo son posibles las clases A y D, sino también la C.

Además del proceso de identificación en sí, al método matricial se le pueden añadir algunas otras características, por ejemplo, Raudys y Jain (1991) exponen seis enfoques para determinar la "probabilidad de error en la clasificación", (PMC: Probability of Misclassification): Distancia Euclideana, Discriminante Linear de Fisher, Función Discriminante Cuadrática, Vecino más Cercano, Ventana de Parzen y Multinomial.

Sneath (1980) proporciona un ejemplo sencillo, en donde se explica un programa que manipula una matriz con probabilidades para cada taxon y estado de carácter.

V. 1. a. i) Incorporación de Negación.

La incorporación de la negación puede realizarse indicándola con un uno negativo en el vector de entrada:

VECTOR V DE ENTRADA:

CARACTERISTICAS

a	b	c	d	e	f
[0	1	0	0	-1	-1]
C1		C2			

Los ceros representan que el usuario no ha indicado afirmación o negación para la correspondiente columna (estado de carácter).

La implementación de negación en estos sistemas se puede lograr reagrupando los renglones; en donde cada grupo esté conformado por los estados de un mismo carácter, y asi decidir en que casos realizar una multiplicación por cero y en cuales no. Para lograr esto se puede realizar una transformación al vector, mas que modificar el algoritmo de clasificación.

Es más sencillo realizar una transformación del vector y operar con el transformado, que operar con el vector original y realizar cambios en la operación de multiplicación del vector con la matriz. Cuando se introduce una negación en un vector de entrada, como en el caso anterior, el vector se transforma mediante la regla de la tabla IV.

Si se aplica la regla de la tabla IV al vector que se dió como ejemplo de negación, se produciría el siguiente Vector Resultado:

VECTOR V2 RESULTADO DE TRANSFORMAR V:

CARACTERISTICAS

a	b	c	d	e	f
[0	1	1	1	0	0]

Si se realiza la operación:

$$V2 \times M = R2$$

El resultado es:

$$R2 = [\ 2 \quad 0 \quad 1 \quad 2 \quad 1]$$

A B C D E

A.- Detectar los caracteres en los que haya al menos un valor negativo. Para cada carácter:

1.- Para los valores que pertenezcan al mismo carácter en que se introdujo la negación:

1.1.- Substituir los ceros por unos.

2.- Substituir las negaciones por ceros en todo el vector.

B.- Producir el Vector de Salida.

Tabla IV. Regla para tratamiento de negaciones en el vector de entrada.

Lo que significa que el ejemplar puede pertenecer a las clases A y D, que son las que comparten 2 estados de carácter con el ejemplar bajo determinación. También se puede decir que pertenece a las clases A, C, D y E, pero no a la B, siempre y cuando se establezca un margen de tolerancia de uno (2-1=1). Se observa que para la característica C2 (en el VECTOR V DE ENTRADA) no se introdujo ningún uno, es decir no se dijo explícitamente qué estados de carácter podría presentar el ejemplar, en cambio si se indicó qué estados de carácter (para el caracter C2) definitivamente no presenta (indicados con -1).

Otra alternativa para incorporar la negación, es agregar renglones. Los renglones que se agregan son las negaciones de las características, por ejemplo, si la característica A es Tiene Espinas, se agrega la característica (o renglon) no A que

representa a **No tienen espinas.** La matriz ejemplo 2 contiene como caracteristicas a las negaciones de las caracteristicas de la matriz ejemplo 1:

	CLASES				
CARACTERISTICAS	A	B	C	D	E
a	0	1	0	1	1
no a	1	1	1	0	1
b	1	0	1	1	0
no b	1	1	0	1	1
c	0	0	0	0	1
no c	1	1	1	1	1
d	1	0	0	1	0
no d	0	1	1	1	1
e	1	1	1	1	0
no e	1	1	1	1	1
f	0	1	0	1	1
no f	1	1	1	0	1

MATRIZ EJEMPLO 2

La matriz ejemplo 2 contiene más información que la matriz ejemplo 1. Por ejemplo, el Taxa D tiene las caracteristicas a y b, y no tiene las caracteristicas b y c, es decir, puede o no presentar la caracteristica b.

A continuación se muestran algunas consecuencias de incluir caracteristicas negadas en una policlave. Respecto al manejo de la negación en el proceso de diagnosis, consúltese a Salazar (1990). El problema de la negación también es interesante en la generación de descripciones botánicas. Obsérvese que combinaciones son las posibles para un taxon cuando se incluye una caracteristica y su negación, por ejemplo A y no A:

	A	no A
Combinación 1	1	1
Combinación 2	1	0
Combinación 3	0	1
Combinación 4	0	0

En la combinación uno, se expresa el hecho de que un taxa pueda o no presentar la característica A. En la combinación 2 se representa el hecho de que el taxon siempre presenta la característica A. En la combinación 3 el taxon nunca presenta la característica A y la combinación 4 es un error, pues dada cierta caracteristica el taxon la debe o no presentar, también puede ser interpretado como el que la caracteristica A no es aplicable al taxon. A continuación se presenta un ejemplo:

Tiene Espinas	No tienen Espinas	Interpretación
1	1	¿ en qué porcentaje ?
1	0	siempre tiene espinas
0	1	nunca tiene espinas
0	0	contradicción o no aplicable

Esta situación de las cuatro posibilidades fué utilizada para la construcción de claves simópticas por Leenhouts (1966). En las claves simópticas se da al usuario la libertad de elegir los caracteres, de acuerdo al número de taxa que los presenten. El usuario tiene una visión global del "valor" que tienen los diferentes caracteres en la clave y respecto al ejemplar bajo determinación.

V. 1. b) Tarjetas Perforadas

Las claves en que su estructura de datos es una matriz de los taxa contra sus características, pueden ser representadas en tarjetas perforadas.

Existen básicamente dos alternativas. Una posibilidad es representar cada característica en una tarjeta y las perforaciones representan la presencia de esa característica en los taxa. La ausencia de perforación indica ausencia de esa caracteristica. Cada taxon tienen asignada una determinada área de la tarjeta y que es el mismo lugar en cada tarjeta.

La elección de una tarjeta indica la presencia en el ejemplar de la caracteristica representada por la tarjeta. Asi, cuando se superponen varias tarjetas, se está indicando la presencia en el ejemplar de cada característica representada por cada tarjeta. Quedarán areas en donde todas las tarjetas contienen perforación, lo que indica que ese taxon presenta todas esas caracteristicas. Las áreas que cuando menos contengan una tarjeta sin perforar, indicarán taxa que ya no pueden pertenecer al ejemplar bajo determinación, por no presentar todas las caracteristicas que se seleccionaron.

Ejemplos de este tipo de alternativas se describen en Murguía (1988), Simpson y Janos (1972) y Hansen y Rahn (1969).

La otra alternativa es al reves, es decir, que cada tarjeta represente a un taxon y no a una característica, y cada perforación represente la presencia de características en el taxon. Un ejemplo de este tipo de claves se da en Little (1968). La diferencia estriba en que siendo ahora las tarjetas las que representan taxa, y las áreas a las características, se deben elegir todas las tarjetas (todos los taxa son candidatos) y seleccionar las áreas de las tarjetas. Por lo tanto estas claves contienen las perforaciones al borde de las tarjetas y la presencia de una característica se indica haciendo una muesca al borde de la tarjeta. La ausencia de una característica se indica con una perforación que no tiene muesca. Así, todas las áreas deben estar perforadas, pero no todas con muesca. La elección de las características observables en el ejemplar se indican en la clave eligiendo áreas, lo que equivale a seleccionar todas aquellas tarjetas (Taxa posibles) que tengan una muesca (característica presente) en el área correspondiente a la característica deseada.

V. 2) Enfoque de Teoría de Conjuntos

El proceso de determinación puede ser abordado como un proceso de operaciones de conjuntos.

Para explicar este proceso se utilizarán los siguientes símbolos:

- V: Universo. Características consideradas por la policlave.
- $\cap$: Intersección.
- U: Unión.
- $\subseteq$: Inclusión.
- $\in$: Pertenencia.
- $\emptyset$: Conjunto vacío.
- | |: Cardinalidad
- EJ: Características del ejemplar bajo determinación.
- ej: Características del ejemplar bajo determinación conocidas por la clave.
- CAR: Característica.
- $\langle Ti \rangle$: Taxa considerados por la clave.
- DIAG: Taxa posibles.

Definiciones:

EJ $\subseteq$ V.
Para todo Ti: Ti $\subseteq$ V,

En la determinación mediante conjuntos, se realizan las operaciones mostradas en la tabla V.

I) Inicialización:
   a) Asignar ∅ a ej.
   b) Asignar {Ti} a DIAG.
II) Preguntas al usuario:
   Repetir
      c) Elegir una CAR | CAR ∈ V y CAR ∈ EJ.
      d) Asignar EJ ∪ {CAR} a ej.
      e) DIAG = DIAG ∩ {Ti para los que ej $\underline{C}$ Ti}
   hasta que |DIAG|=1 o hasta que no se puedan elegir mas CAR.
III) Diagnosis final.
   f) Mostrar al usuario los Ti ∈ DIAG.

Tabla V. Algoritmo de diagnosis mediante interseccion de conjuntos.

Este enfoque es simple de implementar, pues existen lenguajes de programación que incorporan las operaciones y manejo de conjuntos entre sus estructuras de datos, así como procedimientos y funciones predefinidas.

En la tabla VI se muestran los operadores para conjuntos en Pascal, tomados de Schneider et al. (1982).

Notación usual	Notación Pascal	Operador
∩	*	Intersección.
∪	+	Unión.
$\underline{C}$	<=	Inclusión.
-	-	Diferencia.
=	=	Igualdad.
<>	<>	Desigualdad.
∅	[]	Conjunto vacio.
∈	in	Pertenencia.

Tabla VI. Operadores del lenguaje Pascal para conjuntos.

A continuación se describe el proceso de diagnosis, mediante operación de intersección entre conjuntos sobre la matriz ejemplo 2 (de la sección anterior). Los conjuntos contienen como elementos a los taxa, y cada uno equivale a una

característica o combinación de ellas. Los conjuntos que intervienen al inicio de la diagnosis son:

U = (I, II , III, IV)
A = (I, II)
B = (I, III)
C = (II, IV)
no A = (III, IV)
no B = (I, II, IV)
no C = (I, III, IV)

El usuario ha indicado que el ejemplar tiene la característica B y no tiene la característica C;

B ∩ no C = (I, III)

si el usuario indica a continuación que el ejemplar no presenta la característica A, entonces la diagnosis final es el taxon III:

(B ∩ no C) ∩ no A = (III)

V. 3) Enfoque Lógico

Aunque en realidad todos los enfoques tratados en este trabajo tienen un trasfondo lógico, en esta sección se hace explícito el tratamiento de diagnóstico mediante la lógica matemática. Como la lógica matemática es el principal sistema de símbolos en el que reposa la ciencia actual, es natural que en muchos casos resulte el esquema de representación y tratamiento de información y conocimientos más adecuado.

En este enfoque se trata al proceso de diagnosis desde un punto de vista declarativo, en donde el conjunto de declaraciones debe ser congruente y no contener inconsistencias.

La técnica de los programas de policlaves que utilizan este esquema, deben eliminar proposiciones lógicas, con el fin de que las que introduzca el usuario como verdaderas sean congruentes con el sistema. Se puede abordar desde un punto de vista de lógica proposicional, o más estructurado desde otro de lógica de predicados, en donde la información se lleva a una "granularidad" mayor, es decir, la información se representa con mayor detalle.

