Acta Cryst. (1976). A32, 927

Formal Aspects of the Interpretation of Fourier Maps

BY JAN C. J. BART AND A. BUSETTI

Montedison Research Laboratories 'G. Donegani', Via del Lavoro 4, Novara, Italy

(*Received* 2 *March* 1976; *accepted* 10 *April* 1976)

General set and graph theory have been applied to formalize the process of recognition of a set of electron density maxima in Fourier maps as a chemical (sub)structure in order to avoid human intervention. The information content of F and E maps is discussed in relation to the formulation of hypotheses on the labelling function relating the valued chemical structure graph to the non-valued Fourier graph. Various combinatorial methods based on graph theory, which find application in chemical structure information retrieval techniques, are indicated to establish relations between (sub)structures, leading to one-to-one mapping of the Fourier maxima to specific entities of the known organochemical structure.

1. Introduction

Recent theoretical advances (*e.g.* direct methods), progress in software techniques (*e.g.* fast Fourier methods) and improved accessibility to hardware features (*e.g.* computer graphics systems) have provided most crystallographic laboratories with the essential tools for modern crystal structure determination. In this process, human intervention is still normal in the interpretation of Fourier (F) and E maps; actually, the latter analyses are in a vague and relatively unsystematic state. In manual topological structure analysis, old-fashioned shadowing techniques have gradually been replaced by pictorial pattern recognition methods and prints or drawings expressing interconnexions between true and false maxima. The crystallographer uses intuitively the concept of similarity among structures in the survey of F maps, *i.e.* in the attempts to relate structure and electron density maxima. It appears desirable now to minimize such human effort and intuition by development of algebraic solutions to the F map. We describe here the principles of correlation of a stored computer version of the topological chemical structure with the stored electron density map. Obviously, such procedures are of great help especially for multiple Fourier calculations (*cf.* direct methods) and when automatic crystal structure determination is being pursued.

To achieve the goal, F maps and chemical structures are to be described in a form amenable to further elaboration. Graphs and Boolean matrices are the best mathematical tools for this purpose. For direct inspection, Fourier patterns represented in matrix form offer little advantage over a graphic representation. However, matrices may be arranged according to certain principles so as to reveal the presence of subgroupings. Also, matrix algebra allows analysis of the structure of a set, *e.g.* by operations of squaring and cubing. In the former case, each entry in the resulting matrix stands for the number of two-step connexions between the specified two members of the group. Diagonal entries indicate the number of two-step connexions existing from a node to itself, thus the coordination number. Elements in the cubed matrix indicate walks of length three between any two nodes, *etc.* It is thus possible to obtain such information as which atoms are indirectly connected to each other, how indirect is this connexion, which nodes are connected to the greatest number of nodes, *etc.* Being able to handle efficiently such aspects of group structure should make it feasible to handle more adequately the interpretation of F maps.

2. General considerations

2.1. *Crystallography and mathematics*

Whereas the use of mathematics in crystallography has been oriented mainly towards the development of theory and generation of models which account for processing of experimentally observed data, little effort has been put into formalizing the concept of structure. Modern approaches of mathematics make extensive use of algebraic structures, ordered sets, graph theory and topological spaces (Bourbaki, 1966). Graph theory and combinatorial analysis are, finding increasing applications as a tool of analysis in widely differing areas of science and technology (Seshu & Reed, 1961; Flament, 1963; Sussenguth, 1964), have found wide use also in strictly chemical problems (Balaban, 1967; Lederberg *et al.*, 1969; Rouvray, 1971), but much less in X-ray crystallography. A systematic exposition of properties of networks and of relevant techniques may lead to the construction of networks capable of expressing mathematical relations in a new fashion. Network geometry allows for carrying out automatic calculations in any problem that can be represented in this manner. It is our intention to develop a mathematical concept of chemical structure identification, confined and prejudiced as little as possible by our experience. The formalism is based on the algebraic theory of sets and relations and can be used as a model for the definition of some characteristic properties of crystallographic reality.

2.2. *The combinatorial Fourier problem*

The determination of the nature of the maxima in electron density maps may intuitively be considered as a combinatorial problem. A formal treatment of this subject is given below. Here we just wish to observe that experience with such problems (Unger, 1964) teaches that a broad set-up of testing a variety of incomplete routes is often more efficient than an exhaustive approach in depth. But even such simple 'breadth' search is usually computationally expensive and prohibitive for big structures with high branching ratios. In fact, consider the critical permutation problem. Suppose that the F map consists of m peaks and the molecule of n atoms ($m \leq n$). Disregarding the structure (*i.e.* neighbourhood relations), the true atom-by-atom correspondence is one out of

$$\binom{n}{m} m! = \frac{n!}{(n-m)!} \tag{1}$$

sets. Direct comparison between all possible subsets of equal order of the atoms of the chemical structure and the Fourier maxima would therefore be a tremendous task if we set aside the concept of 'structure' and especially when we consider that (1) diverges rapidly both with the increasing number of atoms n of the molecule under investigation and the growth of the F map.

