This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6” x 9” black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
STUDIES IN COMPUTATIONAL BIOCHEMISTRY:
COMPUTER PREDICTION OF XENOBIOTIC METABOLISM AND
THE THREE-DIMENSIONAL SOLUTION STRUCTURE OF
RESIDUES 1-28 OF THE ALZHEIMER'S DISEASE
AMYLOID BETA-PEPTIDE

by

JOSEPH TALAFOUS

Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy

Thesis Advisor: Gilles Klopman

Department of Chemistry
CASE WESTERN RESERVE UNIVERSITY
January 1995
We hereby approve the thesis of

Joseph Taliesins

candidate for the Ph. D.
degree.*

(signed)  

(chair)  

Michael B. Zogorski

date 8/26/94

*We also certify that written approval has been obtained for any proprietary material contained therein.
STUDIES IN COMPUTATIONAL BIOCHEMISTRY:
COMPUTER PREDICTION OF XENOBIOTIC METABOLISM AND
THE THREE-DIMENSIONAL SOLUTION STRUCTURE OF RESIDUES 1-28
OF THE ALZHEIMER'S DISEASE AMYLOID BETA-PEPTIDE.

Abstract

by

JOSEPH TALAFOUS

META is a new knowledge-based expert system that simulates the
biotransformation and metabolism of xenobiotics. By application of a collection of
rules which supply knowledge, the program recognizes key functional groups within
the complete xenobiotic structure and substitutes them to predict possible
metabolites. META employs established methodologies to predict lipophilicity,
stability, and reactivity of the metabolites. A comprehensive knowledge base was
constructed to model mammalian xenobiotic metabolism with META methodology
by consulting experts in medicinal chemistry. Metabolic pathways are often
experimentally observed to compete for substrates, which is modeled by the
prioritization of the rules. The META-4 algorithm optimizes the prioritization
knowledge base development. Another algorithm, called Graphsort, is used to assign a unique name to chemical entities that can be represented by graphs. Uniqueness is accomplished by canonical ordering of the nodes using novel node invariants based on distance distributions.

The β-peptide is the major proteinaceous component of amyloid deposits in Alzheimer's disease. The three-dimensional solution structure of residues 1 to 28 of the amyloid β-peptide was determined using nuclear magnetic resonance spectroscopy, distance geometry, and molecular dynamics techniques. The nuclear magnetic resonance data used to derive the structure consisted of nuclear Overhauser enhancements, vicinal coupling constants, and temperature coefficients of the amide-NH chemical shifts. In membrane-like media, the peptide folds to form a predominately α-helical structure with a bend centered at residue 12. There is strong backbone homology with mellitin and alamethicin, which may provide a structure-based explanation to the β-peptide ion-channeling and aggregation properties. The side-chains of histidine-13 and lysine-16 reside on the same face of the α-helix and their proximity may constitute a binding site for the heparan sulfate proteoglycans. The molecular details of this structure could assist in the design of rational treatments to curtail the binding of heparan sulfate proteoglycans, or inhibit an α-helix → β-sheet conversion that may occur during the early stages of amyloid plaque formation in Alzheimer's disease.
This dissertation is dedicated to my family:
Judith, James, Julie, Joseph Sr., and Anne
whose love and care made this possible
and to dear ol' Rose.

To Kevin Dale Krause, in memorium.
Preface

As a rule, new medicines are usually developed only with great expense. Rational drug design uses information from diverse disciplines to reduce this expense. Design efforts traditionally have been focused on the determination of structure-activity relationships of molecular structures that are mainly described by their two-dimensional connectivity. However, it is becoming increasingly clear that this is not a complete picture. The bioactivity attributed to a single structure may actually be due to one or more of its metabolites or it may originate from its three-dimensional shape. The characterization of such chemical entities is a common thread running through this dissertation.

Another theme is the prevalent use of computers. One might argue that the simple problems in biological chemistry have already been solved. The remaining complex problems have so many parts with subtle interactions that good software and fast computers are needed to solve them. Thus, it is critical to continually develop powerful computer methodologies that reflect the needs of such research.
Acknowledgment

Above all, I thank Professor Gilles Klopman for the best education available and Mario Dimayuga, Ph.D. for many helpful discussions.

I thank those who contributed to the content and production of this dissertation:

A. B. Anderson, Ph. D.** J. J. Mieyal, Ph.D.*
D. Balthasar, Ph. D. R. Pearlman, Ph. D.
A. Brünger, Ph. D. D. S. Ptchelintsev, Ph. D.
A. M. Chaka, Ph. D. C. J. Rooney
D. A. Demeter, Ph. D. L. M. Sayre, Ph.D.*
D. Fercu, M. S. R. Scaringe, Ph.D.
J. M. Fleischer, Ph. D. C. Simmonds
A. M. Gelormini S. Srivastava, Ph.D.
R. Henderson, Ph. D. W. P. Stratten, Ph.D.
A. L. Hoffmeyer H. Ton-That, Ph.D.
M. Homyak, M. D. S. Wang, Ph.D.
J. Jacob, M. S. E. L. Wang, Ph.D.
J-Y. Li M. G. Zagorski, Ph.D.*
X. Liu, Ph. D. R. C. Zaniewski
G. McMillan, Ph. D.* Z. Zhang

** Committee chairman
* Committee member
# Table of Contents

Abstract .................................................................................................................. i

Dedication ............................................................................................................... iii

Preface .................................................................................................................. iv

Acknowledgment .................................................................................................. v

Table of Contents ................................................................................................... vi

Table of Figures ..................................................................................................... vii

Table of Tables ....................................................................................................... xii

**CHAPTER 1: Basic META Methodology** ............................................................ 1

  INTRODUCTION ...................................................................................................... 2

  METHOD .................................................................................................................. 4

    Problem analysis .................................................................................................. 4

    Relative biological activity ............................................................................. 5

    Tissue distribution of metabolic enzymes .................................................. 5

    Inter-species and inter-individual variations ............................................. 6

  Basic META methodology .................................................................................. 6

  Management of metabolites ............................................................................ 11

  IMPLEMENTATION ................................................................................................. 12

  DISCUSSION .......................................................................................................... 14

  CONCLUSION .......................................................................................................... 17

**CHAPTER 2: Modeling Mammalian Metabolism with META Methodology** .... 18

  INTRODUCTION ...................................................................................................... 19

  METHOD .................................................................................................................. 20

  THE DICTIONARY .................................................................................................. 22

    Phase 1 biotransformations ........................................................................... 24

    Phase 2 biotransformations ........................................................................... 32

  EXERCISING THE DICTIONARY .......................................................................... 36

  CONCLUSIONS ........................................................................................................ 39

vii
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity theory and analysis of algorithms</td>
<td>109</td>
</tr>
<tr>
<td>Graph invariants</td>
<td>112</td>
</tr>
<tr>
<td>Node invariants</td>
<td>113</td>
</tr>
<tr>
<td>Differentiation of nodes</td>
<td>116</td>
</tr>
<tr>
<td>Distance distributions</td>
<td>118</td>
</tr>
<tr>
<td><strong>METHOD</strong></td>
<td>120</td>
</tr>
<tr>
<td>Development cycle of the algorithm</td>
<td>120</td>
</tr>
<tr>
<td>Testing graph theoretic conjectures experimentally</td>
<td>122</td>
</tr>
<tr>
<td>The GraphSort algorithm</td>
<td>123</td>
</tr>
<tr>
<td>Propagation via distance distributions</td>
<td>126</td>
</tr>
<tr>
<td>Cycle distributions</td>
<td>129</td>
</tr>
<tr>
<td>Geodesics</td>
<td>131</td>
</tr>
<tr>
<td>Demonstration of differentiation of nodes</td>
<td>131</td>
</tr>
<tr>
<td>Final reordering</td>
<td>138</td>
</tr>
<tr>
<td>Implementation details</td>
<td>138</td>
</tr>
<tr>
<td><strong>RESULTS</strong></td>
<td>139</td>
</tr>
<tr>
<td>Performance with easy and difficult graphs</td>
<td>139</td>
</tr>
<tr>
<td>Average performance</td>
<td>141</td>
</tr>
<tr>
<td><strong>DISCUSSION</strong></td>
<td>141</td>
</tr>
<tr>
<td>Comparison to other algorithms</td>
<td>141</td>
</tr>
<tr>
<td>Concerns for robustness</td>
<td>143</td>
</tr>
<tr>
<td>Concerns for efficiency</td>
<td>147</td>
</tr>
<tr>
<td>Concerns for aesthetics of final line code</td>
<td>148</td>
</tr>
<tr>
<td><strong>CONCLUSION</strong></td>
<td>149</td>
</tr>
</tbody>
</table>

**CHAPTER 6: The Three-dimensional Solution Structure of Residues 1-28 of the Alzheimer’s Disease β-peptide**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INTRODUCTION</strong></td>
<td>152</td>
</tr>
<tr>
<td><strong>MATERIALS AND METHODS</strong></td>
<td>155</td>
</tr>
<tr>
<td>Sample preparation</td>
<td>155</td>
</tr>
</tbody>
</table>

$\chi$
Procedures for NMR data acquisition and analysis .......... 135
Distance geometry and simulated annealing computations ... 157

RESULTS ............................................................................. 159
NMR measurements ..................................................... 159
Three-dimensional structure computations ...................... 160
Consistency of the structure with the NMR constraints ....... 161

DISCUSSION ....................................................................... 167
Physiological relevance of the solution conditions .......... 167
Relationship between bioactivities and structures of β-Peptide... 168
General features of the β-(1-28) structure ......................... 171
Homologies with ion-channel forming peptides ............... 171
Exploring the homology with simple modeling ............... 174
Binding motif of proteoglycans and other proteins .......... 177
Fate of β-peptide in vivo .................................................... 178

EPILOGUE ......................................................................... 181

BIBLIOGRAPHY ................................................................. 184

APPENDIX ......................................................................... 223
Table of Figures

1. P-450 hydroxylation of N-methyl-N-ethyl-nitrosamines ........................................ 7
2. Example of a spontaneous reaction ................................................................. 9
3. Metabolites of N-methyl-N-ethyl-nitrosamine ............................................... 10
4. An interactive session with META .................................................................. 13
5. Major primary metabolic pathways of imipramine .......................................... 14
6. Some major secondary metabolic pathways of imipramine ............................ 15
7. Example of the generalization abilities of META methodology .................... 23
8. Metabolites of phenacetin ................................................................................. 36
9. Major metabolites of phenanthrene .................................................................. 45
10. Epoxidation of aromatic compounds .............................................................. 49
11. Nucleophilic internal hydride shift .................................................................. 50
12. Consensus mechanism for hydroxylation by P-450 ......................................... 51
13. Examples of tautomersisms which require quantum mechanical treatment ...... 52
14. Relevant tautomersisms .................................................................................... 53
15. Parameterization of hydroxylation and epoxidation ......................................... 57
16. Key to the quantum mechanical indices .......................................................... 62
17. Correlation of $\Delta H_{rxn}$ with the partial bond character ................................ 63
18. Example of an actual META transform ........................................................... 70
19. Training set of xenobiotics used in the demonstration of META-4 ................. 71
20. Synopsis of basic META methodology ........................................................... 73
21. How to represent prioritization of transforms with directed graphs ............... 80
22. Outline of the META-4 algorithm .................................................................... 82
23. Digraph of literature database corresponding to training set of Figure 19 ...... 84
24. Examples of different types of graphs .............................................................. 98
25. An example of a spatial pharmacophore .......................................................... 99
26. Example of coding aniline by three different routes ....................................... 102
27. Example of an isomorphism and an automorphism ........................................ 106
28. The graph isomorphism problem .................................................................... 109

xii
29. Flowchart of testing a version of Graphsort .......................................... 119
30. Summary of methodology to uniquely name a graph ............................. 121
31. Flowchart of node differentiation .......................................................... 123
32. Example of calculation of two types of distance distributions ............... 127
33. Example of calculation of the cycle distribution ..................................... 129
34. Example of the use of geodesics ............................................................. 130
35. Demonstration of node differentiation with a simple graph ...................... 134
36. 1-2. Demonstration of node differentiation with a difficult graph ............. 136
37. Results of route tracing with isonicotinic acid ...................................... 138
38. Overview of the formation of β-peptide ............................................... 153
39. Portion of a NOESY contour plot of the amide-amide region .................. 156
40. The final accepted family of 95 structures of β-(1-28) peptide .............. 163
41. The average rmsd of Figure 40 as a function of residue ......................... 165
42. Abbreviated β-(1-28) structure emphasizing charge distribution ............ 169
43. Overlap of three ion-channeling peptides ............................................. 172
44. Electric dipole caused by side chains at physiological pH ...................... 173
45. Gross packing in analogy to the mellitin dimer ..................................... 174
46. Formation of the tetramer .................................................................. 175
47. Hypothetical model of aggregation of β-peptide ..................................... 176
48. β-peptide in cross β-pleated sheet conformation .................................... 177
49. Charge complementarity of cross β-pleated sheet conformation ............. 178
50. Hypothesis for the overall fate of the β-peptide .................................... 179
# Table of Tables

1. Training set of xenobiotics ......................................................... 22
2. Phase 1 biotransformations ......................................................... 29
3. Phase 2 biotransformations ......................................................... 33
4. Major biotransformations of test compounds ............................ 38
5. Major routes of biotransformations ............................................ 39
6. Epoxidation/hydroxylation training set ..................................... 58
7. Quantum mechanical indices for enol-keto tautomerisms ............... 60
8. Quantum mechanical indices for isoaamide-amide tautomerisms ....... 61
9. Quantum mechanical indices for enamine-imine tautomerisms ........ 61
10. Example literature database .................................................... 75
11. Hypothetical examples of substrate binding lists ....................... 76
12. Contingency table used to evaluate substrate binding lists ............ 77
13. Evaluation of substrate binding with many transforms .................. 79
14. Contingency table used to evaluate masking of binding lists .......... 88
15. Input parameters and results of application of META-4 ............... 90
16. Prioritization results of the demonstration training set ................ 91
17. 1-3. Representative test structures for Graphsort ..................... 145
18. Build-up times of selected peaks ............................................. 160
19. Characterization of the final accepted 95 structures .................... 162
20. Comparison of primary sequences of mellitin, alamethicin, and β-(1-28) ..... 170
CHAPTER 1

Basic META Methodology
INTRODUCTION

The ever-increasing use of chemicals for industrial solvents, cosmetics, food additives and preservatives, herbicides and pesticides, drugs and medicines has brought about a tremendous increase in the possible routes to exposure of humans, animals, and crops to xenobiotics (i.e., chemicals foreign to normal physiology and metabolism). The effects of xenobiotics on the organism may be caused not only by the parent agent, but also by its metabolites which are the result of biotransformations. Accordingly, methodologies have been developed that are capable of elucidating the metabolic pathways of xenobiotics. In conjunction with quantitative structure activity relationships (QSAR), the exact chemical entities causing the therapeutic or toxic effects attributed to the parent agent can be uncovered.

Xenobiotics can be absorbed into the organism through many routes. In the case of environmental xenobiotics (e.g., pollution), the routes are likely to be inhalation or direct contact with the skin, rather than oral ingestion or intravenous injection. Once absorbed, the organism must eliminate the xenobiotic, otherwise it would accumulate to toxic levels. Elimination can occur by successive biotransformations until the unwanted xenobiotic is excreted. In mammals, most xenobiotics are converted concurrently or consecutively to multiple metabolites by hepatic microsomal and cytosolic enzymes. Other tissues (e.g., plasma, kidney, lung, and the gastrointestinal tract) also contribute significantly to the biotransformation of xenobiotics. However, the same metabolic pathways may

---

* This chapter is based on an accepted publication in the Journal of Chemical Information and Computer Sciences.
actually cause the xenobiotic to become much more deleterious to the organism. For example, relatively low levels of enzyme activity in the skin cells may contribute significantly to local metabolic toxic activation.

In many cases, it has been possible to identify the bioactive metabolites of particular chemical agents and, based on this knowledge, to predict the metabolic conversion and bioactivity of analogous compounds. In a prospective sense, this approach has led to the development of successful pro-drugs (inactive parent drug/therapeutic metabolite), and the avoidance of certain types of drug interactions and toxicity.

As new chemical agents are designed for the variety of purposes alluded to above, there is an increasing need to predict accurately and completely the nature and bioactivity of their metabolites. Hence, the Klopman laboratory is developing computer software (called META) that predicts the metabolic products of xenobiotics along with the possible bioactivity of these metabolites. Its knowledge originates from an expert system evaluation of their chemical structure which does not require prior knowledge about the actual metabolism of the compound.

Some progress has been accomplished by researchers in the computer-assisted prediction of xenobiotic metabolism, but the literature does not offer much background in this area. Indeed, while certain pharmaceutical companies may have engaged in such a task, (e.g. Upjohn Company), they have not published their results in any substantial detail, nor have they made their software publicly available. To our knowledge, only one computer program is commercially available that claims to predict metabolites. However, there are a number of programs developed for the purpose of computer-assisted synthesis planning which may employ principles similar to those used in META.
This chapter outlines the basic operation of the META program. Subsequent chapters describe the development of the knowledge base employed by META to actually predict the metabolic transformation of chemicals.

**METHOD**

*Problem analysis*

The major objective in the development of this program is to provide an interactive means of predicting possible metabolic transformations of any given compound, regardless of whether its metabolites are experimentally known to exist.

A sequential approach was used to generate the stable metabolites, taking into consideration the fact that experimentally observed metabolites may be the result of several intermediate metabolic steps. Knowledge of biotransformation steps is translated to computer-readable rules and stored in a *dictionary* file. Each rule expresses distinctive chemical structural information relevant to the specificity of each metabolic biotransformation. In addition, a dictionary of spontaneous reactions was constructed to detect and process unstable intermediates generated by any individual biotransformation step.

To construct an effective dictionary, the major types of reactions involved in the biotransformations of xenobiotics must be identified. Fortunately, there is ample research and literature devoted to this task which can be condensed to rules describing the essential activity of each reaction type. META must be able to systematically and exhaustively apply these rules and display the primary, secondary, tertiary, etc. potential metabolites of a parent compound, usually leading to a series of metabolites with various pharmacological importances.
For certain applications, such as when predicting the identity of an undetectable trace toxic species, the ability to predict very minor metabolites will be required. In most cases, however, it will only be necessary to classify the metabolites initially according to their relative rates of formation and susceptibility to further metabolism (i.e., their relative abundance and rate of elimination). The classification will be derived from "hard data" obtained from the pharmacological literature concerning the relative likelihood of a given biotransformation and concentrations of the metabolites.

Relative biological activity

The relatively abundant metabolites can be further subdivided according to their potential for biological activity using various criteria, for example:

- therapeutically active
- inactive and excretable
- inactive and retained
- potentially toxic

In this mode, the user will be alerted to these special cases. For example, certain compounds that are conjugated with glutathione and taken up by the kidney may become renal toxins by the action of the \( \beta \)-lyase enzyme, rather than normal excretion eventually as mercapturic acid conjugates.\(^\text{13}\) Certain metabolites excreted into the intestine via the bile as glucuronide conjugates may be cleaved by \( \beta \)-glucuronidase, reforming the parent compound, and followed by reabsorption.

Tissue distribution of metabolic enzymes

Different local relative concentrations of metabolites will arise in different tissues depending upon the route of administration, oil/water partition coefficient,
protein binding, and active transport of the xenobiotics in question. Moreover, the relative amounts of cytochrome P-450 isozymes and phase 2 conjugation enzymes vary from tissue to tissue not only with respect to their constitutive levels, but also with respect to their relative inducibility and degree of repression caused by exposure to various chemical agents. Thus a complex decision tree was developed to rank the importance of potential metabolites relative to specific bioactivities.

Inter-species and inter-individual variations

The majority of the literature data on metabolic pathways and relative abundance of metabolites is for animal species. Numerous examples are known of differences in relative abundance of primary metabolites, relative activities and inducibility of enzymes, etc. among the animal species and humans.\textsuperscript{13} Thus, the initial steps of literature excavation and dictionary development will be based upon similarities among the species, but finer-detailed prediction for a particular species will require scrutiny of the differences specific to that species. The prediction of genetically-based inter-individual variations in metabolism, such as the polymorphisms associated with N-acetylation\textsuperscript{14} and certain P-450 mediated oxidations,\textsuperscript{15} will be treated similarly.

Basic META methodology

The objective of the META program is to determine the potential metabolites that could be produced from the molecular structure of a xenobiotic entered into the program. The molecular structures can be entered by a variety of methods. Simple and complex molecules can be entered through either KLN line code,\textsuperscript{16} SMILES line code\textsuperscript{17} (Chapter 5, p. 103), direct graphical input with a mouse or through a MOL file. The META program was developed to be consistent with another
Figure 1. One possible pathway of P-450 hydroxylation of N-methyl-N-ethyl nitrosamine.

program developed in the Klopman laboratory, the Computer Automated Structure Evaluation (CASE) program."\(^{18}\)

META operates in conjunction with a variety of dictionaries, each containing biotransformation information, compiled independently by the program MEDIC, that specify recognition sites on xenobiotic substrates, and the biotransformations that the virtual enzyme will catalyze. Each dictionary models a particular metabolic model, for example, individual animal species or different organs and pharmacological compartments.

The target fragment is the molecular fragment believed to be recognized by an actual enzyme or enzymes, and the product fragment replaces the target fragment, thus simulating biotransformation. Each target-product fragment pair is called a transformation operator or simply a transform, which can be viewed as a "virtual enzyme."

The product of a biotransformation may be an unstable intermediate and will react spontaneously. To model this phenomenon, another type of dictionary is consulted that contains transforms that model spontaneous reactions. The operation is repeatedly applied to the product until the program is satisfied that the resulting molecule is stable, that is, no transforms in the spontaneous dictionary apply.
Let us follow an example of a sequence of biotransformations. Figure 1 illustrates a pathway followed by the mammalian P-450 metabolism of N-methyl-N-ethyl nitrosamine. The CH2N group of 1 is recognized as a target fragment by a P-450 transform. The product fragment is CH(OH)N. The transform that encodes this biotransformation is:

Find: \( \text{CH2-N} \) -
Replace with: \( \text{CH -N - <1-OH>} \)

This transform might be entitled "the hydroxylation of an aliphatic carbon \( \alpha \) to a nitrogen atom." The molecule resulting from this transform is 2, CH3CH(OH)N(N=O)CH3.

The spontaneous reactions dictionary (MESP.DIC) will then be consulted and it will recognize that the CH(OH)NN=O moiety is intrinsically unstable and should be rearranged into CH=O and N=N-OH yielding 3a, CH3N=NOH and acetaldehyde, 3b. This rearrangement is modeled by the following transform:

Find: \( \text{OH -CH -N -N =O} \)
Replace with: \( \text{O =CH N= N -OH} \)

It should be noted that since this rearrangement is carried out transparently, the net effect of the original hydroxylation appears to the user as N-dealkylation.

A facility is needed to recognize the instability in a molecule and to transform it to the corresponding molecular structure. The dictionary of spontaneous reactions partially satisfy this need. For example, if by some metabolic
process a hydroxyl group is attached to a carbon atom already bonded to another hydroxyl group, the program must be able to recognize that the molecule is intrinsically unstable and is a candidate for spontaneous rearrangement, as shown in Figure 2.

Further consultation of the spontaneous dictionary will indicate that $3a$ is still inherently unstable and another transform is used to produce $4a$, CH$_3^+$; $4b$, N$_2$; and $4c$, H$_2$O. Since CH$_3^+$ is reactive, the process is repeated until no further unstable products are detected. The application of spontaneous reaction rules operates in the background by default and only the stable products are displayed. However, unstable intermediates can be displayed if desired. Upon request for further biotransformation, the program will submit user-selected products for further processing and will continue to accumulate the products into a metabolic pathway map until it is determined that the final products can be excreted.

The program maintains a history of all the metabolic steps previously taken, and will upon request, identify the enzyme class (reaction type), title of the transform, relevant literature references, and a graphic depiction of the transform used to generate any of the metabolites. Duplications are culled from the product list by comparing every new metabolite to those already generated so as to simplify metabolite management as far as possible. In this way metabolic cycles may be discovered.
Figure 3. Metabolites of N-methyl-N-ethyl nitrosamine as produced by the META program.

Figure 1 shows the metabolic fate of N-methyl-N-ethyl nitrosamine when the ethyl moiety is initially hydroxylated. Since the ethyl and methyl groups can be hydroxylated using slightly different transforms, the analogous products of methyl hydroxylation are also produced by the program as shown in Figure 3. The carbonium ions, $\text{CH}_3^+$ and $\text{CH}_2\text{CH}_2^+$ are seen to result from this process. Consulting a third dictionary (CARC.DIC) of fragments that are known to elicit biological activity (e.g. structural attributes suggesting genotoxicity) enables the
program to issue a warning that the input molecule may be carcinogenic because it produces a metabolite suspected to be conducive to cancer.

**Management of Metabolites**

A straightforward application of transforms such as illustrated in Figure 3 will yield, even for very simple molecules, a large number of possible products. Furthermore, with the ability to perform multiple metabolic steps, each of the products generated from the parent structure is available for further metabolism, resulting in more products. To manage the combinatorial explosion of information generated by the program, the following operations are performed.

First, a prioritization of the products by pharmacological significance identifies the major and minor products (Chapter 4). Unless otherwise requested, minor products are not generated thereby reducing the number of metabolic pathways.

Second, duplicate structures at each metabolic step are eliminated. The recognition of duplicates is done by storing and comparing various graph invariants of increasing complexity, but this may be accomplished by Graphsort (Chapter 5) eventually. A simple filter using the total number of atoms and the number of hydrogens is first performed. If there is a match with a previously encountered structure, then the molecular weights are compared. If there is still a match, the determinant of a modified adjacency matrix is compared. The modification to the standard adjacency matrix is to use an index for each atom representing the atom type, hybridization, and number of nonhydrogen ligands, e.g., \( \text{CH}_3(\text{sp}^3) = 4 \), \( \text{CH}_2(\text{sp}^3) = 5 \), \( \text{CH}(\text{sp}^3) = 6 \), \( \text{C}(\text{sp}^3) = 7 \), \( \text{CH}_2(\text{sp}^2) = 8 \), etc. as the diagonal element.
This method is effective in detecting and eliminating most duplicates from the potentially large number of structures generated by the program.

Third, a criterion for terminating further biotransformations was developed. Essentially, each product is evaluated for excretion by the kidneys. At each metabolic step, the octanol/water partition coefficients for the new metabolites are estimated (Chapter 3, p. 56). Compounds with values under the predetermined limit (currently 0.0) are presumed to be eliminated. A number of programs capable of estimating log P for diverse molecules developed by the Klopman laboratory,\textsuperscript{19,20} are used in conjunction with this task.

Fourth, the development of the numerous transforms includes cataloging relevant literature references and the archetypal reactions and substrates. This data is made available by the program on demand. Hence, the applicability of a given transform to a particular situation may be evaluated by the user independently.

**IMPLEMENTATION**

The dictionary will be discussed in detail in Chapter 2. The user interface has been fully developed for the VAX platform. Figure 4 shows the computer screen of the implementation of the actual META program. The program is interactive, allowing the users to explore metabolic pathway possibilities of interest. Major and/or minor metabolites can be displayed. Hard copies of the results can be produced automatically.
Figure 4. An example of the computer screen during an interactive session with META. The first window (upper left) is the metabolite management window where the xenobiotic is entered into the program.* A typical session may generate many metabolites, so the pathways explored are organized into a tree-like structure. Notice the menu options along the top of the window. If menu option "Display" is selected, the highlighted structure "U3" (unstable intermediate) appears in a separate window (lower left). The metabolites of each step can be displayed in another window (lower right). If menu option "Rules" is selected, the transform rule that generated each step is listed and the target fragment is highlighted in the structures.

* For further implementation details, see Dimayuga, M. A. L. Ph. D. Thesis, Case Western Reserve University, 1991.
DISCUSSION

Imipramine, a commercially important antidepressant drug, provides a good example of the potential complexity and extent of xenobiotic biotransformation as illustrated in Figures 5 and 6. Its metabolites display a remarkable spectrum of physicochemical and pharmacological properties. Imipramine has become the classic model for the study of complex drug metabolism situations and their
pharmacokinetic and pharmacodynamic implications. Differences in the metabolism of imipramine are a major cause of variability in clinical response among individuals.

Figure 5 shows the major primary routes of metabolism of this relatively simple molecule in mammals. A large fraction of a dose of imipramine is N-dealkylated to the therapeutically active product desipramine by the cytochrome P-450 system. Imipramine undergoes aromatic hydroxylation to 2-OH-imipramine, which produces high cardiac toxicity and is responsible for teratogenic effects in rabbits. Toxicity effects are of particular concern because depressed patients have a propensity to overdose their medication. Imipramine is also N-oxidized by the hepatic flavoprotein, amine oxidase, to imipramine-N-oxide.

There are notable inter-species differences in the metabolism of imipramine. N-oxidation is the major initial reaction in pigs, while N-dealkylation predominates in rats and guinea pigs. Desipramine is rapidly metabolized in mice and rabbits, but accumulates in rats and humans.

As shown in Figure 6, the metabolic profile for imipramine expands broadly as secondary and subsequent pathways are considered. Desipramine undergoes
N-dealkylation and aromatic hydroxylation. The production of amine and alcohol functionalities in phase 1 metabolism allow *O-glucuronidation* in phase 2 metabolism, the products of which are the major excretion components in the urine. Figure 6 is not complete in that minor secondary metabolites are not shown. Indeed, more than 24 different metabolites have been detected and identified.

It is important to note that the N-dealkylation transform described earlier in this paper for N-methyl-N-ethyl nitrosamine also applies to imipramine. This same transform can remove the side chain of desipramine. Furthermore, a minor variant of this transform, i.e., the recognition of CH$_3$N as well as CH$_2$N as potential hydroxylation targets, can be used for the demethylation of both imipramine and 2-OH-imipramine. N-dealkylation is one example of many META transforms that powerfully generalizes to apply to many different chemical structures.

Despite the complexity of imipramine metabolism, the major metabolic pathways of imipramine can be modeled successfully by only a few META transforms that correspond to the four biotransformations discussed above. These transforms may hit sequentially in various combinations, producing more than thirty distinct structures. However, some combinations, such as hydroxylated N-oxide metabolites, have never been experimentally observed (to the best of my knowledge). This points to the need to carefully design the transforms to ensure that only relevant metabolic biotransformations are predicted.

The metabolism of imipramine provides a good yardstick for the evaluation of META performance. The complex metabolism of molecules such as imipramine will be used to determine the predictive accuracy of META and guide the development of the dictionary. META acquires "intelligence" by developing
transforms for specific biotransformations that will apply (even unexpectedly) to new biotransformations.

CONCLUSION

The breadth of this incomplete metabolic profile for a single simple molecule (Figure 6) highlights the formidable task of presenting a comprehensive listing of potential metabolites of environmentally important chemicals and drugs.

The computerized approach described here is intended to make it possible not only to review the sequential formation of all conceivable metabolites from a given precursor molecule, but also to scrutinize the relative importance of the various metabolites on the basis of priority rankings (Chapter 4). These priorities can be developed according to relative abundance, tissue selectivity, isozyme selectivity, species variation, etc.

The predictive power of the program is predicated on the breadth of the information that is provided in the dictionary. Thus, a comprehensive data base from the literature is required. It was important to consider not only the bioactivities of the individual metabolites, but also the potential for interactions (enhancement or inhibition of bioactivity) among the metabolites and the parent chemical agent or other chemical agents and their metabolites in cases of coincident exposure. The commitment to the utility of the program required extensive dictionary development and thorough literature research, which is the subject of the next chapter.
CHAPTER 2

Modeling Mammalian Metabolism with META Methodology
INTRODUCTION

In the first chapter, an overview of the underlying methodology of the META program was described as well as the concept of a dictionary which acts as the knowledge base. Used with the appropriate dictionary, the META program applies biotransformation rules, called transforms, which recognize and substitute chemical functional groups. The inputs to META are line-coded\(^1\) chemical structures and the output is a graphical display of possible metabolites together with an overall map of the metabolic pathways traversed. The topic of this chapter is to outline the design and construction of a dictionary that accurately models the pathways known to be involved in the metabolism of xenobiotics in mammals.

Many pharmacologic effects of foreign compounds, especially drug efficacy and toxicity, are intimately related to metabolic biotransformations.\(^2\) The identification of the structures of the metabolites is among the first steps in delineating the molecular basis for pharmacological effect. The use of META methodology coupled to a comprehensive dictionary provides the needed collection of possible metabolites. META may predict metabolites that are overlooked or even unexpected by the human expert, especially if finer-detailed and less-encountered metabolic knowledge is included in the dictionary. An accurate apportionment of pharmacological efficacy can be accomplished only after all the metabolic species are defined.

Mammalian metabolism lends itself to modeling by META methodology for two reasons. First, examples of metabolic activation of xenobiotics in mammals are

\(^*\) This chapter is based on an accepted publication in the Journal of Chemical Information and Computer Sciences.
abundant and provide a great amount of knowledge from which to design transforms. Second, the enzymes involved in xenobiotic metabolism must be fairly non-specific toward their substrates, otherwise the number of specific enzymes required would be enormous and solvent capacity would be exceeded. This broad substrate specificity allows META transforms to extrapolate accurately, providing the user with a powerfully predictive tool.

METHOD

Since the utility of META is largely determined by the contents of the dictionary, the philosophy of its design was carefully considered. Comprehensiveness was a primary design objective, in contrast to other published programs in this field which model a limited and specific knowledge domain. Our dictionary incorporated only well-established metabolic data from reviews, textbooks, and monographs. Available metabolic literature pertains to any of several mammalian species. Rodent-specific pharmacological data was preferred and rodent liver metabolism was the most common reference, but data from any mammalian source was substituted if the rodent data was unobtainable. Hence, we do not claim that the resulting dictionary is an exhaustive collection of biotransformation rules for any particular mammal. Rather, our goal was to model a theoretical "average" mammal, in which the resulting dictionary would provide an excellent starting point for the development of more specific dictionaries that pertain to various species and organs.

An initial dictionary was carefully constructed that closely parallels the present consensus of the xenobiotic metabolic knowledge of Profs. J. J. Mieyal and L. M. Sayre (both from Case Western Reserve University), who are experts in the
field of medicinal chemistry. The META transforms were organized around pharmacological reaction type and focused on chemical substructures, i.e., O-dealkylation, N-hydroxylation, etc. The reaction types were associated with enzyme systems\textsuperscript{16} that are known to catalyze such reactions. In many cases, more than one type of enzyme can carry out the same reaction type, and each enzyme system may include several different reaction types. This method of classification allows, but does not require, parameterization of enzyme-substrate lipophilicity relationships, pharmacological compartment location, etc. It continues to serve as an adaptable framework for the addition of new metabolic data. Each enzyme system belongs either to phase 1 (functionalization) or phase 2 (conjugation) metabolism. The transforms in the initial dictionary were then intuitively prioritized.

The consensus knowledge provides general guidelines of possible metabolic pathways to follow, but the actual pathways traveled are far more specific to the structure of the substrates. Accordingly, the initial dictionary was subsequently improved by assimilation of specific biotransformations that had not been predicted accurately. An iterative refinement of the dictionary using test xenobiotics was performed. The initial dictionary was tested with fifty xenobiotics (Table 1) that were selected for their well-defined metabolism, which served as a training set for guiding dictionary refinement. For each xenobiotic, the performance of individual transforms was determined by comparing the actual experimental metabolic status (observed vs. not observed) against the META predictive status for each transform (hit vs. miss).
Table 1. Training set of xenobiotics.

