1831 (1959).

Mountain, R. D., "Spectral Distribution of Scattered Light in a Simple Fluid," *Revs. Mod. Phys.*, 38, 205 (1966).

Mountain, R. D., and J. M. Deutch, "Light Scattering from Binary Solutions," *J. Chem. Phys.*, 50, 1103 (1969).

Oliver, C. J., and E. R. Pike, "Laser Scattering Measurement of the Thermal Diffusivity of Organic Liquids Using a Fabry-Perot Etalon," *Phys. Letters*, 31a, 90 (1970).

Riedel, L., "Neue Wärmeleitfähigkeitsmessungen an Organischen Flüssigkeiten," *Chem. Ing.-Techn.*, 13, 321 (1951).

Searby, G. M., Ph.D. thesis, Oxford (1971).

Shaw, R., "Heat Capacities of Liquids," *J. Chem. Eng. Data*, 14, 461 (1969).

Touloukian, Y. S., P. E. Liley, and S. C. Saxena, "Thermal Conductivity, Nonmetallic Liquids and Gases," in *Thermophysical Properties of Matter*, vol. 3, Y. S. Touloukian and C. Y. Ho (eds.), The TPRC Data Series, IFI/Plenum, New York (1970).

Touloukian, Y. S., and T. Makita, "Specific Heat, Nonmetallic Liquids and Gases," *Thermophysical Properties of Matter*, vol. 6, Y. S. Touloukian and C. Y. Ho (eds.), The TPRC Data Series, IFI/Plenum, New York (1970).

Manuscript received April 12, 1973; revision received July 18 and accepted July 20, 1973.

Reaction Path Synthesis Strategies

A system is offered for the mathematical representation of the functional and structural features of organic molecules and their reactions. Two basic synthesis strategies, an antithetic method and a synthetic method, are defined and compared. The use of heuristic information in guiding synthesis is evaluated. A synthetic method which uses dynamic programming is presented and applied to the synthesis of bihelical DNA. Possible extensions to other classes of compounds are presented.

GARY J. POWERS
and
RUSSELL L. JONES

Department of Chemical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

SCOPE

The ability to invent chemical reactions which have desired properties is central to the practice of chemical engineering. Reactions are used to transform raw materials into desired products, to solve separation problems, and to heat and cool process streams. At present there exists very little in the way of systematic techniques to synthesize reaction paths which solve chemical processing problems. In this paper we review the current status of research in reaction path synthesis and present a new representation and search strategy for organic chemical reaction paths. Several of these techniques have been programmed for digital computers and allow the rapid search for optimal reaction paths.

CONCLUSIONS AND SIGNIFICANCE

It is possible, using a state-space representation, to generate systematically all possible reaction paths leading to a desired molecule. If evaluation data are available for the reactions involved, a dynamic programming search strategy can be used to select the optimum reaction path. This approach gives a much more systematic means for generating, evaluating, and selecting reaction paths. It also indicates that there now exists a straightforward way of teaching reaction path synthesis.

The invention of reaction paths with desired properties is central to the synthesis of chemical processing systems. Chemical reactions are used to transform raw materials into desired products, separate mixtures, and heat and cool process streams (Rudd, 1973; Powers, 1973; Lauer, 1952).

In order to put chemical synthesis into a more systematic format it is necessary to collect and classify the vast number of organic structures, functionalities, and reactions into a more useful system. It is also necessary to develop more formal methods for generating, evaluating, and selecting chemical reaction paths.

Such a system must be built around an operational definition of organic synthesis. Organic synthesis is concerned with the structure of molecules and the net transformations in structure which occur during chemical reaction. The current organization of knowledge about reactions is based on mechanism and is not particularly well suited for use in synthesis. (This classification, based on mechanism, undoubtedly contributes greatly to the confusion that engulfs students in introductory organic chemistry courses. They are taught mechanism and then asked to do synthesis.) The system selected for representing the states and operators which define chemical synthesis must be general so that new structures and reactions can be rapidly and easily included. Furthermore, the system should be defined so that all possible reaction paths, even those not presently known, could be generated.

Since synthesis is a goal-oriented activity, the system must include information on the context of the synthesis, that is, the target molecules and possible starting materials.

A system aimed at meeting these criteria is developed here and its application to the synthesis of bihelical DNA illustrated. A computer program, DINASYN, which is based on these concepts, is briefly described. It is hoped that this approach will have value in the area of industrial reactions. However, the detail of information required to evaluate industrial reaction paths is great, and this approach would therefore be best suited, initially, for small

R. L. Jones is at the University of California, Berkeley, California.

classes of compounds (that is, substituted aromatics, halogenated hydrocarbons, etc.) for which more detailed information exists.

This approach may also have future value in other areas, such as organizing organic structure and reactions for teaching synthetic industrial chemistry to engineers and other chemical synthesis oriented professionals.

STATE-SPACE REPRESENTATION OF ORGANIC SYNTHESIS

The state-space representation (Nilsson, 1971; Powers, 1972) is particularly well suited to the problem of organic synthesis. In problem-solving using the state-space representation, states and operators are defined. To define a state, it is necessary to select a set of descriptors which characterize the state. Thus an important part of any state-space problem formulation is the selection of some particular form of description for the states of the problem. Virtually any kind of data structure can be used to describe states. These include symbol strings, vectors, two-dimensional arrays, trees, and lists. The form of the data structure selected often bears a resemblance to some physical property of the problem being solved. In the case of organic synthesis the description must represent the structure—the molecular skeleton—and the functionality—the type and placement of functional groups on that skeleton. Several forms for describing structure and functionality have been advanced. Corey et al. (1969) have developed a technique whereby the graphical representation commonly used to describe molecules is automatically translated into a linked list. Hendrickson (1971) has developed a notation which represents molecules by a linear list of symbols. Several workers in the area of chemical documentation have developed similar linear codes suitable for computer use (Wiswesser, 1954).

The other part of the state-space representation is the definition of operators. Operators change one state into another. Thus they can be considered functions whose domain and range are sets of states. Actually they are partial functions since an operator may not be applicable in all states. The arguments of the operators are the descriptors of each state. The value of the function is the descriptor for the output state. Hence, operators simply transform one state into another. In organic synthesis, the operators are chemical reactions which change structure or functional groups, or both.

As an example of descriptors and operators for organic synthesis, consider the following approach due to Hendrickson (1971a):

Define a carbon site S_i by its four attachments:

R = σ bond to carbon
Π = π bond to carbon
H = bond to H or less electronegative atom
Z = bond to more electronegative atom ($Z = N, O, X, P, S$).

The valence or number of attachments to a carbon site is four, hence the sum of the types of attachments can be defined. Let:

σ = number of σ bonds to the site (the number of attached carbons),
π = the number of π bonds to carbon,
h = number of bonds to H or less electronegative atoms,
z = number of bonds to heteroatoms (N, O, X, P, S).

where

$$0 \leqq (\sigma, h, z) \leqq 4 \quad \text{and} \quad 0 \leqq \pi \leqq 2$$

Since π bonds can be considered as both a bond to carbon and a functionality like z, it is convenient to define the functionality f as the sum of functional heteroatom attachments (Z) and carbon-carbon π bonds (Π).

$$f = \pi + z$$

For every carbon site:

$$\sigma + \pi + h + z = \sigma + f + h = 4$$

The character (C) of a site is defined by Hendrickson (1971) as

$$C = 10\sigma + f$$

Hence C is a two-digit number (base 10) in which the first digit (σ) shows the skeletal nature (σ = 1, primary; 2, secondary; 3, tertiary; 4, quaternary) and the second (f) shows the functionality level (f = 1, alcohol, ether, halide, olefin; 2, ketone, aldehyde, acetylene; 3, carboxylic acid family, nitrile; 4, derivative of carbon dioxide) (Hendrickson Cram, and Hammond, 1970).

