Starting Material Oriented Retrosynthetic Analysis in the LHASA Program. 2. Mapping the SM and Target Structures

A. Peter Johnson* and Chris Marshall

The Maxwell Institute for Computer Applications in the Molecular Sciences, University of Leeds, Leeds LS2 9JT, U.K.

Received January 16, 1992

This paper describes a method for matching two chemical structures in which the degree of similarity may be small. The combined use of iterative set division and a backtracking algorithm is described. Rapid identification of the largest areas of carbon skeleton common to the starting material and target structures is followed by a more rigorous analysis of the nonidentical areas to complete the atom-to-atom mapping of the structures.

INTRODUCTION

General. This paper is the second in a series entitled Starting Material Oriented Retrosynthetic Analysis in the LHASA Program.[1] The papers describe a strategy incorporated into the LHASA retrosynthetic analysis program which enables the program to select an appropriate starting material (SM) for a target structure and to direct retrosynthetic analysis toward that structure. In order to plan the conversion of a starting material or set of starting materials into a desired target compound, it is necessary first to establish which part of the target structure is to be furnished by the selected SM. This paper describes a method for mapping two chemical structures in which the degree of similarity may be small.

Once a SM has been selected, it is often a trivial matter for a trained organic chemist to see where a SM fits best onto a required target structure (Figure 1). More complex situations do arise in which the best fit is not obvious, particularly in cases where a carbon skeleton rearrangement occurs or reactions take place at many sites (Figure 2). In some cases there may be several approximately equal ways of placing the SM in the target with no overriding reason to describe one as better than another.

Often the starting point for the process of finding the best match is the identification of an area of similarity in the carbon skeletons of the structures. Closely allied to this is the perception of similarities in the location of functional groups and stereocenters and their relationship to sites where structural alterations are required to convert the SM to the target. At the same time the chemist uses a knowledge of both general reaction types and specific reactions to estimate feasibility in the laboratory. The parallel consideration of graphical mapping information and specific chemical knowledge allows the best chemists to select the most promising mapping between the SM and the target in a single process.

In taking this approach chemists use chemical information at many levels of abstraction. They use detailed knowledge of the likely course of specific reactions with which they are familiar, often resulting from years of research in the areas of interest. They use less detailed knowledge of the regio- and stereochemistry of a large range of "standard" reactions, as well as general observations and heuristic rules, such as the knowledge that it is difficult to effect chemical changes at a carbon atom remote from any functionalizing group. The selection of the largest common carbon skeleton as the first basis for similarity is itself based on the rule that it is generally

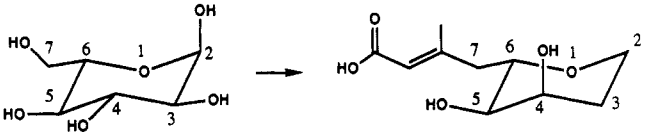

Figure 1. Obvious mapping between target and starting material.

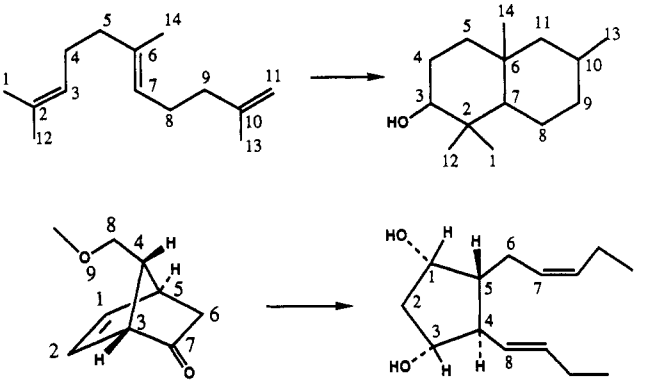

Figure 2. Less obvious mappings between targets and starting materials.

easier to make or break a carbon–heteroatom bond than a carbon–carbon bond.

The diverse but incomplete nature of the information available to the chemist means that although it is usually possible to arrive at a reasonable solution to the mapping problem, it cannot always be guaranteed that the choice is the best one (however "best" might be defined).

Computer Methods. The problem of determining whether two structures are identical has been tackled by workers in the field of chemical structure matching and in the more general field of graph matching. Algorithms have been developed for this problem and for the related problem of substructure searching, i.e., the recognition that the graph of one structure is wholly contained within the graph of another.[3-5,9] Work on the wider problem of finding the best fit between two structures, neither of which is a substructure of the other, has also been described.[6-8] This work has mainly been oriented toward reaction indexing, where the structures of the SM and product are normally more similar than different, because only the sites at which reaction takes place are modified. Usually, the best mapping between them can be assigned unequivocally by pairing off nearly all the SM atoms with those of the product.

In the field of reaction indexing, matching by atom set manipulation[7,8] is generally preferred to using a backtracking

0095-2338/92/1632-0418$03.00/0

method since it consumes much less computer time. Its limitation is that it is only effective for structures having substantial identical fragments. This paper describes the combined use of an atom set manipulation method and a backtracking algorithm to achieve rapid mapping of structures which may be widely dissimilar.

