An. Inst. Biol. Univ. Nal. Autón. México 42, Ser. Biol. Exp. (1): 1-22, figs., 5 tablas (1971)

SELECCIÓN DE UN GRUPO ÓPTIMO DE CARACTERÍSTICAS PARA LA IDENTIFICACIÓN TAXONÓMICA AUTOMATIZADA

Guillermina Yankelevich *
Irene Vasconcelos **

RESUMEN

Se propone un método cuantitativo, basado en la teoría de información, para optimizar el procedimiento de identificación taxonómica. Se sugiere que, la selección de características de acuerdo con su "contenido de información" sobre un grupo taxonómico dado, permite la conformación de un conjunto pequeño de ellas que conduce a un diagnóstico más eficiente a la vez que confiable.
Se discuten comparativamente los resultdaos de la identificación taxonómica usual de las familias de diversos ejemplares de fanerógamas y la identificación realizada a través del método automatizado propuesto.

ABSTRACT

A quantitative method, based on information theory, to optimize the taxonomic identification procedure, is proposed.
It is sugested that the selection of atributes according to its "information content", about a certain taxonomic group, provides a small set of them, necessary for an efficient and reliable diagnosis.
The method is illustrated with a problem of family identification from a group of individuals belonging to *Fanerogama*. A comparative discussion with the current procedures is included.

INTRODUCCIÓN

Es considerable el número de estudios que sobre taxonomía cuantitativa han aparecido en la literatura. Algunos de ellos fueron elaborados en el siglo pasado, como el de Adanson, contemporáneo de Lineo, quien propuso un conjunto de propiedades básicas que los individuos deberían reunir para ser agrupados en unidades taxonómicas. En la actualidad, todavía algunos procedimientos de clasificación taxonómica cuantitativa, se basan en los principios Adansonianos (taxonomía numérica) (Sokal y Sneath, 1963).

A pesar de que la taxonomía incluye en su estudio, además de los procedimientos de clasificación, los de nomenclatura, los de identificación y los correspondientes al estudio de las bases teóricas que rigen la clasificación, es sorprendente observar que la mayor parte de los estudios en taxonomía cuantitativa se han dirigido fundamentalmente al proceso de clasificación taxonómica

* Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México.
** Departamento de Radiobiología, Instituto Nacional de Energía Nuclear.

propiamente dicha. Así, la aplicación de técnicas de cromatografía, serología, biología molecular, como también las de taxonomía numérica y algunos otros procedimientos basados en principios estadísticos, han generado métodos cuantitativos para evaluar el grado de semejanza, o la magnitud de las diferencias entre individuos o unidades taxonómicas, con objeto de agruparlos en forma objetiva.

Probablemente, entre los procedimientos cuantitativos mencionados, la taxonomía numérica solamente se ha ocupado, aunque de manera somera, de analizar el problema de la nomenclatura (Sneath y Sokal, 1962) y también en forma importante del establecimiento de las bases teóricas del proceso de clasificación taxonómica (Sneath, 1957; Cain y Harrison, 1958).

Sorprende, sin embargo, encontrarse con la circunstancia de que casi ninguno de los métodos cuantitativos se ha ocupado de abordar el problema de la identificación de los organismos, con respecto a los grupos taxonómicos ya establecidos, o de la identificación de un grupo taxonómico dentro de otras jerarquías más elevadas. Afirman algunos autores, que la identificación es un proceso lógico derivado del proceso de clasificación, de tal forma que, una vez logrado el establecimiento de clasificaciones naturales, objetivas y reproducibles, la identificación se convierte en un procedimiento inmediato derivado de los mismos conocimientos (Sokal y Sneath, 1963).

Las consideraciones resumidas en el párrafo anterior explican por qué, en la actualidad, una gran parte de los taxónomos continúan trabajando con procedimientos de identificación que no han sido modificados en varias décadas; esto significa, que no ha sido suficientemente aprovechada la tecnología moderna en este tipo de actividad que, en otros campos, ha elevado considerablemente el nivel de eficiencia con la cual se realizan trabajos rutinarios similares.

Las técnicas mencionadas, fundamentalmente las de computación, ya han sido utilizadas en los aspectos de almacenamiento y recuperación de datos, lo cual, representa una gran ayuda para el taxónomo; sin embargo, es necesario hacer hincapié, en que su uso, se ha hecho empleando las claves para identificación ya establecidas que, en algunos casos, son en sí mismas ineficientes (Metcalf, 1954) y no han sido optimizadas con respecto al número de características que emplean para el propósito.

Siendo el problema de identificación taxonómica, en un buen número de los casos una etapa de rutina de la cual el investigador en taxonomía no puede sustraerse, sería deseable que a ella se dedique el menor tiempo posible mediante el empleo de las técnicas de computación ya mencionadas, pero también utilizando procedimientos de identificación más eficientes en cuanto a la selección del mínimo número de características necesarias para tal fin y, sobre todo, aquellas que son las más apropiadas para ese propósito.

El simple hecho de postular la selección de un conjunto de características como el más adecuado, implica el conceder distinta importancia a cada una de ellas. Este procedimiento ha sido muy ampliamente discutido por diversos autores (Cain y Harrison, 1958 y Michener y Sokal, 1957) y lo han juzgado inadecuado en el proceso de clasificación taxonómica: el seleccionar las características para formar grupos taxonómicos implica un prejuicio por parte del investigador, que puede influir en el tipo de agrupamiento generado. Sin embargo, en el diagnóstico taxonómico, como los mismos Sokal y Sneath (1963) señalan, es deseable conceder "pesos" diferentes a las características tomando en consideración aquella o aquellas que son ex-

cluyentes con respecto a la identificación de un grupo taxonómico.

La utilización de métodos cuantitativos en la identificación taxonómica ya aparece en la literatura, aun cuando los reportes son escasos. Así, Rescigno y Maccacaro (1960) y Moller (1961), describen ensayos de procedimientos cuantitativos para la identificación taxonómica, basados en la teoría de información, el primero, y en aspectos estadístico-probabilísticos, el segundo.

El presente trabajo, propone el empleo de un método cuantitativo, basado en la teoría de información, para optimizar el procedimiento de identificación taxonómica. En él se sugiere que la selección de características, de acuerdo "con su contenido de información" sobre un grupo taxonómico dado, permite la conformación de un conjunto pequeño de ellas que conduce a un diagnóstico más eficiente a la vez que confiable. Este método ha sido sugerido con anterioridad por uno de los autores en una nota preliminar en la que se publican resultados utilizando estadísticos subjetivos (Yankelevich y Negrete, 1969).

En virtud de que el trabajo que a continuación se presenta emplea, como ya ha sido indicado, conceptos de la teoría de información, se ha considerado pertinente incluir un pequeño apéndice al cual el lector puede recurrir para consultar algunos de estos aspectos.

MATERIAL Y MÉTODOS

Las plantas utilizadas en el presente trabajo, fueron fanerógamas colectadas en la zona del Pedregal de San Ángel, dentro de los límites de la Ciudad Universitaria, y específicamente, del área comprendida en el costado izquierdo del Jardín Botánico Exterior del Instituto de Biología; éstas pertenecen exclusivamente a la parte baja de la asociación *Seneacionetum praecosis*, que es una de las más extendidas en el Pedregal (Rzedowky, 1954).