The oxidation state (x) at any single carbon site is further calculated by

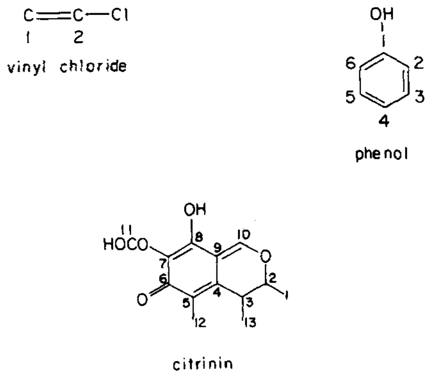
$$x = z - h = 2z + \sigma + \pi - 4$$

The connectivity and functionality of a structure can be defined by numbering its n atoms and giving the functionality of each site. This may be written as the site number with its character ($C = 10\sigma + f$) as a superscript. Each site number is linked by a dash to the next site number if they are bonded or by a slash if they are not.

Sites that are not attached sequentially to a given site are listed in parentheses after it. Bonds through heteroatoms are denoted by double slashes. This type of notation has full connectivity information and allows the functionality class of each site to be noted. Some examples are shown for vinyl chloride, phenol and citrinin in Figure 1. The exact functionality of each site and its heteroatoms could be saved in a separate list so that a complete description of the molecule would be possible.

For a reaction at a carbon site defined as above, an operator may be defined as the replacement of one descriptor by another.

Hence, S_iS_j is an operator which forms S_i as S_j is removed. Thus HZ is the replacement of a heteroatom by hydrogen, that is, reduction of a halide. $H\Pi$ indicates



vinyl chloride:	$1^{11} - 2^{12}$
Phenol:	$1^{22}(6) - 2^{21} - 3^{21} - 4^{21} - 5^{21} - 6^{21}(1)$
Citrinin:	$1^{10} - 2^{21} - 3^{30}(13) - 4^{31}(9) - 5^{31}(12) - 6^{22} -$ $7^{31}(11) - 8^{22} - 9^{31}(4) - 10^{12}/11^{13}(7)/12^{10}(5)/13^{10}(3)$

Fig. 1. Nomenclature examples.

saturation at one carbon of an olefin bond, either by hydrogenation, or by addition of HX, H_2O, etc. ΠZ is the formation of a π bond by the loss of a heteroatom as in dehydrohalogenation, etc.

With this definition there are 16 possible transformations that can occur at a single site. The reaction types and *symbols* are illustrated in Table 1.

The total change in all the carbon sites during reaction is obtained by listing the changes at each carbon site sequentially. For example, $H\Pi \cdot Z\Pi$ represents the addition of HBr to an olefin. The reactions representing all the possible reactions involving just two carbon sites are shown in Table 2. The first group involves the construction or cleavage of σ bonds between carbon. The second group involves only π bonds. The isohypsic term denotes reactions which involve no change in oxidation state ($\Delta X = 0$). The dehydrohalogenation of 1,2-dichloroethane to form vinyl chloride ($\Pi H \cdot \Pi Z$) is isohypsic.

For reactions involving more than one site the reactions are listed for each carbon site. For example, a common fragmentation is $\Pi Z \cdot \Pi R \cdot ZR$. Several characteristics of this system can be noted:

1. The order of writing the reaction pairs (that is, $H\Pi$ or ΠZ, etc.) does not change the sense of the overall reaction.

2. Changing the order of the symbols within a reaction pair reverses the reaction (ΠH is the reverse of $H\Pi$).

3. The reactions involving $\pm R$ (construction or cleavage) and $\pm\Pi$ (addition or elimination at π bond sites) must occur at adjacent sites.

4. Reactions involving two or more sites must always terminate in $\pm H$ and/or $\pm Z$ unless the sites are cyclic. The central (nonterminal sites) exhibit only $\pm R$, $\pm\Pi$ symbols.

The states and transformations for all the possible carbon sites can be expressed in a triangular graph with σ, f, and h axes. The graph is illustrated in Figure 2 and is constrained for $\sigma + f + h = 4$. The graph is plotted for the case where Π and Z are grouped into one class of functionality F ($f = \Pi + Z$).

Each vertex on the graph represents a carbon site type. The character $C = 10\sigma + f$ is shown below each site type. The edges between vertices represent lines of constant σ, f, or h. The $\sigma = 0$ family represents carbon sites that are not bonded by carbon to a larger carbon skeleton. Carbon sites with $\Pi = 1$ or 2, that is, alkenes or alkynes, can only appear at sites with $f \geq 1$ and $1 \leq \sigma \leq 3$. This structural limitation is shown on Figure 2 for alkene, alkyne, and aromatic carbons by shading of the appropriate regions.

Table 1. Reaction Symbol and Type at One Carbon Site (After Hendrickson, 1971)

Type	Symbol
I. Substitution	
Proton exchange	HH
Carbon interchange	RR
π Rearrangement	$\Pi\Pi$
Nucleophilic substitution	ZZ
II. Oxidation-reduction	
Oxidation	ZH
Reduction	HZ
III. Construction-cleavage	
Oxidative construction	RH
Reductive cleavage	HR
Reductive construction	RZ
Oxidative cleavage	ZR
Constructive addition	$R\Pi$
Fragmentation	ΠR
IV. Elimination-addition	
Oxidative elimination	ΠH
Reductive addition	$H\Pi$
Reductive elimination	ΠZ
Oxidative addition	$Z\Pi$

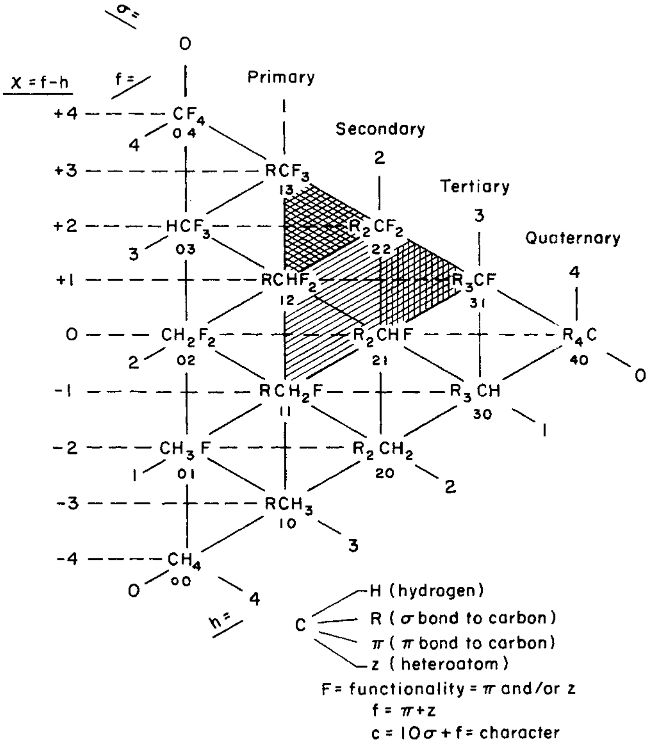


Fig. 2. Character triangle for carbon sites and interconversions The character C, is shown beneath each carbon site. The shaded areas indicate possible double bond sites (C = C),▨ ; possible triple bond sites (C ≡ C),▧; and possible aromatic sites▥.

Table 2. Reactions that Occur at Two Carbon Sites (Hendrickson, 1971)

		$+R$	$-R$	ΔX
σ bond construction and cleavage	Oxidative ($-H$ or $+Z$)	$RH \cdot RH$	$ZR \cdot ZR$	+2
	Isohypsic	$RH \cdot RZ$	$ZR \cdot HR$	0
	Reductive ($+H$ or $-Z$)	$RZ \cdot RZ$	$HR \cdot HR$	−2
		$+\Pi$	$-\Pi$	ΔX
π bond reactions	Oxidative ($-H$ or $+Z$)	$\Pi H \cdot \Pi H$	$Z\Pi \cdot Z\Pi$	+2
	Isohypsic	$\Pi H \cdot \Pi Z$	$Z\Pi \cdot H\Pi$	0
	Reductive ($+H$ or $-Z$)	$\Pi Z \cdot \Pi Z$	$H\Pi \cdot \Pi H$	−2

Table 3. Single Site Reaction Classification (Hendrickson, 1971)

