

◇本日の講義スケジュール

序章 13:00～13:20

- ・自己紹介
- ・時代の流れと、研究環境および研究内容や研究スタイルの変化

第一章 13:20～13:40 (10分休憩)

化学データサイエンスおよび人工知能の適用イメージ

第二章 13:50～15:00 (10分休憩)

化学データサイエンスおよび人工知能の適用事例：様々なアプローチが可能

第三章 15:10～15:50 (10分休憩)

化学データサイエンスおよび人工知能の適用手順やパターン

第四章 16:00～16:50

化学データサイエンスおよび人工知能の適用上での留意事項

- ①化学分野の留意点
- ②データサイエンス実施上での留意点

まとめと提案 16:50～17:00

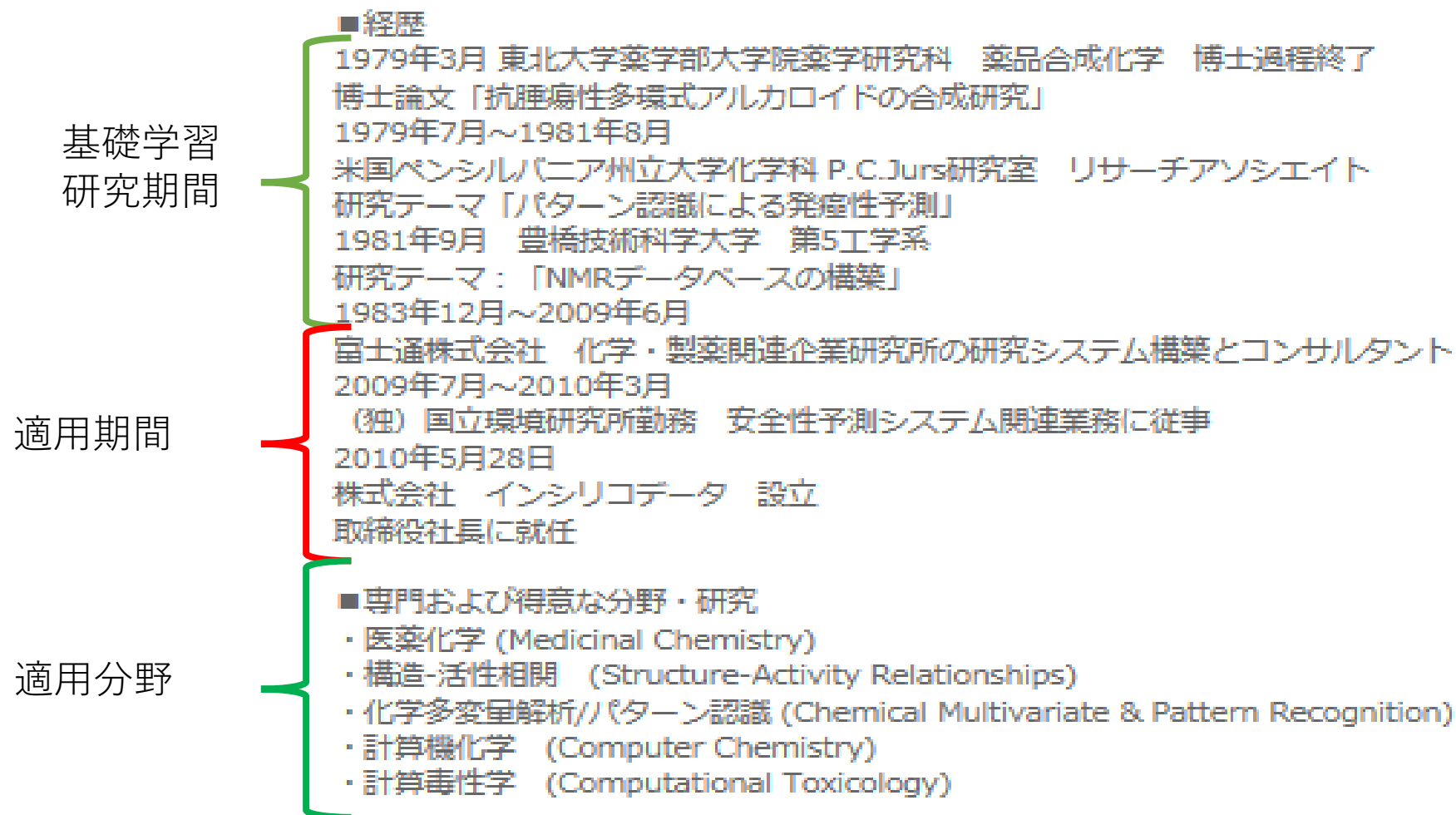
- ①化学データサイエンスおよび人工知能のまとめ
- ②今後の展開についての「オートノマス創薬」の提案
 - ・KY法の展開（クラス分類と重回帰型）
- ③自由討論

序章：

- ・ 自己紹介
- ・ 化学データサイエンスおよび人工知能の時代的な流れと、変化
- ・ 化学データサイエンスおよび人工知能の歴史
(実は古くから研究は始まっていた)

自己紹介

挨拶：株式会社 インシリコデータ 湯田浩太郎



Computer-assisted struc...