DISCUSSION

For many organic syntheses, the greatest challenge is the construction of the required carbon skeleton. As it is generally more difficult to make or break bonds between carbon atoms than bonds to heteroatoms, chemists usually prefer a starting material with a suitable carbon skeleton over one having a similar heteroatom substitution pattern to the target but a less similar carbon skeleton. In the same way, the method described in this paper begins by attempting to identify similarities between the carbon skeletons of the target and the proposed SM. The position and nature of each functional group and heteroatom is considered in a second step designed to order the matchings.

Chemical structures can be regarded as simple graphs in which the atoms are the nodes of the graph and the bonds are the arcs joining the nodes. The chemical graph is said to be *colored* because atoms may be of different types, indicated by their atomic number, charge, etc., and bonds may be of different orders. To map the SM onto the target, LHASA uses a combination of a rapid iterative set division algorithm with a slower, but more mismatch-tolerant, backtracking algorithm. The combination produces a self-balancing system which achieves rapid identification of similarity in closely related compounds but is still capable of detecting similarity between less closely related structures.

This procedure begins with an exploration of the similarities of the carbon skeletons of the SM and target. Graphs corresponding to these skeletons are generated. The graphs may be disconnected since only the bonds between carbon atoms are included (heteroatoms are included in the mapping at a later stage). In the first step, a set division algorithm is used to find the atoms central to the largest identical regions (LIR) in the carbon subgraphs. Starting with a single set of all carbon atoms, the algorithm uses their connectivities iteratively to divide the set into subsets until the central atoms are found. These are called the ROOT ATOM pairings.

Thereafter, the sets of matched atoms from preceding, submaximal iterations of the set division process are used together with connectivity information to find the further atom-to-atom correspondences within identical regions. Differences between the symmetry of the carbon subgraph and the full structure may produce several matchings from one root atom pair. Each unique matching is called a PARTMAP.

This process is then repeated starting from submaximal atom pairings so that an ensemble of partmaps is built up.

The spatial relationship between the partmaps is investigated, and favorably disposed partmaps are combined to give larger partmaps. Sets of symmetrically equivalent partmaps are reduced to one representative. The partmaps are sorted in order of decreasing size.

Each partmap is expanded into a full match by mapping unmapped areas as far as possible using a backtracking algorithm. At this stage heteroatoms are included in the mapping process. Each partmap may give rise to several full maps, each of which is rated according to a heuristic estimate of synthetic proximity (HESP).[2] After all partmaps have been processed the full maps are sorted by their HESP values and displayed to the chemist.

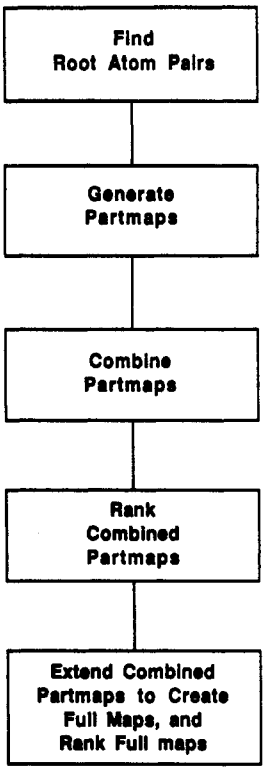

Figure 3. Stages in the mapping process.

The procedures are described in more detail below and are further explained by a worked example at the end of this paper. The overall sequence of steps is outlined in Figure 3.

(a) Finding Root Atom Pairs. The full connection tables for the SM and target structures are reduced to their carbon subgraph by removing all non-carbon atoms and setting all bond orders to unity.

In a process similar that used by other workers,[6-9] atoms are collected into sets according to their connectivity. Set division is achieved by iteratively considering a progressively larger shell of environment for each atom, first the α atoms, then the β atoms, and so on, until the atom is found to be unique or the edge of the structure is reached. The set of atoms which have contributed to the label at each iteration is also recorded. Each iteration of the loop is termed a LEVEL. The lower levels contain atom sets which correspond to small areas of identity, the higher levels to atoms at the center of larger areas of identity.

Initially, each atom is labeled according to the number of carbon atoms attached to it. Identically labeled atoms are collected into sets for the SM and target. This produces a maximum of 5 label sets for each structure, containing carbons bearing 0–4 adjacent carbon atoms. The set of atoms which has contributed to the label for each atom is the atom itself and its α atoms. Copies of the sets are stored for use during the regrowing phase of the algorithm.

The label sets for the SM and target structures are compared, and any sets which contain both SM and target atoms are reprocessed. Subsequent labels for each atom are generated from the labels of the adjacent atoms by the formula

$$\text{NEW LABEL} = 5(\text{OLD LABEL}) + \sum(\text{OLD ADJACENT ATOM LABEL})^2$$

Multiplication of the old label and squaring of adjacent atom labels causes the function to diverge rapidly and provides a canonical label for each atom. The rapid divergence of the function means that the probability of two atoms coincidentally having different environments but the same label is reduced.

The new labels describe the β environment of each atom, since the old labels described the α environment and the labels of the adjacent atoms were used to construct the new labels. The set of atoms which has contributed to the label at this level is the union of the old set and the atoms α to the old set. To avoid the chance of coincidental label identity, atoms are only placed in the same label set in the current iteration if they were common to a label set in the previous iteration.