Forward reactions			*f* classes*			Reverse reactions
Substitution ($\Delta C = 0$)	f_{44}	CO_2	⇄	CZ_4		
(FF)	f_{33}	COOH	⇄	CZ_3		Same
	f_{22}	$C = 0$	⇄	CZ_2		
	f_{11}	COH	⇄	CZ		
Reduction ($\Delta\sigma = 0$)	f_{43}	CO_2	⇄	HCOOH	f_{34}	Oxidation ($\Delta\sigma = 0$)
(HF)	f_{32}	COOH	⇄	CHO	f_{23}	(FH)
$\Delta C = -1$	f_{21}	$C = 0$	⇄	CHZ	f_{12}	$\Delta C = +1$
$\Delta f = -1$	f_{10}	CZ	⇄	CH	f_{01}	$\Delta f = +1$
Reductive construction	f_{43}	CO_2	⇄	RCOOH	f_{34}	Oxidative cleavage ($\Delta h = 0$)
($\Delta h = 0$)						
(RF)	f_{32}	COOH	⇄	$RC = 0$	f_{23}	(FR)
$\Delta C = +9$	f_{21}	$C = 0$	⇄	RCZ	f_{12}	$\Delta C = -9$
$\Delta f = -1$	f_{10}	CZ	⇄	CR	f_{01}	$\Delta f = -1$
Oxidative construction	f_{33}	HCN	⇄	R − CN	f_{33}	Reductive cleavage ($\Delta f = 0$)
($\Delta f = 0$)						
(RH)	f_{22}	CHO	⇄	$RC = 0$	f_{22}	(HR)
$\Delta C = +10$	f_{11}	CHZ	⇄	RCZ	f_{11}	$\Delta C = -10$
	f_{00}	CH	⇄	CR	f_{00}	

* Simple generalizations of the transformations (usually with oxygen groups) but any heteroatom may be replaced by another without changing the *f* class (for example, f_{33} ≡ HCN, HCOOR, etc). Unlabeled bonds are to *R* or *H*, but not to Z.

Table 4. Two Carbon Site Reactions: Classification and Enumeration (After Hendrickson, 1971)

	Reaction type	Total* number of operators	Trivial** operators	Significant operators
Construction (or cleavage)	$RH \cdot RH$	55	−10	45
	$RH \cdot RF$	100	−16	84
	$RF \cdot RF$	55	−10	45
		210	−36	174
Elimination (or addition)	$FH \cdot FH$ ($\pi H \cdot \pi H$)	21	−6	15
	$FH \cdot FF$ ($\pi H \cdot \pi Z$)	36	−9	27
	$FF \cdot FF$ ($\pi Z \cdot \pi Z$)	21	−6	15
		78	−21	57

* All possible transformations for both sites combined, from Figure 2.

** Trivial transformations are those which combine two one-carbon units into a C_2 unit (that is, $RH \cdot RH$, $RH \cdot RF$, $RF \cdot RF$).

There are 70 possible transformations for single carbon sites represented in Figure 2. These transformations define the operators which change the state of molecules during synthesis. These reactions are further classified in Table 3. The classification and enumeration for two site reactions is given in Table 4.

The classifications define all the existing and potential chemical reactions in terms relevant to synthesis. It is also interesting to note that several reactions suggested by this classification have received very little if any recognition in the present literature (Hendrickson, 1971). One example is the oxidative coupling of aldehyde derivatives (RCHO + R′CHO → R—CO—CO—R′); $\Delta X = +2$ or reductive coupling of ketones with a CO_2 derivative ($CO_2 + R_2CO \rightarrow R_2C(OH)COOH$; $\Delta X = -2$).

With the classification given in Figure 2, it is possible to generate all the possible routes to any given type of site.

The total number of reaction paths to a site can be determined by graph theory. A 15 × 15 adjacency matrix *A* can be constructed for the 15 site types defined on Figure 2. The elements in the matrix are 1 or 0 depending on whether the sites are directly linked (1) or not (0). The sum of a row (or column due to symmetry) is the number of single paths to the site corresponding to that row (or column). The elements in the squared matrix (A^2) represent the number of binary paths (length = 2 reaction steps) between any two types of carbon sites. The sum of any row or column in A^2 gives the total number of binary paths leading to that site type. In general, the number of paths of length *n* (that is, *n* reaction steps) to a given site type is given by the sum of the row (or column) corresponding to the site in the matrix A^n. Table 5 gives the number of single and binary paths for Figure 2.

Table 5. Number of Reaction Paths on Figure 2.0* (After Hendrickson)

Group	Sites	Number of sites	Single paths	Binary paths
I	00, 04, 40	3	2	6
II	01, 03, 10, 13, 30, 31	6	4	12
III	02, 20, 22	3	4	16
IV	11, 12, 21	3	6	22
		15	16	56

* Trivial paths (that is, $A \rightarrow D \rightarrow A$) have been omitted.

The trivial binary paths have been omitted from Table 5 even though they can have chemical significance when used as blocking, protecting, or activating groups which are later removed.

This scheme can be used to determine the total number of ways to produce a given carbon site. For example, a carboxylic acid has the character $C = 13$. Single paths to it must then arise from sites with the character 03, 04, 12, or 22. Transformations from sites 03, and 04 are reactions involving the attachment of single carbons: 03 site carbons being HCN or CN— and the 04 sites being CO_2, $COCl_2$, or derivatives varying F. The other two represent aldehyde (12)oxidation (reaction class FH) (or oxidation of another $C = 12$ derivative such as $C \equiv CH$, $> C = CHOR$, etc.) and cleavage (FR) of $C = 22$, that is, ketones and their derivatives. The number of single paths to a secondary amine ($C = 21$; group IV) is six. Hence if we wished to make an α-amino acid utilizing one step to reach each site type, there would be $4 \times 6 = 24$ different reaction paths—four for the carboxylic acid group and six for the secondary amine for a total of 24 combinations.

The total number of synthesis paths to even a simple molecule such as vinyl chloride can be surprisingly large. Figure 3 illustrates the structure of vinyl chloride and its connectivity list. The total number of last single site steps in any synthesis of vinyl chloride is the sum of the operators applicable to each of its two sites: Carbon — 1 ($C = 11$), 6 paths; Carbon — 2 ($C = 12$), 6 paths, for a total of 12 last operator applications in the synthesis of vinyl chloride. Figure 3 illustrates a few of these last steps. Precursor types of the same functionality as the target sites have been included in Figure 3.

By applying all possible operators to each site in the molecule, it is possible to generate all the precursors to the target. Similarly, if all possible operators are applied to each precursor it would be possible to determine all the possible precursors two applications removed from the target. In this way, the complete synthesis tree could be generated (Corey, 1969). Of course, the tree will be very large for even simple molecules.

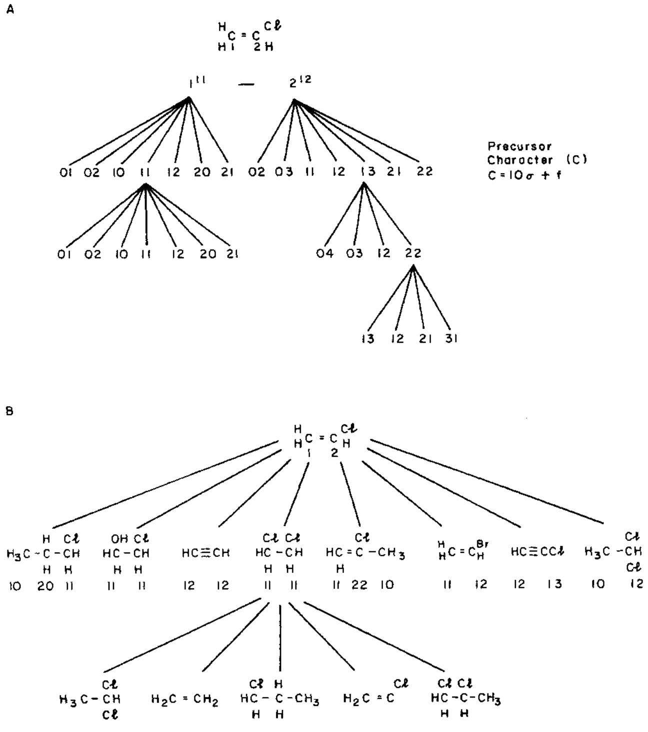


Fig. 3. A. Synthesis tree for vinyl chloride in which only the character of precursors is given. B. A partial synthesis tree for vinyl chloride for various starting materials.