Journal of Medicinal Chemistry

J. Med. Chem. ACS Med. Chem. Lett.

Article

Computer-assisted structure-activity studies of chemical carcinogens. Aromatic amines

Kohtaro Yuta, Peter C. Jurs

J. Med. Chem., 1981, 24 (3), pp 241-251
 DOI: 10.1021/jm00135a003
 Publication Date: March 1981

ACS Legacy Archives

Note: In lieu of an abstract, this is the article's first page.

Click to increase image size

J. Med. Chem. 1981, 24, 241-251

Articles

Computer-Assisted Structure-Activity Studies of Chemical Carcinogens. Aromatic Amines

Kohtaro Yuta and Peter C. Jurs*

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802.
 Received August 18, 1980

Studies of molecular structure-carcinogenicity relations for a set of 157 aromatic amines are reported. A computer-assisted approach using pattern-recognition methods was used to develop a series of discriminants for aromatic amino carcinogenic potential. The 157 compounds were divided into subsets according to tumor site, route of administration, and activity. Sets of calculated molecular structure descriptors were generated that could support linear discriminant functions able to separate sets of active carcinogens from inactive compounds. Prominent among the important structural descriptors were those coding sizes and shapes of the amines. The pattern-recognition results were not strongly affected by differences in active site, and the study showed that mixed data sets could be used in computer-assisted structure-carcinogenicity studies.

Retrieve Detailed Record of this Article

Retrieve Substances Indexed for this Article

Retrieve All References Citing this Article

Computer-assisted struc...

Computer-Assisted Structure-Activity Studies of Chemical Carcinogens. Aromatic Amines

Kohtaro Yuta and Peter C. Jurs*

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802.
 Received August 18, 1980

Studies of molecular structure-carcinogenicity relations for a set of 157 aromatic amines are reported. A computer-assisted approach using pattern-recognition methods was used to develop a series of discriminants for aromatic amino carcinogenic potential. The 157 compounds were divided into subsets according to tumor site, route of administration, and activity. Sets of calculated molecular structure descriptors were generated that could support linear discriminant functions able to separate sets of active carcinogens from inactive compounds. Prominent among the important structural descriptors were those coding sizes and shapes of the amines. The pattern-recognition results were not strongly affected by differences in active site, and the study showed that mixed data sets could be used in computer-assisted structure-carcinogenicity studies.

Aromatic amino compounds have been studied and a great deal of data accumulated over the years. The first report on the carcinogenicity of aromatic amines was by Rehn¹ in 1895 who described bladder cancers among employees in a Swiss dye factory. Many studies have followed, in part because aromatic amines are strongly related to industrialization and humans have been exposed to them widely. For example, aniline is employed world-wide in numerous industrial processes, fluorene compounds were originally developed as insecticides, and compounds such as butter yellow have been used as food additives. Primary aromatic amines are commonly used in industrial azo dye syntheses and as antioxidants in rubber products. Thus, the study of the carcinogenic potential of this class of compounds is of widespread interest.

A variety of hypotheses regarding the relationships between structure and activity of aromatic amines have been advanced. These compounds usually produce tumors at a site remote from the site of administration, including the liver, intestine, and bladder. The effects of aromatic amine exposure are species dependent. Activation through metabolic change is an essential step in aromatic amine carcinogenesis. The most prominent theory involves N-hydroxylation of aromatic amines.²⁻⁶ Conversion to N-hydroxy compounds appears to be a requirement for carcinogenic activity. However, not all N-hydroxy compounds are carcinogenic. Although N-hydroxylation seems to be necessary for carcinogenic activity, it is not sufficient to explain the effect. The "para principle" involves the proposition that an amine should have a long, uninterrupted conjugated system with the amino group attached to one of the para carbons.⁷⁻⁹ The presence of such a conjugated system lends stability to the ultimate reacting species. However, these theories do not explain all available biological data, and they are still under development. The carcinogenic activity of aromatic amines and some structure-activity relationships have been presented and discussed in several books and reviews.¹⁰⁻¹³

The present work involves the computer-assisted study of a set of 157 aromatic amines using chemical structure information handling and pattern-recognition methods to attempt to develop structure-activity relationships. The methodology of this approach to SAR studies has been described elsewhere and will be summarized here.

Methodology

The fundamental premises involved in applying pattern-recognition methods to SAR studies of chemicals with genetic toxicity are as follows: (1) Molecular structure and biological activity (genetic toxicity) are related. (2) The structures of compounds having genetic toxicity and compounds of similar structural classes that are nontoxic can be adequately represented by a set of molecular structure descriptors. (3) A relationship can be discovered between the structure and activity by applying statistical and/or pattern-recognition methods to a set of tested compounds. (4) The relation can be extrapolated to untested compounds to provide predictive ability.

