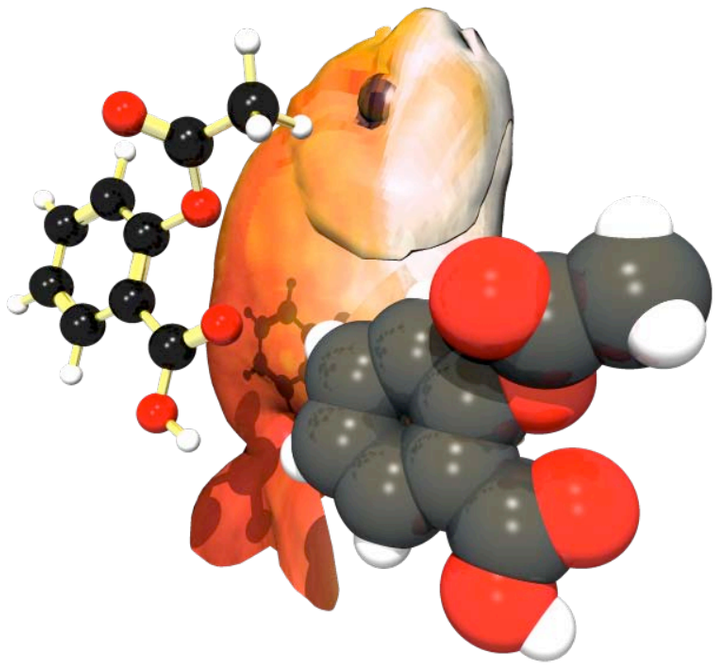
Open Babel: File Translation for Computational Chemistry and Nanoscience



Dr. Geoffrey Hutchison

Department of Chemistry & Chemical Biology
Cornell University
Ithaca, NY 14853-1301

NNIN/CNF Fall Workshop
October 11, 2005

Open Babel: The Questions...

Answers to the six important questions:
(not just Frequently Asked Questions…)

- Why?
- When?
- Where?
- Who?
- What?
- How?

Open Babel: The Questions...

Answers to the six important questions:

- Why?
- When?
- Who ? Where?
- What? How?

Challenges:
A Plethora of File Formats

Currently supported input types

alc -- Alchemy file
prep -- Amber PREP file
bs -- Ball & Stick file
caccrt -- Cacao Cartesian file
ccc -- CCC file
c3d1 -- Chem3D Cartesian 1 file
c3d2 -- Chem3D Cartesian 2 file
cml -- Chemical Markup Language file
crk2d -- CRK2D: Chemical Resource Kit 2D file
crk3d -- CRK3D: Chemical Resource Kit 3D file
box -- Dock 3.5 Box file
dmol -- DMol3 Coordinates file
feat -- Feature file
gam,gamout -- GAMESS Output file
gpr -- Ghemical Project file
mm1gp -- Ghemical MM file
qm1gp -- Ghemical QM file
hin -- HyperChem HIN file
jout -- Jaguar Output file
bin -- OpenEye Binary file
mmd,mmod -- MacroModel file
out,dat -- MacroModel file
car -- MSI Biosym/Insight II CAR file
sd,sdf -- MDL Isis SDF file
mdl -- MDL Molfile file
mol -- MDL Molfile
mopcrt -- MOPAC Cartesian file
mopout -- MOPAC Output file
mmads -- MMADS file
mpqc -- MPQC file
bgf -- MSI BGF file
nwo -- NWChem Output file
ent,pdb -- PDB file
pqs -- PQS file
qcout -- Q-Chem Output file
ins,res -- ShelX file
smi -- SMILES file
mol2 -- Sybyl Mol2 file
unixyz -- UniChem XYZ file
vmol -- ViewMol file
xyz -- XYZ file

