P- 18

化学システムを正しく利用するための連携および統一技術

Coordination and unified technology for proper use of chemical systems

Kohtaro　Yuta

In Silico Data, Ltd. (http://www.insilicodata.com)

Introduction: 動物実験代替法分野でも様々な化学システムやデータベースを利用し、開発することが多くなってきた。化合物を扱うシステムに入力される情報は同一かつ同一基準であることが前提で構築される。例えば化合物については、化合物名、立体／幾何／配座異性、共鳴構造、互変異性体、他に関して様々な表記法が存在する。即ち、一つの化合物が多種多様の表記法により表現され（一元多項対応）、この点が化合物をコンピューター上（デジタル）で扱うことを困難なものとする。システムを効率よく、信頼性高く使うためには化合物に起因する様々な問題に注意することが必要である。また、化学のみならず、システムで適用される様々な基本事項を守ることも求められる。システムを用いれば常に最高の結果が得られるケースは少なく、様々な点に留意して使うことが必要である。

In the field of alternatives to animal experiments, various chemical systems and databases are being used and developed more and more. The information input to the system that handles compounds is constructed on the assumption that they are the same and have the same criteria. For example, for compounds, there are various notations for compound names, conformations / conformations, resonance structures, tautomers, and others. That is, one compound is expressed by a wide variety of notations (one-dimensional multinomial correspondence), which makes it difficult to handle the compound on a computer (digitally). It is necessary to solve various problems caused by such problems.

データ解析実施上での留意点：Focal points on execute data analysis

□データ解析前や実施時での留意事項

サンプル関連での留意点

○サンプル分布：クラスサンプル数に大きな偏りがあってはならない
- ニクラス分類　⇒　サンプル数の多いクラスに判別関数が支配される
- 強化学習　⇒　サンプル数が偏ると、正しい判断ができなくなる

○サンプル数とパラメータ数との関係：ニクラス分類、重回帰
- 解析信頼性　⇒　ニクラス分類　４＜サンプル数/パラメータ数、　重回帰　５～６＜サンプル数/パラメータ数
- クラスサンプル数とパラメータ数　⇒　ニクラス分類　小さなクラスのサンプル数　≧　パラメータ数

パラメータ関連：データ解析に用いられるパラメータに関する留意点

○欠損データの扱い形式：システム単位で異なるので、異なるシステム間でのデータ共有で留意が必要
- データ補完　⇒　別の基準の値を入れる（０，１，平均値、最大/最小値、他）、スパースモデリングの適用、他
- パラメータ消却　⇒　パラメータそのものをデータ解析に用いない

○パラメータの桁数の扱い：化合物関連のパラメータは桁数の違いが極めて大きい
- 桁数の違いはそのままにしてデータ解析を行う
- オートスケーリングにより、パラメータ間の桁数の違いを解消する

□既成の判別関数等を用いて予測等を行う時の留意事項

サンプルデータ関連：判別関数作成時のサンプル分布

○サンプル分布　⇒　予測対象化合物が判別関数作成時に用いた化合物の補償範囲内か外か　⇒　保証範囲内であれば予測精度は高いが、範囲外であれば予測精度は低い

○パラメータ　⇒　欠損パラメータの割合　⇒　少ないと予測精度が高い、多いと予測精度が低い

○三次元パラメータ　⇒　適切な三次元構造から算出されたパラメータであるか（一元一項対応の問題がクリアされているか否か）

一元一項対応の重要性：Importance of canonicalization

□化合物を扱う場合常に留意すべき極めて重要な概念

○一元一項対応と一元多項対応：
- 一元一項対応　⇒　一つの化合物を指定する表記は一つ（同一形式の表記法において）しか無い
- 一元多項対応　⇒　一つの化合物を指定する表記は多数存在する

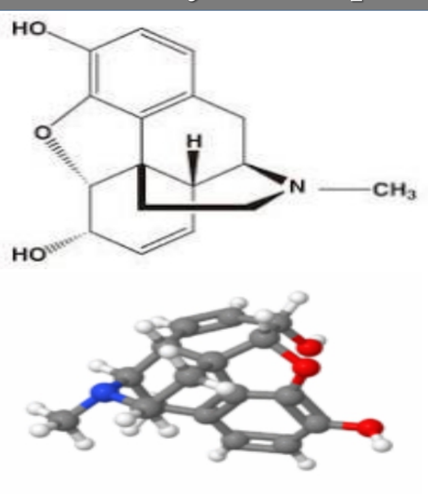
□化合物の一元一項対応とパラメータの一元一項対応

○化合物の場合　⇒　一つの化合物が様々な表記手法と表記形式を有し、化合物が一意に決まらない　⇒　データベースの統一が困難。データベースの串刺し検索が困難。
⇒　化合物のシステムへの入力の違いにより、同一化合物が異なる化合物として扱われる可能性が高い

○パラメータの場合　⇒　一つの化合物を指定する表記が多数存在する

化合物操作上での問題：Problems related to compound manipulation / storage

◇ Diversity of compound notation: No unified information



■Chemical ID Number

CAS number:57-27-2
ATC code:N02AA01 (WHO)
PubChem:CID: 5288826
DrugBank:APRD00215
ChemSpider:4450907
KEGG:D08233