The structure-activity studies were done using the ADAPT (automatic data analysis using pattern-recognition techniques) computer software system. This system has been developed over the past few years and has been described previously.¹⁴

The fundamental steps involved in performing an SAR study using the ADAPT system are as follows: (a) Identify, assemble, input, store, and describe a data set of structures for chemicals that have been tested for the biological activity of interest. (b) Develop computer-generated molecular structure descriptors for each of the members of the data set. The descriptors may be derived directly from the stored topological representations of

(1) L. Rehn, *Arch. Klin. Chir.*, **50**, 588 (1895).
 (2) J. L. Radomski and E. Brill, *Science*, **167**, 992 (1970).
 (3) J. L. Radomski and E. Brill, *Arch. Toxicol.*, **28**, 159 (1971).
 (4) G. M. Conzelmann, Jr., A. A. Rey, and E. Brill, *J. Natl. Cancer Inst.*, **50**, 989 (1973).
 (5) F. F. Kadlubar, J. A. Miller, and E. C. Miller, *Cancer Res.*, **37**, 805 (1977).
 (6) E. K. Weisburger, *Annu. Rev. Pharmacol. Toxicol.*, **18**, 395 (1978).
 (7) H. Druckrey, *Arzneim. Forsch.*, **2**, 503 (1952).
 (8) N. P. Buu-Hoi, *Arzneim. Forsch.*, **4**, 531 (1954).
 (9) J. C. Arcos and M. F. Argus, "Chemical Induction of Cancer", Volume 2B, Academic Press, New York, 1974.

(10) J. C. Arcos and M. Arcos, in "Progress in Drug Research", E. Jucker, Ed., Wiley-Interscience, New York, 1962.
 (11) W. C. Hoeper and W. D. Conway, "Chemical Carcinogenesis and Cancers", Charles C. Thomas, Springfield, Ill., 1964.
 (12) D. B. Clayton and R. C. Garner, in "Chemical Carcinogens", C. W. Searle, Ed., American Chemical Society, Washington, D.C., 1976.
 (13) J. L. Radomski, *Annu. Rev. Pharmacol. Toxicol.*, **19**, 129 (1979).
 (14) A. J. Stuper, W. E. Brugger, and P. C. Jurs, "Computer Assisted Studies of Chemical Structure and Biological Function", Wiley-Interscience, New York, 1979.

0022-2623/81/1824-0241\$01.00/0 © 1981 American Chemical Society

富士通勤務時代の業務（１）

１９８３年から２００９年（２６年間）

◇コンピュータケミストリー関連業務

■コンピュータケミストリー関連分野での主たる業務

1. 化学関連システムの開発／設計／販売／教育／メンテナンス実施および支援業務
 - ・ CADD (Computer Aided Drug Design) 関連システム構築、教育支援
 - ・ ケモメトリクスおよび構造－活性相関支援システム構築、教育支援
 - ・ 化合物／反応等のデータベースシステム(MACCS／REACCS：現ISIS)の汎用機への移植
2. データ解析支援業務
 - ・ ADAPT, TSAR, DIWA, SARD, **ADMEWORKS** システムを用いた構造－活性相関解析支援
 - ・ ADAPT, TSAR, **MODELBUILDER** システムを用いたケモメトリクス解析支援
 - ・ TOPKAT, **PREDICTOR** システムを用いた化合物のADME／毒性予測支援
3. コンサルタント関連業務
 - ・ 計算機化学関連一般教育
コンピュータケミストリー全般（データベース、化合物表記法、他）
 - ・ コンピュータによる構造－活性相関教育／解析実施
Hansch-Fujita法、パターン認識による構造－活性相関、毒性／分解性／代謝予測、
コンビナトリアルケミストリー／HTS全般、インシリコドラグデザイン支援
 - ・ ケモメトリクス教育／解析実施
多変量解析／パターン認識による化学およびバイオ関連データ解析
4. 情報収集業務
 - ・ コンピュータケミストリー、化学関連データ解析の全般的情報収集
 - ・ 富士通製品の競合製品調査

◇化学関連システムの設計・開発関連業務

■化学関連支援システムの設計・開発および支援実績

1. パターン認識による構造－活性相関・ケモメトリクス支援システム；
ADAPT（米国ペンシルバニア州立大学（基本部分構築）、富士通（GUI部分構築））
2. 人工知能によるバイオアナログウス化合物創出システムの設計支援と立ち上げ；
EMIL（京都大学農学部との共同開発）
3. 生体内代謝予測支援システム構築（北里大学薬学部との共同開発）
4. 合成ルート創出支援システム構築；CASINO（通産省物質研究所との共同開発）
5. 化学スプレッドシートシステム設計支援；KAMELO（田辺製薬様との共同開発）
6. 化学ツールキット構築；CHECK（富士通開発）
7. 3次元構造－活性相関支援システム構築；SARD（富士通開発）
8. JOISオンラインシステムの化合物データ移行設計支援