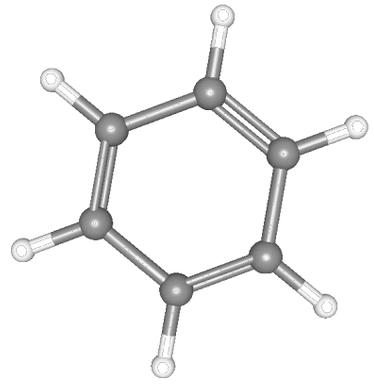
Currently supported output types

alc -- Alchemy file
bs -- Ball & Stick file
caccrt -- Cacao Cartesian file
cacint -- Cacao Internal file
cache -- CAChe MolStruct file
c3d1 -- Chem3D Cartesian 1 file
c3d2 -- Chem3D Cartesian 2 file
ct -- ChemDraw Connection Table file
cht -- Chemtool file
cml -- Chemical Markup Language file
crk2d -- CRK2D: Chemical Resource Kit 2D file
crk3d -- CRK3D: Chemical Resource Kit 3D file
cssr -- CSD CSSR file
box -- Dock 3.5 Box file
dmol -- DMol3 Coordinates file
feat -- Feature file
fh -- Fenske-Hall Z-Matrix file
gamin,inp -- GAMESS Input file
fchk -- Gaussian Cartesian file
gau -- Gaussian Input file
gpr -- Ghemical Project file
gr96a -- GROMOS96 (A) file
gr96n -- GROMOS96 (nm) file
hin -- HyperChem HIN file
jin -- Jaguar Input file
bin -- OpenEye Binary file
mmod,dat,mmd -- MacroModel file
sd,sdf -- MDL Isis SDF file
mdl,mol -- MDL Molfile
mopcrt -- MOPAC Cartesian file
mmads -- MMADS file
bgf -- MSI BGF file
csr -- MSI Quanta CSR file
nw -- NWChem Input file
ent,pdb -- PDB file
pov -- POV-Ray Output file
pqs -- PQS file
report -- Report file
qcin -- Q-Chem Input file
fix,smi -- SMILES file
mol2 -- Sybyl Mol2 file
txyz -- Tinker XYZ file

At last count... Open Babel supports 67 formats with 27+ more requested by users

PLUS: Multiple software versions
Non-standard implementations!

Challenges: Many Representations of Chemical Data



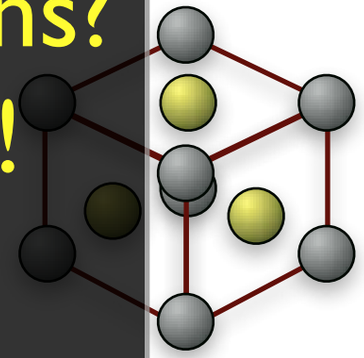
- Molecular Mechanics:
  Atom & bond types,
  No orbitals
- Quantum Mechanics:
  Atoms (no typing),
  No "bonds"
- Crystallography
  Fractional coordinates
- Solid State Codes:
  Unit cells / translation
- Daylight SMILES
  Connectivity only
  No coordinates!

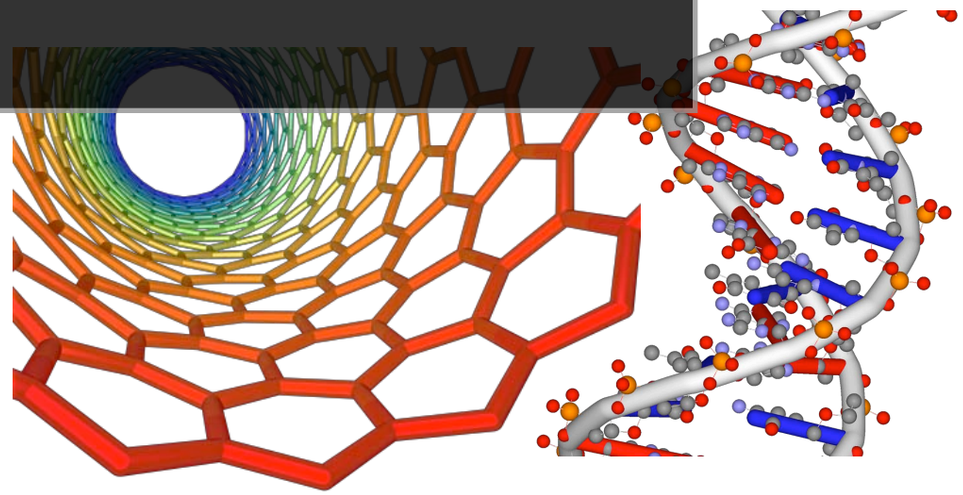
PLUS: Explicit or implicit hydrogens?

Different atom typing rules!