Again, the sets of identically labeled atoms are collected, and those which include atoms from both the SM and the target are reprocessed. Atoms for which the set of adjacent atoms has not changed since the previous iteration are removed from further consideration. The complete carbon subgraph environment of these atoms has been explored.

The process of set division is continued until one of the following criteria is met:

1. The set of adjacent atoms has not changed for any atom in either structure. This occurs when the carbon skeletons of the target and the SM are isomorphic—the whole skeleton is the LIR.

2. There is only one set left containing atoms from both the target and the SM, and it contains only one atom from one of them. There is a unique LIR in one of the structures (although it may match several positions in the other structure).

3. There is no set left with members in both structures. There is more than one LIR for which correspondences are contained in the set of labels stored after the previous iteration. The algorithm recovers this set.

Each combination of an atom from the target with an identically labeled atom from the SM constitutes a root atom pairing which can be used to generate a partmap. Those in the last sets correspond to the LIRs.

(b) Generating Partmaps. The procedure used in LHASA from this point has not been described before. A breadth-first search using the connectivity of the structure, the symmetry of the whole molecule, and the sets of equivalent atoms identified during the previous section is used to construct the partmaps. Each partmap contains the atom-to-atom correspondences between the SM and target structure for a particular root atom pairing. During this process as much information as possible about the symmetrical equivalence of the atoms is used to eliminate symmetrically redundant solutions.

Determination of atom-to-atom correspondences is begun by selecting a root atom from the SM as the first atom in a partmap and assigning the target atom from the root atom pair to it. The atoms adjacent to each of the root atoms are then examined. Since the root atoms are equivalent there must be the same number of carbons adjacent to each of them. If there is only one, the adjacent carbon atom in the SM is added to the partmap and the adjacent carbon in the target is assigned to it. More generally there will be several carbon atoms adjacent to each atom, and it is necessary to establish the correspondences between them. First, the carbon skeleton matches generated in the preceding iteration of the LIR detection routine are checked. Atoms can only be assigned to others from the same set since these are the ones with an identical environment. If for an atom adjacent to the root atom there is exactly one atom in the set of equivalent atoms, then a unique assignment can be made and placed in the partmap.

If atoms still remain undifferentiated, they are equivalent within the carbon subgraph, but this does not necessarily mean that they are equivalent within the full target and SM structures. The full symmetry of the structures must be examined. If the atoms are equivalent with respect to the full structural symmetry in either the target or the SM, then it does not matter which atom from one structure is paired with one from the other and a single, arbitrary assignment is made and placed in the partmap. If this is not the case, then every possible combination must be recorded. The partmap is duplicated and each possible assignment is added to a separate copy of the partmap. One of these partmaps is then processed further. The others are stored to be processed later.

Once all the α atoms have been assigned, the correspondences of the β atoms are recovered from the list of sets at the previous level. This process is repeated until the first level of sets is reached. This marks the edge of the identical area and the partmap growing stops. If there are any partmaps stored from duplication, they are now processed to completion in the same way.

Each pair of root atoms is investigated in the same way. If the root atom from one structure is symmetrically identical to one which has already been paired, then no partmap is generated, since it would produce matchings symmetrically equivalent to those already found.

The completion of this section of the process gives the complete set of LIRs, but smaller areas of identity between the structures may be important. To find these, partmaps are also constructed around the atoms from preceding levels of sets, one by one. Atom-to-atom correspondences which are adjacent to root atom pairs are ignored since they would merely produce subsets of the larger partmaps already constructed. The process is repeated for each preceding level until enough identical areas have been generated according to the following criteria:

1. The process stops if the partmap array size has been reached (this is a parameter defined within the program and currently set to 60).

2. If more than 20 partmaps have been generated the process stops when all root atoms at the current level have been processed.

3. If any partmaps have been generated by higher levels the process terminates when level 2 is reached (the level relating to information about atoms and those α and β to them).

4. If the only root atom pairs found are at level 1, these are used (atoms with the same connectivity).

(c) Combining the Partmaps. Once all the partmaps have been constructed the possible overlap between them is investigated. Partmaps which overlap favorably are combined to give larger ones.

The overlap between two partmaps is considered to be good if no atom in the target is mapped to two different atoms in the SM or vice versa. If one or more atoms are mapped in both partmaps, the partmaps are connected. If no atoms are shared between the partmaps, the quality of the overlap depends on the path between them: the paths connecting the partmaps in the SM and target structures are grown and compared. If a path exists between any pair of atoms in one partmap and another in the target and one of the same length exists between the equivalent atoms in the SM then the combination of the partmaps is considered to be good and they are connected. If a pair of paths exists but they are of different lengths, they are connected but the combined map is considered to be poor and it is penalized during the sorting stage. (The approach is simplistic in that it does not take account of the atom types on the path: this is left to later parts of the matching process.)

Within the limitations of the permitted size of the partmap array, all possible combinations of partmaps are considered. Each combination is checked against those already made to

ensure that it is unique: if the atom-to-atom mapping is identical for every atom, then the new partmap is rejected; if atoms are matched to different atoms in the two partmaps but these atoms are symmetrically identical, then the new partmap is rejected.

Once the overlap of the partmaps has been fully investigated they are sorted on the number of atoms which have been allocated partners. A score of 1 is given for each atom successfully mapped. The scores for poor combined partmaps are reduced by the difference in length of the most similar paths between their constituent partmaps.

