match no. 21 pp. 3-31 1986

MATHEMATICAL STATEMENTS ABOUT THE REVISED CIP-SYSTEM

Roland H. Custer
Koechlistrasse 3, CH-8004 Zurich

(Received: November 1986)

ABSTRACT
========
The CIP-System is investigated from a mathematical point of view (problems of consistence and completeness). After an exact and complete definition of the rules including the domain of applicability, its deficiencies are demonstrated and supplementary rules for their removal are suggested. Some of the requirements that the CIP-System fulfills are proved mathematically.

1. INTRODUCTION

1.1. The CIP-System

The CIP-System enables the chemist to describe different stereoisomers when the constitution is given (the non-chemist may consult Section 1.2). This is done by a number of stereodescriptors which are derived by rules and then attached to certain atoms (or bonds). The system is estimated by virtue of its applicability for those molecules the chemist happens to describe. But it is not equally suitable for all molecules; it is generally appropriate for

molecules with constitutions such that the differences between the stereoisomers originate exclusively from different ligand positions at atoms with tetrahedral geometry, or at double bonds (wether cumulated or not) with tetrahedral geometry or CIS-TRANS-isomerism.

As the rules have been successively developed by adding supplements for new cases, most (simple) molecules are easily treated. However, for the remaining cases, difficulties soon arise and the general case is extremely intricate. Recently, V. Prelog and G. Helmchen have revised the CIP-System [2] to overcome some of the difficulties experienced with the older rules [1]; nevertheless they respected the grown character of the system. Both papers contain, beyond partly unclear statements of the rules, many individual cases and discussions of other problems in connection with chirality, stereochemistry, and rules. But they lack mathematical exactness and systematic presentation.

As far as I know, neither the CIP-System of 1966 [1] nor the revised rules of 1982 [2] have ever been analysed mathematically. Not even an actual definition (in the mathematical sense) of the rules seems to have existed, although especially V. Prelog himself and O. Weissbach made great efforts to detect any weak points in the CIP-System, as I know from personal contacts, e.g. the discovery of the first examples of non-reconstructible molecules (according to the revised rules of 1982) by O. Weissbach.

Thus, before creating a computer program which applies the CIP-System (described in [3,4]), I looked at the rules in the way mathematicians do (definitions - theorems - proofs) supplying:

1. An exact and complete definition of the rules including the domain of applicability, hereby putting them in a sensible order, complementing them when necessary, and consulting V. Prelog personally when his published description seemed unclear to me. The domain - valid for the theoretical investigation and for the computer program [3,4] - has been defined making the best use of the few good clues available in [1] and [2].

2. A description of cases where the CIP-System fails and suggestions for supplements to the rules to remove these

shortcomings.

3. Proofs that the CIP-Rules (with some supplements) fulfill some of the requirements which are usually taken for granted.

The next section is to acquaint the non-chemist with the chemical problem of different spatial arrangements of the atoms as well as with the nomenclature that goes with it. Chapter 2 contains the basic definitions, the rules, and some mathematical definitions which are needed for the proofs. Chapter 3 treats the recognition of constitutional differences by the CIP-System, Chapter 4 the recognition of stereochemical differences. Whereas the theorems and proofs in Chapter 4 are quite simple, the result presented in Chapter 3 (Theorem 1) is not. This is due to the fact, that the hierarchical digraph (see Section 2.2) may represent the same atom several times, which makes an 'induction' of mappings from the hierarchical digraph to the original molecular graph difficult.

1.2. Three-Dimensional Chemical Structures for Non-Chemists

(This section is based on the conviction that all terms refer to models and never to the essence, and that any model is identical to its structure.)

The fundamental structures of organic chemistry are aggregates, called molecules, of a number of atoms. These molecules are described by listing the bonds between the specified atoms or by giving a systematic name which reflects this bond structure. Such a description is called a constitution. Unfortunately, there can be several different molecules, so called stereoisomers, with the same constitution, due to several possibilities of geometric arrangements of atoms and bonds.

The most common cases in organic chemistry are the bonds between a carbon atom and its 4 neighbouring atoms, and the bonds between a double-bonded pair of carbon atoms and their 4 neighbouring atoms. In the former case, the 4 neighbours shape a tetrahedron and if they are all different, then there are 2 non-congruent possibilities of such a

tetrahedron. If we define an order among these atoms, we can connect them either by a right-handed or a left-handed screwthread line; this screwthread line can be used to define a descriptor. In the latter case there are different distances between the 4 ligands - always arranged in a plane - which also gives 2 principle possibilities of arrangement, called CIS-TRANS-isomerism.

A description of a molecule taking account of all these geometric possibilities is called a configuration.

Some other terms are used in this connection: Differences in the configuration but not in the constitution are called stereochemical differences. Enantiomers are molecules which are the mirror images of each other. Diastereomers are stereoisomers (as defined above) which are not enantiomers. With parts of a molecule e.g. ligands, one speaks about stereotopic, enantiotopic, or diastereotopic parts, the definition of these terms being perfectly analogous. Of course, stereochemistry is the branch of chemistry which deals with these problems.

2. RULES AND DEFINITIONS

2.1. Molecules, Graphs, Stereoelements

A molecule is not a mathematical object, so the first requirement is the construction of an abstract model. This step is normally done by the chemists, though mostly as an unconcsious habit. For a molecule, this mathematical object is (or is based on) the chromatic graph.

def A graph is a pair of sets (X,E), where E is a relation on X (a set of ordered pairs of elements of X). The elements of X are called nodes, the elements of E edges.

There is an obvious way to visualise a graph on paper by points and arrows.

def A chromatic graph is a graph with at least one function f: X⟶A or f: E⟶A, where A is an arbitrary set.

