Use of the Cambridge Structural Database in Study of Single and Partial Double C-X (X=C,N,O) Bonds in Organic Molecules in Crystalline State

Rudolf Kiralj and Márcia M. C. Ferreira

Laboratório de Quimiometrica Teórica e Aplicada, Instituto de Química, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, 13084-971, Brazil

E-mails: marcia@iqm.unicamp.br, rudolf@iqm.unicamp.br,
URL: http://lqta.iqm.unicamp.br
Contents

1) Introduction to the CSD – general comments

2) The CSD use in Brazil – some statistical parameters

3) The CSD in study of bond lengths: our study

4) Data mining strategy

5) General C-C bond

6) Univariate Structure Correlation for Bond Lengths in π-Systems: PB-PAHs, aza-PAHs, diaza-PAHS, polyaza-PAHs, picrates, nucleobases

7) Multivariate Approach to Bond Length Prediction in π-Systems: PB-PAHs, nucleobases

8) Future Perspectives: CSD Use With Other Methods
Introduction to the CSD – general comments

What is the Cambridge Structural Database (CSD)?
Cambridge Crystallographic Data Centre (CCDC), University of Cambridge, UK
http://www.ccdc.cam.ac.uk/

-a database containing structural information (atomic parameters for a crystal of known cell dimensions and space group)

-includes detailed information from X-ray, neutron and synchrotron diffraction studies

-covers organic compounds, organo-metallic compounds, and metal-organic coplexes

-does not include: proteins, high polymers, inorganic compounds, purely inorganic carbon compounds (carbides, carbonates, carbonyls and cyanates)

How many crystal structures does it contain?

-established in 1965: a few thousand structures
-current annual increase is > 20 000 structure
-the last version (November 2002) has 272066 structures
The number of chemical compounds in the CAS (Chemical Abstract Service) Registry Database, >50% of which are peptides and proteins, is growing superexponentially. Exponential trend is observed for the CSD (Figure 1).

Figure 1. Left: Cumulative growth of the CAS-Registry Database. http://www.cas.org/casdb.html#regdb
Contents on 14-March-2003: 47 120 998 chem. compounds.
http://www.ccdc.cam.ac.uk/annrep2001/report.html
How and how much scientists use the CSD?

3 basic modes: 1D, 2D, 3D information (Figure 2)

Figure 2. Types of information in the CSD.
Source: the CSD home page.
Use of the CSD (database & software)

Direct

- Level I
  - check for compounds or structures
  - references

Semi-direct

- Level II
  - retrieval of some structures
  - simple data mining
  - qualitative structure correlation

- Level III
  - systematic retrieval of structures
  - use of other CSD modules
  - simple statistics or 3D search
  - structure-based drug design of some compounds

- Level IV
  - systematic retrieval of many structures
  - use of advanced CSD modules
  - advanced statistics on 3D search, chemometrics
  - QSAR, molecular modeling, drug design of statistically big set of compounds

Indirect

- papers, tables, results from CCDC
The CSD use in Brazil – some statistical parameters

What are the CSD products free for Brazilian scientist?

http://www.ccdc.cam.ac.uk/prods/

CCDC Products

- **CSD**
  - Cambridge Structural Database
- **ConQuest**
  - New Interface to the CSD
- **QUEST**
  - Search and Retrieval Program for the CSD
- **VISTA**
  - Statistical Analysis of Geometric and Other Data
- **PreQuest**
  - Creation of In-house Databases
- **Mercury**
  - Crystal Structure Visualisation
  - Available for free download
- **RPluto**
  - Graphical Display of Molecular and Crystal Structures
- **DBUse**
  - Database of Publications using the CSD and Other CCDC Products
- **SuperStar**
  - Predicting Protein-Ligand Interactions
- **GOLD**
  - Protein-Ligand Docking
- **Relibase +**
  - Easy Searching of Protein-Ligand Complexes
- **DASH**
  - Structure Solution from Powder Diffraction Data

- \(\square\) -free for Latin America (Affiliation centre: Instituto "Rocasolano" -CSIC, Madrid, Spain)
- \(\square\) -free to download
- \(\square\) -public
- \(\square\) -commercial
How many CSD licenses are in Brazil today?