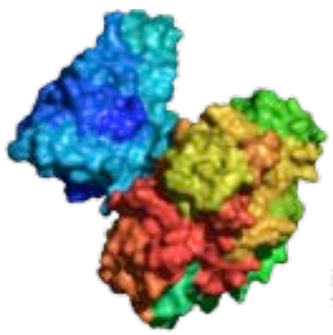
2D vs. 3D

Proteins and biomolecules

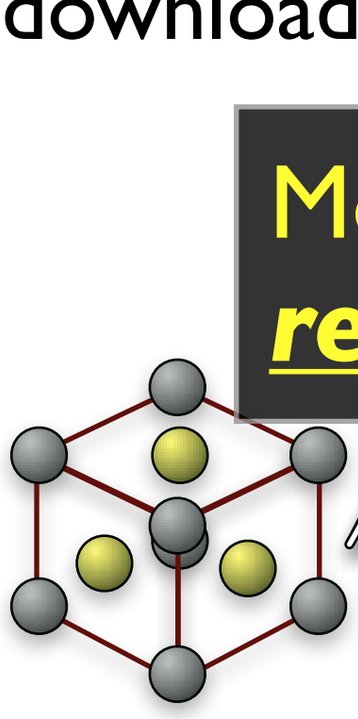




Challenges:
Multiple Programs Needed



download



database

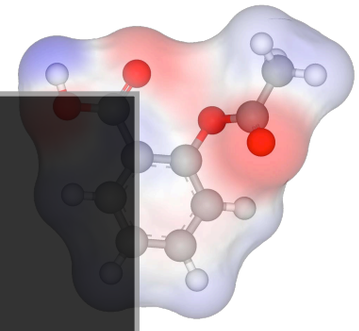
A

B

C

D

calculation



visualization

Modern computational modeling
requires easy interconversion of data…

Open Babel: The Questions...

Answers to the six important questions:

- Why?
- When? How Long?
- Who? Where?
- What? How?

A Brief History of Babel?

- Babel: 1992-1996, Pat Walters & Matt Stahl (U. Arizona)

  With this program we hope to implement a general framework for converting between file formats used for molecular modeling.

  Additional options: center molecule, slice multi-molecule files, add/delete hydrogens

Open Babel 2.0 planned for release in "fall" 2005

- OBabel: Pat Walters
- OELib: ~2000–2001 Matt Stahl, OpenEye
- Open Babel: 2001–

  Open Babel is a project designed to pick up where Babel left off, as a cross-platform program and library designed to interconvert between many file formats used in molecular modeling and computational chemistry and related areas.

Open Babel: The Questions...

Answers to the six important questions:

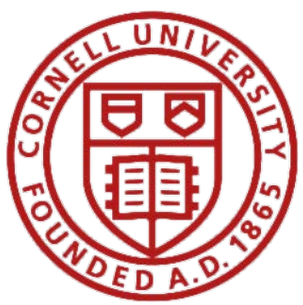
- Why?
- When? How Long?
- Who? Where?
- What? How?

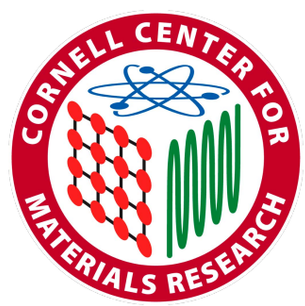
Acknowledgments: A Cast of Many

- Pat Walters
- Matt Stahl
- Roger Sayle
- Anthony Nicholls
- Joe Corkery

Open Eye Scientific

- Michael Banck
- Chris Morley
- Peter Murray-Rust
- Francesco Bresciani
- Jean Bréfort
- Alex Clark
- Nick England
- Vincent Favre-Nicolin
- Fabien Fontaine
- Malcolm Gillies
- Richard Gillilan
- Brian Goldman
- Tommi Hassinen
- Bryan Herger
- Stefan Kebekus
- Erik Kruus
- Eugen Leitl
- David Mathog
- Sergei Pachovsky
- Steffen Reith
- Louis Richard
- Ajay Shah
- Chris Swain
- Bob Tolbert
- Egon Willighagen
- Pawel Wolinski
- Jörg Wegner





Open Source for Open Science

Open source promotes software ***reliability*** *and quality by supporting* ***independent peer review*** *and rapid evolution of source code. To be OSI certified, the software must be distributed under a license that* ***guarantees*** *the right to* ***read***, ***redistribute***, ***modify***, *and* ***use*** *the software* ***freely***.

