Machine learning in chemoinformatics and drug discovery

Chemical graph theory

Chemical descriptor, fingerprints

Machine learning

Y.-C. Lo, S. E. Rensi *et al,* Drug Discovery Today (if=6.88), Aug. 2018.

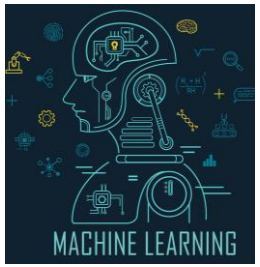
MODNBIO

Overview

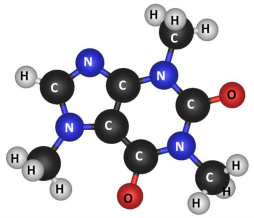
- CADD의 개요
- Overview of chemoinformatics
  - Chemical graph theory
  - Chemical descriptor
  - Chemical fingerprints
  - Machine learning in QSAR
- QSAR modeling
- Conclusion

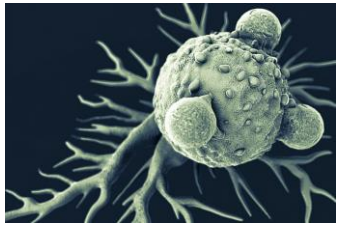
CADD의 개요

- Computer-aided drug discovery (CADD)
  - physical model (quantum chemistry, molecular dynamics simulation etc.)
  - Machine learning (pattern recognition)
    - 경험의 연장, 그를 통한 예측 -〉 효율성이 좋고, 큰 데이터에 유리함
    - Chemoinformatics에서 ML의 역할



understand





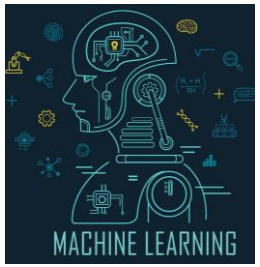
for

toxicity
metabolism
drug-drug interaction
similarity

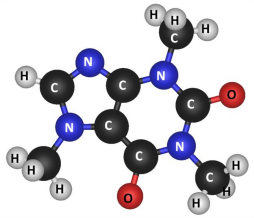
QSAR

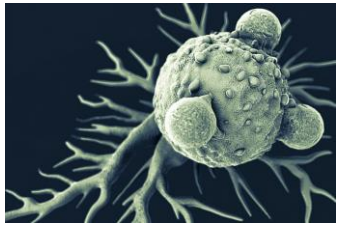
CADD의 개요

- Quantitative Structure-activity Relationship (QSAR)?
- Quantitative Structure-property Relationship (QSPR)?
  - 물질이 가지고 있는 화학 구조적 특징으로부터 반응성 혹은 독성과 같은 특징을 유추하는 방법
  - Hansch and Free-Wilson analysis (선형 모델, 데이터 접근의 어려움)
  - 깊고 복잡한 비선형적 모델, 많은 양의 bigdata가 요구됨