Para colectar la vegetación del lugar, se recurrió al "método del cuadro" (Oosting, 1951; Weaver, 1950). Se escogió un lugar al azar y se procedió al muestreo correspondiente; para ello se delimitó un "cuadro", colocando también al azar, con una orientación NNE y SSE. Con este método se obtuvo el área mínima, cuya curva se presenta en la Fig. 1. Con base en estos datos se continuó el muestreo y se concluyó con la delimitación de un segundo "cuadro"; en él, se colectaron y registraron todas las fanerógamas presentes. El material fue manipulado con las técnicas botánicas convencionales y pasó a formar parte del Herbario Nacional del Instituto de Biología en la Ciudad Universitaria. La colecta se realizó durante la primavera y verano (del 23 de abril al 4 de septiembre de 1969).

Una vez conocida la naturaleza del material, se procedió a la determinación de dos estadísticos: a. La abundancia relativa de las familias localizadas en el área, y b. La abundancia de los géneros en cada una de las familias.

Con base en los estadísticos mencionados en el párrafo anterior, se elaboraron dos matrices que aparecen en la Tabla 1 (véase al final del artículo). Una de ellas, la matriz columna, señala las probabilidades de encontrar cada una de las familias dentro del área y la otra, matriz rectangular, muestra la probabilidad de observar las características (anotadas en el encabezado de la matriz rectangular), dadas las familias correspondientes en la matriz columna. En la Tabla 2 (véase al final del artículo) se

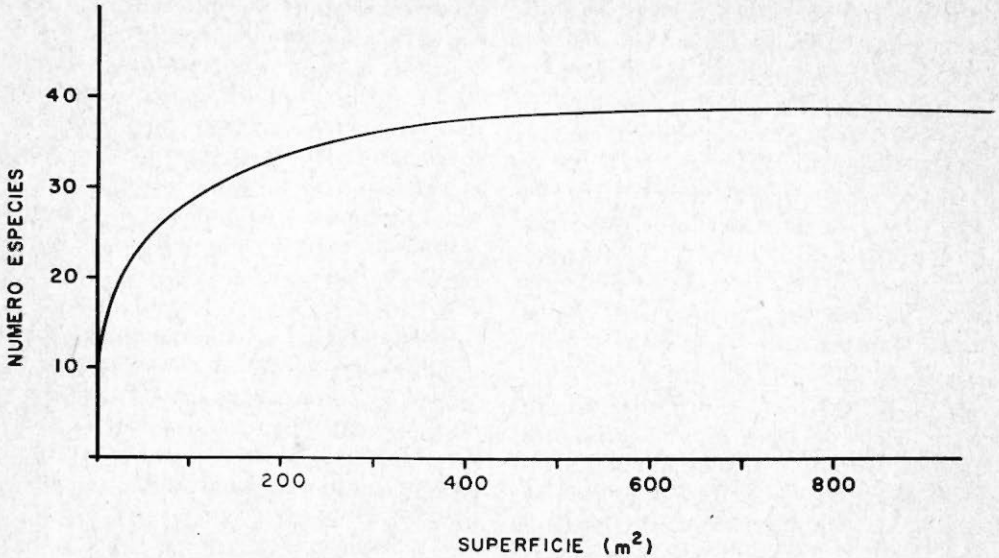


Fig. 1. Gráfica de la determinación del área mínima para las fanerógamas del Pedregal de San Ángel (parte baja de la asociación *Senecionetum praecosis*).

enlistan las características utilizadas, numeradas de acuerdo con la misma clave que se les adjudica en la Tabla 1.

La determinación de las probabilidades condicionadas en cuestión (probabilidad de observar una característica dada la familia), se realizó tomando como base el número de géneros en cada una de las familias que presentaban una característica dada, en relación al total de géneros de la familia.

El valor de la probabilidad condicionada se obtuvo a través de la ecuación de Bayes que a continuación se transcribe:

$$P(Yi/X) = \frac{P(Yi) \prod P(X/Yi)}{\Sigma (P(Yi) \prod (P(X/Yi)}$$

en donde P (Yi/X) es la probabilidad condicionada de una familia (Yi) dado un conjunto de características (X); P (Yi) representa la probabilidad a priori de una familia (Yi); Π, el producto sobre todos los valores de X (características); y Σ, la suma sobre todos los valores i (familias).

Utilizando los valores de probabilidad a priori y condicionadas obtenidas con los cálculos anteriores, se determinó la incertidumbre a priori de las familias empleando la siguiente ecuación:

$$H(Yi) = -\sum_{i=1}^{n} \pi \log_2 \pi$$

donde (pi) es la probabilidad a priori de una familia.

Con los valores de probabilidad condicionada obtenidos con la ecuación de Bayes, se procedió a calcular la incertidumbre condicionada de las familias, dadas las características que las describen; este cálculo se llevó a cabo con las siguiente ecuación:

$$H(Yi/X) = -\sum_{x=1}^{n} \sum_{y=1}^{m} P(Y/X) \log_2 P(Y/X) P(X)$$

La diferencia entre los valores de incertidumbre a priori e incertidumbre

condicionada, *se consideró como un índice de la información que aporta la característica en cuestión*. Con los valores de información obtenidos de la manera descrita, para cada una de las características, se construyó el histograma de distribución de cantidad de información de las características que aparece en la Fig. 2.

Para el proceso automatizado de identificación taxonómica, se elaboró un programa para la máquina computadora B 5500, empleando como lenguaje fuente el Algol.

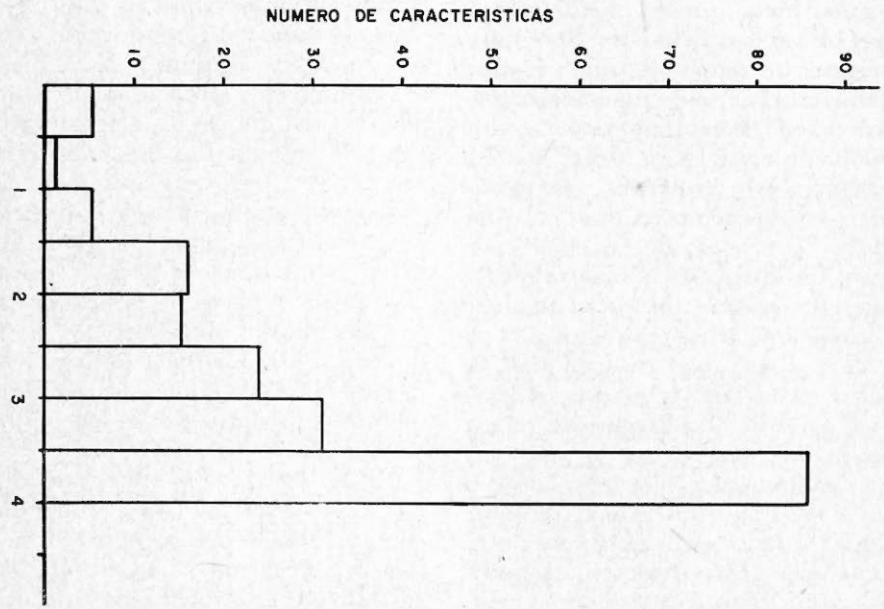


Fig. 2. Histograma de la distribución de la "cantidad de información" contenida en las características de los individuos pertenecientes a las familias de fanerógamas. Abcisas: información contenida en una característica (bits).