V. 3. a) Sistema Lógico Proposicional

Notación:

P: proposición
V: Verdadero
F: Falso
¬: Negación
^: Conjunción
v: Disyunción.
->: Implicación
Ti: Taxon

Reglas:

a) Todas las características son una P.

P1, P2, P3, ... Pj

así, se tienen j proposiciones.

b) Toda proposición es o falsa o verdadera, pero no ambas cosas.

c) El ejemplar está representado por un conjunto de proposiciones falsas y verdaderas.

d) Cada taxon está representado por una conjunción de las j proposiciones, sólo que unas negadas y otras afirmadas:

Ti: Pn ^ Pn+1 ^ Pn+2 ^ ... ^ Pm ^ ¬Pk ^ ¬Pk+1 ^ ¬Pk+2 ^ ... ^ Pj

Se dice que las fórmulas lógicas que tienen la estructura anterior están en Forma Normal Conjuntiva (FNC).

e) La policlave consta de t conjunciones del tipo anterior (t=Número de Taxa que considera la Clave).

Taxa considerados por la clave = T1, T2, T3, ... , Tt

El proceso como opera la policlave se muestra en tabla VII.

I) Inicialización:
   a) Adicionar todas las Ti al sistema
II) Preguntas al usuario:
   Repetir
      b) Preguntar al usuario por la falsedad o verdad de una Pi.
      c) Si Pi es falsa, eliminar todas las Ti que contenga a Pi (afirmada)
         Si Pi es verdadera, eliminar todas las Ti que contenga a ¬Pi (negada).
   hasta que no puedan responderse mas Pi o hasta que sólo quede una Ti.
III) Diagnosis final.
   d) Mostrar al usuario los Ti que permanecen.

Tabla VII. Algoritmo de diagnosis mediante congruencia con Forma Normal Conjuntiva (FNC).

Uno de los problemas con este método, es que muchas veces el significado de una proposición afirmada no es estricto, es decir, significa que puede o no ser verdadera. Por lo tanto, el elemento clave para ir eliminando los Ti son aquellas proposiciones que estan negadas cuando el usuario indicó su afirmación. Por ejemplo, si se tiene la siguiente proposición:

P = Espinas presentes.

El significado a esta proposición es: Si P está negada en Ti, entonces ningún elemento de Ti tiene espinas; Si P está afirmada en Ti, entonces los elementos de Ti pueden tener espinas, pero no estrictamente todos.

Si esta proposicion aparece negada en una Ti y el usuario indicó su afirmación, entonces esa Ti sera eliminada, pero, por el contrario, no se pueden eliminar aquellas Ti que contengan a P afirmada cuando el usuario niegue P.

Este problema se soluciona adicionando a la policlave otra proposición:

Q = Espinas ausentes

Así, el significado de P ahora puede ser estricto, es decir en aquellas Ti en que P aparezca afirmada y Q negada (todos los elementos de Ti tienen espinas) podrán eliminarse cuando el usuario indique que P es falsa.

Símbolos en Ti			Significado Taxonómico
P	^	Q	Ejemplares de Ti con y sin espinas.
¬P	^	Q	Ejemplares de Ti sin espinas.
P	^	¬Q	Ejemplares de Ti siempre con espinas.
¬P	^	¬Q	Error.

V. 3. b) Sistema Lógico de Predicados

La información en lógica de predicados tiene un nivel mayor de "granularidad" que en la lógica proposicional, es decir, se desglosan más los elementos del sistema. Mediante esta representación, se puede manejar el esquema de carácter - estado de carácter que el esquema Lógico proposicional simple no es capaz de manipular. Kowalski (1986) y Genesereth y Nilson (1988) explican a la Lógica de Predicados con algunos ejemplos de diagnóstico.

Los predicados representan a los caracteres y los argumentos a los estados de carácter. Definamos el predicado espinas, que contiene un argumento:

ESPINAS().

Este predicado puede tener como argumentos a alguno de los estados de carácter. Así, se pueden tener dos instancias diferentes para el predicado ESPINAS:

ESPINAS(presentes).
ESPINAS(ausentes).

En este sistema también se representan a los taxa mediante fórmula tipo Ti, pero en lugar de estar conformadas por proposiciones, están conformadas por predicados. Además se incluye la disyunción.

La presencia de dos o más instancias de un mismo predicado en una misma Ti sólo es permitida como disyunción. Se permiten conjunciones sólo entre predicados distintos:

```
Ti:  (P1(A11) v P1(A12) ... v P1(A1j)) ^
     (P2(A21) v P2(A22) ... v P2(A2k)) ^
     .
     .
     .
     (Pn(An1) v Pn(An2) ... v Pn(Anl))
```

donde Aij son argumentos de los predicados o estados de carácter y cada instancia de predicado puede estar afirmada o negada. Así, en el Ti anterior se tienen n predicados o características y cada característica puede tener j, k ... o l diferentes estados de carácter.

En este esquema, el usuario, más que responder con verdadero o falso, indicará el estado de carácter para los caracteres. El sistema se debe encargar de eliminar aquellos Ti que no contengan a ese predicado con ese argumento.

Con este esquema se pueden eliminar algunos Ti cuando el usuario responda una instancia de predicado como falsa: se eliminan todas aquellas Ti en que aparezca afirmada y las demás instancias del mismo predicado aparezcan negadas. No se podrán eliminar aquellas Ti en las que aparezca afirmada más de una instancia, a menos que el usuario indique que también son falsas:

Analícese el caso para una respuesta A11 a P1:

Si todas las instancias están negadas menos la que el usuario responde como falsa, entonces ese Ti se elimina:

$$P1(A11) \vee \neg P1(A12) \ldots \vee \neg P1(A1j)$$

Si no todas las instancias están negadas y el usuario responde sólo una como falsa, entonces ese Ti no se elimina:

$$P1(A11) \vee P1(A12) \ldots \vee \neg P1(A1j)$$

En general, si todas las instancias están negadas menos las que el usuario responde como falsas, entonces ese Ti se elimina.

Una ves más se observa como el tratamiento de la negación es algo más complejo que el de la afirmación, pero su incorporación adiciona información que hace arribar mas rápido o con menos respuestas, a una diagnosis final.

V. 4) Enfoque Probabilistico

El enfoque probabilistico tratado en esta sección, se refiere a la incorporación de valores de incertidumbre en el enfoque matricial, así como un esquema de probabilidad condicional.

El enfoque de Redes Neuronales también puede ser visto como un enfoque probabilistico, pues existe mucha relación entre estas dos metodologías, como asi se observa en Wan (1990). Más adelante se dedica un subcapitulo a Redes Neuronales.

V. 4. a) Probabilidad en matrices

Recordando el esquema de tratamiento mediante matrices (sección V.1), se observa que la información contenida en la matriz son valores binarios, o uno o cero, y no se aceptan valores intermedios. El aceptar valores intermedios, significa, en el enfoque de probabilidades en matrices, precisamente la probabilidad de que ocurra ese hecho. Así, cuando se tienen valores binarios, solo se dice que un hecho ocurre o no ocurre.

Dado un ejemplar para determinación y una matriz de datos de los posibles taxa, se pueden obtener valores probabilisticos para los posibles taxa. Esto se logra integrando a la matriz de datos un valor de probabilidad para cada elemento de la matriz, por ejemplo, dada la matriz M:

		CLASES				
CARACTERISTICAS		A	B	C	D	E
C1	a	0	1	0	1	1
"	b	1	0	1	1	0
C2	c	0	0	0	0	1
"	d	1	0	0	1	0
"	e	1	1	1	1	0
"	f	0	1	0	1	1

se puede incorporar un valor probabilistico a cada elemento de la matriz, para obtener la matriz MP:

		CLASES				
CARACTERISTICAS		A	B	C	D	E
C1	a	0	1	0	0.8	1
"	b	1	0	1	0.2	0
C2	c	0	0	0	0	0.5
"	d	0.3	0	0	0.2	0
"	e	0.7	0.4	1	0.5	0
"	f	0	0.6	0	0.3	0.5

Obsérvese que los valores cero en la matriz M también tienen probabilidad cero en la matriz MP. Además, la suma de las probabilidades de un mismo carácter es uno, por ejemplo, el carácter C2, que está compuesto por los estados c, d, e y f, suma uno para todos los taxa.

V. 4. b) Probabilidad Condicional

Otro método probabilístico es mediante probabilidad condicional, utilizando el Teorema de Bayes. En el contexto de la identificación y suponiendo que el conjunto POS de posibles clases a las que puede pertenecer el ejemplar Ej es exhaustivo y excluyente, la fórmula de Bayes es:

$$P(Ti/Ej) = \frac{P(Ti)/P(Ej/Ti)}{\Sigma\ i->n\ (\ P(Ti)\ *\ P(Ej/Ti)\ )}$$

donde:

- P(Ti/Ej) es la probabilidad de que el ejemplar Ej pertenezca al taxon Ti.
- P(Ti) es la probabilidad "a priori" de que el ejemplar Ej pertenezca al taxon Ti.
- P(Ej/Ti) es la probabilidad de que el taxon Ti presente las caracteristicas del ejemplar Ej.
- n = | POS |
- Σ i->n es la suma de los elementos i hasta el n.

Aplicar el Teorema de Bayes en identificación significa encontrar el Ti que tenga la probabilidad "a posteriori" mayor.

Pankhurst (1978) propone calcular a P(Ej/Ti) mediante la fórmula:

$$P(Ej/Ti) = \prod_{j=1}^{n} (P(Ci/Ti))$$

donde:

- $\prod_{j=1->n}$ es el producto de j hasta n.
- P(Ci/Ti) es la probabilidad de que el taxon Ti presente el carácter Ci.

Desde este punto de vista, aplicar el Teorema de Bayes al proceso es una especialización del enfoque de matrices, pues para cualquier problema de diagnosis, el conjunto de los taxa a los que puede pertenecer el ejemplar son los mismos, con la diferencia de que en el método de Bayes, a cada taxon se le asigna una probabilidad.

En realidad, la aplicación del Teorema de Bayes puede ir más lejos, pues hasta aquí se ha dicho que la probabilidad de los estados de carácter (Ci) son un número para cada taxon, pero estas probabilidades dependen del nivel taxonómico inferior al que intenta diagnosticar la clave. Por ejemplo, si la clave es para géneros, entonces las probabilidades de los estados de carácter varían según la especie, es decir, las probabilidades de los estados de carácter dependen de la presencia de otro estado de carácter. Es interesante, entonces, que este proceso pueda ser implementado mediante un sistema dinámico con retroalimentación probabilística, en donde se tienen que definir los criterios del orden de los cálculos de la probabilidades.

El principal problema de la probabilidad condicional, es que es muy difícil encontrar o asignar las probabilidades para cada elemento de la matriz de datos. Además, los errores en la aproximación de cada valor hacen que el error final o acumulado sea, en ocasiones, muy alto.

Las probabilidades "a priori" dependen de muchos factores, entre ellos el lugar de colecta. Como México es un país relativamente poco colectado, no se tienen elementos suficientes para asignar estas probabilidades en muchos casos. Las probabilidades de P(Ej/Ti) dependen de diversos factores pero pueden ser calculadas a partir de la matriz de datos por simple comparación entre los taxa candidatos.