understand





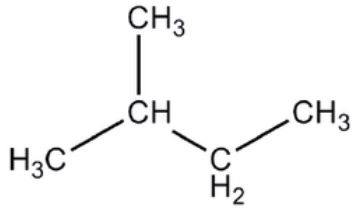
for

toxicity
metabolism
drug-drug interaction
similarity

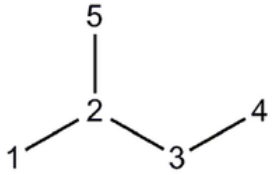
QSAR

Overview of chemoinformatics

- Chemical graph theory



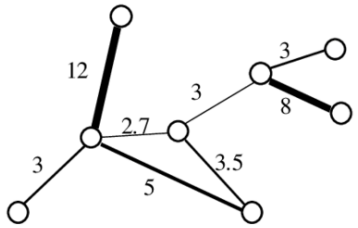
Molecule



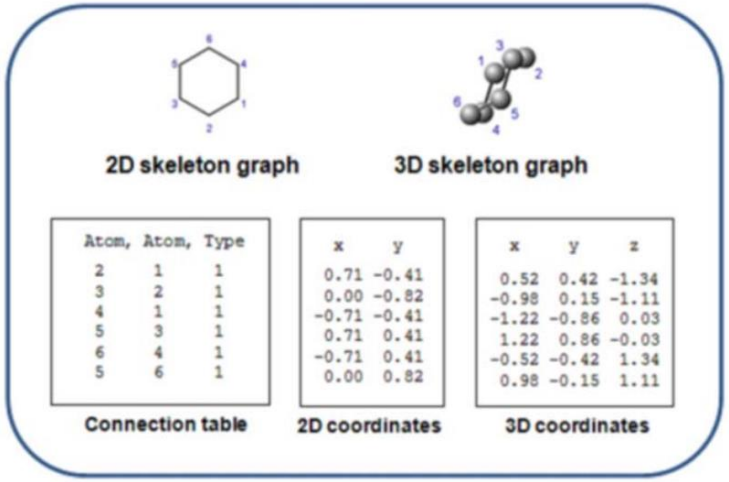
Graph

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

adjacency matrix

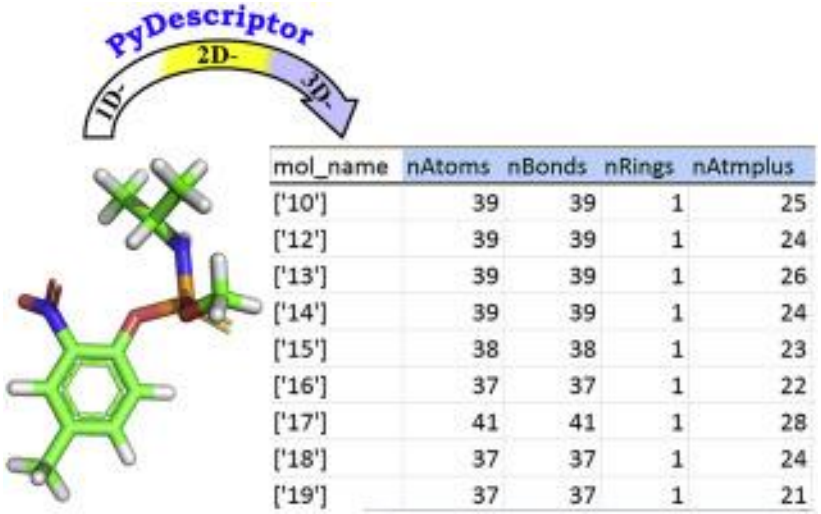


	1	2	3	4	5	6	7	8
1		3	0	0	0	0	0	0
2	3		12	2.7	5	0	0	0
3	0	12		0	0	0	0	0
4	0	2.7	0		3.5	3	0	0
5	0	5	0	3.5		0	0	0
6	0	0	0	3	0		3	8
7	0	0	0	0	0	3		0
8	0	0	0	0	0	8	0	