■compound properties

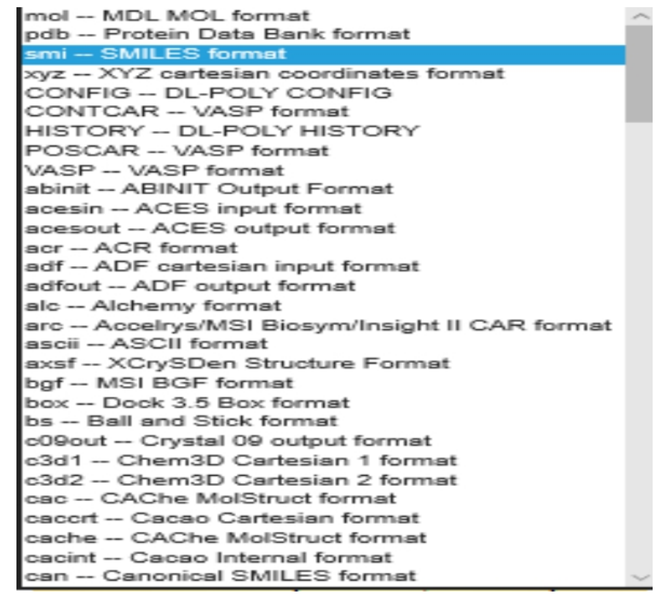
Chemical formula:C17H19NO3

■Reproducibility of chemical compounds: Linear notation of compounds

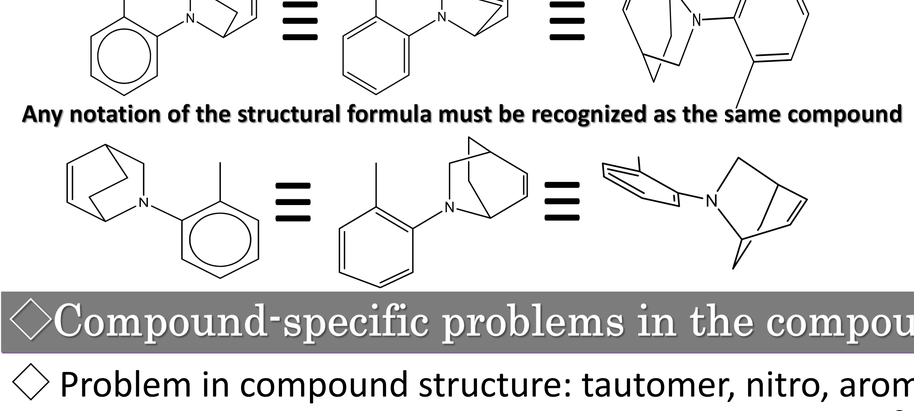
Compound name：Morphine
IUPAC: (5α,6α)-7,8-didehydro-4,5-epoxy-17-methylmorphinan-3,6-diol
SMILES: OC(C=CC1CC2N3C)=C(OC4C(O)C=5)C1C4(CC3)C2C5
InChIKey: InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,13,16,19-20H,6-8H2,1H3/t10-,11+,13-,16-,17-/m0/s1

■Reproducibility of chemical compounds: Notation by connection table

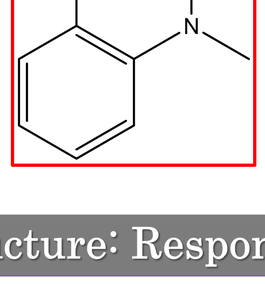
List of file formats handled by the “OpenBabel system”



◇Necessity of canonicalization of compounds: Response to compound diversity



Any notation of the structural formula must be recognized as the same compound



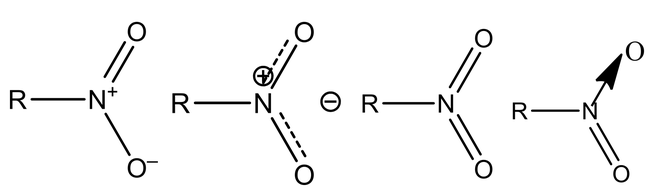
□Canonicalization is required to correctly perform compound searches

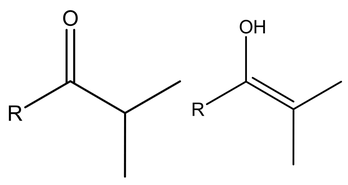
There are many structural patterns in one compound. Compound does not hit in search.

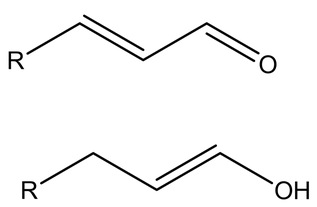
SMILES 1: OC1=C(N(C)C)C=CC=C1 ;by **ChemDraw**
2:c1(O)c(N(C)C)cccc1 ;by **Ecosar**
3:C1=CC(=C(C=C1)N(C)C)O ;by **QSAR Toolbox**
4:CN(C)c1ccccc1O ;by **OpenBabel**
5:C1=CC(O)=C(N(C)C)C=C1 ;Manual Input by Yuta
6:C1(O)=C(N(C)C)C=CC=C1 ;Manual Input by Yuta

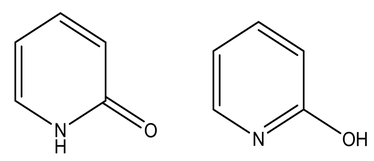
◇Compound-specific problems in the compound structure: Response to compound diversity is required

◇ Problem in compound structure: tautomer, nitro, aromatic









Many others