# Starting Material Oriented Retrosynthetic Analysis in the LHASA Program. 3. Heuristic Estimation of Synthetic Proximity

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

A procedure is described for deriving a numerical representation of the synthetic proximity of two compounds, that is, the expected ease with which one could be converted to the other. The procedure has been used for the starting material oriented retrosynthetic strategy in the LHASA program. Account is taken of the number and types of atoms already mapped by another procedure,[2] the number of atoms which must be changed to interconvert the compounds, and the changes which must be made to functionality, bonding, and stereochemistry. Where changes are necessary, account is taken of the proximity of facilitating functional groups.

## INTRODUCTION

This paper is the third in a series entitled Starting Material Oriented Retrosynthetic Analysis in the LHASA Program.[1] These 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. The previous paper[2] described a method for matching two chemical structures: the target and starting material(s) of a proposed synthesis. Normally more than one possible matching is found. This paper describes the method used to rank the possibilities using a heuristic estimate of synthetic proximity (HESP).

## DISCUSSION

To select the most promising SM from a set of compounds mapped onto a target compound it is necessary to judge what is or is not a chemically rectifiable difference and to compare the ease of making the chemical changes that the different mappings would require, that is, a measure is needed of the proximity, in chemical terms, of a SM to a target to which it is mapped. For example, there is little difference between the proximities for the mappings in Figure 1: both require the retrosynthetic breaking of one carbon–carbon bond, the making of another (shown in bold), and three functional group interchanges (FGI, marked with an asterisk). The mapping in Figure 1a should perhaps be given slight preference because it involves the breaking of a carbon–carbon double bond rather than a carbon–carbon single bond. The mapping in Figure 2a should be more strongly preferred over the mapping in Figure 2b—each mapping requires the making and breaking of two $\sigma$ bonds, but the second mapping also requires functional group interconversions at atoms 6 and 8. Calculation of a numerical HESP value allows LHASA to apply this kind of reasoning quickly.

A measure of *synthetic proximity* is also used for other purposes in the LHASA Starting Material Oriented Strategy module, such as the evaluation of alternative synthetic sequences, for which the synthetic proximities of intermediates to the SM must be compared.[3] Values derived from the HESP calculation must be suitable for comparisons of different mappings between the same structures and of mappings between one target and a range of SMs.

The concept of synthetic proximity, or chemical distance, has been discussed before,[4,5] and a method using the graphical differences between the structures to determine the number of bonds which change in the conversion of the SM to the target has been described.[5] The procedure described in this paper uses additional criteria of a more chemical nature to take account of the chemical manipulations of the SM which are needed to convert it to the target.

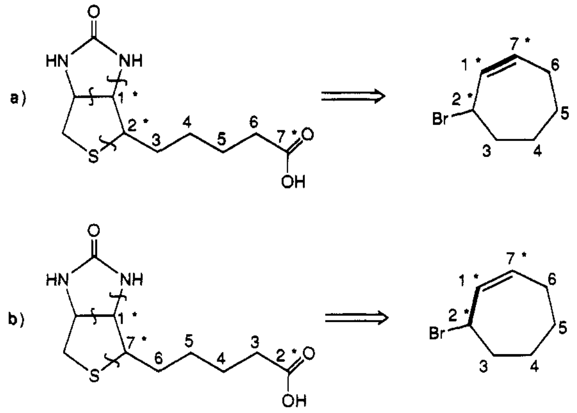


Figure 1. Similar mappings with similar ratings.