— Open Source Definition (by Open Source Initiative)

Keys:

- Access to source code
- Built-in peer review (like science!)
- Flexibility — users can modify freely
- Broad community of developers
- *Standardizing* — promotes software reuse

Additional Benefits of Open Source

- Code reuse: stop reinventing the wheel!
  No need to write code for import/export
- Public verification and testing:
  Both algorithm and implementation in code
  Open peer review

No restrictions on **use**
Only "restrictions" on **distribution**
(Most scientists don't distribute code)

- **User flexibility:**
  Open file formats ⇒ no vendor "lock-in"
  Easily use multiple programs
- **Access to source code:**
  Anyone can customize, fix bugs, add features, port to new architectures...

Open Babel: The Questions...

Answers to the six important questions:

- Why?
- When? How Long?
- Who? Where?
- What? How?

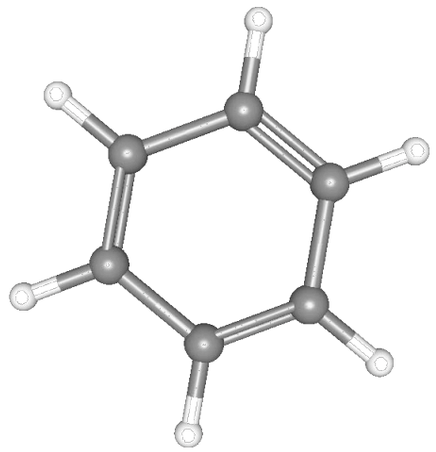
Current Features in Summary

- Huge variety of chemical file formats
  with thorough testing and bug-fixing!
- Daylight SMARTS pattern matching
- Connectivity & bond order perception
- Chirality perception
- Gasteiger partial charge calculation
- Hydrogen addition/deletion
- Isotopes & common chemical data
- Fractional coordinates & unit cells
- Batch conversion, merging, slicing, etc.
- Cross-platform: Windows, UNIX, Mac...
- *More to come…*

Keys to implementation:
"Lazy perception" of data
Flexible representation of properties

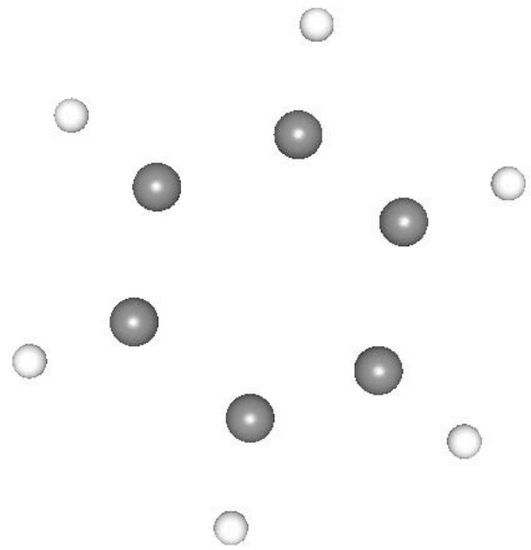
Lazy Perception in Action...

Sybyl Mol2



OBMol

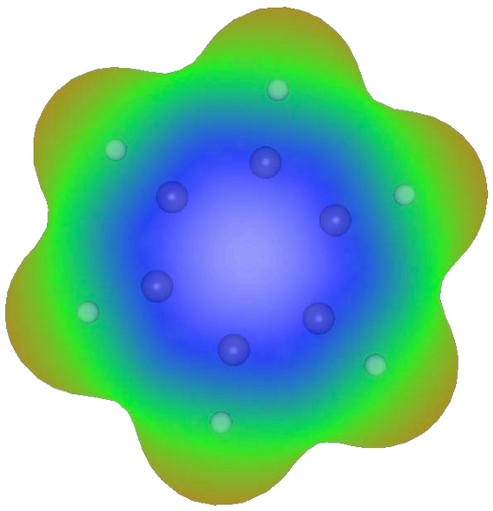
XYZ



- XYZ format doesn't require partial charges
  Why compute them?
- No residue information, no chains, no bonds...
  No atom type translation needed!
- Fast output