The number of the licenses follows curvilinear growth (Figure 3). There is a characteristic geographical distribution of the licenses (Figure 4).

![Figure 3. The growth of the CSD potential use in Brazil.](image1)

![Figure 4. Geographical distribution of the 2003 CSD licenses in Brazil.](image2)

- **4 licenses per state:** SP
- **2 licenses per state:** MG
- **1 license per state:** AL, GO, PE, RS, SC, SE
- **Area around the CSD site (300 km radius)**
How much the CSD is used in Brazil?
-no reliable parameters
-probably the CSD is not much used in Brazil
-a tentative for the CSD use statistics for Brazil:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>7</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Brazil</td>
<td>12</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Chile</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Colombia</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cuba</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mexico</td>
<td>11</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Peru</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Uruguay</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Venezuela</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>F. YU Countries</td>
<td>5</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Hungary</td>
<td>3</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Japan</td>
<td>75</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>Germany</td>
<td>59</td>
<td>7</td>
<td>43</td>
</tr>
<tr>
<td>UK</td>
<td>38</td>
<td>14</td>
<td>120</td>
</tr>
<tr>
<td>Russia</td>
<td>53</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>USA</td>
<td>228</td>
<td>5</td>
<td>91</td>
</tr>
</tbody>
</table>
The CSD in study of bond lengths: our study

Definition of the problem:
-the CSD enables data mining for many high-quality crystal structures of many classes of compounds
-questions: What is a chemical, especially covalent bond? In organic molecules? What is a C-C bond, especially partial double, aromatic? What does it depend on? Can it be predicted in various chemical problems?

Typical classes of organic (C-H-N-O) compounds:

-simple organic compounds mainly with single C-C bonds
-planar benzenoid polycyclic aromatic hydrocarbons (PB-PAHs) and their aza-, diaza- and polyaza-derivatives
-picrate-like systems (picrate derivatives, picrates)
-nucleobases (nucleic acid bases)

Search for quantitative relationships (equations):

-quantitative relationships between experimental bond lengths in crystalline state and bond descriptors (the Pauling $\pi$-bond orders, other bond orders, topological descriptors, crystal packing parameters)

Aza-PAHs:

Diaza- and polyaza-PAHs:
Nucleobases 2003:
Simple organic compounds, carbon allotropes and molecular complexes from the CSD, spectroscopy, proper *ab initio* calculations, and other literature sources: M. M. C. Ferreira, R. Kiralj, in preparation.

Hydrocarbons:
- ethane
- ethylene
- acetylene
- egzohydrogenated (zig-zag) nanotubes (7,0) and (12,0)
- $n$-nonane
- allene
- dodecahedrane
- cyclopentadienyl anion
- cyclobutane
- cyclohexane
- hexabenzocoronene (its central ring)
- benzene
- (PB-PAHs)

Carbon allotropes:
- fullerenes $C_{60}$ and $C_{70}$
- carbyne
- hexagonal and rhombohedral graphite
- cubic diamond
- FCC metallic carbon
- acetylide anion

Complex species:
- hexachlorobenzene
- (pyrene)$_2$, pyrene-TCNQ, pyrene-(DOP)$_2$
- (TCNQ)$_2$, (TCNE)$_2$, (DDQ)$_2$
# Data mining strategy