With molecules, this A is some set of chemical information, such as atom types or bond types. The projection of molecules on chromatic graphs is so common that I omit any discussion of it (see e.g. [5]); a formal description of the involved mathematical structures can be found in Section 2.3. There is one point to mention: The CIP-System requires more than just one piece of information to a node: atom type, isotope, perhaps some geometric features. But a quick look at the rules (Section 2.2) reveals that we need only one kind of information at a time, so we may assume just one function to some scalars. Note that the necessary geometrical information is also procured by such functions; there are actually only a small, finite number of situations (i.e. ligand positions at the most common atoms in organic chemistry), which can be characterised by some descriptors (inclusive CIP-Descriptors). Extensions are possible but cumbersome (there is a complete description of the octahedral case in the Appendix of [3]). In the same way, the situations which CIP-Descriptors can describe are limited. Thus, the suitability of descriptors determines the domain where the CIP-System can be used. V. Prelog uses the so-called stereoelements but gives only vague hints how to find them. The choice of stereoelements determines directly the domain of the CIP-System. As border cases only complicate the argumentation (as well as any computer program), I confined myself to the most common cases, thus skipping the stereoplanes and keeping only a small part of the stereoaxes (see Section 2.2).

2.2. The CIP-Rules

In this section, I restate (with 2 small changes) the revised CIP-Rules [1,2] in a concise way, which can be used as a basis both for the mathematical proofs and the programs. The procedure to derive a CIP-Descriptor has 3 steps: choice of a stereoelement, ordering the ligands, determination of a descriptor. The 2nd step (here subdivided in 2 parts) is by far the most complicated.

Step 1: Choice of Stereoelements

As stereoelements, choose all atoms with quasi-rigid, tetrahedral ligand arrangement (potential stereocenters) and all double bond axes - with, in principle, angles of 180 degrees between the double bonds - with quasi-rigid, 120-degree-triligant ligand arrangements at the ends of the double bond axis, respectively (potential stereoaxes).

Step 2, Part 1: Construction of the Hierarchical Digraph

Construct a hierarchical digraph from the designated atom, i.e. the potential stereocenter or the atom at the end of the double bond axis; this is a connected, acyclic, and directed graph, and in graph theory called a tree.

The designated atom is represented by the first node, called the root. All edges should be directed away from the root. Now, all neighbour atoms are represented by a new node and the bond to them by a new edge between the new node and the root. When the passed bond has multiplicity m (m>1), m-1 additional nodes (they are called 'duplicate atoms' but are nevertheless nodes not atoms) and edges are added at each of the concerned nodes. The duplicate atoms represent the other atom of the multiple bond than the node with which they have a common edge. (With aromatic bonds, one has to take the average of the characteristics of the concerned atoms.) No further edge is added to duplicate atoms.

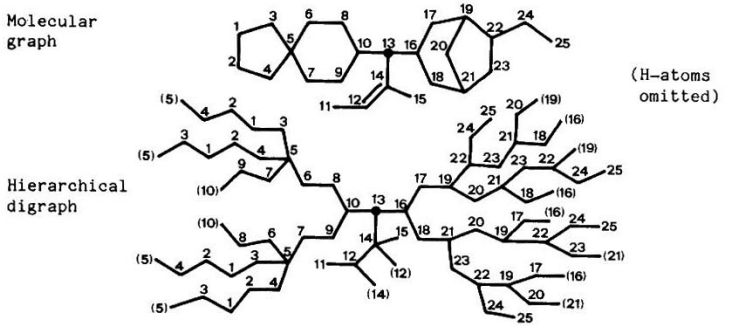


FIGURE 1

After the neighbours of the root have been represented in this way, the same procedure is carried out with the root's neighbours and the root's neighbour's neighbours excluding the root itself. Then proceed to the next neighbours, etc. The procedure stops when no neighbour, other than the original one, exists or when the latest added node represents an atom, which has already been represented between this latest node and the root (root inclusive). Those nodes are also called duplicate atoms and no further edges are added to them (see Fig. 1, the duplicate atoms are those with the number in parantheses).

Step 2, Part 2: Ordering Rules

The nodes of the whole tree can be ordered. One speaks of rank of priority (short: priority) of a node, or of node x ranks higher than node y.

Before starting the actual comparison of nodes described below, it should be noted that all nodes closer to the root rank higher than those farther away, nodes at the same distance from the root having the same priority (distance = number of edges between two nodes).

The nodes are compared with each other using some characteristics of the atoms they represent:

SR1. Larger atomic number ranks higher than smaller atomic number.

SR2. Larger atomic weight ranks higher than smaller atomic weight.

SR3. CIS-node at a double bond ranks higher than TRANS-node *.

SR4 a. R or S descriptor ranks higher than r or s descriptor which ranks higher than 0 descriptor.
b. Two equal R or S descriptors rank higher than

---

* A CIS-node represents that ligand atom at the far end of a double bond (or an odd number of cumulated double bonds) which lies in CIS-position (___/) to the atom represented by the node in the chain to the root. TRANS-node is the other one.

two non-equal R or S descriptors.
The following characteristics determine hierarchically the order of comparison:

i. Higher rank of first descriptor in the compared pair.
ii. Higher rank of second descriptor in the compared pair.
iii. Lower rank of the least common ancestor in the graph (for the definition of 'least common ancestor', see Section 2.3).

c. r descriptor ranks higher than s descriptor.

SR5. R descriptor ranks higher than S descriptor.

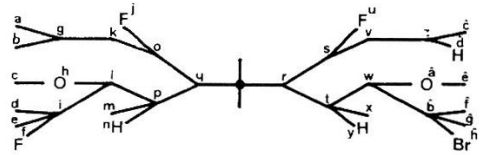
SR1 - SR5 are called Sequence Rules. I have introduced the subrule SR4biii (to remove a hole, see Section 4.2) and made rule 5b in [2] to SR4c to avoid an exception in the hierarchy of the Sequence Rules.

