“With the End in Mind” – Compound Auto Verification

Identify those compounds that truly need to be looked at without unduly burdening chemists with faulty information, thus saving time.
System Philosophy

- Designed for customer acceptance
  - Minimal additional user effort to produce result

- False Positive Tolerant
  - Allow time to improve and train DB
  - Can be tuned as system improves to increase accuracy …
    become more stringent … and more correct

- Identify those compounds that truly need to be looked at
  without unduly burdening chemists with faulty information,
  thus saving time
- Allow for a feedback loop to generate good assignment DB
- Build a Spectral DB as a “fringe benefit”
Success Criteria

- **Installation Performance**
  - Re-installation not needed
  - Site visit for installation
  - Operational within 2 weeks of visit
  - Ability to train and monitor DB remotely

- **System Performance**
  - Reboot $< 1/1000$ MTBF
  - Can’t read 2D data or FID $< 1/100$ MTBF

- **Verification Performance**
  - For compounds and data that are:
    - Pharmaceutical like (rule of 5)
    - Not salts, rotomers or tautomers
    - Data is collected with sufficient S/N
    - Distinguishable structure *
  - Training $<=1$ per project upon identified/confirmed false negative

  - Reference Compounds
    - False positives $< 20\%$
    - False negatives $< 8\%$
    - Ambiguous Results $< 24\%$

  - Proprietary Compounds
    - False Positive + Negatives $< 17\%$
    - False Negatives $< 6\%$
    - Ambiguous Results $< 24\%$

* Sum of the shift 13C differences $> 4.5$ ppm from false structure
The Implementation Plan

- Provide “Combined” 1D-Proton, 2D-HSQC NMR Auto verification of Identity
  - Greater precision from 13C – NMR data
  - Benefit from higher resolution Proton info with use of
    - Multiplicity
    - Integrals
    - Knowledge of exchangeables

- Implement system on a “voluntary” basis to
  - Fine tune system
  - Gain “customer” trust
  - Demonstrate payback
  - Grow Database of structure linked spectra rapidly
  - Begin to achieve a cycle of prediction improvement with the benefit of easily added spectra to structure linkage.
1. Run Open Access
2. Sweep Data
3. Initiate Registration
4. Select My Data
5. Register
6. Generate MetaData
7. Queue Verification
8. Verify Data
9. Write Results
10. Read/Link Results
11. Email
12. Supervisor Review

Just add the HSQC from the Drop Down
1. Run Open Access – HPLC/LC–MS

1. Run Open Access
2. Sweep Data
3. Initiate Registration
4. Select My Data
5. Register
6. Generate MetaData
7. Queue Verification
8. Verify Data
9. Write Results
10. Read/Link Results
11. Email
12. Supervisor Review

© 2007 Lexicon Pharmaceuticals, Inc.
2. Sweep Data to LIMS

1. Run Open Access
2. **Sweep Data**
3. Initiate Registration
4. Select My Data
5. Register
6. Generate MetaData
7. Queue Verification
8. Verify Data
9. Write Results
10. Read/Link Results
11. Email
12. Supervisor Review
3–5. Register the Compound

1. Run Open Access
2. Sweep Data
3. Initiate Registration
4. Select My Data
5. Register
6. Generate MetaData
7. Queue Verification
8. Verify Data
9. Write Results
10. Read/Link Results
11. Email
12. Supervisor Review

Chemist must select the Analytical Data
6–7. Trigger Meta Data Generation

1. Run Open Access
2. Sweep Data
3. Initiate Registration
4. Select My Data
5. Register
6. **Generate MetaData**
7. Queue Verification
8. Verify Data
9. Write Results
10. Read/Link Results
11. Email
12. Supervisor Review

```
<Verification Meta data file>
<Instrument>Av400
<Structure File Name>2809108.mol
<Verification Type Requested>Combined Verification
<26 Spectrum Directory name>:\instdata\Av400\2007\chemist\Apr12-2007\26\data\1ð
<26 Spectrum Directory name>:\instdata\Av400\2007\chemist\Apr12-2007\26\data\1º
```
8–9. Auto – Verification !!!