RESULTADOS

En la Tabla 3 (véase al final del artículo) se enlistan las familias de fanerógamas encontradas en la zona estudiada del Pedregal, así como también los géneros correspondientes a cada una de ellas. Puede observarse, que este cuadro no incluye a todas las familias ni a todos los géneros que otros autores han descrito para dicha zona (Sánchez, 1969); sin embargo, como en este trabajo se precisa calcular los estadísticos ya nombrados en la parte correspondiente a métodos, se empleó exclusivamente la vegetación que fue colectada. Puede observarse, además, que en la matriz columna de la Tabla 1, no aparecen las familias Portulacaceae, Anacardiaceae y Loganiaceae, en vista de que su probabilidad de aparición en la colecta fue extraordinariamente baja, y por lo tanto, se valoró como cero.

Las características enlistadas en la Tabla 2, comprenden tanto las observadas en los ejemplares colectados, como las

descritas por algunos autores (Lawrence, 1958; Sánchez, 1969) como pertenecientes a las familias con las que se trabajó y que por la época del año en que se llevó a cabo la colecta, no pudieron observarse directamente en ellos (como lo son las flores y frutos de algunos ejemplares).

Es interesante notar que la matriz rectangular está formada fundamentalmente por características de tipo binario (valores de cero y de uno), incluso hay características que se pudieran considerar como "claves" que presentan una probabilidad de 1, para cierta familia y cero para todas las demás; ejemplo de este tipo son las características números 6, 10, 13, 27 ... etcétera, las cuales permitirían un diagnóstico determinístico en algunos casos. Se presentan también otros caracteres como por ejemplo: 1, 9, 2, 22 y otros, cuya probabilidad de aparición en la familia, es de uno para casi todas ellas (características redundantes).

En la columna del lado derecho de la Tabla 2, se anotan los valores de información calculados para cada una de las características. Estos valores de información son los que se tomaron para construir el histograma de la Fig. 1. Puede notarse la distribución tan amplia de valores de información con que contribuyen las características utilizadas. Estos valores oscilan desde casi cero bits hasta 4 bits de información. Se observa también que el intervalo de clase con la columna más alta en el histograma, coincide con aquellas características que aportan al máximo contenido de información; esas características corresponden a las denominadas "características clave".

Las columnas del histograma, con escaso contenido de información (las que van entre 0 y 1.5 bits) corresponden a aquellas que aparecen con la misma probabilidad en casi todas las familias (características redundantes).

Para la identificación automática, se emplearon únicamente las características cuyo contenido de información está entre 2 y 3 bits (quinta y sexta columna del histograma). Esto se debió a que se deseaba excluir las características "clave" (8ª columna) y las características entre 3 y 3.5 bits de información (7ª columna) mostraron ser insuficientes para la identificación de todos los géneros. Así mismo, la 6ª y 7ª columnas excedían el número de características suficientes. Puede observarse que la "optimización" fue realizada únicamente por ensayos sucesivos. Con este reducido número de características, se logró un 97% de aciertos en los diagnósticos.

En la Tabla 4, se muestran algunos ejemplos obtenidos con el programa de computadora; éstos se obtuvieron utilizando la ecuación de Bayes, que permitió calcular la probabilidad de que un individuo pertenezca a cada una de las familias consideradas. Nótese que el diagnóstico no es de tipo determinístico, es decir *se identifica el individuo como perteneciente al grupo taxonómico que presenta más alta probabilidad.*

En la Tabla 5, se muestra una lista, en donde aparecen todos los géneros trabajados y las 3 familias a las cuales cada uno mostró más alta probabilidad de pertenecer.

DISCUSIÓN

La decisión de estudiar las plantas fanerógamas del Pedregal para llevar a cabo el presente trabajo, se hizo en función del fácil acceso para el muestreo del material.

Conviene señalar que el método propuesto, posee generalidad suficiente para ser aplicado a cualquier otro conjunto de vegetales o animales de cualquier zona. De hecho, este procedimiento, fue

propuesto inicialmente por Yankelevich y Negrete (1969), quienes valoraron su aplicabilidad con datos obtenidos por otros autores en el orden Orthoptera.

El empleo del teorema de Bayes para diagnóstico diferencial fue postulado inicialmente por Takahashi (1965), en diagnóstico médico para características no exclusivas (síntomas); en este estudio también se hizo la misma consideración: las características que describen una familia pueden, en parte, presentarse en individuos de otras familias. Posteriormente, este mismo método ha sido aplicado por otros autores, también en diagnóstico clínico automatizado (Negrete *et al*, 1966), los cuales obtuvieron un nivel de confiabilidad en sus resultados del 75%.

El uso de diagnóstico probabilístico como el que aquí se propone, pudiera ser utilizado por el taxónomo como una estimador que proporciona información preliminar sobre el grupo taxonómico al que un individuo pertenece. El hecho de poder obtener una escala de probabilidad para todos los grupos manejados, permite al investigador hacer una evaluación preliminar, aun en el caso en el que las probabilidades obtenidas fueran muy semejantes para varios grupos. Este último resultado, además puede orientar sobre la insuficiencia en el número de características utilizadas o sobre la inadecuada selección de ellas (Yankelevich y Negrete, 1969).

El método propuesto para identificación emplea también la probabilidad *a priori* de encontrar un individuo de una cierta familia; el conocimiento de este estadístico es de gran importancia para el investigador, cualquiera que sea su procedimiento de identificación y muy probablemente, lo utiliza en forma subjetiva en el proceso habitual.

Como puede observarse en este trabajo, las características empleadas por el taxónomo para propósitos de identificación no representan el mínimo de las que pudieran ser utilizadas. En nuestro caso, el uso de 52 características, únicamente, de las 193 descritas, es suficiente para lograr la identificación correcta de 63 géneros de los 65 utilizados aun cuando se excluyeron las características "clave". De manera similar Yankelevich y Negrete (1969) lograron diagnósticos confiables en ortópteros con sólo 33 características de las 93 que describen los subórdenes trabajados. La adición de un par de características "clave" correspondientes a las familias incorrectamente diagnosticadas originó un grupo suficiente, para la identificación correcta de todos los géneros.

Las características de tipo "clave" fueron excluidas del estudio ya que, su presencia permite realizar simplemente un diagnóstico de tipo lógico y se pretende mostrar que el método es particularmente útil cuando no es posible llevar al cabo una identificación de este tipo. Sin embargo, la automatización del sistema de identificación propuesto permite incluir características valoradas en cualquier tipo de escala. Más aún, circunstancias en las cuales este tipo de características son difíciles de observar (ya sea porque la época del año no ayude para tal fin, o por el arduo trabajo que en sí su observación implique) o la confiabilidad de su observación sea muy baja, un método probabilístico de identificación se ve ampliamente justificado.