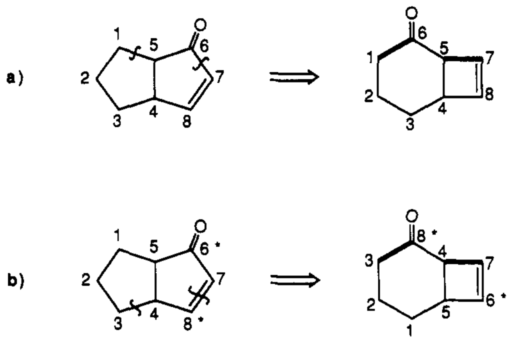
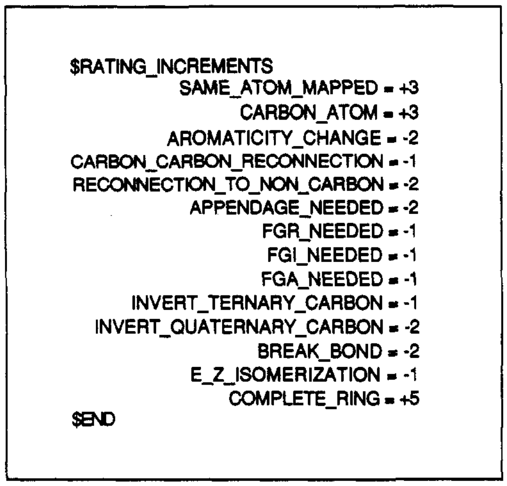


Figure 2. Similar mappings with different ratings.

The result of mapping a starting material to a target is a set of atom-to-atom correspondences. The correspondences are used to identify chemically meaningful relationships between the structures from which to calculate HESP values, and they also provide the goal list of mismatches which must be rectified during the retrosynthetic analysis. The nature of the changes needed (e.g., functional group modification, carbon appendage removal, etc.) is recorded for later use.[3]

In the LHASA system, the overall proximity of a mapping between two structures is measured by the sum of the prox-

0095-2338/92/1632-0426$03.00/0



```
$RATING_INCREMENTS
                    SAME_ATOM_MAPPED = +3
                         CARBON_ATOM = +3
                  AROMATICITY_CHANGE = -2
       CARBON_CARBON_RECONNECTION = -1
      RECONNECTION_TO_NON_CARBON = -2
                    APPENDAGE_NEEDED = -2
                          FGR_NEEDED = -1
                          FGI_NEEDED = -1
                          FGA_NEEDED = -1
             INVERT_TERNARY_CARBON = -1
        INVERT_QUATERNARY_CARBON = -2
                          BREAK_BOND = -2
                  E_Z_ISOMERIZATION = -1
                       COMPLETE_RING = +5
$END
```

**Figure 3.** User-definable file containing incremental values for the HESP.

imities for each mapped atom plus an increment for each fully mapped ring

$$\mathrm{HESP} = \sum_{i=1}^{n}(a_i - d_i) + \sum_{j=1}^{m} b_j$$

where $a$ = size of atom increment, $d$ = size of atom decrement, $b$ = size of ring increment, $n$ = number of mapped atoms, $m$ = number of mapped rings.

The contribution of each atom consists of two components, a positive contribution because the atom has been mapped and a negative one for any mismatching of the features of the atom. The positive contribution means that mappings of large portions of the structure are more highly rated than mappings of smaller portions. The decrements ensure that mappings which require a large amount of rectifying chemistry are given low HESP values. They take account of changes that are needed to functionality, bonding, and stereochemistry by checking each mapped SM atom in turn and comparing it with the corresponding atom in the target. The additional positive contribution for rings is included because of the extra synthetic value many chemists place on a fully matched ring compared to the equivalent number of matched acyclic atoms.

The program is provided with a default set of increment and decrement values which can be overridden by the user (see Figure 3). The default values were selected to give results consistent with the expectations of a chemist. They were derived by manual stepwise refinement so that the highest rated mapping between structures corresponded to the chemically "best" mapping obtained by visual inspection for a diverse selection of compounds.

The user can change any or all of the values to reflect particular considerations that are important for the problem under investigation. For example, a chemist designing a synthesis in which the control of stereochemistry is important might increase the value of the decrement for mismatched stereofeatures to give maps that correctly match the stereogenic features in the structures a higher rating than those that map more of the structure but less of the stereochemistry correctly. Similarly, the decrement for mismatched aromaticity might be increased to force the matching of aromatic ring systems by a user more concerned with aromatic substitution chemistry than with building the aromatic rings from nonaromatic precursors.