Evaluation and Selection of Reaction Paths

Given that the above method could be used to generate all possible synthesis paths to a target molecule, the problem remains of selecting the best pathway. In order to select the best pathway, it is necessary to define a measure of merit (I) which can be used to compare the candidate reaction paths. What constitutes a proper merit figure depends greatly on the environment in which the synthesis is to be carried out. Very different criteria are used to evaluate potential reactions for laboratory synthesis than are used in the evaluation of industrial reactions (Rudd, Powers, and Siirola, 1973). In the laboratory the cost of reagents is not usually important, low yields (<50%) are acceptable, many solvents may be used, sophisticated separation techniques are available, and hazards associated with the materials can be minimized by using small amounts of material in safe enclosures. The range of pressure readily available in the laboratory is also confined to slightly above atmospheric to moderate vacuum. Hence, laboratory syntheses are usually evaluated on the basis of the time required to perform the synthesis and the reliability of the synthetic operations (ambiguity of products).

If the reliability of a particular reaction path can be estimated it would be possible to determine how much more time it would require to sufficiently resolve the ambiguity. Hence the reliability of a pathway can be expressed in terms of a time requirement. The use of time as a merit measure is discussed in more detail in the following DNA synthesis example.

In industrial reaction path evaluation the measure of merit commonly is cost, in dollars. The major determinants of the costs associated with a given reaction path are:

1. Product price
2. Raw material costs (yield)
3. Waste disposal costs (byproducts)
4. Capital for equipment
   a. Reactors (kinetics/thermodynamics)
   b. Separators (separation difficulties)
   c. Heaters, coolers (energy balancing)
   d. Pumps, compressors
5. Utilities
   a. Steam
   b. Electricity
   c. Fuel
6. Labor
7. Insurance and Safety.

In order to accurately assess the capital costs, it is necessary to perform a rather detailed design of the chemical processing system. Several techniques for automatically designing these systems have been developed (Siirola, 1971). It is feasible, however, to screen reaction paths with a merit measure based solely on the properties of the reaction path itself, (for example, the product price, raw material costs, byproducts produced, heats of reaction, number of reactions, etc). The exact form of the merit measure will also depend on the conventions used for economic analysis. One crude evaluation function commonly used to screen reaction-paths is

$$I = \text{Price of Product} - \text{Cost of Raw Materials} - \alpha \text{ (number of reaction steps)}$$

where α is a weighting coefficient.

The objective is then to select the reaction path which maximizes I.

Simply stated,

$$\underset{S_i,\ O_k}{\text{Max}}\ (I) \tag{1}$$

where S_i is the set of possible starting materials and O_k is the set of operators (chemical reactions).

The evaluation of the raw material costs depends on the cost of the raw material and the yield of the reaction step. Hence, in order to evaluate reaction paths it will be necessary to predict reaction yield. The yield of a reaction depends on the kinetics of competing reaction paths and the thermodynamic equilibrium for the reaction system. There is no formal method to accurately predict these yields. There have been several attempts to correlate experimental data. Linear free energy relationships (LFER) have provided a quantitative means for correlating relative activities. This approach has become a well-defined section of physical organic chemistry. Wells (1968) reviews the application of LFER in organic chemistry. LFER's have been developed to predict the effects of reagent structure, functionality, and reaction medium. LFER's have been developed for aromatic substitution reactions, acid dissociation, hydrolysis of substituted aliphatic esters, ketals, and acetyls, saponification of esters, acetates, and alkyl benzoates, and substitution reactions of trans-3-substituted acrylic acids and 2-substituted maleic acids.

The prediction of yields in heterogenously catalyzed reactions is very difficult and little exists in the way of formal methodology. Hence the outlook for predicting reaction yields is not good at this time. It is possible, however, for small families of compounds, that is, substituted aromatics, or halogenated aliphatics, to correlate data on reactivity. This approach has been used in the prediction of yields for the DNA reactions discussed in the final example.

If the yield is to be used in preliminary screening of reaction paths, it may only be necessary to determine if the yield is high ($> 90\%$), medium ($\sim 50\%$), or low ($< 30\%$). It is possible to correlate a greater amount of reactivity data with these classes of yield.

Note also that for certain reactions the raw materials will cost more than the product is worth even if 100% yields are obtained. In these cases, the prediction of yield is not necessary.

Synthesis Strategies

Given the above statement of the problem, it is necessary to define a search strategy for determining the optimum synthesis reaction path. There are three primary methods for searching for the optimal reaction path (Powers, 1972):

1. Antithetic (working backwards) method
2. Synthetic (working forwards) method
3. Hybrid method.

The antithetic method (Corey, 1969) has been proposed as an effective method for generating and searching for optimal reaction paths. Ireland (1969) has indicated that it is the central issue in effective organic synthesis. He states:

> "If there is any key to success in planning a synthesis, it is to work the problem backwards. This is really the cardinal rule of synthesis. Beginning with the total concept of the molecule desired and all of its structural ramifications, we methodically break it apart, piece by piece, in such a way that we can best predict success in reassembling the pieces. We tackle each problem as it is presented and work toward the solution that will leave the resulting synthetic tasks simpler than the ones just solved. When we have successfully unraveled the complex molecular network and proposed how each component may be rejoined in its turn to reconstruct our objective, the synthetic plan is in hand. In designing a synthesis, we must think back from the complex to the simple so that, in practice, we may rationally work from the simple to the complex with a suitable map to follow."

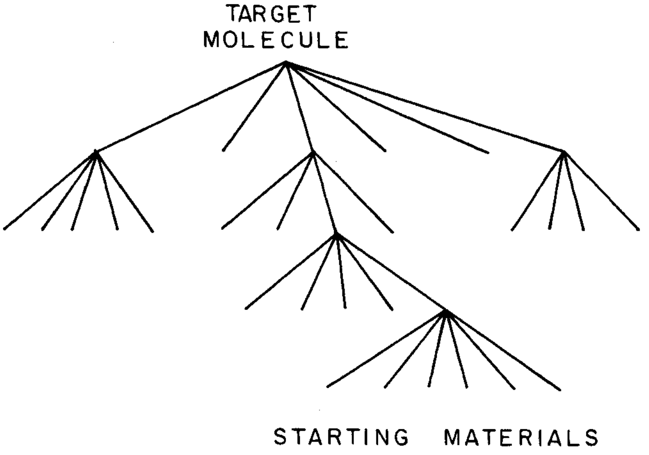


Fig. 4. A partial synthesis tree.

The main problem with the antithetic method is that if we work backwards, generating all possible precursors the breadth of the synthesis tree becomes overwhelming. If we try to evaluate the reaction paths as we work backwards, it is clear that complete and accurate evaluation is not possible. In order to evaluate reaction paths, it is necessary to know all the steps in the paths. When working backwards, complete intermediate evaluation is not possible since the complete synthesis paths have not yet been generated. Heuristic techniques are commonly invoked to reduce the breadth of the synthesis tree. Heuristics are rules of thumb which have, in the past, proved to be useful means for reducing the scope of the search. The rules do not guarantee optimal solutions in a mathematical sense. Rules based on the simplicity of intermediates have been advanced to reduce the size of the search (Corey, 1969, 1972a).

The antithetic approach has the advantage that the intermediate states (molecules) are easy to generate. Merely applying the operators to the target molecule generates the intermediates. Furthermore, every reaction path leads to the target molecule.

The size and shape of a typical synthesis tree is shown in Figure 4. In working backwards, it is necessary to reduce the scope of the synthesis to the point where evaluation of the remaining paths is feasible. Notice in Figure 5 how the number of intermediates decreases as the depth of the synthesis is extended to more and more primitive starting materials. The number of intermediates would indicate that the application of heuristics should be dependent on the level in the synthesis. More stringent heuristics that eliminate possible intermediates are required in the middle of the tree generation to produce a manageable search.