If we discard the atom-by-atom correspondence and impose the sole condition that Fourier peaks must be mapped onto the appropriate atomic species (this requirement corresponds to the attribution of the correct f potentials), we need to perform

$$\sum \left(\frac{\Pi \binom{n_i}{m_i}}{\Pi (m_i)!} \right) m! \tag{2}$$

permutations (*cf.* also Feller, 1968), where n_i is the number of atoms of kind i in the molecule and the sum extends over all the partitions $\{m_i\}$ of m in subsets m_i such that $\sum m_i = m$, $m_i \leq n_i$ and $\sum n_i = n$. The numerator is in fact that of a hypergeometric distribution and expresses the number of ways to extract $m_i \leq n_i$ objects once fixed $\{m_i\}$, such that $\sum m_i = m$; however, when these are considered to be the same, as is the case when we do not take into account the fact that the atom i may not belong to the substructure represented in the F map, the numerator equals 1.

We observe that (1) is insensitive to whether atoms are all of the same kind or not, whereas (2) equals the numerator $\binom{n}{m}$ for an equal-atom structure and approaches 1 for $m \to n$. Although this may seem to be an obvious result, (2) has the property that the better the quality of the F map, the easier the matching process; this is different from the procedure according to (1), which determines the atom-by-atom correspondence and where $n!$ checks are necessary for $m \to n$.

Let us now consider an addendum of (2) and suppose we know the number m_i of atoms in set i of the F map on the basis of peak heights (in this case the sum reduces to one term). When $m \to n$ and $m_i \to n_i$ the formula reduces to

$$\frac{n!}{n_1! n_2! \cdots} = \frac{n!}{\Pi (n_i!)}. \tag{3}$$

But (2) also reduces to $n!/\Pi(n_i!)$ when $m \to n$ because if $n = m$ and $\sum n_i = n$, the only partition $\{m_i\}$ of m such that $m_i \leq n_i$ is $\{n_i\}$. This means that with a perfect F map (all and only true peaks are present), disregarding structural features such as bonds and peak distances, the desired match (*i.e.* the allocation of the proper scattering factor to each maximum in the F map) is contained in $n!/\Pi(n_i!)$ permutations. In case just one class $\bar{j}$ has been found (heavy atoms), (2) becomes

$$\frac{(n-n_{\bar{j}})!}{\Pi(n_i!)/n_{\bar{j}}!} = \frac{(n-n_{\bar{j}})! n_{\bar{j}}!}{\Pi(n_i!)} \tag{4}$$

which is much less than $n!/\Pi(n_i!)$. [Notice that the ratio (3)/(4) is exactly $\binom{n}{n_{\bar{j}}}$, *i.e.* the number of attempts to be performed to recognize class $\bar{j}$]. But if n is high and there is a great variety in electron densities in an incomplete F map (2) becomes prohibitive. Therefore, a more selective approach is necessary. Techniques are thus needed which perform a cursory inspection of the myriad of possible combinations so that the majority of sets not satisfying the search requirement will be rejected at an early stage. In this respect, it is useful to introduce the concept of 'structure' in the F map, relying upon known distance requirements and neighbourhood relations. We therefore introduce the formal concepts of set and structure in the next section.

2.3. *Sets and structures*

Definition 1. We call a set any collection of distinguishable objects, which are called the elements, members or objects of the set.

A set may be defined either by enumeration of its elements between braces or by a common property of its members, *e.g.* the set of even numbers $S = \{0,2,4,6,\ldots\} = \{\text{integers } n \text{ such that } n = 2m\}$. We shall write $s \in S$ for 's belongs to S'.

Definition 2. A set A is a subset of a set B if $a \in A$ implies $a \in B$.

The empty set $\varnothing$ is the set with no elements. Given n sets A_i, $i = 1,\ldots,n$, we consider the following sets:

$\bigcap_{i=1}^{n} A_i = \{x | x \in A_i \text{ for any } i\}$ (intersection $A_1 \cap A_2 \cap \ldots \cap A_n$).

$\bigcup_{i=1}^{n} A_i = \{x | x \in A_i \text{ for at least one } i\}$ (union $A_1 \cup A_2 \cup \ldots \cup A_n$).

$\prod_{i=1}^{n} A_i = \{(x_1,\ldots,x_n) | x_i \in A_i \text{ for any } i\}$ (product set $A_1 \times A_2 \times \ldots \times A_n$); if $A_1 = A_2 = \ldots = A_n = A$ the product set is A^n.

Definition 3. An n-ary relation between n sets A_i, $i = 1,\ldots,n$, is a subset $\varrho \subset \prod_{i=1}^{n} A_i$.

