機械学習 ケモインフォマティクス

樋口 帆乃香

奈良県立奈良高等学校

序論

ケモインフォマティクスとは化学に関連するデータをコンピュータで解析し，様々な課題を解くために用いる方法であり，製薬業界で頻繁に利用されている．例えば化合物の特性と薬剤としての効果の関係を解析することや，自社の化合物データにある化合物を化合物間の類似性に基づいてクラス分けを行ったりすることが可能である．また近年ではこれらのアプリケーションに加え新規化合物のデザインや合成活性化予測等もなされている．

本論

動機：本論文の著者が元々化学分野に関心を持っていたことから，機械学習と化学が融合されたケモインフォマティクスについてELCASを通して詳しく学習することにした．具体的には，化合物の類似性を評価するという点からビタミンと表される化合物群を用いて2種類の試行を行った［4］.

方法：ケモインフォマティクスを実行するためにはコンピュータで化合物の構造・性質などを取り扱う必要があるが，人間とコンピュータでは化合物の認識方法が異なる．認識方法を統一させるための専用のソフトウェアとしてRDKitを使用した［1］．RDKitにおいてMolオブジェクトとして扱われる分子の読み込みは「SMILES」形式から行った．「ChemSpider」［5］というデータベースから化合物の「SMILES」形式を検索し化学構造を文字列で表現した.

本試行に用いたビタミンを描画したものは以下の通りである.

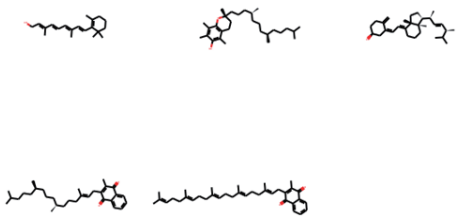

図1（左上段から右に：ビタミンA，レチノール，ビタミンD，エルゴカシフェノール，ビタミンE，トコフェノール，ビタミンK，フィロキノン，メナキノン）

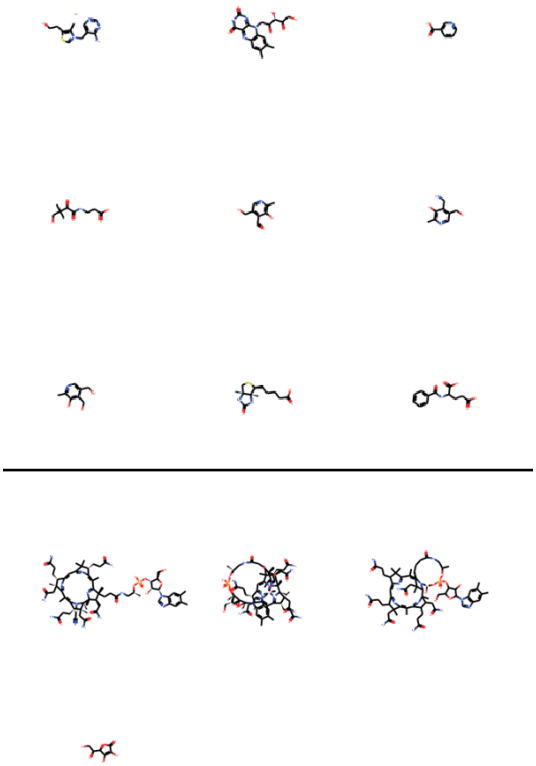

図2（左上段から右に:ビタミンB1，チアミン，ビタミンB2，リボフラミン，ビタミンB3，ナイアシン，ビタミンB5，パントテン酸，ビタミンB6，ピリドキサール，ピリドキサミン，ピリドキシン，ビタミンB7，ビオチン，ビタミンB9，葉酸，ビタミンB12，シアノコバラニン，メチルコバラニン，ヒドロキソコバラニン，ビタミンC，アスコルビン酸）

連絡先：
myamada@i.kyoto-u.ac.jp（山田　誠，京都大学）

［試行1］
同機能ビタミンの類似性を解析する［2］.