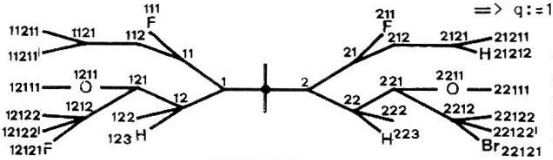
The sequence of comparison is determined by the following rules:

1. One Sequence Rule or subrule is applied to all pairs of nodes of the tree to be compared before the next one is applied.
2. Only nodes of equal priority are compared.
3. Taking the numbers of edges on the path between two nodes as distance, nodes closer to each other are compared before nodes farther away from each other (for the definition of 'path', see Section 2.3).
4. Pairs of nodes with higher priority are compared before nodes with lower priority. Several pairs with equal priority and distance to each other are compared simultaneously.

It may happen, that of 2 or more nodes which until now have the same priority, one has less successors than the other (or even none at all; for the definition of 'successor', see Section 2.3). In this case imaginary nodes are added so that the named nodes have the same number of successors; the imaginary nodes rank lowest in all respects.

The comparison of nodes affects the ranks of priority in the following way: The nodes which have been compared change priority according to the Sequence Rule applied. Then all mutually corresponding neighbouring nodes of the two

1)	r compared to q	=	
2a)	s compared to t	=	
	o compared to p	=	
b)	s,t compared to o,p	=	
3a)	u compared to v	u>v	u:=??1, v:=??2
	w compared to x compared to y	w=x>y	y:=??3
	j compared to k	j>k	j:=??1, k:=??2
	l compared to m compared to n	l=m>n	n:=??3
b)	u,v,Ø compared to w,x,y	u,v,Ø>w,x,y	u:=?11, v:=?12, w:=?2?, x:=?2?, ... ==> s:=?1, t:=?2
	j,k,Ø compared to l,m,n	j,k,Ø>l,m,n	j:=?11, k:=?12, l:=?2?, m:=?2?, ... ==> o:=?1, p:=?2
c)	u,v,Ø,w,x,y compared to j,k,Ø,l,m,n	=	
4a)	â compared to b̂	â>b̂	â:=?2?1, b̂:=?2?2
	h compared to i	h>i	h:=?2?1, i:=?2?2
	- lone successors -		g:=?121, z:=?121
b)	â,b̂ compared to Ø,Ø compared to Ø,Ø	â,b̂>Ø,Ø=Ø,Ø	â:=?211, b̂:=?212 ==> w:=?21, x:=?22, y:=?23
	h,i compared to Ø,Ø compared to Ø,Ø	h,i>Ø,Ø>Ø,Ø	h:=?211, i:=?212 ==> l:=?21, m:=?22, n:=?23
c)	-(already ranked)-		
d)	z,â,b̂ compared to g,h,i	=	
5a)	ĉ compared to d̂	ĉ>d̂	ĉ:=?1211, d̂:=?1212
	f̂ compared to ĝ compared to ĥ	ĥ>f̂=ĝ	ĥ:=?2121, f̂:=?2122, ĝ:=?2122'
	a compared to b	=	a:=?1211, b:=?1211'
	d compared to e compared to f	f>d=e	f:=?2121, d:=?2122, e:=?2122'
	- lone successors -		c:=?2111, ê=?2111
b)	-(already ranked)-		
c)	-(already ranked)-		
d)	-(already ranked)-		
e)	ĉ,d̂,ê,ĥ,f̂,ĝ compared to a,b,c,f,d,e	a,b,c, .. > ĉ,d̂,ê, ...	a:=11211, b:=11211', ..., ĉ:=21211, ... ==> g:=1121, h:= ..., z:=2121, ... ==> j:=111, k:= ..., u:=211, ... ==> o:=11, p:=12, s:=21, t:=22 ==> q:=1, r:=2



The numbers indicate priority: smaller number ranks higher than larger number

FIGURE 2

compared nodes change priority correspondingly, if they were equal in priority. Then the neighbour's neighbours change priority in the same way, etc., until the whole tree has been altered (inducing changes of priority in all directions). Note that a reversal in the order between two nodes induced by the priority can never occur; there are only refinements to the order in the tree.

I have tried to demonstrate the procedure of comparing two ligands in Fig. 2. We start with the hierarchical digraph at the top (without imaginary nodes; these are referred to by Ø in the text of the figure). On the left side, I have noted the steps and substeps of the comparison, followed by its result. On the right side, the rank of priority so far established is indicated by digit strings. The digraph at the bottom of the figure shows the final result. The procedure is done without implicit H-atoms; this does not affect the rank of priority anyway.

SR4 is special, in that the descriptors used there have first to be derived (they are called subsidiary descriptors and do not generally correspond to the final descriptors). This is done by taking the node in question as the new root, changing the directions of the edges correspondingly and for the rest applying the procedure described in this section. Note that this may result in a recursive nesting (Theorem 2 in Section 4.1 shows that this nesting is never endless).

Step 3: Assignment of Descriptors

With potential stereocenters, project the ligand positions, following the order of priority in a natural way onto a line shaped as a screwthread (screwthread line); the symbols R and r are to be used if a right oriented screw is needed, S and s being used for the others. When one but only one pair of ligands can be distinguished by SR5 and not by SR1 - SR4, then r and s are to be used, otherwise R and S. If two ligands are equal the descriptor is 0 (except in the case explained below).

The even-number-of-double bonds axes are treated in the same way but with the additional rule that the two ligands at one end of the axis (it does not matter which end) are always followed first. The descriptor is attached to the middle atom in the chain.

The middle bond of the other axes receive the descriptor Z, if those ligands which are first in priority on its respective side of the axis are in CIS-position to each other (___/); otherwise the descriptor is E.

Symmetry

There are potential stereoelements with several equal ligands where, nevertheless, a projection of ligand positions on a screwthread line is unambiguous and rational. The most common cases are molecules with C2-, C3-, or D2-Symmetry (symmetry groups defined by A. Schoenflies [6]). There is a special rule for these cases: One of the ligands is arbitrarily declared to be of higher rank which then induces priority as described above (the priority differences determined otherwise have to be respected). If the same descriptor results independently of the arbitrary initial choice of ligand, it can be attached to the stereoelement in question.

