The Spirit Of The Undertaking: Origins In Macsyma And Dendral

6.871-- Lecture 3

MACSYMA: Symbolic Mathematics

- Goals of the Project
- System Description
- Lessons

Goals of Project

To help applied mathematicians in solving problems

$$\int \frac{x^4}{\left(1-x^2\right)^{\frac{5}{2}}} =$$

Symbolic Mathematics: AI Approaches

- Slagle: SAINT
- Moses: SIN
- Moses and Martin: MACSYMA
- Reduce-II
- Mathematica, Matlab

SAINT: Symbolic Automatic Integrator

$$\int \frac{x^4}{\left(1-x^2\right)^{\frac{5}{2}}} dx$$

Try *y = arcsin x*, yielding:

$$\int \frac{\sin^4 y}{\cos^4 y} dy$$

$$\int \frac{\sin^4 y}{\cos^4 y} dy$$

three possible ways to deal with this:

$$\int \tan^4 y\, dy$$

$$\int \frac{z^4}{1+z^2} dz$$

$$\int \left(-1+z^2+\frac{1}{1+z^2}\right) dz$$

$$-z+\frac{z^3}{3}+\int \frac{dz}{1+z^2}$$

try w = arctan z

$$\arcsin(x)-\tan(\arcsin(x))+\frac{1}{3}\tan^3(\arcsin(x))$$

$$\int \cot^{-4} y\, dy$$

$$-\int \frac{dz}{z^4\left(1+z^2\right)}$$

$$\int 32 \frac{z^4}{\left(1+z^2\right)\left(1-z^2\right)^4} dz$$

(from z = tan(y/z))

SAINT

- Steps
  - 26 standard forms (1-step solutions, tables)
  - 8 Algorithmic transforms (eg. sum of integrals)
  - 10 Heuristic transforms, of which derivative divides is "the most successful"
    - » Goals evaluated on depth of integrand
    - » Ex., xe^{x^2} is of depth 3

SAINT

- Worked like the average engineer, i.e., lots of search and backtracking
- Conceived of in terms of search, worked *because* of that. The power comes from:
  - Problem decomposition
  - Methodical exploration of alternatives
  - Looking far, wide, and deep
  - Speedy tree construction, search, backtracking
- Success is just a matter of trying enough alternatives

SAINT

Some interesting statistics:

Saint's Average Performance

	Subgoals	Unused Subgoals	Level	Heuristic Level
32 Author problem	6.4	2.0	3.5	1.0
52 MIT Problems	4.7	0.8	2.9	.8
84 Problems	5.3	1.25	3.0	.9

The Mindset Shift

SAINT will frequently [need to] explore several paths to a solution … because it lacks the powerful machinery that SIN possesses.

One of the striking features of these programs is how little knowledge they require in order to obtain a solution. Persson in his recent thesis dealing with "sequence prediction" seems to feel that placing a great deal of context dependent information in a program would be "cheating." This emphasis seems to be useful when one desires to study certain *problem solving mechanisms* in as pure a manner as possible.

We, on the other hand, intended no such study of specific problem solving mechanisms, but mainly desired a powerful integration program which behaved closely to our conception of expert human integrators.

SIN, we hope, signals a return to an examination of complex problem domains.

-- Moses, 1963.
[emphasis added]

Sin

- Steps
  1. Derivative divides
  2. 11 specific methods
     - Substantial effort in deciding which to apply
     - Largely organized around recognizing the form of the problem
  3. General purpose methods (e.g., search)
- Note the sequence.
- "We feel that too few AI programs employ the fact that in many problem domains there exist methods which solve a large number of problems quickly."