富士通勤務時代の業務（３）

◇バイオ／蛋白インフォマティクス関連業務

■バイオ／蛋白インフォマティクス関連の業務内容

★本業務は１９９９年１０月より始めたものであり、多くの実績はない。

★現在行っている業務中心にまとめる。

- １．バイオ（特に解析系）関連研究の全般的調査
- ２．ポストゲノムにおけるバイオ解析システムの検討／評価

・バイオ関連支援システムの設計・開発および支援実績

- １．モチーフ探索システム設計（国立遺伝学研究所）；初期の参加のみ（１９９－年）
モチーフ探索へのパターン認識適用の可能性の検討
- ２．SNP解析関連業務
 - ・徳島大学医学部ゲノム解析センター 板倉光夫教授による重点領域研究
「ヒト遺伝子多型系の体系的解析による疾患遺伝子の解析」に参加
 - ・鎌谷直之教授（東京女子医科大学膠原病リウマチ痛風センター センター長）との
遺伝子多型解析システム開発関連作業に従事
- ３．発現プロファイル解析関連業務
- ・NEDOプロジェクト
“遺伝子発現頻度解析（通称 高原プロジェクト）”への参加（２０００年度）
発現プロファイルデータベースシステムと連携する薬理活性データベース構築

・蛋白関連支援システムの設計・開発および支援実績

- １．蛋白シミュレーション関連業務
 - ・NEDOプロジェクト（２００３より参加、２００７年終了）
“生体高分子構造情報利用技術開発（通称 中村プロジェクト）”への参加
- ２．創業バリューチェーン（大阪大学坂田教授主催）への参加（２００５より参加）
 - ・産官学の３共同体制で行う新規医薬品開発を目指したコンソーシアム

富士通勤務時代の業務（４）

◇環境関連業務

■環境関連の業務内容

1. 生体毒性（発ガン性）関連研究；J u r s研究室での研究
 - ・ A D A P Tシステムを用いた多変量解析／パターン認識による発ガン性予測
2. 環境ホルモン特性に関する研究
 - ・ A D A P TおよびT s e rシステムを用いた環境ホルモン特性に関する研究
 - ・ 環境関連支援システムの設計・開発および支援実績；
 1. 財団法人 化学物質評価研究機構のW E B上での毒性予測システムの設計および開発支援
 2. 財団法人 製品評価研究機構および通産省（現；文部通産省）からの調査依頼
「諸外国における化学物質規制に関する構造－毒性相関の適用」に関する調査
3. 環境研究所（筑波）との化学物質毒性予測に関するシステム開発共同研究開始
4. 国立医薬品食品衛生研究所との共同研究
 - ・ A D M E W O R K Sで用いる化審法および薬事法関連化合物規制予測モデルの開発
 - ①A m e s試験予測モデルの構築（T A 9 8およびT A 1 0 0）：2 0 0 4年度
 - ②染色体異常試験（小核試験）予測モデル構築：2 0 0 5年度
5. 大阪府立環境研様との、皮膚感作性予測に関する共同研究実施：2 0 0 6年度
6. 国立環境研究所様、環境毒性予測に関する共同研究及びシステム構築
 - ・ 2007年度；甲殻類遊泳阻害試験48時間EC50、藻類生長阻害試験72時間EC50
 - ・ 2 0 0 8年度；LogP 推算式の構築支援
7. （財）日本自動車研究所様への研究業務支援
 - ・ 廃棄ガスと発がん性の相関解析支援
8. 厚生労働省、感作性物質分類小委員会での皮膚／気道感作性予測モデルの構築