El resultado cualitativo puede estar controlado por un umbral que indique cuál es la probabilidad mínima para aceptar a un taxon como miembro del resultado de la diagnosis. Si este umbral es cero, entonces el resultado es cualitativamente igual si se utilizan probabilidades o si se utilizan matrices con valores binarios.

V. 5) Enfoque de Teoría de Preferencias

Otro enfoque interesante y poco recurrido en identificación biológica, es el de la Teoría de Preferencias, muy utilizada para construcción de modelos descriptivos y predictivos en las ciencias sociales.

A continuación se analizan algunos puntos de este enfoque, que se trata en Fichefet et al. (1984) bajo el nombre de "Modelos de Decisión de Multicriterio".

Si

A : es un conjunto de taxa
C1..Cn: son diversos criterios
R1..Rn: son relaciones de preferencia/indiferencia sobre A

a Ri b, con a,b ∈ A, significa que el taxon **a** se prefiere o es indiferente al objeto **b** con respecto al criterio Ci.

Se define a la relación de preferencia Pi y a la Indiferencia Ii por:

a Pi b <-> a Ri b ^ ¬(b Ri a)
a Ii b <-> a Ri b ^ b Ri a

Cuando A es un conjunto finito no vacío y cuando Ri es un orden débil (i.e. una relación que es transitiva: si **a Ri b ^ b Ri c -> a Ri c**; y conectada: **a Ri b o b Ri a**) entonces existe una función Ui sobre A tal que para todo **a, b ∈ A:**

a Ri b <-> Ui(a) >= Ui(b)
a Pi b <-> Ui(a) > Ui(b)
a Ii b <-> Ui(a) = Ui(b)

A Ui se le llama "función valor" o "función de utilidad ordinal" y al valor de Ui(a) se le llama la utilidad de a con respecto al criterio Ci.

Cuando Ri es de orden débil para cada i, los modelos de identificación se pueden clasificar en dos tipos:

I) Encontrar una "regla de selección colectiva", i.e. una relación funcional F, tal que para todo conjunto de relaciones débiles Ri, ..., Rn, se determina una y sólo una "relación de preferencia global" R = F(R1, ..., Rn) dado que a R b <-> a se prefiere o es indiferente a b con respecto al criterio R.

II) Encontrar una "función valor global" en W1 x W2 x ... Wn donde Wj= {Uj(a)|a∈A}, y tal que a se prefiere o es indiferente a b con respecto a todos los criterios C1,... ,Cn. donde

U(a) = U[U1(a),...,Un(a)] y U(b)= U[U1(b),...,Un(b)].

En I) las Ri representan a las características observables en el ejemplar, con las que se puede calcular la función F. El significado de R=F(R1,...,Rn) es que la diagnosis final es más probable que el ejemplar pertenezca a a que a b (a R b).

Las relaciones de preferencia se pueden representar mediante matrices cuadradas de taxa vs. taxa para cada estado de carácter:

Matriz de Preferencia para un estado de carácter X:

	T1	T2	T3	T4	...
T1	0	0	1	1	
T2		0	0	-1	
T3			0	-1	
T4				0	
.					
.					
.					

en donde el cero (0) representa indiferencia, uno (1) representa que se prefiere el renglón respecto a la columna y el menos uno (-1) que se prefiere la columna al renglón. Asi, la interpretación de la matriz anterior es:

Cuando se observe el estado de carácter X:

se prefiere el taxon T1 al T3
se prefiere el taxon T1 al T4
se prefiere el taxon T4 al T2
se prefiere el taxon T4 al T3

las demas relaciones son indiferentes.

Morales (1988) expone una variante de la teoría de preferencias bajo el nombre de Teoria de Test. En ese contexto

se habla de testores mínimos, como una combinación de característica con un valor alto para el proceso de diagnóstico.

V. 6) Enfoque de Arboles

Los árboles son una estructura de datos utilizada frecuentemente para representar decisiones y, en general, clasificaciones de diversos tipos. Existen varios dominios en los que se utilizan a los árboles como herramienta para la diagnosis.

Aunque las claves dicotómicas presentan algunas desventajas sobre las claves de multientrada o policlaves, también tienen características muy interesantes para representar conocimiento. De hecho, los árboles de decisión se plantean como una herramienta metodológica para la construccion de bases de conocimientos en Sistemas Expertos.

Argüello (1990) presenta algunas de las razones de porqué se han utilizado o no los árboles de decisión, además de presentar la línea de investigación que sigue el grupo que dirige el autor. Es interesante el algoritmo que describen para la construcción de árboles de decision y que difiere en algunos aspectos de los presentados en el campo de la taxonomía biológica para la construcción de claves dicotómicas.

En general se puede decir que los árboles de decisión constituyen una de las mejores herramientas para la identificación en aquellos casos en que son observables todos los caracteres que manipula la clave. A manera de propuesta, aquí se menciona que los árboles de decision, o claves dicotómicas, puedan agruparse para conformar "sistemas de claves", en donde el usuario puede accesar las diferentes claves de manera interactiva e interrelacionada. Así, se podrían concebir como sistemas intermedios entre las claves dicotómicas y las policlaves, pues serían sistemas de "semi-entradas". Estos sistemas permitirían no responder algunos de los nodos de las claves, tratando de avanzar automaticamente en otra clave y preguntar los nodos por los que no se habían recorrido en la clave anterior. A este respecto, el sistema ECLAD (Enseñanza de Claves Dicotómicas) construido por el autor (Murguía, 1990b), da algunas ideas de las ventajas de automatizar el recorrido de las claves, aunque no se muestran todas las ventajas de interrelacionar claves dicotómicas.

Otro grupo interesado en la automatizacion de recorridos de claves dicotomicas, es el dirigido por Rolando Hernández en Cuba (comunicación personal). El ha tenido la amabilidad de proporcionarme copias de sus trabajos a este respecto.

Es natural intentar recuperar la información y el conocimiento almacenado en las claves dicotómicas, entre ellos se encuentran conocimientos como:

- las características importantes para la diagnosis de los elementos del grupo,
- uno de los conjuntos de condiciones necesarios para diagnosticar cada elemento del grupo,
- una jerarquización de las características del grupo, pues el taxónomo sigue alguna estrategia para colocar unas características en nodos iniciales de la clave y otras en nodos finales.

Las Policlaves y Claves dicotómicas

A partir de una matriz de datos se pueden construir claves dicotómicas del grupo. Esto podría insinuar que la información que tiene una policlave es mayor que la que contiene una clave dicotómica. En términos generales se puede afirmar lo anterior, pero el hecho es que la información estructurada en una clave dicotómica genera información adicional.

En la figura 8 se muestra una clave dicotómica, construida a partir de la información de una matriz. En esa clave dicotómica se observa cómo se pierde información; por ejemplo, no se dice nada acerca de la característica C para los TAXA III y IV. También se observa cómo el taxon I aparece bajo el camino B y no B, pues en la matriz de datos contiene un uno en la característica B y también en la característica no B, puede o no presentar B.

Tschudi (1988) y Argüello (1990) discuten la relación entre árboles y matrices, que puede analogarse a la relación entre claves dicotómicas y policlaves. En general, también la relación entre policlaves y matrices puede extraerse de las metodologías para construcción automatica de claves dicotómicas, como en Morse (1971).

V. 7) Enfoque de Sistemas Expertos

Los orígenes de los Sistemas Expertos se remontan al año de 1965, cuando un grupo de investigadores dirigidos por Feigenbaum construye el programa DENDRAL, que auxilia en la elucidación de

		ESTADOS DE CARACTER					
		A	B	C	no A	no B	no C
TAXA	I	1	1	0	0	1	1
	II	1	0	1	0	1	0
	III	0	1	0	1	0	1
	IV	0	0	1	1	1	1

A partir de la matriz de datos anterior se puede construir la siguiente clave dicotómica:

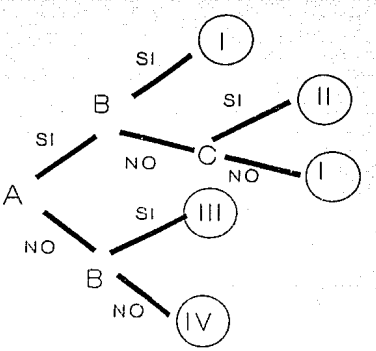


Fig. 8: Pérdida de información en una clave dicotómica.

la estructura molecular de compuestos orgánicos (Dussauchoy y Chatain, 1988). El primer sistema experto formal fue precisamente un sistema de diagnóstico: MYCIN, que diagnostica enfermedades infecciones de la sangre.

El enfoque de Sistemas Expertos actualmente es muy recurrido para los procesos de diagnóstico. Esta tecnología es muy utilizada para resolver problemas en donde interviene conocimiento especializado, que se captura en un programa que lo sabe manipular para resolver problemas concretos. Uno de los principales grupos de aplicación es el de diagnóstico, desde diagnóstico médico, hasta el de averías en locomotoras.

Existen publicaciones sobre Sistemas Expertos a diversos niveles de profundidad. Además, existen revistas especializadas que presentan aplicaciones y otras que presentan mejoras o nuevas técnicas en el tratamiento de problemas de diagnóstico usando Sistemas Expertos.

En general, la arquitectura de un sistema experto consta de una Base de Conocimientos, una Máquina de Inferencias y una Interfase con el usuario. En la figura 9 se muestra el arreglo de estos componentes en el sistema, de acuerdo con Dussauchoy (1988). Los Sistemas Expertos más completos tienen un módulo para explicar el razonamiento seguido en la solución de los problemas y otro módulo para la adquisición de más conocimientos.

La interfase con el usuario puede ser basada en menús o mediante un diálogo en lenguaje natural. En ocasiones ya se incorporan reconocedores de voz.

La base de conocimientos contiene los conocimientos especializados que los expertos en el dominio poseen. Aquí se almacenan los objetos que intervienen en la solución de problemas y sus relaciones; además, se consideran las excepciones y casos particulares. Estos conocimientos pueden almacenarse en diversos tipos de estructuras. La estructura más recurrida son las reglas.

La máquina de inferencias se encarga de hacer operativo el conocimiento almacenado. Realiza deducciones e inferencias a partir de hechos básicos, simples, que el usuario proporciona al sistema. También existen diferentes estrategias para tratar el conocimiento. En caso de que el conocimiento esté representado mediante reglas, las estrategias se pueden dividir en búsquedas de solución de los hechos a las hipótesis y búsquedas de las hipótesis a los hechos.