1. Run Open Access
2. Sweep Data
3. Initiate Registration
4. Select My Data
5. Register
6. Generate MetaData
7. Queue Verification
8. Verify Data
9. Write Results
10. Read/Link Results
11. Email
12. Supervisor Review

ACD / Automation Server

The “Black Box”

- C + H NMR Predictor
- 1D and 2D NMR Manager Processing and Databasing
- 2D NMR Expert

1 Advanced Chemistry Development, Inc., Toronto, ON, Canada

© 2007 Lexicon Pharmaceuticals, Inc.
12. Supervisor Review (Verification)

1. Run Open Access
2. Sweep Data
3. Initiate Registration
4. Select My Data
5. Register
6. GenerateMetaData
7. Queue Verification
8. Verify Data
9. Write Results
10. Read/Link Results
11. Email
12. Supervisor Review
Initial Trial Data Sets

• Meaning
  – False Positive vs True Positive
  – False Negative vs True Negative
  – Valid vs Invalid Data
    low S/N
    rotomers / tautomers
    referencing problems

• Size of trial Sets
  – 13 Aldrich Compounds
  – 32 Proprietary Compounds, Fully characterized by Analytical
  – 30 ACD/Labs provided compounds (100mg each)
  – 10 building blocks
  – 150 + compounds and growing … weekly/daily
Each analysis score is composed of two parts. Conformance and confidence. Conformance identifies whether the results of the analysis are consistent with the proposed structure. Confidence is an indication of instrument performance and the meaningfulness of the result.
Prediction of Correct vs. Incorrect Structures

No overlap with HOSE code prediction DB entries

- Shift of bond
- Regio–isomers
- Bond order/state maintained

© 2007 Lexicon Pharmaceuticals, Inc.
Poorly differentiated structures

- 2 false positives
- 4 false cautions
- 0 false negatives

Formulas and weights do not match in Aldrich catalog.

- Aldrich 515523
- Aldrich 658030
- Aldrich 406368
- Aldrich 659029
- Aldrich 551023
- Aldrich 417777

Drawn wrong in ACD-FIND.

Tautomer likely drawn wrong in Aldrich catalog.
Auto Verification scores of Aldrich Reference Compounds

- Correct Structures
  - Run HSQC
  - Auto verify
  - Plot Scores

- Incorrect Structures
  - Intentionally similar
  - Generally regioisomers
  - Auto verify
  - Plot scores

False Positives 15%
False Negatives 0%
Ambiguous
- Correct 23% (15%)
- Incorrect 39%

Tautomer = 81
Aldrich 515523

© 2007 Lexicon Pharmaceuticals, Inc.
Scoring threshold selection for verification

- Based on observation of trends from benchmark compounds
  - 13 Aldrich Compounds
  - 31 Proprietary Compounds
  - Prerequisites
    - LC
      - 95% purity by UV
    - LCMS
      - Presence of expected m/z or adduct(s) of at least 20% in SIC for the major UV peak

- NMR
  - >= 70 Consistent
  - >= 50 Possible/Examine
  - >= 35 Ambiguous (or Error condition)
  - 35 – 0 Failure

Dual Scoring System

© 2007 Lexicon Pharmaceuticals, Inc.
Reference Compound Auto Verification Results (N = 50)

30 ASDI and 20 Aldrich Compounds

Good

Bad

Trained DB

Correct Structure

Incorrect Structure

Result/Answer %

Right Answer

Ambiguous

Wrong Answer

False Negative

False Positive

© 2007 Lexicon Pharmaceuticals, Inc.
Untrained DB

Particularly challenging set since … these are the compounds chemist submit for structure determination work

Result/Answer %

<table>
<thead>
<tr>
<th></th>
<th>Correct Structure</th>
<th>Incorrect Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right Answer</td>
<td>63% (12)</td>
<td></td>
</tr>
<tr>
<td>Ambiguous</td>
<td></td>
<td>37% (7)</td>
</tr>
<tr>
<td>Wrong Answer</td>
<td></td>
<td>False Negative</td>
</tr>
<tr>
<td></td>
<td></td>
<td>False Positive</td>
</tr>
</tbody>
</table>

© 2007 Lexicon Pharmaceuticals, Inc.
Correct Score vs. Incorrect Score for 50 Reference Compounds

30 ASDI + 20 Aldrich

Correct Structure Score

Incorrect Structure Score

Indistinguishable Area

False Negative Area

Good

Bad
Induced Workflow
Result

Structure identified as incorrect by NMR auto verification software

False Negative?  
Definition:  
**False Negative:** A correct structure for the compound that has been reported to be incorrect

Analyst

.bootstrap 0cm 0cm 0 0

Diversity

Prediction DB Training

Bug

Invalid Data

Reason?