Definition 4. A function is a binary relation $\varrho \subset A \times B$ such that if $(a,b) \in \varrho$ and $(a,c) \in \varrho$ then $b=c$, and for any $a \in A$ there is a $b \in B$ such that $(a,b) \in \varrho$ [usually written $b=\varrho(a)$].

Definition 5. A structure $\mathscr{S}=(S,\varrho)$ is a couple constituted of a set S and a binary relation $\varrho \subset S^2$.

Example 1. A molecule is a structure. In fact, let S be the set of atoms, then for $\varrho \subset S \times S$ so defined $(a,b) \in \varrho$ if and only if there exists a bond between atoms a and b.

Definition 6. Given two structures $\mathscr{S}=(S,\varrho)$ and $\mathscr{S}'=(S',\varrho')$, we write $\mathscr{S}$ is a substructure of $\mathscr{S}'$ ($\mathscr{S} \subset \mathscr{S}'$) if $S \subset S'$, $\varrho \subset \varrho'$.

N structures $\mathscr{S}_i=(S_i,\varrho_i)$ are said to be disjoint if $\bigcap_{i=1}^{n} S_i = \varnothing$; obviously, $\cap S_i = \varnothing \Rightarrow \cap \varrho_i = \varnothing$, but the converse is not true.

Definition 7. We define a union of n structures $\mathscr{S}_i=(S_i,\varrho_i)$ the couple $\mathscr{S}=(\bigcup_{i=1}^{n} S_i, \bigcup_{i=1}^{n} \varrho_i)$.

It is a straightforward exercise to show that $\mathscr{S}$ is a structure. A structure is said to be disconnected if it is the union of disjoint structures, connected if not.

Definition 8. A function $\varphi \subset A \times B$ (A,B could be product sets) is called one-to-one if $a \neq b$ implies $\varphi(a) \neq \varphi(b)$.

Definition 9. Two structures $\mathscr{S}=(S,\varrho)$, $\mathscr{S}'=(S',\varrho')$ are monomorphic if there is a one-to-one function $\varphi \subset S \times S'$ (so-called monomorphism) such that $(a,b) \in \varrho$ if and only if $[\varphi(a),\varphi(b)] \in \varrho'$; they are called isomorphic if φ^{-1} is also a one-to-one function (*i.e.* φ is also onto).

We will call automorphism any isomorphism φ from a structure $\mathscr{S}=(S,\varrho)$ onto itself (there is always at least one automorphism, the identity from S onto itself).

A structure is called symmetric when the relation is symmetric, *i.e.* when $(a,b) \in \varrho$ if and only if $(b,a) \in \varrho$. (Obviously, a chemical structure is symmetric.)

As is well-known (Lynch, Harrison, Town & Ash, 1971), a molecule may be stored in the computer as a structure (adjacency matrix). But, given a structure, it is quite difficult to distinguish such objects as cycles and chains. This is because they are geometrical rather than algebraic entities. We need therefore a *geometrical* description of a structure; such a description is usually called a graph.

2.4. *Graphs*

Definition 10. Given a connected structure (S,ϱ) and an ordered n-tuple $(a_i,\ldots,a_n) \in S^n$ (that is $a_i \in S$), we recognize a path of length $n-1$ if $(a_i,a_{i+1}) \in \varrho$ for any $i=1, n-1$.

It is often called a path of length n from a_1 to a_n. By definition, (a_1) is a path of length 0.

Definition 11. We define metric space as the couple (R,δ) where R is a set and $\delta \subset [(R \times R) \times \mathbb{R}]$ ($\mathbb{R}$ = set of real numbers) is a function such that: (i) $\delta(a,a)=0$; (ii) $\delta(a,b)=\delta(b,a) \geq 0$; (iii) $\delta(a,b)+\delta(b,c) \geq \delta(a,c)$ for any $a,b,c \in R$; δ is called a distance.

Definition 12. Given a connected symmetric structure (S,ϱ), we call a connected undirected graph the couple (S,δ) where $\delta \subset (S \times S) \times \mathbb{R}$ is the function so defined that $\delta(a,b)$ = minimum length of paths between a and b for any $a,b \in S$.

Definition 13. Given a structure $\mathscr{S}$ and the graph G generated from it, we call a connected subgraph of G any graph G' generated from a connected substructure $\mathscr{S}' \subset \mathscr{S}$. (We shall write $G' \subset G$.)

It is straightforward to show:

Proposition 1. A connected graph is a metric space.

Usually no distinction is made between a structure and its associated graph. Consequently, both geometric (topological and/or metric) and algebraic aspects are often involved in chemical structure matching.

Definition 14. Two metric spaces (S,δ), (S',δ') are called monometric (isometric) if there exists a one-to-one function $\varphi \subset S \times S'$ such that $\delta(x,y)=\delta'[\varphi(x),\varphi(y)]$ for any $x,y \in S$ [and conversely for any $x',y' \in S'$ there is $x,y \in S$ such that $\delta'(x',y')=\delta(x,y)$ and $\varphi(x)=x'$, $\varphi(y)=y'$].