Las ventajas de una representación de conocimiento mediante reglas es que no es indispensable definir la totalidad de relaciones entre los objetos, como comunmente se tiene que hacer

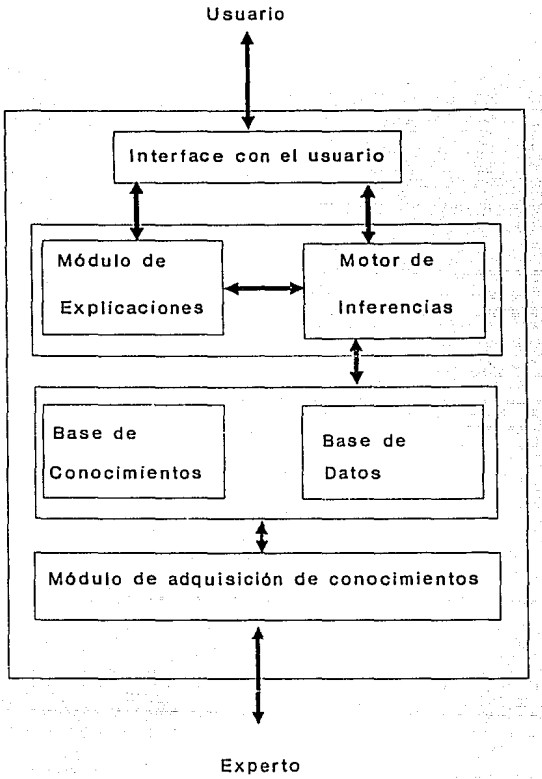


Fig. 9: Arquitectura básica de los Sistemas Expertos.

en el enfoque matricial. Solamente se indican aquellas relaciones importantes en el sentido de que son las más valiosas para realizar inferencias y deducciones. Además, las reglas son independientes entre sí; así, las reglas se pueden adicionar, quitar o modificar sin que esto repercuta en la arquitectura del conocimiento. Otra ventaja es que las reglas se pueden adicionar poco a poco y el sistema puede ser funcional antes de que esté totalmente terminado, o bien, el sistema se puede ir mejorando, considerando nuevos casos o casos no previstos en la construcción del sistema prototipo.

Explicaciones accesibles sobre Sistemas Expertos se encuentran en textos generales sobre Inteligencia Artificial, como en Rich (1983) y Frenzel (1986). Además, para implementar pequeños modelos, se puede recurrir a libros de texto sobre lenguajes de programación para la Inteligencia Artificial, que casi siempre brindan ejemplos de taxonomia biológica, aunque caricaturescos, como en Ford (1989) y en Winston y Horn (1984). Una perspectiva en general de cómo incorporar metodologías de la Inteligencia Artificial a la Botánica se discute en Murguía (1990a).

V. 8) Enfoque de Redes Neuronales

La tecnología de redes neuronales es utilizada para resolver problemas de diagnóstico y clasificación. Surge a partir de la cibernética y son Wiener y Rosenblueth a quienes se les atribuye el surgimiento y creación de esta disciplina, híbrida de la Informática y las Neurociencias (Arbib, 1987).

La arquitectura de redes neuronales es variable y las metodologías para su construcción, así como los cálculos que se realizan dentro de ellas varían, pues existen varias escuelas y enfoques.

En términos generales, una red neuronal consta de una capa de neuronas de entrada, una o más capas intermedias y una capa de neuronas de salida.

Las neuronas de la capa de entrada se comunican por una o varias sinapsis a las capas intermedias, y las neuronas de la capa intermedia se comunican, también por una o varias sinapsis a las neuronas de la capa de salida.

La función de una red neuronal, desde el punto de vista de diagnóstico, es asignar un vector de salida (mediante la activación o inactivación de las neuronas de la capa de salida) a cada vector de entrada. Las conexiones entre las neuronas tienen asignadas fórmulas, con pesos que calculan un valor de

actividad para cada neurona y que dependen de los valores de las señales que llegan a cada neurona.

La red neuronal debe ser entrenada con un conjunto de vectores de entrada y salida deseada para cada vector. Las salidas reales se comparan con las salidas deseadas y se realiza un ajuste automático de los pesos de cada sinapsis.

Es interesante notar que las redes neuronales pueden proporcionar un vector como salida. El vector de salida puede interpretarse como una clase o grupo taxonómico. Se pueden reservar ciertas posiciones del vector de salida para indicar la presencia de propiedades emergentes del vector de entrada.

Si se representa al ejemplar como un vector de información presencia-ausencia de sus estados de carácter, por ejemplo un vector de e elementos:

[0 0 0 1 1 0 ... 1 0 1 0]

Así, la entrada de información será un vector de este estilo mediante las neuronas de la capa de entrada.

Las neuronas de la capa de salida pueden representar a cada clase. Si se tienen c clases posibles, la red contendrá c neuronas en la capa de salida:

Notación:

e número de neuronas de la capa de entrada
c número de neuronas de la capa de salida
n número de capas intermedias
Ei i-ésima neurona de la capa de entrada,
Si i-ésima neurona de la capa de salida,
Nji i-ésima neurona de la capa intermedia j

Considerando la notación anterior, la red puede tener el siguiente aspecto:

E_1 E_2 E_3 E_4 ... E_m

N_{11} N_{12} N_{13} N_{14} ... N_{1j1}

N_{21} N_{22} N_{23} N_{24} ... N_{2j2}

.
.
.

N_{n1} N_{n2} N_{n3} N_{n4} ... N_{njn}

S_1 S_2 S_3 S_4 ... S_n

La red neuronal puede ser alimentada con un vector de entrada de e elementos y reportará un vector de n elementos:

vector de entrada: [0 1 0 0 1 1 0 1 0]
vector de salida: [0 0 0 1 0 0]

lo que indica que el ejemplar pertenece a la cuarta clase.

El número de neuronas de cada capa es determinado, generalmente de forma arbitraria, así como el número de capas intermedias.

La red neuronal es construida fijando un cierto número de capas intermedias y su respectiva cantidad de neuronas, también se fija el número de neuronas de la capas de entrada y salida, así como la interpretación de cada neurona (presencia o ausencia de cierto estado de carácter para las neuronas de la capa de entrada y una clase para las neuronas de la capa de salida).

El número de sinápsis tambien es fijado de manera arbitraria y pueden conectar neuronas de capas contiguas o de capas no contiguas.

Cada sinápsis tiene una fórmula asociada con parámetros específicos, la actividad de una neurona depende directamente de la actividad de sus neuronas presinápticas. Uno de los parámetros de cada neurona de la capa de salida es un valor umbral que determina su actividad o inactividad.

El proceso de construcción de una red es el siguiente: Se establece un conjunto de "vectores de entrenamiento" que son una muestra representativa de los casos que identificara la red. Para cada vector de entrenamiento se debe conocer el vector

resultado deseado. El usuario introduce cada vector de entrenamiento y observa el valor del vector de salida e indica al programa si es la salida deseada o no, en éste último caso debe indicar cual es la salida correcta. En este proceso, la red reajusta los parámetros de las sinápsis de la red. Es necesario un cierto número de iteraciones hasta obtener salidas correctas. Al terminar ese proceso la red está debidamente entrenada y podrá identificar de manera correcta la mayoría de los casos.

Existen diferentes algoritmos para ajustar los pesos de las sinápsis, pero en general, el proceso de entrenamiento de la red es el mismo.

Existen paquetes para generar redes neuronales, como Neural Networks (NeuralWorks Professional II, 1989), a través de sólo indicar el conjunto de entrenamiento. Algunos dan acceso a modificar los pesos en las conexiones y a elegir el tipo de arquitectura de la red y algoritmo de aprendizaje.

Una de las desventajas de esta tecnología es que se requiere del conjunto de vectores de entrenamiento, es decir, ejemplos de conjuntos de características con su correspondiente diagnosis correcta.

Además, como en el proceso de construcción de la red se realizan muchas operaciones aritméticas, generalmente se requieren computadoras rápidas y con mucha memoria. Como cada neurona generalmente se conecta con más de una neurona, el número de sinapsis crece más que aritméticamente, con el número de neuronas. Engel y Cran (1990) dan un ejemplo sencillo de diagnóstico de siete patrones, en donde las entradas son vectores de 10 valores binarios y se expone una manera simple de comparar los resultados de la red con los de los humanos.

Shoemaker (1991) propone a las Redes Neuronales como medio para aproximar las probabilidades condicionales en contextos de clasificación. Stirling y Morrell (1991) exponen un método interesante que utiliza Teorema de Bayes, pero incorporan la ayuda externa del usuario.

Geva y Sitte (1991) realizan una comparacion entre un método estadistico-espacial y el enfoque de Redes Neuronales, en particular con el algoritmo de retropropagación.

Erkmen y Stephanou (1990) exponen conceptos interesantes que analogan la Termodinámica a la Teoria de la Información mediante el concepto de entropía. Mediante Teoría de Fractales se expone cómo medir la relevancia de una pieza de evidencia en una base de conocimientos. Además, es interesante cómo se pueden tratar diversos niveles de granularidad de la información en un mismo sistema.

resultado deseado. El usuario introduce cada vector de entrenamiento y observa el valor del vector de salida e indica al programa si es la salida deseada o no, en éste último caso debe indicar cual es la salida correcta. En este proceso, la red reajusta los parámetros de las sinápsis de la red. Es necesario un cierto número de iteraciones hasta obtener salidas correctas. Al terminar ese proceso la red está debidamente entrenada y podrá identificar de manera correcta la mayoría de los casos.

Existen diferentes algoritmos para ajustar los pesos de las sinápsis, pero en general, el proceso de entrenamiento de la red es el mismo.

Existen paquetes para generar redes neuronales, como Neural Networks (NeuralWorks Professional II, 1989), a través de sólo indicar el conjunto de entrenamiento. Algunos dan acceso a modificar los pesos en las conexiones y a elegir el tipo de arquitectura de la red y algoritmo de aprendizaje.

Una de las desventajas de esta tecnología es que se requiere del conjunto de vectores de entrenamiento, es decir, ejemplos de conjuntos de características con su correspondiente diagnosis correcta.

Además, como en el proceso de construcción de la red se realizan muchas operaciones aritméticas, generalmente se requieren computadoras rápidas y con mucha memoria. Como cada neurona generalmente se conecta con más de una neurona, el número de sinapsis crece más que aritméticamente, con el número de neuronas. Engel y Cran (1990) dan un ejemplo sencillo de diagnóstico de siete patrones, en donde las entradas son vectores de 10 valores binarios y se expone una manera simple de comparar los resultados de la red con los de los humanos.

Shoemaker (1991) propone a las Redes Neuronales como medio para aproximar las probabilidades condicionales en contextos de clasificación. Stirling y Morrell (1991) exponen un método interesante que utiliza Teorema de Bayes, pero incorporan la ayuda externa del usuario.

Geva y Sitte (1991) realizan una comparación entre un método estadístico-espacial y el enfoque de Redes Neuronales, en particular con el algoritmo de retropropagación.

Erkmen y Stephanou (1990) exponen conceptos interesantes que analogan la Termodinámica a la Teoría de la Información mediante el concepto de entropía. Mediante Teoría de Fractales se expone como medir la relevancia de una pieza de evidencia en una base de conocimientos. Además, es interesante como se pueden tratar diversos niveles de granularidad de la información en un mismo sistema.

V. 9) Análisis comparativo de los enfoques presentados

Todos los enfoques presentados contienen tanto elementos de tipo lógico-deductivo como estadístico-inductivo. Los enfoques de Lógica Matemática y Teoría de Conjuntos pueden combinarse con enfoques estadístico-probabilísticos; el enfoque de Sistemas Expertos, es un enfoque lógico pero que incorpora ponderaciones de tipo estadístico y el enfoque de Redes Neuronales es un enfoque totalmente inductivo, pues los métodos para construir la red utilizan "ejemplos" y no se basan en análisis deductivos del conocimiento. El enfoque de Probabilidades está más inclinado hacia el lado estadístico-inductivo. Los enfoques de Teoría de Preferencias y Arboles pueden verse como intermedios.