The synthetic method (Powers, 1972b) solves the synthesis problem by working forwards from the starting materials to the target molecule. In other words, the starting materials are transformed, by repeated application of the operators, into the target molecule. This planning is done in the same direction as the synthesis is executed. This approach has one great advantage:

The search for the optimum path can be guided by the

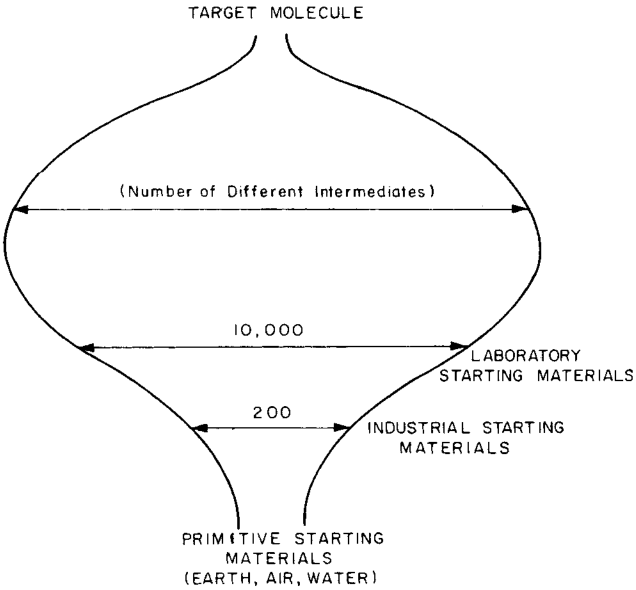

TARGET MOLECULE
(Number of Different Intermediates)
10,000
LABORATORY STARTING MATERIALS
200
INDUSTRIAL STARTING MATERIALS
PRIMITIVE STARTING MATERIALS (EARTH, AIR, WATER)

Fig. 5. The number of intermediates in organic synthesis.

principle of optimality.

The principle of optimality (Bellman, 1957) is based on the intuitively obvious principle that

> An optimal set of decisions has the property that whatever the first decision is, the remaining decisions must be optimal with respect to the outcome which results from the first decision.

In molecular synthesis, the key is the fact that the best way to produce any molecule is to select the optimum way of carrying out the immediately preceding reaction on the best of all possible precursors. For example, if we are using time to evaluate synthesis pathways, the minimum time to produce any molecule is made up of the minimum times to join two groups which have also been made in the minimum time. Hence, if we are going to determine the optimum way to make any given molecule, we must know the optimum way to produce its precursors. The only place to start this process is at the starting materials.

This search technique, also called dynamic programming, has been used in a wide range of decision problems in inventory control, material allocation, process control, and chemical process design. An application of dynamic programming to distillation column sequencing by Hendry (1972) is very similar to the method used here.

The major disadvantages of the synthetic method are in defining the correct sets of starting materials, and forcing the reaction paths to converge on the target molecule. The first problem can be partially solved by defining starting materials in a general fashion. The classification scheme given in the previous sections is very useful in defining general types of starting materials.

The second problem, of forcing the reaction paths to converge on the target molecule, can be solved by focusing on the carbon skeleton. Operators are applied to starting materials to produce molecular skeletons which are contained in the target molecule.

The synthetic or working-forward method portends a powerful method for selecting optimum synthesis reaction paths. The technique is capable of handling very large synthetic problems. An example of applying this approach to the synthesis of bihelical DNA will be given in the next section.

It is possible to combine the antithetic and synthetic methods into a hybrid technique. Corey (1972) has suggested a method in which the synthesis planning works backwards from the target and forwards from the starting material, meeting in the middle of the synthesis tree.

COMPUTER-AIDED SYNTHESIS OF DNA

An example of the synthetic (working forwards) method applied to the synthesis of bihelical deoxyribonucleic acid is now given. This example was chosen to illustrate the power of the dynamic programming search method in solving very large synthesis problems. The linear nature of the DNA molecules also makes the representation of the molecule much easier than for general organic molecules. Only two different classes of reaction (operator) are used in this type of work further simplifying the synthesis problem.

The chemical synthesis of bihelical DNAs of specific nucleotide sequence by Professor H. Gobind Khorana and his coworkers has lead to unique studies of biological processes such as DNA replication and transcription. The availability of such DNA molecules also suggests possible means for controlling genetic diseases (Khorana, 1972).

Structure of DNA

Deoxyribonucleic acid is a chain of nucleotides joined together by a deoxyribose sugar chain. There are four of these nucleotides formed from the bases adenine, cytosine, guanine, and thymine. Cytosine (C) and thymine (T) belong to the pyrimidine molecular class (an aromatic compound of the formula $C_4H_4N_2$ or a derivative) while adenine (A) and guanine (G) are purines (an aromatic compound of the formula $C_5H_4N_4$ or a derivative). The

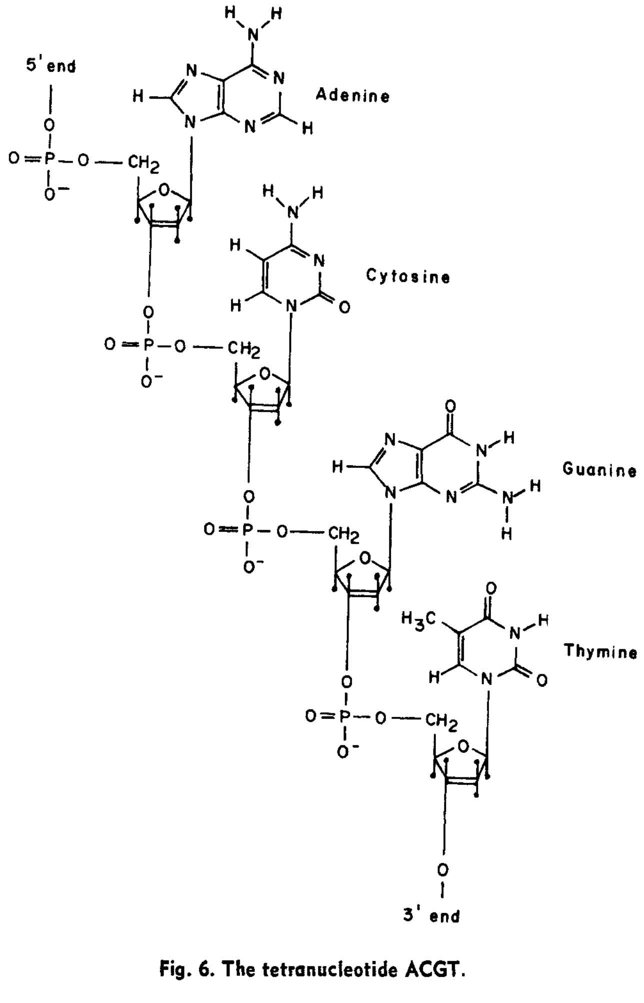


Fig. 6. The tetranucleotide ACGT.

structure of the tetranucleotide ACGT is shown in Figure 6. One should note the chain is highly directional. When considering a single strand of nucleotides, the naming convention is to start the naming from the 5′ end.

The most important structural feature of DNA is that it consists of two long nucleotide chains twisted about each other in the form of a regular double helix, as shown in Figure 7. The diameter of the helix is about 20Å and the helix makes a complete turn every 34Å (or 10 nucleotide lengths). The two chains are held together by hydrogen bonds between pairs of bases. This results in the complementary arrangement between the two chains—A is always paired with T and C is always paired with G. These are the only arrangements possible because two purines occupy too much space and two pyrimidines occupy too little space to allow formation of a regular helix. Figure 8 shows the hydrogen bonds present in the base pairs. Bihelical strands have opposite polarity so the custom is to have the top chain listed from 3′ to 5′ while the second chain follows the usual convention. Numbering of the duplex chains, however, is from right to left. A schematic representation of the yeast alanine DNA sequence is given in Figure 9.