<table>
<thead>
<tr>
<th>Acetaminophen</th>
<th>Chloramphenicol</th>
<th>Hydralazine</th>
<th>Pentobarbital</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-acetyl-</td>
<td>Cimetidine</td>
<td>Ibuprophen</td>
<td>Phenacetin</td>
</tr>
<tr>
<td>amino-fluorene</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amphetamine</td>
<td>Coumarin</td>
<td>Isoniazid</td>
<td>Phenol</td>
</tr>
<tr>
<td>Aniline</td>
<td>DDT</td>
<td>Lidocaine</td>
<td>Phenylacetic acid</td>
</tr>
<tr>
<td>Aspirin</td>
<td>Diazepam</td>
<td>Lindane</td>
<td>Phenyltrim</td>
</tr>
<tr>
<td>Carmustine</td>
<td>Diazinon</td>
<td>Methotrexate</td>
<td>Promotil</td>
</tr>
<tr>
<td>Benzo(a)pyrene</td>
<td>Dichlorvos</td>
<td>Methyl Isocyanate</td>
<td>Pyridine</td>
</tr>
<tr>
<td>Benzene</td>
<td>Dioxin</td>
<td>Naphthalene</td>
<td>Styrene</td>
</tr>
</tbody>
</table>
| Benzidine       | Dimethyl-
| Nitrosamine     | Nitrobenzene   | Thiouracil    |
| Benzoic acid    | Diphenyl-
| Methylphosphine | Nitroglycerin  | Toluene       |
| Caffeine        | Disulfiram      | Parathion    |               |
| Captopril       | Ethanol         |              |               |
| Carbon tetrachloride | Halothane   |              |               |

Depending upon performance, transforms were debugged as necessary in order to increase the predictive accuracy of the latest dictionary version. Since generalizability of the transforms is the source of the predictive power of META, transforms were constructed to recognize the smallest substrate fragment, yet still generate products that are consistent with experimental observations. This struck a good balance in the context of predicting chemical toxicity and drug efficacy, where a significantly higher cost would result for not predicting an actual metabolite as compared to suggesting a non-existent metabolite. Reducing the size of transforms increases processing speed and lowers memory requirements. Transform prioritization was adjusted as necessary on the basis of observed pharmacokinetics of several test molecules. For enzymes whose transform efficiency changes with substrate lipophilicity (e.g. cytochrome P-4508), we used a log P function to prioritize the transforms.

THE DICTIONARY

Both phase 1 and phase 2 of xenobiotic metabolism were modeled with META methodology. A given xenobiotic usually passes through both phases, but
Figure 7. Example of the generalization abilities of META methodology. a.) Biotransformations of pentobarbital and hexane, in this case aliphatic hydroxylation. b.) A META transform that models the biotransformation. It recognizes the ethyl substrate fragment and replaces it with the corresponding hydroxylated product fragment. The substrate fragment is kept as small as possible so that the transform will apply to xenobiotics outside the training set. (Branching groups are located between angle brackets. The first numeral inside indicates the atom of the main chain to which the branching group is connected.)

xenobiotics may undergo either phase or neither phase before elimination. Given the breadth of both phases of mammalian metabolism, our comprehensive META dictionary was required to be extensive, so a complete listing here is beyond the scope of a dissertation. Instead, we will discuss the essential activity of each reaction type that has been encoded into META transforms, providing xenobiotic archetypes. A representative transform for each reaction type is provided. Biotransformation reaction types for each phase of xenobiotic metabolism are listed in Tables 2 and 3. Presently, we have constructed over 900 fundamental transforms based on pharmacological data of more than 150 xenobiotics.

The reaction type of aliphatic hydroxylation (a phase 1 biotransformation, see below) is now used to demonstrate the mechanics of META (Figure 7). The principal metabolite of the sedative pentobarbital is that obtained by hydroxylation...
of its sec-pentyl side chain at the ω-1 position. The solvent hexane undergoes an analogous reaction which is modeled by the same transform. Note that a single META transform models both biotransformations.

**Phase 1 biotransformations**

Occasionally referred to as "non-synthetic" reactions, phase 1 biotransformations usually metabolize the xenobiotic first. However, the metabolites of phase 1 biotransformations are frequently not excreted, even though they may be less lipophilic. Phase 1 biotransformations include oxidative, hydrolytic, and reductive reactions, and generally introduce polar functionalities to lipophilic species. A product of phase 1 biotransformations may have higher or lower pharmacological effect than the substrate, or it may exhibit an altogether different effect, such as toxicity. Table 2 lists phase 1 reaction types that have been modeled by META transforms.

The cytochromes P-450 monooxygenase isozyymes17-21 are central to phase 1 xenobiotic metabolism, catalyzing several reaction types that are represented by a total of 197 transforms. They are perhaps the most versatile biological catalysts known and are the most powerful in vivo oxidizing agent. Cytochromes P-450, which are often the first line of attack on lipophilic, non-polar substances, have extremely broad substrate specificity. There are thousands of chemical compounds that are known to be metabolized by cytochrome P-450. The overall behavior of cytochrome P-450 is well-defined,22 allowing for simple coding by META transforms. All the reactions of cytochrome P-450 insert a single atom of oxygen into the xenobiotic, presumably by a free radical mechanism, to form a hydroxy or epoxide moiety.23
Aliphatic hydroxylation by P-450 is represented by forty-three transforms and one type (ω-1) was discussed above (Figure 7). Our dictionary contains transforms for hydroxylation of methyl, methylene, and methine sp³ carbons in an aliphatic chain. Aliphatic substituents on aromatic rings (e.g. toluene, tolbutamide, and methaqualone) are generally hydroxylated at the methylene adjacent to the aromatic ring because the phenyl ring activates this position.

Aromatic hydroxylation generally proceeds via an epoxide intermediate,²⁴ but direct hydroxylation via an electrophilic aromatic substitution mechanism occurs for some aromatic substrates.²⁵ The epoxide intermediate may either be relatively stable or spontaneously rearrange to the phenol, as indicated by the NIH shift. The electrophilic epoxide may be hydrolyzed, or it may be adducted with glutathione, or it may bind covalently to macromolecules via reaction with nucleophilic sites on proteins or DNA.²⁶-²⁷ The latter is one type of covalent binding to DNA that has been implicated in chemical carcinogenesis.²⁸ Specific transforms for the generation of transannular oxygenation, such as in the generation of naphthalene 1,4-oxide, were also written. The analgesic phenacetin,²⁹ the antidepressant imipramine,³⁰ benzene,³¹,³² and particular molecular positions on polycyclic aromatic hydrocarbons³³ (e.g. benzo[a]pyrene³⁴) are epoxidized by P-450. The sensitivity of aromatic hydroxylation/epoxidation to substituent effects prompted us to extend META modeling to use quantum mechanical indices and graph indices (Chapter 3).

For aliphatic ethers, sulfides, halides, and amines, hydroxylation by P-450 generally occurs at the carbon adjacent to the heteroatom which is usually followed by spontaneous dissociation of the hydroxylation product to release the substituted heteroatom and generate an aldehyde or ketone as the second product. The net biotransformation is oxidative cleavage of C-X bonds, where X can be oxygen,
sulfur, halogen or a nitrogen. A hydrogen α to the heteroatom in the substrate is essential. Twenty-eight transforms model O-dealkylation of aromatic and aliphatic ethers (e.g., the conversion of codeine to morphine and phenacetin to acetaminophen). Sixteen transforms model S-dealkylation with the thyroid inhibitor 6-methylthiouracil being an example substrate. Dehalogenation\(^{35}\) of insecticides and commercial solvents is a common metabolic pathway which is modeled by thirty transforms. Halogenated alkanes\(^{36}\) such as the anaesthetic halothane\(^{37}\) are dehalogenated by P-450. Eighty transforms were constructed to model N-dealkylation, a frequently encountered biotransformation which acts upon substrates such as ethylmorphine, dimethylformamide, and imipramine.

Deamination is the same as N-dealkylation except that the former term pertains to primary amines and can occur through at least three enzyme systems: microsomal cytochrome P-450, mitochondrial flavin-dependent monoamine oxidase (MAO),\(^{38}\) and the copper amine oxidases.\(^{39}\) The P-450 reaction involves α-hydroxylation, whereas the MAO reaction proceeds through dehydrogenation to an imine which is subsequently hydrolyzed.\(^{40}\) Since there must be a hydrogen α to the nitrogen, substituted anilines are not deaminated. Catecholamine neurotransmitters are deactivated via deamination by MAO. MAO mainly dehydrogenates primary amines, but secondary and tertiary allylic amines are also metabolized. Fifteen META transforms modeled deamination by MAO. Two other transforms represent the copper amine oxidases, an enzyme system including plasma amine oxidase, semicarbazide-sensitive amine oxidase, and diamine oxidase whose substrate specificity overlaps with MAO. The copper amine oxidases contain a quinone cofactor and metabolize strictly unbranched alkyl and aralkyl primary
amines. Histamine, putrescine, and cadaverine undergo deamination by diamine oxidases.

\[ N\text{-hydroxylation} \] and \[ N\text{-oxygenation} \] are catalyzed by flavin monooxygenase (FMO) and/or cytochrome P-450, with overlapping substrate specificities. Xenobiotics with no α-hydrogen, such as tert-alkylamines (e.g. phentermine), primary arylamines (e.g. aniline, benzidine), primary arylamides (e.g. phenacetin), and hydrazines (e.g. isopropylhydrazine, procarbazine), are principally N-hydroxylated by P-450. Aromatic heterocyclic amines (e.g. 3-methylpyridine) are also N-oxygenated by P-450. FMO N-oxidizes strictly aliphatic amines. Virtually all tertiary amines (e.g. trimethylamine, cocaine, nicotine, the antihypertensive pargyline, and tranquilizer perazine) are potential substrates for N-oxygenation by FMO, generating amine N-oxides. Secondary amines are N-oxidized by both enzymes. The hydroxylamine products can be subjected to repeated FMO N-oxygenation to give nitrones (from secondary amines) and oximes (from primary amines). 2-Acetylaminofluorene undergoes N-hydroxylation which is required for activation to the ultimate carcinogen. Sixty-two META transforms were constructed which model N-hydroxylation and N-oxygenation.

Sulfoxidation is almost exclusively catalyzed by FMO, which oxidizes a broad variety of divalent sulfur compounds. Sulfoxidation is more common than S-dealkylation. Thirty-six transforms modeled the biotransformation of alkyl sulfides (e.g. chlorpromazine, cimetidine); disulfides (e.g. disulfiram); thiols (e.g. thiophenol); sulfinic, sulfonic, and sulfonic acids; thioamides and thiocarbamides (e.g. thionicotinamide, methimazole). Thiocarbamides and thioamides (e.g. thioacetamide) undergo sequential oxidation by FMO, the products of which are
subject to tautomerization.\textsuperscript{50} Several pesticides (e. g. disulfoton) are excellent substrates for S-oxidation by FMO in vitro. Desulfuration, which occurs when a sulfur atom is replaced by oxygen, is catalyzed by both P-450 and FMO. The insecticide parathion\textsuperscript{15} and the solvent carbon disulfide undergo desulfuration.

\textit{Alcohol oxidation}\textsuperscript{55} is catalyzed mainly by the enzyme alcohol dehydrogenase, which is modeled with four META transforms. Alcohol dehydrogenase has broad specificity and readily oxidizes most primary and secondary aliphatic alcohols and aralkyl alcohols,\textsuperscript{56} while tertiary and sterically hindered secondary alcohols are poor substrates.\textsuperscript{57} In particular, ethanol is oxidized to acetaldehyde.\textsuperscript{58} Alcohol dehydrogenase is also known to reduce aldehydes and aromatic ketones.\textsuperscript{49}

\textit{Aldehyde oxidation} of aliphatic and aromatic aldehydes is catalyzed by aldehyde dehydrogenases.\textsuperscript{59} Metabolites of MAO and alcohol dehydrogenase usually undergo aldehyde oxidation.\textsuperscript{60} Seven META transforms were needed to model the broad specificity of aldehyde dehydrogenase. Acetaldehyde\textsuperscript{61} and 5-hydroxyindoleacetaldehyde are examples of substrates. Many pesticides (e. g. aldicarb) are oxidized to acids by aldehyde dehydrogenases.\textsuperscript{60}

\textit{Purine oxidation} is catalyzed by xanthine oxidase and aldehyde oxidase.\textsuperscript{62} These two activities actually represent several different species of enzymes that incorporate an atom of oxygen into the product from water rather than from molecular oxygen, as do P-450 and FMO. This is a characteristic of molybdenum flavoheme enzymes that use FAD.\textsuperscript{63} Both enzymes oxidize nitrogen-containing heterocyclic compounds as well as aliphatic and aromatic aldehydes. Four META transforms were needed to model their broad and overlapping substrate specificities.
### Table 2. Phase 1 biotransformations. Special characters in the transforms are as follows. The caret symbol (^) represents an atom in a three-membered ring, the double quote (") represents an sp^2^ carbon, and X represents a halogen.

<table>
<thead>
<tr>
<th>Reaction type</th>
<th>Example of META transform</th>
<th>Example compounds</th>
<th>Enzyme enzyme system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aromatic hydroxylation</td>
<td>CH =CH -&lt;1-OH&gt;</td>
<td>benzene</td>
<td>P-450</td>
</tr>
<tr>
<td></td>
<td>C -CH -&lt;1-OH&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aliphatic hydroxylation</td>
<td>CH3=CH2 -&lt;2-OH&gt;</td>
<td>hexane</td>
<td>P-450</td>
</tr>
<tr>
<td></td>
<td>CH3=CH -&lt;2-OH&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epoxidation</td>
<td>CH =CH -&lt;1-O2&gt;</td>
<td>benzo[a]pyrene</td>
<td>P-450</td>
</tr>
<tr>
<td></td>
<td>CH-CH =&lt;1-O2&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-oxygenation</td>
<td>CH2-N -CH3=&lt;2-C&gt;</td>
<td>imipramine</td>
<td>flavin monooxygenase</td>
</tr>
<tr>
<td></td>
<td>CH2-NO -CH3=&lt;2-C&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-hydroxylation</td>
<td>NH -CO -N</td>
<td>chloramphenicol</td>
<td>P-450</td>
</tr>
<tr>
<td></td>
<td>NH -CO -&lt;1-OH&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O-dealkylation</td>
<td>CH&quot;-O -CH2=</td>
<td>phenacetin</td>
<td>P-450</td>
</tr>
<tr>
<td></td>
<td>CH&quot;-O -CH=</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-dealkylation</td>
<td>CH&quot;-S -CH3</td>
<td>6-methyl-thiopurine</td>
<td>flavin monooxygenase</td>
</tr>
<tr>
<td></td>
<td>CH&quot;-S -CH2=&lt;3-OH&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-dealkylation</td>
<td>CH3-N -CH3</td>
<td>dimethylformamide</td>
<td>P-450</td>
</tr>
<tr>
<td></td>
<td>CH3=N -CH2=&lt;3-OH&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deamination</td>
<td>NH2-CH -NH</td>
<td>amphetamine</td>
<td>monoamine oxidase</td>
</tr>
<tr>
<td></td>
<td>NH =C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sulfoxidation</td>
<td>S -S -SO</td>
<td>disulfiram</td>
<td>flavin monooxygenase</td>
</tr>
<tr>
<td></td>
<td>S -S -</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Desulfurization</td>
<td>CH2-O -P =S =2-O   -&gt;</td>
<td>parathion</td>
<td>P-450</td>
</tr>
<tr>
<td></td>
<td>CH2-O -P =O =2-O -&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oxidative dehalogenation</td>
<td>X -C -X X -C =O</td>
<td>halothane</td>
<td>halogenase</td>
</tr>
<tr>
<td>β-oxidation</td>
<td>CH2-CH2-CO -OH</td>
<td>phenylalkyl acids</td>
<td>mitochondrial β-oxidation</td>
</tr>
<tr>
<td></td>
<td>CO CH3-CO -OH =1-OH &gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ester hydrolysis</td>
<td>CO -O -CH2=</td>
<td>procaine</td>
<td>esterases</td>
</tr>
<tr>
<td></td>
<td>CO -OH CH2=&lt;3-OH &gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amide hydrolysis</td>
<td>NH -CO -NH2 CO =&lt;2-OH &gt;</td>
<td>phenacetin</td>
<td>amidases, peptidases</td>
</tr>
<tr>
<td></td>
<td>NH =CO -N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epoxide hydrolysis</td>
<td>CH -O -CH =</td>
<td>naphthalene</td>
<td>epoxide hydratase</td>
</tr>
<tr>
<td></td>
<td>CH -OH CH =&lt;3-OH &gt;</td>
<td>1,2-oxide</td>
<td></td>
</tr>
<tr>
<td>Alcohol oxidation</td>
<td>OH -CH2- O =CH -</td>
<td>ethanol</td>
<td>alcohol dehydrogenase</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aldehyde oxidation</td>
<td>CH3=COH =&lt;2-OH &gt;</td>
<td>acetaldehyde</td>
<td>aldehyde oxidase</td>
</tr>
<tr>
<td></td>
<td>CH3-CO =&lt;2-OH &gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purine oxidation</td>
<td>N =CH -NH =N</td>
<td>theophylline</td>
<td>xanthine oxidase</td>
</tr>
<tr>
<td></td>
<td>N =C -NH =&lt;2-OH =</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carboxyl reduction</td>
<td>COH-C =</td>
<td>trichloro-</td>
<td>carbonyl reductase</td>
</tr>
<tr>
<td></td>
<td>CH2-C =&lt;1-OH &gt;</td>
<td>acetaldehyde</td>
<td></td>
</tr>
<tr>
<td>Azoreduction</td>
<td>C&quot;-N =N -C =</td>
<td>prontosil</td>
<td>flavin reductase</td>
</tr>
<tr>
<td></td>
<td>C&quot;-NH2 NH2-C =</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nitroreduction</td>
<td>NO2-C =</td>
<td>chloramphenicol</td>
<td>flavin reductase</td>
</tr>
<tr>
<td></td>
<td>NO -C =</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The antineoplastic agent 6-mercaptopurine and the anti-asthmatic theophylline undergo purine oxidation to their uric acid derivatives. In a more unusual example, the nicotine Δ1(5')-iminium ion is converted to the lactam cotinine by aldehyde oxidase.64

β-oxidation of xenobiotic straight-chained alkyl carboxylic acids as well as endogenous fatty acids occurs through the same metabolic pathway. One META transform was required to model the oxidative cleavage of two carbon units from the substrate. Phenylalkyl acids undergo β-oxidation.49

Ester hydrolysis65 is catalyzed by a wide range of enzymes (e.g. acetylcholinesterase, carboxylesterase) with different substrate specificities. Twenty META transforms represent hydrolysis of aromatic and aliphatic esters. The pharmacological effects of procaine and atropine are terminated upon hydrolysis of their ester bond. Thioesters (e.g. phenyl thioacetate66) are also subject to hydrolysis.

Amide hydrolysis65 is similarly catalyzed by numerous enzymes with broad specificity. Amides are generally more stable to hydrolysis than their analogous esters.49 Two META transforms represent the hydrolysis of all amides. The amide bond in the antiepileptic phenytoin and the anesthetic lidocaine are subject to hydrolysis by amidases. Hydrolysis of the amide bond in phenacetin generates products that may induce methemoglobinemia and renal toxicity.67,68

Epoxide hydrolysis, catalyzed by epoxide hydratases,69 generates 1,2-diols. Eleven META transforms were needed to model this highly non-specific reaction type. Most substrates are the epoxide products of cytochrome P-450 monooxygenase which may react with nucleophilic centers in RNA and DNA.27
Hydrolysis of benzo[a]pyrene 7,8-oxide produces a precursor for a mutagen/carcinogen.70

_Carbonyl reduction_71 of aldehyde and ketones was simulated with twenty META transforms. Carbonyl reductases include both aldehyde72 and ketone reductases73,74 which do not overlap in substrate specificity. Carbonyl-containing xenobiotics which are not oxidized are reduced to alcohols and thereby supply a site for conjugation75 (see phase 2 reactions below). Aldehyde reductases do not appear to be involved in endogenous metabolism and show broad specificity for aromatic, aliphatic, and sugar aldehydes. Ketone reductases reduce aliphatic, unsaturated, and aromatic ketones.76 Common substrates are quinones, aromatic aldehydes, and ketoaldehydes.

_Nitroreduction_ is catalyzed by flavin-containing reductases, with cytochrome c (P-450) reductase being an example.77 Cytochrome c reductase reduces nitro compounds (e.g., 2-nitrofluoranthene,78 chloramphenicol,79 and nitrobenzene80) to their respective nitroso form. META transforms have also been included that model the further reduction of nitroso to hydroxylamine, then to amine.

A META transform representing _azoreduction_81,82 of the dye prontosil and sulfasalazine exists in our dictionary. The biotransformation of the azo functionality to aromatic primary amines occurs by both the intestinal microfluora and by cytochrome c (P-450) reductase.83 DT diaphorase is another reductase which reduces the red food dye amaranth.84
Phase 2 biotransformations

Also termed conjugation reactions\textsuperscript{85} or synthetic reactions, phase 2 biotransformations are listed in Table 3. Phase 2 biotransformations attach hydrophilic groups such as carbohydrate, amino acid, peptide, or sulfate groups which ordinarily remove bioactivity and convert the xenobiotic to more excretable forms. However, methylation and acetylation of heteroatom-containing substrates also occur, which do not necessarily enhance water solubility or excretability. The limited number of conjugation centers (-COOH, -OH, -NH\textsubscript{2}, -SH, epoxides, and other electrophilic centers) makes prediction of phase 2 reactions by META reliable. At least one phase 2 reaction will probably occur if a conjugation center is present.

\textit{N- and O-glucuronic acid conjugation}\textsuperscript{86,87} is the most common phase 2 detoxication pathway. Seventeen transforms were written for phenols, alcohols, and carboxylic acids substrates. Salicylamide is excreted mainly as the glucuronide.\textsuperscript{88} The resulting adduct is quite hydrophilic and is usually excreted in the urine and bile immediately, unless first deconjugated by intestinal flora with subsequent enterohepatic cycling.

If not glucuronidated, the hydroxyl group usually undergoes \textit{O-sulfation}.\textsuperscript{89,90} Sixteen transforms were written for \textit{O-sulfation} of aliphatic and aromatic alcohols\textsuperscript{91} and hydroxylamines.\textsuperscript{92} The N,O-sulfate product of the carcinogen N-hydroxy-2-acetylaminofluorene is much more toxic than its substrate.\textsuperscript{93} Many substituted phenols\textsuperscript{8} such as 3-hydroxycoumarin\textsuperscript{94} undergo \textit{O-sulfation}. One transform models aromatic amine \textit{N-sulfation} with aniline and sulfanilamide being example substrates. Sulfation reactions are catalyzed by several sulfotransferases.
Table 3. Phase 2 biotransformations in mammals. Z2 represents glutamine, Z3 represents glutathione, and Z4 represents glucuronic acid. (See Table 2 for explanation of other special symbols.)

<table>
<thead>
<tr>
<th>Reaction type</th>
<th>Example of META transform</th>
<th>Example compounds</th>
<th>Enzyme enzyme system</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-glucuronide</td>
<td>NH₂-C =</td>
<td>aniline</td>
<td>N-glucuronyl transferase</td>
</tr>
<tr>
<td>conjugation</td>
<td>NH -C =&lt;1-Z4 &gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O-glucuronide</td>
<td>OH -C =</td>
<td>salicylamide</td>
<td>O-glucuronyl transferase</td>
</tr>
<tr>
<td>conjugation</td>
<td>O -C =&lt;1-Z4 &gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O-sulfate</td>
<td>OH -C =</td>
<td>acetaminophen</td>
<td>phenol sulfortransferase</td>
</tr>
<tr>
<td>conjugation</td>
<td>O -C =&lt;1-SO₂-OH &gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-sulfate</td>
<td>NH₂-C =</td>
<td>2-naphthylamine</td>
<td>arylamine sulfortransferase</td>
</tr>
<tr>
<td>conjugation</td>
<td>NH -C =&lt;1-SO₂-OH &gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-acetylation</td>
<td>NH₂-N =</td>
<td>isoniazide</td>
<td>N-acyltransferase</td>
</tr>
<tr>
<td>conjugation</td>
<td>NH -N =&lt;1-CO -CH₃ &gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N,O-</td>
<td>OH -NH -CH -</td>
<td>N-acetylamino-</td>
<td>N,O-acetyltransferase</td>
</tr>
<tr>
<td>transacetylation</td>
<td>O -NH -CH -&lt;1-CO -CH₃ &gt;</td>
<td>fluorene</td>
<td></td>
</tr>
<tr>
<td>Glutamine</td>
<td>OH -CO -C =</td>
<td>phenylacetic acid</td>
<td>glutamyl transferase</td>
</tr>
<tr>
<td>conjugation</td>
<td>NH -CO -C =&lt;1-Z2 &gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycine</td>
<td>OH -CO -C =</td>
<td>salicylic acid</td>
<td>glycyl transferase</td>
</tr>
<tr>
<td>conjugation</td>
<td>NH -CO -C =&lt;1-CH₂-CO -OH&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-methylation</td>
<td>NH₂-CH₂-</td>
<td>phenyl-ethanolamine</td>
<td>N-methyl transferase</td>
</tr>
<tr>
<td>conjugation</td>
<td>NH -CH₂-&lt;1-CH₃ &gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O-methylation</td>
<td>C&quot; -OH</td>
<td>catechol</td>
<td>COMT</td>
</tr>
<tr>
<td></td>
<td>C&quot; -O &lt;2-CH₃ &gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-methylation</td>
<td>SH -C =CH -</td>
<td>thiouracil</td>
<td>S-methyl transferase</td>
</tr>
<tr>
<td></td>
<td>S -C =CH -&lt;1-CH₃ &gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glutathione</td>
<td>CH -O&quot; -CH -</td>
<td>styrene oxide</td>
<td>glutathione S-transferase</td>
</tr>
<tr>
<td>conjugation</td>
<td>CH -Z₃ CH -&lt;3-OH &gt;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**N-acetylation**\(^{95}\) is the major route of aromatic amine (e.g. procainamide, sulfisoxazole) metabolism\(^{75}\) for most species. N-Acetylation of the sulfonamido group in sulfanilamide results in a compound that is less soluble in both water and lipid, which may precipitate in the kidney as it is concentrated in the urine. Fourteen transforms were used to model N-acetylation of amines. N-acetyltransferase has a broad specificity toward compounds with an amino group connected to unsaturated rings or carbonyl groups. Secondary amines are not acetylated.\(^{49}\) Hydrazines\(^{46}\) such as isoniazid undergo N-acetylation. Six transforms represent **N,O-transacetylation**,\(^{96}\) which is catalyzed by N,O-acyltransferase. Aromatic amines are thought to induce tumors as a consequence of their N-oxygenation
followed by O-acetylation. Hydroxylamines and hydroxamides are the primary substrates.

*Glycine conjugation*\(^{97}\) forms an amide bond involving the \(\alpha\)-amino group of glycine and an aliphatic or aromatic carboxylic acid that has been activated by conversion to the thioester with coenzyme A. Five transforms were encoded for biotransformation of primary, secondary, and tertiary aliphatic acids and aromatic acids. *Benzoic acid*,*salicylic acid*,*nicotinic acid*, as well as medium chain alkyl carboxylic acids are conjugated with glycine. Glycine conjugation products are usually excreted through the bile.

*Glutamine conjugation*\(^{97}\) forms an amide bond with the \(\alpha\)-amino group of L-glutamine and seems to occur exclusively in primate species.\(^{75}\) In humans, phenylacetic acid is conjugated with L-glutamine instead of glycine as in other species. Eight META transforms were used to model conjugation of aliphatic and aromatic carboxylic acids. Mescaline is known to conjugate with glutamine.\(^{99}\)

*Glutathione (GSH) conjugation*, catalyzed by glutathione S-transferase, is the principal route of phase 2 metabolism of reactive electrophiles (e.g. halides, Michael acceptors, epoxides). Following GSH conjugation, the glutamic acid and glycine termini of the GSH tripeptide are hydrolytically removed, and the resulting cysteine conjugate is N-acetylated to afford mercapturic acids, which are excreted. Although glutathione conjugation is usually not quantitatively significant in xenobiotic elimination, it protects proteins and nucleic acids from electrophilic addition. Reactive epoxides that are generated by cytochrome P-450 are removed in this way. Seventy-one META transforms were used to model GSH conjugation of epoxides (e.g. styrene epoxide\(^{100}\)), \(\alpha,\beta\)-unsaturated compounds (e.g. iminoquinone\(^{101}\)), alkyl halides (e.g. methyl iodide), vinyl halides (e.g.
tetrachloroethylene\textsuperscript{102}, and isocyanates (e. g. methyl isocyanate\textsuperscript{103}). Glutathione conjugate formation is an important bioactivation mechanism for several groups of compounds.\textsuperscript{104} Two META transforms were written for cysteine conjugate \(\beta\)-lyase,\textsuperscript{105}, which liberates acetic acid from the cysteine conjugate. Two META transforms were written for glyoxylase\textsuperscript{106} which conjugates GSH as an intermediate in reduction of dicarboxyl compounds.

Methylations\textsuperscript{107} are important because they often significantly change the pharmacological activity of the substrate. Most \textit{N-methylations} are catalyzed by a family of specific \(N\)-methyltransferases, such as indoleethylamine \(N\)-methyltransferase. \(N\)-methylation occurs mainly in the biosynthesis or metabolism of naturally occurring endogenous compounds such as histamine, estradiol, thyroxine, norepinephrine, dopamine, and serotonin.\textsuperscript{108} Several META transforms were written to model \(N\)-methylation for aliphatic and aromatic primary, secondary, and tertiary amines, including cyclic amines such as normorphine. Pyridines (e. g. nicotinamide, nicotinic acid\textsuperscript{109}) are \(N\)-methylated to quaternary \(N\)-methylpyridinium salts.

Six META transforms represent \textit{O-methylation} of phenols catalyzed by a family of monohydric, dihydric, and trihydric phenol \(O\)-methyltransferases, which have broad specificity with regard to \(o,p\)-substitution on the aromatic ring.\textsuperscript{110} Caffeic acid is \(O\)-methylated.\textsuperscript{111} \(O\)-methylations of endogenous catechols such as epinephrine and other neurotransmitters occur primarily via catechol \(O\)-methyltransferase (COMT).
EXERCISING THE DICTIONARY

After its development and refinement, we needed to test the dictionary under conditions consistent with actual pharmacological use. The chief use of META is anticipated to be the prediction of all significant metabolites of a given xenobiotic. The first exercise demonstrated that META could accurately predict the metabolism of the relatively simple drug molecule phenacetin in mammals.\textsuperscript{29,68} (Figure 8). Its major primary metabolite is the therapeutically active product acetaminophen formed via O-dealkylation by the P-450 enzyme system. Deacetylation via
hydrolysis by amidases is usually second in prominence. The other two P-450-mediated reactions, epoxidation\textsuperscript{112} and N-hydroxylation,\textsuperscript{113} are minor pathways leading to potentially toxic metabolites.

The metabolic profile for phenacetin expands broadly as secondary biotransformation pathways are considered. Examples of enzyme-mediated reactions as well as non-enzymatic reactions are illustrated in Figure 8. Acetaminophen and N-hydroxyphenacetin are hydrolysed by amidases to N-hydroxyphenetidine and $p$-aminophenol, respectively. Acetaminophen and $p$-aminophenol undergo autooxidation to their iminoquinone form, which is subsequently conjugated by glutathione. Phenacetin-2,3-epoxide spontaneously rearranges to 2-hydroxyphenacetin, which is modeled by a separate spontaneous reaction dictionary (MESP.DIC). These two reactions illustrate the potential coupling between spontaneous reactions (ring opening, autooxidation) and enzyme-mediated reactions (glutathione conjugation). Both the primary and secondary metabolites of phenacetin were correctly predicted by META.

The second pharmacological exercise had a larger focus, insofar as we were interested to determine whether META could generate (among others) the most prominent metabolites for a large set of "classic" xenobiotics. Table 4 lists the compounds randomly selected from R. T. Williams' book\textsuperscript{75} and review\textsuperscript{114} which were mostly outside the training data. The META program was run with our dictionary and correctly predicted the major biotransformation for all selected chemicals with one exception, diethyl ether, which is mainly excreted via the lungs due to its volatility.
<table>
<thead>
<tr>
<th>Compound</th>
<th>Biotransformation</th>
<th>Compound</th>
<th>Biotransformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetaldehyde</td>
<td>Oxidation</td>
<td>Meprobamate</td>
<td>Demethylation</td>
</tr>
<tr>
<td>Acetoxime</td>
<td>Hydrolysis and reduction</td>
<td>Meprobamate</td>
<td>Aliphatic hydroxylation</td>
</tr>
<tr>
<td>Aniline</td>
<td>N-acetylation and aromatic</td>
<td>Naphthalene</td>
<td>Epoxidation</td>
</tr>
<tr>
<td></td>
<td>hydroxylation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anisole</td>
<td>O-demethylation</td>
<td>Nicotinic acid</td>
<td>Glycination</td>
</tr>
<tr>
<td>Aspirin</td>
<td>Hydrolysis</td>
<td>Nitrobenzene</td>
<td>Reduction and aromatic hydroxylation</td>
</tr>
<tr>
<td>Benzamide</td>
<td>Hydrolysis</td>
<td>Oxalic acid</td>
<td>Excreted unchanged</td>
</tr>
<tr>
<td>Benzanthracene</td>
<td>Aromatic epoxidation</td>
<td>Phenacetin</td>
<td>O-dealkylation</td>
</tr>
<tr>
<td>Benzyamine</td>
<td>Oxidative deamination</td>
<td>Phenobarbital</td>
<td>Aromatic hydroxylation</td>
</tr>
<tr>
<td>Bromobenzene</td>
<td>Epoxidation</td>
<td>Phenol</td>
<td>Sulfation</td>
</tr>
<tr>
<td>Cadaverine</td>
<td>Oxidative deamination</td>
<td>Phenylacetamide</td>
<td>Hydrolysis</td>
</tr>
<tr>
<td>Camphor</td>
<td>Aliphatic hydroxylation</td>
<td>Phenylbutazone</td>
<td>Aromatic and aliphatic hydroxylation</td>
</tr>
<tr>
<td>Chloral hydrate</td>
<td>Reduction</td>
<td>Phenylhydroxylamine</td>
<td>Reduction</td>
</tr>
<tr>
<td>Cyclohexane</td>
<td>Aliphatic hydroxylation</td>
<td>Procaine</td>
<td>Hydrolysis</td>
</tr>
<tr>
<td>Erythriol</td>
<td>Excreted unchanged</td>
<td>Proniosil</td>
<td>Reduction</td>
</tr>
<tr>
<td>Ethanol</td>
<td>Oxidation</td>
<td>Quinone</td>
<td>Glutathione conjugation</td>
</tr>
<tr>
<td>Ethyl ether</td>
<td>Excreted unchanged</td>
<td>Salicylamide</td>
<td>Glucuronidation</td>
</tr>
<tr>
<td>Ethylene glycol</td>
<td>Oxidation</td>
<td>Sulfanilamide</td>
<td>N-acetylation</td>
</tr>
<tr>
<td>Isoniazid</td>
<td>N-acetylation</td>
<td>Thiopental</td>
<td>Aliphatic hydroxylation</td>
</tr>
<tr>
<td>Meperidine</td>
<td>Hydrolysis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the third pharmacological exercise, we checked whether the major metabolic route to elimination could be accurately predicted. Levine\textsuperscript{115} characterizes the major multi-step pathways given for the four common functionalities listed in Table 5. META followed the observed pathways for test structures containing each functionality.

Prodrugs\textsuperscript{116,117} can be used to increase duration of action and/or reduce side-effects of particular pharmacological agents. Esterification is often used in the design of prodrugs. META generated epinephrine from its ester prodrug dipivaloylpepinephrine which is an antiglaucoma agent.\textsuperscript{118} Morphine is generated from both deesterification of heroin and O-demethylation of codeine, which also were predicted correctly by META.
<table>
<thead>
<tr>
<th>Functionality</th>
<th>Major route to elimination</th>
<th>Predicted?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydroxyl</td>
<td>Oxidation, glucuronide conjugation</td>
<td>Yes</td>
</tr>
<tr>
<td>aliphatic</td>
<td>Glucuronide, sulfate conjugation, or methylation</td>
<td>Yes</td>
</tr>
<tr>
<td>aromatic</td>
<td>Glycine, or glucuronide conjugation</td>
<td>Yes</td>
</tr>
<tr>
<td>Carboxyl</td>
<td>Oxidation, glucuronide conjugation</td>
<td>Yes</td>
</tr>
<tr>
<td>aliphatic</td>
<td>Deamination, glucuronide conjugation</td>
<td>Yes</td>
</tr>
<tr>
<td>aromatic</td>
<td>Acetylation, glucuronide conjugation, methylation</td>
<td>Yes</td>
</tr>
<tr>
<td>Amino</td>
<td>Hydroxylation, epoxidation</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**CONCLUSIONS**

We have successfully modeled the majority of mammalian xenobiotic biotransformations with a dictionary of META transforms and demonstrated that the dictionary is useful and predictive. The next level of sophistication requires three aspects of metabolism to be addressed: improvement of metabolite prioritization, species differences, and organ distribution.