◇医療関連業務

■医療関連の業務内容

1. 日本医科大学様との医療を目指したN M Rメタボノミクス関連研究支援
2. 皮膚および気道感作性予測に関する支援作業：感作性物質分類小委員会（福井大学様）

序章：

- ・ 自己紹介
- ・ 化学データサイエンスおよび人工知能の時代的な流れと、変化

化学データサイエンスやAIの適用という観点からの 理解すべき点

今回は手法の**基本原理の理解**よりも、**運用上でのテクニックや事例**を学ぶ
特徴、制限事項、長所、短所、適用限界、他

時代が大きく変化し、昔と現在では化学データサイエンスやAIを
実施するための全体的な環境は大きく変化している

時代の変化に応じて、同じ研究をするにしても
ひと昔前と現在では実施環境が大きく変化したことに伴い
学ぶべき内容も異なり、新たに学ぶべき事項も出てくる

時代の変化により、学ぶべき内容も大きく変化している。
今回はこの事実がわかるように**自動車を用いて仕事をする**時の
手続きを例にしてまとめて行く

「化学データサイエンスやAI（人工知能）を学びたい」



「学びたい」には二つのアプローチが存在する

基本と原理の
展開

データサイエンスやAIの
基本や原理を学ぶ
専門家を目指す

今
回

自分の研究への
適用技術の展開

データサイエンスやAIを
自分の研究に適用し、新たな
発見や展開を目指したい

時代の流れと、研究環境および研究内容や研究スタイルの変化

◇ 1990年代のデータサイエンス実施環境（コンピュータ初期の時代）

研究しようにも、データサイエンス実施関連ソフトが無い。また、コンピュータの能力も低く、メモリー等の問題で多数のサンプルの扱いは難しく、計算速度も遅くて時間がかかっていた。また、データベースも殆どなく、自分でデータの収集が必要だった。この時代に化学データサイエンスやAI研究を実施しようとする場合、以下のような**研究の実施前提**条件のクリアが必要となる。

- ・ データサイエンス（データ解析手法）等の数学（基本）を理解できる
- ・ コンピュータ関連での知識と実施環境がある
- ・ 数式をプログラミングに変換できること
- ・ プログラミングできること
- ・ 化学とコンピュータの連携技術（コンピュータケミストリ）を理解していること
- ・ 化学分野へのデータ解析処理（ケモメトリクス）を理解していること
- ・ 化学関連の基礎知識を有すること
- ・ 解析目的である薬理活性、毒性、物性等に理解があること

◆ 化学データサイエンスおよびAIの実施プロセス

1. データ解析をコンピュータで実施するための環境構築（ケモメトリクス研究支援ソフト）
2. 先に構築したケモメトリクス研究支援ソフトを使って、データ解析を実施
3. 目的とする研究のためのデータ解析の実施

時代の流れと、研究環境および研究内容や研究スタイルの変化

◇ 1990年代のデータサイエンス実施環境

研究しようにも、**データサイエンス実施関連ソフトが無い**。また、**コンピュータの能力も低く**、メモリ等の問題で多数のサンプルの扱いは難しく、計算速度も遅くて時間がかかっていた。

1. データ解析と化学を融合した環境構築（ケモメトリクス研究支援ソフト）

手順1：データ解析ソフトの構築

- ①データ解析手法の基本原理や数式を学んで理解
- ②数式をコンピュータに乗せる工夫がいる
- ③データ解析ソフトの構築

手順2：化合物構造式をコンピュータに乗せるソフトか、手入力

手順3：化合物構造式を数値データ（パラメータ）に変換するソフトの構築

手順4：パラメータの前処理プログラム構築

2. 先に構築したケモメトリクス研究支援ソフトを使って、データ解析を実施

手順1：サンプル化合物の収集とシステムへの入力

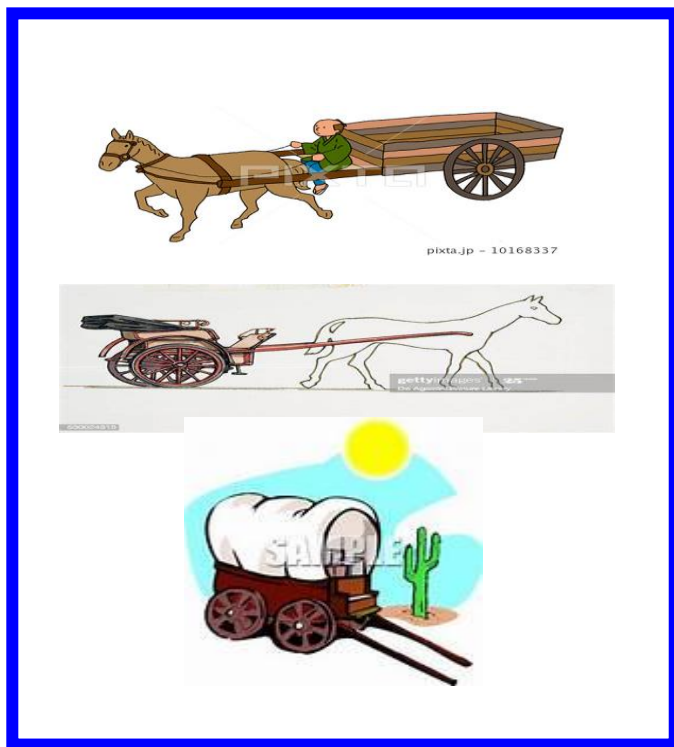
- ①研究解析目的を実現するに十分な内容を有する化合物サンプルと薬理活性等のデータ
- ②集めたサンプルのチェックと吟味（化合物の異性体、立体情報、他）、構造の間違いチェック等
- ③サンプルと数値データのシステムへの入力