These options are particularly useful when the starting material is one automatically selected from a pool of available starting materials rather than a single, user-selected SM. In effect, manipulation of the increments and decrements allows the user to guide the selection of the SM toward compounds which have a particular type of relationship to the target structure.

Each SM atom is checked, and as mentioned above, the HESP value is incremented if the atom is correctly mapped to an atom in the target (i.e., if both atoms are the same chemical element). This increment recognizes that the most important measure of the goodness of a map is its size. The value is further incremented if the atoms are carbon atoms: maps maintaining large unchanged carbon skeletons throughout a synthesis are preferred over maps of similar size but including heteroatoms, since carbon–carbon bonds are harder to make or break than carbon–heteroatom bonds.

If the SM atom is the origin of a functional group (i.e., a carbon atom to which a functional group recognized by LHASA is attached) but the corresponding target atom is not, a functional group must be removed during the synthesis. Retrosynthetically a functional group addition (FGA) is required. The HESP value is decremented. Similarly, if the target atom is also the origin of a functional group but the group is not identical to the one on the SM atom, then a functional group interconversion is needed, and the HESP value is decremented.

If the target atom is the origin of a functional group but the SM atom is not, then the atom must be functionalized during the course of the synthesis. The ease with which this can be done normally depends on the proximity of other functional groups. Strictly, because of the differing possibilities of remote functionalization, it also depends on the nature of the proximal groups, but to take this into account would necessitate the inclusion of too much specific chemical knowledge in the evaluation routine. In practice, it has been found that ignoring this information does not seriously impair the performance of the routine. Shells of atoms neighboring the SM atom are checked until the closest functional group is found. In the current implementation of the program, the HESP value is reduced by the square of the number of bonds separating the SM atom from the nearest functional group, up to a maximum decrement of 16.

Next the stereochemistry at the pair of atoms under examination is checked. If both atoms are stereocenters, the dispositions of the mapped atoms around them are compared. It is not possible to use definitions derived from the Cahn–Ingold–Prelog rules for this comparison because reactions may have changed the weightings of appendages. The check is based on the rotation lists generated by LHASA for the target and SM. For each stereocenter, LHASA creates a list based on the positions of atoms drawn by the user. If the stereo-defining bond is a wedge, the atoms attached to the center are listed in clockwise order; if it is a dashed bond they are listed in anticlockwise order.

To do the stereochemical check, a new list is made in which each atom in the target is located according to the position of the atom to which it is mapped in the SM. The new list is compared with the original list for the target, and pairs of atoms in the new list are swapped until the two lists are identical. If the number of swaps is zero or even, then the two centers are stereochemically equivalent. If the number of swaps is odd, then a stereoinversion is required in the course of the synthesis.

For example, suppose that in the mappings shown in Figure 4 atoms represented by the same letters of the alphabet have been mapped together. The rotation list for the target in Figure 4a is ABCD and the rotation list for the SM is adcb. A new rotation list is created, placing target atoms in the

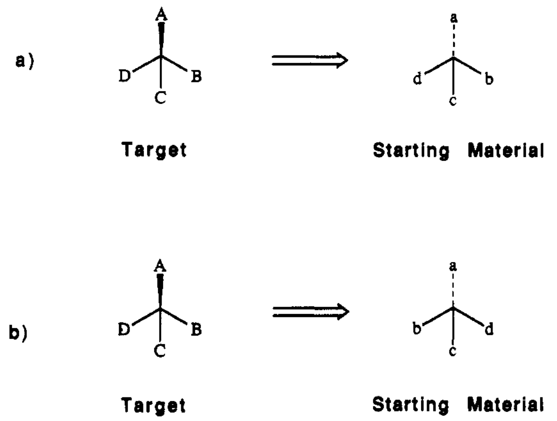


**Figure 4.** Mapping of stereocenters.

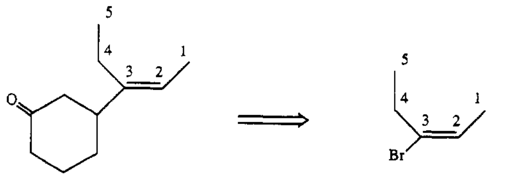


**Figure 5.** Good mapping of an asymmetrical alkene.

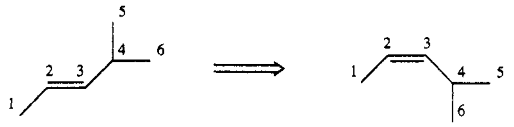


**Figure 6.** Mapping of an asymmetrical alkene requiring an isomerization.

**Table I.** Example of the Calculation of a HESP Value Using the File of Incremental Values Shown in Figure 3 for the Mapping Shown in Figure 7