2.3. Mathematical Notations

For mathematical reasoning it is nowadays indispensable to define and name the sets and functions used concisely.

def A graph homomorphism is a mapping $h:(X,E) \longrightarrow (X',E')$ such that: $(x,y) \in E \Leftrightarrow (h(x),h(y)) \in E'$.

def A graph isomorphism is an invertible graph homomorphism.

def A graph automorphism is an isomorphism of a graph on itself.

def Two graphs are isomorphic if there exists a graph isomorphism from one to the other.

Note, that by this definition a graph homomorphism - or isomorphism or automorphism - need not preserve the chromatism, and that an isomorphism is bijective.

def A sequence of edges e_1, e_2, ... , e_k such that $e_i=(n_{i-1}, n_i)$ or $e_i=(n_i, n_{i-1})$, and $n_0=x$ and $n_k=y$, all n_i different, is called a path between x and y.

def A graph is called connected, if there exists at least 1 path between any pair of different nodes.

def A tree is an acyclic connected graph, i.e. between any pair of different nodes there is one and only one path.

def A rooted tree is a tree with one selected node called the root, and all edges directed away form it.

def GM is the set of chromatic graphs representing molecules with two numbers attached to each node. A GM-graph $G^\wedge$ (i.e. $G^\wedge \in GM$) is a quadruple $G^\wedge = (X^\wedge, E^\wedge, t^\wedge, l^\wedge)$, $X^\wedge$ being the set of nodes, $E^\wedge$ the the set of edges, $t^\wedge: X^\wedge \longrightarrow I$ representing the CIP-relevant information (see below), and $l^\wedge: X^\wedge \longrightarrow \mathbb{Z}^+$ standing for an arbitrarily chosen but fixed labeling (numbering of the nodes). The CIP-relevant information assigned to the nodes by the function $t^\wedge$ is: the atomic number (SR1), the atomic mass (SR2), CIS, TRANS, and INDIFFERENT (INDIFFERENT = no CIS-TRANS-distinction possible) (SR3), and R, S, r, s, and O (O=no chirality) (SR4 and SR5).

def The values of l are called labels.

Note that GM-graphs are symmetric, i.e. $(x,y) \in E^\wedge \iff (y,x) \in E^\wedge$, and that $l^\wedge$ is injective.

def G2 is the set of rooted trees (in Section 2.2 called: hierarchical digraphs) derived from GM-graphs by the CIP-System (see Section 2.2). A G2-graph G" is also a quadruple, G" = (X",E",t",l"), where X", E", and t" are: the set of all nodes, the set of all edges, and the function representing the CIP-relevant information. $l": X" \longrightarrow \mathbb{Z}^+$ assigns the same labels as in the original GM-graph.

Note that several nodes may receive the same label.

def G1 is the set of graphs which arise from G2-graphs when the function l" is disregarded, thus just picking out the triple (X",E",t"), i.e. $G"=(X",E",t",l") \in G2$ ==> $G':=(X",E",t") \in G1$.

Note that G1- and G2-graphs are non-symmetric; as rooted trees, they are even antisymmetric. The function l is not necessarily injective. As described in Section 2.2, the ordering and the Sequence Rules are applied to G1-graphs, i.e. labels are disregarded.

The relation between the nodes of the graphs in the three sets GM, G1, and G2 will be expressed as follows:

def The nodes x and y mutually represent each other if x of $G^\wedge$ (or of G") goes over to y of G" (or of G') when G" (or G') is derived from $G^\wedge$ (or G").

As G1- and G2-graphs are rooted trees, there are many expressions with obvious meanings. Referring to a graph $G"=(X",E",t",l") \in G2$ or $G'=(X",E",t") \in G1$ (they are all rooted trees), we define:

def y successor of x: <==> $(x,y) \in E"$

def x predecessor of y: <==> $(x,y) \in E"$

Except for the root, every node has exactly one predecessor.

def A successor chain of x is a sequence $x=x_1,x_2, \ldots ,x_n$ of nodes, where $(x_i, x_{i+1}) \in E"$.

def A predecessor chain of x is a sequence $x_1,x_2, \ldots ,x_n=x$ of nodes, where $(x_i, x_{i+1}) \in E"$.

def y descendant of x:
<==> y is contained in a successor chain of x.

def x ancestor of y:
<==> x is contained in a predecessor chain of y

def A branch is a subgraph, containing all descendants of one single node and this node as well. This single node is called head of the branch.

Note that, taking the head as root, a branch is again a rooted tree.

def The distance between x and y is the number of edges of the path between x and y.

def A generation is a set of all nodes which have the same distance from the root (in [2]: sphere).

def The least common ancestor of x and y is the ancestor of both x and y, which has the smallest distance to x of all ancestors of x and y.

def The family of x and y (x neither ancestor nor descendant of y) is the union of the two branches having as heads the ancestors of x and of y which are successors of the least common ancestor of x and y, respectively.

def A leaf is a node which has no successor.

3. CONSTITUTIONAL INCOMPLETENESS

3.1. CIP's Failing

It seems intuitively clear, that the CIP-System will grasp all constitutional differences. Nevertheless, a closer investigation shows that this is not true and that a considerable change of the rules is necessary to guarantee constitutional completeness.

I would like to give 2 examples. Suppose there is an atom type A which allows ligancies of 1 and 3 (all single bonds). Consider then the following molecule and its hierarchical digraph:

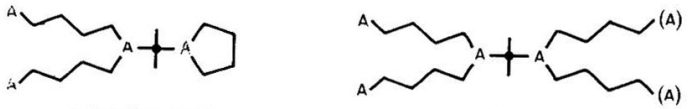


FIGURE 3

In the hierarchical digraph, the left and the right ligands cannot be distinguished.