Yes

Chemist

No

Diversity

DB Training

Registration Verification

Supervisor

Reject

Approve

Start

End

LexChem

LexChem

ID and Fix Structure

Analyst

G o v e r n m e n t

Analytical Chemistry
Project Fringe

Benefits
Spectral DB grown as a fringe benefit
Assignment Starting Point ... automatically derived!

- With Proposed Assignments
2D assignments automated as well

- Peaks
- Volumes
- Assignments
How we got here

Early vision from 2004 ... one step at a time

• Phase I – Instrument Integration (Desktop Access Project) 2004–2005

• Phase II – LexAIMS / LexChem Integration (PDF Project) 2006
  Open format data viewing (PDF)
  Registration PDF availability

• Phase III – LC and HPLC 2006
  Data Interrogation (Parsing) and Reporting

• Phase IV – NMR Auto Verification Project 2007

• Phase V – Database Training with Assigned Structures Now

• Phase VI – Database Seeding and Harvesting 2008 ??
Instrument Integration

- Allocated Storage Space
- Sweep Data

- Storage Area Network (SAN)
  - 6 x 238GB disk drives
  - 0.7 terabytes storing all HPLC, LCMS and NMR data since 2004

- 2 x 1.4 GHz Xeon CPU’s
  - 1 GB RAM

- 2 Analytical Servers
  - Scalable Virtual Machines
    - Data Storage
    - Auto verification

Phase I

Verification System Stats
Peak memory use
450 MB

Peak CPU usage
very often 100%

4 GB Hard Drive

© 2007 Lexicon Pharmaceuticals, Inc.
Storage Paradigm

Phase I

Mass Storage Media Costs (Log Plot)

First Hard Drive:
IBM 1956 5MB
$10,000
($2 Million / GB)

Centralized Instrument Management

- Centralization of Instrument Management
- Standardization of Instrument Platforms (within needs)
- Harmonization of training and interfaces

Phase I

Efficiency
Cost, Operation, Use
Scale of economy
The Push for PDF ... provide access to data

Phase II

When no PDF available →

Spreadsheet View
No PDF Available
View Spectrum (Re-process)
Electronic Access Requires Standardization  Phase II

- Voluntary Notebook Convention conformance
  - Properly formed NB# in spectra (%)
    - 2003 0%
    - 2004, Jan 35%
    - 2004, Jun 71%
    - 2005 94%
    - 2006 98%
    - 2007 99%

- Voluntary Linking of Spectra to registered Compounds
  - Minimization of residual reasons for manual entry key
  - Gentle persuasion ... but with payback..
    - Mid 2006 30%
    - Nov 2006 74%
    - Jan 2007 85%
    - Mar 2007 94%
    - May 2007 95%

- True Indicator of successfully engineering program ...
  user use / acceptance
HPLC and LCMS Verification

- Conceptually simple ... single data point info
  - Can be done before structure to Spectrum Link

- Read HPLC Chromatogram
  - Extract Purity from the Shimadzu “.ARE” file with integration

- Read LCMS data
  - Extract M/Z adduct info
  - Identify presence or absence of compound parent ion/adducts
NMR Auto-verification System

- Act as “Sponsor” of the system
  - Explain pay back for the system
  - Limit impact on workflow

- Run blind tests to validate potential
  - (Benchmark compounds and Proprietary)