El contenido de información que aportan las características es un parámetro que pudiera ser utilizado en su selección para propósitos de identificación, a través del establecimiento de un umbral de información por parte del especialista, o sea que el umbral representaría la mínima cantidad de información que una característica cualquiera debiera contener para ser tomada en consideración; así, por ejemplo, en nuestro caso, un umbral adecuado sería 2.5 bits, y todas aquellas características

cuyo contenido de información sea menor a este valor, son automáticamente desechadas.

El sistema descrito no está aún optimizado. Se encuentra en estudio el proceso de determinación del "grupo mínimo" de características para un diagnóstico óptimo.

Por último, es necesario señalar que el valor de información calculado por este procedimiento, es exclusivamente un estimador de información que para el propósito ha mostrado ser suficiente. Su cálculo real requeriría de las probabilidades condicionadas correspondientes a todas las posibles combinaciones de características que generan grupos exclusivos, lo cual representa un trabajo casi imposible de realizar, aun contando con máquinas de cálculo rápido.

CONCLUSIONES

1. El conocimiento de la información que aportan las características de los individuos, al grupo taxonómico al que pertenecen, permite la selección de un grupo de ellas, pequeño, pero suficiente para una identificación taxonómica confiable.

2. El conocimiento de dicho parámetro, permite además seleccionar las características más apropiadas para los propósitos de identificación ya que las seleccionadas son aquellas que proporcionan el máximo de información, acerca del grupo taxonómico.

3. La ecuación de Bayes permite un diagnóstico taxonómico que puede representar una ayuda al taxónomo en sus trabajos de identificación. Este procedimiento tiene además la ventaja de evitar la apreciación subjetiva del que identifica.

4. El programa para el proceso de identificación taxonómica automatizada que en este trabajo se propone, permite un aumento en la eficiencia del investigador en taxonomía, con base en la economía del tiempo que dedica a dicha labor, además de la posibilidad del manejo simultáneo de una gran cantidad de datos, que de otra manera no sería posible.

5. Se insiste en la necesidad de evaluar el grado de dificultad en la obtención de las características; este parámetro, junto con el que aquí se maneja (información aportada por las características) permitiría la selección de un conjunto suficiente para un diagnóstico confiable.

APÉNDICE

ALGUNOS CONCEPTOS DE LA TEORÍA DE INFORMACIÓN

Existen varias definiciones sobre el concepto de "cantidad de información". Éstas han variado de acuerdo con los diversos campos de trabajo en los que han sido manejados.

En virtud de que en el presente trabajo únicamente se aborda el concepto de "cantidad de información" derivado a partir de la medida de cantidad de variedad en un conjunto, la revisión se llevará al cabo exclusivamente bajo este aspecto.

La cantidad de información contenida en un conjunto (Edwards, 1964), puede ser medida como una función de la dificultad con la que se puede identificar un elemento de dicho conjunto, y por consiguiente, también es función de tamaño del mismo.

En el caso de que el conjunto estuviera

formado por un solo elemento, no existiría dificultad para identificarlo y la información sería cero. Considérese ahora el caso de un conjunto formado por dos elementos; se requeriría de una pregunta para determinar de cuál de los dos elementos se trata. Si los elementos fueran cuatro en el conjunto, el número mínimo de preguntas necesarias para identificar a alguno de ellos serían dos.

Puede observarse que el número de preguntas, que pudieran considerarse como función directa del "grado de dificultad en la identificación de un elemento", aumenta en uno a medida que el conjunto duplica sus elementos; es decir, existe una relación logarítmica entre ambos, que puede expresarse de la manera siguiente:

$$n = 2^I$$
$$I = \log_2 n$$

Siendo (n) el número de elementos de conjunto e (I) el número de preguntas necesarias para identificar un elemento o "contenido de información del conjunto". Estas consideraciones son válidas únicamente si se emplea una estrategia óptima en la selección. (División del conjunto en dos subconjuntos iguales y preguntar a cuál de los dos subconjuntos pertenece el elemento en cuestión. Este procedimiento se repite hasta quedar el elemento identificado.)

La información contenida en un conjunto, puede expresarse como contenido total, o como información promedio por elemento. A esta última, se le ha denominado incertidumbre del conjunto y se denota con la letra H.

Cuando los elementos del conjunto, no poseen la misma abundancia dentro de él, el cálculo de incertidumbre debe tomar en cuenta el valor de probabilidad ya que el número de preguntas para identificar un elemento no es función exclusiva del número de elementos diferentes, sino también de su abundancia relativa en el conjunto. En este caso la incertidumbre es un promedio pesado por el factor probabilidad, como se expresa en la siguiente ecuación:

$$H = - \sum_{i=1}^{n} pi \log_2 pi$$

En caso de que los n elementos fueran equiparables; p = 1/n, y por lo tanto:

$$H = - \sum_{i=1}^{n} 1/n \log_2 1/n$$
$$= - n \, (1/n \log_2 1/n)$$
$$= - \log_2 1/n$$
$$= \log_2 n$$

lo cual muestra que para eventos equiprobables, los valores de I y H en un conjunto son iguales.

Además de las dos variables ya mencionadas, la incertidumbre es también función de las dependencias entre los elementos; esto es, si éstos no son independientes entre sí, además de su probabilidad *a priori* de presentarse en el conjunto, hay que considerar en los cálculos de H, la probabilidad condicionada de encontrar un elemento con respecto a la presencia de los otros. En este caso, la ecuación para el cálculo de incertidumbre se transforma de la manera siguiente:

$$H\ (y/x) = - \sum_{x=1}^{n} \sum_{y=1}^{m} p\ (y/x)\ \log_2 p\ (y/x)\ p\ (x)$$

Siendo p (x) la probabilidad *a priori* de que se encuentren los elementos en el conjunto y p (x/y), que se calcula como p (y, x) / p (y) es la probabilidad condicionada de que se presente (y), dado que se encontró (x).

Existen varias unidades para medir cantidad de información (Goldman, 1955); ellas también son dependientes del campo de trabajo en donde el concepto sea usado.

En el presente artículo se emplea como unidad el bit o binit (palabra nemotécnica derivada de inglés "binary unit") que es la más comúnmente usada y se deriva de la definición de cantidad de información que involucra el concepto de variedad. El bit o binit, como medida de cantidad de información tiene la ventaja de que, cuando en un conjunto se tiene la mínima variedad (dos elementos diferentes) la información es igual a 1 bit ($\log_2 2 = 1$ bit), esto es, la mínima cantidad posible. Por otra parte, tratándose de un solo elemento en el conjunto o sea, siendo cero la variedad, la información contenida también es cero ($\log_2 1 = 0$ bit).

AGRADECIMIENTOS

A los biólogos Nelly Diego, Francisco González Medrano y Hermilo Quero que en forma desinteresada y amable asesoraron a los autores de este artículo en los aspectos de taxonomía botánica.

LITERATURA

CAIN, A. J. y G. A. HARRISON, 1958. An analysis of the taxonomist's judgment of affinity. *Proc. Zool. Soc. London, 131*: 85-98.