3. 目的とする研究のためのデータ解析の実施

研究の実施には、数か月から一年以上かかる

自動車の利用を考えた時の実施手順

コンピュータ初期の時代における作業環境イメージ



移動手段の自作が必要



交通インフラ不十分



仕事の種類も多くない

時代の流れと、研究環境および研究内容や研究スタイルの変化

◇ 2021年代のデータサイエンス実施環境（現在でのコンピュータ環境）

データサイエンス実施関連ソフトが多数存在する。また、コンピュータの能力が高くなった。

化合物関連のデータベースも多様多種で、サンプル数も大きくなり、ビッグデータに近いものも出ている。ひと昔前のように、サンプルデータの収集に時間がかかることは少なくなった。計算速度も速くなり、扱えるサンプル数も増え、高度で複雑な計算も高速で実施可能となっている。

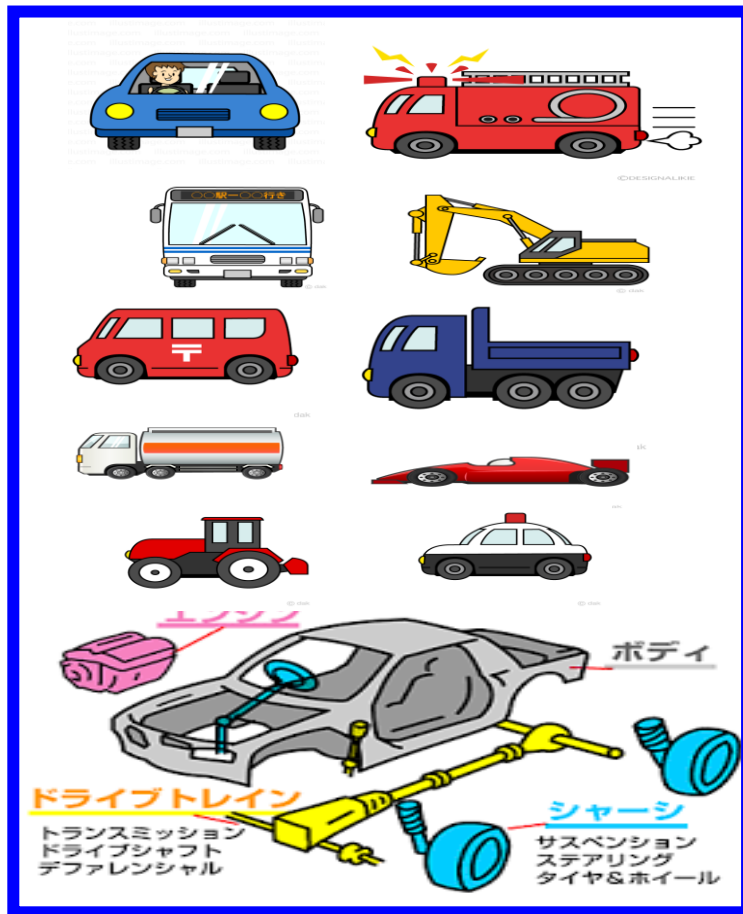
現在における研究の実施前提条件は以下のように大きく変化してきた。解析ソフト関連の構築技術の重要性が急速に低下してきた。今後は解析の運用関連技術が急速に重要性を拡大している。

- ~~データサイエンス（データ解析手法）等の数学を理解できる~~
- ~~コンピュータ関連での知識と実施環境がある~~
- ~~数式をプログラミングに変換できること~~
- ~~プログラミングできること~~
- ・ 化学とコンピュータの連携技術（コンピュータケミストリ）を理解していること
- ・ 化学分野へのデータ解析処理（ケモメトリクス）を理解していること
- ・ 化学関連の基礎知識を有すること
- ・ 解析目的である薬理活性、毒性、物性等に理解があること

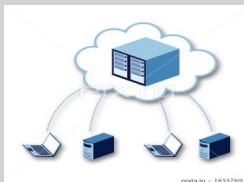
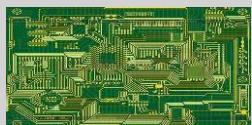
◆ 化学データサイエンスおよびAIの実施プロセス

1. データ解析をコンピュータで実施するための環境構築（ケモメトリクス研究支援ソフト）
2. 先に構築したケモメトリクス研究支援ソフトを使って、データ解析を実施
3. 目的とする研究のためのデータ解析の実施