Macsyma Organization

- 5000 operations
- User-driven
- Independent operations

Macsyma Lessons

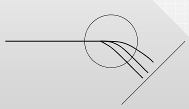
- Keep the system modular and loosely coupled
  - It is sometimes cheaper to translate one representation to another in order to solve the problem more efficiently
  - Use of a common language for communication makes this approach tractable (eg, dense and sparse polynomials)
- Do not duplicate knowledge
  - leads to unmanageable system

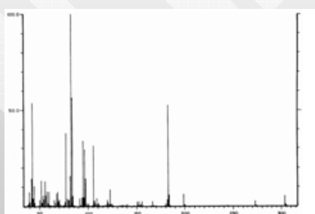
Symbolic Math Lessons

- Character of the problem changes as knowledge evolves
  - SAINT
    - » Worked as people *appeared* to: extensive search and backtracking
  - SIN
    - » Almost always correct on the first guess: found the sources of power in the domain
  - RISCH: Algorithmic Integration
    - » Guaranteed to succeed if the expression is integrable
      - Uses very special representation
      - Computationally complex and expensive
      - Process not understandable to users but provably correct.

Dendral: Structure Elucidation

- Given:
  - Empirical Formula: $C_9H_{18}O$ (total MW = 142)
  - Known Structure Constraints
  - Mass Spectrum



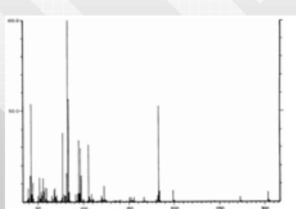


Result

```
        O
        ||
C – C – C – C – C – C – C – C – C
```


How to Proceed?

- Given:
  - Empirical Formula: $C_9H_{18}O$ (total MW = 142)
  - Known Structure Constraints
  - Mass Spectrum



- Catalog?

Generate and Test

```
   |
— C —          H —          — O —
   |
```

For C_9 H_{18} O two possible structures are

```
    O                                       O
    ||                                      ||
C – C – C – C – C – C – C – C – C    C – C – C – C – C – C – C – C – C
```

Difficulties in Generate & Test

The Generator Should Be:

Complete
Irredundant
Informed

The overall paradigm should be:

PLAN
GENERATE
TEST

The need in structure elucidation:
Empirical formula $C_{20}\ H_{43}\ N$
Possible Structures: 14, 715, 813

How Can the Program Plan Its Attack?

What should the program *know?*
Rules: spectrum features ⇒ molecule class

IF There are peaks at M1 and M2 such that
M1 + M2 = MW + 28 and
M1 is high and M2 is high
THEN The structure is one of the ketones

IF There is a high peak at 44 and
there is a high peak at M1 – 44
THEN The structure is one of the aldehydes

Knowledge Representation

- Efficiency vs. Comprehensibility
  vs. Additivity
  vs. Modifiability
- Level of representation

Efficiency and …

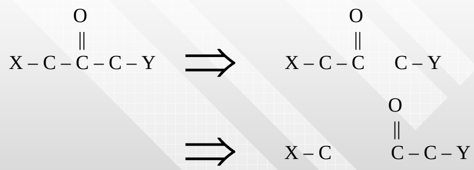
If high peak at 57 and high peak at 113
Then ketone

If high peak at 57 and high peak at 98
Then ether

Level of Representation

IF
There are peaks at M1 and M2 such that
M1 + M2 = MW + 28 and
M1 is high and M2 is high
THEN
The structure is one of the ketones

Level of Representation



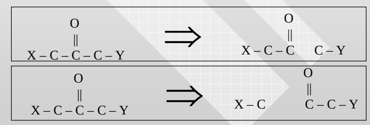
IF There are peaks at M1 and M2 such that
M1 + M2 = MW + 28 and
M1 is high and M2 is high
THEN The structure is one of the ketones

Representation Punchline

Lesson:
Use the

Highest level
Most Transparent
Easily modified

representation you can find



In the Knowledge Lies the Power

Empirical formula: $C_{20} H_{43} N$

Information Sources	Possible Structures
Topology	42,867,912
Chemistry	
Mass Spectrum	
Chemist's Information	
NMR	

In the Knowledge Lies the Power

- Lesson:
  Knowledge can obviate the need for search.
  (If you know where to look you don't have to search)
- Lesson
  Knowledge migrated from the tester to the generator.
  (It's often better to have a smart generator)

Building the Program Advances The Field

- The SAINT, SIN, MACSYMA, Risch progression
- Dendral's accumulation, rationalization and development of chemistry knowledge.