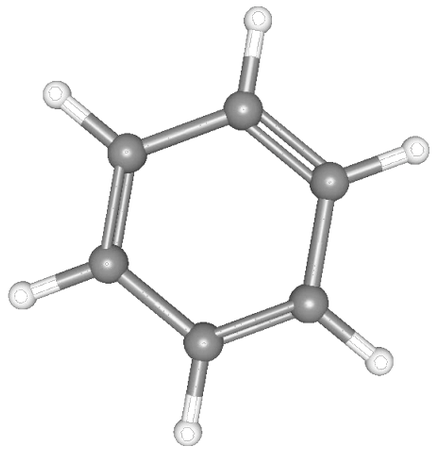
Lazy Perception in Action...

Gaussian 98/03 Output



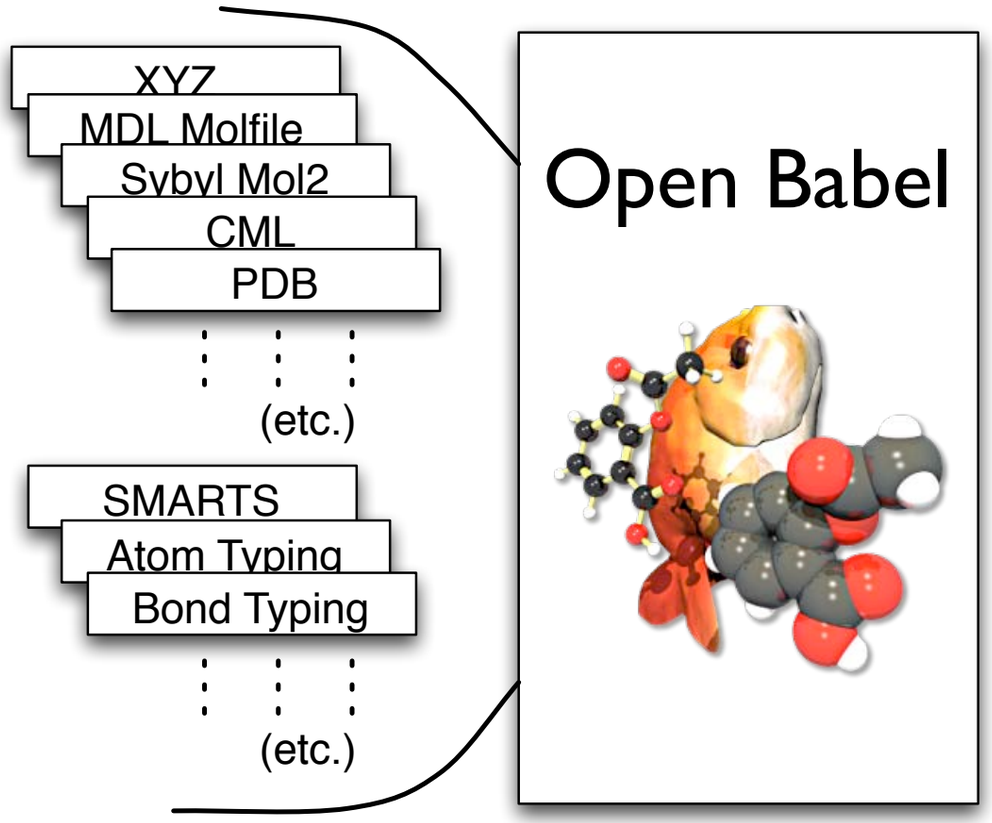
OBMol

Sybyl Mol2



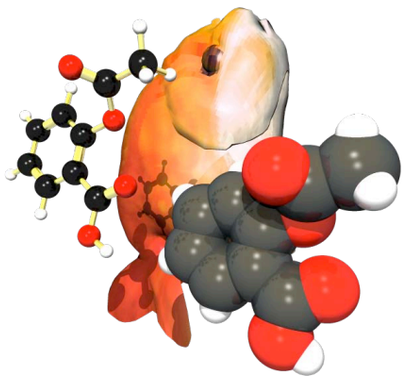
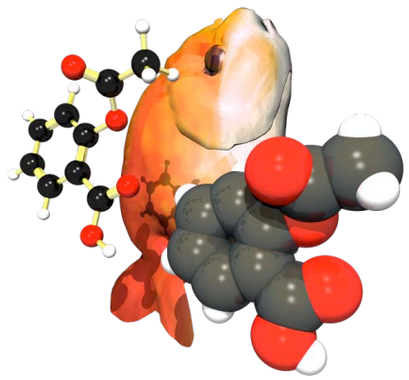
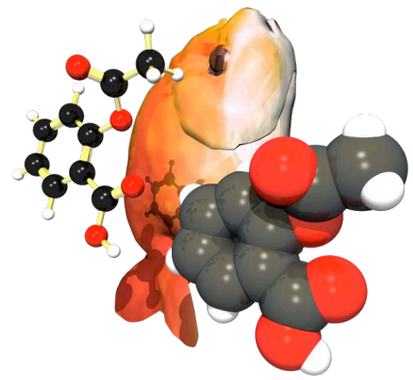
- **Connectivity assignment**
- **Bond perception needed:**
  double bonds, functional groups, aromaticity
- **Atom typing & partial charges assigned**

Solving the Chemical Representation "Problem"



- **Whole is greater than the sum of all parts:**
  No one person handles all file formats
- **Key goal reflected in "lazy evaluation"**
  Leave no data behind, but "perceive" as little as possible — conversion should not create data!

Solving the Chemical Representation "Problem"

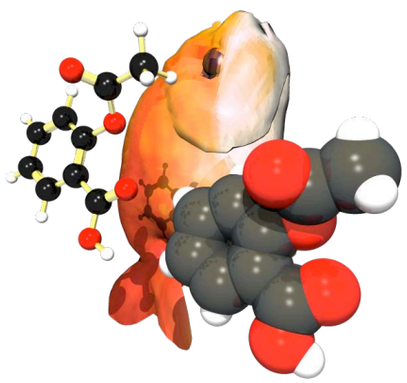
Editor / Viewer		Analysis	Database
Open Babel 	OpenGL Graphics	Open Babel 	Open Babel 

- **Code reuse through open source code:**
  *Focus on problems **beyond** the basics*
  New science, not new software development
- **Rapid development**
- **Reduce non-standard file formats & bugs**

Code-Reuse Example: *obgrep* Match Molecular Patterns

Database

Open Babel



- Total 216 lines of C++ code:
  Includes blank lines & comments!
- Contributed code, not originally part of Open Babel library
- Matches SMARTS molecular patterns in database file(s)
- Import/Export handled by Open Babel
  "Database" can be any file format, any computer, any drive

Example Workflows in Nanoscience

- **Custom Monte Carlo program (385 lines)**
  Read Gaussian output, calculations, write XYZ
- **Batch Conversion**
  MM optimization ⇒ DFT ⇒ INDO excitations