自動車の利用を考えた時の実施手順



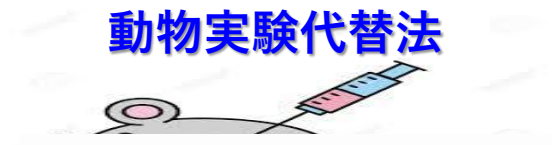
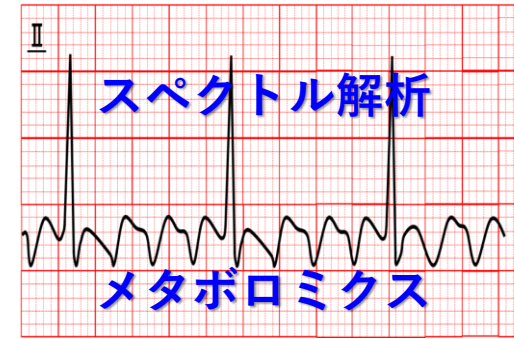
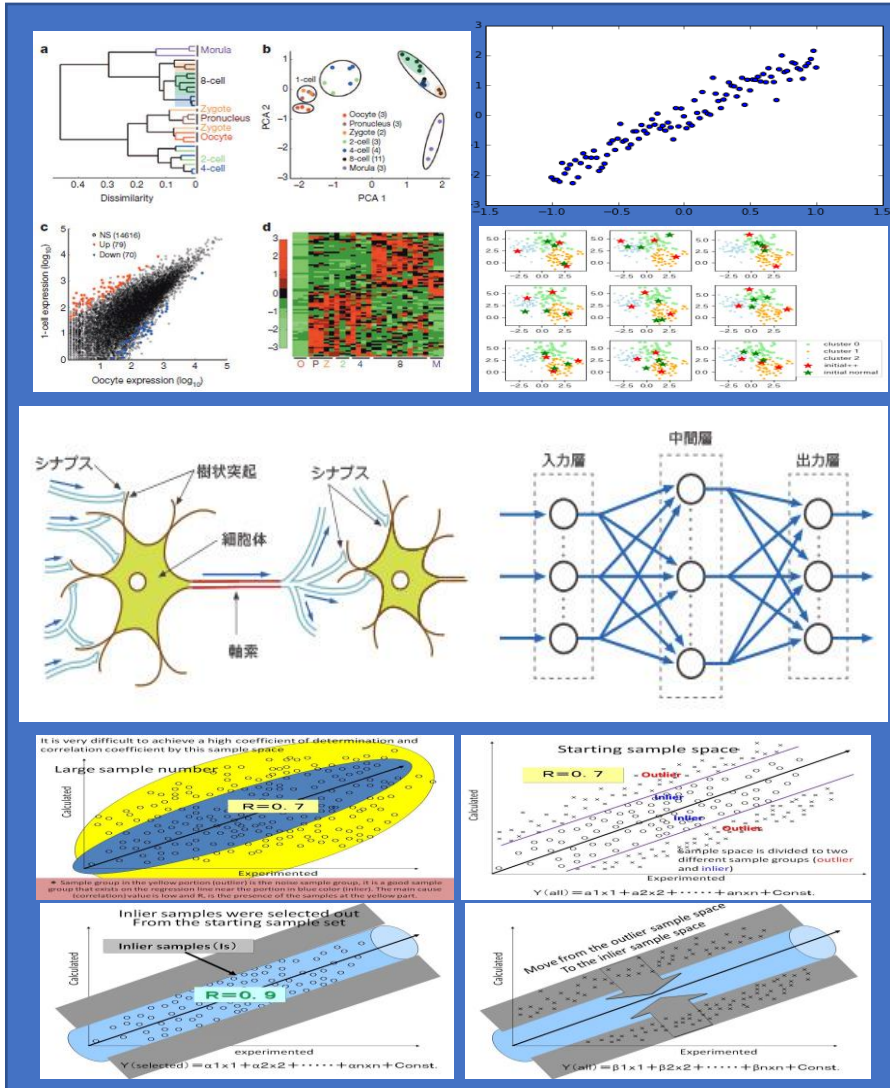
全体のインフラ関連



今回のチュートリアル対象範囲

適用分野の研究

データサイエンスやAI手法 ：目的実施に必要な機能



■コンピュータ関連およびインフラ関連技術

①種々データベース関連技術

- ・リレーショナルDB、ネットワーク型、その他

②ハード関連

- ・スマートフォン、タブレット、PC、WS、グラフィクスカード、HPC、他

③種々ソフトウェア

- ・ケモメトリクス支援、データサイエンスや人工知能関連ソフト

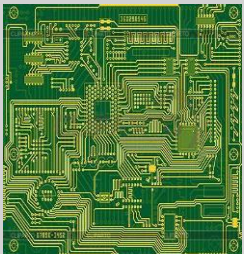
④関連ソフト

- ・API (Application Programming Interface)
IBM Watson、Microsoft Azure、他
- ・開発言語 (C、Java、Python、TensorFlow、他)

■データベースのサイズ、種類、検索形式関連

- ・通常のデータベース：ある分野や研究目的のように十分なサイズ
- ・ビッグデータベース：データ量が極めて大きなデータベース
- ・専門データベース (化合物、タンパク、薬理活性、毒性、物性、他)
- ・特殊データベース (機器スペクトルデータベース、医療画像データベース)
- ・汎用データベース (文献データベース、科学ニュース関連データベース、他)