An immediate consequence of the definition of a graph is the following:

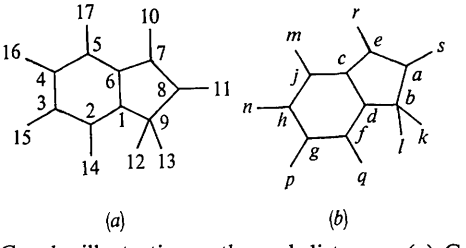
Proposition 2. Two graphs are monometric (isometric) if and only if the structures they are generated from are monomorphic (isomorphic).

We notice that there is always a monomorphism between a substructure $\mathscr{S}' \subset \mathscr{S}$ and a structure $\mathscr{S}$; a subgraph $G' \subset G$ and the graph G are thus always monometric.

The last proposition is the main reason why we are allowed to identify graphs with structures. In fact, there is a one-to-one correspondence between undirected connected graphs and connected symmetric structures.

One can visualize a graph (and a structure, *cf.* proposition 2) in the following way: (i) draw points on a surface, each point representing an element s of the object set S of the graph; these points are called vertices of the graph; then (ii) draw a continuous line between two vertices a,b if and only if $\delta(a,b)=1$ [*i.e.* $(a,b) \in \varrho$ which is the relation of the structure generating the graph]; such a line is called an edge.

Terms like path and distance (defined before) now become meaningful. Note that in Fig. 1 G and F are isometric but they are not the same graph since their object sets are different.



(*a*) (*b*)

Fig. 1. Graphs illustrating paths and distances. (*a*) Graph *G*. (*b*) Graph *F*.

Other important concepts are (*cf.* Sutherland, 1967): *Local* or *node degree* $\varrho(x)$ of a vertex x is the number of edges of the graph incident with it. For any subset $A, \varrho(A)=\sum\varrho(x)$ represents the total number of edges issuing from the vertices in A, counting twice those which have both their endpoints in A. *Gamma set* of a node x is the list of nodes reachable from x by a path of specified length. The Γ^1 set of x, $\Gamma^1(x)$, is the list of all nodes immediately connected to x; the $\Gamma^m(x)$ set is the list of all nodes connected to the nodes in the Γ^{m-1} set of node x. *Gamma degree*: the gamma$=m$ degree of a node x is the number of nodes of the gamma$=m$ set for node x. *Connectivity* of order m of a set of nodes $\{A\}$ is another set of nodes $\Gamma^m\{A\}$ representing all nodes which can be reached by a path of given length (*i.e.* traversing m bonds) from any one of the original nodes. The connectivity of a set of nodes $\{A\}$ is thus the union of the gamma sets of all its nodes. Connectivities are calculated for paths until redundancies are generated. The connectivity is usually found for paths of length 1 or 2. Examples of a gamma set and connectivity of the first order are $\Gamma^1(1)=\{2,6,9\}$ and $\Gamma^1\{h,f\}=\{d,g,j,n,q\}$ in Fig. 1. *Quasi-order* or *node order* of node x is the number of connexions that must be traversed before getting back to the given x avoiding direct backtracking; null order is attributed if x is not in a ring. In set generation (Sussenguth, 1965) the concept of quasi-order is used in connexion with the smallest rings in the structure.

2.5. *Chemical structures and Fourier graphs*

Although it is not strictly necessary to use chemical knowledge in crystal structure determinations (*cf.* superposition methods in Patterson analysis), a variety of methods for (partial) deconvolution of the Patterson function take *direct* advantage of this information in the form of the conformation of (part of) the molecule (Nordman & Nakatsu, 1963; Huber, 1965; Braun, Hornstra & Leenhouts, 1969; Hornstra, 1970; Sirota, Galiulin & Simonov, 1974). This paper utilizes the information content of the topological chemical structure only.

In analogy to example 1, both the chemical constitution of a molecule and the topology of an F map can be defined as neighbourhood relations between atoms and between electron density maxima, respectively; both are visualized in topical maps in terms of bonds and bonded neighbours, *i.e.* as structures.

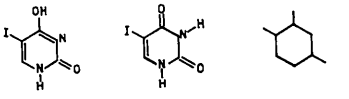
In the following treatment of structural diagrams and Fourier maps as connected and undirected graphs, the skeleton of the hydrogen-suppressed chemical structure will be called the inquiry graph G; the Fourier graph F denotes the (set of un-)connected graph(s) based on the assemblage of maxima in the electron density map connected by presumed bonds (derived from distance criteria). Apart from node values in G no other chemical information is included. The object sets in the graphs G and F are indicated by $\mathscr{G}$ and $\mathscr{F}$.

As to the details of these graphs, we notice:

(i) Stereochemical detail is not normally expressed. Thus, graphs differing only by the spatial relations of the nodes and the orientations of the branches (molecular geometry), are equivalent, *e.g.* stereoisomers or crystallographically independent molecules.