方法：Morganフィンガープリントを用い，タニモト係数で類似性を評価した．フィンガープリント作成には色々なルールがあるが，原子からある距離に位置する部分構造を数え上げ（radius=2の設定が主に使われる）ビット配列に0か1を埋め込んだ形式にハッシュ化したものがMorganフィンガープリントである．タニモト係数は以下の式で表され（式1）分子A，分子Bの類似性を0から1の値で計算する．

$$S_{AB} = \frac{c}{a + b - c}$$

式1（aは分子Aのビット配列で1が立っている数，bは分子Bのビット配列で1が立っている数，cは分子AとBで共通に1が立っている数を指す）

結果：

vitamin B6	Pantothenic acid × Pyridoxal
	0.560975609756097

Pyridoxal × Pyridoxine
0.729729729729729

vitamin K	Phylloquinone × Menaquinone
	0.441176470588235

Pantothenic acid × Pyridoxamine
0.560975609756097

vitamin B12	Cyanocobalamin × Methylcobalamin
	0.466076696165191

Cyanocobalamin × Hydroxocobalamin
0.556250000000000

Methylcobalamin × Hydroxocobalamin
0.546875000000000

図3（同機能ビタミン間の類似度）

同機能ビタミン間の類似度はおおよそ45%から70%であることが解析された.

［試行2］
Mol Log P関数を用いて疎水性を評価する.

方法：Log Pを計算によって予測するアルゴリズムMol Log P関数を用いた．Log Pとは分子の性質を決める指標となる記述子の一種であり，疎水性を評価するものである.
　また分配係数と呼ばれ，用いる2相が水と油の場合において特定物質がそれぞれに溶ける割合を表している．したがってLog Pの数が大きいほど疎水性が高いことが分かる.

結果：一般にビタミンA，ビタミンD，ビタミンE，ビタミンKは脂溶性ビタミンと定義されており，またビタミンB1，ビタミンB2，ビタミンB3，ビタミンB5，ビタミンB6，ビタミンB7，ビタミンB9，ビタミンB12，ビタミンCは水溶性ビタミンと定義されている.
　Log Pを解析し順に並び替えたものを以下に示す.

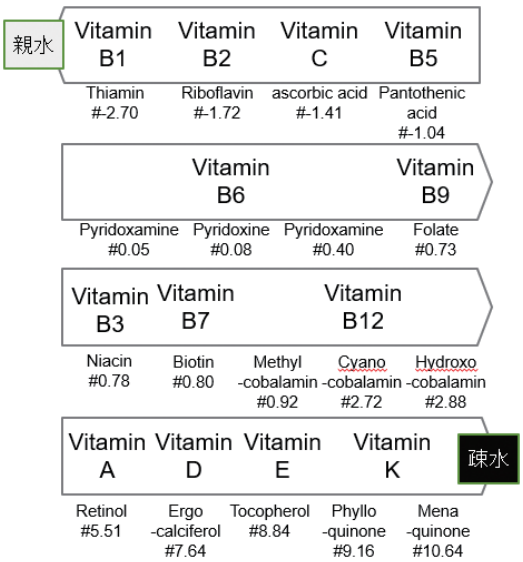

図4（ビタミンのLog P値）

脂溶性ビタミンが油に溶けやすく水溶性ビタミンが水に溶けやすいことが明らかに示された．また同機能ビタミンは同じ程度の疎水性があることが明らかとなった.

結論

　ELCAS実習を通して，ケモインフォマティクスの問題に取り組んだ．具体的には，Morganフィンガープリントを用い，タニモト係数で化合物間の類似性を評価することを行った．その結果，データから機械学習を用いることで，脂溶性ビタミンが油に溶けやすく水溶性ビタミンが水に溶けやすいことが明らかに示された．今後は，膨大なデータを用いた化合物の解析を行うことを目標としたい.

謝辞

本活動においてご指導ご協力してくださった京都大学情報学研究科山田誠准教授, また大学生・大学院生の方々に深く御礼申し上げます.

参考文献

［1］RDFkit, https://www.rdkit.org/
［2］https://github.com/Mishima-syk/py4chemoinformatics
「化合物の類似性を評価してみる」
［3］https://future-chem.com/
「科学の新しいカタチ」
［4］https://ja.wikipedia.org/wiki/ビタミン
「ビタミン」
［5］http://www.chemspider.com/Default.aspx
「ChemSpider」
［6］https://qiita.com/Mochimasa/items/4e34ceb8eb9519513a94
「化合物データの取り扱いのイロハ」