| atom number | features taken into account | change to HESP value |
|---|---|---|
| 1 | a carbon atom mapped to a carbon atom; the only bond on it is also correctly mapped | +6 |
| 2 | a carbon atom mapped to a carbon atom | +6 |
| | site of a FGI | −1 |
| | site of a C–C bond formation but with a functional group at the same site | −2 |
| 3 | a carbon atom mapped to a carbon atom; the stereochemistry is also correct | +6 |
| 4 | a carbon atom mapped to a carbon atom | +6 |
| | a functional group addition is required in the synthetic direction and the nearest functional group is two bonds away | −4 |
| 5 | a carbon atom mapped to a carbon atom | +6 |
| | the double bond to atom 6 needs to be isomerized | −1 |
| 6 | a carbon atom mapped to a carbon atom; the double bond has already been considered through its connection to atom 5 | +6 |
| 7 | a carbon atom mapped to a carbon atom | +6 |
| | one carbon carbon bond must be broken in the synthetic direction and the nearest functional group is one bond away | −2 |
| | final HESP value | 32 |

order defined by their relationship to the atoms in the SM. Thus the new list is ADCB. If the second and fourth entries in this list are exchanged, the list becomes ABCD, which is identical to the original list for the target. An odd number of exchanges was required, and it can be concluded that stereoinversion is required. The target list for Figure 4b is ABCD. The SM list is abcd, and so the new list created for checking is ABCD. This is identical to the original list for the target. No exchanges of atoms are necessary and so the stereocenters are recognized to be identical.

If more than one atom attached to the stereocenter in either the target or the SM is unmapped, there is insufficient information to determine the stereochemical relationship between the two centers and so they are not checked. If just one atom pair is unmapped, there is still sufficient information to proceed with the analysis using the three mapped pairs of atoms and to discover whether inversion or retention of configuration is required.

If inversion is required at a center, then this becomes a goal in itself. Ease of stereochemical isomerization depends on both the nature of the center and the proximity of the closest functional group (including any group at the stereochemical center). The parameter file containing default values for increments and decrements allows for stereocenters carrying a hydrogen atom (ternary centers) to be distinguished from those that do not (quaternary centers) since the latter are normally harder to invert than the former. The decrement currently made is the product of the value contained in this file and the square of the number of bonds separating the stereocenter from the closest functional group in the SM.

When control of stereochemical configuration is required in a synthesis, the use of some reactions may be prevented or restricted, for example, by the need to avoid reaction conditions that would cause unwanted inversion or racemization. The need for stereochemical control is therefore recorded and taken into account later in the retrosynthetic analysis, when transforms are selected from the knowledge base.

Next, bonds from each SM atom to adjacent carbon atoms are checked. (Bonds to heteroatoms from the atoms under examination are implicitly considered by the functional group checking described above.) If all the carbon atoms adjacent to the SM atom map to carbon atoms adjacent to the target atom, no changes are required. If there is no carbon atom adjacent to the target atom corresponding to one adjacent to the SM atom, then a bond must be broken during the synthesis, and so the HESP value is decremented. For example, in Figure 1a there are two mapped carbon atoms bonded to atom 1 in the SM: atom 2 and atom 7; in the target, none of the atoms bonded to atom 1 is mapped to atom 7, and so the bond between atom 1 and atom 7 must be broken during the synthesis. Similarly, if there is no atom adjacent to the SM atom corresponding to an atom adjacent to the target atom, then a bond must be made during the synthesis (i.e. a bond is broken in the retrosynthetic sequence). The proximity of functional groups to SM atoms at which carbon–carbon bonds are made or broken is checked. If there is no functional group on the SM atom, the decrement is currently multiplied by the square of the number of bonds separating the closest functional group from the SM atom up to a maximum decrement of 32.