EDWARDS, E., 1964. *Information transmission* Chapman S. Hall Londres, 133 pp.

GOLDMAN, S., 1955. *Information theory*. Prentice Hall Nueva York, 385 pp.

LAWRENCE, H. M., 1958. *Taxonomy of vascular plants*. McMillan Nueva York, 730 pp.

METCALF, Z. P., 1954. The construction of Keys. *Systematic Zool., 3*: 38-45.

MICHENER, C. D. y R. R. SOKAL, 1957. A cuantitative approach to a problem in clasification. *Evolution, 11*: 130-162.

MÖLLER, F., 1962. Quantitative methods in the systematics of Actinomycetalis. IV. The theory and application of a probabilistic identification key. *Giorn. Microbiol., 10*: 29-47.

NEGRETE, M. J., OLIVARES, L. y C. P. SOLÍS, 1966. El uso de estadísticos subjetivos en la simulación del diagnóstico médico por medio de computadoras. *Bol. Inst. Estud. Méd. Biol., 24*: 107-108.

OOSTING, H. J., 1951. *Ecología vegetal*. Aguilar, Madrid, 436 pp.

RESCINGNO, A. y G. A. MACCACARO, 1960. The information content of biological classifications. In Cherry C (Ed.).

REZDOWSKI, J., 1954. Vegetación del pedregal de San Ángel. Tesis profesional. IPN. México.

SÁNCHEZ, S. O., 1969. *La flora del Valle de México*. Herrero, México.

SNEATH, P. H., 1957. The application of computers to taxonomy. *J. gen. microbiol., 17*: 201-226.

SNEATH, P. H. y R. R. SOKAL, 1962. Numerical taxonomy. *Nature, 193*: 855-886.

SOKAL, R. R. y P. H. SNEATH, 1963. *Principles of numerical taxonomy*. Freeman, Londres 359 pp.

TAKAHASHI, K., 1965. Methodological aspects of computer diagnosis. Gordon Research Conferences on Biomathematics. Andover, Nueva Hampshire.

WEAVER, J. E., y F. E. CLEMENTS, 1950. *Ecología vegetal*. Acme Agency, Buenos Aires, 667 pp.

YANKELEVICH, G. y M. J. NEGRETE, 1969. El uso del contenido de información de las características en la identificación taxonómica automatizada. *Biol. Estud. Méd. Biol.* México, *26*: 73-79.

Tabla 1

Matriz Columna

Matriz Rectangular →

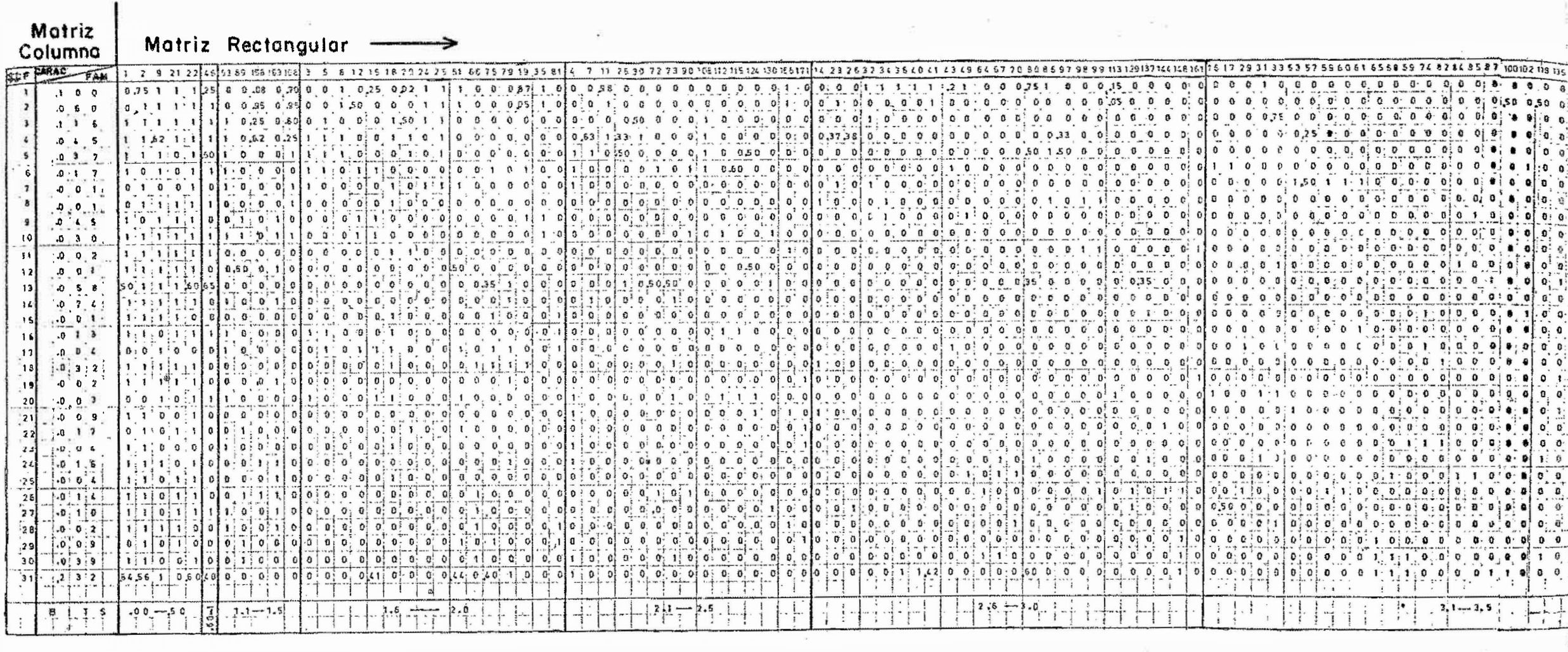
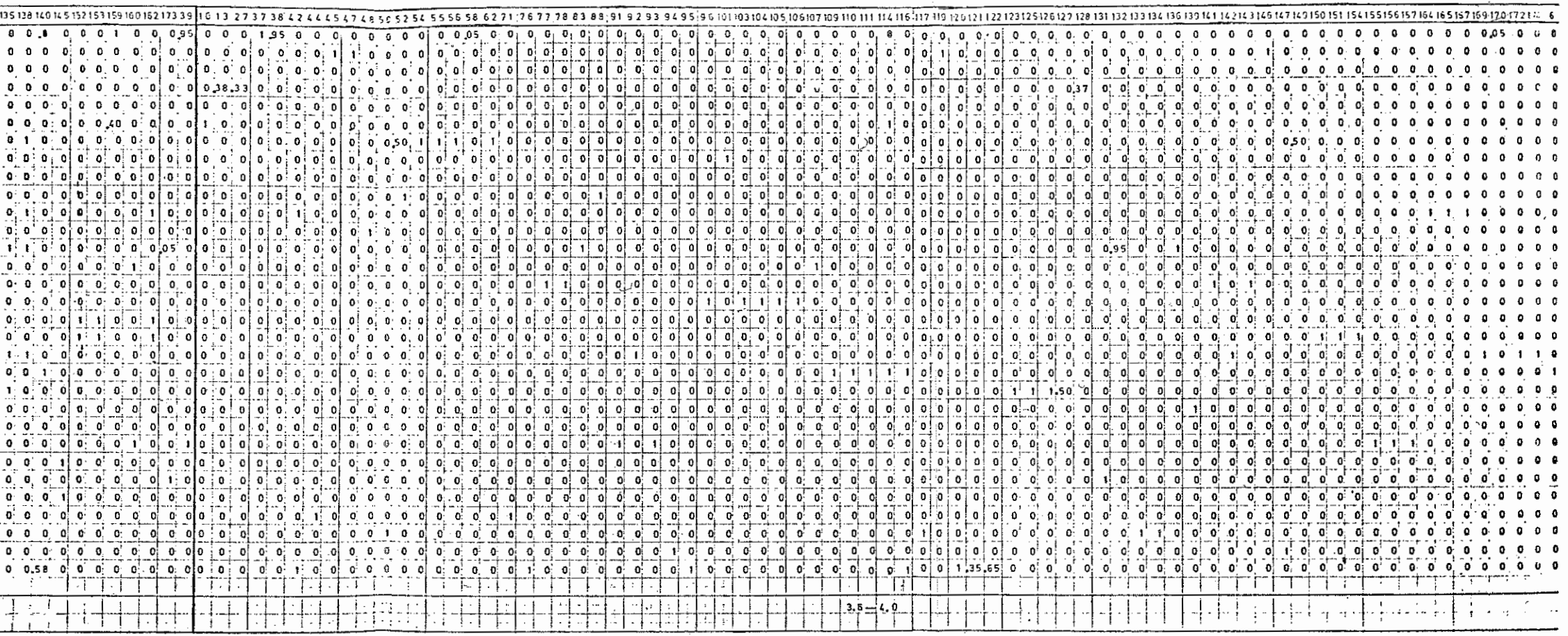