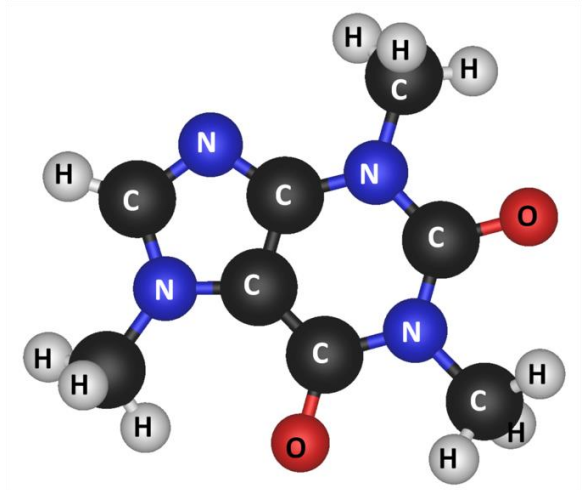


Overview of chemoinformatics

- Chemical descriptor (0d, 1d, 2d, 3d)
  - 컴퓨터로 입력이 가능한 화학구조의 특징

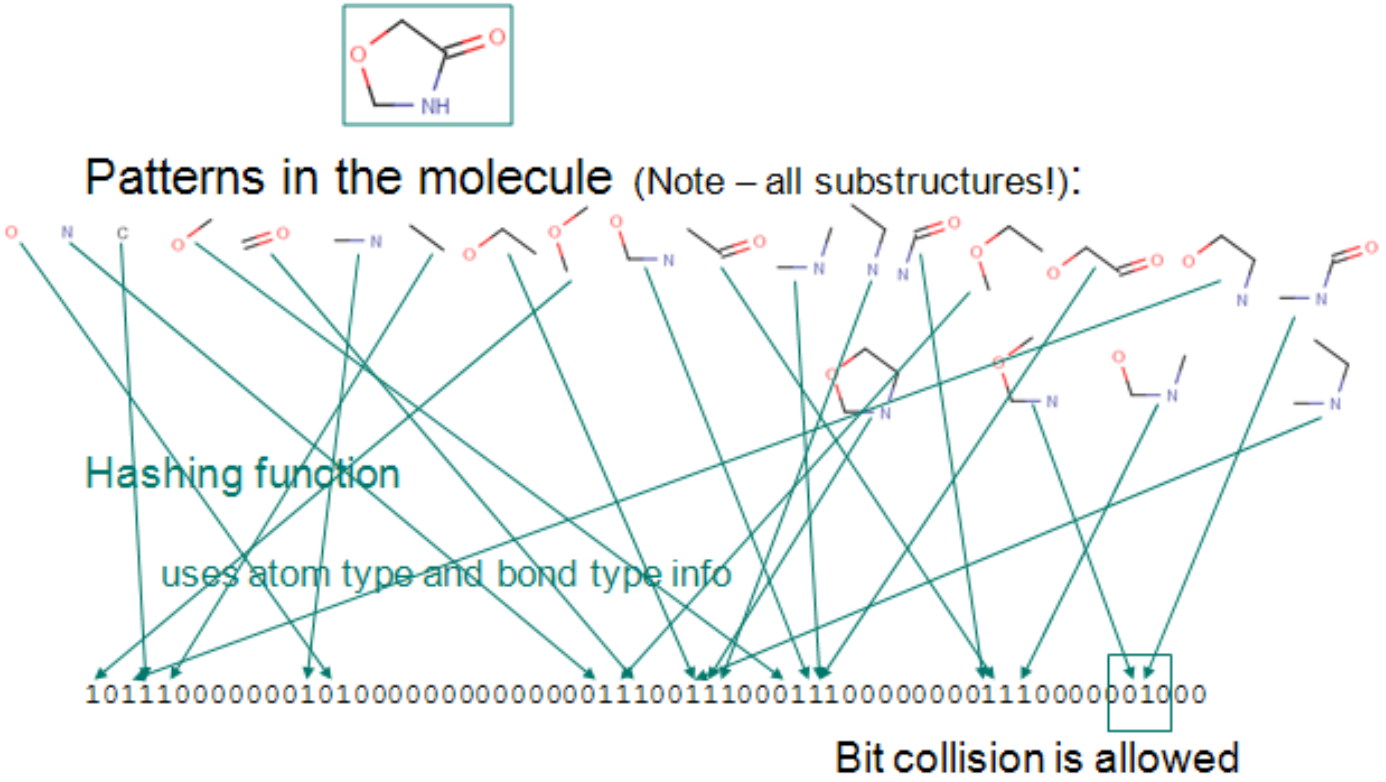


mol_name	nAtoms	nBonds	nRings	nAtmplus
['10']	39	39	1	25
['12']	39	39	1	24
['13']	39	39	1	26
['14']	39	39	1	24
['15']	38	38	1	23
['16']	37	37	1	22
['17']	41	41	1	28
['18']	37	37	1	24
['19']	37	37	1	21