全体のインフラ関連



今後の化学データサイエンスやAIの変化と適用

□ 今後の化学データサイエンスやAIを取り巻く環境の変化

- ・ コンピュータ関連の様々な技術の向上と展開
- ・ データサイエンスやAI自体の発展
- ・ 多種多様な関連データベースの構築やビッグデータ
- ・ ケモメトリクス研究支援システムの発達と開発
- ・ 研究パターンの変化
- ・ 適用分野の拡大と融合研究の増加
- ・ その他

三段階での化合物評価／予測システムの展開と実現化

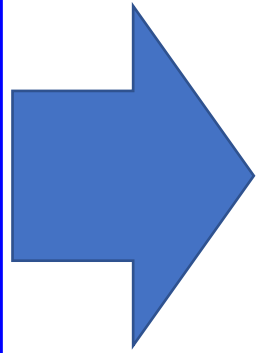
自動車の利用からTVの利用に変化した時の実施手順

三段階操作

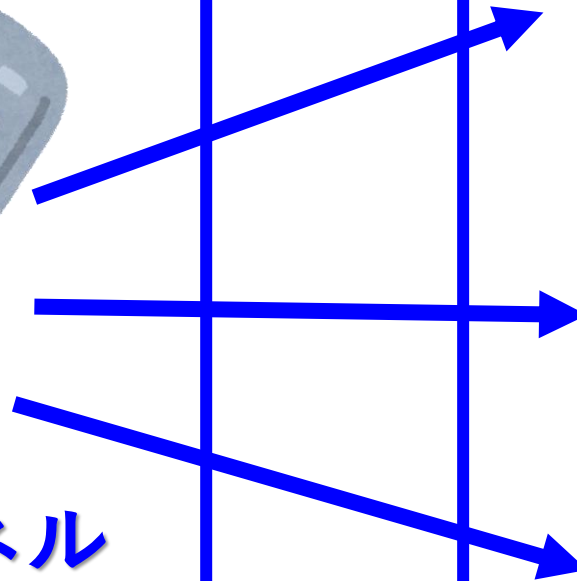
今後の時代的な進歩による予測／評価研究実施イメージ



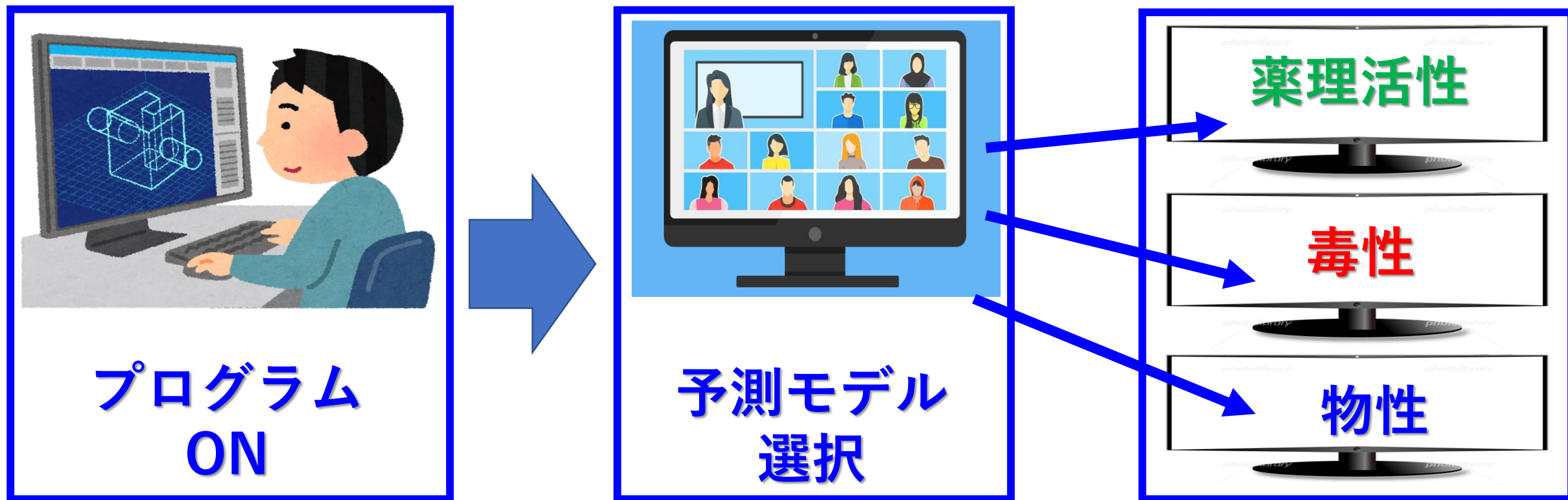
スイッチ
ON



チャンネル
選択



自動車の利用からTVの利用に変化した時の実施手順 三段階操作で完了



予測や評価プロセスの完全自動化による、
目的化合物の種々特性を三段階の操作で取得
⇒⇒⇒⇒⇒「スマート創薬」の実現

TS-02 第一章 時代の変化に続く