Tabla 1. Matriz de probabilidad *a priori* (matriz columna) de las familias de fanerógamas del Pedre
de las características señaladas en las columnas (clave Tabla 2) dada la familia correspondiente a ca
intervalos de clase de acuerdo con su "contenido de información"



edregal de San Ángel. Matriz de probabilidad condicionada (matriz rectangular
cada uno de los renglones. Las características se han ordenado y agrupado en
" que se señala en el último renglón de la matriz.

TABLA 2

Clave de las características	CARACTERÍSTICAS	*"Contenido de Información"*
22	Hierbas	0.0110
46	hierbas anuales	0.9532
80	hierbas perennes	2.7189
72	hierbas volubles	2.3685
75	plantas leñosas	1.8885
78	árboles balsamíferos, corteza con canales balsamíferos	3.8779
104	plantas laticíferas	3.8779
124	plantas suculentas	2.4857
125	plantas sin hojas	3.8779
126	grupos de aguijones presentes	3.8779
152	arbustos trepadores con los troncos flexibles	3.4037
1	flores actinomorfas	0.2768
51	flores cigomorfas	1.8353
106	flores anodinas	3.8779
108	flores trímeras	2.2009
139	flores hexámeras	3.8779
162	flores tetrámeras	3.1073
163	flores pentámeras	1.3915
29	flores polígamo-dioicas	3.1145
2	flores hermafroditas	0.0608
15	flores unisexuales	1.8986
43	flores dioicas	2.9771
140	flores monoicas	3.4339
115	flores femeninas con perigonio	2.4903
109	flores masculinas con estambres numerosos y dispuestos en cabezuela	3.8779
159	flores masculinas con perianto infundibuliforme	3.5667
3	perigonio corolino	1.8761
10	perigonio tubular (corto)	3.8779
27	perigonio de pétalos separados hasta la base	3.8779
28	perigonio infundibuliforme	2.5894
53	sépalos y pétalos del mismo color	3.1115
83	quilla presente	3.8779
96	hipsófilas presentes	3.8779
97	perigonio tubular infundibuliforme	2.8902
116	vilano presente	3.8779
141	sepálos de prefloración valvada	3.8779
147	cáliz de 4-5 divisiones, adherentes al ovario	3.8779
171	cáliz de sépalos soldados entre sí	2.2562
54	pétalos de igual longitud que los sépalos	3.8779
64	corola gamopétala tubular	3.0457
55	labelo presente	3.8779

Clave de las características	CARACTERÍSTICAS	*"Contenido de Información"*
68	corola gamopétala en forma de campana	3.1895
93	canales resiníferos presentes	3.8779
117	corola bi o unilabiada tubular	3.8779
120	corolas periféricas alargadas en lígulas	3.8779
123	pétalos y sépalos unidos formando un eje largo	3.8779
130	corola acampanada e infundibuliforme	2.5723
142	corola de prefloración contorneada	3.8779
150	corola de prefloración valvada	3.8779
159	glumas presentes	3.5667
172	corola de pétalos unidos por su base a una columna estiminal	3.8779
160	2-5 estilos	3.2048
4	ovario ínfero	2.1092
5	ovario trilocular, tricarpelar	1.8912
6	ovario con muchos o pocos óvulos en cada celda	3.8779
11	3 estambres presentes	2.3711
12	estilo tripartido	1.8196
20	6 estambres presentes	1.8536
21	ovario súpero	0.3598
30	estaminodios presentes	2.5529
35	ovario tricarpelar, unilocular, con un solo óvulo	2.0179
36	2 estigmas presentes	2.9984
44	2-3 estigmas	3.8779
56	ginostegio corto sin pie	3.8779
59	ginostegio presente	3.5272
60	polinia presente	3.5272
61	ovario tricarpelar, unilocular, con varios óvulos presentes	3.5084
62	3 placentas parientales	3.8779
65	4-5 estambres	3.1055
66	ovario bicarpelar, bilocular	1.9701
42	Ovario bicarpelar, unilocular	3.8107
70	estambres unidos a la base de la corola	2.9606
76	ovario con óvulos péndulos en cada división	3.8779
81	disco presente	2.0149
82	ovario excéntrico	3.1768
84	estambres opuestos a las divisiones del perigonio	3.1575
85	ovario con funículo manifesto	3.4723
88	placentación central	3.8779
94	ovario semi-ínfero	3.8779
99	anteras biloculares	2.9606
101	estigma peltado	3.8779
110	dorso del carpelo estriado	3.8779
111	placenta bipartida	3.8779
112	3 estilos cortos	2.5170