Los enfoques difieren en cuanto al tipo de representación de la información y del conocimiento taxonómico y en consecuencia también en el tipo de algoritmos que operan sobre dichas representaciones. Los algoritmos que operan sobre representaciones estructuradas de conocimiento son más complejos que aquellos que operan sobre representaciones no estructuradas; pero los beneficios de usar algoritmos que operen sobre representaciones estructuradas son grandes. Así, es más sencillo construir algoritmos para operar matrices que para operar fórmulas de lógica de predicados, pero la representación lógica es más sencilla y práctica, pues una fórmula lógica adquiere más significado para el taxónomo que una matriz. La mayoría de los esfuerzos encaminados a automatizar operaciones lógicas se agrupan bajo los términos de "Programación Lógica" y de "Demostración Automática de Teoremas"; un logro grande en este sentido es el lenguaje de programación Prolog.

Aunque en la actualidad la representación mas difundida y utilizada son matrices, en Sistemas Expertos se encuentra quizá una forma más adecuada o entendible por los taxónomos de representar el conocimiento taxonómico. La ventaja de la representación en el enfoque de Sistemas Expertos es que el método fué concebido para dominios de aplicación con información incompleta.

Algunos de los enfoques presentados cuentan con paquetes en el mercado específicamente para la automatización de procesos de identificación (tabla VIII). Aunque algunos enfoques han sido poco o nulamente considerados en la práctica de la identificación taxonómica, éstos pueden ser experimentados a corto plazo utilizando "esqueletos" de Sistemas Expertos o paquetes para construir Redes Neuronales.

Enfoque	Capacidad para manejar información incompleta	Representación estructurada	Paquetes para su manejo
Matrices	n	n	s
Teoría de Conjuntos	n	n	n
Lógica Matemática	s	s	n
Probabilidad	n	n	n
Teoría de Preferencias	n	n	n
Arboles	s	n	n
Sistemas Expertos	s	s	s
Redes Neuronales	s	n	s

Tabla VIII. Algunas caracteristicas de los diferentes enfoques metodológicos para abordar el problema de la automatización de la identificación taxonómica.

Aunque cada uno de los enfoques presentados tiene sus caracteristicas propias que lo identifican. La ventaja o desventaja de automatizar con cada enfoque depende del contexto. La elección de un método particular no sólo depende de las cualidades y cantidades de la información a automatizar, sino también de los recursos humanos, económicos y tecnológicos involucrados, así como de las metas que persiga la institución y los individuos al construir una clave.

A continuación se listan preguntas cuya respuesta ayuda a la elección del método a usar en la construcción de una clave en computadora:

¿Qué recursos consume cada método?
¿Quienes pueden participar en la construcción de la clave?
¿Qué experiencia tiene la institución y los individuos en el uso de cada una de las diferentes aproximaciones?
¿Qué conocimientos se requieren para construir la clave?
¿Qué conocimientos requerirá el usuario para usar la clave?
¿De qué información se dispone para construir la clave y para identificar?
¿Qué objetivos se persiguen al construir la clave?
¿Dentro de qué contexto funcionara la clave?
¿De qué maneras se pude automatizar la relación entre la identificación y otros procesos dentro del proyecto?

B) Seleccionar aquellas características que estén contestadas para la mayoría de los taxa en la base de datos. Es decir, dar un peso mínimo a las características que estén indeterminadas en la base de datos:

$$- q_i$$

q_i = fracción de especies que permanecen como posibles y que tienen indeterminado el carácter i.

C) Seleccionar aquella característica que la mayoría de los taxa la tengan como verdadera o falsa, pero un mínimo de ambas:

$$- \frac{| Yi - Ni |}{Yi + Ni}$$

Yi = Número de taxa que tiene si para la característica i
Ni = Número de taxa que tienen no para la característica i

Existen otros criterios que dependen del uso previo que se le haya dado a la clave. En estos programas, se almacenan todas las sesiones previas y con base en esas experiencias, el programa decide cuál es la mejor pregunta a realizar. Por supuesto que este tipo de estrategias sólo funciona para programas que se utilizan en proyectos formales, en donde los ejemplares determinados van guiando al programa, ayudándolo a generar probabilidades.

Una estrategia de este tipo es no preguntar aquellas características que se han utilizado en determinaciones erróneas previamente. Esto es posible, si en las diagnosis se lleva un registro de los errores que se han cometido, pudiendo preguntar al usuario directamente, o detectándolos automaticamente, cuando ningún taxa permanece como posible:

$$- \frac{ei}{Sr}$$

ei = El número de errores que se han encontrado para el carácter i, tomando en cuenta los taxa que permanecen como posibles.

Sr = Número de taxa que permanecen como posibles.

También pueden evitarse las preguntas que menos se han podido responder por el usuario:

$$- \frac{di}{Sr}$$

di = El número de veces que la pregunta se ha respondido como desconocida: "?"

Sr = Número de taxa que permanecen como posibles al responder la pregunta i.

También pueden detectarse pares de preguntas que se responden a la vez como desconocido, lo que detecta que los ejemplares están incompletos o no observables en partes relacionadas o en la misma estructura:

$$- \frac{dij}{Sr}$$

di = El número de veces que las preguntas j e i se han respondido como desconocida "?" a la vez. Sabiendo que para el ejemplar en determinación se ha respondido j como desconocido.

Sr = Número de taxa que permanecen como posibles al responder la pregunta i.

Chou (1991) expone un algoritmo para construcción de árboles de decisión de mas de 10 características y explica algunos métodos para ir partiendo el universo en mitades.

VI.2) Propuesta: Teoria de Test para elegir la pregunta al usuario

Considerando una matriz de características vs. objetos, un testor es una combinación de características que permiten discriminar entre todos los objetos de la matriz. Pueden existir más de un testor para una sola matriz de datos. Un Testor Tipico es un Testor en el que todas las características que lo componen son necesarias para mantener su condición de Testor.

Es fácil comprender que un buen criterio para elegir las preguntas al usuario es escogiendo un conjunto de características que sean un testor típico.

Como para una matriz pueden existir más de un testor típico, un criterio puede ser elegir el testor típico más pequeño. Otro criterio puede ser elegir el testor que contenga características con mayor peso informacional, ese peso informacional puede ser asignado directamente por el especialista o bien calculado a partir del conjunto de testores típicos. En general, algunos de los criterios explicados en la sección anterior pueden utilizarse para el cálculo de peso informacional de las variables.

Se observa como pueden definirse estrategias de recorrido de los testores, típicos o no, según información del peso de las variables o de los mismos testores o conjuntos de variables.

Morales (1988) expone un método para calcular los testores típicos mediante un algoritmo que utiliza "backtrak" y que se basa principalmente en el concepto de error de discriminación, es decir dada una combinación de características, su error es el número de pares de objetos que no se discriminan.

Capítulo VII

Propuesta: Claves Dinámicas

Los individuos son los elementos que se observan de manera concreta y son susceptibles de ser estudiados y analizados, son sólo parte de un continuo en el devenir del organismo. Además de que los organismos se manifiestan en diferentes formas, variando en más de una dimensión, las poblaciones de organismos también son dinámicas. Respecto a identificación, surge la pregunta ¿Cómo debe cambiar la definición de las clases si las clases también son cambiantes? o bien, ¿Cuales son los atributos que no varían, para con ellos definir a las clases?. Es sabido que las características morfológicas o anatómicas de las clases (especie) varían, también las relaciones dentro y entre las clases varían, cuando menos, con el tiempo y el espacio. Quizá en niveles de abstracción superiores, el cambio sea menor, y la definición de clasificaciones más estables sea posible. Quizá sean las meta-relaciones, o estructuras de relaciones las más constantes.

Mientras no se encuentre una estructura estable, será necesario que el concomimiento participe en el cambio de la naturaleza. Quizá sea la manera de establecer el cambio de las clasificaciones lo que sea más estable, es decir, considerar herramientas dinámicas que varíen acorde a como también varían los objetos de estudio.

La automatización de sistemas dinámicos de la naturaleza planteada es una tarea compleja, pero por lo pronto es posible hacer explícita la necesidad de métodos adaptables a procesos dinámicos en la clasificación y la identificación biológica, y dar al usuario las herramientas para que controle o pueda seguir esos cambios.

VII. 1) Definición empírica

El concepto de clave dinámica dado aquí no es en el sentido expresado por Morse (1971), que lo utiliza como sinónimo de policlave "on-line". En ese entonces, la identificación automática se observaba desde dos puntos de vista: proceso en "batch" o proceso "on-line". El proceso en batch consiste en alimentar al programa y esperar resultados, no se permite una interacción de información entre el programa y el usuario durante el tiempo de ejecución del programa. Con la posibilidad de incorporar procesos "on-line", la identificación se vió beneficiada por el hecho de que las computadoras permitían establecer un diálogo, es decir, que el usuario pudiera introducir información al programa, el programa brinda resultados parciales y el usuario puede de nuevo introducir datos, proceso que en la actualidad es más común.

Esa etapa ha sido superada y es casi implícito el hecho de que los procesos de identificación automática se realizan "on-line" y no en "batch".

Aquí se aborda el concepto de "dinámico" desde un punto de vista, también de interacción del usuario con el programa, pero poniendo especial atención en la estructura de la información de la policlave.

Como se explicó en el capítulo II, en un sistema de identificación existen varios elementos importantes: el ejemplar bajo determinación, el sistema de clasificación, el nivel taxonómico de identificación, los taxa candidatos, los métodos y herramientas de identificación y el identificador.

El Sistema de Clasificación y las características observables en el ejemplar, son los elementos en que se fundamenta la identificación, ya que es mediante los caracteres que se inicia la identificación, es decir, el usuario indica las características que presenta el ejemplar. Entonces la identificación se realiza de las características a los taxa.

Sin embargo, en varias ocasiones el proceso de identificación se realiza en sentido inverso, es decir, se propone a uno o varios taxa como probables y se intentan verificar estas hipótesis. La introducción de características ahora está guiada por las hipótesis sobre los taxa. Por lo que se realiza una identificación de los taxa a las características.

El proceso de identificación de taxa a características puede ser observado cuando se intenta utilizar una clave dicotómica al revés, cuando se intenta corroborar que el ejemplar pertenece a algún taxon que el identificador cree más adecuado.

La propuesta es, entonces, brindar al identificador mas flexibilidad en el manejo de la información durante el proceso.

Aún no es posible incorporar mucho del conocimiento sobre identificación en programas de computadoras, pero sí es factible brindar al usuario las mayores facilidades para accesar y procesar la información que se pueda almacenar.

Así, una clave dinámica debe permitir al usuario una identificación de las características a los taxa, como tradicionalmente se hace, pero también una identificación de los taxa a las características. Más aún, debe permitir una identificación mixta, en donde se van introduciendo características presentes en el ejemplar y también se pueden indicar los taxa que el usuario crea mas probables. En el proceso de identificación, más que "brincar" de una estrategia a la otra con facilidad, se establece una estrategia híbrida, en donde las características que introduce el usuario están guiadas por las mismas hipótesis que se plantean y viceversa.

En el sentido anterior, ya algunos autores indican que la intervención del usuario en el proceso de identificación automatizada es valiosa (Stirling y Morrell, 1991).

VII. 2) Modelo Formal

En una clave dinámica, durante el proceso de determinación se tiene acceso al sistema de clasificación de una forma más directa y flexible que en otras herramientas para determinación.

De acuerdo al conocimiento del grupo que la clave contiene y que el usuario posee, las características pueden clasificarse en:

- Características que presenta el ejemplar
- Características que no presenta el ejemplar

- Características que presentan las clases (taxa)
- Características que no presentan las clases

- Características que el usuario sabe que presenta el ejemplar.
- Características que el usuario sabe que no presenta el ejemplar.

