Methods for tautomer enumeration, -searching and -duplicate filtering

J Szegezdi, Z Mohácsi, T Csizmazia, S Dóránt, Á Papp, G Pirok, S Csepregi, F Csizmadia

May 2010, European UGM
Contents

• Tautomerization plugin
• Methods for generating tautomers
• How tautomers are used in property predictions
• Tautomer duplicate search
• Generic tautomer generation
• Handling of stereochemistry
• Tautomer substructure search
• Custom tautomerization: Standardizer
**What is tautomerization?**

Tautomers are isomers of organic compounds that readily interconvert by a chemical reaction called **tautomerization**. It is common that this reaction results in the formal migration of a hydrogen atom or proton, accompanied by a switch of a single bond and adjacent double bond. The concept of tautomerizations is called **tautomerism**. Tautomerism is a special case of structural isomerism and can play an important role in non-canonical base pairing in DNA and especially RNA molecules.
Tautomerization Plugin
Can generate:

- **Dominant tautomer distribution with estimated ratio:** those likely to exist (e.g. at a given pH).
- **Major tautomer:** the most dominant one
- **All tautomers:** all theoretically possible
- **Generic tautomer:** used for duplicate tautomer searching
- **Canonical tautomer:** canonicalization based on empirical rules
1. Identify H donors and acceptors.
2. Filter them depending on parameters and method.
3. Consider bond paths between donors/acceptors.
4. Process results of 2 & 3:
   - Combinatorially enumerate,
   - Rank results (dominant & canonical),
   - Calculate distribution (dominant),
   - Filter & select (canonical)
   - Combine paths into regions (generic)
Enhancing property predictions

ChemAxon property predictions have option to consider tautomers:

• Log D: whole dominant tautomer distribution is used, weighted by the ratio of isomers.

• pKa, major microspecies, logP: the single major tautomer is used for the calculation.
Tautomer searching methods in JChem
Table/index option

• „Duplicate search uses tautomers” – check box at table / index creation.
• It makes all duplicate searches „tautomer search” on that table
• Algorithm:
  Convert query to generic tautomer
  Tautomer hash code screening
  Graph matching of generic tautomers
Generic tautomer calculation is accessible in Marvin for understanding/checking behavior:
Generic tautomer

Example:

\[
\text{Generic tautomer}
\]

May 2010, European UGM
Generic tautomer – algorithm

H donor and acceptor atoms are identified

Excluded parts are located. (e.g. Protected stereo, aromaticity, etc.)

Tautomer regions are identified based on alternating single/double bonds, etc. between H donor and acceptor atoms.

Assignment of tautomer regions
1. Variable bonds are replaced by any bonds.
2. Bonding electron count, number of D and T in the region is attached to the region as data. (Needed for graph matching.)
Stereochemistry

Theory:

- When spontaneous tautomerization happens in nature, the specific stereo configuration is lost easily in two steps of back and forth tautomerization. (E.g. racemization.)

- For this reason, in general we ignore stereochemistry in tautomer regions, but there are options to protect different types of stereochemistry.
Stereochemistry

Current practices in JChem (considering structure registration):

• Tetrahedral stereo centers (which are marked by wedge bonds) are excluded from tautomerization (= they are never included in tautomer regions).
  Explanation: wedge stereo drawing is usually not accidental. This way the chemists express that a separate stereoisomer can be isolated.

• Double bond(E/Z) stereochemistry is allowed to tautomerize.
  Explanation: It is very easy to „accidentally” define E/Z stereochemistry. – Just by the atom coordinates. This way it is not easy to decide what the original intention was.
Examples:

\[
\begin{align*}
\text{H}_3\text{C} & \text{CH}_3 \quad \text{H}_3\text{C} & \text{\text{\text{E}}} \text{CH}_3 \quad \text{H}_3\text{C} & \equiv \text{CH}_3 \\
\text{CH}_3 \quad \text{CH}_3 & \equiv \text{CH}_3
\end{align*}
\]

\[
\begin{align*}
\text{H}_3\text{C} & \text{CH}_3 \quad \text{H}_3\text{C} & \text{\text{\text{E}}} \text{CH}_3 \quad \text{H}_3\text{C} & \equiv \text{CH}_3 \\
\text{CH}_3 & \text{CH}_3 \equiv \text{CH}_3
\end{align*}
\]
Example 2: As a result, each row below will be recognized as different compounds. Molecules in the same row are recognized as tautomers of each other:

- (“protect double bond stereo” option is off)
- (“protect all tetrahedral stereo centers” option is on)
- (“protect all tetrahedral stereo centers” option is on)
Tautomer search option

Query molecule is enumerated

Query:

non-tautomeric search:

tautomer search:
Standardization methods

- Canonical tautomer
  - Custom transforms – examples:
Standardization method

- Custom transform example for ring-chain tautomerism:

## Comparison of the methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Available search types</th>
<th>Registration speed</th>
<th>Tautomer search speed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>tautomer</td>
<td>non-tautomer</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Duplicate</td>
<td>Substructure</td>
<td>Duplicate</td>
</tr>
<tr>
<td>Duplicate tautomer table option</td>
<td>OK</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tautomer search option</td>
<td>-</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>Canonical tautomer in standardization</td>
<td>OK</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Custom standardization transforms</td>
<td>OK</td>
<td>OK</td>
<td>-</td>
</tr>
</tbody>
</table>
Summary

- Tautomer calculation plugin offers various methods
- Tautomers can improve property predictions
- JChem offers 4 ways for handling tautomers
Acknowledgements

• József Szegezdi – development of the tautomerization algorithms

• Zsolt Mohácsi – plugin framework, standardizer

• Szilárd Doránt, Tamás Csizmazia, Ali Baharev – tautomer search

• Ákos Papp, György Pirok, Ferenc Csizmadia, Szabolcs Csepregi – concepts
References