Our dictionary has proven to rarely miss the prediction of possible metabolites. It is important to realize that the accuracy of some predictions is indeterminate because the pharmacological data needed to verify the prediction is not available. Pharmacological studies seldom determine that a possible metabolite is not produced in any quantity. Indeed, the prediction of these minor, possibly unrecognized metabolites is one reason why META was developed. In contrast, most experimental studies report the extent to which detected metabolites occur. Consequently, the metabolites can be ordered in terms of their pharmacological significance. Examples are plentiful. Cytochrome P-450 oxidizes methylene groups preferentially over methyl groups if both are present in the substrate; alkyl side chains are usually hydroxylated on the penultimate carbon atom. Lipophilicity
affects whether an epoxide substrate is conjugated with glutathione or hydrolyzed by epoxide hydratase. Esters that are not sterically hindered are more likely to be hydrolyzed. Carboxylic acids are subject to competing reactions of glycine and glutamine conjugation. The META methodology contains a parameterization which permits relative adjustment of transform priorities, such as to reflect a substrate's biotransformation preference. Experimental observations such as the above will continue to be incorporated into the dictionary by refining our initial transform prioritization.

Our dictionary is a starting point for prospective species-specific or individual-specific dictionaries. Biological factors (e.g. species, sex, genetics, xenobiotic exposure history, and age) are known to affect metabolism and disposition. Malathion is hydrolyzed in mammals to produce readily excretable compounds, but it is oxidized in insects to produce a cholinesterase inhibitor, which confers its insecticidal effects. The degree of N-deacetylation is species-dependent; whereas humans experience the side effect of methemoglobinemia associated with the deacetylated products of phenacetin, rabbits do not show this side effect because their amide hydrolysis is relatively low. Furthermore, the ratio of N-hydroxylation to O-dealkylation in hamsters is much larger than the corresponding ratio in mice, whereas guinea pigs have a very low N-hydroxylase activity. Regarding individual-differences, the distribution and induction of P-450 isozymes depends on the xenobiotic exposure and genetic background of the individual, which accounts for interindividual differences in response to some therapeutic drugs. There are differences in carcinogenicity and pharmacokinetics of aromatic amines and hydrazine drugs because N-acetyltransferases are products of multiple genes.
Perhaps the main value in META application will be the prediction of metabolite toxicity, and there are many important cases of organ-specific toxicity. This results from the selective accumulation of a particular metabolite, an unusually high activity of a certain enzyme, and/or an unusual sensitivity of the tissue to the toxic effect of the metabolite. For example, the hepatic toxicity of trichloroethylene is probably due to its increased metabolism to trichloroacetic acid, dichloroacetic acid and trichloroethylene epoxide.\textsuperscript{124} Also, 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) is metabolized by monoamine oxidase B in the brain to the pyridinium derivative MPP\textsuperscript{+} which accumulates to toxic levels in dopaminergic neurons.\textsuperscript{125} Prioritization of transforms (Chapter 4) could be used to develop tissue-specific META dictionaries. Such separate dictionaries are a prerequisite for modeling the distribution and transport of xenobiotics and their metabolites among various compartments.\textsuperscript{126}
CHAPTER 3

Extended META Methodology
INTRODUCTION

Basic META methodology (Chapter 1) entails the use of a dictionary of prioritized transforms that recognize and replace key functional groups in xenobiotics, generating possible metabolites. While META methodology can accurately capture the general behavior of most enzymes, some exhibit very diverse chemistry. As a result, it may be difficult to model some enzymes with the basic methodology because no all-encompassing transform exists; any single transform would have too many exceptions. The root of the problem is that some reaction types depend not only on the target fragment, but also on the surrounding structure that is not encoded by the transform.

Accordingly, the specificity of the transform can be increased, but this introduces two difficulties. First, there is a significant computational cost associated with increasing the number of transforms in the dictionary. Second, since the META transforms are restricted by the conventions of CASE fragment syntax, certain molecular structural features such as multiple branching and rings cannot be fully described. An alternative approach to increasing the specificity of the transforms is to utilize the fact that the enzymes associated with the biotransformations obey physicochemical laws, and that sophisticated methodologies that accurately model these laws can be applied to such enzyme systems. This approach is discussed in this chapter.

Incorporation of such established methodologies into META will model difficult reaction types more accurately and efficiently. Generally, the external methodology is evoked only when it is known that basic META methodology needs assistance. The use of special transforms and indicators (i.e., the class number or
special switches associated with the transform) promotes efficiency by treating the more difficult reaction types in greater depth while leaving the simpler reaction types to the usual treatment. Thus, this extension to META methodology will be computationally "transparent."

Cytochrome P-450 is an excellent example of an enzyme that catalyzes broad and diverse biotransformations. The primary function of cytochrome P-450, which is central to phase 1 metabolism, is to introduce reactive functionalities to non-polar substrates\(^1\) so that they can be more easily excreted. Accordingly, P-450 has evolved a hydrophobic binding site\(^2\) and resides in the endoplasmic reticulum,\(^3\) where non-polar molecules are likely to be encountered. As a result, a lipophilic molecule has a higher probability of being metabolized by P-450 than a hydrophilic molecule.\(^4\) Since the goal of META is to predict the biotransformations of enzymes like P-450, the lipophilicity of the entire structure of each substrate must be predicted.

Cytochrome P-450's chief enzymatic activity, hydroxylation, is rather insensitive to molecular structure. Aromatic rings and heterocycles may be hydroxylated. As previously discussed in Chapter 2, various META transforms were written to model hydroxylation, which generate enols, enamines, and isoamidine products. The energetic stability of these unstable intermediates are accordingly determined by a novel decision process utilizing quantum chemical calculations.

Besides hydroxylation, P-450 is known also to epoxidize alkenes and aromatic substrates. The oxidative metabolism of phenanthrene (Figure 9) is such an example. The prediction of epoxidation products of polycyclic aromatic hydrocarbons (PAH's) is very important to the field of toxicology. For example, Jerina's bay region theory is used to predict carcinogenicity of epoxidized PAH's.\(^5\)
The epoxide is an electrophilic functionality that is known to covalently bind nucleophilic centers in biological macromolecules,\textsuperscript{6,7} which is assumed to account for its carcinogenic properties.\textsuperscript{8,11} Such patterns of epoxidation and hydroxylation are difficult to explain on the basis of Molecular Orbital theory. Nevertheless, we found that a successful parameterization based on generalized perturbation (GP) theory can be made.

While this chapter is rooted in the modeling of P-450, the methodologies will be applied to the modeling of other relevant enzyme systems in META. For example, an unforeseen combination of META transforms other than P-450 might produce unexpected unstable intermediates which can be handled with the same methodologies.
BACKGROUND

Lipophilicity

Measurement of lipophilicity

The octanol-water partition coefficient \( P \) of a xenobiotic, which is defined as the ratio of the concentration of the xenobiotic in octanol over the concentration of the xenobiotic in water,\(^{12}\) is used to quantify lipophilicity. Since \( P \) can range over many orders of magnitude, the logarithm is customarily taken and \( \log P \) reported, just as \( \text{pH} \) is the logarithm of hydronium ion concentration. \( \log P \) ordinarily ranges from -3 (hydrophilic) to +6 (lipophilic). The state of the art of prediction of \( \log P \) is quite advanced; methodologies can be categorized in two ways: group contribution and surface area.\(^{13,14}\)

Lipophilicity is required to model P-450 and distribution of xenobiotics

As discussed above, the binding site of P-450 possesses an intrinsic hydrophobicity.\(^{15}\) The fact that binding and lipophilicity of the substrate are observed to be directly related\(^2\) can be used to fine-tune the priority (Chapter 4) of transforms concerned with P-450.

The transport and distribution characteristics of the xenobiotic, which are to be modeled by META, are strongly dependent upon its lipophilicity.\(^{16-18}\) Many physiological functions are actually due to concentration gradients of chemicals across biological barriers that form compartments in which chemicals are sequestered.\(^{19}\) Since the pharmacological effect of the xenobiotic is a sensitive function of the concentration of the chemical at the site of action, the impermeability of the compartment walls greatly affects the pharmacologic activity. The most
pharmacologically important compartments are: the brain, stomach, small intestine, large intestine, liver, lungs, circulatory system, kidneys, muscles, and skin. Some compartments do not directly contact each other (e.g. liver and brain) while others do: the circulatory system contacts all compartments so that oxygen and nutrients can be exchanged. In order to predict distribution accurately, the compartment configuration must be modeled.

Administration of the xenobiotic corresponds to setting an initial concentration of the drug in its respective compartment. Distribution of the chemical to other compartments can then occur. For a chemical to be transported to a different compartment, a lipophilic biological barrier must be crossed. There are three general classes of transport observed in biology.\textsuperscript{20,21} The first is passive transport, where the molecule moves down a chemical gradient spontaneously. It is observed mainly in lipophilic chemicals and usually follows a first order kinetic rate. Almost all biological barriers are composed of bilayer phospholipid membranes which are very lipophilic. Thus, the permeability of these membranes to passive transport is proportional to the lipophilicity of the chemical. The other two types of diffusion are facilitated transport, where a carrier escorts the chemical \textit{down} a chemical gradient, and active transport, where a carrier escorts the chemical \textit{up} a chemical gradient, are not as relevant to xenobiotics.

Passive transport can be accurately modeled with a first order rate constant, which is composed of two factors: the lipophilicity of the chemical and the surface area of the barrier.\textsuperscript{22} Thus a molecule that is not highly lipophilic might be easily absorbed if it is in a compartment with a high surface area, such as the small intestine. Since the lipophilicity of any chemical can be accurately calculated by existing methodologies,\textsuperscript{23} and the relative surface area of the compartment can be
estimated by using Fick's first law of diffusion, the rate constant of the transport process can be estimated, yielding a collection of first-order rate laws. Any system of first order kinetic equations can be solved numerically or analytically.\textsuperscript{24} Thus, given the compartment configuration and any specific transport rules, the distribution of the chemical as a function of time can be accurately modeled. This approach has been accomplished in the past for at least three specific systems: GABA/Glu compartmentalization in the brain,\textsuperscript{25} styrene metabolism via inhalation,\textsuperscript{26} and benzene metabolism in rats.\textsuperscript{27}

A specific example of distribution occurs in the excretion of xenobiotics by the kidney. If the xenobiotic is distributed to the blood compartment, it will be transported to the kidneys and removed by glomerular filtration. Passive diffusion will drive reabsorption of lipophilic substances, thus prohibiting their excretion, and hydrophilic xenobiotics will remain in the urine. Hydrophilic endobiotics (e.g. Na\textsuperscript{+}, H\textsubscript{2}O, and glucose) will be absorbed by active transport mechanisms.

**P-450 epoxidation/hydroxylation of alkenes and aromatic compounds**

Epoxide intermediates may be unstable and form hydroxylation products

Posner et. al.\textsuperscript{28} performed early investigations of hydroxylation. The hydroxylation of aromatic products has in the past been thought to proceed through an epoxide (arene oxide) intermediate.\textsuperscript{29} "It is now generally accepted that arene oxides are the principal activated intermediates of benzene and substituted benzene bioactivation..."\textsuperscript{30} "Arene oxides appear to be the principal primary oxidative metabolites of B[a]P formed by cytochrome P-450-dependent monooxygenases."\textsuperscript{31}
For most aromatic compounds, enzymatic hydrolysis and conjugation of the intermediate by glutathione provide evidence for the epoxide intermediate (Figure 10). Further evidence of an epoxide intermediate was obtained by the observation of hydroxylation-induced migration of deuterium and tritium labels, coined the NIH shift (Figure 11),\textsuperscript{32,33} which provided indirect evidence of the epoxide intermediate.

Other aromatic compounds pass through unstable epoxides, which undergo ring openings with potential NIH shift to form the phenol.\textsuperscript{29} However, direct hydroxylation cannot be ruled out. Recently, Kurata et. al.\textsuperscript{34} used the Fenton reaction to conclude that an arene oxide intermediate is not required for the NIH shift. The aromatic hydroxylation of toluene does not produce an observable epoxide intermediate — only the o, p-phenols are detected.\textsuperscript{35,36} Bromobenzene undergoes the usual pathway of epoxidation.\textsuperscript{4} Korzekwa\textsuperscript{37} eliminated initial epoxide formation and initial electron abstraction (charge transfer) as viable mechanisms for cytochrome P-450 catalyzed hydroxylation of chlorobenzene.
Rietjens et. al.\textsuperscript{38,39} have found that P-450 probably functions through electrophilic attack for fluorobenzenes.

However, this uncertainty can be modeled with "virtual" direct hydroxylation. For META purposes, the actual pathway(s) that is followed is irrelevant as long as the correct metabolites are formed. The direct hydroxylation pathway generates the same compounds as ring opening. The validity of aromatic hydroxylation in META will be maintained even as the exact mechanism becomes elucidated.

Mechanism of hydroxylation

Exactly how cytochrome P-450 accomplishes its great diversity of chemical reactions is not generally known. The general mechanism of oxygen atom incorporation by cytochrome P-450 is far from being firmly established. It is important to be somewhat familiar with the mechanism in order to discuss the parameterization later. Tomaszewski et. al.\textsuperscript{40} categorized the possible mechanisms as abstraction, insertion, addition-rearrangement, or direct addition, with the latter being the only viable primary mechanism. Ortiz de Montellano\textsuperscript{41} maintains that P-450 functions by sequential one-electron steps rather than by a single, concerted transfer of the ferryl oxygen to the substrate; the radical intermediates undergo rearrangement. Koymans et. al.\textsuperscript{42} also argue for a sequential one-electron oxidation mechanism. Guengerich\textsuperscript{43,44} proposes a mechanism which unifies disparate P-450
Figure 12. Consensus mechanism of hydroxylation by cytochrome P-450. 1.) Binding of substrate (camphor) 2.) One-electron reduction of heme group. 3.) Binding of O₂ to reduced heme. Notice the electron transfer from the iron atom to the oxygen species. 4.) One-electron reduction of bound oxygen species. The cleavage of the dioxygenase bond now may follow two pathways. 5.) Homolytic cleavage forms a hydroxy radical which is the attacking species. 6.) Abstraction of hydrogen by the attacking radical species with release of water. 7.) Proton addition. 8.) Heterolytic cleavage via addition of a proton and release of water. The attacking radical species is then FeO⁺⁢³. 9.) Abstraction of proton. 10.) Radical recombination. 11.) Extrusion of product for both pathways.

Reactions which involve stepwise electron abstraction and oxygen rebound to electropositive centers. The generally accepted mechanism of cytochrome P-450 for aliphatic hydroxylation⁴⁵,⁴⁶ is shown in Figure 12

**Energetic stability of hydroxylation products**

P-450 potentially generates unstable hydroxy intermediates.

Metabolic pathways such as those involving P-450 may lead to unstable hydroxyl intermediates. The energetic stability is controlled by factors such as conjugation and tautomerism as illustrated in Figure 13. Transforms must be
Figure 13. Examples of tautomerism which require quantum mechanical treatment to complement normal META transforms. Left: A reaction that should proceed to the keto form, that is, the transform should hit. Right: A structure that should remain in the enol form, that is, the transform should not hit.

Included in the spontaneous reaction dictionary that can recognize and convert these species to their stable product.

Calculation of free energy for all possible metabolites is not feasible.

The equilibrium constant of spontaneous reactions might be estimated by calculation of the ΔG of the candidate reaction. If the ΔG is negative, then the transform would proceed to form product(s). While a single thermodynamic calculation might not be computationally expensive, the ΔG of all possible pathways would need to be calculated. The great number of metabolic possibilities makes this thermodynamic approach unreasonable.

Alternatively, a parameterization using quantum mechanical coefficients of the substrate can be used to estimate the stability of the intermediates. The Huckel molecular orbital method partitions the π-electron density between the atoms of a molecule. The resulting distribution may provide some indication of the position of equilibrium, without requiring each and every possible metabolite to be tested. This method identifies substrates that require QM treatment, and applies the transform if and only if the result is a stable product.
Tautomerism

Three types of tautomers that are relevant to P-450 hydroxylation or other enzyme systems are listed in Figure 14. P-450 hydroxylation of aromatic compounds may produce enol intermediates which may be more stable in their keto-form. The $\Delta H_{\text{rxn}}$ of the tautomeric reaction is $-10 \text{ kcal/mol}$, as calculated by bond additivity of the META transform. However, if the fragment is involved in an aromatic ring, the equilibrium may shift toward the left because the $\Delta H_{\text{rxn}}$ of the fragment is offset by the resonance energy (about $36 \text{ kcal/mol}$ for benzene).

Isoamides (iminols) may result when P-450 hydroxylates a heterocyclic ring. For acyclic structures, the equilibrium is strongly shifted toward the side of the amide. Beak reports that equilibria between a wide variety of isoamides and the corresponding amides have been shown to favor the amide. However, the $\Delta H_{\text{rxn}}$ of the corresponding transform is $+9 \text{ kcal/mol}$, indicating that the isoamide form should be favored. This apparent inconsistency can be resolved by taking solvent effects and intermolecular H-bonding into account. Interestingly, even resonance energy, which has been intensively investigated experimentally and computationally by the
archetypal 4-hydroxypyridine / 4-pyridone system,\textsuperscript{49-56} cannot overcome such effects.

Monoamine oxidase (Chapter 2, p. 26) can generate an imine by the dehydrogenation of primary or secondary amines. Like keto-enol tautomerism, equilibrium lies strongly on the side of the imine for acyclic structures, unless there is no hydrogen on the nitrogen.\textsuperscript{57,58} Examples of a tautomeric shift toward the imine form are scarce among conjugated heterocyclic forms,\textsuperscript{59,60} probably because the resonance energy offsets $\Delta H_{\text{rxn}}$.

\textit{Hückel method}

Since the three-dimensional geometry of the molecules submitted to the program is not generally available and its determination would be highly computationally time consuming, the possibility of using a simple connectivity-based quantum mechanical method was explored. Since the Hückel theory accurately predicts aromaticity, it might be used to detect the resonance for the enol form and decide whether or not to proceed to the keto form. In practice, most of the structures input to META are relatively simple conjugated moieties, thus lending to the suitability of a simple Hückel-type molecular orbital (HMO) method which usually provides sufficient information to handle these types of problems. Van Catledge's parametrization, which is self-consistent, gives good results for this purpose.\textsuperscript{61}

The HMO method approximates the Hamiltonian of the organic molecular structures with $n$ atoms by focusing on the $\pi$-electrons. The $2p$ atomic orbitals are linearly combined into $\pi$ molecular orbitals. The energy levels of the LCAO molecular orbitals are solved by substitution into the Schrödinger wave equation.
Integrals are produced that are named \( \alpha \) (coulomb) and \( \beta \) (resonance). The coulomb integral approximates the energy of the atomic orbital, and the resonance integral approximates the energy difference of bonding two atomic orbitals together. Using linear algebra techniques to solve the system of \( n \) equations with \( n \) unknowns, the eigenvectors yield the coefficients of the molecular orbitals and the eigenvalues yield the energies.

Van Catledge parameterized \( \alpha \) and \( \beta \) for all combinations of C, B, N, O, F, Si, P, S, and Cl by fitting them to a semi-empirical method developed by Parisier-Parr-Pople which in turn used the Beveridge-Hinze parameter set. The usual Hückel approximations were used: 1.) off-diagonal effective Hamiltonians that are not bonded are parameterized as zero. 2.) on-diagonal overlap between identical 2p orbitals is one 3.) off-diagonal overlap between different 2p orbitals is zero. The HMO/PPP method was accomplished with the META subroutines MATX, MATMO, and MDESC.

**Perturbation theory of reactivity**

In the late 1960's, Klopman formulated a simple and powerful quantitative view of chemical reactivity. A chemical reaction always involves a movement of at least one electron from a donor orbital to an acceptor orbital. However, the problem is to predict which orbitals in the reactants are involved. Clearly, the movement which results in the greatest reduction in the energy will be most favored. Klopman recognized that quantum mechanical perturbation theory could be used to predict how the orbitals involved in the reaction interacted. According to first order perturbation theory, the smaller the energy difference in the orbitals, the more the resulting molecular orbitals "repel" each other as they approach (assuming the
overlap of the wavefunctions of the orbitals is not zero due to symmetry), forming lower energy molecular orbitals that may accommodate electrons. Fukui\(^{64}\) recognized that the HOMO of one reactant and the LUMO of the other reactant (collectively called the *frontier* orbitals) are usually the orbitals involved in the electron movement. The closer that the HOMO and LUMO are in energy, then the stronger the perturbation, and the resulting molecular orbital will be much lower in energy, and a covalent bond can form. If the HOMO and LUMO are distant in energy, then only an ionic bond can form as the electrons are transferred.

This continuum in perturbation was coined by Klopman to be *charge control* vs. *orbital control*. The Klopman-Salem equation quantitates the above argument for all pairings between all orbitals. Klopman\(^{65}\) used generalized perturbation theory to successfully explain the chemical reactivity of most types of organic reactions, including supermoleculelphilicity, captodative stabilization, free radicals, and most notably, Pearson's empirical categorization of hard and soft acids and bases. The hydroxylation and epoxidation reactions by cytochrome P-450 should obey the Klopman-Salem equation.

**METHODS AND RESULTS**

*Lipophilicity is predicted using the group contribution method*

The LOGP program written by Klopman and Wang,\(^{23}\) estimates the partition coefficient of every molecule submitted to, or generated by, the META program. The LOGP program is based on the group contribution of star-centered fragments that were determined with a learning set of 935 compounds. The LOGP program can predict the log \(P\) within or very near to experimental error (about 0.5 units). In
Figure 15. Parameterization of hydroxylation and epoxidation. Each point represents a double bond from the training set of compounds (Table 6). The ordinate is the nucleophilic index. The abscissa is the epoxidation knowledge, which is +1 if it is known to epoxidize, -1 if it is known not to epoxidize, -0.5 if it is assumed not to epoxidize, and 0 if its status is indeterminate. The cutoff of −0.8 discriminated the substrates that undergo epoxidation; exceptions are listed in the text.

In the present version of META, the priorities of transforms involved in cytochrome-P-450 hydroxylation are fine-tuned (in subroutine MCOM) in the following manner:

- if \( \log P > 1.25 \), then priority = base priority
- if \(-0.75 < \log P \leq 1.25\), then priority = base priority +2
- if \( \log P \leq -0.75 \), then priority = base priority +4

Excretion by the kidneys is simulated by terminating further metabolism if the \( \log P \) of the xenobiotic is negative (Chapter 1).
Table 6. Epoxidation/hydroxylation training set.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Epoxidation Product</th>
<th>Hydroxylation Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2-difluorobenzene</td>
<td>benzo[a]pyrene</td>
<td>dibenzo[a,h]pyrene</td>
</tr>
<tr>
<td>1,3-difluorobenzene</td>
<td>benzo[a]pyrene-7,8-dihydrodiol</td>
<td>dibenzo[a,i]pyrene</td>
</tr>
<tr>
<td>1-nitropyrene</td>
<td>benzo[b]fluoranthene</td>
<td>dibenzo[a,e]fluoranthene</td>
</tr>
<tr>
<td>12-methylibenz[a]anthracene</td>
<td>benzo[c]phenanthrene</td>
<td>fluoranthene</td>
</tr>
<tr>
<td>2-nitrofluorene</td>
<td>benzo[e]pyrene</td>
<td>fluorobenzene</td>
</tr>
<tr>
<td>5-methylchrysene</td>
<td>benzo[f]quinoline</td>
<td>isoprene</td>
</tr>
<tr>
<td>6-methylchrysene</td>
<td>benzo[j]fluoranthene</td>
<td>naphthalene</td>
</tr>
<tr>
<td>6-nitro-5-methylchrysene</td>
<td>benzo[k]fluoranthene</td>
<td>nitrobenzene</td>
</tr>
<tr>
<td>7,12-dimethylibenz[a]anthracene</td>
<td>bromobenzene</td>
<td>phenanthrene</td>
</tr>
<tr>
<td>aniline</td>
<td>cholanthrene</td>
<td>phenol</td>
</tr>
<tr>
<td>anthanthrene</td>
<td>chrysene</td>
<td>pyrene</td>
</tr>
<tr>
<td>anthracene</td>
<td>dibenz[a,h]anthracene</td>
<td>styrene</td>
</tr>
<tr>
<td>benzene</td>
<td>dibenz[a,j]anthracene</td>
<td>toluene</td>
</tr>
<tr>
<td>benzo[a]anthracene</td>
<td>dibenz[a,j]acridine</td>
<td>triphenylene</td>
</tr>
</tbody>
</table>

**Epoxidation and hydroxylation are predicted with GP theory**

Two special transforms, one for epoxidation and one for hydroxylation, were included in the META dictionary. Each transform recognizes a conjugated double bond from the substrate and evokes a computational facility that assesses the nucleophilic character of each of the conjugated double bonds. Regular transforms for the generation of bridging epoxides, such as in the case of naphthalene 1,4-oxide, were included in the META dictionary.

A training set of xenobiotics that undergo epoxidation or hydroxylation was compiled from literature precedent.\textsuperscript{66,67} A database of 42 compounds was compiled (Table 6) which contained PAHs, substituted benzenes, and alkenes whose experimental epoxidation and hydroxylation status was known. To parameterize successfully, double bonds that are known not to epoxidize must be included in the database. Since such negative data is rare (Chapter 2), some reasonable assumptions were made in accordance with the literature.

A procedure based on Klopman's previous description of charge and orbital control\textsuperscript{63} was developed, whereas the ability of a bond to undergo a concerted nucleophilic addition is mostly determined by its contribution to the HOMO of the
molecule. The following nucleophilic index $N_{xy}$ was therefore defined for each bond:

$$N_{xy} = 2 \sum \frac{(c_{ix} + c_{iy})^2}{0.1 + E_i}$$  \hspace{1cm} (Eq. 1)

where $c_{ix}$ and $c_{iy}$ are the Huckel coefficients of the atomic orbital of the bonded atoms $x$ and $y$ in the occupied molecular orbital $i$ and $E_i$ is the HMO energy of the molecular orbital $i$. The advantage of this index is that it incorporates the orbital symmetry constraints needed to evaluate the potential for concerted addition to the bond.

$N_{xy}$ values were calculated for a total of 253 CH=CH double bond substrate fragments derived from the training set and their $N_{xy}$ values are graphed in Figure 15. Bonds with a calculated index value greater than ~0.80 correlated very well with the experimental epoxidation of the partial double bonds. The epoxidation patterns of the 31 PAH's in the database are successfully predicted, except for one partial double bond in benzo[j]fluoroanthene. Substituted benzenes were clustered near the cutoff of 0.8; benzene, phenol, and toluene were on the borderline. Two alkenes also epoxidized as expected. The $N_{xy}$ of the alkenes are similar and well above the cutoff. This corresponds to the experimental fact that double bonds are almost always epoxidized. Ortiz de Montellano\textsuperscript{68} states, "The only general pathway for the metabolism of an olefin function is its oxidation to an epoxide." When $N_{xy}$ is less than 0.80, hydroxylation becomes possible and is predicted by using the HOMO charge densities on each atom, thus modeling direct hydroxylation. In this way the hydroxylation patterns for substituted benzenes have been largely reproduced. Nine substituted benzenes (e. g. aniline) hydroxylated according to experiment.
<table>
<thead>
<tr>
<th>Enol name</th>
<th>S</th>
<th>$B_{12}$</th>
<th>$B_{23}$</th>
<th>$Q_1$</th>
<th>$Q_2$</th>
<th>$Q_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-naphthol</td>
<td>+</td>
<td>-0.141</td>
<td>0.348</td>
<td>1.961</td>
<td>0.967</td>
<td>1.018</td>
</tr>
<tr>
<td>picric acid</td>
<td>+</td>
<td>-0.087</td>
<td>0.434</td>
<td>1.948</td>
<td>0.875</td>
<td>1.016</td>
</tr>
<tr>
<td>norizalpinin</td>
<td>+</td>
<td>-0.110</td>
<td>0.447</td>
<td>1.957</td>
<td>0.929</td>
<td>1.069</td>
</tr>
<tr>
<td>m-cresol</td>
<td>+</td>
<td>-0.133</td>
<td>0.482</td>
<td>1.961</td>
<td>0.967</td>
<td>1.058</td>
</tr>
<tr>
<td>norizalpinin</td>
<td>+</td>
<td>-0.110</td>
<td>0.486</td>
<td>1.957</td>
<td>0.929</td>
<td>1.073</td>
</tr>
<tr>
<td>o-cresol</td>
<td>+</td>
<td>-0.159</td>
<td>0.490</td>
<td>1.963</td>
<td>0.998</td>
<td>1.026</td>
</tr>
<tr>
<td>phenol</td>
<td>+</td>
<td>-0.133</td>
<td>0.491</td>
<td>1.961</td>
<td>0.969</td>
<td>1.027</td>
</tr>
<tr>
<td>m-cresol</td>
<td>+</td>
<td>-0.133</td>
<td>0.499</td>
<td>1.961</td>
<td>0.967</td>
<td>1.049</td>
</tr>
<tr>
<td>hydroquinone</td>
<td>+</td>
<td>-0.153</td>
<td>0.504</td>
<td>1.962</td>
<td>0.987</td>
<td>1.025</td>
</tr>
<tr>
<td>3-fluoro-4-hydroxy</td>
<td>+</td>
<td>-0.164</td>
<td>0.522</td>
<td>1.963</td>
<td>1.000</td>
<td>1.006</td>
</tr>
<tr>
<td>phenylacetic acid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>norizalpinin</td>
<td>+</td>
<td>-0.119</td>
<td>0.530</td>
<td>1.956</td>
<td>0.926</td>
<td>1.069</td>
</tr>
<tr>
<td>o-cresol</td>
<td>+</td>
<td>-0.159</td>
<td>0.531</td>
<td>1.963</td>
<td>0.998</td>
<td>0.989</td>
</tr>
<tr>
<td>2-naphthol</td>
<td>+</td>
<td>-0.141</td>
<td>0.635</td>
<td>1.961</td>
<td>0.967</td>
<td>1.038</td>
</tr>
<tr>
<td>1-naphthol</td>
<td>+</td>
<td>-0.164</td>
<td>0.643</td>
<td>1.958</td>
<td>0.961</td>
<td>1.038</td>
</tr>
<tr>
<td>norizalpinin</td>
<td>+</td>
<td>-0.230</td>
<td>0.734</td>
<td>1.963</td>
<td>1.037</td>
<td>0.872</td>
</tr>
<tr>
<td>3-hydroxyhexatriene</td>
<td>+</td>
<td>-0.195</td>
<td>0.915</td>
<td>1.956</td>
<td>0.955</td>
<td>1.057</td>
</tr>
<tr>
<td>hydroxystilbene</td>
<td>0</td>
<td>-0.190</td>
<td>0.931</td>
<td>1.956</td>
<td>0.955</td>
<td>1.063</td>
</tr>
<tr>
<td>enol of phenylacetaldelyde</td>
<td>-</td>
<td>-0.214</td>
<td>0.975</td>
<td>1.949</td>
<td>0.940</td>
<td>1.076</td>
</tr>
<tr>
<td>enol of acetone</td>
<td>-</td>
<td>-0.151</td>
<td>1.044</td>
<td>1.955</td>
<td>0.913</td>
<td>1.185</td>
</tr>
<tr>
<td>enol 2 of methyl ethyl ketone</td>
<td>-</td>
<td>-0.151</td>
<td>1.044</td>
<td>1.955</td>
<td>0.913</td>
<td>1.185</td>
</tr>
<tr>
<td>enol 2 of methyl propyl ketone</td>
<td>-</td>
<td>-0.151</td>
<td>1.044</td>
<td>1.955</td>
<td>0.913</td>
<td>1.185</td>
</tr>
<tr>
<td>enol 2 of methyl butyl ketone</td>
<td>-</td>
<td>-0.151</td>
<td>1.044</td>
<td>1.955</td>
<td>0.913</td>
<td>1.185</td>
</tr>
<tr>
<td>enol 2 of methyl isobutyl ketone</td>
<td>-</td>
<td>-0.151</td>
<td>1.044</td>
<td>1.955</td>
<td>0.913</td>
<td>1.185</td>
</tr>
<tr>
<td>methyl amyl ketone enol 2</td>
<td>-</td>
<td>-0.151</td>
<td>1.044</td>
<td>1.955</td>
<td>0.913</td>
<td>1.185</td>
</tr>
<tr>
<td>cyclohexyl acetone enol 1</td>
<td>-</td>
<td>-0.151</td>
<td>1.044</td>
<td>1.955</td>
<td>0.913</td>
<td>1.185</td>
</tr>
<tr>
<td>pentafluorocacetone</td>
<td>+</td>
<td>-0.239</td>
<td>1.221</td>
<td>1.960</td>
<td>0.991</td>
<td>1.128</td>
</tr>
<tr>
<td>cyclopentanone enol</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>cyclohexanone enol</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>enol 1 of methyl ethyl ketone</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>enol 1 of methyl propyl ketone</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>enol 1 of methyl butyl ketone</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>enol 2 of methyl butyl ketone</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>methyl amyl ketone enol 1</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>diethyl ketone enol</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>diisopropyl ketone enol</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>ethyl propyl ketone enol 1</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>ethyl propyl ketone enol 2</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>ethyl butyl ketone enol 1</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>ethyl butyl ketone enol 2</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>ethyl amyl ketone enol 1</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>ethyl amyl ketone enol 2</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>cyclohexyl acetone enol 2</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>camphor enol 1</td>
<td>-</td>
<td>-0.262</td>
<td>1.259</td>
<td>1.960</td>
<td>1.003</td>
<td>1.126</td>
</tr>
<tr>
<td>disobutyl ketone enol</td>
<td>-</td>
<td>-0.401</td>
<td>1.503</td>
<td>1.965</td>
<td>1.089</td>
<td>1.071</td>
</tr>
</tbody>
</table>
Table 8. Hückel coefficients for some unstable isomides. Structures identified by roman numerals refer to Bodor.69

<table>
<thead>
<tr>
<th>Isoamide name</th>
<th>B_{12}</th>
<th>B_{23}</th>
<th>Q_1</th>
<th>Q_2</th>
<th>Q_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIIIa</td>
<td>0.231</td>
<td>0.626</td>
<td>1.953</td>
<td>0.851</td>
<td>1.269</td>
</tr>
<tr>
<td>VIIIa</td>
<td>0.231</td>
<td>0.626</td>
<td>1.953</td>
<td>0.851</td>
<td>1.269</td>
</tr>
<tr>
<td>XLVIIIb</td>
<td>0.237</td>
<td>0.634</td>
<td>1.952</td>
<td>0.821</td>
<td>1.270</td>
</tr>
<tr>
<td>XLVIIIb</td>
<td>0.232</td>
<td>0.634</td>
<td>1.952</td>
<td>0.848</td>
<td>1.253</td>
</tr>
<tr>
<td>VIIa</td>
<td>0.218</td>
<td>0.634</td>
<td>1.956</td>
<td>0.895</td>
<td>1.240</td>
</tr>
<tr>
<td>XLVIIIb</td>
<td>0.237</td>
<td>0.636</td>
<td>1.952</td>
<td>0.822</td>
<td>1.244</td>
</tr>
<tr>
<td>2-hydroxypyrimidine</td>
<td>0.237</td>
<td>0.637</td>
<td>1.952</td>
<td>0.823</td>
<td>1.226</td>
</tr>
<tr>
<td>VIIa</td>
<td>0.237</td>
<td>0.637</td>
<td>1.952</td>
<td>0.820</td>
<td>1.269</td>
</tr>
<tr>
<td>XLVIIIb</td>
<td>0.237</td>
<td>0.638</td>
<td>1.952</td>
<td>0.822</td>
<td>1.253</td>
</tr>
<tr>
<td>IVa</td>
<td>0.218</td>
<td>0.639</td>
<td>1.956</td>
<td>0.895</td>
<td>1.222</td>
</tr>
<tr>
<td>XLVIIIb</td>
<td>0.237</td>
<td>0.641</td>
<td>1.952</td>
<td>0.820</td>
<td>1.289</td>
</tr>
<tr>
<td>VIIib</td>
<td>0.252</td>
<td>0.642</td>
<td>1.946</td>
<td>0.786</td>
<td>1.337</td>
</tr>
<tr>
<td>XLVIIIc</td>
<td>0.256</td>
<td>0.656</td>
<td>1.945</td>
<td>0.777</td>
<td>1.319</td>
</tr>
<tr>
<td>XLIXc</td>
<td>0.236</td>
<td>0.659</td>
<td>1.950</td>
<td>0.848</td>
<td>1.305</td>
</tr>
<tr>
<td>XLVIIIe</td>
<td>0.276</td>
<td>0.680</td>
<td>1.940</td>
<td>0.712</td>
<td>1.293</td>
</tr>
<tr>
<td>DXa</td>
<td>0.227</td>
<td>0.688</td>
<td>1.953</td>
<td>0.869</td>
<td>1.252</td>
</tr>
</tbody>
</table>

Table 9. Training set enamines. Structures identified by roman numerals refer to Bodor.69 (S=stability)

<table>
<thead>
<tr>
<th>Enamine name</th>
<th>S</th>
<th>B_{12}</th>
<th>B_{23}</th>
<th>Q_1</th>
<th>Q_2</th>
<th>Q_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>XXXIXb</td>
<td>-</td>
<td>1.812</td>
<td>0.796</td>
<td>1.183</td>
<td>-0.078</td>
<td>0.799</td>
</tr>
<tr>
<td>lb</td>
<td>-</td>
<td>1.646</td>
<td>0.841</td>
<td>1.084</td>
<td>0.0035</td>
<td>0.762</td>
</tr>
<tr>
<td>XXXVIIIb</td>
<td>-</td>
<td>1.684</td>
<td>0.893</td>
<td>1.182</td>
<td>0.186</td>
<td>0.660</td>
</tr>
<tr>
<td>XXXVIIIa</td>
<td>+</td>
<td>1.653</td>
<td>0.930</td>
<td>1.202</td>
<td>0.355</td>
<td>0.744</td>
</tr>
<tr>
<td>XXXVIIIa</td>
<td>+</td>
<td>1.883</td>
<td>0.930</td>
<td>1.202</td>
<td>-0.275</td>
<td>0.744</td>
</tr>
<tr>
<td>3-methyladenine</td>
<td>+</td>
<td>1.851</td>
<td>0.932</td>
<td>1.158</td>
<td>-0.512</td>
<td>1.259</td>
</tr>
<tr>
<td>1-methyl-2-phenyl enamine</td>
<td>+</td>
<td>1.883</td>
<td>0.939</td>
<td>1.185</td>
<td>-0.219</td>
<td>1.011</td>
</tr>
<tr>
<td>XXXa</td>
<td>+</td>
<td>1.703</td>
<td>0.971</td>
<td>1.249</td>
<td>0.382</td>
<td>0.958</td>
</tr>
<tr>
<td>XXXa</td>
<td>+</td>
<td>1.891</td>
<td>0.971</td>
<td>1.249</td>
<td>-0.503</td>
<td>0.958</td>
</tr>
<tr>
<td>XXVIIIa</td>
<td>+</td>
<td>1.687</td>
<td>0.974</td>
<td>1.218</td>
<td>0.284</td>
<td>1.018</td>
</tr>
<tr>
<td>XXVIIIa</td>
<td>+</td>
<td>1.892</td>
<td>0.974</td>
<td>1.218</td>
<td>-0.460</td>
<td>1.018</td>
</tr>
<tr>
<td>Pyrazole</td>
<td>+</td>
<td>1.606</td>
<td>0.995</td>
<td>1.120</td>
<td>0.268</td>
<td>0.665</td>
</tr>
<tr>
<td>XXXa</td>
<td>+</td>
<td>1.885</td>
<td>0.995</td>
<td>1.190</td>
<td>-0.869</td>
<td>1.545</td>
</tr>
<tr>
<td>XLIIa</td>
<td>+</td>
<td>1.911</td>
<td>1.023</td>
<td>1.149</td>
<td>-0.403</td>
<td>1.046</td>
</tr>
<tr>
<td>Imidazole</td>
<td>+</td>
<td>1.635</td>
<td>1.048</td>
<td>1.083</td>
<td>0.112</td>
<td>0.839</td>
</tr>
<tr>
<td>XXXa</td>
<td>+</td>
<td>1.917</td>
<td>1.060</td>
<td>1.208</td>
<td>-0.486</td>
<td>1.237</td>
</tr>
<tr>
<td>XXXa</td>
<td>+</td>
<td>1.917</td>
<td>1.064</td>
<td>1.146</td>
<td>-0.441</td>
<td>1.012</td>
</tr>
<tr>
<td>XLIIa</td>
<td>+</td>
<td>1.679</td>
<td>1.096</td>
<td>1.085</td>
<td>-0.067</td>
<td>1.016</td>
</tr>
<tr>
<td>XXVIIIa</td>
<td>+</td>
<td>1.687</td>
<td>1.107</td>
<td>1.122</td>
<td>-0.015</td>
<td>0.962</td>
</tr>
<tr>
<td>XXXa</td>
<td>+</td>
<td>1.668</td>
<td>1.146</td>
<td>1.064</td>
<td>-0.003</td>
<td>1.012</td>
</tr>
<tr>
<td>XLIIa</td>
<td>+</td>
<td>1.645</td>
<td>1.149</td>
<td>1.023</td>
<td>-0.064</td>
<td>1.046</td>
</tr>
<tr>
<td>XXXa</td>
<td>+</td>
<td>1.703</td>
<td>1.208</td>
<td>1.060</td>
<td>-0.281</td>
<td>1.237</td>
</tr>
</tbody>
</table>
Figure 16. Key to the quantum mechanical indices corresponding to the target fragments in the spontaneous dictionary for a.) enol b.) isoaamde (iminol) and c.) enamine tautomerization.