Overview of chemoinformatics

- Chemical fingerprints
  - 컴퓨터로 입력이 가능한 화학구조의 특징을 1차원 벡터 형태로 변환한 것



Overview of chemoinformatics

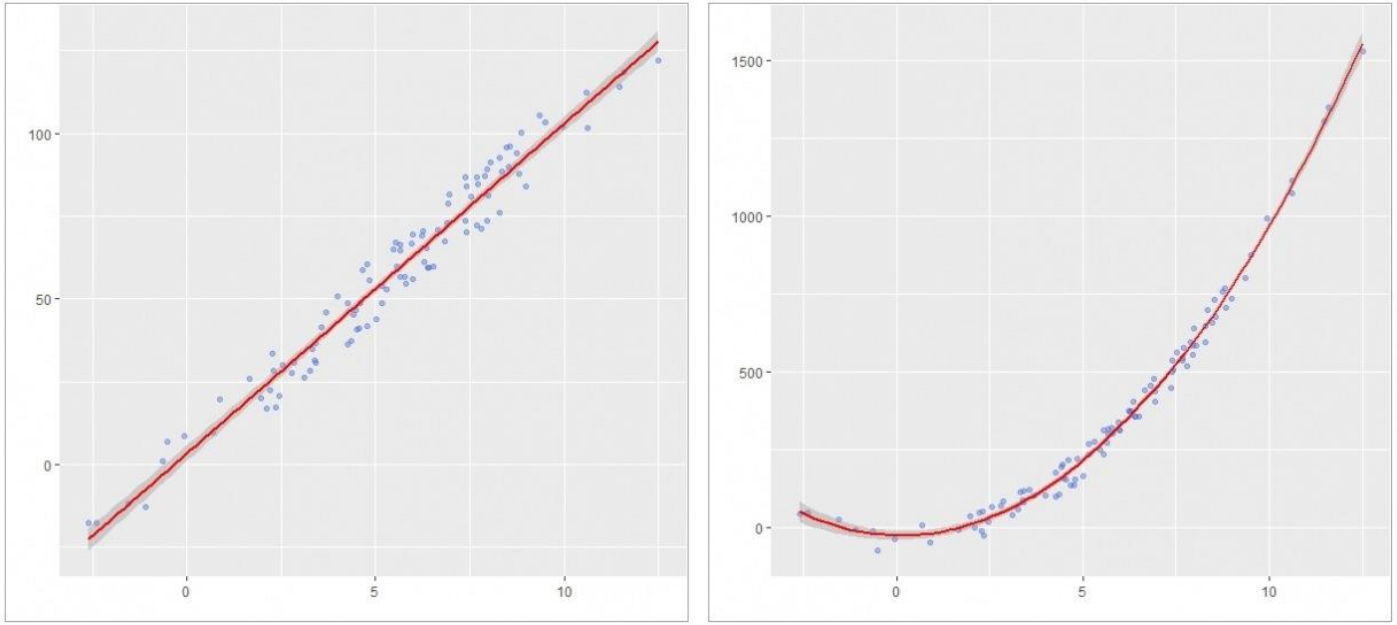
- Machine learning in QSAR
  - Naive Bayes