Las cuatro primeras clases dependen de la construcción de la clave, de qué tan validada esté la información que posee la

clave. Las dos últimas dependen del usuario y del ejemplar bajo determinación.

La clave dinámica debe brindar al usuario herramientas y facilidades para poder observar y manipular las relaciones entre esas seis clases de características.

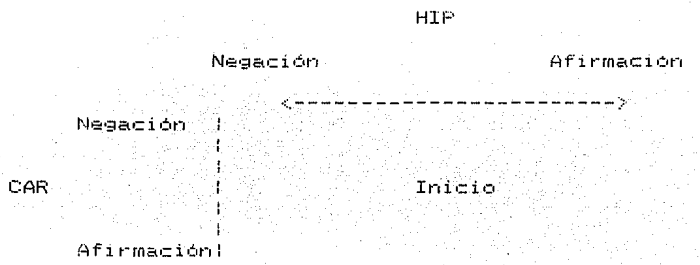
Otro tipo de información que el usuario posee es acerca de las clases (taxa). Un taxon puede tener ciertas intersecciones con la información contenida en la clave, pero pueden existir áreas de información que posea el usuario y que no estén representadas en la clave. Así, se debe permitir explotar al máximo la información relevante en la determinación del ejemplar.

Una forma en que la clave dinámica auxilia en ese sentido, es permitir al usuario indicar sus sospechas respecto a qué clases puede pertenecer el ejemplar. Así, la clave tendrá más información sobre el problema y podrá auxiliar al usuario en mayor medida. A su vez, el usuario, al recibir más información por parte de la clave, podrá avanzar más rápido hacia una diagnosis final y correcta.

De esta manera, se plantea un espacio de dos dimensiones, cada una representa una clase de información:

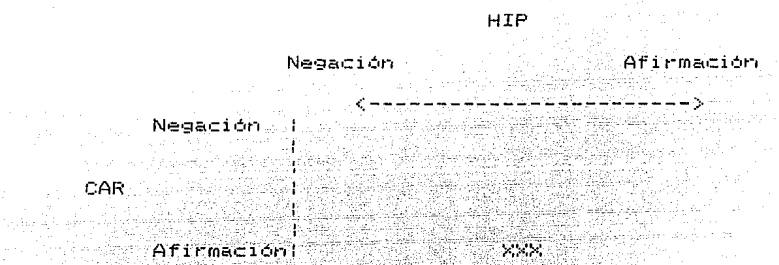
dimensión CAR: características del ejemplar
dimensión HIP: características de las clases

Cada dimensión tiene dos direcciones, la afirmativa y la negativa, es decir, para el caso de CAR, existen las características que presenta el ejemplar y las que no presenta el ejemplar, también para HIP existen las características de las clases y las que no presentan las clases.

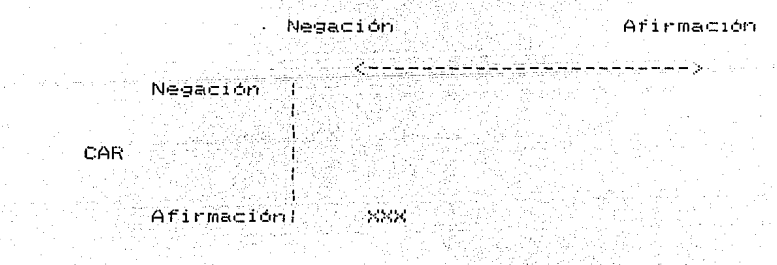


El punto intermedio entre Negación y Afirmación es un punto Neutro. Así, al iniciar la determinación, el usuario está ubicado en el centro (Inicio), pues no ha adicionado información sobre características en el ejemplar, ni sobre sospechas de a qué clases puede pertenecer el ejemplar.

En el proceso de determinación, el usuario se mueve en este espacio de dos dimensiones. Si el usuario indica la presencia de ciertas características en el ejemplar y no ha indicado ninguna clase como tentativa, entonces se moverá hacia abajo:



El usuario puede indicar a qué clases definitivamente no considera que pertenece el ejemplar, por lo que se movería a la esquina inferior izquierda:



Como HIP es un conjunto de más de un elemento, las características afirmadas y negadas son el conjunto unión de las características de todas las HIP que el usuario indique (afirmadas o negadas).

La combinación que el usuario realiza al seleccionar características e hipótesis, afirmándolas o negándolas, puede representarse mediante el recorrido de la figura 10. El punto de partida es el cuadro central. El usuario se mueve en direcciones ortogonales. Las diagonales son resultado de dos movimientos ortogonales de diferente sentido (fig. 11). Como el usuario puede retractarse de alguna afirmación o negación indicada a la clave, entonces pueden existir flechas que van hacia el cuadro central.

La navegación en este espacio representa el proceso de determinación. La tarea del sistema es mantener la congruencia, es decir, cuando se adicione información se deben eliminar todas aquellos taxa que no cumplan con las restricciones puestas por el usuario. Además, el sistema debe mostrar qué elementos son los que conforman el sistema en cada paso. Así, cuando el usuario retire ciertas afirmaciones hechas con anterioridad, el sistema deberá incorporar los taxa que puedan entrar al sistema cumpliendo las restricciones restantes.

VII. 4) Implementación del modelo

La propuesta anterior se ha llevado a la práctica en ciertos aspectos. La implementación se realizó programando un ambiente que manipula las características y las hipótesis de forma interactiva con el usuario.

El ambiente está dividido en cuatro áreas (fig. 12):

1) **Enunciados:** Son las características que el usuario selecciona según las presente el ejemplar bajo determinación. Los enunciados (características) que se seleccionan, cambian de color en la pantalla de gris a blanco, mientras que los que no se han seleccionado permanecen de color gris. Esto facilita al usuario la elección de características, pues en realidad el usuario es el que indica a la computadora, sin que la computadora tenga que preguntar.

2) **Diagnosis:** Son los taxa a los que puede pertenecer el ejemplar bajo determinación, dados los enunciados que se han seleccionado en el área 1.

3) **Resumen de la Diagnosis:** Es un renglón en donde se indica cuantos enunciados se han seleccionado y cuantos taxa son los posibles. Al inicio de la diagnosis ese número es grande y a medida que se van seleccionando enunciados, disminuye. El estado óptimo es sólo un taxon posible.

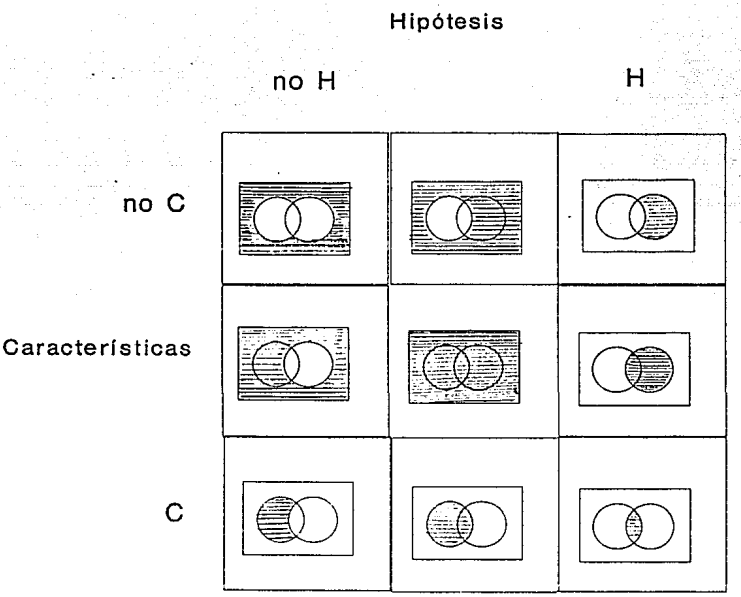


Fig. 10: Las nueve áreas del espacio de la Clave Dinámica.

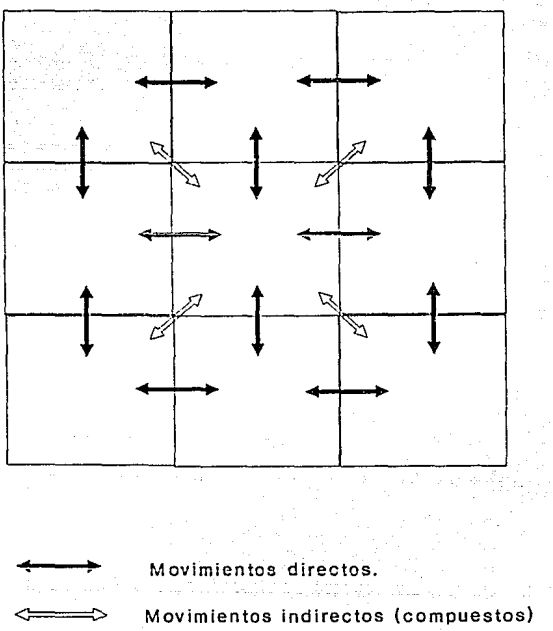


Fig. 11: Navegación lógica en la Clave Dinámica.

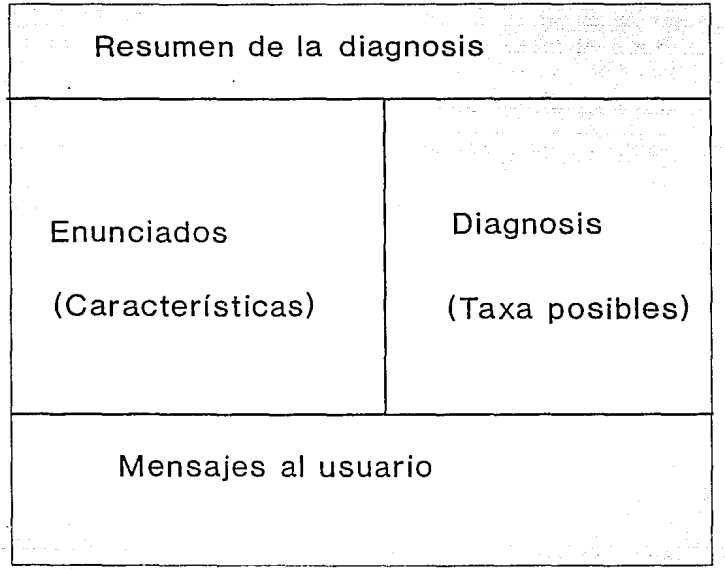


Fig. 12: Areas del ambiente de la Clave Dinámica.

4) **Mensajes al usuario:** Se proporciona mayor información al usuario sobre el enunciado que se desea seleccionar. Aún no se ha implementado el módulo que brinde más información para los taxa.

El sistema no está implementado en su totalidad respecto al modelo teórico planteado en la sección anterior, pues sólo considera afirmaciones y aún falta por implementar las negaciones. Pero aún esta versión limitada ha mostrado ser una herramienta agradable, fácil de usar y práctica en el laboratorio.

El ambiente está diseñado para que el usuario no tenga que cambiar de pantalla, de manera que en una sola pantalla tenga acceso a la información de cualquiera de los cuatro componentes explicados anteriormente.

Cuando el usuario selecciona una característica, ésta cambia de color gris a blanco y el área de Diagnosis muestra el nuevo conjunto de taxa que cumplen con esa restricción. El usuario puede retractarse de su afirmación volviendo a seleccionar la característica, lo que hará que el enunciado cambie de color blanco a gris. Cualquier enunciado seleccionado puede ser de-seleccionado de esta manera.