- Establish success criteria

- Negotiate Implementation

- One Virtual Machine (VM) server
  - Running ACD/Labs’ Automation Server

- Trigger built in Corporate compound registration system

- Only possible after structure to spectrum link
### Performance

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Spec %</td>
<td>Actual %</td>
<td>Spec %</td>
<td>Actual %</td>
<td>Actual %</td>
</tr>
<tr>
<td>False Positives</td>
<td>20</td>
<td>15</td>
<td>20</td>
<td>17</td>
</tr>
<tr>
<td>(Incorrect Structures)</td>
<td>20</td>
<td>15</td>
<td>20</td>
<td>17</td>
</tr>
<tr>
<td>False Negatives</td>
<td>8</td>
<td>5</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>(Correct Structures)</td>
<td>8</td>
<td>5</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>Ambiguous</td>
<td>24</td>
<td>10</td>
<td>21</td>
<td>24</td>
</tr>
<tr>
<td>(Correct Structures)</td>
<td>24</td>
<td>10</td>
<td>21</td>
<td>24</td>
</tr>
<tr>
<td>Invalid</td>
<td>n/a</td>
<td>–</td>
<td>–</td>
<td>n/a</td>
</tr>
<tr>
<td>(poor S/N, phasing, NMR purity etc.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prerequisites</td>
<td>n/a</td>
<td>–</td>
<td>–</td>
<td>n/a</td>
</tr>
<tr>
<td>Absent linked data</td>
<td>n/a</td>
<td>–</td>
<td>–</td>
<td>n/a</td>
</tr>
</tbody>
</table>

* Corrected for 2 invalid compounds
** random set of 50 checked against intentionally incorrectly drawn similar structures

Installation Performance Met
System Performance Met
Verification Performance … close, but

© 2007 Lexicon Pharmaceuticals, Inc.
Identified Shortcomings

- **2D Communication**
  - assignments in 2D not carried over to 1D ... more weight on 1D
  - Exchangeable Protons
    - Not linking back to 2D for full advantage
    - Absence of correlation in 2D should be more fully used to interrogate 1D–NMR
    - Exchangeables not taken into account with multiplicity (fixed in v 11)

- **Dark Regions (solvent and water exclusion)**
  - Overlap with real peaks
  - Water moves ...
  - CDCl₃ ... not handled very well ...

- Occasional assignment of singlets to protons that should be multiplets
Identified Shortcomings (cont.)

- Dealing with trivial organic impurities such as acetone etc.

- Prediction weighted based on solvent / conditions needed
  - Currently possible to use solvent specific training DB
  - Need underlying vendor DB to favor reported solvent

Operator and Instrumental Issues

- Referencing
  - Sensitive to 0.1 ppm errors in 1H
    (silane impurities in MedChem compounds)

- Selection of the actual solvent at run time
  - Chemists select CDCl3 and run CD3OD on occasion
  - This often results in a failed spectrum
Difficult Challenges

• Indistinguishable Structures
  – Requires greater reliance on auxiliary data such as
    • Multiplets …
    • COSY correlations ??

• The real world of MedChem
  – S/N
  – Organic solvent impurities
  – The so what factor
  – Evaluation of salts (HCL, maleate, acetate etc.)
  – Speed !!
Future Plans … Next Steps

- Parameter Optimization … currently defaults used

- COSY

- 1D – 13C … originally began with this in 1996

- Tautomer form analysis (analyze all use best)

- Broadband decoupled Proton spectra

- High Throughput HSQC probes

\(^1\) Andrew J. Pell, Magn. Reson. Chem 2007, 45: 296-316
Conclusion

• Automated compound auto-verification is plausible as part of a multidisciplinary approach as we have demonstrated.

• Many MedChem and Reference (Aldrich, ASDI) compounds ‘met Specs’.

• Structurally more complex MedChem compounds did not ‘meet specs’ but training and ‘script’ improvements are improving acceptance criteria.

• Integrating a system to perform automated compound verification provides value by highlighting compounds for which structural data is complex and subject to interpretation.
Acknowledgements

• All the chemists that have produced compounds and ran HSQC’s

• Management
  – Dave Kimball
  – Alan Main
  – Mike Messinger

• Cheminformatics
  – Jim Robinson
  – Igor Antonyak
  – Roger Luo
  – Prentice Bisbal

• EBS – Information Systems
  – Brian Dworak
  – Brad Lewis
  – Julie Gervolino

• Analytical
  – Leonard Hargiss
  – Vince Caruso
  – Gonzalo Hernandez
  – Lisa Imbrogno
  – Julita Kendys