A restricted procedure similar to that used by META was used to prioritize competing epoxidations on the same PAH (Figure 9). The priority of the transform is fine-tuned from its base level by a procedure that is currently under development in the Kloppman laboratory (Chapter 4).

**Tautomerization**

The HMO coefficients of bond and charge density were correlated to the thermodynamic stability for each tautomeric form of a database of molecules that might be formed by P-450 hydroxylation. The bond orders $B_{xy}$ (Figure 16) are computed as follows with the following equation:

$$B_{xy} = 2 \sum \left( c_x c_y \right)$$  \hspace{1cm} (Eq. 2)

The bond order $B_{23}$ of the partial double bond was most successful at discriminating whether the enol is stable or should be transformed into the keto form. Figure 17 independently shows that there is a good correlation of $B_{23}$ with the $\Delta H_{rxn}$
Figure 17. Correlation of $\Delta H_{rxn}$ with the partial double bond character as calculated from the bond order $B_{23}$. The $\Delta H_{rxn}$ was calculated by finding the difference in potential energy of the enol structure and the keto structure using the AM1 single point energy with Hyperchem (Autodesk, Inc., Sausalito CA) software defaults. Each structure was first minimized with MM+ molecular mechanics force field. Data points are: 5 tautomeric reactions (left to right, enol forms only are listed) 2-isopropen-1-ol, cyclohexen-1-ol, 1-naphthol, phenol, $p$-chlorophenol.

determined through molecular modeling of some example tautomerisms. Other possibilities for a good separation were tested including Brown's index\textsuperscript{65} as well as the charge densities for all bonds and atoms, for a total of ten different indices.

The training compounds for keto-enols tautomerization (Table 7) were taken from various sources.\textsuperscript{48,70-72} Partial bond order $B_{23}$ between the carbons was completely discriminating. Of the 38 enols evaluated, those with $B_{23}$ less than 0.95 are stable as enols, while the remaining are stable in their keto form. This cutoff was used in META.

The Mulliken-type charge densities $Q_x$ (Figure 16) were calculated as follows:
\[ Q_c = 1 - 2 \sum_{i} c_i^2 \] (Eq. 3)

Table 7 shows that the charge density \( Q_3 \) on the non-carbonyl carbon was also discriminating. If the charge density was less than 1.06, then the enol was stable.

A similar training set for isoamide-amide tautomerization was compiled. Due to the scarcity of experimentally observed data for isoamides, quantum mechanically modeled isoamide data was introduced into the database. Fifteen structures containing the isoamide functional group were selected from a paper by Bodor addressing heterocyclic molecules. Vinylogous amides were excluded. No stable isoamides could be found (experimental or calculated) but their corresponding amides are known to be stable. Due to this fact, QM treatment was not required but the indices are listed in Table 8 for general interest.

The training set for enamine-imine tautomerization is listed in Table 9. Eighteen structures contained the enamine functional group, with most being selected from Bodor's paper. Bond density \( B_{23} \) between the carbons was highly discriminating; if the \( B_{23} \) was calculated to be less than 0.92, then the enamine was assumed stable.

**DISCUSSION**

*Epoxidation/hydroxylation parameterization is consistent with the consensus mechanism*

The generally accepted mechanism for aliphatic hydroxylation (Figure 12) probably applies at least partially to the hydroxylation and epoxidation of unsaturated compounds. An interesting similarity exists between the aliphatic mechanism and the hydroxylation and epoxidation of aromatic compounds. The heterolytic pathway forms an oxygen species ("oxene") that is isoelectronic to
carbene, whereas the homolytic pathway forms a hydroxy radical. Both are bound to a positively charged iron atom. The putative singlet oxene may add to a double bond, just as carbene does, by forming a three-membered ring. There is ample literature precedence for the hydroxylation of double bonds by the hydroxy radical. The pathway divergence in the aliphatic hydroxylation mechanism may correlate with the epoxidation vs. hydroxylation dichotomy.

The structure of the substrate may influence the nature of the cleavage of the perferryl intermediate in cytochrome P-450 by electronic effects. Upon binding, the substrate may donate electrons and favor heterolytic cleavage, or withdraw electrons and favor homolytic cleavage. This would imply that a higher partial double bond density in the substrate would encourage heterolytic cleavage, formation of the oxene, and epoxidation. Lower partial bond density in the substrate would encourage homolytic cleavage, formation of the hydroxyl radical, and hydroxylation.

Both attacking species are electrophilic if for no other reason than their association with the positively charged iron. Viewed in the context of GP theory, the π-molecular orbitals of the double bond would transfer at least one or two electrons from its HOMO to the SOMO of the hydroxyl radical, or the LUMO of the oxene. Since the SOMO and the HOMO are very similar in energy, the HOMO-LUMO interaction with the oxene would be more orbitally controlled than the HOMO-SOMO interaction with the hydroxyl radical. Thus, epoxidation would be orbitally controlled, and hydroxylation would be charge controlled, which is consistent with the parameterization using GP theory as described above.
Modeling the hydrophobic pocket with graph indices

As the mechanism of P-450 hydroxylation is becoming clearer, biochemists are focusing their attention on the binding site of cytochrome P-450, which is probably the main source of the variance of the substrate specificity of different P-450 isozymes. The mechanism of P-450 (Figure 12) is most likely the same for the isozymes. A series of substituted camphors are known to have quite different hydroxylation regio-specificities, which may be due to varying orientation of the substrate with respect to the activated perferryl intermediate. White et. al. studied the stereochemistry of aliphatic hydroxylation and found that a discrete tricoordinate carbon intermediate was formed, implying that hydroxylation stereospecificity must be enforced by the steric forces of P-450. Schwarze et. al. have performed detailed molecular modeling of the binding site. As studies like these continue to define the hydrophobic binding site, it may become possible to model the binding site with graph indices (Chapter 5, p. 112).

Tautomerism

Comparison of $B_{23}$ in tables 7 and 8 might lead to the expectation that the isoamides should be stable when they actually are not. The HMO calculations for isoamide-amide tautomerism appears not to be helpful. Beak concurs: "Our studies show that the quantum mechanical theories which have been applied to the determination of the relative stabilities of such isomers do not give useful results." This is due to solvent effects and intramolecular H-bonding. In gas phase, the QM indices developed in this chapter may be more predictive.

The methodologies developed in this chapter can be extended to other unstable proton tautomeric forms that may be generated by META. Such
tautomeric reaction include thione-thiol,76 aminal-amidine,77 azo-hydrazone,78,79 acetylene-allene, nitrile-ketimine, nitroso-oxide, and phosphorane-phosphate.80

CONCLUSION

Several methodologies have been discussed that extend basic META methodology, allowing accurate modeling of enzymes that catalyze diverse reactions. Methods to predict lipophilicity are used by META to adjust the priority of the transforms of P-450 and to model excretion of the xenobiotic or its metabolites from the organism. A parameterization based on GP theory successfully predicted the epoxidation and hydroxylation of unsaturated compounds by P-450. The bond orders and atomic charge densities were shown to predict the position of tautomeric equilibrium accurately. Certainly, other methodologies such as graph indices can be used to extend META even further.
CHAPTER 4

META-4: An Algorithm that Optimizes Transform Priorities
INTRODUCTION

In the first chapter, an overview of the underlying logic and operation of the META program and the concept of a dictionary (which acts as the knowledge base) was described. Used with the appropriate dictionary, the META program applies transforms (Figure 17) to substrates that are recognized. In the second chapter, a dictionary of transforms was described that models the metabolism of xenobiotics in mammals. As previously discussed, the concentration of observed metabolites of the biotransformations can range from major to barely detectable, allowing intuitive prioritization of the transforms by the experts. Even though this prioritization may be approximate, it is very important to model it as completely as possible due to the combinatorial explosion of potential metabolic pathways. Without any "metabolic compass" that indicates the most important metabolic pathways, the investigator may be overwhelmed with large numbers of irrelevant metabolic predictions.

In this chapter, a novel algorithm called META-4 is described that optimizes the priorities of the transforms directly from specific xenobiotic biotransformation data acquired from the pharmacological literature, such that the number of correct metabolic predictions are increased while sorting the predictions with respect to pharmacological significance. As in the concept of transforms, the new prioritization should generalize to new xenobiotics.

There are at least three reasons for the need for the META-4 algorithm. First, due to the large amount of substrate data and transforms in a normally-sized dictionary, reprioritization is labor-intensive and well-suited for computer automation. Without an algorithm, a dictionary involving hundreds of transforms and xenobiotic data could be manually prioritized only at a low-resolution level.
a.) [8.1] Glycination of aromatic carboxylic acids
b.) R 1p101; 23p54
c.) D -1 2:2 -3 -4 6:1
d.) X 3 8
e.) OH -CO -C =
f.) NH -CO -C =<1-CH2-CO -OH >

Figure 18. An example of an actual transform used in the META dictionary. Line a.) The title of the transform which describes the biotransformation rule being modeled. Line b.) Reference pointers to the literature that are listed in footnotes at the end of the dictionary. Line c.) The D-field which is composed of the hit/miss and P-order information for all the substrates relevant to the transform. Each substrate number refers to SMILES or KLN line codes at the end of the dictionary. Line d.) Marker indicating that the next two lines are a transform, followed by the priority of the transform (v=3) and enzyme class (8). Line e.) Target fragment. Line f.) Product fragment. Similar transforms are encoded for glutamine conjugation, glucuronic acid conjugation, sulfate conjugation, and hydroxylation for our exemplary aromatic acid dictionary (Figure 19).

Second, a computer algorithm assigns priorities on a non-subjective basis, allowing diverse dictionary-development personnel to add new transforms in a methodical manner without needing to know the intent of the other existing transforms. Third, the assignment of priorities may be conceptually difficult because META transforms are strongly interdependent by nature. Ignorance of the interdependencies may cause a well-intentioned change in the dictionary to have far-reaching deleterious consequences such as perturbing or even undoing previous work by oneself or others. Using an algorithm minimizes the requirement for a global understanding of the dictionary (i.e., expert status) during prioritization.

This chapter will develop the META-4 algorithm from META methodology axiom using graph-theoretic concepts. Our goal is to demonstrate the utility and feasibility of the algorithm. Necessary data structures and execution of a prototype will be demonstrated with the aid of a training set of aromatic acids (Figure 19) whose mammalian metabolic routes are well-known. To the best of our knowledge,
Figure 19. A simple but highly interdependent training set of xenobiotics that will be used to demonstrate META-4. These xenobiotics' line codes are stored in the dictionary and referred to by the numbers in the D-field (Figure 18). These compounds are experimentally observed to preferentially undergo one or more of the biotransformations listed in the caption to Figure 18.

META-4 is a novel approach to solving the core problem of the modeling of reaction pathways.

THEORY OF META METHODOLOGY

In order to describe the META-4 algorithm more clearly, a rigorous treatment of basic META methodology is helpful. In META methodology, transforms are virtual enzymes, so enzyme kinetics can provide a biochemical interpretation of the priority values of transforms. The Michaelis-Menten approximation\(^1\) is based on the following scheme:

\[
E + S \xrightleftharpoons[\kappa_r]{\kappa_s} ES \rightarrow E + P \quad \text{(Scheme A)}
\]

Figure 20 represents the processing for substrates \(S\) that are submitted to a dictionary to be METAbolized. The META dictionary is a set of \(m\) transforms \(E\)
with associated priority values \( v \) (Figure 18) that operate upon each particular substrate \( S_i \). If the target fragment of a given transform \( E_j \) is present in \( S_i \), then \( E_j \) is said to bind \( S_i \), in which case the complex \( ES_{i,k} \) is generated and the priority value \( v_j \) is transferred to the complex. A priority value is assigned to each transform, which ranges from \( v_{\text{min}} \) (highest priority) to \( v_{\text{max}} \) (lowest priority). (Presently, the range is \( 1 \rightarrow 9 \), with the special priority value of zero being reserved for spontaneous reactions.) Notice that a lower-valued \( v_j \) means higher priority, analogous to the ranking of, say, football teams. All the complexes are merged to form an ordered set called a binding list. The higher priority complexes on the binding list are more "important" than the lower priority complexes.

"Importance" was equated to the experimentally observed concentration of each product \( P_i \) and which was used it to assign \( v_j \) according to the following rationale. In the context of a real enzyme system (Scheme A), the lack of \( P \) ([\( P \])=0) says nothing about whether or not binding occurred. In the case that binding does not take place, that is, \( K_m = 0 \) and [ES] = 0, no \( P \) is formed. In the case where binding does occur and ES does form, the catalytic step may be inactive (\( k_{\text{cat}} = 0 \)) and the reaction velocity \( k_{\text{cat}}[\text{ES}] \) is zero, and again no \( P \) is formed. Alternatively, the presence of \( P \) shows that the reaction velocity is not zero and binding must occur.

In terms of this biochemical framework, \( v \) can be interpreted as a reaction velocity, especially in the presence of binding. However, the case of non-binding can only be modeled accurately by the absence of a META transform. Now, the matching procedure used by META (of transform recognizing substrate) imposes the assumption that the binding constant \( K_m \) is infinite. In terms of META methodology, binding of a substrate by a transform either occurs or it does not.
Figure 20. Synopsis of basic META methodology. A substrate $S_i$ is submitted for METAbolism to the dictionary which consists of $m$ transforms $E$ with their associated priority $v$. Transforms that recognize and bind $S_i$ generate a binding list of $n$ complexes $ES$, where some of its products may not be experimentally observed. After culling, $ES$ are ordered according to $v$. Only the highest priority (lowest-valued $v$) $ES$ are considered major products $P$ and displayed (shaded box). The purpose of META is to optimize the numerical values of $v$ such as to predict all and only experimentally observed $P$ for all substrates in the correct order.

occur with no middle ground. For reasons to become apparent, the required absence is not always possible. Given this case of "false and forced" binding, an alternative way to achieve $[P] = 0$ is by artificially lowering the priority of the transform, so that it can be hidden from view by a mask. In this way, bad binding is fixed by assuming a slower reaction velocity. This approach, motivated by transforms recognizing too much, offers hidden advantages.

Masking is accomplished with META methodology by including two modes: Major and All. In Major mode, masking is applied such that only the highest priority $ES$, i.e. ones that are $W$ v-units from lowest-valued $ES$ ($a_j$), are displayed. ($W=1$ in the present version of META. For purposes of computational efficiency,
low priority transforms are never applied.) Ideally, only experimentally observed products should be in this category. A product that is predicted in Major mode is said to hit, otherwise, it is a miss. Notice that the major products are still ordered in importance as determined from the literature. The maintenance of this order is half of the purpose of META-4. In All mode, masking is not applied and the entire binding list is displayed.

The pharmacological literature very rarely predicts that a substrate which possesses the appropriate chemical functionality can never bind to the enzyme and never be biotransformed. Usually, it is reported as "not experimentally detectable" (i.e., no information about existence of \( P \)) which admits both the binding and non-binding cases. Accordingly, evaluation of binding of \( S_i \) by \( E_j \) is usually indeterminate. The use of the two modes described above handles this ambiguity. META predictions that are not experimentally observed yet pharmacological conceivable can and should be reduced in priority and masked. Assuming a perfect assignment of \( v \), All mode reveals which substrates may be recognized by the enzyme and form complexes but are not appreciably converted to observed products due to undefined factors. Practically speaking, this means that the user is alerted to binding possibilities and none are arbitrarily discarded. In this way, META provides many metabolic suggestions while giving some indication of probability and therefore can direct experimental attention to possible trace (but pharmacologically reasonable) metabolites. This feature is especially important for toxicological applications.

Masking cannot fix all bad binding META transforms. In the cases where a transform is known to not bind the substrate correctly, another choice is to increase the specificity of the transform by adding more atoms to the target fragment.
Table 10. A common pharmacological example of a literature database that will be used throughout this chapter to illustrate META-4. Note that the actual transform code is irrelevant, so they are not shown here. Numerical entries in this table imply that the substrate is experimentally observed to form a product via this biotransformation; the value gives the pharmacological importance (P-order). Note that these numerical values are relative and meaningful within a single S only. A hyphen ("-") entry means that the substrate is not observed to be biotransformed by the transform. An "x" means that literature information is unavailable.

<table>
<thead>
<tr>
<th>Transform</th>
<th>$S_1$: Phenol</th>
<th>$S_2$: Benzoic acid</th>
<th>$S_3$: Phenylacetic acid</th>
<th>$S_4$: Salicylamide</th>
<th>$S_5$: Catechol</th>
<th>$S_6$: Salicylic acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_1$: Hydroxylation</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$E_2$: Sulfation</td>
<td>1</td>
<td>-</td>
<td>x</td>
<td>2</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>$E_3$: Glucuronidation</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>$E_4$: Glutamination</td>
<td>-</td>
<td>x</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$E_5$: Glycination</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>1</td>
</tr>
</tbody>
</table>

However, a fine balance is involved here. If we increase the specificity mindlessly, the dictionary will grow too rapidly and "memorize" the data. As previously discussed (Chapter 2, p. 22), the predictive power of META originates from the generalization abilities of the transforms, something which we do not want to squander. Masking allows for some overgeneralization errors to be tolerated. Therefore, the other half of the purpose of META-4 is to mask the incorrectly binding and ambiguously binding transforms yet simultaneously expose the correctly binding transforms.

DATA STRUCTURES

The literature database is supplied by the expert

Essentially, META-4 is restricted to fitting its metabolic product predictions of $P$ to experimentally observed metabolic product data by only assigning numeric values to $v$. Consequently, META-4 must have access to an expert-supplied data structure which records the following expectations established from the literature:
Table 11. Hypothetical example of substrate binding lists for each transform in a dummy META dictionary created for the sake of illustration. A plus (+) means binding by the transform and a minus (-) means the substrate was ignored by the transform.

<table>
<thead>
<tr>
<th>Substrate</th>
<th>$S_1$: Phenol</th>
<th>$S_2$: Benzoic acid</th>
<th>$S_3$: Phenylacetic acid</th>
<th>$S_4$: Salicylamide</th>
<th>$S_5$: Catechol</th>
<th>$S_6$: Salicylic acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_1$: Hydroxylation</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>$E_2$: Sulfation</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>$E_3$: Glucuronidation</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$E_4$: Glutamation</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$E_5$: Glycination</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

- The observed hit/miss status of each $E_j$ for each $S_i$. This will answer the question: which substrates should hit which transforms? A metabolic example would be that of both benzoic acid and phenylacetic acid undergoing glucuronidation but not hydroxylation.

- The relative hit order of the $P$ for each $S_i$, called $P$-order. For example, benzoic acid is preferentially biotransformed by glucuronidation over glycination (notice that both products are observed).

Supplying the data about the metabolic fate of any substrate automatically places it in the training set of $S$. Table 10 shows a simple literature database of six aromatic acid compounds. Such a table is written by META-4 to file for bookkeeping and diagnostic purposes. This data structure is input from the $D$-fields in the META dictionary, by preceding the transform with the hit/miss status and $P$-ordering information (Figure 18). This is convenient because much of the pharmacological literature is organized about a specific biotransformation, with xenobiotic archetypes supplied. Incidentally, the excavation of the literature in this per-transform style causes the tendency for contradictions of $v$ between other transforms to be inadvertently entered into the dictionary due to the local level of
Table 12. Contingency table used to evaluate substrate binding with individual transforms. The binding status can provide a course of action and highlight weaknesses in the dictionary or literature database. See text for explanation of symbols.

<table>
<thead>
<tr>
<th>Binding status</th>
<th>Observed</th>
<th>No data</th>
<th>Not observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_j$ binds $S_i$</td>
<td>TP</td>
<td>IP</td>
<td>FP</td>
</tr>
<tr>
<td>$E_j$ does not bind $S_i$</td>
<td>FN</td>
<td>IN</td>
<td>TN</td>
</tr>
</tbody>
</table>

focus. As we shall see, these inconsistencies are detected and removed by META-4. Complete knowledge of hit/miss status and $P$-ordering for each $E_j$ is not required. This robust feature permits the continuous accumulation of literature expectations with concurrent utility.

**Binding lists are calculated by META**

The META-4 algorithm requires binding lists to be calculated for each $S_i$. To do this, META is run in a batch mode which submits all substrates for which it has any literature expectations to the dictionary for binding evaluation. This is accomplished by running the META program in an autofeed mode so that all the $S_i$'s are transformed by one step only. All mode and Show-Intermediates mode is used so that the entire binding process of META is examined. Our example of six aromatic compounds from Figure 19 was used to generate Table 11, which is also output to file by META-4 for diagnostic purposes.

**Binding performance status matrix**

Binding performance is then evaluated for each $S_i$ by comparing the literature database (Table 10) to see if each $E_j$ binded correctly. There are six types of binding status as shown in Table 12, each suggesting definite actions to be taken by the dictionary developer:
• A binding true positive (TP) means that $E_j$ binded $S_i$ and a $P_{ij}$ is reported by the literature to be detected. META-4 will attempt to keep this $S_i$ from being masked.

• A binding false positive (FP) means that the $E_j$ binded $S_i$ but a $P_{ij}$ is reported by the literature to not be observed by the literature. META-4 will attempt to mask this $S_i$ caused by an overgeneralizing $E_j$. The dictionary developer may choose to manually increase the specificity of the transform, creating new "virtual enzymes" that can be prioritized separately.

• A binding false negative (FN) means that $E_j$ did not bind $S_i$ but a $P_{ij}$ is reported to be observed. The transform may be coded incorrectly or the literature information was incorrectly entered into the D-field. (Please note that at least the intent and function of the transform must be preconceived or "allocated" before a false negative can be detected.)

• A binding true negative (TN) means that $E_j$ did not bind $S_i$ and a $P_{ij}$ is reported to not be observed by the literature. No action is required.

• A binding indeterminate positive/negative means that the transform did bind/did not bind but literature expectations have not been established yet.

Here, META-4 is requesting the dictionary developer to query the literature and put the results in the D-fields (Figure 18, Table 10).

Clearly, a successful run of META-4 will mask the binding FP and IP while exposing binding TP. While META-4 does not consider TN, FN, or IN during the reprioritization process, it does calculate and tabulate their binding status (Table 12). This information directs the efforts of the dictionary developer (as a compiler focuses a programmer's attention to erroneous FORTRAN code) to $E$ and $S$ that require manual intervention or literature expectation.
Table 13. Evaluation of substrate binding with many transforms. Each entry is the binding status for each substrate with respect to each transform. The abbreviations are explained in the text.

<table>
<thead>
<tr>
<th>↓ Transform</th>
<th>( S_1 ): Phenol</th>
<th>( S_2 ): Benzoic acid</th>
<th>( S_3 ): Phenylacetic acid</th>
<th>( S_4 ): Salicylamide</th>
<th>( S_5 ): Catechol</th>
<th>( S_6 ): Salicylic acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E_1 ): Hydroxylation</td>
<td>TP</td>
<td>TN</td>
<td>FP</td>
<td>TP</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>( E_2 ): Sulfation</td>
<td>TP</td>
<td>TN</td>
<td>IP</td>
<td>TP</td>
<td>TP</td>
<td>TN</td>
</tr>
<tr>
<td>( E_3 ): Glucuronidation</td>
<td>TP</td>
<td>TP</td>
<td>TP</td>
<td>TP</td>
<td>TP</td>
<td>TP</td>
</tr>
<tr>
<td>( E_4 ): Glutamation</td>
<td>TN</td>
<td>IN</td>
<td>TP</td>
<td>TN</td>
<td>TN</td>
<td>TN</td>
</tr>
<tr>
<td>( E_5 ): Glycination</td>
<td>TN</td>
<td>TP</td>
<td>TN</td>
<td>TN</td>
<td>IN</td>
<td>TP</td>
</tr>
</tbody>
</table>

In turn, Table 13 shows the application of the contingency table (Table 12) to all \( S_i \) and \( E_j \). Individual transforms and individual substrates have statistics compiled that characterize their specificity, sensitivity, OCP, ECP, and \( \phi^2 \), all of which are used to highlight trouble spots in the dictionary. Another statistic, percentage indeterminacy with respect to both \( S \) and \( E \) allows dictionary developers to characterize their lack of expectations per substrate or transform, which directs future literature excavation by establishing an order or "strategy" of attack.

The overall dictionary performance index \( \phi^2 \) measures the binding performance of the dictionary with respect to all the substrates:

\[
\phi^2 = \frac{TP^2}{(TP+FP)(TP+FN)} + \frac{TN^2}{(TN+FN)(TN+FP)} + \frac{FP^2}{(FP+TP)(FP+TN)} + \frac{FN^2}{(FN+TN)(FN+TP)} \tag{Eq. 1}
\]

where \( TP, TN, FP, \) and \( FN \) are derived from the contingency table tabulated from the whole dictionary and training set of substrate data. Note that indeterminacy is not used in deriving this value.
Figure 21. a.) The ordering of two transforms represented graphically. The substrate $S_3$ (phenyl acetic acid) constrains $E_3$ (glucuronidation) to be of higher priority than $E_4$ (glutamation). b.) Left: $S_1$ (phenol) constrains both transforms to be of the same priority, but $S_4$ (salicylamide) and $S_5$ (catechol) constrain $E_3$ (glucuronidation) to be of higher priority than $E_2$ (sulfation). Right: all the substrates "vote" (1 vs. 3) to form an overall S-constraint. c.) An example of a cyclic ordering inconsistency which is detected and deleted by META-4.

PREREQUISITE GRAPH CONCEPTS

Since META-4 is best expressed in terms of graph-theoretical concepts, a brief discussion of their relation to META-4 is in order. Indeed, well-established graph algorithms are employed by META-4. Directed graphs, also called digraphs, can express the ordering relations of data structures such as Table 10 quite naturally. A digraph is a set of nodes with a set of edges indicating directed adjacency. A substrate may require a certain order between a pair of transforms, which we name an S-constraint. Given two transforms $E_a$ and $E_b$ that are represented by two nodes, four types of S-constraints exist which can be represented by edges.
• $E_a$ has greater priority than $E_b$, i.e., $v_a < v_b$, which is represented as an edge directed from the node corresponding to $E_a$ to the node corresponding to $E_b$. Such is the case in Figure 21a.

• $E_a$ has lower priority than $E_b$, i.e., $v_a > v_b$, which is represented as an edge directed from $E_b$ to $E_a$.

• $E_a$ has equal priority than $E_b$, i.e., $v_a = v_b$, which is represented as an edge with no direction (double-headed arrow) between $E_a$ and $E_b$.

• No ordering information exists, i.e., $E_a$ has no relation to $E_b$, in which case no edge is drawn.

Any number of $S$-constraint edges can exist between a pair of $E$-nodes. For the aromatic acid series example, the $S$-constraints between the two transforms $E_2$ (sulfation) and $E_3$ (glucuronidation) are illustrated graphically in Figure 21b. One type of ordering inconsistency immediately becomes obvious: no set of $v_j$ exists that satisfies all these $S$-ordering restraints simultaneously. To resolve such inconsistencies (which may likely occur in leading edge pharmacological research literature), a simple "voting" procedure was employed. Each edge is weighted in a manner that reflects the certainty of knowledge and they are summed to give a resultant edge as shown in Figure 21b. The voting results are retained for later purposes of deleting directed cycles (see below). Naturally, the majority vote is the best ordering that minimizes the violations of the pair of $E$-nodes. In our aromatic acid example with all weighting at unity, the best prioritization is that of glucuronidation being of higher priority than sulfation, i.e., $v_3 < v_2$.

A more insidious type of ordering inconsistency may be encountered in practical use. Suppose a new transform $E_X$ is introduced with the restraints shown in Figure 21c. In a way different than above, all $S$-constraints can not be satisfied
simultaneously by any set of $v$ since a directed cycle is formed. To minimize total ordering error for the three $E$-nodes, the $S$-constraints for one pair of $E$-nodes must be neglected, which translates to deleting the directed cycle by removing an edge. In our example, the edge labeled "(0,1)" would be removed.

As the number of transforms and substrates increases, the number of possible directed cycles increases very quickly. Such directed cycles can be deceptively large and must be detected, otherwise an endless vicious circle of manual prioritization may result. An edge in every directed cycle in the digraph must be removed in order to form a directed acyclic graph or dag. The observed data must be represented as a dag because it needs to be topologically sorted.\(^3\)
THE META-4 ALGORITHM

The purpose of META-4 is to find a set of $v$ that optimizes masking while maintaining as many $S$-constraints as possible. After initialization, the META-4 algorithm proceeds in two phases (Figure 22). The first phase imposes several types of $S$-constraints on $E$, which are represented by a digraph $G$. $G$ is then processed and topologically sorted, which supplies a point in $v$-hyperspace that is closer to a good final answer than a randomly-placed point would be. In this way, the combinatorial optimization procedures of the second phase are more readily accomplished. The second phase assigns the final numerical values to $v$ by simulated annealing.

Initialization

The basic data structures are prepared that were described in the above section. META-4 first reads all input parameters from file (Table 15). The META dictionary file is read, where the transforms, hit/miss status, $P$-order, initial $v$, and substrate names are sustained. At this point, a parsing function checks for common errors in the dictionary. The binding list data structure (Table 11) of the substrates is received from a previous META batch run.

Phase 1

In the first phase of META-4, a directed graph is created from three types of $S$-constraints: initial prioritization, $P$-order, and TP/FP separation, which shall be described now.

Dictionary development is inherently an iterative refinement, so there is usually some initial prioritization of the transforms. For example, the intuition of experts can provide a low resolution prioritization which we want to maintain. The
Figure 23. The digraph $G_I$ constructed from the literature database of Table 10. The voting results are not included, and no cyclic inconsistencies are present. However, an $S$-constraint with $E_j$ of higher priority than $E_i$ due to an initial prioritization, for example, could introduce a cyclic inconsistency (dashed arrow), which would be detected and deleted by META-4.

The first task is to construct a directed graph $G_0$ from the initial priorities $v^0_j$. One dummy $S$-constraint is imposed between each pair of $E$-nodes such as to record the initial ordering of that pair. An initial prioritization is totally ordered, therefore $G_0$ is always fully connected and acyclic. The resulting dag is weighted by the user-defined factor $w_0$.

The partial ordering $S$-constraints imposed on the transforms by $P$-order in the literature data structure was described above. The second step is to assemble a directed graph $G_I$ in an analogous manner that records these partial ordering restraints. As an example, Figure 23 is $G_I$ corresponding to the literature database of the training set (Table 10). If we restrict our attention to one substrate $S_i$, then $P_{i,j}...P_{i,k}$ are totally ordered, and a dag can be necessarily constructed for their transforms. However, the ordering of more than one $P$-lists is not necessarily consistent relative to each other. This might be visualized as merging the individual daggs corresponding to each substrate at their $E$-nodes. This process may produce
cyclic inconsistencies, and a set of $v_j$ may not exist that will reproduce the literature $P$-order of all $P$ simultaneously.

The binding status data structure imposes further $S$-constraints due to masking requirements. In order for phase 2 to be effective, the binding TP must be reasonably separated from the binding FP. The third step is to construct the masking digraph $G_2$, from the binding status data structure (Table 11). $G_2$ is constructed by adding a directed edge for each $S$, directed from the $E$-node that generated the binding TP to the $E$-node that generated the binding FP. All other combinations of binding status are ignored. Like $G_1$, $G_2$ is not necessarily a dag, and simultaneously satisfying all the masking ordering restraints imposed by the substrates may not be possible.