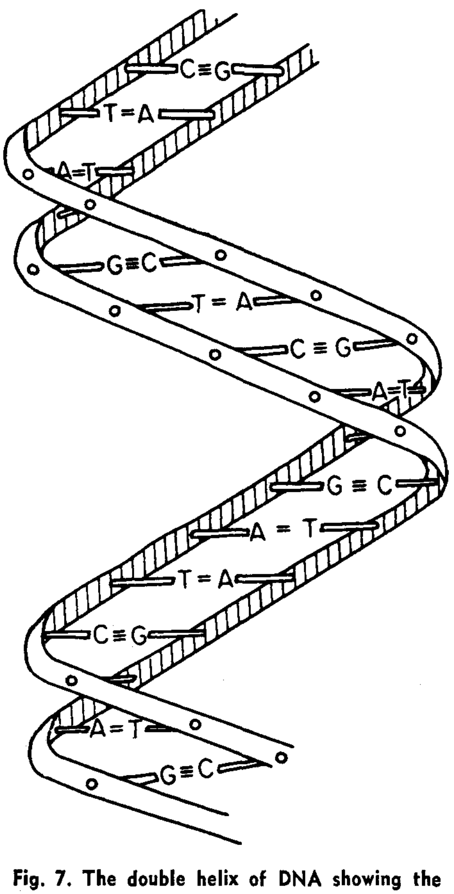


Fig. 7. The double helix of DNA showing the phosphate sugar ribbons on the outside and paired bases bridging the gap between the two strands.

The synthesis problem is to find the optimal way to make a specific DNA from the mononucleotides A, C, G, and T.

The chemical synthesis of DNA molecules is a very complex and time consuming process. The synthesis of the DNA for a yeast alanine transfer ribonucleic acid (RNA), which is shown in Figure 9, required approximately 20 man years of effort by Dr. Gobind Khorana and his research staff. The complexity of the target DNA molecules and the large number of synthetic steps required to produce them demands efficient synthesis strategies.

Although the appearance of the yeast alanine DNA seems simple, many possible pathways from raw materials to finished product exist. Table 6 gives the number of possible reaction paths to form a single strand of DNA as the length of the strand increases. Figure 10 illustrates the possible paths for a tetranucleotide.

Synthesis Procedures

There are three major operators that the synthesis strategist can employ to transform the four mononucleotides A, C, G, and T (from the bases adenine, cytosine, guanine, and thymine, respectively) into a finished DNA strand such as the alanine sequence. These operators are the protection of mononucleotides, single strand condensation, and duplex joining.

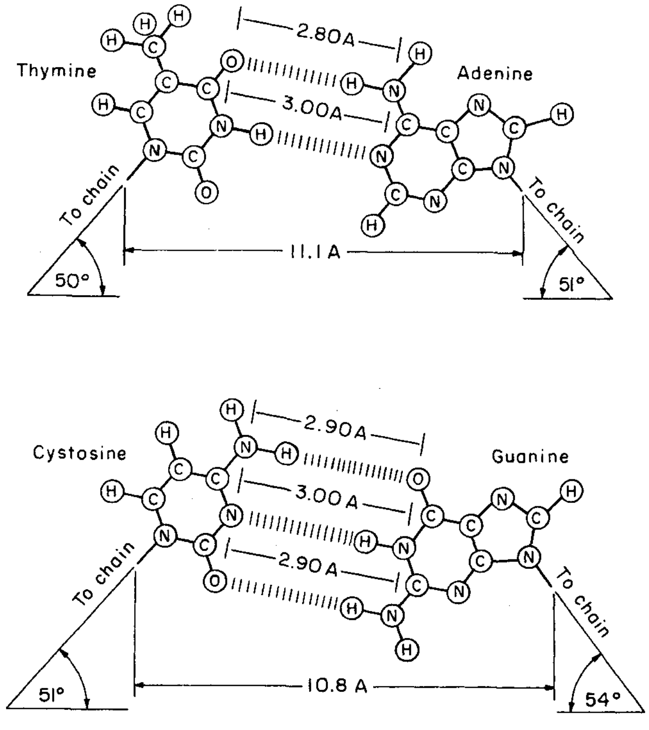


Fig. 8. The hydrogen bonding properties of adenine-thymine and guanine-cytosine base-pairs.

```
77777777666666666655555555554444444444333333333322222222221111111111
76543210987654321098765432109876543210987654321098765432109876543210987654321

CCCGCACACCGCGCATCAGCCATCGCGCGAGGGAATCGTACCCTCTCAGAGGCCAAGCTAAGGCCTGAGCAGGTGGT
GGGCGTGTGGCGCGTAGTCGGTAGCGCGCTCCCTTAGCATGGGAGAGTCTCCGGTTCGATTCCGGACTCGTCCACCA
```

Fig. 9. Schematic representation of the nucleotide sequence for the structural gene for an Alanine transfer ribonucleic acid from yeast. The numbering is from 1 to 77.

Table 6. Total Number of Reaction Paths for Forming a Nucleotide Sequence of a Given Length

Length	Number of paths
2	1
3	2
4	5
5	14
6	42
7	132
8	429
9	1,430
10	4,862
11	16,796
20	10^{10}
50	10^{27}
100	10^{50}
150	10^{84}
200	10^{113}
500	10^{286}
1,000	10^{575}

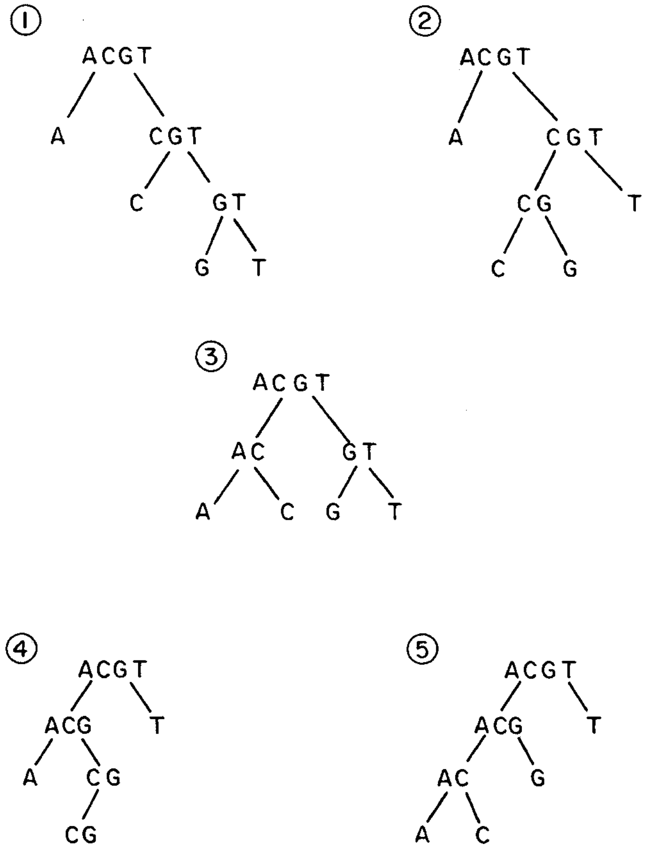


Fig. 10. Five reaction paths for the synthesis of a tetranucleotide.

$A + C \longrightarrow AC$

$AT + C \longrightarrow ATC$

$AGCTACCT + GTACT \longrightarrow AGCTACCTGTACT$

Fig. 11. Examples of single strand condensations.

The first step in DNA synthesis is to protect the amino functions in the heterocyclic ring in the A, C, and G bases by reaction with a standard acylating agent such as benzoyl chloride or anisoyl chloride.

The second step is to synthesize a set of single strands by combining the mononucleotides A, C, G, and T to form polynucleotide chains. Examples of single strand condensations are shown in Figure 11. Each step involves three reactions:

1. Blocking the 5′-hydroxyl group of the nucleotide being joined at the 3′ end with a bulky acid sensitive trityl group or by cyanoethylation. A trityl group is not removed during the following procedures and need not be replaced at each step. However, a cyanoethyl group must be replaced each time. A trityl is used only for blocking a 5′ end that also forms the end of the target single strand.

2. Blocking the 3′-hydroxyl group of the nucleotide (that has the 5′ end that is attacking the other reagent) with alkali-labile group such as esters.

3. Condensation of suitable protected and blocked nucleotides using activating agents such as dicyclohexylcarbodimide and aromatic sulphonyl chlorides.

The final step is to join the single strands together to form the final product. This involves base pairing of unprotected single strands of nucleic acids to form overlapping chains in a bihelical duplex followed by the joining of these chains using the enzyme ligase. Examples of duplex joining reactions are shown in Figure 12.