<table>
<thead>
<tr>
<th></th>
<th>PB-PAHs</th>
<th>Aza-PAHs</th>
<th>DiazaoPAHs</th>
<th>Picrates</th>
<th>Nucleo-bases</th>
<th>Simple organic comp. etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R )</td>
<td>( \leq 7% )</td>
<td>( \leq 7% )</td>
<td>( \leq 7% )</td>
<td>( \leq 7% )</td>
<td>( \leq 6% )</td>
<td>( \leq 7% )</td>
</tr>
<tr>
<td>( \sigma/\AA )</td>
<td>( \leq 0.015 )</td>
<td>( \leq 0.015 )</td>
<td>( \leq 0.015 )</td>
<td>( \leq 0.015 )</td>
<td>( \leq 0.005 )</td>
<td>( \leq 0.005 )</td>
</tr>
<tr>
<td>Year</td>
<td>any</td>
<td>any</td>
<td>any</td>
<td>any</td>
<td>( \leq 1975 )</td>
<td>any</td>
</tr>
<tr>
<td>disorder</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>errors</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>atm type</td>
<td>CH</td>
<td>CHN</td>
<td>CHN</td>
<td>CHNO</td>
<td>CHNO</td>
<td>CHNOX</td>
</tr>
<tr>
<td>org.met.</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>met. com</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>planar</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>any</td>
<td>any</td>
<td>any</td>
</tr>
<tr>
<td>excluded bonds</td>
<td>-</td>
<td>-</td>
<td>N-N</td>
<td>-</td>
<td>N-N</td>
<td>N-N</td>
</tr>
<tr>
<td>retrieved species</td>
<td>PB-PAHs</td>
<td>aza-PAHs</td>
<td>diazaoPAHs</td>
<td>picrates</td>
<td>nucleosides</td>
<td>hydrocarbons C allotr. mol.com.</td>
</tr>
</tbody>
</table>
Figure 5. C-C bond length as a function of the Pauling’s bond multiplicity, from acetylene and acetylide anion to saturated hydrocarbons and diamond, and even more to intermolecular complexes bound by weak C-C bonds. Bond orders of other types give very similar results.
Novoa et al. discovered the longest C-C bond even known: an electron deficient bond in donor-acceptor complexes, with very small bond number, and bond lengths up to the graphite interlayer distance.

Various carbon compounds, allotropes and species are spread over the whole range of the bond number, along the Pauling’s and Brown’s curves, supporting the new findings on the longest C-C bonds.

This discovery shakes the old concept of the C-C bond: “nonbonding interatomic” or “intermolecular interactions” should be considered as intemolecular chemical bonds.

Univariate Structure Correlation for Bond Lengths in $\pi$-Systems: PB-PAHs, aza-PAHs, diaza-PAHS, polyaza-PAHs, picrates, nucleobases

Relationships between structural parameters:
- Structure correlation (SC)
- (Quantitative) Structure-Structure Relationships ((Q)SSR)
- Bond Length-Bond Order Relationships (BLBOR)
- Bond Length-Bond Descriptor Relationships (BLBDR)

Our study: Univariate relationships (linear regression) between bond lengths $D$ and Pauling $\pi$-bond orders $P$ ($P = M - 1$): \[ D/\text{Å} = a + bP \]

$a$ and $b$ exhibit expected similarities and differences:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-C</td>
<td>PB-PAHs</td>
<td>1.468(2)</td>
<td>-0.147(5)</td>
</tr>
<tr>
<td></td>
<td>Aza-PAHs</td>
<td>1.462(6)</td>
<td>-0.143(13)</td>
</tr>
<tr>
<td></td>
<td>Diaza-PAHs</td>
<td>1.458(3)</td>
<td>-0.143(8)</td>
</tr>
<tr>
<td></td>
<td>Polyaza-PAHs</td>
<td>1.421(30)</td>
<td>-0.087(81)</td>
</tr>
<tr>
<td></td>
<td>Picrates</td>
<td>1.497(11)</td>
<td>-0.212(25)</td>
</tr>
<tr>
<td></td>
<td>Nucleobases</td>
<td>1.487(7)</td>
<td>-0.202(16)</td>
</tr>
<tr>
<td>C-N</td>
<td>Aza-PAHs</td>
<td>1.444(10)</td>
<td>-0.184(18)</td>
</tr>
<tr>
<td></td>
<td>Diaza-PAHs</td>
<td>1.415(4)</td>
<td>-0.152(8)</td>
</tr>
<tr>
<td></td>
<td>Polyaza-PAHs</td>
<td>1.431(20)</td>
<td>-0.128(42)</td>
</tr>
<tr>
<td></td>
<td>Nucleobases</td>
<td>1.398(3)</td>
<td>-0.127(7)</td>
</tr>
<tr>
<td>C-O</td>
<td>Picrates</td>
<td>1.326(5)</td>
<td>-0.198(18)</td>
</tr>
<tr>
<td></td>
<td>Nucleobases</td>
<td>1.295(16)</td>
<td>-0.101(26)</td>
</tr>
<tr>
<td>All</td>
<td>Nucleobases</td>
<td>1.429(5)</td>
<td>-0.199(13)</td>
</tr>
</tbody>
</table>
Figure 7. The mean C-C bond length in various π-systems.