But even if we impose some restriction on the ligancy of atoms which forbid such molecules, we can have constitutional differences not perceived by CIP:

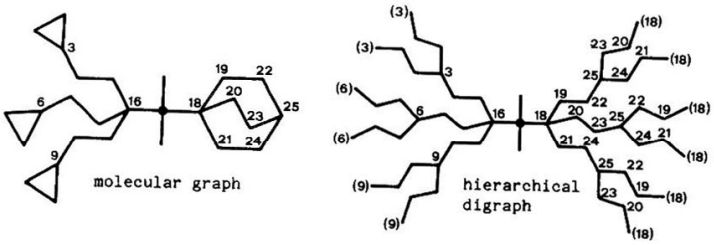


FIGURE 4

From the simplicity of this second example we see, that CIP is fundamentally incomplete.

A possibility to save the system would be to introduce some new 'Sequence Rule', applicable only to the leaves (the outermost nodes) of the hierarchical digraph, e.g. one could consider the distance from the root of the duplicate atom ('SR1b: A duplicate atom with its predecessor node having the same label closer to the root ranks higher than a duplicate atom with its predecessor node having the same label farther from the root, which ranks higher than any non-duplicate-atom-node'). With such a supplement - in the following called the linking supplement - CIP indeed grasps all constitutional differences. This is proved in the following for tetrahedral stereocenters. The extension to

the whole domain, as defined in Section 2.2, is almost trivial.

3.2. G1-Graphs

We define a relation between the nodes and a relation between the bonds on G1-Graph G'=(X",E",t"):

def A node x or a branch B' is indistinguishable from a node y or a branch C', respectively, if there is a graph automorphism f on X" such that:
a) t"(n) = t"(f(n)) for all n X".
b) f(x) = y or n∈B' <==> f(n)∈C', respectively.

Proposition 1:
Indistinguishable is an equivalence relation.
Proof: ... (straightforward) ... .

Proposition 2:
x head of branch B', y head of branch C'.
B' indistinguishable from C' implies x indistinguishable from y.
Proof: ... (trivial) ... .

We can now show that two nodes x and y with the same rank of priority are indistinguishable. In the light of Proposition 2 we prove:

Lemma 1:
x head of branch B', y head of branch C'.
Branch B' is indistinguishable from C', if x has the same rank of priority as y.

Proof:
a) Ancestors of x and ancestors of y belonging to the same generation, up to the least common ancestor, must have pairwise equal priorities, since priorities are induced (see Section 2.2).

b) Let u and v be the heads of the two branches U' and V', respectively, of the family of x and y. By (a), they have equal priority. For every descendant of u, there must

always be a descendant of v which ranks equal: For all successors of u and v there must always be one of u and one of v ranking equal, since any difference in priority would induce a difference in priority of u and v. The same argument can be used for the next generation, i.e. the successors of the successors and through an induction argument for the entire branches U' and V'.

c) These pairs of equal priority can be used to define an automorphism f on G': The nodes of U' will be mapped on the node of V' with the same priority (if there are several, we choose those pairs which have ancestors that represent each other, if there are still several, we choose arbitrarily) and vice versa; neither the nodes of U' nor of V' will be mapped on themselves.

d) Only nodes with the same image under t" can have equal priority. Thus, f satisfies criterion (a) in the definition of indistinguishable and the Lemma is established.

###

def A mapping constructed as in the proof of Lemma 1 is called a Cindi-mapping (Constructed indistinguishable).

3.3. Cindi-Mappings on G2-Graphs

G2-graphs can be considered as special cases of G1-graphs. The difference lies in the function l, included in G2-graphs but not in G1-graphs.

There is no difficulty in constructing Cindi-Mappings on an arbitrary G2-graph G"=(X",E",t",l"). Here, we are mostly interested in how Cindi-Mappings treat the labels, specially those which indicate the cyclic structure of the original molecular graph. But first, two simple general propositions.

Proposition 3:
Cindi-Mappings are involutions.

Proof:
This follows directly from their construction principles.

###

Proposition 4:
f Cindi-Mapping, u,v,x,y nodes.
f(x)=y, u ancestor of x, v ancestor of y, u and v in the same generation implies f(u) = v .

Proof:
This follows from the fact that we have trees and that Cindi-Mappings are graph isomorphisms.

###

def Branching junctions are non-leaf nodes which have the same label as one of their descendants.

def Junction leaves are leaves which have the same label as one of their ancestors.

A moment's thought will convince the reader, that branching junctions and junction leaves are exactly those nodes which represent the atoms which are nearest to the potential stereocenter of the cyclic substructure being determined; these are the atoms where one - as Helmchen and Prelog state [2] - has 'to break the remaining n-1 bonds'.

The next proposition and the following lemma are those statements which need the linking supplement (mentioned in Section 3.1) to be valid. The first example of Section 3.1 (Fig. 3) shows its necessity for Proposition 5.

Proposition 5:
A Cindi-Mapping maps junction leaves on junction leaves.

Proof:
Cindi-Mappings preserve priority. With the linking supplement no junction leaf will have the same priority as a leaf which is not a junction leaf.

###

The example shown in Fig. 4 demonstrates, that Proposition 5 is not strong enough. Fortunately, we can obtain more out of the linking supplement.

Lemma 2:
f Cindi-Mapping, x branching junction, y junction leaf and descendant of x, and $l''(x)=l''(y)$.
Then, $l''(f(x)) = l''(f(y))$.

Proof:
For any junction leaf n, there exists a node m, ancestor of n, such that $l''(m)=l''(n)$. Say, m is in the p-th generation. The linking supplement and the fact that Cindi-Mappings preserve priority enforce not only that $n':=f(n)$ is a junction leaf, but also that n and n' 'point back' at nodes which lie on the same generation; in other words, if m' is the ancestor of n' in the p-th generation, then $l''(m')=l''(n')$. By Proposition 4, $f(m)=m'$.

Each branching junction x has, by definition, at least 1 junction leaf z as descendant with $l''(x)=l''(z)$ (actually at least 2). The first part of this proof yields thus, that branching junctions are mapped on branching junctions, and that $l''(f(x))=l''(f(z))$. But this last statement is valid for all junction leaves which are descendants of x and for which $l''(y)=l''(x)$.