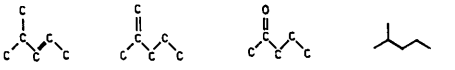
(ii) No positional or orientational parameters are expressed. The relative positions of the nodes in the graphs do not necessarily reflect a configuration relative to any preconceived coordinate or geometric system.

(iii) In the F graph no distinct node and branch values are considered, as these features are absent in electron density maps, which are typically non-valued structures. Hydrogen-suppressed line-graphs of the tautomers (I) and (II) are the same (III):



(I) (II) (III)

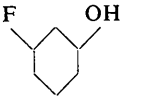
Isomers such as (IV) and (V) are no longer distinguishable and not resolvable from (VI)



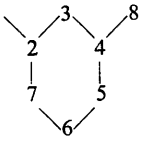
(IV) (V) (VI) (VII)

on the basis of the unweighted topology of F maps, as their molecular skeletons (VII) are equivalent. The equivalence of chemically distinct but topologically identical representations is not restrictive in the following treatment.

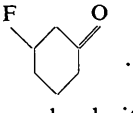
Finally, from consideration of the molecule (*i.e.* the valued graph G):



and the associated F-graph



we notice that they are clearly isometric, although we cannot decide which one of the atoms 1 and 8 corresponds to F. Similarly, in



This is because the F graph admits a non-trivial automorphism* and the information about bonds and atoms is entirely lost. We need therefore a more detailed description of the molecule because its 'geometrical' description is ambiguous.

* This is shorthand for: 'the structure generating F admits a non-trivial automorphism'.

2.6. *Valued structures and valued graphs*

We define a valued (V) structure as the quadruple $(S,\varrho,\lambda,\beta)$ such that: (i) (S,ϱ) is a structure; (ii) $\lambda\subset\varrho\times\Lambda$ (where Λ is the set of edge-labels) is a function; (iii) $\beta\subset S\times B$ (B is the set of point-labels) is a function. For example, the water molecule H–O–H is a valued structure; in fact, let S be the set $\{1,2,3\}$, ϱ the symmetric relation $\{(1,2),(2,3)\}$ (non-ordered couples), if $\Lambda=\{$single bond, double bond, triple bond$\}$, $B=\{$atomic symbols$\}$, $\lambda=\{[(1,2),$ single bond$],[(2,3),$ single bond$]\}$, $\beta=\{(1,\mathrm{H}),(2,\mathrm{O}),(3,\mathrm{H})\}$, then the valued structure $(S,\varrho,\lambda,\beta)$ describes unambiguously the molecule and can easily be stored in a computer.

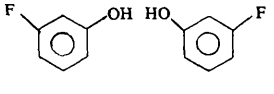
Definition 15. Two valued structures $\mathscr{S},\mathscr{S}'$ are called valued monomorphic (isomorphic) if (S,ϱ), (S',ϱ') are monomorphic (isomorphic) and $\Lambda=\Lambda'$, $B=B'$, $\beta(s)=\beta'[\varphi(s)]$ for any $s\in S$ and $\lambda[(s,r)]=\lambda'[\varphi(s),\varphi(r)]$ for any $(s,r)\in\varrho$. That is, φ must preserve some properties of the elements of S and ϱ, which are described by edge-labelling functions λ,λ' and point-labelling functions β,β'. For instance, the two V structures of Fig. 1 are isomorphic as structures but not as V structures. Thus the concept of a valued structure is necessary if we want our morphisms to preserve some chemical properties. In a similar way valued automorphisms can be defined.

The usual concept of a valued graph follows immediately:

Definition 16. Given a valued structure $\mathscr{S}=(S,\varrho,\lambda,\beta)$, we call a valued graph the quadruple $G=(S,\delta,\bar{\lambda},\bar{\beta})$ such that (S,δ) is the graph generated from (S,ϱ), $\beta=\bar{\beta}$ and $\bar{\lambda}$ is the function assigning to each path (a,b) of length l the label $\lambda(a,b)$. [It would be more correct to state: 'assigning to each path $(a_1,\ldots,a_n)$ a word $\lambda_1\lambda_2\ldots\lambda_{n-1}$ such that $\lambda_i=\lambda(a_i,a_{i+1})$', but this leads to the same results and makes use of concepts from formal languages, which have not been defined here.]

We define valued monometries and isometries between valued graphs in the same way as V monomorphisms and V isomorphisms.

Obviously, two V structures may be isomorphic but not V isomorphic, as we have seen above. There may also be a function which is an isomorphism but not a valued one even if the V structures are V isomorphic. Example:



$\begin{cases}\varphi(\mathrm{F}) = \mathrm{OH}\\ \varphi(\mathrm{OH}) = \mathrm{F}.\end{cases}$

3. Formalization of the problem

3.1. *The interpretation of Fourier maps*

The problem we are facing can now be formulated in a more precise way. There are given two V structures $\mathscr{S},\mathscr{S}'$, one of which is without labelling functions. The structures are supposed to be V isomorphic because they are the 'same' molecule; we are interested in finding the unknown point-labelling functions of the F map, *i.e.* in identification (mapping) of its vertices. (It is fair to deal with V structures since chemical structures and F maps are stored in the computer as structures.)