Figure 8. The C-C bond length variation in various π-systems.
Multivariate Approach to Bond Length Prediction in $\pi$-Systems: PB-PAHs, nucleobases

Figure 9. Degeneration of the $D - P$ graph for PB-PAHs.

Figure 10. Degeneration of the $D - P_w$ graph for nucleobases.
Bond descriptors to decrease the data degeneration:
- Pauling $\pi$-bond order $P$
- weighted Pauling $\pi$-bond order $P_w$
- Pauling $\pi$-bond orders including crystal packing effects (H-bonds, vdW interactions, etc.) $P_x$
- electrotopological index: the sum of atomic numbers $Q$
- topological indices:
  - the number of neighboring bonds $n$
  - the number of neighboring rings $m$
  - the number of neighboring bonds $l$
  around bonds already counted for $n$

PB-PAHs: $P$, various $P_x$, $n$, $m$, $l$
Nucleobases: $P$, $P_w$, various $P_x$, $n$, $Q$

![Chemical structures](image)

Figure 11. Some examples for (electro)topological indices
Results of the Principal Component Analysis (PCA)

Figure 12. 3D scores plots from PCA show similarity and dissimilarity between PB-PAHs and nucleobases. Roman numerals denote various classes of bonds defined by topological indices: \((nml)\) for PB-PAHs, \((Qn)\) for nucleobases.
Results of the Hierarchical Cluster Analysis (HCA)

Figure 13. HCA dendograms for samples (A, B) and variables (C, D) for PB-PAHs and nucleobases. The groups from the 3D PCA scores plots are marked in the same way, adding $s$ (for short bonds) and $l$ (for long bonds) for subgroups of VII and X (PB-PAHs), and VIII and IX (nucleobases).
Partial Least Squares (PLS) regression showed to be more adequate than linear regression for the study of C-C, C-N and C-O bond lengths in PB-PAHs and nucleobases.

Much better PLS regression models were obtained using more bond variables than only Pauling’s bond order \( P \).

PB-PAHs:

\[
D/\text{Å} = 1.431 - 0.060P - 0.063P_x + 0.006n + 0.004m + 0.001l
\]

\( R = 0.94, \Delta = 0.007 \text{ Å} \) \( \Rightarrow \) univariate: \( R = 0.90, \Delta = 0.010 \text{ Å} \)

PLS: 2 principal components (>96% original data)

Nucleobases:

\[
D/\text{Å} = 2.304 - 0.080P - 0.078P_x - 0.068Q - 0.006n
\]

\( R = 0.93, \Delta = 0.017 \text{ Å} \) \( \Rightarrow \) univariate: \( R = 0.65, \Delta = 0.037 \text{ Å} \)

PLS: 3 principal components (>98% original data)
Future Perspectives: CSD Use With Other Methods

DATA MINING +

CHEMOMETRIC ANALYSIS +

STRUCTURAL & COMPUTATIONAL METHODS =

A VERY POWERFUL MEANS TO STUDY BOND LENGTHS IN ORGANIC CRYSTALS +

INTERESTING AND USEFUL RESULTS ON BASIC CHEMICAL CONCEPTS (WHAT IS A BOND?? C-C BOND??) +

DIRECTIONS FOR FUTURE STUDIES (intrinsic \( \pi \)-system properties, substitution effects, crystal packing effects, (hetero)aromaticity, crystal packing, etc.)

THIS IS VALID NOT ONLY FOR BOND LENGTH BUT FOR OTHER STUDIES WHICH USE THE CSD EXTENSIVELY.