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$

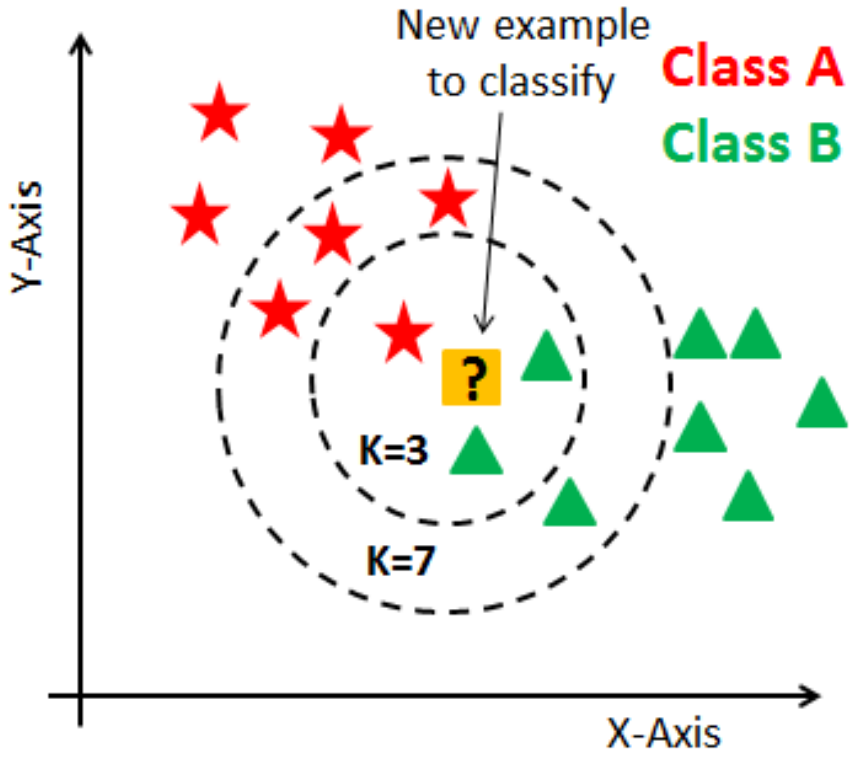
Overview of chemoinformatics

- Machine learning in QSAR
  - Regression analysis



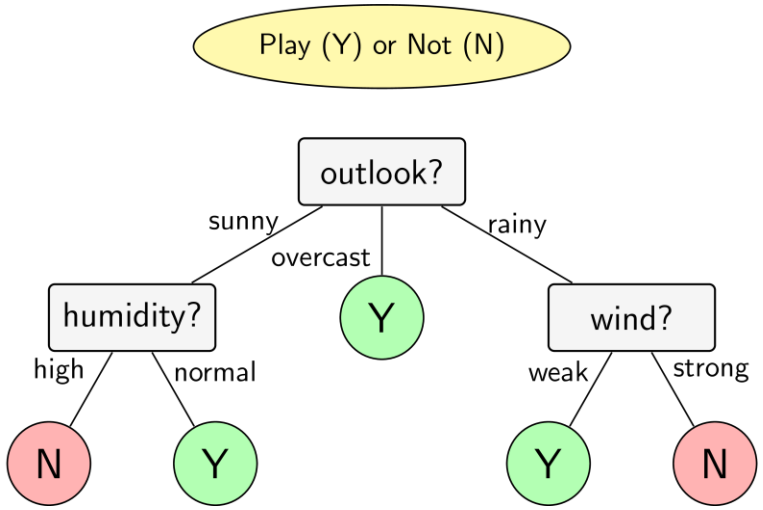
Overview of chemoinformatics

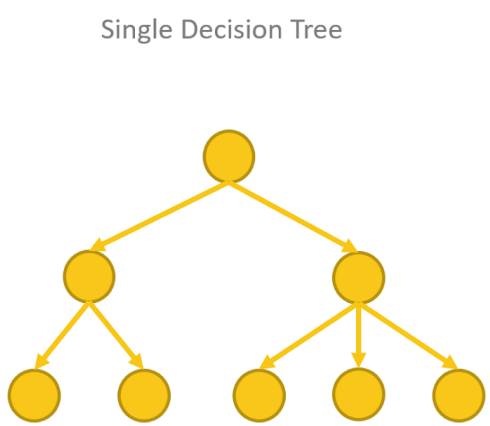
- Machine learning in QSAR
  - k-Nearest neighbors