Since $\mathscr{S}$ and $\mathscr{S}'$ are V isomorphic, they can be supposed to be the same V structure (and conversely). (In the set $\mathscr{V}$ of valued chemical structures the condition of V isomorphism is an equivalence relation and therefore partitions $\mathscr{V}$ in such a way that the members of the same class are indistinguishable from the chemical structural standpoint, *i.e.* they are conformers). The solution to this problem is usually searched for considering only isomorphisms and not V isomorphisms. We have seen that this is ambiguous even if only true peaks are present. Suppose in fact that $\mathscr{S}$ (the V-structure) admits n automorphisms and $m\leq n$ V automorphisms. Since we do not know the labelling functions, the probability of getting an automorphism which is a non-valued one is $p=(n-m)/n$. This means that we have a probability p of error on the kind of atoms and/or bonds. We see that this probability is zero if and only if each automorphism is a valued one ($m=n$), that is if and only if information about the kind of atoms and bonds (peak heights and peak distances) is completely redundant. This is not often the case. It becomes then self-evident that if $p\to 1$ (*i.e.* when there is little information on the geometrical structure as in most organic molecules) one must know at least partially the labelling functions in order to minimize the error probability. This is the case of heavy-atom Fourier maps or of refined Fourier patterns of light-atom structures when the electron densities have taken up reliable values which permit distinction between, say, C and O peaks. It is clear then that in principle different algorithms are to be used as a function of the value of p: if $p\to 0$ algorithms from graph theory are useful; if this is not the case, hypotheses on labelling functions are highly desirable. From a study of the automorphism group of the molecule those atoms may be identified whose recognition would minimize p. In fact, suppose that only n_1 isomorphisms map oxygens onto oxygens, then their identification in the F map and the condition that oxygens must be mapped onto oxygens would reduce p to $(n_1-m)/n_1$. The probability of error p then depends obviously on the information (I) available about the F map, which is the sum of two parts, structural information (in the structure) and chemical information (labelling functions). Since we want to keep p constant in the algorithm, I should be constant. This cannot be done considering only structural information since the latter depends upon the number of isomorphisms and V isomorphisms. When spurious peaks are present in the F map, the probability of error increases because of the possibility of considering wrong vertices. Similarly, p increases when true electron density maxima are missing; in fact, with n automorphisms of the structure and q automorphisms admitted by F, there are $n\leq n_1\leq nq$ homomorphisms from the F-map into the structure.

The general method to establish whether an F graph is relevant to an inquiry graph thus cannot be a simple one. The process which recognizes the Fourier graph homomorphically in the inquiry graph is satisfactory only in the most favourable cases, *i.e.* when the structure has no automorphism other than the identity. Indeed, such a straightforward process mostly fails, especially when the F graph is incomplete. Often there is here even the incidental occurrence of false electron density maxima, leading eventually to false connexions between nodes. The retrieval problem is thus to locate an arbitrary subset of the chemical structure G in the stored F structure.

3.2. *Classification of Fourier graphs*

In order to describe the retrieval problem in more detail, it is useful to classify F maps according to four forms of relatedness to the chemical structure:

(1) $F=G$. Then $G\cap F=G\cup F=G$ and $G-F=\varnothing$. In other words, G and F are isomorphic (*cf.* Fig. 1, where all nodes x of G are uniquely related to x' in F, so that $\Gamma x_G=\Gamma x'_F$).

(2) $F\subset G$. The F graph is a subgraph of the chemical structure graph. Then $G\cap F=F$, $G\cup F=G$; $G-v=F$ is obtained by removing from G the set $\{v_i\}$ and relative incident edges. The set $\{v_i\}$ represents the atoms which are not resolved in F. The previous case is the special case of $\{v_i\}=\varnothing$.

(3) $F\not\subset G$. If $\{s_i\}$ denotes a set of vertices (*i.e.* the set of spurious peaks), then the graph whose vertices are precisely all those vertices of F which are not in $\{s_i\}$ and whose edges are precisely all those edges of F with end-vertices not in $\{s_i\}$, is denoted $(F-s)$. Suppose $(F-s)\subset G$. The F graph and the chemical structure graph have common subgraphs: the F map contains part of the structure, but also spurious peaks. Thus $G\not\subset F$, $G\cap F=(F-s)$, $G\cup F=F+v=G+s$ and $\{\mathscr{G}\}-\{\mathscr{F}\}=\{v_i\}-\{s_i\}$.

(4) $F\not\subset G$, $(F-s)=G$. The chemical structure graph is a subgraph of the Fourier graph: G represents only a fragment of a chemical compound or contains extraneous vertices due to phase errors. Thus $G\subset F$, $G\cap F=G$, $G\cup F=F$ and $\{\mathscr{F}\}-\{\mathscr{G}\}=\{s_i\}$.