Evaluation Function

With this definition of states and operators, it is necessary to select an appropriate evaluation function so that the optimal reaction path can be determined.

The evaluation of reaction paths introduces several problems. First, what are the criteria for evaluation? Second, how can these criteria be combined to yield a single evaluation criterion (or do we have a vector optimization problem)? Third, can we predict, in general, the values of these criteria?

For the synthesis of DNA molecules, it is quite clear that the activities involved in synthesis can be broken into the areas of reaction, separation, and analysis.

The evaluation criteria that are pertinent to each of these areas are:

1. Reaction
   a. Yield of product
   b. Cost of raw materials
   c. Time for reaction
2. Separation
   a. The Yield of separation
   b. The Purity of product
   c. The Time for separation
3. Analysis
   a. The Ambiguity of analysis
   b. The Time for analysis

Of these criteria the most dominant is time. All of the others can be converted into time units. For example, if low reaction yields are encountered they can be overcome by running larger batches which require slightly more time. Hence the time value of yield can be determined. Similarly if a reaction is not reliable it will require a longer time, on the average, to run successful batches. If a difficult separation is encountered due to small differences in charge between the species being separated, then it will require more time to effect the required separation. The additional time may be consumed in operating the separator more slowly or by rerunning the partially separated mixture. In analysis it may require several analyses

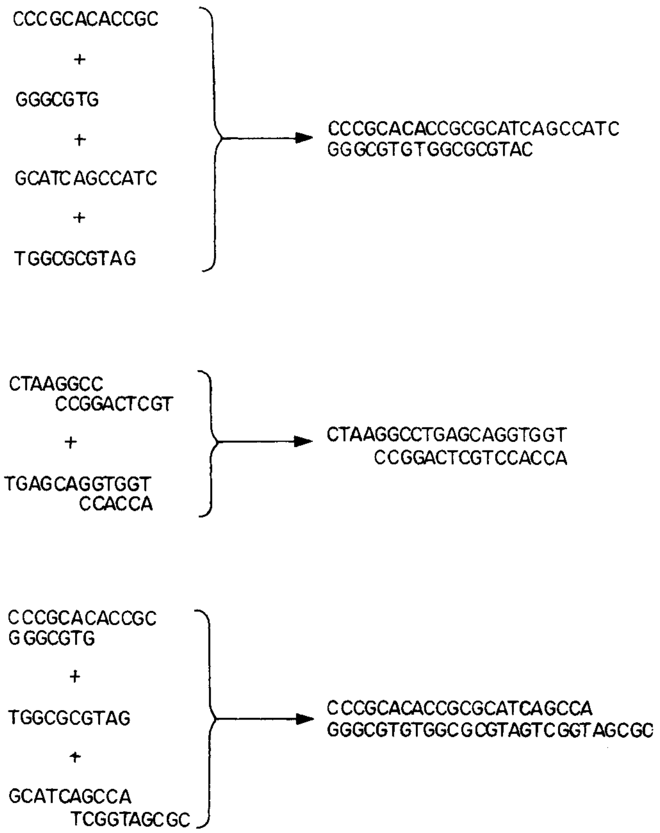


Fig. 12. Examples of duplex joining reactions.

to determine the exact nature of an ambiguous compound. Hence a greater amount of time is needed. For operations which are impossible, it is possible to consider essentially an infinite amount of time to be required.

Hence a uniform means for evaluating nearly all the features of a DNA synthetic pathway can be determined if only the time required to achieve the desired synthesis is considered.

Mathematical models for predicting the time required for the reaction, separation, and analysis steps in DNA synthesis have been developed (Randall, 1972; Powers, 1973).

The synthesis problem is to find the reaction path from mononucleotides to target molecule which minimizes time. The total synthesis of the alanine gene required approximately 20 man-years of laboratory efforts using the synthesis path developed by Khorana et al. (1972).

DINASYN

A computer program DINASYN (DeoxyrIboNucleic Acid SYNthesizer) was written by Jones (1972) for determining the optimal reaction path to any general DNA molecule. The program is based on a dynamic programming approach to the search for optimal synthesis paths.

Working the synthesis problem forward (dynamic programming approach) involves two major steps. The first is to generate all possible subgroups from length one to the final molecule size. Then one must evaluate the cost of making each subgroup, starting with the cost of the raw materials (the four protected mononucleotides).

Initially all the dinucleotides are considered. Optimum reaction conditions, separation time, and analysis time are determined for each of these groups. Next the times for the trinucleotide groups are calculated, optimizing over the joining location (which two subgroups are joined to form the target group). The joining point which gives the lowest time is selected as optimum, and this time is used in the next level of optimization. The evaluation of the subgroups continues until the evaluation of the target molecule (the last subgroup) is completed.

It should be noted that the optimization carried out for all groups of three or more nucleotides is an optimization over the total cost of the reaction sequence. This is possible since the subgroups used in these reactions have already had their optimum synthesis path determined. These minimum subgroup reaction path times are added to the time of the last reaction to form the total time, which must be minimized to give the minimum cost joining point and optimal reaction path for each group. This step by step build-up of optimal reaction subpaths is continued until eventually the optimal synthesis path is generated.

An example of the subgroups and the reaction evaluations necessary for each subgroup optimization in the sequence TCAGCA is shown in Table 7. Note that only 34 evaluations are required compared to 210 (42 paths times 5 reaction evaluations per path) for an exhaustive search procedure. Computational savings greatly increase as the length of the target strand increases.

The DINASYN program has been applied to a large number of different DNA synthesis problems (Powers, 1973). The results have indicated reductions in total synthesis time of 20% to 50% (depending on the size of the target molecule) over the synthetic plans developed by synthetic chemists using a heuristic approach. One brief example follows.

Part of the synthesis path used by Dr. Khorana and his co-workers is illustrated in Figure 13. The DINASYN program predicts a total synthesis time of 40,000 hours or ~ 20 man-years of effort (2000 hr./man-year) for the complete synthesis path. This is very close to the actual

Table 7. Example of the Searching Procedure Using the Dynamic Programming Approach

A total of 34 reaction evaluations is necessary

TACGAC			
T + ACGAC TA + CGAC TAC + GAC TACG + AC TACGA + C			
TACGA	ACGAC		
T + ACGA TA + CGA TAC + GA TACG + A	A + CGAC AC + GAC ACG + AC ACGA + C		
TACG	ACGA	CGAC	
T + ACG TA + CG TAC + G	A + CGA AC + GA ACG + A	C + GAC CG + AC CGA + C	
TAC	ACG	CGA	GAC
T + AC TA + C	A + CG AC + G	C + GA CG + A	G + AC GA + C
TA	AC	CG	GA
T + A	A + C	C + G	G + A
T	A	C	G
START	START	START	START

time required for the synthesis. (This is not very remarkable since this synthesis was used to develop part of the mathematical models in DINASYN.)

What is interesting, however, is the optimal synthesis path to the alanine gene developed using the DINASYN program and shown in part in Figure 14. The total time for this path is 19,000 hours, for a reduction of approximately ten man-years, or 50%.

These kinds of reductions have been found possible in a number of other synthesis plans. This program is now in routine use to plan the synthesis of bihelical DNA molecules.

Work is now underway to extend these concepts to more general classes of organic molecules.

In conclusion, then, a systematic procedure has been developed for defining the states and transformations which are useful in organic chemical synthesis. The system is particularly well suited to the needs of organic synthesis. The functionality and structure of organic species are defined so that all possible sites and transformations are considered. With this system, it is possible to generate all possible reaction paths to any given target molecule. Several search strategies were defined for selecting the optimum reaction path. A synthetic method based on a dynamic programming approach was shown to have several advantages over an antithetic (working backwards) approach to synthesis planning. An example of the dynamic programming approach was illustrated for the synthesis of bihelical DNAs of specific nucleotide sequences.