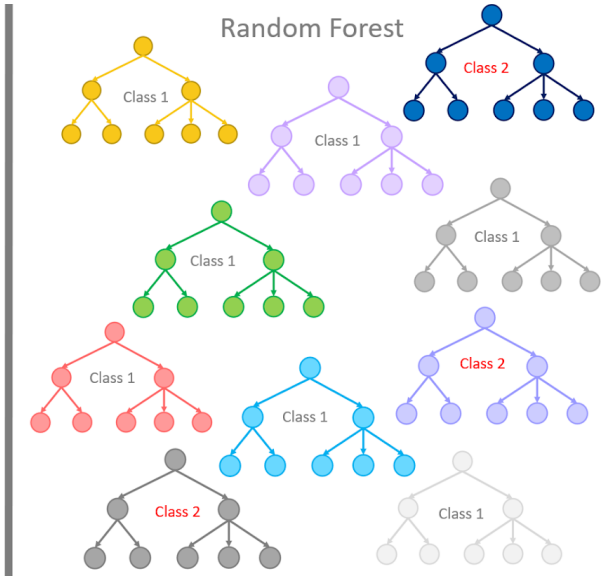


Overview of chemoinformatics

- Machine learning in QSAR
  - Random forest

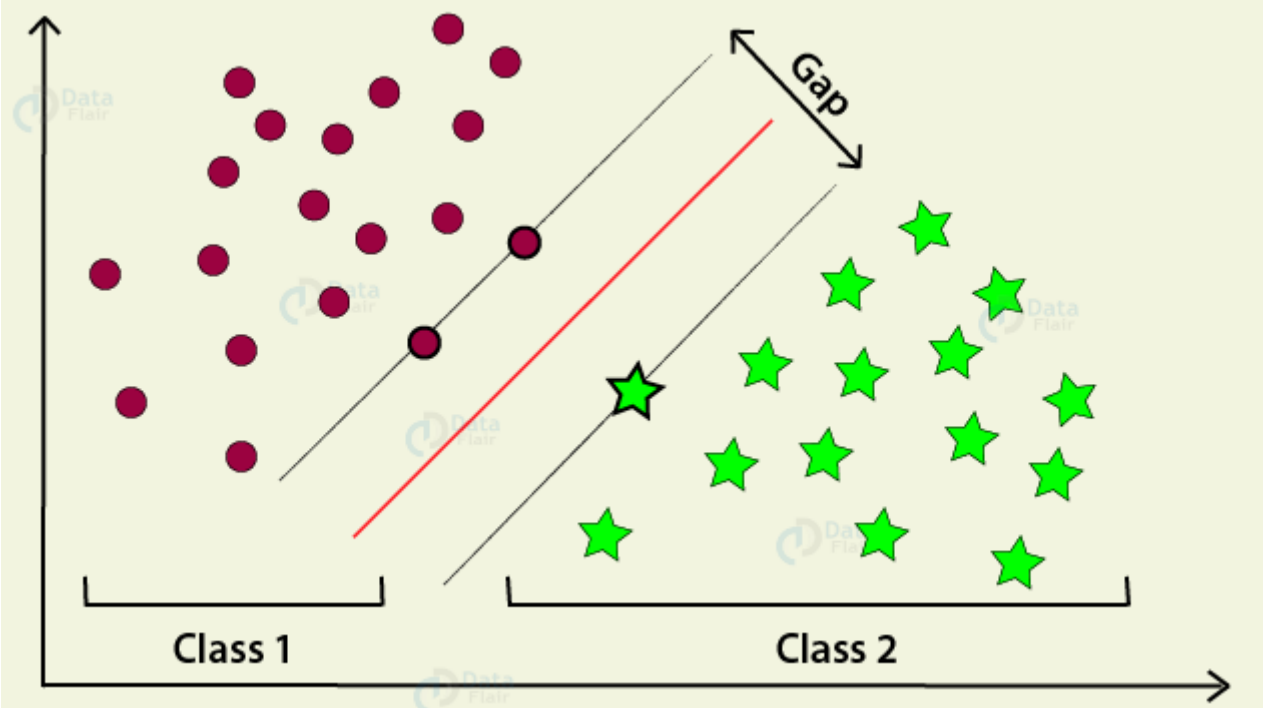






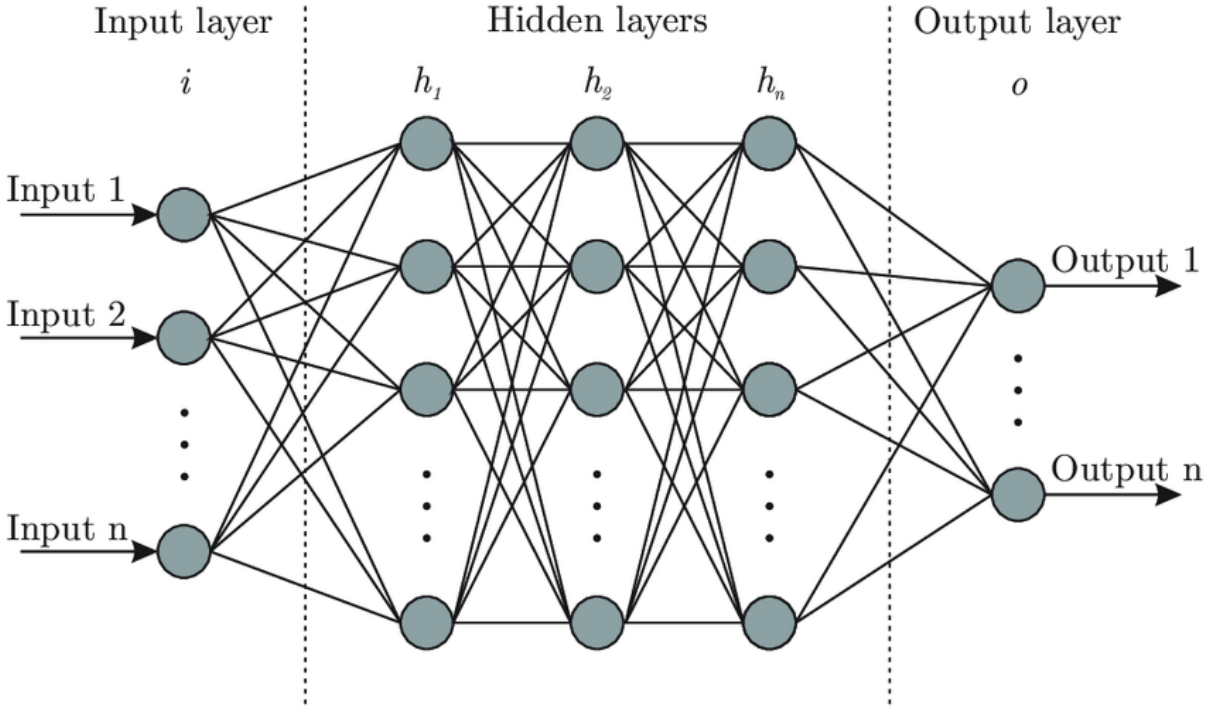
Overview of chemoinformatics

- Machine learning in QSAR
  - Support vector machine



Overview of chemoinformatics

- Machine learning in QSAR
  - Neural network and deep learning



Overview of chemoinformatics

- QSAR modeling
  1. problem definition - 해결하고자 하는 문제가 무엇인지 정의하는 단계
     - similarity, protein-ligand affinity, toxicity, biological response, physiochemical property
  2. molecular encoding - 분자의 특성을 컴퓨터에 입력가능한 형태로 바꾸는 것
     - descriptor, fingerprint, graph vector
  3. feature selection - 비지도 학습 및 domain knowledge를 통한 특징점 선택
     - PCA, ICA, mean shift
  4. model implementation - 예측을 위한 모델 구상 및 구현
     - a variety of machine learning and deep learning architecture

Overview of chemoinformatics

- Conclusion
  - 단순한 protein - ligand interaction으로 임상약물안전기준을 만족시키기 힘들다.
  - 다양한 data type의 통합하는 data fusion 기술이 요구된다.
    - structural, genetic, pharmacological data from molecular and organism level
    - 최신 머신러닝, 딥러닝 기술이 요구됨 - 신약개발을 위한 많은 양의 데이터를 통한 새로운 모델이 효과를 보이고 있음