The process of determining structural relatedness between the chemical formula and the Fourier map is then that of isolating in G and F the *intersection* of the sets G and F $(G\cap F)$.

It would be desirable to be able to identify the type of F map *a priori*; however, this is generally not possible. The problem obviously loses its significance once an efficient generalized problem-solving algorithm has been developed. The complexity of the combinatorial problem for the general case ($p\simeq 1$) is still such that with contemporary equipment other, more indirect, automatic solutions are to be preferred. In equal-atom structures it is possible to increase the information content of the F map. Such an approach has been followed by Koyama & Okada (1975) for crystal structure determinations without any human intervention and is based on a sequence of convergent Fourier calculations after repeated attempts to eliminate spurious peaks by regulating the noise level and through evaluation of chemical physical data (temperature factor). Mapping may then be performed on the basis of a final, rather perfect and thus classified F map. As mentioned before, additional information on the labelling functions is also available when G and F contain vertices which are easily identified, *e.g.* heavy atoms; the consequent decrease in the value of p leads to a simpler mapping procedure, in accordance with manual experience. Automatic analysis of organic compounds from the heavy-atom position without any chemical assumption has in fact been achieved by Koyama & Okada (1970), again by treating all light atoms as C atoms and using the temperature factor to distinguish between false and true atomic sites and peak heights to recognize atomic species. In this way, the combinatorial problem is circumvented.

4. Structure matching procedures

Having formalized the problem, we must try to come up with practical solutions. Most experimental methods for structure matching are based on graph theory and have been developed for structure information retrieval purposes (*cf.* Committee on Chemical Information, 1964–1969). Procedures for mapping V graphs (non-unique representations are sufficient) obviously need considerable modification to suit line graphs (non-valued graphs), as all characteristics exploiting chemical properties should be removed (*cf.* Gould, Gasser & Rian, 1965). Topological substructure comparisons are usually based on atom-by-atom search or more selective methods (Ray & Kirsch, 1957; Ballard & Neeland, 1963; Cossum, Krakiwsky & Lynch, 1965). An essentially iterative but efficient process in a simple topological matching procedure, developed by Gluck (1965) and subsequently modified by CAS (Leiter & Morgan, 1966), does not adequately handle topological graphs.

Algorithms designed to match each node in the F map with an atomic species may start by fitting an atom of G to a node in F, followed by attempts to match some node of the respective Γ^1 sets, up to nodes in higher-order sets. A stage might eventually be reached at which either all nodes in the two structures are matched, or no further successful matches can be made, at which point back-tracking is necessary, eventually several levels, until an untried branch is found. If all possible branches have been tried out without success, a no-match condition has been found. Back-tracking is expensive, *e.g.* the time required to find a no-match condition between two similar structures varies as 2^n, where n is the number of atoms in the smaller structure. Iterative node-by-node searching techniques usually require screening devices and short cuts to minimize non-productive path tracing. In case of the F graph, the probabibilty of finding a match may be

somewhat enhanced by designing a path over nodes with the highest electron density. Although 'educated' selective search has been applied to various combinatorial problems (Walker, 1960; Golomb & Baumert, 1965) yielding the same answer with far fewer than $N!$ trials of the brute force approach, more specialized techniques are frequently more efficient than the exhaustive back-track.

More powerful than node-by-node matching is set reduction, which makes use of various graph theoretical property values: *e.g.* node value, node degree, branch value and gamma set (Unger, 1964; Penny, 1965; Sussenguth, 1965; Figueras, 1972). The space of all feasible solutions is then partitioned into smaller subsets. The time required to determine a match/no match condition between two structures is estimated to be proportional to $(n-1)^2$, where n is the number of atoms in the smaller structure. Other rapid structural retrieval programs have been described (Feldmann, Heller, Shapiro & Heller, 1972).

In searching for suitable algorithms for matching the G and F structures, it was considered imperative to check whether existing mathematical theories offer any efficient solution. In a subsequent paper we will show that network analysis and retrieval techniques enable us to handle F graphs of type 1 (*cf.* § 4.2) by standardization procedures, types 2 and 4 by substructure search and type 3 by a general partial substructure search procedure.

5. Conclusions

The framework of the algebraic solution to the interpretation of Fourier maps of organic structures has been traced on the basis of the topological properties of the chemical graph, along the lines previously expressed by Bart & Giordano (1973) and Koch (1974). Without the introduction of neighbourhood relations (structure) the combinatorial problem of Fourier analysis is typically an explosive problem, which is eventually considerably reduced in size in cases where particular atomic species can somehow be recognized (*cf.* Koyama & Okada, 1970). In other cases, simplification may be achieved by the introduction of structure-oriented mathematics. The results of network analysis and retrieval techniques in combinatorial computations are responsible for the development presented here. Due to the low information content of Fourier structures, highly selective structure matching strategies are required to amplify human logical capacity in this field.