Clave de las características	CARACTERÍSTICAS	*"Contenido de Información"*
118	estilo bipartido	3.1218
132	4 estambres didínamos	3.8779
135	10-code estambres presentes	3.1411
143	Estambres en doble número que los sépalos	3.8779
92	los estambres más o menos unidos en la base	3.8779
145	5 estambres presentes, y opuestos a las divisiones de la corola	3.1145
50	estambres en balancín	3.8779
146	óvulo basal	3.8779
148	estambres insertos en el tubo de la corola	2.8676
151	estambres insertos al lado del disco carnoso	3.8779
155	estilopodio presente	3.8779
156	carpóforo presente	3.8779
164	estambres tetradínamos	3.8779
165	ovario con falso tabique membranoso	3.8779
166	placentación parietal	2.5278
173	los filamentos de los estambres se unen formándose, así un tubo, pero su parte apical queda libre (anteras)	3.3309
174	estilo dividido en tantas ramas como carpelos hay, o en el doble	3.8779
7	plantas con bulbo	2.4244
16	plantas con tubérculo subterráneo	3.2698
170	rizoma rastrero estolonífero	3.8779
149	raíz fasiculada y carnosa	3.8779
23	plantas con rizona	3.0519
38	raíz envuelta en una vaina	3.8779
52	plantas con fibras radicales	3.7371
17	tallos volubles	3.1055
31	tallos nudosos	3.1466
39	tallos huecos	3.3399
47	tallo tricuetro	3.8779
113	tallos carnosos	2.7327
127	plantas acaulecentes	3.8779
133	tallo prismático (cuadrangular)	3.8779
8	hojas en roseta basilar	1.9908
18	hojas alternas	1.7557
24	hojas sentadas	1.9529
25	hojas paralelinervias	1.6776
32	hojas envainadoras	2.8489
40	língula presente	2.9984
48	hojas cilíndricas	3.8779
73	estípula presente	2.3170
79	hojas compuestas	1.9195
89	hojas opuestas	1.4794
90	hojas acorazonadas	2.3752
114	hojas reticulinervias	3.8779

Clave de las características	CARACTERÍSTICAS	*"Contenido de Información"*
137	hojas pinnadas	2.7356
153	zarcillos presentes	3.4037
154	zarcillos e inflorescencias se oponen a las hojas	3.8779
157	pecíolo y vaina	3.8779
168	hojas glabras	1.4403
9	inflorescenia presente	0.5246
19	flores en panícula	2.0218
26	flores solitarias	2.9279
33	inflorescencia cimosa	2.8489
34	flores desnudas	3.4469
41	las glumas con las flores determinan la espiguilla	2.9984
49	flores protegidas por brácteas	3.8779
58	iflorescencia es espádice	3.5745
69	prefloración valvada	3.1895
74	flor panículo-cimosa	3.1575
86	inflorescencia en espiga	2.9833
87	flor en cabezuela o corimbo	3.1604
102	flores verdosas	3.9700
121	cabezuela homógama	3.8779
129	prefloración plegada y contorneada	2.8992
131	flores agrupadas en cincinos	3.8779
138	inflorescencia racimosa	3.3673
144	flores en las axilas de las hojas	2.8857
158	flor dispuesta en umbela	1.5688
119	debajo de la inflorescencia hay hojas involucrales	3.8779
103	flores agrupadas en ciatos	3.8779
13	fruto tricuatro con aristas aladas	3.8779
14	fruto una baya	2.9784
37	fruto cariopsis, con embrión de dilatación lateral	3.8779
45	fruto nuez tricuetra	3.8779
57	cápsula alargada	3.5745
63	fruto capsular	1.1087
77	fruto drupáceo	3.8770
100	fruto aquenio	3.3692
128	cápsula subglobosa	3.8779
136	fruto una legumbre	3.8779
167	fruto una silícua	3.8779
91	hojas papiráceas	3.8779
105	semillas con arilo	3.8779
134	epidermis con pubescencia glandulosa	3.8779
71	epidermis con pubescencia urticante	3.8779
161	pubescencia simple	2.8793
169	tegumento formado por pubescencia estrellada	3.8779

Tabla 2. Características generales utilizadas por el taxónomo en la identificación de los individuos pertenecientes a las familias de fanerógamas; a la izquierda se señala la clave arbitraria que se les ha asignado y a la derecha, su "contenido de información".

TABLA 3

EMDRYOPHYTA SIPHONOGAMA

ANGIOSPERMAE

Órdenes	*Familias*	*Géneros*
	MONOCOTYLEDONEAE	
Glumiflorae	Gramineae	*Stipa sp., Leptochloa sp., Setaria sp., Tripsacum sp., Panicum sp., Paspalum sp.,*
	Cyperaceae	*Cyperus sp.*
Farinosae	Commelinaceae	*Commelina sp., Tinantia sp.*
	Liliaceae	*Calochortus sp., Allium sp. Echcandia sp.*
Liliiflorae	Amarylidaceae	*Agave sp., Menfreda sp.*
	Dioscoreaceae	*Dioscorea sp.*
Microspermae	Orchidaceae	*Habenaria sp., Spiranthes sp.*
	DICOTYLEDONEAE	
Piperales	Piperaceae	*Peperomia sp.*
	Amaranthaceae	*Iresine sp.*
Centrospermae	Portulacaceae	*Portulaca sp.*
	Caryophyllaceae	*Drymaria sp.*
Rhoedales	Cruciferae	*Lepidium sp.*
	Crassulaceae	*Echeverria sp., Tillaea sp.*
Rosales		
	Leguminosae	*Eysenhardtia sp., Phaseolus sp., Crotolaria sp.*
	Oxalidaceae	*Oxalis sp.*
Geraniales	Burseraceae	*Bursea sp.*
	Euphorbiaceae	*Euphorbia sp.*
Sapindales	Anacardiaceae	*Schinus sp.*
	Sapindaceae	*Cardiospermum sp.*
Rhamnales	Vitaceae	*Cissus sp.*
Malvales	Malvaceae	*Anoda sp.*

Órdenes	*Familias*	*Géneros*
Parietales	Begoniaceae	*Begonia sp.*
Opuntiales	Cactaceae	*Opuntia sp., Mammilaria sp.*
	Lythraceae	*Cuphea sp.*
Myrtiflorae		
	Onagraceae	*Oenothera sp.*
Umbeliferae	Umbeliferae	*Arracacia sp.*
Plumbaginales	Plumbaginaceae	*Plumbago sp.*
Contortae	Loganiaceae	*Euddleja sp.*
	Asclepiacaceae	*Gonolubus sp.*
Tubiflorae	Hydrophylaceae	*Wigandia sp.*
	Labiatae	*Salvia sp.*
Órdenes	*Familias*	*Géneros*
Rubiales	Rubiaceae	*Bouvardia sp.* *Parthernium sp., Dahlia sp.* *Baccharis sp., Conyza sp.*
Campanulatae	Compositae	*Gnaphalium sp., Tagetes sp.* *Verbesina sp., Senecio sp.* *Bidens sp., Ageratum sp.* *Eupatorium sp.*

Tabla 3. Géneros y familias de fanerógamas que se encontraron representadas en el Pedregal de San Ángel.