El usuario puede cambiar al área de diagnosis para observar la lista de posibles taxa. Tanto en el área de enunciados como en la de diagnosis, se pueden utilizar las teclas de flechas, página arriba y página abajo, así como la tecla inicio y fin.

El usuario puede indicar los taxa a los que el cree que pertenece el ejemplar con sólo presionar la tecla retorno sobre el taxon deseado. Al seleccionar un taxon, el efecto es que los enunciados que presenta ese taxon cambian a color verde. Los enunciados que el usuario indique como verdaderos para el ejemplar bajo determinación y que presentan los taxa que el usuario indico como tentativos, aparecen de color verde brillante. En general se tienen cuatro posibilidades (tabla IX), cada una se indica en el programa con un color diferente:

gris: enunciado no seleccionado para el ejemplar bajo determinacion y no presente en los taxa tentativos.

blanco: enunciado seleccionado para el ejemplar bajo determinación y no presente en los taxa tentativos.

verde obscuro: enunciado no seleccionado para el ejemplar bajo determinacion y presente en los taxa tentativos.

verde brillante: enunciado seleccionado para el ejemplar bajo determinacion y presente en los taxa tentativos.

	seleccionado para el ejemplar	presente en los taxa tentativos
gris	no	no
blanco	si	no
verde obscuro	no	si
verde brillante	si	si

Tabla IX. Colores para los diferentes estados de los enunciados en la clave dinámica.

El programa contiene ademas las siguientes funciones al presionar la tecla respectiva:

B. **Buscar.** Busca una subcadena en los enunciados o en los taxa. Por ejemplo puede buscar un enunciado que contenga la palabra aquenio.

N. **Nuevos datos.** Comienza la determinación de otro ejemplar. Se eliminan todas las indicaciones que se hayan dado durante la sesión previa.

C. **Contestar.** Estando en el area de hipótesis, contesta los enunciados que presenta la hipótesis seleccionada. Es el equivalente a presionar retorno en todas las caracteristicas que presenta la hipótesis.

Retorno. En el área de caracteristicas selecciona y de-selecciona caracteristicas. En el área de hipótesis, cambia de color a las caracteristicas que presenta la hipótesis seleccionada.

E. **Eliminar Hipotesis.** De-selecciona las caracteristicas de las hipótesis indicadas por el usuario (con retorno).

INS. El usuario puede pedir al sistema un resumen de los enunciados que ha seleccionado para el ejemplar bajo determinación

ESC. Sale del programa. Termina la sesión.

Capítulo VIII

GENCOMEX: Clave Dinámica para Géneros de Compuestas de México

Al programa explicado en la sección anterior se le incorporó una base de datos sobre géneros de Compuestas de México. La información fué proporcionada por el Dr. José Luis Villaseñor, Jefe del Herbario Nacional (MEXU) del Instituto de Biología de la UNAM. Dicha información consta de:

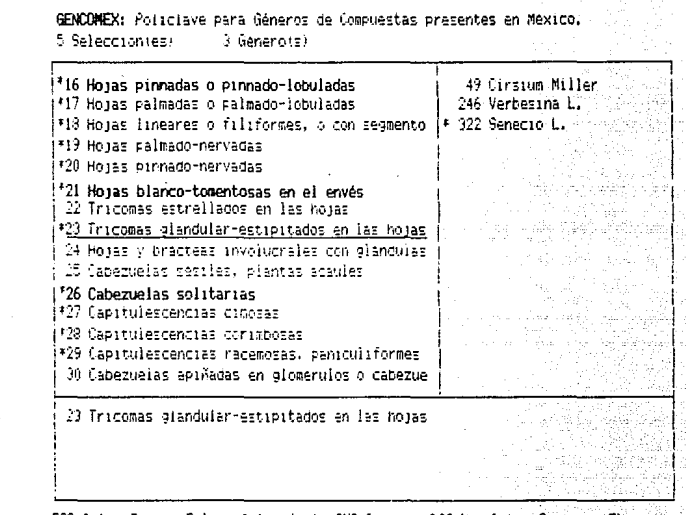
a) 92 enunciados que representan a 92 estados de caracteres.

b) 350 géneros, que representan el total de géneros de la familia Compositae presentes en México.

c) La matriz de datos, que es de 92 x 350.

La metodología para construir la matriz de datos fué la descrita por Murguía (1988). Es de esperar que contenga errores, pues la matriz de datos es extensa y muchos de los géneros no han sido estudiados de manera profunda. Pero al igual que la policlave FAMEX, a la matriz de datos se le harán correcciones con el tiempo, en la medida que se utilice por botánicos que intenten determinar material colectado en México, o incluso en otros países.

La policlave GENCOMEX (fig. 13) es una herramienta adicional para la determinación de Compuestas a nivel genérico. Aunque existen trabajos para la identificación de los géneros de la familia en México, como los de Rzedowski (1978) o Villaseñor (1986, 1989), la policlave GENCOMEX brinda un medio más fácil para la diagnosis. Un hecho evidente es que el uso de las claves dicotómicas que consideran más de 300 géneros es un verdadero reto. Aunque las claves de Rzedowski (1989) están divididas por Tribus, llegan a contener hasta un nivel de sangría de 31, lo que significa que para ciertos especímenes, el uso de las claves obliga al usuario a responder más de 31 preguntas. Las claves de Villaseñor (1987, 1989) son para regiones geográficas restringidas (Península de Yucatán y Tabasco y Cuenca del Río



Notacion:

Enunciado en donde esta posicionado el cursor.

Enunciados seleccionados para el ejemplar en determinación.

* Enunciados de la hipotesis del usuario.

Fig. 13: Ambiente del usuario de la policlave GENCOMEX

Balsas) consideran 100 géneros y 131, respectivamente. La primera contiene un glosario de términos taxonomicos de la familia, la segunda tiene una sangria máxima de 30, ademas, contiene claves para las especies.

Conclusiones

La identificación biológica es susceptible de automatizarse. Los esfuerzos por realizarla se basan en diferentes estrategias.

Para poder avanzar con pasos seguros, es necesario construir marcos formales de representación, tanto del mismo proceso de identificación, así como de las metodologías para su automatización.

La identificación biológica puede realizarse de las características hacia las clases, o menos común, de las clases a las características. Es posible construir herramientas en donde se haga uso explícito de ambas estrategias, para constituir metodologías híbridas.

Bajo el nombre de "Claves Dinámicas" se propone un sistema lógico, surgido de la lógica matemática, teoría de conjuntos y métodos de búsqueda de la Inteligencia Artificial, para facilitar el proceso de la identificación biológica.

Glosario

Arbol: Gráfica no cíclica en la que se agrupa a los nodos en niveles (0, 1, 2, ...). El nivel cero está constituido por un solo nodo llamado nodo **raiz**, del que parten relaciones a los nodos del nivel uno.

Arbol Dicotómico: Arbol en que de cada nodo sólo parten dos relaciones al nivel inferior.

Características Excluyentes: características que no pueden presentarse al mismo tiempo en un ejemplar, por ejemplo, el conjunto (árbol, arbusto, hierba).

Clasificar: Proceso en el que se crean clases o se identifica por vez primera un objeto con una clasificación preestablecida. En Biología la clasificación es un proceso distinto al de la **determinacion** o **identificación.**

Clave: Herramienta utilizada para el proceso de la **Identificación Biológica.**

Clave Dicotómica: **Clave** en la que se utiliza la representacion de **árbol dicotómico** (a veces tri- o tetra-tómico). Cada nodo del árbol representa una decision lógica (verdad o falso) entre dos conjuntos de **caracteristicas excluyentes.**

Clave Dinámica: **Policlave** en la que la identificación se logra mediante un proceso interactivo, no sólo de las características a las clases, sino también de las clases a las caracteristicas.

Clave On-line: Cualquier **Policlave** en computadora en la que a medida que el usuario introduce las caracteristicas del ejemplar el programa informa algo.

Descripción Taxonómica: Conjunto de **características taxonómicas** de un cierto grupo taxonómico o ejemplar.

Determinación: En este trabajo se considera sinónimo de **identificación biológica.**

Ejemplar: Unidad biológica que representa a otras unidades de su mismo tipo. Cada uno de los individuos de una especie.

Entrada: Decisiones que se realizan en una **clave.** Es la pregunta que la clave hace al usuario en un momento determinado del proceso. En las **claves dicotomicas** las entradas son cada nodo de la clave.

Enunciado: Proposición lógica que sólo puede ser falsa o verdadera al aplicarla a un ejemplar, por ejemplo "Plantas con jugo lechoso (látex)"

Estado de Carácter: Instancias de las características, por ejemplo, de la característica color sus estados de Carácter pueden ser verde, rojo y azul.

Hoja: Nodo de un **árbol** del que no parte relación alguna, pero si llegan a él. En un **clave dicotomica**, las hojas corresponden a los **taxa.**

Identificación Biológica: Proceso mediante el que se hace corresponder una instancia (organismo) con una clasificación previamente establecida.

Identificación Parcial: Identificación en la que la diagnosis final consiste en un conjunto de uno o más taxa, aunque el ejemplar pertenezca solo a uno de los taxa del conjunto resultado.

Matriz de datos: características utilizadas en la clave, independientemente de cómo se representen internamente.

Raíz: Primer nodo de un árbol. En una **clave dicotomica** está representado por la primer **entrada** a la clave, generalmente representada por un "1" o una "A".

Policlave: **Clave** en la que su estructura de datos es tal que permite al usuario abordar las decisiones en cualquier orden. Tiene muchas **entradas** a la vez.

Policlave en Tarjetas: Policlave en la que la **matriz de datos** se representa mediante perforaciones en tarjetas. Hay dos tipos básicos: 1) cuando las características corresponden a una tarjeta y las perforaciones a los taxa que la presentan, en este caso cada tarjeta lleva impreso un **enunciado;** y 2) cuando cada tarjeta representa un taxon y las perforaciones, las características que este taxon presenta.

Taxa: plural de taxon.

Taxon: grupo definido (clase) en un sistema de clasificación biológica. Por ejemplo: Aster, Solanaceae.

Bibliografía

Arbib,M.A. 1987. Brains, Machines, and Mathematics. Springer-Verlag. N.Y. 201pp.

Argüello,J.R. 1990. Arboles de Decision. Justificación de su estudio, algoritmos y usos. Memorias del 2o. Congreso Iberoamericano de Inteligencia Artificial. Morelia. pp1-48.

Bachelard,G. 1982. La formación del espiritu científico. Contribución a un psicoanálisis del conocimiento objetivo. Siglo XXI. México. 302pp.

Booker,J.B., D.E.Goldberg y J.H.Holland. 1990. Classifier Systems and Genetic Algorithms. En: J.G.Carbonell (ed.) Machine Learning. Paradigms and Methods. MIT Press. 1990.

Casas,G. y C.J.McCoy. 1979. Anfibios y Reptiles de México. Limusa. México. 87pp.

Chou,P.A. 1991. Optimal Partitioning for Classification and Regression Trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 13(4):340-354.

CIBA-GEIGY. 1979. Maize CIBA-GEIGY Agrochemicals. CIBA-GEIGY. México. 105pp.

Copi,I. 1979. Lógica Simbólica. CESCA. México. 407pp.