###

We could state as a trivial corollary, that in such a case all junction leaves are mapped 'consistently'. However we do not need this fact.

Please note, that f does not map the nodes of the whole tree 'consistently' (Fig. 5 provides an example); even some junction nodes with the same label can obtain images with different labels.

3.4. GC-Graphs

To enhance the readability in this and the next section, sets and elements are denoted with systematic superscripts. Thereby the nodes shall represent each other in the self-evident way (for the definition of 'represent', see below).

Our aim is to construct an induction of a graph automorphism on GM-graphs by a Cindi-Mapping. Properly defined, such a graph automorphism could indicate the constitutionally equivalent atoms. Unfortunately, Cindi-Mappings do not preserve the function l (e.g. Fig. 5), obviously a prerequisite for such an induction.

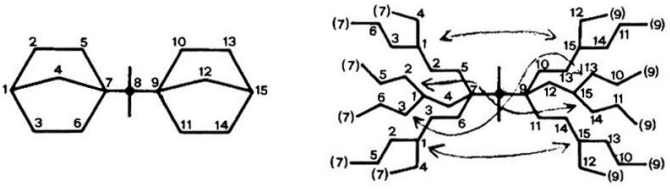


FIGURE 5

That is the reason why we make a detour via GC-graphs. First, two definitions:

def G=(X,E,t,l), H=(Y,F,u,m). G-H:=(X-Y,E-[Y],s,k) where [Y] is the set of all edges containing one or two nodes of Y, and s and k are the restriction of t or l to X-Y, respectively.

def G=(X,E,t,l), H=(Y,F,u,m). $G \subseteq H$ if $X \subseteq Y$, $E \subseteq F$, t is the restriction of u to Y, and l is the restriction of m to Y.

Now, to the definition of GC-graphs:

def GC is the set of all graphs G=(X,E,t,l), such that the restriction of l to all non-leaf nodes is injective and there exists G"=(X",E",i",j")$\in$G2, where G"-G is a union of discrete branches of G", where each branch has at least two nodes.

With other words, you get a GC-graph by taking a G2-graph and cutting off branches (but never just one leaf) until no label is left twice in the tree, except at the leaves.

def G=(X,E,t,l)$\in$GC represents G"=(X",E",t",l")$\in$G2 if $G \subseteq G"$, l(x)=l"(x) for all x$\in$X, and l(X)=l(X").

def G=(X,E,t,l)$\in$GC represents $G\hat{}=(X\hat{},E\hat{},i\hat{},j\hat{})\in$GM, if G represents the hierarchical digraph of $G\hat{}$.

With other words, each atom is represented at least once.

Lemma 3:
G$\in$GC represents G"$\in$G2 and $G\hat{}\in$GM. Then $G\hat{}$ and G" can be reconstructed from G.

Proof:
All labels appear in G. Thus we can sketch G^ (from which G" was derived) in the sense that all atoms are there.

What about the edges ? Because G represents G" and G^, the only edges which one can imagine to be missing are those which were 'cut through'. So suppose, edge e^ between x^ and y^ (y successor of x) is missing because the branch starting with y" was cut off. Then y^ is represented elsewhere in G, and there we find also e (e^ and e in the self-evident relationsship) and once again x, this time as a leaf (by the construction principles of hierarchical digraphs and the prohibition to cut off a single leaf).

To construct G" from G^ is nothing more than constructing the hierarchical digraph.

###

def A set of branches $B"_i \subseteq G" \in \underline{G2}$ (i=1,2,...) is called <u>interconnected</u> if their heads have the same predecessor, the intersection of the sets of their labels is not empty, and every branch has at least 2 nodes.

<u>Proposition 6</u>:
Branches $B"_1$, $B"_2 \subseteq G"$ = (X",E",t",l") interconnected
==> $l"(B"_1)=l"(B"_2)$.

Proof:
In short, $l"(B"_1)$ (and $l"(B_2)$) is just the set of all labels of the nodes in $G \in \underline{GM}$ which form the cyclic system corresponding to $B"_1$ and $B"_2$.
Let G" be the hierarchical digraph of G, the nodes $h"_i$ being the heads of $B"_i \subseteq G"$, and l(b") their common label. From the way hierarchical digraphs are constructed, we conclude:
l(n") is in $B"_1$ <==> there is a path in G^ between the nodes represented by n" and b" not passing through any node represented by an ancestor of $h"_1$.
Since by the definition of interconnected, $h"_1$ and $h"_2$ have the same ancestors, the proposition is established.

###

Lemma 4
$x'' \in X''$, $G''=(X'',E'',t'',l'') \in \underline{G2}$.
There is $G=(X,E,t,l) \in \underline{GC}$ which represents G" with $x'' \in X''$.

Proof:
Start at the root and look at those branches which start with the successors of the root. If there are interconnected ones, cut off all but one of every bunch of interconnected branches (the one which contains x", if any; for the rest: the choice is arbitrary). Then move one generation away from the root and look at the branches starting at the root's successors' successors. The interconnected ones are cut off leaving one of every bunch. Then move one generation away from the root and repeat until no more interconnected branches are left.

In this way we clearly obtain a GC-graph $G \subseteq G''$. No label is lost either, because of Proposition 6. Thus, G represents G".

###

Corollary 1:
$x \in X$, $G''=(X'',E'',t'',l'') \in \underline{G2}$, $G=(X,E,t,l) \in \underline{GC}$.
G represents G". $y'' \in X''$, f Cindi-Mapping on G", $f(x)=y''$.
Then there exists $H \in \underline{GC}$ which represents G", contains y", and is isomorphic to G.

Proof:
With a slight restriction in the choice of the branch not to be cut off, the argument of Lemma 4 can be taken over: With interconnected branches, we choose the one which is isomorphic to the corresponding branch in G (i.e. the one with the same labels). There will be one, since f is an isomorphism.