The highest ranking partmaps are processed first to find their full mapping. This is done for two reasons. Partmaps with highest ranking are likely to lead to mappings which need the least chemistry to convert the SM to the target, since a high ranking implies a large carbon skeleton common to the SM and target. In addition, as partmaps with more atoms already mapped generally have fewer atoms left to map during the backtracking phase, they are likely to be dealt with quicker. The combination of these two factors provides a cut-off parameter as quickly as possible for subsequent mapping.

(d) Finding Further Atom-to-Atom Correspondences. As much information as possible has now been obtained by the manipulation of sets of similar carbon atoms. However, it may still be possible to find chemically meaningful correspondences between atoms which have different connectivities. To devise a synthesis plan which will rectify the differences between the SM and target, these additional atoms must be mapped, and their differences must be defined as accurately as possible.

Each of the partmaps and combinations of partmaps is investigated in turn to see if neighboring atoms that differ in the SM and target can nevertheless be added to create full maps. Heteroatoms are included in the search if they have more than one connection. (Singly-connected heteroatoms need not be included since mismatches can be rectified later by functional group interconversions.)

A backtracking algorithm is used at this stage. Since the search is constrained only by atom type (all other attributes can in principle be changed by chemical reactions) and distinct areas from which to start the mapping are known, pure set manipulation offers no advantages. The order in which atoms are to be mapped is decided first. Since there are usually fewer atoms in the SM than in the target the search is based on the set of SM atoms.

An unmapped atom adjacent to a mapped atom is selected, and its identifier is placed at the head of a search list. Further atom identifiers are added in order to the list by growing out from the first atom until the end of the chain of atoms is reached. If branches have been encountered it is necessary to backtrack to each branch point and add the atoms of the additional branches in the same way. This process is repeated starting from every unmapped atom adjacent to a mapped atom, atom identifiers being added to the bottom of the existing search list. A pointer to the next branch or appendage is associated with each list entry. This allows unnecessary mapping attempts to be skipped as soon as a mismatch is encountered.

An attempt is made to map the first atom in the search list to an atom in the target. (Since the first atom is adjacent to a mapped atom, only atoms adjacent to the equivalent mapped atom in the target need to be considered.) If the mapping is successful, an attempt is made to map the second atom in the list, searching for a corresponding atom adjacent to the one that has just been mapped in the target. This process is repeated until no more correspondences are found or the end of the list is reached. When an atom is successfully mapped it is excluded from further mapping.

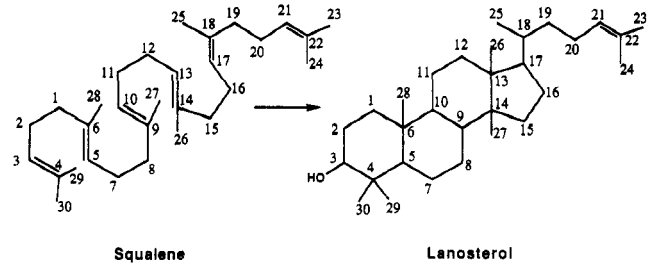

Figure 4. Mapping of squalene onto lanosterol is hard for LHASA to find.

To find the next map, the last successfully mapped atom is unmapped and a search is made for an alternative mapping. If one is found, the search proceeds forward through the list again. If not, the next atom is unmapped. This process continues until all possible maps have been found. If a map is found to repeat the atom-to-atom correspondences of one previously found it is discarded (to eliminate identical maps grown by starting from different partmaps).

The number of maps possible even for moderate-sized structures can be very large. It is not practical to store all of them and in any case many of them will be of little chemical value. As each map is generated the mapped structures are therefore submitted to a routine to calculate a rating, the heuristic estimate of synthetic proximity (HESP).[2] Once 20 maps have been found they are ordered upon their HESP values. The rating for the twentieth is taken as a cut-off. Maps with negative HESP values are automatically deleted. No map subsequently found which receives a lower rating is kept. Maps with higher ratings are added to the set of 20, displacing members at the bottom of the list.

The final list of maps is displayed to the chemist in order of HESP values.

LIMITATIONS OF THE ALGORITHM

The algorithm is well suited to handling the wide diversity of problems encountered in a synthesis design program. However, it will not work in a few cases. In general the cases which fail are those in which the similarity between the structures is very small, or divided into a large number of very small portions by areas of dissimilarity.

Structures for Which Too Many Partmaps Are Generated. Targets and SMs sharing a very large number of small areas of similarity present a combinatorial problem. In Figure 4 the mapping of lanosterol with its biological precursor squalene is depicted. With the exception of the eight-carbon side chain, the areas of similar connectivity are divided into small regions by the bonds forming the ring closures, making the potential number of small overlaps very large. The list of root atom correspondences for areas of two atoms or larger is shown in Table I. All of these correspondences give rise to partmaps, and there is no a priori reason for selecting one over the others.

In cases like this, the program can exhaust its array space for partmaps before the correct selections have been generated. As a result, whether or not the correct mapping is found depends on the order in which the atoms are drawn by the user. The problem could be solved by permitting more partmaps to be stored, but this would greatly increase memory requirements and the time taken to process structures.