Nonaromatic double and triple carbon–carbon bonds are recognized as functional groups by LHASA and are therefore included in the functional group checks described above. Aromatic bonds are not treated as functional groups, and so a check is made for their presence and a decrement is made for each bond which changes from aromatic to nonaromatic or vice versa.

The geometry about double bonds is checked using a similar procedure to the one used for checking stereochemistry at single atoms described above. If isomerization during the synthesis is dictated by the geometry of the mapped atoms, the HESP value is decremented. Thus, for example, the map in Figure 5 is a good one whereas the map in Figure 6 implies isomerization and attracts a decrement.

To avoid counting bonds twice, checks on aromatic bonds and double bonds are made only for bonds from each atom

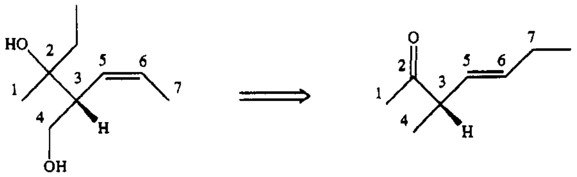


**Figure 7.** Mapping evaluated in Table I.

**Table II.** HESP Values for Some Pairs of Structures Shown in Figures in This Paper

| structures | HESP value |
|---|---|
| Figure 1a | 37 |
| Figure 1b | 37 |
| Figure 2a | 41 |
| Figure 2b | 39 |

to those with higher mapping numbers.

Finally the HESP value is incremented for every ring which maps completely between the SM and the target, because a ring is so commonly and successfully used as the central backbone for a synthesis.

Table I illustrates the calculation of a HESP value for the mapping shown in Figure 7. Table II shows the HESP values for the structures in Figures 1 and 2.

## USEFULNESS OF HESP VALUES

The HESP values calculated by this method are suitable for different mappings of the same structures and for mappings of different structures to a single one. They are used for both these purposes in the LHASA program, in the first case to find the best mapping of a SM onto a target,[2] in the second case to choose between alternative potential SMs[3] and to select nodes from the growing retrosynthetic tree for further processing.[6] The procedure is not designed for the comparison of mappings between different pairs of structures (i.e., the comparison of A mapped to B with C mapped to D): to be useful for this purpose the HESP value would need to be scaled in proportion to the fraction of the target included in the map.

An advantage of the HESP value over the measures of chemical distance used by Ugi[4] and Hendrickson[5] is that it takes greater account of the likely ease of practical, chemical conversion of the SM into the target.

## 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, and the Wolfson 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. 2. Mapping the Starting Material and Target Structures. *J. Chem. Inf. Comput. Sci.*, second of three papers in this issue.

(3) Johnson, A. P.; Marshall, C.; Hopkinson, G. A.; Judson, P. N. Starting Material Oriented Retrosynthetic Analysis in the LHASA Program. 4. Selecting Good Retrosynthetic Routes. *J. Chem. Inf. Comput. Sci.*, in press.

(4) Ugi, I.; Bauer, J.; Brandt, J.; Friedrich, J.; Gasteiger, J.; Jochum, C.; Schubert, W.; Dugundi, J. Computer Problems for the Deductive Solution of Chemical Problems on the Basis of Mathematical Models—a Systematic Bilateral Approach to Reaction Pathways. *Comput. Methods Chem. (Proc. Int. Symp.)* **1979** (Publ. 1980), 275–300.

(5) Hendrickson, J. B.; Braun-Keller, E. Systematic Synthesis Design. 8. Generation of Reaction Sequences. *J. Comput. Chem.* **1980**, *1*, 323–33.

(6) Johnson, A. P.; Marshall, C.; Su, L.; Judson, P. N. Starting Material Oriented Retrosynthetic Analysis in LHASA. 5. Searching a Pool of Starting Materials. *J. Chem. Inf. Comput. Sci.*, in press.