###

N.B. Every bond is represented once and only once in G (similar argument as in the proof of lemma 4).

def $G^\wedge=(X^\wedge,E^\wedge,t^\wedge,l^\wedge) \in \underline{GM}$, $G''=(X'',E'',t'',l'') \in \underline{G2}$,
G" represents $G^\wedge$.
The projection $G''>>G^\wedge$ shall be the mapping $p: X'' \longrightarrow X^\wedge: x'' \longmapsto x^\wedge$, where x" and $x^\wedge$ mutually represent each other.

def $G^\wedge=(X^\wedge,E^\wedge,t^\wedge,l^\wedge)\in \underline{GM}$, $G=(X,E,t,l)\in \underline{GC}$, G represents $G^\wedge$. The projection $\underline{G>>G^\wedge}$ shall be the mapping $p: X \longrightarrow X^\wedge: x \longmapsto x^\wedge$, where x and $x^\wedge$ mutually represent each other.

def $G=(X,E,t,l)\in \underline{GC}$, $G"=(X",E",t",l")\in \underline{G2}$, $G\subseteq G"$. The projection $\underline{G">>G}$ shall be the mapping $p: X" \longrightarrow X$ for which $p(x)=x$ for $x\in G$ and $l(x")=l(p(x"))$.

The projection $G">>G^\wedge$ is clearly the reverse of constructing a hierarchical digraph.

Lemma 5

$G^\wedge=(X^\wedge,E^\wedge,t^\wedge,l^\wedge)\in \underline{GM}$, $G"=(X",E",t",l")\in \underline{G2}$, $G=(Y,E,u,m)\in \underline{GC}$. G representing $G^\wedge$ and G". p projection $G">>G^\wedge$, q projection $G>>G^\wedge$, r projection $G">>G$. f Cindi-Mapping on G", h restriction of f to the nodes of G. H:=h(G).

a) H is isomorphic to a GC-graph representing $G^\wedge$.
b) The mapping g on $G^\wedge$: $x^\wedge \longmapsto q(h(r(p^{-1}(x^\wedge))))$ is well defined ($x^\wedge\in X^\wedge$).
c) $t^\wedge(x^\wedge)=t^\wedge(g(x^\wedge))$.
d) g is a graph homomorphism.

Proof:

a) h is injective, so H is isomorphic to G.
b) $p^{-1}(x^\wedge)$ is a set of nodes with the same labels. By the definition of GC-graphs, G will contain only one of these nodes except for leaves. Cindi-Mappings map a leaf and its ancestors with the same label on a leaf and an ancestor with the same label (Lemma 2), so q projects $r(p^{-1}(x^\wedge))$ on the same node.
c) Evident by the construction of g.
d) $e^\wedge:=(x^\wedge,y^\wedge)$ edge of $G^\wedge$ ==> there is at least one edge (x,y) in G such that $p(x)=x^\wedge$ and $p(y)=y^\wedge$. f is an isomorphism, so there is an edge (f(x),f(y)) in H. q also preserves edges.
On the other hand, $e^\wedge:=(q(x),q(y))$ edge of $G^\wedge$ ==> there is an edge (x,y) in G by the rules for the construction of hierarchical digraphs (q is the restriction of the 'inverse operation'); h and r are mappings between trees which preserve edges, so any 'new edge' would result in a cycle; d":=(v",w") in G" ==> there is a d":=(v",w") in $G^\wedge$ by the rules for the construction of hierarchical digraphs.

###

3.5. Constitutionally Equivalent Atoms

It is time to define the term 'constitutionally equivalent'.

def Nodes $x^\wedge$ and $y^\wedge$ of a GM-graph $G^\wedge$ are called constitutionally equivalent, if a graph automorphism $f^\wedge$ on $G^\wedge$ exists, such that $f^\wedge(x^\wedge)=y^\wedge$.

The reader may convince himself that this definition corresponds to what chemists understand by 'constitutionally equivalent'.

Theorem 1:
$G"=(X",E",t",l") \in G2$, $G^\wedge=(X^\wedge,E^\wedge,t^\wedge,l^\wedge) \in GM$, $p: G" \longrightarrow G^\wedge$ the projection $G">>G^\wedge$.
$x" \in X"$ has the same CIP-priority as $y" \in X"$ <==>
there is an automorphism preserving $t^\wedge$ which maps $p(x")$ on $p(y")$.

Proof:
Having the automorphism, the equality of priority is trivial.

By Lemma 1, there is a Cindi-Mapping f on G" with $f(x")=y"$. By Lemma 4 and its corollary, we have a GC-graph G representing $G^\wedge$ containing x", and a GC-graph H representing $G^\wedge$ containing y", G and H isomorphic. Let q be the projection $H>>G^\wedge$, r the projection $G">>G$, and h the restriction of f to G. By Lemma 5, h(G) is isomorphic to G, hence also to H. Let i denote this isomorphism. Then by a similar argument as in the proof of Lemma 5, $g:G^\wedge \longrightarrow G^\wedge : x^\wedge \longmapsto q(i(h(r(p^{-1}(x^\wedge)))))$ is a graph homomorphism with $t^\wedge(x^\wedge)=t^\wedge(g(x^\wedge))$.

Surjectivity remains to be proved (surjective mappings on finite sets are bijective). f is injective, thus h and i(h) are injective. Since the numbers of nodes of H equals the one of G, i(h) must be surjective. q is trivially surjective (H represents $G^\wedge$), so g is surjective.

###

Corollary 2:
G"=(X",E",t",l") ∈ G2, x",y" ∈ X".
If x" and y" have the same priority, then for any node m" ∈ X" for which l"(m")=l"(x"), there is a node n" with l"(n")=l"(y"), which has the same priority as m".

Proof:
Let G^ ∈ GM represent G" and p shall be the projection G">>G^. Since l"(m")=l"(x") ==> p(m")=p(x") and similarly p(n")=p(y"), by going forward and backward in the statement of Theorem 1 the corollary is yielded.