Structures for Which No Partmaps Are Generated. If the SM does not have any atoms with the same initial connectivity

Table I. Atom-to-Atom Correspondences for Squalene Mapped to Lanosterol

Level 6 Equivalences	
SM	target
23, 24, 29, 30	23, 24

No New Equivalences at Levels 5–3

Level 2 Equivalences		
set	SM	target
1	23, 24, 25, 26, 27, 28, 29, 30	23, 24, 25
2	2, 7, 11, 12, 16, 20	2, 20
3	1, 3, 5, 8, 10, 13, 15, 17, 19, 21	7, 8, 11, 16, 19, 21

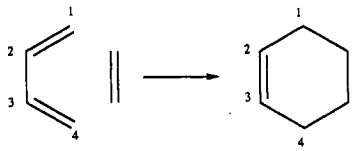

Figure 5. Mapping for the Diels–Alder reaction—ethene remains unmapped.

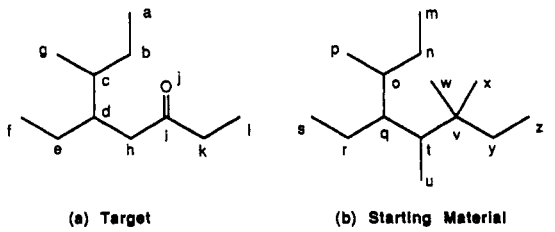

Figure 6. Example of a target and starting material for mapping.

as atoms in the target then there is nowhere from which to begin the LIR generation. This means that no matches can be made between the structures. Such cases are rare and usually involve very small starting materials for which the relationship to the target structure can only be gleaned by use of chemical knowledge. An example is the Diels–Alder reaction of butadiene with ethylene shown in Figure 5. The butadiene moiety is mapped to the diene portion of the structure in six possible orientations, one of which corresponds to the chemically "correct" mapping, but the ethylene structure in the SM cannot be mapped onto cyclohexene as there are no carbon atoms in the target with only one carbon neighbor.

A possible solution to this problem is to permit atoms to match any atom of the same type if no LIR can be found. This effectively reduces the problem to one of an unconstrained backtrack algorithm which is known to be np complete and would be expected to generate very many solutions. It would be difficult to ensure a high rating by the HESP algorithm for the desired mapping since it is dependent on specific chemistry rather than on topological similarity. This option would only be required in very rare circumstances, and it has not been coded.

WORKED EXAMPLES

The target and SM shown in parts a and b of Figure 6 are used to illustrate how the LIRs are found. In practice, atoms are identified by integers, but for the sake of clarity they are identified by letters in this paper.

Identification of Common Subgraphs. Figure 7 shows the atoms in the shells around atom h in the SM: the LEVELS used in the matching process. Similar shells can be built around all the atoms.

First a label equal to the number of adjacent carbon atoms is assigned to every carbon atom in both structures (Figure

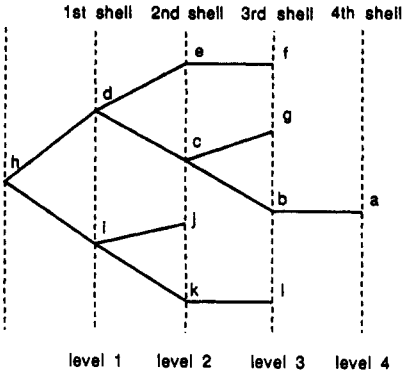

Figure 7. Shells of atoms around atom h in the starting material.

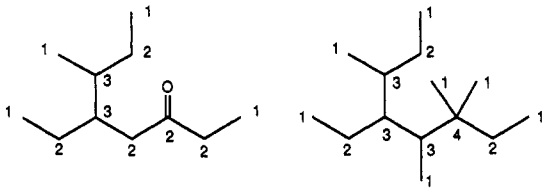

Figure 8. Initial labeling of atoms according to the number of neighboring carbon atoms.

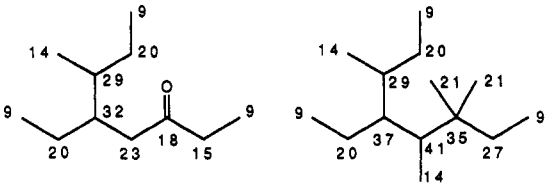

Figure 9. Labeling of atoms at level 1.

Table II. Sets of Atoms at Level 1

set	starting material atoms	target atoms
1	a, f, g, l	m, p. s, u, w, x, z
2	b, e, h, i, k	n, r, y
3	c, d	o, q, t

Table III. Sets of Atoms at Level 2

set	starting material atoms	target atoms
1	a, f, l	m, s, z
2	g	p, u
3	b, e	n, r
4	c	o

Table IV. Sets of Atoms at Level 3

set	starting material atoms	target atoms
1	a, f	m, s
2	b	n
3	g	p

8), and a list is created of sets of atoms bearing each label found in both the SM and the target (Table II).

To further differentiate the atoms, the attributes of their neighbors are now taken into account. A new label for each atom is calculated as the sum of 5 times the current label plus the sum of the squares of the labels of the adjacent atoms (Figure 9). Sets of equivalent atoms are collected again to create the list in Table III, classifying atoms according to their β environment.