- **Crystal Structure**
  Fractional coordinates, convert & add hydrogens
  Conversion of unit cell parameters to vectors
- **Editing & Visualization**
  Editor ⇒ DFT ⇒ View orbitals, vibrations, etc.

XML Formats in Chemistry: CCML

- **Based on Chemistry Markup Language (CML)**
  Extensions for computational modeling input, output
- **Input and Running Programs**
  XML stylesheets format native input
  ⇒ *Calculation Program*
- **Output from programs ⇒ XML/CCML**

  Either use of "XML output" option in program
  or use of Open Babel conversion to CCML file
  Include auxilliary binary files (wavefunc, density...)

New Directions and Future Plans

- **Improve "lazy evaluation"**
  QM ⇒ QM requires no atom or bond typing!
- **Coordinate refinement for SMILES**
  User-request for 2D or 3D structure layout
- **Access to other languages**
  Perl, Python, Java access to code library
- **Support for more chemical data**
  Symmetry, molecular orbitals, charge density, surfaces, calculation results...
- **Support for even *more* file formats**
  Leave no orphaned data! More nano & materials

Links and Other Related Projects

- **Open Babel**
  http://openbabel.sourceforge.net/
- **Ghemical**
  http://www.bioinformatics.org/ghemical/
- **Chemistry Markup Language**
  http://cml.sourceforge.net/
- **Open Source Initiative**
  http://opensource.org/
- **Open Science Project**
  http://openscience.org/
- **Blue Obelisk Movement**
  http://blueobelisk.org/

Open Babel: The Questions...

Answers to the six important questions:

- Why?
- When? How Long?
- Who? Where?
- What? How

- Any more questions?