###

4. STEREOCHEMICAL COMPLETENESS

4.1. Interdependence

A mathematician reading the ordering rules in Section 2.2, asks himself immediately, if the determination of subsidiary descriptors can always be accomplished, since - in cyclic structures - these descriptors may depend on each other. Fortunately, it is not difficult to prove that with the construction of a hierarchical digraph (introduced in the last revision of the CIP-System [2]) termination of the procedure can be guaranteed. I consider this point to be the main success of the revision. N.B. Interdependence of subsidiary stereoelements is the main reason for the reconstruction problem (see [4]).

For the proof, I restricted myself again to the tetrahedral stereocenters as the extension to the whole domain is easy.

Theorem 2:
It is always possible to accomplish the determination of subsidiary stereocenters.

Proof:
The Sequence Rule SR4 is applied only if two or more branches are considered equal by the Sequence Rules SR1 to

SR3. Then - as a first step applying SR4 - one branch is dealt with in order to determine the subsidiary centers. The fact that at least two branches (of the original center) were constitutionally equal, implies that at every subsidiary center, the branch which leads back to the original center is constitutionally different from all the others.

If it happens that other subsidiary centers have to be determined in order to distinguish the branches at the first subsidiary center, the same argument is valid. Generalising this argument, we conclude that in the worst case the jump is from one subsidiary center to the next subsidiary center but always away from the original center. As the graph is finite, this jumping will have an end. The outermost centers can be determined without SR4 and SR5, after which the outermost but one can be determined, and so on, all the way in to the original center.

###

4.2. R,S Versus r,s

It is intended that r and s be used and only used for the so-called pseudochiral stereoelements, i.e. those elements with tetrahedral ligand arrangement which are invariant under reflection (their arrangement of ligands cannot be distinguished from the arrangement of ligands of the reflected structure). This happens if 2 and only 2 ligands at a stereocenter (or at one end of a corresponding stereoaxis) are the mirror images of each other but not equal (these ligands are called enantiotopic ligands, see also Section 1.2).

In the light of the assignement-of-descriptor-rules in Section 2.2, this implies two requirements of the Sequence Rules:

a) SR1 to SR4 must distinguish between all different ligands which are not enantiotopic to each other.

b) SR5 must distinguish between enantiotopic ligands.

A moment's reflection will convince the reader that (a) -

proved below - implies (b): Enantiotopic ligands must always contain at least one stereochemical feature, which directly or indirectly is based on different, non-enantiotopic parts; if by (a) all stereochemical information is represented by (subsidiary) CIP-Descriptors, then a look at SR5 reveals that obviously (b) must be valid too.

As usual, the proof is given for tetrahedral stereocenters only, the necessary extension being easy.

Theorem 3:
By the Sequence Rules SR1 to SR4, all different, non-enantiotopic ligands can be distinguished but no others.

Proof:

1. It is not possible for enantiotopic ligands to be distinguished by the SR1 to SR4c, since they are all completely symmetric with respect to the descriptors R and S.

2. Constitutional differences are recognised by the CIP-System according Theorem 1 (Section 3.5).

3. What remains are different CIP-Descriptors. It can be seen that potentially we might have to compare each descriptor with every other, whilst taking account of different priorities between the subsidiary centers (according to the Ordering Rules, see Section 2.2). What difference can there still be between 2 pairs of subsidiary centers, i.e. where the first descriptors have the same priority as each other and likewise the second descriptors ? Such a difference must lie in the connecting of the 2 centers in the graph. But the ligands being compared are constitutionally equal, thus the only possible difference lies in the length of the common path leading from the original center to the 2 subsidiary centers being compared. Thus, the priority (or even generation number) of the least common ancestor (definitions see Section 2.3) will give a measure for this last variation.

###

The proof shows that the Sequence Rule SR4biii which I have added is necessary.

5. SUMMARY

The rules of Cahn, Ingold, and Prelog for the specification of molecular configuration (CIP-System) have been investigated:

1. Rules and domain of application have been defined in the mathematical sense.

2. The CIP-System as stated in [1] and [2] cannot distinguish between certain constitutionally different ligands. A supplemental rule to remove this defficiency is suggested.

3. There is a minor omission in the Sequence Rule 4. Supplying a SR4biii is enough to fill the gap.

4. Three formal mathematical proofs are given. They state that:
   - the rules, as supplemented by me, can distinguish between any different ligands.
   - the determination of descriptors is a finite procedure.
   - with SR4biii, reflecting a molecule has the expected effect on CIP-Descriptors: R to S and S to R while r and s are left fixed.

I owe a special thank-you to A.S. Dreiding and V. Prelog. Without their stimulation this work would never have been done.
This work has been supported by the Swiss National Science Foundation.

REFERENCES
==========

[1] R.S. Cahn, C.K. Ingold & V. Prelog, Angew. Chem. Int. Engl. Ed. 5(1966), 385.

[2] V. Prelog & G. Helmchen. Angew. Chem. Int. Engl. Ed. 21(1982), 567.

[3] R.H. Custer, "An Investigation of the CIP-System, a Mobile Molecular Model, and Computer Implementations", Diss ETH No 7847, Zurich 1985.

[4] R.H. Custer, "An Investigation of the CIP-System and a Computer Implementation", manuscript in preparation.

[5] J.F. Dubois, "Ordered Chromatic Graph and Limited Environment Concept" in A.T. Balaban ed., "Chemical Applications of Graph Theory", Academic Press, London, 1976, p 333.

[6] W. Nowacki, Z. Kristallogr., Kristallgeom., Kristallphys., Kristallchem., 135(1-2), 145.

[7] W. Schubert & I. Ugi, Chimia 33(1979). 183.

[8] G.L. Lemière & F.C. Alderweireldt, J. Org. Chem. 45 (1980), 4175.

[9] I. Ugi, J. Dugundji, R. Kopp & D. Marquarding, "Perspectives in Theoretical Stereochemistry", Chap. 8, Springer-Verlag, Heidelberg 1984.