The final digraph $G$ is constructed by "adding" the three $S$-constraint digraphs:

$$G = w_0G_0 + w_1G_1 + w_2G_2$$  \hspace{1cm} \text{(Eq. 2)}

Addition is defined as merging common $E$-nodes together, maintaining directed adjacency information. Since the $E$-node labeling is the same for all data structures, it is simple matrix addition. The three ordering constraints are weighted by their corresponding phase 1 factors $w_0$, $w_1$, and $w_2$, which define the importance of each type of $S$-constraint as a starting point in the simulated annealing relative to the other two types.

Since $G$ must be topologically sorted, $G$ must be converted to a dag, which is now described. A digraph contains a directed cycle if and only if we can start and finish at the same node by following the edges in the indicated direction. A strongly connected component (SCC) in a digraph is a set of nodes that are mutually
accessible but mutually inaccessible to all nodes outside the set.³ An SCC with three or more nodes contains at least one directed cycle, and any directed cycle is an SCC. If all SCCs in the digraph consist of one or two nodes, then no directed cycles exist and the digraph is a dag, which can always be topologically sorted. Therefore, conversion to a dag reduces to detecting and reducing all SCCs to one or two nodes by cutting edges between nodes in the SCC. Tarjan's algorithm³ was used to determine the SCCs in $G$. For each SCC, the lowest weight edge is deleted (see above section), and if more than two edges are tied for lowest weight, the order of the $E$-nodes in the META dictionary is used to break the tie. This process is iterated until all SCCs consist of one or two nodes, which defines the dag $A_I$. The computational complexity of Tarjan's algorithm is linear, so repeated application as described above is computationally feasible and efficient.

$A_I$ specifies a good balance of ordering restraints so that the binding TP are of higher priority than the binding FP and so that the P are in observed order. The dag $A_I$ specifies only relative ordering of $E$; hence $A_I$ is topologically sorted by a simple recursive depth-first search (DFS) procedure³ for the purpose of assigning actual numerical values. In general, the values resulting from a topological sort are not unique; there are many sets of $v$ that satisfy $A_I$. However, using the DFS will yield $v$ numbered from 1 to $m$, which are then linearly scaled to range from $v_{min} \to v_{max}$. Unfortunately, these $v$ are not quite adequate for META methodology because topological sorting does not minimize the absolute differences in $v$, but rather only enforces relative ordering. Many "$v$-gaps" remain which preclude most $ES$ from being $P$. 
The goal of phase 2 is to reassign numerical values to \( v \) so that the relative order specified by \( A_I \) is maintained during masking. In other words, we want to simultaneously 1.) minimize the violation of \( S \)-constraints as specified by \( A_I \), 2.) maximize the masking of binding \( FP \), and 3.) minimize masking of binding \( TP \).

Unfortunately, this problem is intrinsically difficult and is computationally intractable to solve exactly for any real-world META dictionary. Our prioritization problem is isomorphic to the traveling salesperson problem which is known to be an NP-complete problem (Chapter 5). Such combinatorial optimization problems are good candidates for simulated annealing (SA), which has shown to give excellent solutions.\(^4\) SA has been used in many fields, one being protein structure determination using geometric distance constraints from NMR data (Chapter 6).

A Monte Carlo algorithm based on the Metropolis algorithm,\(^5\) which is based on thermodynamics, was used to simulate annealing. The newly topologically sorted and linearly scaled \( E \) provides the starting point in the \( v \)-hyperspace. Each SA step consists of generating a new candidate set of \( v \) by changing a random \( v_j \) by \( \pm 1 \) \( v \)-units, unless this yields a \( v_j \) outside the allowed range of \( v_{\text{min}} \rightarrow v_{\text{max}} \). At each SA step, an objective function \( U \) is evaluated, which is thermodynamically analogous to energy. \( U \) is a sum of the \( A_I \) order penalty \( U_{\text{order}} \) and a masking penalty, \( U_{\text{mask}} \):

\[
U = (1 - w_3) U_{\text{order}} + w_3 U_{\text{mask}} \tag{Eq. 3}
\]

A phase 2 weighting factor \( w_3 \) adjusts the balance between \( P \)-order and masking optimization.
To calculate $U_{\text{order}}$, the candidate $v$ are represented as a fully interconnected dag $A_2$. The similarity of $A_1$ and $A_2$ is assessed by summing the arithmetic product of a weighted edge in $A_1$ with the corresponding weighted edge in $A_2$:

$$U_{\text{order}} = \frac{1}{2} \left( \sum_{i=0}^{m} \sum_{j=0}^{m} \delta \cdot A_1(i, j) \cdot A_2(i, j) \right) + 1$$

(Eq. 4)

$$\partial = \begin{cases} 
-1 & \text{if edges in same direction} \\
+1 & \text{otherwise}
\end{cases}$$

$U_{\text{order}}^{\text{max}}$ is first calculated by reversing the direction of all edges of $A_1$ to $A'_1$ and recomputing according to the above equation with the substitution of $A_2$ with $A'_1$. $U_{\text{order}}^{\text{max}}$ is used to scale $U_{\text{order}}$ between 0 (perfect agreement) and +1.

$U_{\text{mask}}$ is based on the statistical measure of concordance:

$$U_{\text{mask}} = 1 - \frac{TP + TN}{TP + FP + TN + FN}$$

(Eq. 5)

where $TP$, $FP$, $TN$, and $FN$ indicate the masking (not binding) status, which is determined in the following manner. The mask position $a_i$ for each substrate $i$, is determined by scanning the binding list for the lowest $v$. The binding $TP$, IP, and FP are then re-evaluated with respect to the mask position as defined in Table 14. $U_{\text{mask}}$ ranges from 0 (perfect masking) to +1.

Table 14. Contingency table used to evaluate masking of binding list. The masking status entries of this table are used to calculate a penalty term that is minimized during the simulated annealing of phase 2.

<table>
<thead>
<tr>
<th>Masking status</th>
<th>Observed</th>
<th>No data</th>
<th>Not observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit</td>
<td>TP</td>
<td>FP</td>
<td>FP</td>
</tr>
<tr>
<td>Miss</td>
<td>FN</td>
<td>TN</td>
<td>TN</td>
</tr>
</tbody>
</table>
The difference $\Delta U$ between the $U$ for the candidate $v$ and the $U$ for the present $v$ decides whether the candidate $v$ is accepted, i.e., the SA step is taken in $v$-hyperspace. In analogy with the Boltzmann distribution, the probability, $p$, that the step is accepted is:

$$p = \exp\left(-\frac{\Delta U}{T}\right)$$  \hspace{1cm} (Eq. 6)

where $T$ is a control parameter that is analogous to thermodynamic temperature.

The above SA process is continued according to a predetermined temperature schedule. When finished, the masking status for all $S$ for all $E$ is calculated and printed to file, along with the same statistical treatment as described above which is used for evaluating binding for highlighting trouble spots.

**IMPLEMENTATION AND RESULTS**

A prototype of META-4 was written in the C programming language and executed on an Alpha DECstation 300, Model 400. Our example of aromatic amino acids (Figure 19) was run and Table 15 lists the parameters used. All initial $v$ were ignored by setting $w_0$ to zero. The importance of $P$-order and masking order were equalized by setting $w_I$ and $w_2$ to 0.5. A three-step annealing schedule was determined by choosing $T$ such that most candidate steps would be accepted early in the schedule while most would not be accepted late in the schedule. $U_{\text{order}}$ decreased from 0.29 to 0.10, indicating that the ordering constraints of $A_I$ were mostly satisfied and $U_{\text{mask}}$ decreased from 0.63 to 0.00, indicating that all good binding complexes proceeded to product and all bad binding complexes were masked. Table 16 lists results of an exemplary execution. Repeated execution with
Table 15. Input parameters and results of application of META-4 to three different META dictionaries. Column 1 list the parameters used in META-4 which must be specified for execution. Column 2 lists the values used to process our aromatic acid example. Column 3 lists the values used in a practical run of the META dictionary.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Demo</th>
<th>META.DIC</th>
<th>DEGR.DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_0$</td>
<td>0.0</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>$w_1$</td>
<td>0.5</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>$w_2$</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>$w_3$</td>
<td>0.5</td>
<td>0.75</td>
<td>0.5</td>
</tr>
<tr>
<td>$T_1$</td>
<td>0.10</td>
<td>0.01</td>
<td>0.002</td>
</tr>
<tr>
<td>Steps at $T_1$ (accepted)</td>
<td>2000 (1490)</td>
<td>2000 (1564)</td>
<td>2000 (1429)</td>
</tr>
<tr>
<td>$T_2$</td>
<td>0.05</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>Steps at $T_2$ (accepted)</td>
<td>2000 (461)</td>
<td>2000 (1564)</td>
<td>2000 (818)</td>
</tr>
<tr>
<td>$T_3$</td>
<td>0.02</td>
<td>0.0001</td>
<td>0.0005</td>
</tr>
<tr>
<td>Steps at $T_3$ (accepted)</td>
<td>2000 (11)</td>
<td>2000 (139)</td>
<td>2000 (141)</td>
</tr>
<tr>
<td>$T_4$</td>
<td>-</td>
<td>-</td>
<td>0.0002</td>
</tr>
<tr>
<td>Steps at $T_4$ (accepted)</td>
<td>-</td>
<td>-</td>
<td>2000 (30)</td>
</tr>
</tbody>
</table>

the same parameters but a different random seed all yielded similar results, indicating robustness. CPU time requirements were negligible.

While efficiency of this META-4 prototype was not a primary concern, its execution provides an upper bound on time requirements. META-4 was run with two real world examples with the parameters shown in Table 15. The META dictionary (version 2.38) consists of 159 priority-grouped transforms and 50 substrates. As reflected by the phase 1 weights, initial order is taken to be most important, and $P$-order least. Phase 1 took 17.1s of CPU time and phase 2 took 0.01s CPU per SA step. $U_{order}$ increased slightly from 0.34 to 0.40 and $U_{mask}$ decreased greatly from 0.40 to 0.22. In other words, the $P$-order was slightly sacrificed in order to obtain much better masking. The overall $U$ plummeted from
Table 16. Prioritization results of the aromatic acid example. Entries in this table are binding status (Table 12). Areas outside the boxed regions for each substrate are masked. Bold type represent the major products (P-list). Notice that the binding FP and IP are masked and the TP are not, and the order of products (P-order) is satisfied, thus fully achieving the purpose of META-4.

<table>
<thead>
<tr>
<th>Substrate →</th>
<th>( S_1 )</th>
<th>( S_2 )</th>
<th>( S_3 )</th>
<th>( S_4 )</th>
<th>( S_5 )</th>
<th>( S_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E_4 ): Glutamnation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( E_3 ): Glucuronidation</td>
<td>1</td>
<td>TN</td>
<td>IN</td>
<td>TP</td>
<td>TN</td>
<td>TN</td>
</tr>
<tr>
<td>( E_5 ): Glycination</td>
<td>2</td>
<td>TP</td>
<td>TP</td>
<td>TP</td>
<td>TP</td>
<td>TP</td>
</tr>
<tr>
<td>( E_2 ): Sulfation</td>
<td>3</td>
<td>TP</td>
<td>TN</td>
<td>TP</td>
<td>TN</td>
<td>IN</td>
</tr>
<tr>
<td>( E_1 ): Hydroxylation</td>
<td>3</td>
<td>TP</td>
<td>TN</td>
<td>FP</td>
<td>TP</td>
<td>TP</td>
</tr>
</tbody>
</table>

0.35 to 0.26, a great improvement from initial expert intuition. The DEGR dictionary (version 1.00) consists of 83 priority-grouped transforms and 45 substrates. \( U_{order} \) increased from 0.30 to 0.37 and \( U_{mask} \) decreased greatly from 0.33 to 0.03. Again, the \( P \)-order was slightly sacrificed in order to obtain a substantial improvement in masking. The overall \( U \) fell from 0.32 to 0.20. It is very likely that the penalties on these real world examples can be lowered still by tweaking the input parameters.

**DISCUSSION**

*Input parameters are determined through trial and error*

The META-4 weighting parameters allow prioritization for special purposes, as well as for finding better prioritizations. The settings for parameters \( w_0, w_1, w_2, \) and \( w_3 \) are determined according to the following guidelines. In the initial stages of dictionary development, \( w_0 \) can be set high so that expert opinion can provide default \( S \)-constraints until the substrate literature data is accumulated. Once fully developed, the dictionary can be used in at least two styles. One style, such as might be used in demonstration or education, requires high confidence in prediction
of $P$. This means that masking is much more important than $P$-order, so $w_I$ is set low and $w_2$ and $w_3$ are set high. The other style, such as might be used in discovery or research where false positives must be tolerated, would set $w_2$ and $w_3$ low.

The Klopman laboratory is in the process of defining the best input parameters for the dictionary at hand. A general observation is that masking is usually easier to achieve than $P$-ordering. Since $U$ must range between 0 and 2, $w_0$, $w_I$, $w_2$, and $w_3$ are relatively insensitive to the size of the dictionary or to its starting point. In contrast, the temperature parameter is very sensitive to the nature of the "energy" surface, which is partially determined by the weighting parameters. As a result, short runs are used to determine the SA schedule before spending larger amounts of computer time on the problem.

The topological sort gives the simulated annealing a head start

Phase 1 provides a dual purpose. First, the results of phase 1 are compiled into one directed graph, which summarizes all three types of ordering constraints needed to be balanced by phase 2. Second, phase 1 renders a good starting point for phase 2. If the topological sort was skipped and the starting point in $v$-hyperspace was instead assigned randomly, then many more steps of SA were needed and in some cases, $U$ would converge to a much higher value than when the topological sort was not skipped. Thus, the topological sort is important because it makes the prioritization problem much easier to solve. Incidentally, note that SA methods do not guarantee a unique global minimum and many very acceptable prioritizations may be possible.
How META-4 aids in dictionary development

The development of a dictionary from start to finish using META-4 proceeds as follows. An area of study is selected and experts or their writings that set the boundaries of the dictionary are located. In consultation with the experts, an initial dictionary is created that models the selected area from a general perspective. A set of substrates that is archetypal is selected and the D-fields (Figure 18) are filled out as far as possible. META-4 is run. The output data structure similar to Table 13 is used to guide the attention of the dictionary developer as described above. Many indeterminates usually result, which will indicate where literature expectations need to be established, and once obtained, the transforms will be prioritized, masked and evaluated automatically by META-4. Troublesome substrates and/or transforms will be noticed and manually fixed. Eventually, the point will be reached where reprioritization will not be helpful because the limits of the transforms' discrimination ability will outweigh the ability to mask. At this point, transform specificity must increase. This can be automated because the transform is composed of a "basis set" of atomic groups (as defined by CASE fragment syntax) that can be added to the bad transform such as to recognize only the substrates that it should recognize. This process is ongoing; new substrates that challenge the dictionary can be added at any time, and META-4 executed again.

META-4 not only assists in literature excavation, but can also assist in the consultation sessions with experts. The experts need not provide transforms; specific xenobiotic examples can be accommodated. For example, experts may supply pieces of information they may recall without reason, and they will be worked into the dictionary with META-4. However, the expert still has the option
of describing general rules, which is often useful in the initial stages of dictionary development.

**Compatibility with META priority fine-tuning procedures**

The Klopfman laboratory is presently experimenting with post-transform $v_f$-fine-tuning procedures that are based on external modules (e.g., quantum mechanical, log P prediction, etc.). Presently, META-4 cannot take such procedures into account, which effectively treats the fine-tuning as an average. Alternative treatments by META-4 may include 1.) ignoring fine-tuned transforms altogether and flag them for manual $v_f$-assignment, or 2.) increasing the specificity of the transform, thus making new virtual enzymes that do not need to be fine-tuned.

**CONCLUSION**

The problem of prioritization of META transforms is intrinsically difficult and computationally complex. In this chapter, the terminology and algorithmic concepts of META-4 using axioms of META methodology were developed. Through the use of a program prototype, META-4 was demonstrated to be a highly feasible and usable tool for META dictionary development and maintenance. It is a first step toward automatic transform generation and knowledge acquisition by META.
CHAPTER 5

Graphsort: An Efficient Algorithm that Uniquely Names Chemical Graphs
INTRODUCTION

Many of the objects that occupy the attention of chemists are described by their connectivity or topology. Indeed, repetitive recognition of such topological chemical entities (TCE's) is required by many tasks in chemistry. An efficient computer algorithm that can be used to automate the recognition of TCE's would provide a solid foundation to advanced chemical software, such as META, the expert system discussed in previous chapters. To this end, a tried and true method was developed that names TCE's uniquely by accepting a line-coded graph and returning its standardized line code. Central to its function is the canonicalization or ordering of the nodes of graphs. An algorithm named Graphsort was developed that achieves canonicalization for all graphs met in normal chemical usage. It employs novel methodology based on distance distributions of graph nodes, edge labels, and cycles. The software described here was extensively tested and determined to be fast and robust, thus providing a novel and useful tool to the computational chemist.

This chapter will discuss the theory and background relevant to Graphsort. Much of my research borrows the best parts of independent projects from different disciplines. Many of the algorithms are developed for special graph classes\(^1-3\) (labeled and not labeled) and for special purposes (graph isomorphism and coding). Implementation and testing of the methodology with difficult graphs are discussed.

THEORY AND BACKGROUND

Graphs are mathematical abstractions

A mathematical graph is any set of nodes with a set of edges which specify which nodes are adjacent or connected.\(^4\) The simplest type of graph is the
undirected (also called linear) graph (Figure 24). Here, the edges do not indicate direction. The nodes and edges can be labeled such as to indicate atomic number, bond order or any arbitrary information. Molecular structure is the most obvious chemical example of an undirected graph.\(^5\) Another type of graph is the directed graph (not considered here, but extensively used in Chapter 4) where the edges can indicate direction, which implies an ordering of nodes. Again, the nodes and edges can be labeled with arbitrary information, even another graph, such as in the case of reaction pathways. It is important to note that the diversity and number of labeled graphs is gigantic.\(^6\) However, as we shall see, it is still an open question as to what is the identity of the graph, that is, what exactly makes any particular graph different from any other.\(^7\)

The word "graph" actually originated from the graphical depiction of molecules by chemists over 100 years ago.\(^8\) Graph theory arose from Cayley's enumeration of organic compounds. Graphs and chemistry are intimately related,\(^9\) because a graph is a mathematical abstraction that is intuitively obvious to the chemist.\(^10\)\(^-\)\(^12\) They are useful because much of chemistry deals with discrete objects (e. g., atoms) and their relation between each other (e. g., bonds, distance, etc.). The fields of graph theory and chemistry nurtured each other as they grew.

For a computer to process, store or manipulate graphs, we must figure out a way to represent graphs numerically. One method that can represent graphs using little space is an **adjacency list**,\(^13\) which is a representation of the graph wherein each node is associated with a list of its adjacent (connected) nodes. Another representation is the **adjacency matrix**,\(^14\) which is generally easier to manipulate due to its predefined size. Spialter\(^15\) was among the first to recognize that the unambiguous representation of molecular structure with a labeled adjacency matrix
Figure 24. Examples of different types of graphs. a.) An unlabeled and undirected graph. The filled circles are the nodes, and the lines connecting them are edges. b.) A labeled and undirected graph. In this case, "Cleveland" is a label for a node and "2800" is a label for an undirected edge. c.) A labeled and directed graph. $n\text{CO} + (2n+1)\text{H}_2 \rightarrow \text{C}_n \text{H}_{2n+2} + n\text{H}_2\text{O}$.

and its use for computer manipulation, which he termed the atom connectivity matrix (ACM). He demonstrated the use of ACM's with delocalized benzene, localized free radicals, linked rings, and transition states.

Our study used a very similar numerical representation of graphs that blurs the distinction between chemistry and traditional graph theory. The adjacency matrix representation $A$ of a labeled undirected graph $G$ with $n$ vertices is a square symmetric $n\times n$ matrix, where off-diagonal element $A_{ij}$ is non-zero only if the nodes are connected by a labeled edge (a real number in our case). The node labels are stored on the diagonal at $A_{i,i}$. Figure 26 shows an example of $A$ for aniline. Three
Figure 25. The spatial pharmacophore of the GABA(B) agonist receptor\textsuperscript{16} is an example of a TCE. The dotted lines indicate allowable ranges of distances in Angstroms between a pair of atoms. Note the dummy atom at "X."

Auxiliary arrays of length $n$ are used to hold the atomic weight, charge, and stereochemistry node labels. We call $A$ the extended adjacency matrix.

**Molecular structure is only one example of a TCE.**

Many examples of TCE’s could be listed. Graphical representation of stable molecular structures is perhaps the most obvious and has been thoroughly explored. Unstable intermediates, fragments, biophores, and reaction centers have been explored to a lesser extent. Conformations of molecules and their fragments (spatial pharmacophores) can also be represented\textsuperscript{17} with graphs (Figure 25).

Why attempt to name all graphs and not just the graphs that represent TCE’s? Graphs were chosen as the input domain to the Graphsort program for two reasons. First, graphs abstract away all but topological information. Many chemical applications are dependent on connectivity,\textsuperscript{18} and graphs supply an excellent basis for expressing TCE’s.\textsuperscript{9} The process of representing the TCE as a graph is automatic and simple. Second, all graphs are not TCE’s but all TCE’s are graphs, so the issue of whether a TCE is "legal" is avoided. Once representation by graphs is admitted, the pre-definition of "TCE" can be avoided if graphs in general are used. It is fair to
say that the number and diversity of mathematical graphs far exceed those of TCE graphs. Hence, if a naming algorithm is general for mathematical graphs, then it will necessarily operate for all TCE’s. Furthermore, the algorithm can fall short of complete generality for mathematical graphs and still be very useful. To my knowledge, this is a novel approach. While there have been many previous studies that start with one limited type of TCE (e.g. molecular structure for storage purposes), the size of the input domain was expanded so that it would automatically include all TCE’s.

**Automated unique naming equals computer recognition**

Why should we bother naming graphs? While the graphical depiction of a graph is comprehensible to humans, it cannot be perceived by computers. The unique naming of an object is a form of recognition, which might be considered a fundamental element of intelligence. Clearly, having the pre-ability to name graphs of any type would free the computational chemist to concentrate on the higher levels of chemical software design. The availability of a recognition facility would allow the designer to devote more time to the application, instead of "reinventing the wheel," which seems to happen often. Therefore, the main goal was to develop a sturdy foundation upon which other software could rely, thus encouraging innovative new chemical software.

A good example of such chemical software is expert systems. Expert systems manipulate various forms of TCE's and must deal with the recognition problem in one form or another. Expert systems that execute a massive search strategy require fast and efficient recognition because the power of the expert system is proportional to the number of possibilities it can test per unit time.
Moreover, it is impossible to anticipate which TCE's will be encountered, so the recognition facility must be robust.

The need for a recognition algorithm was inspired by the CASE expert system, which uses a form of recognition through matching. The training database is fragmented into CASE syntax fragments. The CASE syntax fragments are uniquely named according to a set of rules. The resulting fragment must be "normalized" so that there is a 1-1 correspondence between the actual fragment and its CASE syntax, which effectively provides a name. Likewise, the CASE syntax fragments from the test set molecules are derived. Given that they are also uniquely named, a direct string comparison can determine whether or not the fragments are the same. CASE canonicalization is robust and efficient because it limits its "basis set" of fragments to ones of well-defined topology. Since the fragments are a small subset of all possible graphs, the canonicalization rules are simple and complete. This approximation has been tremendously successful — CASE has explained diverse bioactivities. However, the potential of the CASE approach can be applied to new types of TCE's and Graphsort could provide recognition support.

The metabolism program META (Chapter 1) presents a more formidable challenge because its "basis set" is complete molecular structure. META needs to recognize chemical structure because biotransformation pathways often converge; several transforms may hit and form the same product. Thus META needs to recognize the duplicate products and merge them (Figure 20). Furthermore, Graphsort could be used to expand the CASE syntax fragments used in the transforms in the dictionary.

Expert systems are but one possibility; we cannot, nor should, define all chemical applications. The following are a few possibilities: database storage and
cataloguing, generation and manipulation of structures that are 2D, 3D, partial or complete. Moreover, possible non-chemical uses can also be envisioned. Whatever its final use, clearly the utility of the algorithm depends primarily on two properties. The first is robustness, which is the ability of the algorithm to successfully process a large diversity of graphs. The second is computational efficiency, which is the ability to function relatively quickly. Robustness and efficiency are inversely related.

**Nomenclature conventions**

What is the best naming style for graphs? The usual method of naming chemicals (or any finite set of discrete objects for that matter) is to collect all the

![Chemical structures]

**Figure 26.** Example of coding aniline by three different GRINS line codes (there are many more). Hydrogen-suppressed aniline is numbered such as to indicate the route chosen. The corresponding GRINS code is shown below each structure with its labeled adjacency matrix A that is generated by program "GRINS." Row and column numberings are included for convenience. Note that each A is a different block of numbers. Since computers operate with numbers, a computer would perceive these adjacency matrices as being different when they really represent the same molecular structure graph, unless specifically handled by special software. (As Shakespeare says, a rose is still a rose by any other name...) Graphsort selects one particular A as the canonical representative, and accordingly, a representative GRINS code as the official (canonical) name of aniline.
objects into a set, construct another set of names and preassign them a bijective mapping. The trivial naming of chemicals (benzene, phenol, etc.) is an example of this approach. Such an attempt is practically impossible for the infinitely large set of all graphs. This is also the case for ever-expanding list of compounds in which millions are known and new ones are constantly being synthesized. Moreover, the list of all possible compounds has yet to be enumerated.

Alternatively, IUPAC assigns unique names to stable chemical structures by employing a closed set of rules based on group functionality. This works well because its input domain is restricted to stable compounds only. Maintaining such a closed set of rules would be very difficult for graphs. IUPAC rules generate names easily pronounceable by humans, but it could be argued that this is not the case for some structure names (e.g. tricyclo[3.3.1.1\[3,7\]-decane]!)

If the aesthetics of the name can be ignored, then the adjacency matrix might be used as a name, but for most purposes, it is too bulky and unintuitive, especially for low-valency TCE's found in chemistry. A popular alternative method of naming molecular structure is to use line codes. Line codes are linear strings of characters constructed from chemical structures by a small set of rules.

There are many chemical line code methods in the literature. Most require manual arbitrary selection of a "route" through the molecular structure which results in a sequence of the nodes in the graph. The process of coding begins with selecting a starting node in the molecule and following the adjacencies until no more nodes can be visited. Many routes could be followed.

Route-oriented line codes have the main advantage that they are efficient and easy to use. Line codes provide a natural language for the chemist. Over the years, more powerful computers have reduced the need of brevity of the line code,
permitting less cryptic syntax. Most importantly, line codes are invertible back and forth to structure rather intuitively. Line codes represent chemical graphs more concisely than a sparse matrix does, making them very efficient for storage purposes.

Line codes can be applied to graphs also. However, a line code that can describe labeled undirected graphs has not been developed. Accordingly, the SMILES language\textsuperscript{37} was extended to generate line codes representing graphs. The SMILES language possesses many desirable properties according to Read's criterion\textsuperscript{32} and enjoys widespread acceptance by the chemical community. Thus, my extension of the SMILES language, termed "GRaph INput Syntax" (GRINS) enables line coding of graphs. Essentially, the line code syntax rules for GRINS are similar to SMILES syntax, but GRINS ignores all chemical concepts. First, aromaticity is not specifically coded as it is in SMILES. This enables the user to completely control the node order in $A$. Separate algorithm modules can be developed that modify the bonding based on chemical principles, for aromaticity, tautomerism, partial bond character, etc., by reprocessing the $A$ using standalone chemical software.\textsuperscript{38} Ignoring aromaticity allows GRINS to not require an organic subset; but it can be included if desired. This saves many left and right brackets in the code, which reduces line code length, and clarifies the line code. Second, valency violation is ignored, which is required for fragment and spatial pharmacophore description. Third, normal bond symbols (\textendash,\textendash,\#) are used, but provisions for edge labels that are real numbers (between angle brackets perhaps) can be defined in the future. Fourth, user-defined dummy atoms can be included.
There is one disadvantage to line codes when applied to graphs. The practicality of GRINS is reduced for large highly connected graphs. The resulting code may become longer than A.

The order of nodes in A and the line code are deterministically linked by encoding and decoding algorithms. For example, the order of nodes in A generated by the program GRINS will change if and only if the line code changes. Similarly, the program GROANS, which inverts A back to line code, is also deterministically linked. Given a deterministic algorithm for converting code back and forth to A, the order of the nodes in A is directly determined by the line code. A is sorted to a standard order and converted back to a GRINS line code, which serves as the official name of the TCE. The new line code is algorithmically determined from the sorting of A, effectively selecting the official route. Thus we need to discuss sorting of A to a standard order, which is very difficult, as we shall see.

**Canonical ordering causes the uniqueness**

Nagle was among the first to recognize that the order of nodes in A is the root of the problem to recognizing graphs.\(^{39}\) Given \(n\) nodes in a labeled graph, a set of up to \(n!\) possible adjacency matrices with distinct node sequences exist which represent the same graph. A particular adjacency matrix implies that a distinct sequence of nodes was arbitrarily chosen, that is, a route was selected. The ordering is equivalent to labeling the nodes.\(^{40}\) The selection of the representative ordering is referred to as canonical ordering or labeling of the nodes in the graph.\(^{41}\) The goal of Graphsort is to convert the matrix to a standard canonical order, i.e., sort nodes in the graph.
In the quest to canonically order the nodes, topological symmetry must be taken into account, and topological symmetry is best explained by using elementary group theory.\textsuperscript{42-43} A can be reordered by permuting the order of nodes, which can be expressed mathematically as $A' = gA$, where $g$ is a similarity transformation\textsuperscript{44} that represents the particular permutation. If $A'$ and $A$ are different,\textsuperscript{*} then $g$ is called an \textit{isomorphism} because $A'$ and $A$ represent the same graph. Two graphs which are line coded differently are \textit{isomorphic} if their matrices can be related by any $g$. If $A'$ and $A$ are identical, then $g$ is called an \textit{automorphism}. An automorphism in a graph is defined as a $g$ that preserves adjacency.\textsuperscript{45}

Take the graph in Figure 27, for example. If permutation P1 is applied, edge adjacency is not maintained (e.g., 1 and 3 are not connected anymore), and so it is not an automorphism. However, if permutation P2 is applied, all adjacencies are preserved and so it is an automorphism. Writing out the adjacency matrix will convince you that this is the case. The collection of all automorphisms of the graph forms its automorphism group.\textsuperscript{40} The idea of graph automorphisms is identical to that of symmetry operations in crystallography;\textsuperscript{3} all rules of group theory hold.

\* Two matrices are different if at least one of their elements are not equal.
Automorphisms permute constitutionally equivalent nodes that have the same identity.

Right now, the only way to prove that two graphs $A$ and $A'$ are isomorphic is to directly reorder the matrices. If and only if the results are identical adjacency matrices, then the graphs that they represent are isomorphic.\(^{46}\) (See Figure 26.)

Given a graph, the nodes can be classified into automorphism partitions (also called orbits) which can be uninodeal or multinodal. The automorphism partition consists of cells satisfying that two nodes are in the same cell if and only if there exists an automorphism mapping the two nodes. A good example of the difference is 2,2-dimethylpropane. The hydrogens on each methyl are in a separate cell, but all the hydrogens in the molecule are in one automorphism partition. Again, writing out the adjacency matrices makes this immediately obvious.

Detection of the automorphism partition is imperative because the nodes in each automorphism partition are topologically symmetric. (The hydrogens in ethane are topologically symmetric.) Detecting the automorphism partition is necessary to determine the scalar magnetic equivalencies of a molecule, which would be required in software that automatically assigns the peaks of nuclear magnetic resonance (NMR) spectra to specific nuclei.\(^{42,47}\) The coding of graphs is closely related to symmetry detection also. During the canonicalization of $A$, it would be wrong to expect that all the nodes in the automorphism partition can be ordered arbitrarily and $A'$ would equal $A$. This would be true only for nodes in the same cell. However, any single node in the automorphism partition can be selected arbitrarily, but this act breaks the symmetry.\(^*\) The selection effectively creates a new graph with new

\(^*\) Similar to coloring vertices; for example, see Liu, X.; Klein, D. J. J. Comput. Chem. 1991, 21, 1243-1251.
automorphism partitions, which must be reevaluated because the relationship of the
nodes in the graph is perturbed. To know when to break the topological symmetry
a canonicalizing algorithm must know that it exists. This is why canonicalization
requires detection of the automorphism partitions.

At present, the only certain way to detect the automorphism partition is by
finding a permutation g that makes A'g=A. An obvious approach to the problem is to
apply all g systematically and see if g is an automorphism. This "brute force"
solution of the problem would be impractical because there are too many
possibilities to try. Such algorithms quickly become useless as the problem size
grows. To explore all permutations of n nodes requires n! tests, a number that
grows very fast as n grows.\textsuperscript{48} Even for the size of ordinary molecular graphs of
twenty or more atoms, this is known to be not computationally intractible, taking
hours of computer time for only one graph!

Alternatively, Nagle\textsuperscript{39} suggested a more direct way of canonically ordering
star-type graphs as a large binary integer by writing one row after another. Each
permutation of the unlabeled adjacency matrix would give a potentially new integer.
The integer with the smallest value is defined as the representative ordering.
Randic\textsuperscript{49,50} explored this method further, developing rules for permutation and
minimization of the integer in what is in effect a combinatorial optimization process
(Chapter 3). MacKay\textsuperscript{51} found that local minima exist and gave a graph on
seventeen nodes that was a counterexample to the Randic conjecture, thus
invalidating the generality of the algorithm. Randic modified his conjecture to
cyclically interchange three rows instead of two, but this is not sufficient either.\textsuperscript{52}

Basically, all permutations of A must be tried to ascertain a global minimum
has been reached, which is what is meant by the term "brute force." These types of
Figure 28. An example of the graph isomorphism problem. Are these two graphs the same? One way to answer this question is to name each graph uniquely and then compare the names.

Algorithms can be made "smarter" with heuristics, which help direct the way through the maze of permutations. Several heuristic back-tracking algorithms have been invented, but the use of heuristic cannot guarantee a polynomial solution.

**Complexity theory and analysis of algorithms**

The unique naming problem can be stated as this: given a graph, derive its name. The graph isomorphism problem can be stated as this: given two graphs, are they the same (Figure 28)? Clearly, a solution to the unique naming problem
implies a solution to the graph isomorphism problem; the names could simply be compared to determine isomorphism. The automorphism partitioning problem can be stated as this: given a graph, what nodes are topologically symmetric? Karp showed that graph isomorphism is algorithmically equivalent to the automorphism partitioning problem. Thus, an algorithmic solution to any one of these problems will help solve the other two.

The existence of an algorithmic solution to a problem is not the whole story. Complexity theory studies the inherent computational difficulty of problems. Some problems are known to be solvable only with a non-polynomial (NP or exponential) relationship to the size of the problem while others are known to be solvable with a polynomial (P) relationship to the size of the problem. The time dependence is intimately related to the loop structure of the algorithm. Additionally, some problems are NP-complete, which are known to have non-polynomial solutions, but are not known whether or not they have polynomial solutions. Furthermore, all NP-complete problems have the property of being transformable or reducible to each other, so if a polynomial solution is found for one NP-complete problem, then all NP-complete problems are polynomial.

For example, subgraph isomorphism, used for chemical purposes by Ray, then Sussenguth, requires a brute force matching of the nodes of the graphs that are being compared. The subgraph isomorphism problem is known to be NP, not NP-complete, and not in P. In contrast, the graph isomorphism problem is known to be in NP, but it is not known whether or not it is NP-complete and it is not known whether or not it is P. (One way of viewing this is: we do not know whether graphs have enough information contained in their structure to limit the possibilities for the combinatorial search.) Thus, the unique naming problem appears to be
non-polynomial, but a polynomial algorithm could exist. A polynomial solution to the coding problem would imply a polynomial solution to the graph isomorphism and automorphism partitioning problems.

So what motivates one to work on such an exceedingly difficult problem? First, this challenging problem is academically appealing to some. Finding the unique naming problem to be NP-complete and polynomial would have far-reaching implications for computer science. Polynomial solutions would then be known to exist for all NP-complete problems, many of which have direct chemical applications. Such naive enthusiasm must be tempered by the fact that discovery of an efficient algorithm for graph isomorphism has eluded the finest computer scientists and graph-theorists. However, it is entirely possible that due to its combinatorial nature the graph isomorphism problem requires computer-assistance to "chip away" at the problem until a solution is reached. Such was the case for proving the famous four-color conjecture, which was not known to be true or false until computers assisted in its ultimate proof. At the very least, answering this open question conclusively would put this nagging problem to rest.