One of us (A. B.) is indebted to the Italian Accademia dei Lincei and Sperry Rand for a research fellowship.

References

BALABAN, A. T. (1967). *Rev. Chim. Acad. Rep. Pop. Roum.* **12**, 875–898.

BALLARD, D. L. & NEELAND, F. (1963). *J. Chem. Doc.* **3**, 196–201.

BART, J. C. J. & GIORDANO, N. (1973). *Proc. Int. Conf. on Computers in Chemical Research and Education*, Ljubljana–Zagreb, July 12–17 (1973). Edited by D. HADŽI, pp. 3,1–3,25. Amsterdam: Elsevier.

BOURBAKI, N. (1966). *Elements of Mathematics. General Topology*. Don Mills, Ont.: Addison–Wesley.

BRAUN, P. B., HORNSTRA, J. & LEENHOUTS, J. I. (1969). *Philips Res. Rep.* **24**, 85–118.

COMMITTEE ON CHEMICAL INFORMATION (1964–1969). *Survey of Chemical Notation Systems – Survey of European Non-conventional Chemical Notation Systems – Chemical Structure Information Handling* (1962–1968). Publ. Nos. 1150 (1964), 1278 (1965), 1733 (1969). Washington D.C.: National Academy Sciences.

COSSUM, W. E., KRAKIWSKY, M. L. & LYNCH, M. F. (1965). *J. Chem. Doc.* **5**, 33–35.

FELDMANN, R. J., HELLER, S. R., SHAPIRO, K. P. & HELLER, R. S. (1972). *J. Chem. Doc.* **12**, 41–47.

FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications*. Vol. I, p. 34. New York: John Wiley.

FIGUERAS, J. (1972). *J. Chem. Doc.* **12**, 237–244.

FLAMENT, C. (1963). *Applications of Graph Theory to Group Structure*. Englewood Cliffs, N. J.: Prentice-Hall.

GLUCK, D. J. (1965). *J. Chem. Doc.* **5**, 43–51.

GOLOMB, S. W. & BAUMERT, L. D. (1965). *J. Assoc. Comp. Mach.* **12**, 516–524.

GOULD, D., GASSER, E. B. & RIAN, J. F. (1965). *J. Chem. Doc.* **5**, 24–32.

HORNSTRA, J. (1970). In *Crystallographic Computing*, edited by F. R. AHMED, p. 103–109. Copenhagen: Munksgaard.

HUBER, R. (1965). *Acta Cryst.* **19**, 353–356.

KOCH, M. H. J. (1974). *Acta Cryst.* **B30**, 67–70.

KOYAMA, H. & OKADA, K. (1970). *Acta Cryst.* **B26**, 444–447.

KOYAMA, H. & OKADA, K. (1975). *Acta Cryst.* **A31**, S18.

LEDERBERG, J., SUTHERLAND, G. L., BUCHANAN, B. G., FEIGENBAUM, E. A., ROBERTSON, A. V., DUFFIELD, A. M. & DJERASSI, C. (1969). *J. Amer. Chem. Soc.* **91**, 2973–2976.

LEITER, D. P. & MORGAN, H. L. (1966). *J. Chem. Doc.* **6**, 226–229.

LYNCH, M. F., HARRISON, J. M., TOWN, W. G. & ASH, J. E. (1971). *Computer Handling of Chemical Structure Information*. p. 13. London: Macdonald.

NORDMAN, C. E. & NAKATSU, K. (1963). *J. Amer. Chem. Soc.* **85**, 353–354.

PENNY, H. (1965). *J. Chem. Doc.* **5**, 113–117.

RAY, L. C. & KIRSCH, R. A. (1957). *Science*, **126**, 814–819.

ROUVRAY, D. H. (1971). *R. I. C. Reviews*, **4**, 173–195.

SESHU, S. & REED, M. B. (1961). *Linear Graphs and Electrical Networks*. Reading, Mass.: Addison–Wesley.

SIROTA, M. I., GALIULIN, R. V. & SIMONOV, V. I. (1974). *Kristallografiya*, **19**, 54–59.

SUSSENGUTH, E. H. (1964). *Structure Matching in Information Processing*. Ph. D. Thesis, Harvard Univ., Cambridge (Mass.).

SUSSENGUTH, E. H. (1965). *J. Chem. Doc.* **5**, 36–43.

SUTHERLAND, G. (1967). *DENDRAL – A Computer Program for Generating and Filtering Chemical Structures*. Clearinghouse for Fed. Sci. and Techn. Inform. Rep. AD 676, 126, Washington D.C.

UNGER, S. H. (1964). *Commun. A.C.M.* **7**, 26–34.

WALKER, R. S. (1960). *Amer. Math. Soc. Symp. Appl. Math. Proc.* **10**, 91–94.