• ACD/Labs
  – Brent Lefebvre
  – Joe DiMartino
  – David Fung
BACK–UP SLIDES

FOR

Q & A
Step by Step Process

Verification System Schematics

1. Chemist runs open-access NMR on compound for characterization.
2. LexAims sweeps spectra files into Spectral Archive (N: drive) and catalogs file locations.
3. Chemist submits compound data for registration into LexChem.
4. Chemist selects representative analytical data, through LexAIMS Integration with LexChem, to establish linkage for structure to analytical data for Registered MedChem compounds.
5. Compound data is registered in LexChem.
6. Registration creates metadata files (mol file and text file) if HPLC and LCMS are good.
7. Validation system polls for new metadata files and initiates “COMPARATOR” process.
8. Validation system retrieves NMR Spectra files from N: drive based on paths in metadata file (1H and HSQC). Comparison to NMR predictions performed.
9. Validation System writes Result of conformance of data to COMPARATOR.
10. LexAims polls for results files and captures scores into database.
11. Chemist’s Supervisor receives email that compound is ready for verification.
12. Supervisor initiates Registration Verification process. Verification displays scores from LexAims for all analytical results (including NMR if available). Supervisor Verifies, Rejects or leaves the verification Pending. Supervisor’s response is written back to LexChem. If item rejected, email notification sent to Chemist (not shown).
## HPLC and LCMS Thresholds

<table>
<thead>
<tr>
<th>Technique</th>
<th>Compound Score</th>
<th>Instrument/Reliability Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HPLC</strong></td>
<td>Purity (area%)</td>
<td>Signal Intensity (mv)</td>
</tr>
<tr>
<td>100 - 95 = G</td>
<td>250 = G</td>
<td></td>
</tr>
<tr>
<td>90 - 95 = Y</td>
<td>150 To 250 = Y</td>
<td></td>
</tr>
<tr>
<td>&lt; 90 = R</td>
<td>&lt; 150 mV = R</td>
<td></td>
</tr>
<tr>
<td><strong>LCMS</strong></td>
<td>m/z / adducts when &gt; 20% BPI sum the Intensity</td>
<td>Signal Intensity (count)</td>
</tr>
<tr>
<td>&gt;= 70 = G</td>
<td>&gt; 750000 (7.5 e5) = G</td>
<td></td>
</tr>
<tr>
<td>&lt; 70 = Y</td>
<td>&gt; 100000 (1 e5) = Y</td>
<td></td>
</tr>
<tr>
<td>&lt; 50 = R</td>
<td>&lt; 100000 = R</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt; 40000 = E</td>
<td></td>
</tr>
</tbody>
</table>
Training DB – Requirements for Updating

• SOP with a specific procedure for how to update the Training DB.
  • What is allowed to go in …
    – Well behaved normal spectra /molecules
• What is prohibited
  – Rotomers
  – Tautomers
  – Mixture of diastereomers ??
• pH considerations
• Solvent considerations
• Certainty of assignments
  behavior of the molecules … solvents / pH etc.
• Cross checking / approval or review process etc.
• Only assignment of regio or stereo isomers when complimentary pairs exist … unless incontroversible evidence is available (X–RAY … definitive NOESY etc.)
Select Set of 30 ASDI Compounds

With Training DB set of 19 entries

- Not yet evaluated against incorrect structures
- Invalid spectra or structures not yet removed

Missed 19F Coupling and assignment

10% False Negative
HPLC and LCMS Verification

• Conceptually simple … single data point info
  – Can be done before structure to Spectrum Link

• Read Link

• Extract ASCI info from
  – Shimadzu “.ARE” file with integration info
  – Shimadzu setup file to include column, mobile phase info etc.
  – Micromass “.RPT” files for mass m/z info

• Embed VBA scripts in LIMS to return result

• Preparation for Phase IV pre–requisites
  – Check for Good LC purity
  – Check for m/z in MS
Compound Matched Scores between Correct and Incorrect Structures on Reference Set of 43 Compounds

Inversion Point At n = 39 of 43
MW Range Effects observed on reference set of 43 compounds
Verification Score Trend for 43 Reference Compounds

30 ASDI + 13 Aldrich

Correct Structure Score
Incorrect Structure Score

Performed with our training DB set of 19 entries