Unique naming is a very important real world problem, which provides an additional motivation to basic academic curiosity. What is more important, just the act of attempting this difficult problem has a very practical result. The by-product of the academic pursuit is very useful because the algorithm is honed and tested with challenging examples and counterexamples. Due to the real possibility that no polynomial algorithm for graph isomorphism exists, the balance between efficiency and robustness were always kept in mind, especially from a chemical viewpoint. In this mode, we might think of Graphsort as a type of parameterization in that the data points are graphs and the algorithm, an abstract parameter. At the very least, we
will know which graphs are "legal." Accordingly, we were left with an extremely usable algorithm when practical concerns halted the pursuit of this academic goal.

**Graph Invariants**

It might be hoped that the necessity for canonically ordering $A$ might be bypassed if naming is based on *graph invariants* (also called graph or topological indices), which are any properties that are independent of the labeling. There are many examples of graph invariants: number of nodes, number of edges, diameter of the graph, etc. The literature offers many chemical graph indices that are specialized graph invariants that are mainly used to map chemical structures to a space to predict activity.\textsuperscript{61-65} The META program (Chapter 3) uses graph indices for quantification of steric properties of polycyclic aromatic hydrocarbons. However, graph indices do not offer much promise for providing a unique name due to their low discrimination ability. Many other graph invariants have been tried, but each was found to be insufficiently discriminating for our naming purposes. It was often hoped that combinations of graph invariants were often hoped to uniquely name the graph, but this has not been very successful.

The attempt to find a complete set of graph invariants is alluring and difficult, and might be labeled as a "disease," as Harary tagged the fervor surrounding the 4-color conjecture.\textsuperscript{66} The characteristic polynomial of the unlabeled adjacency matrix was initially presumed to provide a complete graph invariant, but it has since been shown, through ample counterexamples, that this is not the case for the large set of isospectral graphs. Balaban and Harary provided several simple counterexamples.\textsuperscript{67} The eventual direction was to introduce more graph invariants to account for the counterexamples, but then new counterexamples
were found. For instance, Randić suggested eigenvalues and eigenvectors, but Swank showed almost all graphs eventually become *cospectral*, which means that they have the same eigenvalues and eigenvectors, with another graph as \( n \) increases. He observed experimentally that as the number of nodes increases, the amount of uniqueness decreases, that is, more graphs share the same characteristic polynomial.69

This approach of constructing graph names from several graph invariants may be sufficient if the input domain is restricted to chemical graphs, but a complete set of invariants is not presently known that specifies up to isomorphism for graphs in general. Other difficulties with graph invariants acting as names exist. First, the computation of the graph invariants must be efficient, which may not be the case if the indices are not independent and have overlapping discriminating behavior. Second, the graph invariant may not have an inverse, i.e., given the graph, it may be easy to compute the graph invariant, but the opposite may not be true. Turner showed that the outlook was very pessimistic for any type of generalized matrix function to provide a complete graph invariant.70

*Node Invariants*

The cause of the difficulty of graph naming with graph invariants may be that they over-emphasize the compression of information into a single numerical entity.71 *Node invariants* (also called vertex indices) are properties of the node that are independent of how the graph is labeled, and address this concern. Node invariants are to node names what graph invariants are to graph names. This analogy is perhaps equivalent to that of the graph isomorphism problem being algorithmically equivalent to automorphism partition detection.
Since the attempt to construct unique graph names with graph invariants was found to be difficult, the canonicalization problem might be better approached if it is reduced to that of generating unique node names. According to the analogy of node invariants with graph invariants, unique names might be constructed for each node using node invariants. The prospects for success seem greater — more information is available to us because node invariants do not compress structural information to the degree that graph invariants do, especially if allowed to remain as sequences of numbers called distributions. As for graph invariants, node invariants can partially describe the identity of the node. Nodes in the automorphism partition that are constitutionally symmetric necessarily have the same identity. Thus, a complete set of node invariants would detect the automorphism partitioning. As might be expected from analogy with graph invariants, no complete set of node invariants is yet known that entirely represents all possible structural environments of a node.

This unique node naming approach has long been recognized by others such as Schmidt and Druffel and Hopcroft and Wong who used this technique to solve the graph isomorphism problem for planar graphs. Graphsort also employs the design of unique node names. Node invariants are used to construct node names. If a multinode automorphism partition is found, the symmetry is broken, and the process repeated until no further multinode automorphism partitions remain. Once the names of the nodes have been uniquely defined, it is a simple task to sort them according to their identities and achieve canonicalization.

Therefore, we have reduced the problem of identifying the graph to one of fully describing the identity of the nodes. This naturally begs the question: what exactly is the identity of a node? A part of the identity of a node is in terms of local attributes such as node labels (e.g. atomic number, weight, degree sequence, edge
labels, etc.) The remaining part of the identity of nodes in a graph depends on its global relationship to all the other nodes and their identities. This recursive characteristic causes quantifying the global "environment" of the node difficult. If two nodes are in a different environment in the graph, then they must be named differently, and if they are in the same environment in a graph, they must be named the same. The lack of complete representation of global node identity is the sole reason that Graphsort will fail.

To prove topological symmetry, a lack of asymmetry must be demonstrated, which is proved only by showing that a complete set of node invariants does not discriminate. As previously mentioned, a complete set is not yet known, and much of the efforts of this study have been toward designing a set of node invariants that is as complete as possible with emphasis on computational efficiency. Hence, the nature of the node name is much less important than whether they discriminate fully. In other words, we don't care about the details of the node ↔ name mapping, only that it is bijective, which allows a great deal of freedom in the design of node descriptors. The information to construct the node invariants must obviously originate from an arbitrarily ordered graph. The node invariants should be as independent from each other as possible so that discrimination power (and thus CPU cycles) will not be wasted. Independence is difficult to quantify, but basically, nodes that are discriminated by one node invariant should not be discriminated by the other node invariants used in constructing the node name. Naturally, we want to use the invariants in the order that will most likely differentiate for the lowest computational cost of the "average" TCE, if such a thing exists.
Differentiation of nodes

Morgan\textsuperscript{74} proposed an algorithm to be used in the coding of chemical structures (not graphs), which found widespread use in chemical documentation. His algorithm calculates extended connectivity (EC) values by first assigning an initial integer value to each node, which is equal to the number of connected nodes. An iterative process then calculates a new EC value for each node based on the sum-of-squares of the surrounding EC values. The process is continued until no change in the EC values occurs. Notice that the extended connectivity acts as a node invariant that measures how centrally located an atom is in the structure. Wipke and Dyott subsequently improved the Morgan algorithm by including stereochemistry.\textsuperscript{75} The poor discrimination ability of the Morgan algorithm was subsequently improved by Shelly and Munk,\textsuperscript{76} but Carhart still found it to fail. Several others have found that the Morgan algorithm duplicates the existing graph-theoretic idea of walks.\textsuperscript{77-79} Although little attention was paid to the following fact, the Morgan algorithm was one of the first to employ the identities of other nodes in calculating node invariants, as well as progressively differentiate the nodes from coarser to finer intermediate partitions in the hopes of convergence to the automorphism partition.

Whereas the Morgan algorithm is computationally inexpensive, the sum-of-squares function that it employs is ambiguous (e.g. \(1^2 + 4^2 + 4^2 = 2^2 + 2^2 + 5^2\)). Accordingly, Weininger\textsuperscript{80} used the method of corresponding primes (e.g. \(1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 5, 4 \rightarrow 6, \ldots\)) to create an unambiguous function. The Fundamental Theorem of Arithmetic\textsuperscript{45} states that the product of the primes is unique. I tried to apply this method to graphs in general, but, my experience was that, while the function was unambiguous, the resulting prime number quickly becomes unwieldy.
for large graphs, especially highly interconnected ones. The next logical step was to use lists and distributions of the prime numbers, so that no function would be required at all and compression of structural features would be avoided. The prime number distribution for each node was then calculated by recursion out in "shells" (i.e. nodes that are equidistant from the origin node). A difference in nodes was detected by sorting the nodes using their distributions of prime numbers as sort keys. As soon as a difference was found, calculations of further shells was unnecessary. Unfortunately, calculation and maintenance of all these numbers for large symmetric graphs (e.g. buckminsterfullerene) proved to be absolutely intractable.

However, this work was not for naught. The sorting of distributions was the beginning of a novel algorithmic architecture that functioned to differentiate nodes. Sophisticated distance distributions were soon found to be much more computationally efficient at describing the node's identity than the recursion techniques described above. Once differentiated, the identity of the nodes in the intermediate partitions were saved with the use of ranks. Nodes of common rank form intermediate partitions that may or may not contain more than one automorphism partition. (Conversely, automorphism partitions are always contained within one intermediate partition.) Each intermediate partition is sorted and if the nodes in it have different sort keys, then new intermediate partitions are formed. If the nodes pass the complete set of node invariants (i.e. they have the same name), then the nodes are assumed to be topologically symmetric. Once broken, the symmetry can be broken, as previously discussed.

This type of algorithmic architecture offers the following advantages of efficiency. First, the "divide and conquer" architecture of sorting\textsuperscript{13} allows
computational resources to be focused only where appropriate. Sorting need not occur outside the active partition because the relative order of the intermediate partitions is maintained throughout the sort. Ranks allow the identity to be computed only once. Second, the architecture allows a hierarchy of node invariants. Asymmetry may be obvious, in that it is due to attributes more or less local to the node, and it can be easily detected with minimal computation. Or asymmetry may be very subtle, in that it is distributed globally, and much computation must be spent to detect it. Or asymmetry may not be present at all. The hierarchy for detection of topological symmetry starts with less expensive "methods". Third, either differentiation or breaking of symmetry must occur once the nodes in a partition are named. Consequently, the differentiation process never backtracks and computer time is not wasted. Only one node naming is required. Fourth, the final ranking of the nodes can be directly used to canonically order the nodes in $A$.

The ranks progress from state to state as the nodes. The ranks change state whenever nodes are differentiated or topological symmetry is broken. The nature of the node invariants greatly affects the rank progression. The actual progression of rank states are unimportant, but the number of rank states need to be minimized for the sake of efficiency. This gives considerable freedom in the design of node invariants.

**Distance distributions**

The *distance matrix* $D$ of a graph contains off-diagonal elements that represent the minimum number of edges between the nodes $i$ and $j$. The distance is a graph-theoretic distance, not geometric distance. $D$ is a unique representation of the graph and is in some sense a "richer" representation than the adjacency
matrix. Of the several algorithms to generate the distance matrix from the adjacency matrix, the method of Muller was chosen due to its efficiency.

Distance in graphs has only begun to be closely scrutinized by mathematicians and chemists. Results from several graph isomorphism algorithms that employ the distance matrix suggest that the distance-based approach is very promising for our naming purposes. However, the input domain for these algorithms is unlabeled graphs only, which eliminates these algorithms for being used to represent TCE's. Randic conjectures that the set of all paths in a graph probably specify up to isomorphism. Unfortunately, the number of paths grows exponentially with the number of nodes in the graph, which dashes any hopes
for efficiency. Now, the distance matrix selects only the shortest paths, which might contain sufficient information to efficiently specify isomorphism. Capobianco believes that certain distance distributions have a good chance to completely characterize a graph.\textsuperscript{7}

Several topological indices have been devised that utilize the distance matrix. Bonchev used information theory to prove that the distance matrix has superior discriminating ability.\textsuperscript{8} Balaban's \( J \) index, which is based on the distance matrix, is a highly discriminating topological index.\textsuperscript{61,89} Klopman et. al. have met with success in the distance-based node invariants.\textsuperscript{90,91} Klopman et. al. utilize a distance-based node invariant that originates from information theory, which was used to study carcinogenicity of polycyclic aromatic hydrocarbons. Klopman and Henderson\textsuperscript{71} devised a set of distance-based node invariants called \textit{structural environment vectors} that describe the global identity of the nodes. This idea provided a good starting point for experimentation with novel distance-based node invariants.

\textbf{METHOD}

\textit{Development cycle of the algorithm}

Once the basic program architecture was coded, a "workbench" was available to design and test the discriminating ability of various node invariants. The software provides an experimental "laboratory" to test graph theory conjectures, which is like conventional chemistry laboratories providing methodology to test hypotheses. An example-counterexample cycle, not dissimilar to experimental scientific method, was used to develop the algorithm. The cycle started at a point where no counterexamples were known to our conjectures embodied by the
working algorithm. Focus was then turned to hypothesis-disproving by literature search\textsuperscript{92} and/or design of counterexamples. Armed with intimate knowledge of the state of the algorithm, graphs were constructed that were suspected to challenge the current conjectures. These graphs often revealed unanticipated entities and features, which would be extrapolated to the worst-imaginable cases which had to be handled by the next version of algorithm. All counterexamples from previous versions were retained and used in testing each new version of the algorithm. An abbreviated list of the resulting counterexamples are shown in Table 17. These are much more challenging than most of the graphs in the chemical literature and would be highly useful to anyone designing their own algorithm.

It is important to emphasize the human element (Dr. Mario Dimayuga and myself) in the development cycle. Great effort and creative insight were required to form a conjecture that was true for all example graphs simultaneously, and to synthesize new counterexamples. The efficiency and robustness of Graphsort originate mainly from the example graphs, or "torture" graphs as Knuth\textsuperscript{93} might call them. The torture list always contained many chemical structures to maintain efficiency, but also many difficult mathematical graphs to maintain robustness.
Since describing the intermediate versions and the blind alleys we entered is not appropriate, only the final algorithm is described.

**Testing graph theoretic conjectures experimentally**

A graph was determined to be either an example or a counterexample by directly testing Graphsort as schematized in Figure 29. The input GRINS string was converted to an extended adjacency matrix $A$, and the order of nodes was randomly shuffled by program MIXCNM and subsequently Graphsorted. The output $A$ was converted back to a GRINS line code and compared to the output of the previous cycle. If the line codes did not exactly match, then Graphsort failed and a counterexample was found. The matrix was then reshuffled and processed for hundreds of iterations (as much as time permitted), automatically terminating at any time that Graphsort failed.

Because this does not systematically test all permutations, non-failure of this procedure only suggests, albeit very strongly, that the conjectures in Graphsort are valid. However, the testing procedure seemed to have no difficulty finding permutations that failed. Failures in Graphsort are solely caused by non-discriminating node invariants, resulting in a final partition that contains more than one automorphism partition. The probability of swapping the order of two nodes in these two different automorphism partitions for any one of the many iterations in the testing procedure is extremely high, especially if the sizes of the automorphism partitions are not very small compared to the size of the graph.
Figure 31. Node differentiation employed by the Graphsort algorithm. Propagation is in the shaded regions of the figure. "EOL" stands for end-of-list.

The GraphSort algorithm

The complete method of generating a unique GRINS line code is schematized in Figure 30. The user (or calling procedure of an independent program) encodes the graph with GRINS line code by following an arbitrary route. The line code is converted to A which is Graphsorted to canonical order and
converted back to GRINS line code that serves as its unique name. Borrowing from
an object-oriented programming metaphor, this methodology allows the graph to tell
us its name, regardless of the input GRINS line code. Here, the algorithm is
described in detail by first discussing an overview of the differentiation process,
followed by the distance distributions used by the differentiation process.

An input GRINS string is translated to its extended adjacency matrix $A$ by
the program "GRINS." Graphsort reads in $A$, and calculates the distance matrix.\textsuperscript{84}
The diameter of the graph is calculated for purposes of efficiency. The length of all
distance distributions used in Graphsort has the diameter as an upper bound, which
is used to increase efficiency by limiting loop lengths and reducing dynamic memory
allocation requirements.

Graphsort then differentiates the atoms, which is flowcharted in Figure 31. At
the beginning of Graphsort, all nodes are undifferentiated and possess the
common rank of 1. At the end of Graphsort, the nodes must be assigned distinct
sequential ranks from 1 to $n$, which will be used in route tracing and final reordering
of $A$.

The initial stage of differentiation of the nodes begins with the construction
of node names that depend only upon the local neighborhood of the node. Each
node was named by a combination of its label, the labels of the node adjacent to it,
and the labels of the edges connecting them. The following node invariants are
assembled in the following order:
1.) Label of the node (atomic number)
2.) Number of non-terminal (non-hydrogen) nodes connected
3.) Number of terminal (hydrogen) nodes connected
4.) Sorted sequence* of edge label types (bond order)
5.) Secondary node label (stereochemistry)
6.) Tertiary node label (atomic charge)
7.) Quaternary node label (atomic weight)

For example, the oxygen atom in water would have the local node name (8,0,2,1,1,0,0,0,16).

The constructed node names are placed as keys on the sort table that consists of a matrix of \( n \) by \( m \) (where \( n \) is the number of atoms and \( m \) is the length required to accommodate the number of discriminating node invariants). All nodes are selected as the active partition and flagged as the source partition (explained below). The nodes in the active partition (the whole graph at this point in execution) is sorted according to their keys. After sorting, the active partition is reranked by scanning the keys in the sorted table starting at the top. The rank is incremented whenever a difference is found in the keys, thus creating a new partition. Sorting within one partition is computationally more efficient than sorting the entire graph each time.

The differentiation of the active partition (recorded by the new ranking) then is propagated (see below), which may differentiate nodes in and/or below the active partition. Graphsort goes on to activate the partitions individually from lowest rank to highest, never retreating. The result of each activation either differentiates the nodes in the partition, thereby creating finer partitions, or assumes that the partition

*This might be changed to a distribution to be consistent with the style of the other node invariants used in the algorithm.
is automorphic. Each activation proceeds as follows: The first multinode partition is selected as the active partition. If the degree of the nodes in this partition is less than two, then the partition is assumed to be automorphic. In this case, the nodes are completely artificially differentiated, in a process termed dispersing. If the nullity (cyclomatic number) of the entire graph is less than two, and the cycle distributions (see below) are equivalent, the active partition is assumed to be automorphic. In this case, the constitutional symmetry is broken by artificially differentiating one node from the others, in a process called repelling, in which one node is arbitrarily selected and the distance from this node to the others in the active partition is used as the sort key. If the cycle distribution is not equivalent for the nodes, then the partition is differentiated using the cycle distribution as the sort key. In all cases, the active partition is differentiated. The newly created partitions are flagged as source and after being propagated, the next step of differentiation begins by activating the first multinode partition.

Propagation via distance distributions

As previously described in the theory section, differentiation of any intermediate partition perturbs the rank environment of the nodes and is a source of asymmetry that may differentiate other sink degenerate sets. If a sink partition is differentiated, it in turn produces source partitions which may differentiate other sink partitions, and so on, in a process termed propagation. Eventually, propagation ceases because every source partition contains no further information to differentiate any sink partitions.
Figure 32. Example of calculation of the two types of distance distributions used in propagation. The numbers on the edges are labels that indicate three different types (e.g. single, double, and triple bonds). Assume for the sake of illustration that the filled nodes are the source partition and nodes A and B form the sink partition. Nodes A and B cannot be discriminated by their node distribution but can be discriminated by their edge distributions. The edge distributions are calculated by their geodesics. In this example there is only one geodesic for each pair of nodes because this graph is a tree.

Figure 32 illustrates the calculation of the two types of distance distributions used in propagation of differentiation. Information about differentiation is transmitted through distance distributions of both nodes and edges for each node in the sink partition. The distance distribution of nodes (node distribution for short) counts the number of nodes in the source partitions that are $d$-away (for each $d$, up to the diameter) from each node in the sink partition. In contrast to node distance distributions, there are as many distance distributions of edge labels (or edge distributions for short) as there are distinct edge labels types. Each edge distribution counts the edge labels of one type that occur along the geodesic (see below) from the source set to the sink set.

Now that both types of distance distributions have been discussed, propagation can be discussed. Propagation is flowcharted in the shaded box in Figure 31.

1.) After an active partition has been differentiated to finer partitions, these new partitions are flagged as the source partitions, indicating that they may have information to differentiate other partitions.
2.) A partition is selected from the flagged partitions and immediately unflagged. This partition may or may not be automorphic and may be either multinodal or uninodal. The method of selection of the source may be arbitrary, but it must be independent of input node order. We used a simple stack (with respect to rank) approach.

3.) A multinode partition is selected from all partitions as a sink partition, which is a candidate for differentiation by the source. The sink and source may be the same partition.

4.) The node distribution of source nodes for each node in the sink partition is then calculated as previously described. These are placed in the sort table.

5.) The nodes are sorted and reranked. If and only if there is no differentiation, the edge distribution is then calculated. Computational efficiency will be increased if the edge distribution for the last edge label type is ignored. This is because its information is determined entirely by the edge distributions for all the other edge labels, and so does not carry any additional differentiation potential. Edge distributions are prioritized according to edge label type, which is a real number.

6.) If the node distributions are identical for the sink partition, and edge distributions for all types are also identical, then no differentiation occurs and thus no propagation proceeds from these source partitions to this sink partition. After differentiation, the sink partition is flagged as a source partitions since it might break intermediate constitutional symmetry.

7.) Steps 1-6 are repeated for all remaining sink and source partitions.

At this point in execution, non-discriminating RD lists do not imply that the active partition is automorphic.
Figure 33. Example of the calculation of the cycle distribution of node $g$.

Cyclic distributions

Cycles are a part of the identity of a node in a graph\textsuperscript{94} — some cyclic graphs contain topological asymmetry as a result of cycle structure only. Randić enumerated the regular graphs on eight and ten nodes\textsuperscript{95} which clearly demonstrate the need to consider cyclic structure in graphs. His graphs that are not highly regular (as defined by Alavi\textsuperscript{96}) are somewhat difficult. The node distributions and edge distributions alone cannot discriminate some of his structures. This is because distance is a measure of only the shortest path between nodes, resulting in a spanning tree\textsuperscript{13} perspective with respect to the node. Incidentally, if we restrict our attention to the subset of graphs that are trees, the graph isomorphism problem can be solved in polynomial time.\textsuperscript{66} Undoubtedly, part of the identity of the node is characterized by the set of smallest cycles in which the node is a member. (This is distinct from Zamora's SSSR\textsuperscript{94,97}).

The distance (size) distribution of the set of smallest cycles (or cycle distribution for short) is calculated (Figure 33) for each node $g$ in the multinode partition using distance information of a temporary scratch graph, which is constructed from $A$ with $g$ (and any edges to it) deleted. A new distance matrix is calculated for the scratch graph. This is computationally expensive.
Figure 34. Example of the use of geodesics. There are two shortest paths (geodesics) between nodes $a$ and $b$. The numbers on the edges are labels that indicate two different types. Notice that the distribution of edge label types encountered on each geodesic are different. One geodesic has two edge labels of type 1 and one edge label of type 2. The other geodesic has one edge label of type 1 and two edge labels of type 2.

The distance $d(a,b)$ between each pair of each distinct pair of nodes $(a,b)$ that were connected to the now-removed node $g$ (which are not necessarily members of the active partition containing $g$) is fetched from the distance matrix of the scratch graph. If $d(a,b) < n$, then $g$, $a$, and $b$ are members of at least one cycle in the input graph with $(d(a,b)+2)$ members. The count of the geodesics (see below) between $a$ and $b$ gives the number of smallest rings of size $d(a,b)+2$, which is added to the $d(a,b)$th component of the distance distribution of cycles. The count forms the cycle distribution and is used as a sort key.

A standing conjecture in Graphsort is that the edge labels within the cycles have already been discriminated at this point in the algorithm by the propagation process. Considering them with respect to cycles will not provide any further differentiation. Despite substantial effort, no counterexamples to this conjecture have been discovered.
Geodesics

A geodesic is defined as the shortest distance path between two nodes $a$ and $b$. Figure 34 shows an example. Capobianco believes that using geodesics is a promising way to fully characterize a graph. An efficient algorithm to count the number of geodesics between $a$ and $b$ is as follows. An $n$-element array $Q$ is allocated. A node pointer, $p$, is initially set to point at $a$. The $p$th element of $Q$, $Q[p]$, is set to one. $Q[p]$ is added to each $Q[c]$ where $c$ is both connected to $p$ and is closer to $b$ than $p$. $Q[p]$ is then nulled and only then is this value of $p$ saved. Then $p$ is incremented (with wrap-around) to the next node with non-zero $Q[p]$ and $p \neq b$. This process is repeated until all the values of $Q[p]$ are deposited in $b$, which occurs when $p$ is the saved value. At the end of this procedure, $Q[b]$ contains the number of geodesics between $a$ and $b$.

Demonstration of differentiation of nodes

Let's see the differentiation process of Graphsort in action. We will follow the rank progression of two different graphs, one easy and the other difficult. Figure 35 illustrates the rank progression of isonicotinic acid, a metabolite of isoniazid. The graph of isonicotinic acid is drawn without hydrogens (a common practice in the trade) and aromaticity is ignored (i.e., the double bonds in the aromatic nucleus are a different edge label type than the single bonds) and arbitrarily encoded with the GRINS line code of O=C(O)C1=CC=CC=CC=C1. This graph is characterized by 9 nodes, a nullity of 1 and diameter of 5. Initially, all nodes (atoms) are ranked at 1 and local node names are constructed. After step A in Figure 35, the sort table after reranking reads:
<table>
<thead>
<tr>
<th>Ranked</th>
<th>Node</th>
<th>Local node name</th>
</tr>
</thead>
<tbody>
<tr>
<td>partitions</td>
<td>2</td>
<td>8 1 2 0 0 16</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>8 1 1 0 0 16</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>7 2 2 1 0 14</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>6 3 2 1 1 12</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>6 3 2 1 1 12</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>6 2 2 1 0 12</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>6 2 2 1 0 12</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>6 2 2 1 0 12</td>
</tr>
<tr>
<td>7</td>
<td>6 2 2 1 0 12</td>
<td></td>
</tr>
</tbody>
</table>

To expound, the local node name is constructed for node 2 (carbonyl oxygen) by concatenating its atomic weight (8), the number of non-hydrogens attached (1), a sorted sequence of connected edge labels (2,0,0) and the atomic weight (16). All other local node invariants do not discriminate and are not included.

Partitions 1-5 (the entire graph) are flagged as source partitions. The first source partition is partition 1 and the first sink partition is partition 4. The node distribution discriminates the sink partition, so the sort table after reranking (Step B) has split the original partition to partition 4 and 5:

<table>
<thead>
<tr>
<th>Ranked</th>
<th>Node</th>
<th>Local node name</th>
</tr>
</thead>
<tbody>
<tr>
<td>partitions</td>
<td>2</td>
<td>1 0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1 0</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0 1</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0 1</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0 1</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0 1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0 1</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>0 1</td>
</tr>
<tr>
<td>7</td>
<td>2 0 1</td>
<td></td>
</tr>
</tbody>
</table>

This says that node 1 (sp² carbon) has one source node (carboxyl oxygen) 1-away and node 3 has one source node 2-away.
Likewise in step C, source partition 1 differentiates partition 6 using node distributions. The end of the list of sink partitions, and the source partition is changed to 2. However, in step D, the partition 2 source cannot differentiate partition 6 with node distributions. However, edge distributions do discriminate. After sorting and reranking, the sort table reads:

<table>
<thead>
<tr>
<th>Ranked partitions</th>
<th>Node</th>
<th>Local node name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

which says that there is one edge label type 1-away from atom 4 and none 1-away of the same type for atom 8. The same happens in steps F and G; node distributions fail to discriminate the nodes in the sink partition, but edge distributions do not. After step G, all nodes are differentiated and ranked from 1 to 9.
Figure 35. Demonstration of the differentiation of nodes with a simple graph. a.) Left: hydrogen-suppressed molecular graph. Center: input route. The order of the atoms is indicated in the boxes. Right: route that is output from Graphsort. b.) Rank progression during operation. Each graph shows the intermediate rank state after sorting and reranking. The distributions calculated with respect to which partition are listed below each arrow. See text for detailed explanation.
Figure 36.1-2 shows a graph based on a Randic structure\textsuperscript{95} that is much more difficult than the previous example. Local node names fail to discriminate the nodes (step A, Figure 36.1); all nodes are trivalent and connected to two edges with label type 1 and one edge with label type 2. During propagation, the node distribution does not discriminate either (step B); each sink node has a node distribution of (3,4,2). Edge distribution, however, does differentiate the nodes into three partitions (step C), but propagation after this step does not differentiate any further nodes and thus terminates (step U). Notice the 18 undifferentiating steps that waste computer time. With the degree and nullity of the nodes in partition 1 not being of degree less than two, cycle distributions are calculated. However, cycle distributions do not discriminate partition 1; each node is involved in two four-membered cycles and four six-membered cycles. Thus, partition 1 is assumed to be automorphic, and hence are repelled by arbitrarily selecting a node and calculating the distance between it and the nodes in partition 1 which acts as a sort key (Step W). From this point on, a combination of nodes and label distributions suffice to completely differentiate the nodes (Step Y to BB).
Figure 36.1. Demonstration of the differentiation of nodes with a difficult graph modified from Randic.  

a.) Left: The graph with two types of edges, indicated by the thickness of the line. These can be thought of as single and double bonds if desired. Center: Initial order of nodes as determined from input line code. Right: Final order of the nodes after Graphsort. 

b.) Rank progression during operation. Each graph shows the intermediate rank state after sorting and reranking. The distributions calculated with respect to which partition are listed below each arrow. See text for detailed explanation.
Figure 36.2. Continued from the previous page.
**Figure 37.** Route tracing with isonicotinic acid. Compare to Figure 35a. Left: hydrogen suppressed graph. Center: ordering imposed by the differentiation process. Right: Final canonical reordering.

**Final reordering**

At this point in execution, the nodes are completely differentiated and have been ranked from 1 to \( n \), which could act as the canonical ordering. However, \( A \) must be reordered to follow a route for inversion back to the GRINS line code (Figure 37). A depth first search\(^{13} \) was employed. The route starts at the highest numerical value of rank. Thus, given the choice of local node naming, the initial node usually will be a terminal atom (hydrogen) with a low priority node label.\(^{37} \) At each juncture in choosing a route, the node with the smallest numerical value of rank is taken. When all nodes have been visited, the final order is defined. The input \( A \) is then finally reordered to \( A' \) while maintaining adjacencies. The GROANS algorithm converts it to line code.

**Implementation details.**

GRINS and GROANS respectively encode and decode the graph to \( A \). For highly intricate graphs,\(^{98} \) GRINS was bypassed and an ASCII version of \( A \) was converted to a binary input file by programs ACM2CNM. In some cases, the adjacency list was easier to input, and NBR2ACM converted it to an ASCII version of \( A \).
A prototype of Graphsort was written in C and has been ported successfully to VMS, UNIX, and MS-DOS environments. Much of the implementation techniques are described by Press et. al. Computer memory requirements were minimized by dynamically allocating the data structures. Since A is symmetric, one half of its data structure was used to store the distance matrix. Heapsort was used for sorting the sort table. Indirect addressing was extensively used because of the large data structures involved. The code is close to being optimized.

RESULTS

Performance with easy and difficult graphs

Table 17 lists a sampling of the graphs tested, ranging from highly asymmetric chemical graphs, disconnected graphs, subtly asymmetric graphs, and completely symmetric graphs. The most difficult graphs used in testing various other algorithms similar to Graphsort are included. Performance data in Table 17 lists was obtained by Turbo Profiler v2.0 (Borland International, Scots Valley, CA). The CPU index is the time required to differentiate the nodes on a 50 MHz 486DX computer with floating point chip (disk access is not included). The percentage of this time spent calculating distance matrices is tallied.

In addition to Table 17, Graphsort was tested with 10,205 different chemical structures that have been accumulated in the Klopman laboratory over the years. Each chemical structure graph was iterated 2n times as described in Figure 29. Graphsort did not fail for any of these structures.

The most challenging graphs for Graphsort is with subtly asymmetric graphs that are "almost" completely symmetric. While symmetric structures require the maximal amount of computer time, symmetric structures do not test the
completeness of the set of node invariants because there is no asymmetry to detect in the first place — any set of node invariants would assign the same name to topologically symmetric nodes. On the other hand, if two nodes are asymmetric, then only a portion of the list must be computed, thus saving computer time. Thus, execution time is sensitive to the symmetry of the structure. For example, two structures with the same number of nodes, 19 and 44 (Figure 37), significantly vary in CPU time required due to their difference in topological symmetry.

In 1970, Corneil and Gotlieb\textsuperscript{101} claimed to have constructed an efficient algorithm for graph isomorphism which classifies the vertex with degree. Further refinement occurs after artificially separating each node into its own cell. The properties of the final partitioning is represented by a labeled digraph. The authors conjectured that the final partitioning was the automorphism partitioning, but this conjecture has been shown to be false by types of graphs that originate from the study of combinatorics\textsuperscript{102} that are all constructed from association schemes.\textsuperscript{103} Such graphs are exceedingly rare special cases, and they have remained counterexamples to algorithms like Graphsort.

A short description of these combinatoric graphs is appropriate. All nodes in a regular graph have the same degree. Not immediately obvious is the fact that more than one automorphism partitioning can occur in regular graphs. A graph is strongly regular if the graph is regular and any two adjacent nodes have the same number of common adjacent nodes and any two non-adjacent nodes have the same number of common adjacent nodes.\textsuperscript{104} It is known that distance distributions cannot detect the automorphism partition of strongly regular graphs.\textsuperscript{87} Two other types of exceedingly difficult graphs are $n$-level regular and balanced incomplete block design (BIBD), which Mathon devised as a counterexample to the to Corneil and
Godlieb conjecture. Graphsort failed with one of these graphs called A_{25}. Figure 28 shows two representations of the same graph A_{25}.

**Average performance**

Graphsort has been integrated into the program GMK (written by Dr. Dimayuga\textsuperscript{105}) as an option which assigns a unique name to the structures in any D-file which can then be sorted and culled to produce a file that is free of duplicates. A database composed of molecules from many of the databases constructed for use by the CASE program\textsuperscript{20} was used to assess the average performance of Graphsort. The database, which contained 9601 structures, had an even distribution of size and symmetry; it contained no "mathematical" graphs. The entire database was assigned a unique SMILES code in 989 seconds (on an Alpha DECstation 300 model 400), which is about 10 structures per second. Much of this time was probably due to I/O.

**DISCUSSION**

**Comparison to other algorithms**

Graphsort is different from other algorithms in the field. I believe that it has the best balance of robustness (large input domain including labeled nodes and edges) and efficiency (no backtracking). Most algorithms were expressly designed to handle a particular subset of graphs. For example, the Weininger algorithm CANON which is used in CONCORD (Tripos, Inc., St. Louis) appears to be fast, but its input domain is strictly stable molecular structures, using SMILES as its line code. Weininger does not claim CANON to be graph-general.
In contrast, two other algorithms are noteworthy in their robustness. Rucker and Rucker\textsuperscript{106} used the elegant idea of raising the unlabeled adjacency matrix to ever increasing higher powers, \( w \), yielding a matrix whose off-diagonal entries record the number of possible walks\textsuperscript{4} between nodes \( i \) and \( j \) of length \( w \). At each power, the nodes are examined for differentiation. This procedure discriminates most graphs eventually. Their program TOPSYM (used with the IUPAC nomenclature program POLCYC) fails with the combinatorial graphs described above, plus the highly regular graphs of Randic (discussed above). This similarity may be due to the fact that distance matrices and power matrices appear to be closely related; as a matter of fact, the distance matrix is derivable from power matrices.\textsuperscript{81} While direct timing comparisons are not available due to differences in hardware, it appears that Graphsort may be about an order of magnitude faster than TOPSYM. Subsequently, Rucker and Rucker experimented with a procedural "shell" around TOPSYM called TOPSYM2.\textsuperscript{107} Designed to handle Mathon's splendid \( k \)-level regular graphs, TOPSYM2 exacts an extreme computational price. TOPSYM2 iterates TOPSYM for all combinations of pairs of nodes, thus bestowing the ability to handle 2-level regular graphs. Similarly, TOPSYM2 could consider all triplets of nodes in order to handle 3-level regular graphs, but the smallest 3-level regular graph is composed of 139,300 nodes, making this an overkill. In principle, the same procedure could be applied to Graphsort, but I do not believe it to be worth the cost, because Mathon's graphs most likely will not be encountered in practice.

Fowler et. al. also published an algorithm\textsuperscript{87} which uses node distributions and "squeeze tree" search techniques that is claimed to be valid for strongly regular graphs and BIBD's. However, it is not clear whether or not the input domain is
restricted to only these graphs. No direct timing comparisons are available. I must emphasize that this algorithm, as well as TOPSYM2, are for unlabelled graphs only, thus limiting their application to practical chemical situations.