Crisci, V. 1983. Introducción a la Teoria y Práctica de la Taxonomía Numérica. OEA. Programa Regional de desarrollo científico y tecnológico. 132pp.

Dallwitz,M.J. 1980. A general system for coding taxonomic descriptions. Taxon 29:41-46.

Davis, P.H. y J.Cullen. 1979. The Identification of Flowering Plant Families. Cambridge University Press. UK. 113pp.

Duncan,T. y C.A.Meacham. 1986a. Meka and Mekaedit. A General Propose Multiple-Entry Key Algorithm and Editor. University of California. Berkeley. 17pp. y disco de 5 1/4".

Duncan,T. y C.A.Meacham. 1986b. Multiple-Entry Keys for the Identification of Angiosperm Families Using a Microcomputer. Taxon 35(3):492-494.

Dussauchoy,A. y J.N.Chatain. 1988. Sistemas Expertos. Métodos y Herramientas. Paraninfo. Madrid. 238pp.

Edmondson,W.T. (ed). 1963. Fresh Water Biology. University of Washington. 1248pp.

Ehrlich,P.R. y A.H.Ehrlich. 1961. How to Know the Butterflies. WM. C. Company. Pub. Iowa. 261pp.

Engel,C.W. y M.Cran. 1990. Pattern Classification. PC AI 4(3):20-23.

Erkmen,A.M. y H.E.Stephanou. 1990. Information Fractal for Evidential Pattern Classification. IEEE Transaction on System, Man, and Cybernetics. 20(5):1103-1114.

Fichefet,J.,J.P.Leclercq,P.Beyne y F.Rousselet-Piette. 1984. Microcomputer-assisted identification of bacteria and multicriteria decision models. Comput. y Ops. Res. 11:361-372.

Font-Quer, P. 1985. Diccionario de Botánica. Labor, Barcelona. 1244pp.

Ford,N. 1989. Prolog Programming. John Wiley & Sons. UK. 279pp.

Frenzel,L.E. 1986. Crash Course in Artificial Intelligence and Expert Systems. Howard W. Sams & Co. Indianapolis. 358pp.

Gama,S., S.Arias y M.Murguia. 1990. Clave por computadora para géneros de cactáceas de Norte y Centro América. Memorias del XI Congreso Nacional de Botánica, Oaxtepec, Mexico. Octubre, 1990.

Genesereth,M.R. y N.J.Nilson. 1988. Logical Foundations of Artificial Intelligence. Morgan Kaufman Pub. California. 403pp.

Geva,S. y J.Sitte. 1991. Adaptive Nearest Neighbor Pattern Classification. IEEE Transactions on Neural Netwroks. 2(2):318-322.

Gilbert,R. 1988. House Plants. Anyone Can Grow. Dorling Kindersley. Londres. 144pp.

González,L. 1984. Teoría de los Grafos en las Ciencias Sociales. UNAM. México. 210pp.

González,J. 1991. Los procesos transformados y los procesos alterados: fundamentos para una teoría procesual del conocimiento biológico. Urobros 1(2):45-90.

Goodal, D.W. 1968. Identification by Computer. Bioscience 18(6):485-488.

Graf,A.B. 1986. Tropica. Color Cyclopedia of Exotic Plants and Trees. Roehrs Company. New Jersey. 1152pp.

Guzmán,G. 1977. Identificación de los Hongos Comestibles, Venenosos y Alucinógenos. Limusa. México. 452pp.

Hansen,B. y K.Rahn. 1969. Determination of angiosperm families by means of a punched-card system. Dansk Botanisk Arkiv. 26:1-44. (mas 172 tarjetas).

Kowalski, R. 1986. Lógica, Programación e Inteligencia Artificial. Díaz de Santos. Madrid. 412pp.

Leenhouts,P. 1966. Keys in Biology. A survey and a proposal of a new kind. Proceedings Koninkligke Neederlands. Academie der. Wetenschappen. 69 Series C. 571-596.

Little,E.L.Jr. 1968. Clave con fichas perforadas de las familias de los árboles mexicanos. Turrialba 18:45-59.

López,A. 1975. Textos de Medicina Náhuatl. UNAM. Instituto de Investigaciones Históricas, México. 230pp.

Margulis,L. y K.V.Schwartz. 1981. Cinco Reinos. Guía ilustrada de los phyla de la vida en la Tierra. 335pp.

Morales,R. 1988. Un sistema de Clasificación y de Reconocimiento de Patrones. Tesis (Matemático), Fac. de Ciencias, UNAM. México. 83pp.

Moreno,N.P. y R.Allkin. 1988. Métodos Computarizados y algunas de sus aplicaciones al estudio de la flora de México. Boletín de la Sociedad Botánica de México. 48:69-74.

Morse,L.E. 1971. Specimen identification and key construction with time-sharing computers. Taxon (20(2/3):269-282.

Murguía,M. 1988. Clave para familias (Magnoliophyta) presentes en México. Tesis (Biólogo), Fac. de Ciencias, UNAM. México. 96pp.

Murguía,M. 1990a. **Aplicaciones potenciales de la Inteligencia Artificial a la Botánica.** Memorias del XI Congreso Nacional de Botánica, Octubre, 1990.

Murguía,M. 1990b. **Determinación botánica auxiliada por computadora.** Memorias del XI Congreso Nacional de Botánica, Oaxtepec, México. Octubre, 1990.

Murguía,M. 1990c. **SATAX-II: Determinacion y generación de descripciones botánicas auxiliadas por computadora.** Memorias del V Congreso Latinoamericano de Botánica, La Habana, Cuba. Junio, 1990.

Murguía,M. 1992. **Identificación Automatizada.** Notas del Seminario: "La Computadora en el Herbario". Organizado por la Asociación de Biologos Amigos de la Computación, A.C. Jardín Botánico, C.U. México, Enero, 1992.

Murguía,M. y O.Téllez. 1988. **SATAX: Sistema de ayuda al taxónomo, Reunión de trabajo: La computadora en la Flora de México.** (Consejo Nacional de la Flora de México) Jardín Botanico Exterior, CU, México. Julio 1988.

NeuralWorks Professional II. 1989. **User's Guide.** NeuralWare Inc.. PA. 405pp.

Oliver,D. y M.D.González. 1979. **Introducción a la Teoria de Gráficas.** ANUIES. México. 105pp.

Pankhurst,R. (ed) 1975. **Biological Identification with Computers.** Academic Press. N.Y. 333pp.

Pankhurst,R.J. 1978. **Biological identification.** Edward Arnold. Londres. 104pp.

Pankhurst,R.J. y R.R.Atchison. 1975. **An online identification program.** In "Biological Identification with Computers" Ed. R.J.Pankhurst. Academic Press. London.

Phillips,R. 1989. **Los Arboles.** Blume. Barcelona. 223pp.

Peláez, A. e I.Vargas. 1990. **CAPDELTA: Programa para facilitar la captura de datos del sistema DELTA.** Memorias del V Congreso Latinoamericano de Botánica, La Habana, Cuba. Junio, 1990. p238.

Radford,A.E., W.C.Dickison, J.R.Massey y C.R.Bell. 1974. **Vascular plants systematics.** Harper & Row. New York. 891pp.

Raudys,S.J. y A.K.Jain. 1991. Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. IEEE Transactions on Pattern Analysis and Machine Intelligence 13(3):252-264.

Raynal,J. 1974. Un exemple d'application du traitement electronique de l'information a la construction des clefs dichotomiques. Adansonia. 14:459-467.

Rich,E. 1983. Artificial Intelligence. McGraw-Hill. Singapur. 436pp.

Ross,H.H. 1978. Introducción a la Entomologia General y Aplicada. Omega. Barelona. 536pp.

Rzedowski,J. 1978. Clave para la Identificación de los Géneros de la Familia Compositae en México. Acta Científica Potosina 7(1,2):145.

Salazar,E.J. 1979. Modelos Esquematicos para la Elaboracion de Planes en la Educación Superior. ANUIES. México. 116pp.

Salazar,E.J. 1990. Lógica y Expertos. UAM-I. Mexico. 195pp.

Schneider,G.M., S.W.Weingart y D.M.Perlman. 1982. An Introduction to Programming and problem Solving With Pascal. John Wiley & Sons. Singapur. 468pp.

Shoemaker,P.A. 1991. A Note on Least Squares Learning Procedures and Classification by Neural Network Models. IEEE Transactions on Neural Netwroks. 2(1):158-160

Simpson,D.R. y L.James. 1972. A punched card key to the families of dicotyledons of the western hemisphere south of the United States. Field Museum of Natural History. Chicago, Illinois. (mas 60 tarjetas)

Smith,G.M. 1950. The Fresh Water Algae of the United States. McGraw Hill. Nwe York. 719pp.

Smith,H. y E.H.Taylor. 1966. Herpetology of México: Annotated Cheklists and Keys to the Amphibians and Reptiles. Ashton Maryland.

Sneath,P. 1980. BASIC program for determining the best identification scores possible from the most typical examples when compared with an identification matrix of percent positive characters. Computers & Geosciences 6:27-34.

Stirling,W.C. y D.R.Morrell. 1991. **Convex Bayes Decision Theory.** IEEE Transaction on System, Man, and Cybernetics. 21(1):173-183.

Szilasi,W. 1949. **¿ Qué es la Ciencia ?.** Fondo de Cultura Económica. México. 142pp.

Tschudi,F. 1988. **Matrix Representation of Expert Systems.** AI Expert 3(10):44-53.

Turing, A.M. 1950. **Máquinas, Computadoras e Inteligencia.** Traducción al español de "Computers Machinery and Intelligence". Mind, 59(236). En: Mentes y Máquinas. A.R.Anderson (Traductor) 1970. UNAM. México. 168pp.

Villaseñor,J.L. 1987. **Clave genérica para las compuestas del rio Balsas.** Boletín de la Sociedad Botánica de México. 47:65-86.

Villaseñor,J.L. 1989. **Manual para la Identificación de las Compuestas de la Península de Yucatán y Tabasco.** Technical Report No. 4. Rancho Santa Ana Botanic Garden. 122pp.

Villaseñor, J.L. y M.Murguía. 1992. **La Computadora en la Identificación Botánica.** Ciencia y Desarrollo 18(104):130-137.

Villaseñor, J.L. y M.Murguía. (En prensa). **Clave para familias de Magnoliophyta de México.** Consejo Nacional de la Flora de México- Asociación de Biólogos Amigos de la Computación, A.C. México.

Walter,S.M. 1975. **Traditional Methods of Biological Identification.** En: Ed. R.J.Pankhurst. "Biological Identification with Computers" Academic Press. London.

Wan,E.A. 1990. **Neural Network Classification: A Bayesian Interpretation.** IEEE Transactions on Neural Networks 1(4):303-305.

Watson,L. y M.J.Dallwitz. 1983. **The Genera of Leguminosae-Caesalpinioideae. Anatomy, Morphology, Classification, and Keys.** Canberra. The Australian National Universtity Research School of Biological Sciences. Canberra. 95pp.

Watson,L. y P.Milne. 1972. **A flexible system for automatic generation of special purpose dichotomous keys and its applications to Australian grass genera.** Aust.J.Bot. 20:331-352.

Wilson,J.B. y T.R.Partridge. 1986. Interactive plant identification. Taxon 35:1-12.

Winston,P.H. y B.K.Horn. 1984. LISP. Addison Wesley. London. 434pp.

Indice de Figuras.

Indice de Tablas.