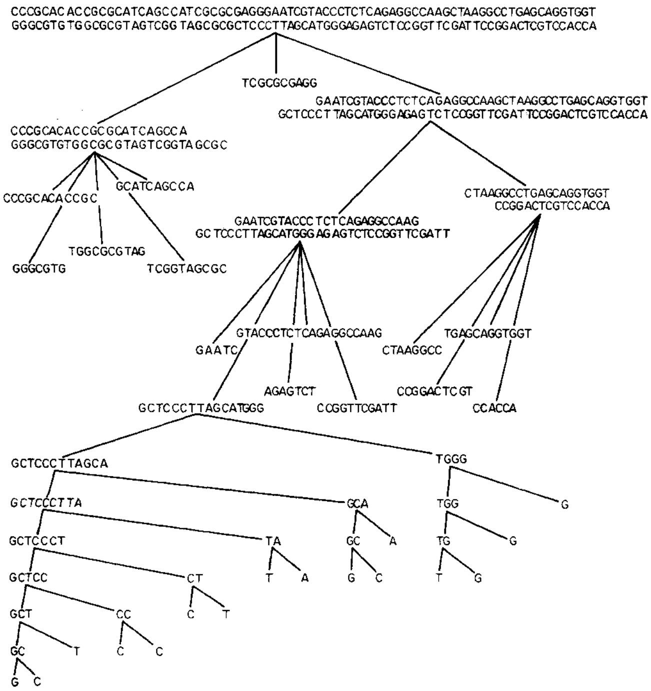

CCCGCAC ACCGCGCATCAGCCATCGCGCGAGGGAATCGTACCCTCTCAGAGGCCAAGCTAAGGCCTGAGCAGGTGGT
GGGCGTG TGGCGCGTAGTCGG TAGCGCGCTCCCTTAGCATGGGAGAGTCTCCGGTTCGATTCCGGACTCGTCCACCA
TCGCGCGAGG
GAATCGTACCCTCTCAGAGGCCAAGCTAAGGCCTGAGCAGGTGGT
GCTCCCT TAGCATGGGAGAGTCTCCGGTTCGATTCCGGACTCGTCCACCA
CCCGCACACCGCGCATCAGCCA
GGGCGTGTGGCGCGTAGTCGGTAGCGC
GCATCAGCCA
CCCGCACACCGC
TGGCGCGTAG
GGGCGTG
TCGGTAGCGC
CTAAGGCCTGAGCAGGTGGT
CCGGACTCGTCCACCA
GAATCGTACCC TCTCAGAGGCCAAG
GC TCCCTTAGCATGGG AG AGTCTCCGGTTCGATT
GTACCCTCTCAGAGGCCAAG
GAATC
AGAGTCT
GCTCCCTTAGCATGGG
CCGGTTCGATT
TGAGCAGGTGGT
CTAAGGCC
CCGGACTCGT
CCACCA
GCTCCCTTAGCA
TGGG
GCTCCCTTA
GCA
TGG
G
GCTCCCT
TA
GC
A
TG
G
GCTCC
CT
T
A
G
C
T
G
GCT
CC
C
T
GC
T
C
C
G
C

Fig. 13. Part of the synthesis path used by H. G. Khorana in the synthesis of a gene for alanine t-RNA from yeast.

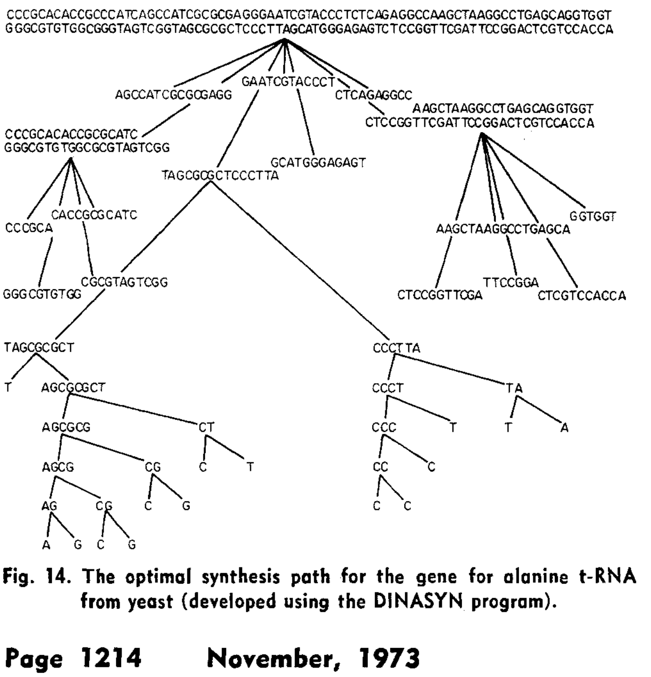


Fig. 14. The optimal synthesis path for the gene for alanine t-RNA from yeast (developed using the DINASYN program).

LITERATURE CITED

Bellman, R., *Dynamic Programming*, Princeton University Press, New Jersey (1957).

Corey, E. J., and W. T. Wipke, "The Computer-Assisted Synthesis of Complex Organic Molecules," *Science*, **166**, 179 (1969).

Corey, E. J., W. Todd Wipke, R. D. Cramer III, and W. J. Howe, "Computer-Assisted Synthetic Analysis: Facile Man-Machine Communication of Chemical Structure by Interactive Computer Graphics," *J. Am. Chem. Soc.*, **94**, 421 (1972a).

———., "Techniques for Perception by a Computer of Synthetically Significant Structural Features in Complex Molecules," *ibid.*, 431 (1972b).

Corey, E. J., R. D. Cramer III, and W. J. Howe, "Computer-Assisted Synthetic Analysis for Complex Molecules. Methods and Procedures for Machine Generation of Synthetic Intermediates," *ibid.*, 440 (1972c).

Corey, E. J., and G. A. Petersson, "An Algorithm for Machine Perception of Synthetically Significant Rings in Complex Cyclic Organic Structures," *ibid.*, 460 (1972d).

Hendrickson, J. B., "A Systematic Characterization of Structures and Reactions for Use in Organic Synthesis," *J. Am. Chem. Soc.*, 93, 6847 (1971a).

———., "Synthesis Design for Substituted Aromatics," *ibid.*, 6854 (1971b).

Hendrickson, J. B., D. C. Craur, and G. S. Hammond, *Organic Chemistry*, 3rd edit., McGraw-Hill, New York (1970).

Hendry, J., and R. Hughes, "Optimal Synthesis of Separation Systems," *Chem. Eng. Progr.*, **68**, No. 6, 20 (1972).

Ireland, R. E., *Organic Synthesis*, Foundations of Modern Organic Chemistry Series, Prentice-Hall, Englewood Cliffs, N. J. (1969).

Jones, Russell, "Computer-Aided Synthesis of DNA," Masters thesis, Mass. Inst. Technol., Cambridge (1972).

Khorana, H. G., et al., "Studies on Polynucleotides CIII, The Total Synthesis of the Structural Gene for an Alanine Transfer Ribonucleic Acid from Yeast," *J. Am. Chem. Soc.*, **94**, 1120 (1972).

Lauer, B. E. and R. F. Hechman, *Chemical Engineering Techniques*, Reinhold, N. Y. (1952).

Nillson, N. J., *Problem-Solving Methods in Artificial Intelligence*, McGraw-Hill, 1971.

Powers, G. J., "Heuristic Synthesis in Process Development," *Chem. Eng. Progr.*, **68**, No. 8, 88 (1972a).

———., "Non-Numerical Problem Solving Methods in Computer-Aided Design," Proc. IFIPS Conf. on Computer-Aided Design, Eindhoven, N. V. (1972b).

Powers, G. J., R. Jones, M. Caruthers, H. van de Sande, and H. G. Khorana, "DNA Synthesis Strategies," *Am. Chem. Soc.*, submitted.

Randall, George, "Reaction Models for DNA Synthesis," Masters thesis, Mass. Inst. Technol., Cambridge (1972).

Rudd, D. F., G. J. Powers, J. J. Siirola, *Process Synthesis*, Prentice-Hall, Englewood Cliffs, N. J. (1973).

Siirola, J. J., and D. F. Rudd, "Computer-Aided Synthesis of Chemical Process Designs," *Ind. Eng. Chem. Fundamentals*, **10**, 353 (1971).

Wells, P. R., *Linear Free Energy Relationships*, Academic Press, London (1968).

Wiswesser, W. J., *A Line-Formula Chemical Notation*, Crowell, New York (1954).

Manuscript received May 29, 1973; revision received July 19 and accepted July 20, 1973.