**Concerns for robustness**

Graphsort was experimentally determined to be very robust. The torture list of Table 17, provided no counterexamples. As described above, 10,205 chemical compounds encountered in normal practice did not provide a counterexample either.

Graphsort relies heavily on the distance matrix of the input graph.* Moreover, the distance matrix of a derivative graph is used to determine the cycle distributions, which allows many more difficult graphs, accounting for its very large input domain. The distance representation is a richer representation than the adjacency matrix, which means that while the same connectivity information is present in an adjacency matrix, the information is more "accessible" in the distance matrix. Fortunately, this level of accessibility happens to be reasonably efficient to compute. Exceeding this level requires more computation, which is the case for the combinatorial graphs discussed above, where the identity of pairs of nodes must be calculated. A method that detects if an input graph is exceeding the capabilities of Graphsort, and reject graphs that can not be handled would be enormously useful. This method would not have to detect the exact topological symmetry in the graph, but rather that just a certain type of topological symmetry exists. Perhaps the characteristic properties of the distance matrix could be used in formulating this

---

*Incidentally, because Graphsort is based so heavily on the distance matrix, any hardware that would calculate the distance matrix would enormously speed up execution times.
method. In this way Graphsort would be completely reliable because its input domain would be clearly defined.

Does an efficient solution to the graph isomorphism problem really exist? The good news is that it appears to be for almost all graphs. The combinatorial graphs that we previously discussed are at best computable in $n^{k+3}$ (exponential) time, according to Mathon. I believe that these graphs will eventually lead to the proof that the graph isomorphism problem is actually in NP. However, an interesting fact about the graph isomorphism problem is that it has been solved for planar graphs in linear time. While this algorithm may be another example of making the problem easier by restricting the input domain to a subset of graphs, why would the fact that something is embeddable in a certain dimensionality affect isomorphism? Kuratowski showed that only two subgraphs (or their homeomorphs) are responsible for non-planarity: $K_{3,3}$ and $K_5$. It is tempting to say that these two subgraphs (or combinations thereof) are causing the difficulty.

The input domain of Graphsort does not include directed graphs. Extention of the input domain to include directed graphs is trivial. Of course, this would have to be accompanied by an extention of the GRINS language.
Table 17.1 Archetypal test graphs for Graphsort. The first column is the number referring to the pictorial depiction of the graph in the Appendix. The second column contains either the reference number to the Bibliography where the graph was obtained, or an asterisk, which indicates that it was selected from other sources or invented for this study. The third column is the input line code which was coded from an arbitrary route. The fourth column is time in seconds to differentiate the nodes. The fifth column is the standard, canonical line code.

<table>
<thead>
<tr>
<th>Graph no.</th>
<th>Ref no.</th>
<th>Input GRINS line code</th>
<th>Time (sec)</th>
<th>Output GRINS line code (canonical name)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>C=C(O)(C)(=O)C</td>
<td>0.0297</td>
<td>CC(=O)C(O)=C</td>
</tr>
<tr>
<td>2</td>
<td>39</td>
<td>X23=C4C3C2C=CC4</td>
<td>0.0312</td>
<td>X=12C3C=CC1C23</td>
</tr>
<tr>
<td>3</td>
<td>*</td>
<td>X123=XX2=XX3=X1</td>
<td>0.0321</td>
<td>X=2X1=XX13=XX23</td>
</tr>
<tr>
<td>4</td>
<td>*</td>
<td>X14=X2X3=X4X123</td>
<td>0.0330</td>
<td>X12=X4X3=X1X234</td>
</tr>
<tr>
<td>5</td>
<td>72</td>
<td>X23X45X26X4X3X56</td>
<td>0.0335</td>
<td>X1X5X24X13X2X345</td>
</tr>
<tr>
<td>6</td>
<td>65</td>
<td>XXX(X)(X)(X)(X)(X)</td>
<td>0.0336</td>
<td>XXX(X)(X)(X)(X)(X)</td>
</tr>
<tr>
<td>7</td>
<td>65</td>
<td>XXX(X)(X)(X)(X)(X)(X)</td>
<td>0.0348</td>
<td>XXX(X)(X)(X)(X)(X)(X)</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>X123X456%14%15%7%1</td>
<td>0.0341</td>
<td>X10%7%15%X148%29%36</td>
</tr>
<tr>
<td>9</td>
<td>*</td>
<td>X=X123=456X=X6X47X5</td>
<td>0.0371</td>
<td>X6=X2X15=XX123=47=XX</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>X2X(X)(X)X4XX3XX2XX4</td>
<td>0.0393</td>
<td>X2XX1XXX3XX1XX23</td>
</tr>
<tr>
<td>11</td>
<td>101</td>
<td>X1X2X34XX56X7X2X3789X9</td>
<td>0.0397</td>
<td>X3X28X9X146X5X17(XX34</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>X23XX4X2XX3XX4</td>
<td>0.0398</td>
<td>X2XX1XXX3XX1XX23</td>
</tr>
<tr>
<td>13</td>
<td>76</td>
<td>X23X4X3X4X5X6X5X6X2</td>
<td>0.0400</td>
<td>X4X1X2X1X2X3X5X3X45</td>
</tr>
<tr>
<td>15</td>
<td>50</td>
<td>X2X3XX4X3XX3XX4X2</td>
<td>0.0404</td>
<td>X2XX1XXX3XX1XX23</td>
</tr>
<tr>
<td>16</td>
<td>80</td>
<td>X2X3X6X5X2X4X5X6X34</td>
<td>0.0404</td>
<td>X4X3X2X1X5XX2XX3X45</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>X26X3X4X5X4X3X6X25</td>
<td>0.0411</td>
<td>X4X5X2X1X3X2X3X45</td>
</tr>
<tr>
<td>18</td>
<td>49</td>
<td>X2X3X6X5X(X2)X4X3X4X4</td>
<td>0.0412</td>
<td>X3X2X5X4X1XX3X45)X</td>
</tr>
<tr>
<td>19</td>
<td>*</td>
<td>X(X)XOXXXXXXXOXXXXXX</td>
<td>0.0414</td>
<td>XXXXXXXXXOXXXXXXXOXXXXXX</td>
</tr>
<tr>
<td>20</td>
<td>111</td>
<td>X2X3X4X3XX2XX6X3X5X46</td>
<td>0.0428</td>
<td>X24X2X5X3XX1X2XX3X45</td>
</tr>
<tr>
<td>21</td>
<td>43</td>
<td>X1X2XX(X)(X)X(X4)(X)X4X3</td>
<td>0.0445</td>
<td>X1X2X3XX(X)(X4)(X)X4X3</td>
</tr>
<tr>
<td>22</td>
<td>*</td>
<td>N1(C3)CCC2CCC(C(C3)</td>
<td>0.0451</td>
<td>C1CCN3CCC2CCC(C(C3)12</td>
</tr>
<tr>
<td>23</td>
<td>40</td>
<td>C2(O)=CC(O)=CC(O)=C2</td>
<td>0.0460</td>
<td>C1=C(C=C(C=C10))O</td>
</tr>
<tr>
<td>24</td>
<td>*</td>
<td>C2(C)C=CC(C)=C=CC1=</td>
<td>0.0490</td>
<td>C1=C=CC2=C=CC1=C=CC</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>X2X4X3X4X3X5X6X5X5X5X5</td>
<td>0.0505</td>
<td>X4X3X5X1XX2XX1X2XX2X</td>
</tr>
<tr>
<td>26</td>
<td>*</td>
<td>N12CCCN3CC(C4C23)CC</td>
<td>0.0507</td>
<td>C4CC1CCN3CCCN2CCC(C4)C1C23</td>
</tr>
<tr>
<td>Graph no.</td>
<td>Ref no.</td>
<td>Input GRINS line code</td>
<td>Time (sec)</td>
<td>Output GRINS line code (canonical name)</td>
</tr>
<tr>
<td>----------</td>
<td>--------</td>
<td>-----------------------</td>
<td>------------</td>
<td>-----------------------------------------</td>
</tr>
<tr>
<td>27</td>
<td>97</td>
<td>X2X3X4X5X7X5X6X3X3</td>
<td>0.0511</td>
<td>X6X15X3X6X1X(X6)X24X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5X6X6X4X7X2</td>
<td></td>
<td>XXX2X(X5)X3X3</td>
</tr>
<tr>
<td>28</td>
<td>106</td>
<td>X2=XX2X3X=XX4X3X=4X4</td>
<td>0.0569</td>
<td>X1=XX1X2X4X5X6X=XX4X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5X6X(X67)X=XX5X=X7</td>
<td></td>
<td>5X=X6X3X=XX23</td>
</tr>
<tr>
<td>29</td>
<td>21</td>
<td>X23X4X2X3X(X3)X4X2X5</td>
<td>0.0664</td>
<td>X5X1X2X(X6X5X3X6)X3X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X6XX(X6)X7XX5X7</td>
<td></td>
<td>2X3X4X(X3X4)X2X1</td>
</tr>
<tr>
<td>30</td>
<td>*</td>
<td>X123X4X5X3X6X7X9X%10X</td>
<td>0.0710</td>
<td>X17%13X6X9X18X%11%15X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5X%11X4X%10X9X11X8X%</td>
<td></td>
<td>23X4%12X25X3X4X5%10X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12X7X6X%13X4X%12</td>
<td></td>
<td>6X%10X14X%11%12%13X%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>%15X1X%14X%13X%15</td>
<td></td>
<td>13X%14X%15X7</td>
</tr>
<tr>
<td>31</td>
<td>49</td>
<td>C1CCCC1C(C)C2CCCCC2</td>
<td>0.0716</td>
<td>C(C2CCCCC2)C1CCCCC1</td>
</tr>
<tr>
<td>32</td>
<td>*</td>
<td>X1=X=7X=2X=3X=X=5X4X</td>
<td>0.0732</td>
<td>X=XX6X3X=X7X=X4X1=X5X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>=6X2X5X7X6X31</td>
<td></td>
<td>2X3=X4X5=X67</td>
</tr>
<tr>
<td>33</td>
<td>*</td>
<td>X15=X6X=X2X3=X6X4X=3</td>
<td>0.0787</td>
<td>X=XX1X2X3X1X=6X4X=X2X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X1=XX4X52</td>
<td></td>
<td>3=X4X56</td>
</tr>
<tr>
<td>34</td>
<td>*</td>
<td>C1C2C3C(C)C2CC3CC3(C)C</td>
<td>0.0803</td>
<td>C(C1CC2C3C4CC3C1C2C=C(C3OC4C(C)C5C)O)</td>
</tr>
<tr>
<td>35</td>
<td>112</td>
<td>X23X6X4X5X6X7X8X9X%10</td>
<td>0.0891</td>
<td>X13%13X2%12X%10X15X7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>%11X%12%13X%14X5</td>
<td></td>
<td>11X5X9X4X8X16X2X%14X3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>%15X2%13X%11%14X9</td>
<td></td>
<td>4X56X7X8X9X%10X%11%12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>%12X6%10X4X8X7%15X%</td>
<td></td>
<td>X%13%14X%15X</td>
</tr>
<tr>
<td>36</td>
<td>*</td>
<td>OC1CC2CN3CC=4C=5C =CC=CC5NC4CC3(C)CC2(C)C1C(=O)OC</td>
<td>0.1072</td>
<td>CC3CC2=NC=1C=CC=CC=CC1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C2CNC3C4CCC(C)C4(C)C</td>
<td></td>
<td>C(C)C(=O)OC</td>
</tr>
<tr>
<td>37</td>
<td>51</td>
<td>X12X8X4X5X6X7X8X9X%1</td>
<td>0.108</td>
<td>X12%16X3%18X1%15X2%1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0X8%11X3%12X%13%1</td>
<td></td>
<td>1X3%10X5X6X4X9X%13X4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8X2%12X1%15X%10%1</td>
<td></td>
<td>%12X5%17X6X7X%14X89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3X%14%15X%5X%9X%16X</td>
<td></td>
<td>X%10X%11X%12X%13X%12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6X%17X%14X%11X%17X%</td>
<td></td>
<td>4%15X%16X%17X%18</td>
</tr>
<tr>
<td>38</td>
<td>106</td>
<td>X23X4X5X6X7X8X9X%1</td>
<td>0.1090</td>
<td>X39X2X4X1X8X5X6X1X7X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0X4X7X5X3X6X%11X8X</td>
<td></td>
<td>%10X2X(X(X9%10X)X6X78X)X3X45</td>
</tr>
<tr>
<td>39</td>
<td>66</td>
<td>O=C2C=C3C4=C5C2=C=CC</td>
<td>0.1215</td>
<td>C(C2C=C3C4=C5C2=C=CC=C6C(C7=C(C(C5=C6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C6C(C=CC=C(C(C5=C6</td>
<td></td>
<td>7)=CC4)=OC(=O)C=C(=C)3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6)C4=C8C7=C5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>*</td>
<td>C(C1C(C)C2C(C)CN3CC</td>
<td>0.1415</td>
<td>C(C3C(C)C2C(C)C1C(=O)O)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>=4C=5C=CC(C)C=CC5N</td>
<td></td>
<td>C2C3CC1C4C=NC=5C=C(C(O)=C(C(C(O)=C(C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C4CC3CC2(C)C1C(=O)O</td>
<td></td>
<td>=C(C=O)C(C(O)=C(C(O)=C(C(O)=C(OC(O)=</td>
</tr>
<tr>
<td>41</td>
<td>43</td>
<td>X1=XX2X3X=XX1X23X4(</td>
<td>0.1802</td>
<td>X1=XX1X2X4X5X6X=XX4X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X5X=XX4X</td>
<td></td>
<td>5X=X6X3X=XX23</td>
</tr>
</tbody>
</table>
### Table 17.3. Archetypal test graphs for Graphsort. (Continued from the previous page.)

<table>
<thead>
<tr>
<th>Graph no.</th>
<th>Ref no.</th>
<th>Input GRINS line code</th>
<th>Time (sec)</th>
<th>Output GRINS line code (canonical name)</th>
</tr>
</thead>
<tbody>
<tr>
<td>42</td>
<td></td>
<td>C1=CC=CC=C123=(CC=C-C</td>
<td>0.2075</td>
<td>C3=CC=C=1=2(C=C3)(C=C-C=C-C=C)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C=C,C) )=(CC=C=C=C-3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td></td>
<td>N1=NC=CC=NN=CC=CN=</td>
<td>0.3663</td>
<td>C1=IC=NN=CC=CN=NC=CC=</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NC=CC=NN=CC=NC=C1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td></td>
<td>X15X2X3X4X5X6X7X4X8</td>
<td>0.6431</td>
<td>X1%10X4X3X2X1X9X7X2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X3X9X2X%10X1X6X%1</td>
<td></td>
<td>X6X3X5X4X11X8X5X6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1X%10X9X8X7%11</td>
<td></td>
<td>X7X8X9X%10%11</td>
</tr>
<tr>
<td>45</td>
<td></td>
<td>X23X4X5X6X7X8X9X%1</td>
<td>0.6441</td>
<td>X1%10X5X8X2X7X4X3X1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0X%11X2X%12X9X6X3</td>
<td></td>
<td>X2X9X6X3X%11X(X(X8X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X%10X7X4X%11X8X5%12</td>
<td></td>
<td>9X%10%11X67X%45</td>
</tr>
<tr>
<td>46</td>
<td>110</td>
<td>X23X4X5X6X7X2XX89X</td>
<td>0.822</td>
<td>X1X3X7XX28XX%11X%12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>%10X%11X(X)X%12X</td>
<td></td>
<td>XX12X4X13XX3X6XX45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X%13X(X)X9X(X)XX%14</td>
<td></td>
<td>XX(X%12X%13)X%10XX</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X(X8X)(X(X%10)(X(X%11)</td>
<td></td>
<td>59XX6X7XX89XX%10%1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X(X12)%X%14X%13</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>47</td>
<td>113</td>
<td>X19X2X3X4X1X5X4X(3</td>
<td>1.5290</td>
<td>X12%16X3%18X1%15X2%1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6)X7X5X8X7X6X2X1X2</td>
<td></td>
<td>12X3%10X5X49X7%13X4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X3X4X1X5X4X(X36)X7</td>
<td></td>
<td>%12X5%17X68X7%14X89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X5X%10X7X4X2X1X2X</td>
<td></td>
<td>X%10%11X%12X13X%14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3X4X1X5X4X(X36)X7X</td>
<td></td>
<td>%15X%16%17%18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5X(X8X%10)X7X6X29</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Concerns for efficiency

The efficiency of Graphsort is due to the the individual nature of the node invariants and their economical use. The algorithm is designed in three procedural levels: local node names, propagation, and cycle detection. Each procedural level describes features of graphs that are "independent." As a result, their associated node invariants are independent in discriminating power. Local node names are independent of rank and so cannot change for a given graph; thus they only need to be calculated once. Labeled tree graphs appear always to be discriminated by the first two levels: local node names and propagation. The relationship between the number of types of edge labels and efficiency is a balance between two factors. A greater number of types of edge labels implies less symmetry, but requires more
edge distributions to be calculated. This may result in a reasonable efficiency for the "average" labeled graph. The order of the procedural levels lends to efficiency. Local nodes is very quick and discriminates many chemical structures immediately. Propagation discriminates edges and nodes without recalculating a new distance matrix. Only if the structure can contain cycles, and all other options are exhausted, are the computationally expensive cycle distributions calculated. Finally, a large amount of efficiency is gained by sacrificing a small amount of robustness with regards to the combinatorial graphs that we discussed already.

The distance matrix calculation is where Graphsort spends most of its time, by far. The worst possible case for performance is where the distance matrix calculation is repeated $2n$ times. Using the method of Mueller, a $n^3 \log n$ algorithm, the algorithmic complexity of Graphsort would be of the order of $n^4 \log n$.

I suspect that there are many theorems in graph theory (in addition to the ones already employed) for special cases or classes of graphs. For example, there are probably several theorems that are based solely on enumeration of graphs below a certain number of nodes. Future versions of Graphsort could incorporate these theorems as "short cuts," which could drastically increase performance for the average case.

**Concerns for aesthetics of final line code**

In a sense, the canonical rules that select a single order of nodes to represent all possible permutations are arbitrary. The properties of the canonical rules is far less important than the requirement that they always select the same order. However, the properties of the canonical rules can greatly affect the aesthetics of the resulting line code. Given that robustness and efficiency are placed at a higher
premium than aesthetics, the development of Graphsort was never constrained by aesthetics. However, now that the canonical rules have been determined, aesthetics might be considered. If Graphsort was to be used exclusively by other chemical software, then aesthetics might be unimportant. However, if a human chemist will be reading or writing the GRINS line code, then some names might be preferable due to their brevity or their clarity.

Accordingly, one may wish to experiment with the optimization of various internal routines. The local node names perform the first differentiation in most TCE's, so their order of construction from node names greatly influence the aesthetics of the resulting line code. Presently, the sort table is always sorted in descending order, the distance distributions are ordered in increasing distances. In practice, local node names significantly reduce the degeneracy of most TCE's. The first node can be either a terminal node or an interior node. We use a depth-first search, but a breadth-first search changes the aesthetics because it tends to maximize the number of parentheses - this might be used to bring out the symmetry of tree graphs. A priority queue could even be established for finer control of the route tracing (Figure 36).

CONCLUSION

The unique naming problem, first defined in 1966, is a vexing problem which continues to this day. The most important finding of my research in the area is that the problem can be solved for all graphs in which the distance matrix can express the identity of the node, including its global environment, which is for the vast majority of graphs. The distance matrix is well worth its computational cost given its high discrimination power. While not general for all graphs, the discrimination power of
the distance matrix can be "trusted" for use in chemical algorithms. Counterexamples to Graphsort are known to exist, but the likelihood of encountering these in normal chemical usage is extremely remote. Graphsort was extensively tested and demonstrated reasonable execution times (Table 17 attests to this), making it a reliable foundation for high level software development. The Graphsort algorithm is a highly useful computational tool.
CHAPTER 6

The Three-dimensional Solution Structure
of Residues 1-28 of the Alzheimer's Disease β-peptide
INTRODUCTION

The most common cause of adult onset dementia is Alzheimer's disease (AD), now affecting approximately 12.5 and 47.2% of the population in the United States over the age 65 and 85, respectively.\textsuperscript{1} The brain tissue in patients with AD contains large quantities of insoluble amyloid plaques. A major component of amyloid plaque is the β-peptide,\textsuperscript{2,3} a small (39 to 43 amino acids) polypeptide with heterogeneous termini that is generated from cleavage\textsuperscript{4,5} of a larger amyloid precursor protein (APP) (Figure 38). The major β-peptide component of amyloid plaque contains 42 amino acid residues\textsuperscript{6} and is referred to as β-(1-42). The β-(1-28) and β-(29-42) peptides (Figure 38) occupy the extracellular and transmembrane regions of APP and β-(1-42).

Amyloid deposition is likely a critical step in the neurodegenerative processes associated with AD.\textsuperscript{4,7} Amyloid plaques are invariably associated with areas of nerve cell death, and injection of synthetic β-peptides directly into rat brain tissue produced cytotoxic effects.\textsuperscript{8} In addition, a direct association exists between the aggregational state of β-peptide and neurotoxicity.\textsuperscript{1} It has also recently been established that soluble extracellular β-peptide is normally produced in cultured cells and human biological fluids,\textsuperscript{9-12} which was predicted by Zagorski. Therefore, it is now critically important to identify the factors that cause soluble, extracellular β-peptide to form insoluble amyloid,\textsuperscript{9} and understand the mechanism of β-amyloidosis in AD.\textsuperscript{11}

\textsuperscript{*} This chapter is based on previously published work: Biochemistry, 1994, 33, 7788-7796.
Figure 38. Overview of the formation of β-peptide from the amyloid precursor protein, which contains 695 residues. The amino acid sequences of the β-peptides discussed in the text are also shown. This diagram emphasizes that, depending upon the conditions, residues 1-28 and 29-42 can exist in distinct conformations and aggregational states in solution. By contrast, in the amyloid deposit the β-peptide adopts an oligomeric β-pleated sheet structure.

Recent work has also established that in membrane systems, such as planar lipid bilayers and solvent-free lipid bilayers, the β-peptide does not precipitate as amyloid, but instead produces voltage-dependent ion channels.\textsuperscript{13,14} In AD a defect of normal channeling of potassium and calcium exists,\textsuperscript{15} together with a direct relationship between the aggregational state of β-peptide and the loss of calcium homeostasis.\textsuperscript{16} It was speculated that such alterations in ion homeostasis, along with β-peptide deposition as amyloid, may both contribute to the neurodegenerative processes associated with AD.

In vitro studies\textsuperscript{17-22} indicate that synthetic β-peptides can adopt either monomeric α-helical or oligomeric β-sheet conformations in solution. The
α-helical conformation is very soluble, whereas the β-sheet conformation is less soluble and eventually precipitates as amyloid. The α-helical conformation is favored in membrane-like conditions at high and low pH, while the β-sheet conformation is favored in water at mid-range pH values (Figure 38). The variance of conformations is largely due to the 1-28 peptide region of β-peptide, since the β-(1-28) peptide can produce soluble monomeric α-helical structures, as well as plaque-like oligomeric β-sheet structures, similar to those found in natural amyloid plaques. In contrast, the region composed of residues 29-42 is hydrophobic, sparingly soluble and aggregates into non-amyloid-like β-sheet structures.

Nevertheless, a molecular basis of amyloid and channel formation remains undefined due to the lack of definitive structural data for the β-peptide. The three-dimensional structure of the β-(1-28) peptide in solution is the subject of this chapter. This structure represents the first high-resolution structural model of a β-peptide fragment, and was derived using well-established two-dimensional (2D) NMR, molecular dynamics, and distance geometry techniques. Unlike the oligomeric β-pleated sheet structure in amyloid plaques, the solution structure in membrane-like media is almost entirely α-helical. The three-dimensional structure shown here will help to establish a molecular basis for amyloid formation, which may ultimately lead to the design of rational treatments to either block the β-peptide ion-channels or curtail amyloid formation in AD.
MATERIALS AND METHODS

Sample preparation

The β-(1-28) peptide was synthesized, purified, and characterized as described previously26 by Barrow et. al. A summary of their work is provided here. Samples were prepared for NMR measurements by dissolving the peptide (3.6 mg, 0.001 mmol) in a solution (0.50 mL) containing either: (1) perdeuterated sodium dodecylsulphate (450 mM) in a mixture by volume of H₂O/D₂O (9:1), or (2) H₂O:trifluoroethanol-d₃ (4:6). The perdeuterated sodium dodecylsulphate (SDS-d₂₅) and trifluoroethanol-d₃ (TFE-d₃) were purchased from Cambridge Isotopes, Inc. All solutions also contained 0.5 mM Na₂EDTA and 0.05 mM NaN₃. The pH of a solution was adjusted with a pH meter (Omega Engineering, Inc.) equipped with an electrode (Microelectrodes, Inc.) that fit inside the 5 mm NMR tube. The desired pH values were obtained at room temperature by adding microliter amounts of either dilute DCI or dilute NaOD. No corrections of pH readings were made for isotope effects or for the presence of the SDS-d₂₅ or TFE-d₃, since control experiments showed that these substances did not significantly alter the pH.

Procedures for NMR data acquisition and analysis

All proton (¹H) NMR spectra were obtained at 500 and 600 MHz using General Electric GN-500, Bruker AM-500 or Bruker AM-600 spectrometers. Two-dimensional NMR data were transferred to Indigo XS24 (Silicon Graphics, Inc.) computer workstations and processed using the FELIX program (version 2.05, Biosym, Inc.). Chemical shifts are referenced to internal sodium
Figure 39. Portion of a NOESY contour plot of the amide-amide (NH-NH) region. The NH(i) to NH(i+1) through-space interactions from Tyr10 to Asp23 are shown with the NH assignments along the diagonal. NOEs from the aromatic 2,6H's of Tyr10 to the NH's of Tyr10 and Glu11, together with NOEs from the aromatic 2,6H's of Phe20 to the NH's of Phe20 and Ala21 are shown at the top of this plot. Spectra were recorded at 500 MHz using a 2 mM sample of β-(1-28) and a 450 mM sample of SDS-d25 dissolved in water (9:1, H2O:D2O) at pH 3.0 and 25.0°C.

3-(trimethylsilyl)propionate-2,2,3,3-d4 (TSP). Probe temperatures were calibrated using dry methanol.27

Two-dimensional nuclear Overhauser enhancement spectroscopy (NOESY28) was run in the phase sensitive mode with quadrature detection in both dimensions.29 The carrier was placed in the center of the spectrum at the position
of the H$_2$O signal, and the H$_2$O signal was suppressed by irradiation with the proton decoupler. For the NOESY experiment, the irradiation was carried out during the recycle delay (which varied from 2.0 to 5.0 s) and during the mixing time (which varied from 25 to 200 ms). Reduction of coil lead pick-up$^{30}$ was accomplished by modification of the pulse sequences to include 90°$_x$, 90°$_y$, -x, -y pulses for both preparation and detection.$^{31}$ For spectral assignments, before Fourier transformation spectra were multiplied by a Lorentzian-to-Gaussian weighting factor in F$_2$, and a 60-90° phase-shifted sine bell in F$_1$. For quantitative measurements of NOE build-up rates,$^{32}$ spectra were instead multiplied by a 90° phase-shifted sine bell in both the F$_2$ and F$_1$ dimensions as this avoided biasing peaks with narrower linewidths. Baseline roll was reduced by careful adjustment of the intensity of the first points in F$_2$ and F$_1$,$^{33}$ and by the application of a cubic-polynomial baseline correction of the rows in the final 2D matrix. The spectral widths in both dimensions were ± 2604 Hz, and a total of 512 increments (each consisting of 64-128 scans) was acquired with 2048 complex points for the F$_1$ and F$_2$ dimensions.

**Distance geometry and simulated annealing computations**

All computations and graphics visualizations were performed on Indigo XS24 (Silicon Graphics, Inc.) computer workstations using the programs X-PLOR$^{34}$ (version 3.0) and INSIGHT II (Biosym, Inc.). The distance constraints were obtained from NOESY data by measuring NMR cross-peak volumes with the FELIX program and were normalized for the number of atoms for a particular interaction. Of the 204 inter-residue NOEs, 152 distance restraints ($r_{ij}$) were quantitatively derived from the cross-peak volumes ($v_{ij}$) of the NOESY spectrum with mixing time ($t_m$) of 150 ms using the equation $r_{ij} = r_{kl} (v_{kl}/v_{ij})^{1/6}$ where $k$ is the
Hβ's of Tyr10, l is the He's of Tyr10, and \( r_{kl} \) was 2.49 Å. The final restraints employed were 1.8 Å as the lower bound and \( r_{ij} + 0.5 \) Å as the upper bound. The remaining 52 inter-residue NOE cross-peaks were classified as medium (1.8 to 3.3 Å) or weak (1.8 to 5.0 Å).\(^{35}\) These latter restraints were derived from a NOESY spectrum obtained with a \( t_m \) of 400 ms. The 47 intra-residue distance restraints were quantitatively obtained from cross-peaks in the NOESY spectrum (\( t_m \) of 150 ms) using the above equation, except that \( k \) and \( l \) were defined as the closest pair of atoms that formed a fixed distance. The final intra-residue restraints employed were 1.8 Å as the lower bound and \( r_{ij} \) as the upper bound.

The previous NMR data showed that the amide protons (NH's) for residues 5 to 27 were hydrogen bonded within an \( \alpha \)-helical structure.\(^{31}\) Using the NH temperature coefficients (\( \Delta \delta / \Delta T \)) which were previously published,\(^{31}\) the 23 hydrogen bonds were classified into three categories: 6 weak (\( \Delta \delta / \Delta T \geq 8.0 \)), 9 medium (\( 8.0 > \Delta \delta / \Delta T \geq 5.0 \)) and 8 strong (\( \Delta \delta / \Delta T < 5.0 \)). On the basis of these categories, the distances between the nitrogen of residue \( i \) and the carbonyl oxygen of residue \( i-4 \) were restrained to 2.3 to 3.2 Å for the weak, 2.5 to 3.2 Å for the medium, and 2.6 to 3.2 Å for the strong bonds. The distance restraint between the NH of residue \( i \) and the carbonyl oxygen of residue \( i-4 \) was 1.6 to 2.2 Å for all three categories of hydrogen bonds. This procedure kept the angles closer to ideality for the stronger hydrogen bonds. Repeating the X-PLOR distance geometry/simulated annealing (DGSA) computations without the inclusion of any hydrogen bonds produced an essentially identical backbone structure within residues 9 to 28.

The vicinal coupling constants\(^{36}\) (\( J \)) were separated into three categories on the basis of their magnitude: small (3 to 5 Hz); medium (6 to 7 Hz); and large (8 to 10 Hz). The three \( J \) categories were then converted into torsion angles with error
limits: $\pm 30^\circ$ (small $J$), $\pm 40^\circ$ (medium $J$), and $\pm 50^\circ$ (large $J$). The harmonic force constant was 10 kcal mol$^{-1}$ rad$^{-2}$. Using the above restraints, the DGSA protocol was applied within the X-PLOR program. The 100 initial structures generated with metric matrix distance geometry were then subjected to molecular dynamics. For each structure, this involved an initial temperature of 3000 K, 5000 heating steps, 5000 cooling steps and 2000 steps of Powell energy minimization. The timestep was 0.003 ps and the dielectric constant was 30. The remaining parameters employed in this study are described in the DGSA protocol in the manual supplied with X-PLOR.

RESULTS

NMR measurements

A nearly complete $^1$H NMR assignment of the $\beta$-(1-28) peptide dissolved in aqueous TFE-d$_3$ solution using standard 2D NMR techniques and sequential assignment procedures was previously accomplished by Zagorski. Using similar methods, the complete $^1$H NMR assignment of the $\beta$-(1-28) peptide in SDS-d$_{25}$ solution was obtained to elucidate its structure when micelle-bound. This approach, which adequately mimics the molecular environment of biological membranes, has been successfully applied to many polypeptides. To form uniform micelles and ensure that the $\beta$-(1-28) peptide was indeed micelle-bound, a relatively high concentration of SDS-d$_{25}$ (450 mM) was used, which is well above the critical micelle concentration of 8 mM.

The expanded region of the 2D NOESY spectrum shown in Figure 39 demonstrates that numerous, close (2.4 to 3.6 Å) through-space interactions or nuclear Overhauser effects (NOE's) are present between neighboring NH protons.
Table 18. Build-up times of selected peaks.

<table>
<thead>
<tr>
<th>Peak assignment</th>
<th>$t_m=50$ ms</th>
<th>$t_m=100$ ms</th>
<th>$t_m=150$ ms</th>
<th>$t_m=200$ ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: YH3→YH2.6</td>
<td>3.99</td>
<td>7.33</td>
<td>9.20</td>
<td>11.2</td>
</tr>
<tr>
<td>2: AαH→AαCH3</td>
<td>0.902</td>
<td>2.96</td>
<td>3.84</td>
<td>4.71</td>
</tr>
<tr>
<td>3: K16αH→F19βH</td>
<td>0.319</td>
<td>1.66</td>
<td>2.64</td>
<td>1.82</td>
</tr>
<tr>
<td>4: F19αH→E22βH</td>
<td>0.0532</td>
<td>0.240</td>
<td>0.00903</td>
<td>0.288</td>
</tr>
<tr>
<td>5: H13αH→K16βH</td>
<td>0.0297</td>
<td>0.232</td>
<td>0.0348</td>
<td>0.893</td>
</tr>
<tr>
<td>6: A21αH→V24γH</td>
<td>0.0297</td>
<td>0.0479</td>
<td>0.111</td>
<td>0.311</td>
</tr>
<tr>
<td>7: Q15αH→V18γH</td>
<td>0.0780</td>
<td>0.668</td>
<td>0.403</td>
<td>0.153</td>
</tr>
<tr>
<td>8: Q15αH→V18βH</td>
<td>0.101</td>
<td>0.767</td>
<td>0.507</td>
<td>1.90</td>
</tr>
<tr>
<td>9: E11αH→H14βH</td>
<td>0.0170</td>
<td>0.201</td>
<td>0.865</td>
<td>1.43</td>
</tr>
<tr>
<td>10: L17αH→F20βH</td>
<td>0.574</td>
<td>6.57E</td>
<td>11.4</td>
<td>13.8</td>
</tr>
<tr>
<td>11: V18αH→A21βH</td>
<td>0.0327</td>
<td>0.932</td>
<td>0.427</td>
<td>1.15</td>
</tr>
<tr>
<td>12: A21αH→V24βH</td>
<td>0.0264</td>
<td>0.00824</td>
<td>0.0313</td>
<td>0.0314</td>
</tr>
<tr>
<td>13: F20αH→D23βH</td>
<td>0.0921</td>
<td>0.159</td>
<td>0.139</td>
<td>0.162</td>
</tr>
</tbody>
</table>

This particular type of NOE, NN(i, i+1), together with other weak-sized αΝ(i, i+1) NOEs and medium-sized αΝ(i, i+2), αΝ(i, i+3), αΝ(i, i+4) and αβ(i, i+3) NOEs observed in other spectral regions, establishes that the peptide exists in a well-defined α-helical conformation.\(^{35}\) The presence of αΝ(i, i+4) NOE's indicates that a 3₁₀-helix is not present. In addition, no alteration in the intensity of the majority of the NOEs was seen upon raising the temperature from 25 to 50°C. This demonstrates that the α-helix is stable and free from rapid conformational averaging. Essentially identical NOE and circular dichroism data (CD) were observed in aqueous solutions containing either SDS or TFE,\(^{17,31}\) indicating that the β−(1-28) peptide adopts similar α-helical structures under both solution conditions.

**Three-dimensional structure computations**

The three-dimensional structure of the α-helical conformation was determined using the NMR data, and the well-established DGSA protocol.\(^{34,38}\) The
NMR data consisted of 251 NOEs, 23 hydrogen bonds, and 27 dihedral angle constraints. Of the 251 NOEs, 204 were inter-residue and 47 were intra-residue. NOE build-up rates (Table 18) obtained from NOESY data acquired with mixing times of 25, 50, 100, 150, and 200 ms, relaxation delay of 4.3 s, showed that spin diffusion was negligible. Initially, 100 structures were generated by DGSA, of which 95 possessed very similar conformations and energies, demonstrating a very high degree of convergence to a single family of structures (Figure 40a). The remaining 5 structures were significantly strained, conformationally different, and were discarded by the X-PLOR ACCEPT procedure. All accepted structures were close to ideal geometry and had no distance restraint violation greater than 0.3 Å (Table 19). The averaged structure, which was obtained from the 95 accepted structures and minimized with the restraints, had a low mean X-PLOR potential energy of 8.1 kcal mol\(^{-1}\) (Table 19). This demonstrates that virtually no strain was induced by the restraints.