The process is repeated to give the sets shown in Tables IV and V. There is only one atom each from the SM and target in level 4 (Table V), and these must therefore mark the limit of the single LIR. Atoms a and m are a root atom pair, and they are mapped onto each other. There is only one atom α

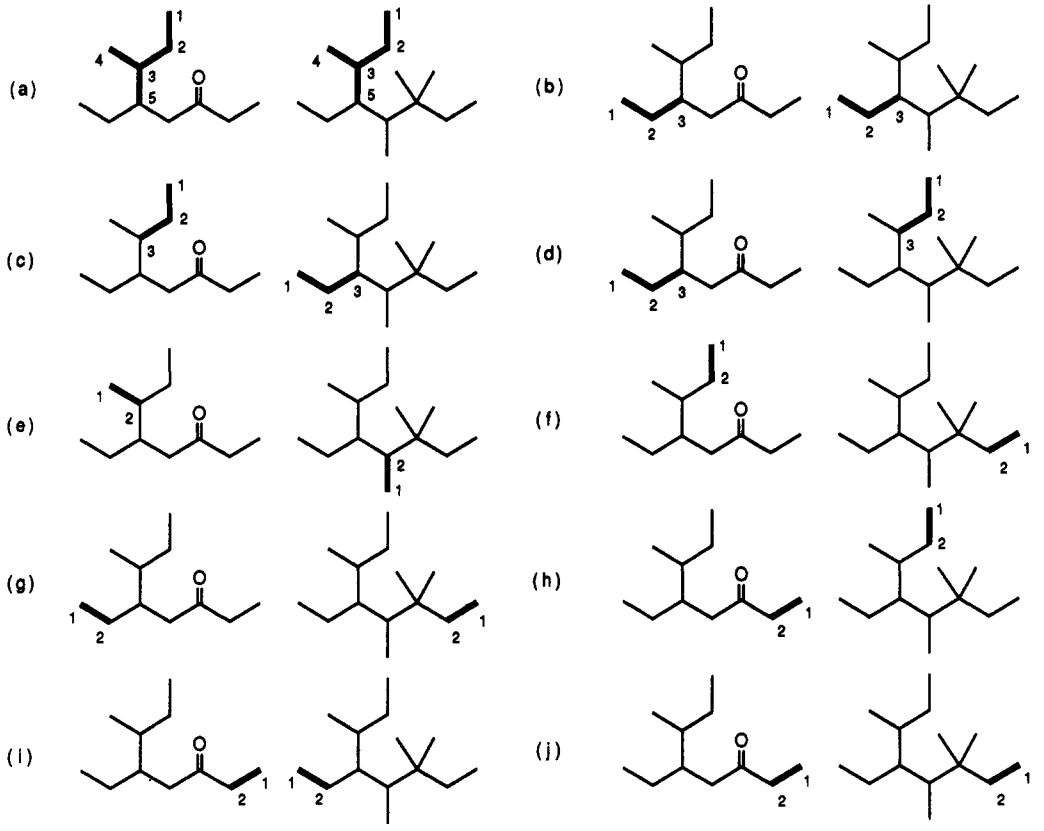

Figure 10. Examples of partmaps.

Table V. Sets of Atoms at Level 4

starting material atoms	target atoms
a	m

to each of a and m (b and n, respectively), and they can therefore also be mapped onto each other. Similarly atoms c and o can be mapped immediately.

There are two atoms adjacent to atoms c and o (d,g and p,q). The traverse from atoms a,m at level 4 to atoms d,g,p,q took three steps, and so level 1 contains information which may help to determine whether, for example, atom d should be mapped to atom p or atom q. Reference to Table II shows that d is in the same set as q, and g is in the same set as p, and so they can be assigned to each other accordingly. No further information about this map can be extracted from the tables, and this is a completed partmap (Figure 10a).

The process is now repeated for the sets at level 3. The pairs g,p and b,n simply create submaps of the first partmap and they are ignored. The mapping of a to m gave the first partmap, but the mapping of f to s is new, and the mappings of a to s and f to m have to be considered, giving the partmaps in Figure 10b–d.

Finally, the process is repeated for sets at level 2, giving the further partmaps in Figure 10e–j, inclusive. As some partmaps have already been generated no attempt is made to generate any partmap from level 1 and the partmap generation terminates.

Overlap of Partmaps. The partmaps are examined to see if there is overlap between them. Any sets of partmaps in which no atom in the SM is mapped to different atoms in the target, or vice versa, can be combined. Partmaps which share common atoms can be connected immediately. Thus the partmaps in Figure 10, parts a and b, are connected to give the combined partmap in Figure 11a. In addition, the partmap in Figure 10j overlaps favorably with the new partmap (i.e., no mappings in one partmap conflict with mappings in the other), and this can be added to create the disconnected map in Figure 11b. Similar disconnected maps can be constructed from several other pairs of partmaps: an illustration is the one in Figure 12, constructed from the partmaps in parts c and j of Figure 10.