TABLA 4

69-28 0 0 DIAGNÓSTICO	61-22 0 0 DIAGNÓSTICO	32-53 0 DIAGNÓSTICO	36 42 65 68 DIAGNÓSTICO	87100118-38	160-49 0 0 DIAGNÓSTICO
10.00002	10.00007	10.00001	10.00000		10.00001
20.00001	20.00004	20.00000	20.00000		20.00001
30.00003	30.00008	30.99999	30.00000		30.00001
40.00001	40.00003	40.00000	40.00000		40.00000
50.00001	50.00003	50.00000	50.00000		50.00000
60.00000	60.00001	60.00000	60.00000		60.00000
70.00000	70.07138	70.00000	70.00000		70.00000
80.00000	80.00000	80.00000	80.00000		80.00000
90.00001	90.00003	90.00000	90.00000		90.00000
100.00001	100.00002	100.00000	100.00000		100.00000
110.00000	110.00000	110.00000	110.00000		110.00000
120.00000	120.00000	120.00000	120.00000		120.00000
130.00001	130.00004	130.00000	130.00000		130.00001
140.00002	140.00005	140.00000	140.00000		140.82214
150.00000	150.00000	150.00000	150.00000		150.00000
160.00000	160.92792	160.00000	160.00000		160.00000
170.00000	170.00000	170.00000	170.00000		170.00000
180.00001	180.00002	180.00000	180.00000		180.00000
190.00000	190.00000	190.00000	190.00000		190.00000
200.00000	200.00000	200.00000	200.00000		200.00000
210.00000	210.00001	210.00000	210.00000		210.00000
220.00000	220.00001	220.00000	220.00000		220.00000
230.09300	230.00000	230.00000	230.00000		230.00000
240.00000	240.00001	240.00000	240.00000		240.17775
250.00000	250.00000	250.00000	250.00000		250.00000
260.00000	260.00001	260.00000	260.00000		260.00000
270.00000	270.00001	270.00000	270.00000		270.00000
280.00000	280.00000	280.00000	280.00000		280.00000
290.00000	290.00001	290.00000	290.00000		290.00000
300.90677	300.00003	300.00000	300.00000		300.00000
310.00005	310.00017	310.00000	311.00000		310.00003

Tabla 4. Listado de computadora que muestra el diagnóstico probabilístico de 5 individuos de familias pertenecientes a las fanerógamas. El encabezado de cada columna indica, en clave, (tabla 2) las características de los ejemplares que fueron utilizadas para llevar a cabo la identificación. Las dos primeras cifras (de izquierda a derecha) de cada columna, representan también en clave, las familias y las 6 siguientes cifras, la probabilidad de que el individuo, con las características señaladas en el encabezado de la columna, pertenezca a cada una de las familias.

TABLA 5

Géneros	*Familias*	*Probabilidades*
Stipa sp.	Gramineae	1.00000
		0.00000
		0.00000
Leptochloa sp.	Gramineae	1.00000
		0.00000
		0.00000
Setaria sp.	Gramineae	1.00000
		0.00000
		0.00000
Panicum sp.	Gramineae	1.00000
		0.00000
		0.00000
Paspalum sp.	Gramineae	1.00000
		0.00000
		0.00000
Tripsacum sp.	Gramineae	1.00000
		0.00000
		0.00000
Cyperus sp.	Cyperaceae	0.00000
		1.00000
		0.00000
Calochortus sp.	Orchidaceae	0.00000
	Liliaceae	0.99998
	Convulvulaceae	0.00002
Echcandia sp.	Orchidaceae	0.00000
	Liliaceae	0.99998
	Convulvulaceae	0.00002
Alliumm sp.	Orchidaceae	0.00000
	Liliaceae	0.99998
	Convulvulaceae	0.00002
Manfredas sp.	Gramineae	0.00008
	Amaryllidaceae	0.99992
	Dioscoreaceae	0.00000
Agave sp.	Gramineae	0.00008
	Amaryllidaceae	0.99992
	Umbeliferae	0.00000
Dioscorea sp.	Dioscoreaceae	1.00000
		0.00000
		0.00000
Habenaria sp.	Orchidaceae	1.00000
		0.00000
		0.00000

Géneros	*Familias*	*Probabilidades*
Spiranthes sp.	Orchidaceae	1.00000
		0.00000
		0.00000
Peperomia sp.	Piperaceae	1.00000
		0.00000
		0.00000
Iresine sp.	Amaranthaceae	1.00000
		0.00000
		0.00000
Lepidium sp.	Cruciferae	1.00000
		0.00000
		0.00000
Echeverria sp.	Commelinaceae	0.00087
	Crassulaceae	0.99904
	Plumbaginaceae	0.00004
Tillaea sp.	Commelinaceae	0.00087
	Crassulaceae	0.99904
	Plumbaginaceae	0.00004
Phaseolus sp.	Leguminosae	0.00000
		1.00000
		0.00000
Crotalaria sp.	Leguminosae	1.00000
		0.00000
		0.00000
Eysenhardtia sp.	Leguminosae	1.00000
		0.00000
		0.00000
Euphorbia sp.	Euphoribiaceae	0.92792
	Orchidaceae	0.07138
	Compositae	0.00017
Cardiospermum sp.	Sapindaceae	1.00000
		0.00000
		0.00000
Cissus sp.	Vitaceae	1.00000
		0.00000
		0.00000
Begonia sp.	Begoniaceae	1.00000
		0.00000
		0.00000
Cuphea sp.	Lythraceae	0.99999
	Asclepiadaceae	0.00001
	Convolvulaceae	0.00000

Géneros	*Familias*	*Probabilidades*
Oenothera sp.	Rubiaceae	0.90677
	Onagraceae	0.09300
	Compositae	0.00005
Arracacia sp.	Umbeliferae	1.00000
		0.00000
		0.00000
Plumbago sp.	Plumbaginaceae	0.99986
	Rubiaceae	0.00010
	Umbeliferae	0.00004
Gonolobus sp.	Asclepiadaceae	1.00000
		0.00000
		0.00000
Ipomeea sp.	Convulvulaceae	1.00000
		0.00000
		0.00000
Evoluvulus sp.	Convulvulaceae	0.00000
		1.00000
		0.00000
Wigandia sp.	Hydrophyllaceae	0.99953
	Commelinaceae	0.00043
	Plumbaginaceae	0.00002
Bouvardia sp.	Rubiaceae	1.00000
		0.00000
		0.00000
Oxalis sp.	Oxalidaceae	0.82214
	Umbeliferae	0.17776
	Compositae	0.00003
Bursera sp.	Burseraceae	1.00000
		0.00000
		0.00000
Anoda sp.	Malvaceae	0.99970
	Leguminosae	0.00029
	Compositae	0.00001
Mammilaria sp.	Cactaceae	1.00000
		0.00000
		0.00000
Opuntia sp.	Cactaceae	1.00000
		0.00000
		0.00000
Verbesina sp.	Compositae	1.00000
		0.00000
		0.00000

Géneros	*Familias*	*Probabilidades*
Dahlia sp.	Compositae	1.00000
		0.00000
		0.00000
Parthenium sp.	Compositae	1.00000
		0.00000
		0.00000
Senecio sp.	Compositae	1.00000
		0.00000
		0.00000
Baccharis sp.	Compositae	1.00000
		0.00000
		0.00000
Bidens sp.	Compositae	1.00000
		0.00000
		0.00000
Conyza sp.	Compositae	1.00000
		0.00000
		0.00000

Tabla 5. Diagnóstico de familia con las 3 probabilidades más elevadas obtenidas con el programa de computadora para los ejemplares colectados en el Pedregal de San Ángel. A la izquierda se señalan los géneros de los ejemplares obtenidos con un procedimiento usual de identificación.