A similar three-dimensional structure was obtained using a different protocol,\(^{39}\) which included the distance geometry algorithm implemented in the program DGII\(^{40}\) as part of the INSIGHT II software package (Biosym, Inc.). This result suggests that the DGSA X-PLOR approach did not introduce computational artifacts.

**Consistency of the structure with the NMR constraints**

The structure consists of two right-handed α-helices (residues 2-11 and 13-27), which are connected by a bend centered at Val12. The bend is consistent with the observed periodic nature of the αH chemical shifts,\(^{41}\) and also with the NMR-derived NH temperature coefficients. Repeating the DGSA computations
### Table 19. Structural characterization for the 95 β-(1-28) structures

<table>
<thead>
<tr>
<th></th>
<th>SA</th>
<th>&lt;SA&gt;&lt;r&lt;sub&gt;1&lt;/sub&gt;</th>
<th>&lt;SA&gt;&lt;r&lt;sub&gt;2&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rmsd from experimental restraints</strong>&lt;sup&gt;2&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance restraints (Å)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inter-residue (204)</td>
<td>0.044 ± 0.002</td>
<td>0.027</td>
<td></td>
</tr>
<tr>
<td>Intra-residue (47)</td>
<td>0.090 ± 0.002</td>
<td>0.082</td>
<td></td>
</tr>
<tr>
<td>Hydrogen bonds (46)</td>
<td>0.012 ± 0.007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total (297)</td>
<td>0.052 ± 0.001</td>
<td>0.040</td>
<td></td>
</tr>
<tr>
<td><strong>Torsion angle restraints (degrees)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ϕ (27) + χ (2)</td>
<td>0.016 ± 0.002</td>
<td>1.160</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>X-PLOR potential energies (kcal/mol)</strong>&lt;sup&gt;3&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond</td>
<td>5.3 ± 0.3</td>
<td>2.5</td>
<td>0.6</td>
</tr>
<tr>
<td>Angle</td>
<td>15.1 ± 1.2</td>
<td>4.8</td>
<td>0.1</td>
</tr>
<tr>
<td>Improper</td>
<td>3.4 ± 0.4</td>
<td>0.8</td>
<td>0.5</td>
</tr>
<tr>
<td>Total</td>
<td>23.8 ± 1.8</td>
<td>8.1</td>
<td>0.6</td>
</tr>
<tr>
<td><strong>Rmsd from idealized geometry</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bonds (Å)</td>
<td>0.005 ± 0.000</td>
<td>0.003</td>
<td>0.001</td>
</tr>
<tr>
<td>Angles (deg)</td>
<td>0.56 ± 0.02</td>
<td>0.31</td>
<td>0.10</td>
</tr>
<tr>
<td>Dihedrals (deg)</td>
<td>34.1 ± 1.6</td>
<td>35.2</td>
<td>35.0</td>
</tr>
<tr>
<td>Impropers (deg)</td>
<td>0.45 ± 0.03</td>
<td>0.22</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>Rmsd of coordinates (Å)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Backbone</td>
<td>0.69 ± 0.24</td>
<td>1.47 ± 0.27</td>
<td></td>
</tr>
<tr>
<td>Non-hydrogens</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA vs &lt;SA&gt;</td>
<td>0.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;SA&gt; vs &lt;SA&gt;&lt;r&lt;sub&gt;1&lt;/sub&gt;</td>
<td>1.25</td>
<td>1.98</td>
<td></td>
</tr>
<tr>
<td>&lt;SA&gt;&lt;r&lt;sub&gt;1&lt;/sub&gt; vs &lt;SA&gt;&lt;r&lt;sub&gt;2&lt;/sub&gt;</td>
<td>0.12</td>
<td>0.15</td>
<td></td>
</tr>
</tbody>
</table>

<sup>1</sup>Rmsd represents the root-mean-squared-deviation. SA is the family of 95 accepted structures resulting from the X-PLOR/DGSA procedure, <SA><r<sub>1</sub> is the averaged family structure minimized with distance and torsion restraints applied, and <SA><r<sub>2</sub> is <SA><r<sub>1</sub> minimized without distance and torsion restraints.

<sup>2</sup>The van der Waals and electric potential energies were zero within the experimental errors.

<sup>3</sup>Rmsd was calculated between distinct pairs in SA and ranged from 0.18 to 1.99 Å for the backbone atoms and from 0.72 to 2.80 Å for the non-hydrogen atoms.

Without the hydrogen bond restraints produced an essentially identical bent α-helical backbone structure, so the bend alone is not a computational artifact of differences in hydrogen bond restraints.
Figure 40. Left: The final accepted family of 95 structures of the β-(1-28) peptide superpositioned using the backbone atoms (dark). The N-terminus is shown at the top and the C-terminus is shown at the bottom. The side-chains (light) of His13 and Lys16 are labeled to emphasize their proximity. Right: An expanded view is shown along the helical axis between Tyr10 to Val18. This segment, which corresponds to the region within the box, is part of an averaged and minimized structure that was generated from the 95 structures. The ζNH₃⁺ of Lys16 is solvent exposed, consistent with the NMR data.31
The bend is consistent with the weak-sized $\alpha N(i, i+2)$ NOE's from Val12 to His14, from Val18 to Phe20, and from differences in other NOE intensities along this region of the peptide. The temperature dependence of the NH chemical shifts showed a weak NH bond for Val12, and a cluster of strong NH bonds within the His13-His14-Gln15-Lys16-Leu17 segment, whereas the next four residues (Val18-Phe19-Phe20-Ala21) have weaker hydrogen bonds. The latter four residues may also act as a hydrophobic shield to the NH protons of His13 to Leu17. These data, together with several side-chain NOE's among Leu17, Phe20, and Val24, suggest that slight packing of the hydrophobic residues occurs, in accordance with the observed bend of the helix.

Several of the measured $J$-coupling constants between the NH and $\alpha H$ protons indicate that the helical structure may be in rapid equilibrium with an extended chain structure. Although the majority of the residues have $J$-values less than 6 Hz, only Val12, Phe19, and Phe20 have values under 4 Hz that are normally associated with rigid helical backbone structures. The $J$-values greater than 6 Hz are mostly found at the ends of the molecule, where more deviation from helicity is to be expected due to the increased mobility or possible “fraying.” However, the relevance of the $J$-coupling constants to the structure may be dubious, since $J$-coupling constants usually are one of the least sensitive NMR parameters for the detection of folded conformations in flexible linear peptides. Large populations of helical structure in peptides can have $J$-values of 6 to 7 Hz. Moreover, a recent report established that $J$-coupling constants can be inaccurate, and the discrepancies occur between those measured from 2D DQF-COSY and resolution-enhanced 1D spectra. To compensate for the possible $J$-value errors in the structure
Figure 41. The average rmsd between the final 95 energy minimized structures and the averaged/minimized structure of the β-(1-28) peptide. Two values are given for each residue: the backbone C, Cα, N, and O subset of atoms (filled bars) and the side-chain atoms (unfilled bars). The rmsd values were determined using the modified AVERAGE procedure within the program X-PLOR.

computations, a reduced force constant of 10, rather than a more standard value of 50, was used for the 27 φ/ψ backbone dihedral angle constraints.

In contrast to the J-coupling constant values, the NOE data obtained at 50°C were very similar to the data seen at 25°C, demonstrating that the β–peptide adopts an ordered α-helical backbone structure that is not subject to conformational averaging. This is supported by the observation of numerous medium-range αN(i, i+2), αN(i, i+3), αN(i, i+4) and αβ(i, i+3) NOE's, together with the small temperature coefficients (less than 5). In addition, the majority of the αH proton chemical shifts are upfield relative to random coil values, indicating helical structure.

For the mean atomic positions of the 95 structures, the root-mean-squared deviation (rmsd) is 0.48 Å for all backbone atoms and 1.04 Å for all non-hydrogen atoms. Residues 4 to 25 are the most defined with a backbone rmsd of 0.41 and a non-hydrogen atom rmsd of 0.83 Å (Figure 41). The side-chains of the majority of
the residues are also well-defined, particularly in the middle region of the peptide (residues 12 to 24) except for Lys16. This high-degree of structural definition can be rationalized by the numerous NMR restraints (greater than 15 restraints per residue), including stereospecific assignments of the prochiral pairs of protons. The three valine residues (Val12, Val18, and Val24) have rmsd below 0.5 Å, and all exist in a trans configuration, in accordance with the coupling constant and intra-residue NOE data. Val12, Val18 and Val24 displayed large vicinal coupling constants (11.2, 11.8, and 9.6 Hz) between the α and β protons.\textsuperscript{31} Except for Val18, in which both γCH\textsubscript{3} groups are chemically-shift equivalent, all valine residues showed separate strong and weak NOEs from each NH to each γCH\textsubscript{3}. The side-chains of His13, His14, and Gln15 likewise adopted unique and well-defined conformations with rmsd values less than 0.45 Å (Figure 40, left). The Gln15 side chain displayed separate weak- and medium-sized intra-residue NOEs from each εNH\textsubscript{2} proton to each γCH\textsubscript{2} proton. The Leu17 δCH\textsubscript{3} group had NOEs to the aromatic 2H of the His13 and the 4H of His14; also, the Val18 γCH\textsubscript{3} groups displayed NOEs to both the 2H and 4H of His14.

By contrast, the side chains of Arg5, Lys16, and Lys28 do not adopt distinct conformations, with somewhat higher rmsd of 2.2, 1.1, and 2.1 Å, respectively. This weaker definition results from the lack of NOE-distance restraints. The lack of NOE data observed at the termini; notably, Asp1, Asn27 and Lys28, also cause substantially high side chain rmsd in the range 1.2 to 2.9 Å (Figure 41). This may be caused by "fraying" at the ends of the peptide, explaining the apparent disorder seen in the overlaid structures (Figure 40, left).
Physiological relevance of the solution conditions

The objective of this study was to determine the membrane-mediated structure of the β-(1-28) peptide using 2D NMR data. Difficulties with high resolution NMR studies of peptide conformations in membranes arise from line broadening due to the high molecular weight of the peptide-membrane aggregates. A useful alternative is to select an organic solvent or micellar system in which the conformation of the peptide would resemble that of the membrane. SDS micelles have been reported to adequately mimic the environment of membranes.\textsuperscript{37,48,49} In addition, TFE is a membrane-mimicking solvent that promotes either inter- or intra-molecular hydrogen bonding, analogous to SDS micelles or phospholipid vesicles, and stabilizes α-helices and β-sheets that are in equilibrium with random coil structures.\textsuperscript{50} Studies using CD and NMR established that the propensity for helix formation in a peptide is a prerequisite for the induction of an ordered helix by TFE; thus TFE will only stabilize (not induce) helical structures in peptide segments that have the propensity to do so based on their amino acid sequences.\textsuperscript{51,52} Similar conclusions were reached by others,\textsuperscript{53,54} where only those peptide segments that correspond to helical segments in the native protein (as determined by X-ray diffraction in the solid state) adopt helical structure in TFE-water solutions. On one hand, a potential pitfall is that TFE and SDS may provide an environment to which the β-peptide adopts a non-native conformation. On the other hand, when used as co-solvents with water, these solvents may mimic a more natural and realistic environment that exists for a given peptide in vivo, such as found in lipid/water interfaces. For instance, for the peptide gastrin, the degree of secondary structure in
TFE/water solutions correlated well with its biological potency. Given, the general difficulty in solubilizing the β-(1-28), these solution conditions should give the best available picture of the β-(1-28) three-dimensional structure.

Many studies indicate that only amphiphilic peptides form helical structure in SDS solution. Thus, because the β-(1-28) peptide is not amphiphilic, the observation of helical structure in SDS solution may be exceptional. The β-(29-42) peptide adopts either a β-sheet or a random coil structure rather than an α-helical structure in TFE and SDS solutions and similar conclusions were obtained with other peptides.

Numerous reports have revealed membrane-like components in preamyloid deposits. Additionally, there are extensive experimental data which suggest that alterations of membrane integrity in brain cells and endothelial cells occur in AD, and many reports have correlated these changes to the occurrence of amyloid deposits. Thus, the rationale for NMR studies in membrane-like environments is that the β-peptide exists in membrane-like compartments within preamyloid deposits, and the β-peptide is found on the basement membrane of cerebral blood vessels, which are preferred sites for early stages of amyloid deposition. Residues 8 to 17 of β-peptide can associate and bind to lipid receptors; also, the 29-42 segment should interact favorably with lipids or membranes, as this segment is hydrophobic and naturally membrane-bound in APP (Figure 38).

**Relationship between bioactivities and structures of β-peptide**

Although the biological role of the amyloid β-peptide has yet to be determined, it is thought to contribute to the progressive neuronal loss of AD. The introduction of purified β-peptide from amyloid plaque cores causes focal neuronal
Figure 42. Schematic $\beta$-(1-28) structure indicating charge distribution extrapolated to neutral conditions. A "+" was superimposed on the nitrogen of the ionized amino/imino group and a "-" on the ionized carboxylic group.

damage,\textsuperscript{68} and injection of synthetic $\beta$-(1-40) peptide into rat brain tissues resulted in $\beta$-amyloid deposits with surrounding nerve cell degeneration.\textsuperscript{8} However, regarding the neurotoxicity tests, considerable discrepancies exist across different laboratories.\textsuperscript{*} Several groups have reported that synthetic $\beta$-peptides show either trophic or toxic responses on neurons in vitro. The reasons for these deviations are unknown, although they are believed to arise in part from differences in the aggregation states and the solution structures of the $\beta$-peptides.

It was recently established that for the $\beta$-(1-40) peptide the random coil structure is non-toxic, while the $\beta$-sheet structure is neurotoxic.\textsuperscript{69} In addition, the amount of $\beta$-sheet structure correlated with the levels of neurotoxicity. Over time, in aqueous solution, the $\beta$-(1-40) and $\beta$-(1-42) peptides adopt greater amounts of oligomeric $\beta$-sheet structure,\textsuperscript{17,18,26} and these results are consistent with the ability of "aged" or aggregated $\beta$-peptide to be neurotoxic, and "fresh" or monomeric $\beta$-peptide to be non-toxic and enhance neurite outgrowth.\textsuperscript{1,70} Interestingly, if the aggregation process is reversed by dissolving "aged" $\beta$-(1-42) in

\textsuperscript{*} For a review, see Neurobiol. Ag. 1992, 31, 535-623.
hexafluoroisopropyl alcohol, which promotes α-helical structure for residues 1-28 and random coil structure for residues 29-42,26 the neurotoxicity disappears.1 One possibility is that the monomeric α-helical conformation is the neurotrophic species, and when an α-helix (monomeric) → β-sheet (oligomeric) transformation occurs the β-peptide becomes neurotoxic. To evaluate this hypothesis, additional studies examining the effects of the α-helical structure on nerve cell cultures are required.

Table 20. Comparison of primary sequences of mellitin, alamethicin, and β-(1-28). Ace, Aib, and Phi corresponds to N-acetate, α-aminoisobutyric acid and L-phenylalaninol, respectively.

<table>
<thead>
<tr>
<th>Residue</th>
<th>Mellitin</th>
<th>β-(1-28)</th>
<th>Alamethicin</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gly</td>
<td>Asp</td>
<td>Ace</td>
</tr>
<tr>
<td>2</td>
<td>Ile</td>
<td>Ala</td>
<td>Aib</td>
</tr>
<tr>
<td>3</td>
<td>Gly</td>
<td>Glu</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Ala</td>
<td>Phe</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Val</td>
<td>Arg</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Leu</td>
<td>His</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Lys</td>
<td>Asp</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Val</td>
<td>Ser</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Leu</td>
<td>Gly</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Thr</td>
<td>Tyr</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Thr</td>
<td>Glu</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Gly</td>
<td>Val</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Leu</td>
<td>His</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Pro</td>
<td>His</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Ala</td>
<td>Gln</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Leu</td>
<td>Lys</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Ile</td>
<td>Leu</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Ser</td>
<td>Val</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Trp</td>
<td>Phe</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Ile</td>
<td>Phe</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Lys</td>
<td>Ala</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Arg</td>
<td>Glu</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Lys</td>
<td>Asp</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Arg</td>
<td>Val</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Gln</td>
<td>Gly</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Gln</td>
<td>Ser</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>NH₂</td>
<td>Asn</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
General features of the β-(1-28) structure

Overall, the three-dimensional structure of the β-(1-28) peptide consists of a bent α-helix with the bend centered at Val12. At mid-range pH, five positively-charged and six negatively-charged residues would reside on opposite faces of the helix, imparting a large electrostatic dipole perpendicular to the helical axis. Except for the disorder at the termini (Figure 40, left), the 95 computed structures showed a high degree of convergence with relatively low rmsd values (Figure 41). This high degree of structural definition results from the large number of NMR constraints, which included at least 15 restraints per residue for the region between residues 4 and 25. It should be kept in mind that a low rmsd does not prove a small conformational space was sampled for the peptide in solution, but rather establishes that a high degree of precision was obtained for the solution coordinates. To examine these possibilities, additional NMR relaxation time studies using 13C and 15N enriched β–peptides must be performed to compare the dynamics and local mobilities in different regions of the peptide.

The helical structure is consistent with the NMR data, except for some of the J-coupling constants between the αH and NH protons.

Homologies with ion-channel forming peptides

The β–(1-40) peptide produces cation-selective channels in bilayer membranes.13,14 Recent 2D NMR studies of the β–(1-39) peptide in TFE solution demonstrated that residues 1 to 28 are helical and residues 29 to 39 are disordered (Zagorski et al., unpublished). This latter interpretation is consistent with CD data, in which the β–(29-42) peptide is primarily random coil in TFE solution, and the β–(1-39) and β–(1-28) peptides adopt 45% and 65% α-helix structure.
Figure 43. Overlap of the backbones of three peptides that have ion-channeling as a common property. Color code: black is β-(1-28), blue is alamethicin, red is mellitin.

respectively. These results suggest that, in TFE solution and perhaps in bilayer membrane-like solutions similar to SDS, the same residues have α-helical conformations in both peptides.

The structure of the β-(1-28) peptide was compared to the structures of alamethicin and mellitin, which are small peptides that likewise form ion channels in bilayer membranes. Superpositioning of the backbone conformation of the β-(1-28) peptide, with the X-ray crystallographic backbone structures of alamethicin and mellitin revealed that, despite the lack of any obvious sequence homology (Table 20), the backbone tertiary structures of these three peptides are very similar, particularly in the degree of helical bending. The rmsd between residues 2-21 of alamethicin and 1-20 of mellitin is 1.84 Å; the rmsd of regions 1-20 of the β-(1-28) peptide to alamethicin and mellitin are 1.74 Å and 1.47 Å, respectively. In solution or the crystalline state, melittin folds to form two α-helices joined by a bend between Thr11 and Gly12. The β-(1-28) peptide has a helix-helix angle of 21°, which is almost identical to the 20° value of melittin in methanol solution. This angle was determined by a least-squares fit between the β-(1-28) peptide and two appropriately-sized ideal right-handed α-helices.
Figure 44. Electrostatic dipole caused by side chain ionization state extrapolated to neutral pH. The cylinder represents the α-helix of Figure 42. Overall, one face of the α-helix is negatively charged while the other is positively charged. The vector sum of the charges induces a large electrostatic dipole moment perpendicular to the helical axis. Note that the strand polarity is indicated by the N-terminus and C-terminus.

The bend in β-(1-28) cannot be ascribed to amphipathicity, nor by the absence of a hydrogen bond. Many α-helical segments in proteins are curved, bent, or kinked.77 The bends are usually amphipathic or associated with a proline residue, as in the case of melittin and alamethicin. The amphipathicity generally causes a bend by allowing the hydrophobic face to be concave and the hydrophilic face to be convex, with the centrally located proline promoting helix bending by the absence of a hydrogen bond. DeGrado et. al. have found that a synthetic melittin analog which does not contain proline forms a tetramer in aqueous solution.

The β-(1-28), alamethicin, and melittin likewise share several biophysical properties.* First, the β-peptide and melittin both induce lipid phase separation and decrease the fluidity of membranes.78 Second, replacement of Pro14 in melittin dramatically alters its channel forming properties.72,79 By analogy, amino acid substitutions near the bend region of β-peptide may also destroy its channel activity. For example, substitution of the His13-His14 segment in β-peptide promotes its

---

* It has recently come to my attention that the antibiotic zervamicin is a short peptide with backbone bending and ion-channeling properties similar to alamethicin. See Karle, I. L. et. al., Proc. Natl. Acad. Sci. 1991, 88, 5307-5311.
precipitation as a $\beta$-pleated sheet structure.\textsuperscript{19} Thus, the bend may be important for stabilization of the $\alpha$-helical structure as well as for ion channel activity. The third similarity is a $\beta$-sheet $\rightarrow$ $\alpha$-helix conversion, which occurs for $\beta$-peptide\textsuperscript{17,31} and was proposed to occur within the alamethicin ion channel.\textsuperscript{80} Given the wealth of information concerning both the structural and functional properties of alamethicin and melittin,\textsuperscript{72} these peptides provide good starting points in designing models of ion channel formation in $\beta$-peptide.

\textit{Exploring the homology with simple modeling}

These structural homologies suggested a unifying hypothesis that may explain simultaneously the ion-channeling and aggregation properties of the $\beta$-peptide, assuming that it is valid to extrapolate the $\beta$-(1-28) structure reported here to the entire peptide in neutral pH conditions. Mellitin is an amphipathic $\alpha$-helix which crystallizes in space group C222\textsubscript{1} with dimers as the asymmetric unit. The asymmetric unit of mellitin is composed of a dimer with a nearly perfect two-fold non-crystallographic symmetry. The pathway of crystallization is probably partial $\alpha$-helix formation $\rightarrow$ dimer formation $\rightarrow$ tetramer formation.\textsuperscript{74} This process of aggregation is driven by forces that maximize the entropy of the solvent, which Eisenberg\textsuperscript{81} summarized as the hydrophobic dipole moment. I speculate that the \textit{electrostatic} dipole moment (Figure 44) that is perpendicular to the helical axis
Figure 46. Formation of a tetramer by parallel translation of Figure 45 with charges on the surface. This view is parallel to the axes of the helices. The channel in the center might form a pore in a membrane that conducts ions. Provides the driving force of the packing of the β-peptide into similar α-helical bundles.

Accordingly, I overlaid the β-peptide on the crystal structure of mellitin dimer obtained from the Brookhaven Protein Data Bank. Many orientations about the helical axis in conjunction with a two-fold rotation will produce a good match of electrostatics between the monomers with negligible steric hindrance. Figure 45 illustrates one possible packing. This pair of dimers can easily associate in an antiparallel manner to form a tetrameric 4-α-helical bundle as in mellitin (Figure 46). The exterior of the resultant tetramer is hydrophilic and charged at physiological pH, providing a site of nucleation/predilection for further association into repeating periodic units as shown in Figure 47. This aggregate may be the species responsible for the ion-channeling properties. A synthetic peptide based on a similar hydrophobic packing of tetramers was synthesized by Lear et. al, and it exhibited ion-channeling properties.82

This model of quaternary structure also qualitatively predicts the shape of the β-peptide solubility vs. pH function. The actual state of ionization of the
carboxyl and amino moieties of the side chains strongly depend on pH. Clearly, the electrostatic forces promoting aggregation of the β-peptide would be maximized in the region of neutral pH. Oppositely, the α-helices would possess overall like charges at pH extremes, thus repelling and solublizing the aggregate.

The β-peptide in the neuritic plaques are known to be in the cross-β-pleated sheet conformation similar to silk, where the side chains alternately point up and down. For normal β-pleated sheet structures, hydrogen bonding occurs between individual strands, forming sheets that associate through hydrophobic forces. Kirshner reports that the β-sheets are stacked in antiparallel mode, not parallel, due to hydrophobic and charge interactions. As in the α-helix, there is significant spatial charge separation perpendicular to the strand axis in the β-sheet conformation: +3 vs. -2 net charge at neutral pH (Figure 48). Once hydrogen bonded, the newly formed sheet could have a two-dimensional pattern of charges that is self-complementary such as to promote β-sheet antiparallel stacking (Figure 49).
Figure 48. β-peptide in the cross β-pleated sheet conformation, which occurs in the amyloid plaques. Hydrogen bonding between antiparallel strands would occur perpendicular to this page.

Additionally, the polarity of the strands in the β-sheet aggregate (Figure 49) is the same as the polarity of the α-helical aggregate (Figure 47). Thus, an uncoiling of the α-helices to β-sheet would not incur a high entropic cost. An interesting modeling experiment would be to determine if a pathway exists that is energetically favorable, especially by maintaining the charge complementarity between monomers during the transition from α-helix→β-sheet. Such a slow transition, which might partially explain the 30-year preclinical phase\textsuperscript{84} of AD, might be inhibited by a drug designed from knowledge of this quaternary structural model. Such a drug would probably not need to be highly specific because, as far as I am aware, α→β transitions are rare in normal physiology.

**Binding motif of proteoglycans and other proteins**

Amyloid plaques are compact and their formation is selective to particular anatomical features,\textsuperscript{85} which may suggest that something is providing a nucleation site for aggregation. Binding of the β-peptide to the highly negatively charged heparan sulfate proteoglycans (HSPG) enhances the aggregation of β-peptide.\textsuperscript{86} By comparing consensus sequences from other known HSPG-binding proteins, the Val12-His13-His14-Gln15-Lys16-Leu17 segment was identified as a likely site for binding with HSPG.\textsuperscript{87} Regarding the α-helical structure, the side-chains of His13
Figure 49. Charge complementarity of cross-β-sheet conformation. The strand of β-peptide in the cross β-pleated sheet conformation was abbreviated to emphasize charge distribution. This view illustrates how charge complementarity between the antiparallel sheets could stabilize the structure. Note that continued antiparallel packing in this way would generate a structure with the same strand polarity as the structure shown in Figure 47.

and Lys16 are proximate and reside on the same face of the molecule (Figure 40, right). Both His13 and Lys16 are important residues for amyloid fibril production,18,23 and rodents with the His13 → Arg mutation do not develop mature amyloid deposits.88 There is some evidence that addition of heparin to a solution of the β-(1-28) peptide causes the 2H and 4H NMR signals of His13 to shift and broaden.89 This result, together with the proximal location of His13 and Lys16, is consistent with the proposition that these residues may constitute a binding motif with glycosaminoglycans.86,87,90 In fact, an analogous structural motif may be involved in the binding of the β-peptide to transthyretin, a normal protein component of plasma.91 A structural model for the complex of transthyretin and the α-helical structure of the β-peptide showed that the side chains of His13 and Lys16 form part of a positive potential binding surface that makes contact with the complementary negative potential surface of transthyretin.

Fate of β-peptide in vivo

Outlined in Figure 50 is a hypothetical scheme that correlates the various biological activities with the different structures of β-peptide. Barrow and Zagorski originally proposed that the β-peptide normally exists in human biological fluids in a
Figure 50. Scheme for the fate of β-peptide after being generated by proteolysis of APP and eventually precipitating as amyloid in a β-sheet structure. This updated scheme incorporates an earlier version by Barrow and Zagorski\textsuperscript{26} but now incorporates more recent results about the ability of β-peptide: (1) to exist normally in soluble forms\textsuperscript{9-12} and (2) to produce ion channels in membrane-like media\textsuperscript{14}. Our results suggest that the β-peptide ion channel may consist of a tetrameric α-helical structures.

soluble form\textsuperscript{17}. Subsequently, this proposition was supported using both in vivo and in vitro studies\textsuperscript{9-12}.

Once released by proteolytic cleavage of APP, the β-peptide may exist in solution as either a random coil or an α-helical structure that are both monomeric and very soluble. At physiological pH, the α-helices can associate to produce tetramers\textsuperscript{26,31}. The structural similarities with melittin and alamethicin suggest that a tetrameric assembly of α-helices for β-peptide may constitute its ion channel\textsuperscript{13,14}. Zagorski previously postulated that an α-helix → β-sheet conversion occurs for
β-peptide during amyloid formation in AD, and this pathway may be mediated by the tetramer. At mid-range pH, five positively-charged and six negatively-charged residues would reside on opposite faces of the helix, and this charge configuration could account for the helical association that occurs. Related α-helix (soluble, membrane-bound) → β-sheet (insoluble, aqueous solution) conversions of other peptides and proteins, including the scrapie prion proteins, are well-known.92-94

Many theories about amyloid deposition in AD support the notion that local environmental factors in the human brain promote the precipitation of β-peptide from a soluble, non-toxic form to an insoluble, toxic form. In the amyloid plaque, the β-peptide exists in an oligomeric β-pleated sheet structure that is resistant to further proteolysis and degradation.95 One possible environmental factor is the binding of HSPG to β-peptide, as discussed above. In addition, the observation of the α-helical structure for β-peptide in membrane-like environments can account for some of the features during β-amyloidosis. It is thought that amyloid deposition occurs in stages, where initially the more soluble and non-toxic preamyloid deposits57,96,97 form prior to the insoluble, toxic amyloid plaques. In preamyloid deposits, β-peptide exists in an unpolymerized state83 and is not stained by cross β-pleated sheet specific dyes such as Congo red and thioflavin-S.98 These facts, together with the knowledge that numerous membrane-like components are present in preamyloid deposits,58,97 are consistent with an α-helical structure for β-peptide in preamyloid deposits.
EPILOGUE
Results of research often focus attention to new possibilities and the work described in this dissertation is no exception. A few of them will now be discussed.

The META program can be interfaced with other layers of software, such as the CASE program. If integrated into commercial drug design packages, metabolic concerns could be addressed immediately during a session of molecular modeling. By performing META methodology "backwards," prodrug candidates could be designed automatically from existing drugs.\(^1\) Incorporation of goal-seeking behavior into META using heuristics could assist in the exploration of metabolic pathways, allowing the program to find a specific metabolite or pathway among the myriad of possibilities. A NCSA Mosaic home page using forms would provide a front-end to the new "information superhighway," allowing expertise on xenobiotic metabolism to be traded over the Internet.

Another methodological approach to modeling metabolism is to construct an object-oriented model of metabolism. Object-oriented programming provides many advantages for such data-intensive problems. The most important advantage is the correspondence of the program architecture and the system being modeled, which alleviates the "growing pains" usually encountered in enlarging the model and code.\(^2\) Since there are relatively few types of objects in metabolism (e.g. enzymes, compartments, compounds) but many instances of each, the advantage of inheritance is great. The initial construction of object base classes would be required but worth the time because the classes would be extensively reused.

---

\(^1\) Incidentally, the legal precedent for this will be set in the near future. See *Chemical and Engineering News*, August 29, 1994, p. 9.

\(^2\) See *Advanced Systems*, August 1994, p. 81.
As discussed, the generality of Graphsort allows it to be applied to practically any programming situation that manipulates topological structure. Many different software applications can be built around it, allowing the automation of chemical "possibility thinking" (like META) which would otherwise be too difficult to do without computer recognition. Graphsort is not limited to chemical applications. For example, graphs generated from the edge-detection of a photographic scene could be uniquely named, and if the name is present in a table of known objects, then that object in the scene would be recognized. Expert systems that recognize topological objects can be constructed and point the inquirer to further data.³

The structure of the β-(1-28) peptide provides some understanding of plaque formation in Alzheimer's Disease, which is probably caused by the aggregational properties of the β-peptide. As previously discussed, the three-dimensional conformation of the β-(1-28) peptide may explain the aggregation of the β-peptide. Recently, the activity of prions, which are proteins that are infectious without any genetic material, was found to be caused by an aggregational process that is essentially a conformationally dependent seeding event.⁴ In general, the basic mechanism of protein aggregation is not clearly understood at a molecular level. NMR spectroscopy techniques are best-suited to provide non-invasive information about the aggregational process of the β-peptide. Such studies might tell us whether aggregation is caused by a general phenomenon, or rather by many subtle interactions, as occurs in protein folding.

³ A similar effort, but for combinatorial objects, was accomplished by mathematicians. See Science. 1994, 265, 473.
(1) Hansch, C. A Quantitative Approach to Biochemical Structure-Activity

(2) Levine, R. R. Pharmacology: Drug Actions and Reactions: Little, Brown

(3) Smith, D. H. In Artificial Intelligence Applications in Chemistry; T. H.
Pierce and B. A. Hohne, Eds.; American Chemical Society: Washington,

(4) Spann, M. L.; Chu, K. C.; Wipke, W. T.; Ouchi, G. Use of Computerized
Methods to Predict Metabolic Pathways and Metabolites. J. Environ. Pathol.
Toxicol. 1978, 2, 123.

(5) Wipke, W. T.; Ouchi, G. I.; Chou, J. T. In Structure-Activity Correlation as
a Predictive Tool in Toxicology; L. Goldberg, Ed.; Hemisphere: New York,

(6) Tinker, J. F.; Gelernter, H. Computer Simulation of Metabolic

(7) Johnson, M. A.; Gifford, E.; Tsai, C.-c. In Concepts and Applications of
Molecular Similarity; M. A. Johnson, Ed.; John Wiley & Sons, Inc.: New

(8) Mavrovounitis, M. L.; Stephanopoulos, G.; Stephanopoulos, G.
1990, 36, 1119-1132.

(9) Darvas, F. Predicting metabolic pathways by logic programming. J. Mol.
Graphics 1988, 6, 80-86.


Spritzer, G. A.; Searleman, J. E. Empirical Explorations of SYNCHEN.
Science 1977, 197, 1041-1049.


(23) Aeppli, L. Teratologic Studies on Imipramine in Rats and Rabbits. A contribution to layout and interpretation of teratologic studies with special


**CHAPTER 2**


(69) Oesch, F.; Schladt, L.; Knehr, M.; Dohmer, J.; Thomas, H. In *Molecular Aspects of Monooxygenases and Bioactivation of Toxic Compounds*; E.


(76) Leibman, K. C. Reduction of Ketones in Liver Cytosol. Xenobiotica 1971, 1, 97-104.


(78) Fu, P. P. Metabolism of NitroPAHS. Drug Metab. Rev. 1990, 22, 209-268.


(122) Gillette, J. R. Effects of Induction of Cytochrome P-450 Enzymes in the Concentration of Foreign Compounds and Their Metabolites and on the Toxicological Effects of These Compounds. *Drug Metab. Rev.* 1979, 10, 59-87.


(125) Sayre, L. M. Biochemical mechanism of action of the dopaminergic neurotoxin 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP). *Toxicol. Lett.* 1989, 48, 121-149.


**CHAPTER 3**


(64) Klopman, G. In Chemical Reactivity and Reaction Paths; G. Klopman. Ed.: John Wiley & Sons; NY, 1974; pp 55-166.

(65) Klopman, G.; Macina, O. In Advances in Molecular Modeling; Jai Press Inc.: 1988; Vol. 1; pp 1-40.


**CHAPTER 4**


(4) Harary, F. *Graph Theory*; Addison-Wesley: Reading, Massachusetts, 1969.


(85) Buckley, F. *Distance in Graphs*; Addison-Wesley: Redwood City, CA, 1989.


CHAPTER 6


(16) Mattson, M. P.; Tomaselli, K. J.; Rydel, R. E. Calcium-destabilizing and neurodegenerative effects of aggregated β-amloid peptide are attenuated by basic FGF. Brain Res. 1993, 621, 35-49.


(39) Marcinowski, K.; Zagorski, M. G., Notes from DGII execution.


(64) Blusztajn, J. K.; Gonzalez-Coviella, I. L.; Longue, M.; Growdon, J. H.; Wurtman, R. J. Levels of phospholipid catabolic intermediates, glycerophosphocholine and glycerophosphoethanolamine, are elevated in brains of Alzheimer’s Disease, but not of Down’s syndrome patients. *Brain Res.* 1990, 536, 240-244.


(73) Fox, R. O.; Richards, F. M. A voltage-gated ion channel model inferred from the crystal structure of alamethicin at 1.5-Å resolution. Nature 1982, 300, 325-330.


(76) Ikura, T.; Go, N. Proteins 1991, 9, 81-89.


(89) Keane, C.; Zagorski, M., Notes from heparin experiments.


APPENDIX
APPENDIX

Pictorial depiction of the test graphs from Table 17. Different edge types are depicted by light and bold lines. Different node labels are included or implied by filled or unfilled circles.