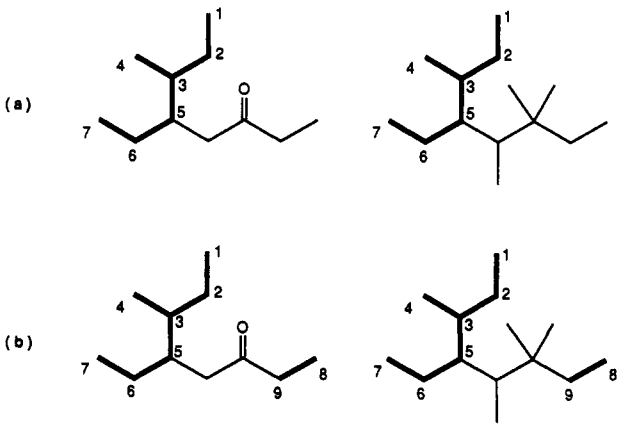

Figure 11. Combined partmaps.

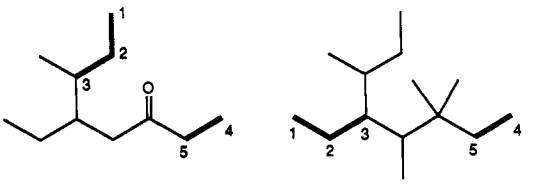

Figure 12. Combined partmaps.

Now the disconnected combined maps are examined to find the connections between the partmaps within them. First, the shortest path is grown between the partmaps in the SM. Then the shortest path between the partmaps in the target is grown. If the path lengths are equal, then the atoms in the paths are added to the combined map to create a full map

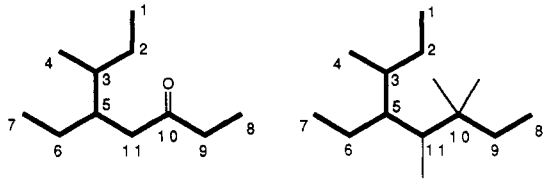

Figure 13. Completed full map.

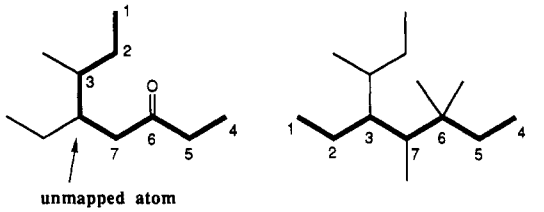

Figure 14. Alternative full map.

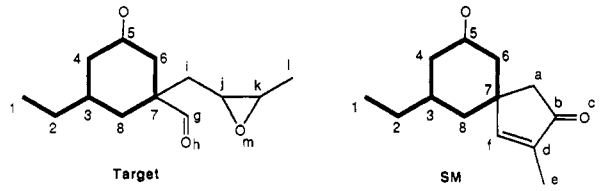

Figure 15. Example of target and starting material for mapping by backtracking.

Table VI. Example of a Search List for Backtracking

list position	atom identifier	pointer	list position	atom identifier	pointer
1	a	6	7	d	11
2	b	6	8	b	10
3	d	6	9	a	10
4	e	5	10	e	11
5	f	6	11	end	0
6	f	11			

(e.g., the full map in Figure 13, created from the combined map in Figure 11b). If the path lengths are not equal, other, longer paths are compared. If a pair of equal paths is found, this is used to create a full map. If there is no equal path, this overlap will still be used later to create a full map which will be penalized in its ranking against other maps because it requires a carbon chain expansion or contraction in the synthetic sequence. An example is the full map in Figure 14, created from the combined map in Figure 12.

Backtracking. In the above example all of the carbon atoms in the SM are mapped onto atoms in the target (see Figure 13). The job of mapping is complete and no backtracking is needed.

Backtracking is illustrated by considering the example in Figure 15. At the end of the above mapping process the partmap shown in bold has been built. Now the SM is used as the basis for the search list in Table VI. Appendages to the partmap are built starting from every unmapped atom adjacent to it. The first adjacent atom found to be unmapped is atom a. The only unmapped atom adjacent to it is atom b, and this is placed second on the list. Atom c is ignored because it is a heteroatom with only one connection. Atom d is placed third in the list.

There is a choice of atoms for the fourth position, and atom e is chosen arbitrarily. There are no further atoms attached to atom e, and so it marks the end of an appendage. A failure pointer for this atom is set, pointing to the next position in the list. (If there had been other atoms in the chain leading from the branching point to atom e, the failure pointer for those atoms would also have been set to the same value, since there is no reason to continue the search along an appendage following failure at any atom in it.) The algorithm returns to the last branching point and looks for the start of another appendage. This time atom f is added to the list. There are no more unmapped atoms attached to atom f. Now there are no branch points right back to atom a, and the failure pointer for all atoms back to that point can be set to the next position in the list, position 6.

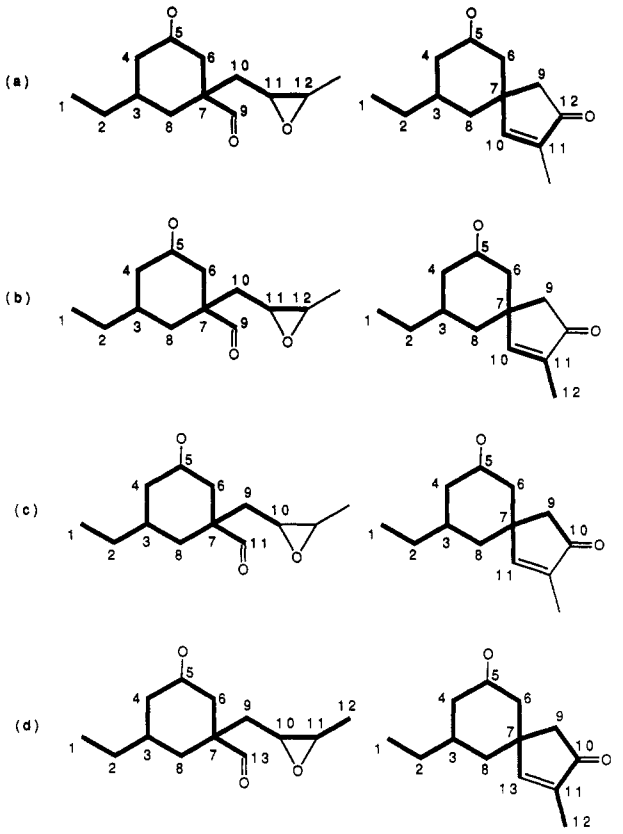

Figure 16. Extended maps generated during backtracking.

The process is repeated for the next atom adjacent to the partmap which has not been mapped: atom f. Atoms are listed again even though they have already appeared in the search list so that the description of each appendage is complete. Any redundancy which might arise is dealt with during the following mapping phase.

After this no more unmapped atoms can be found in the SM and so the list is complete and position 11 is flagged as the end of the search list. (As explained above, letters have been used as atom identifiers instead of integers for the purposes of these examples. In the actual system the end of the list is flagged by setting the atom value to zero.)

Mapping to the target now commences. An atom of the same type as the first atom in the list, atom a, and adjacent to the corresponding mapped atom is chosen arbitrarily. In this case, let it be atom g. The next atom in the search list is atom b, another carbon atom. There is no unmapped carbon atom attached to atom g, and so this mapping ends.

The failure pointer switches the search to position 6 in the list, atom f. The only remaining unmapped atom in the target to which f can be mapped is atom i. Therefore atom d is mapped to atom j. A previous attempt has been made to map the next atom in the list, atom b, but the attempt failed and a new attempt is therefore allowed. This time a mapping to atom k is successful. The next atom in the list, at position 9, atom a, has already been mapped successfully in this run, and it is therefore excluded. The atom in position 10 is atom e. Since this carbon atom is adjacent to atom d which has already

been mapped to atom j, only an unmapped carbon atom adjacent to j would be a candidate. There is no such atom and mapping therefore ends.

The next position is the end of the list. This map is therefore complete (see Figure 16a). Now the last mapped atom is unmapped and an alternative mapping for it is sought. The map in Figure 16b is found. After this no alternatives are possible until all the atoms have been unmapped, allowing a tentative mapping of atom a to atom i. This is successful, and mapping proceeds along the appendage until a failed attempt is made to map atom d to atom m. The map is completed with the mapping of atom f to atom g, giving the map in Figure 16c. Finally, the algorithm backtracks until the remapping of atom d to atom k is found, leading to the map in Figure 16d.

All atoms adjacent to the partmap have now been used as starting points for the growing of appendages, and the complete set of full maps has been created. For a simple structure like this one the number of maps is small. For more complex structures the number could potentially be large. As each map is created it is therefore evaluated, and only the 20 best maps found are retained. The evaluation process will be discussed in the following paper in this issue.[2]

ACKNOWLEDGMENT

We extend thanks to E. J. Corey, A. K. Long, and S. Rubenstein for useful discussions and to P. N. Judson who assisted in the preparation of this paper. This work has been supported by the Science and Engineering Research Council, the Department of Education and Science, Imperial Chemical Industries, the Wolfson Foundation, and the Maxwell Foundation.

REFERENCES AND NOTES

(1) Johnson, A. P.; Marshall, C.; Judson, P. N. Starting Material Oriented Retrosynthetic Analysis in the LHASA Program. 1. General Description. *J. Chem. Inf. Comput. Sci.*, first of three papers in this issue.

(2) Johnson, A. P.; Marshall, C. Starting Material Oriented Retrosynthetic Analysis in the LHASA Program. 3. Heuristic Estimation of Synthetic Proximity. *J. Chem. Inf. Comput. Sci.*, third of three papers in this issue.

(3) Sussenguth, E. H. A Graph-Theoretic Algorithm for Matching Chemical Structures. *J. Chem. Doc.* **1965**, *5*, 36–43.

(4) Ming, T.-K.; Tauber, S. J. Chemical Structure and Substructure Search by Set Reduction. *J. Chem. Doc.* **1971**, *11*, 47–51.

(5) Figueras, J. Substructure Search by Set Reduction. *J. Chem. Doc.* **1972**, *12*, 237–44.

(6) Cone, M. M.; Venkataraghavan, R.; McLafferty, F. W. Computer-Aided Interpretation of Mass Spectra. 20. Molecular Structure Comparison Program for the Identification of Maximal Common Substructures. *J. Am. Chem. Soc.* **1977**, *99*, 7668–71.

(7) Lynch, M. F.; Willett, P. The Automatic Detection of Chemical Reaction Sites. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 154–9.

(8) McGregor, J. J.; Willett, P. Use of a Maximal Common Subgraph Algorithm in the Automated Identification of the Ostensible Bond Changes Occurring in Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 137–40.

(9) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures–A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–13.