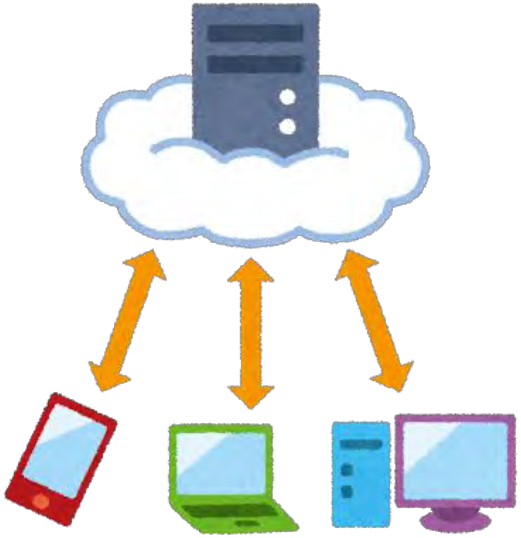
SG7

自然科学における統計サンプリングのプログラミング・シミュレーションの実践

サンプリング

- ビッグデータの時代
  - PCの能力の向上…ムーアの法則
  - 大量のデータを解析できるようになった



- 統計データ…サンプリングが必要…どうするか?

機械学習

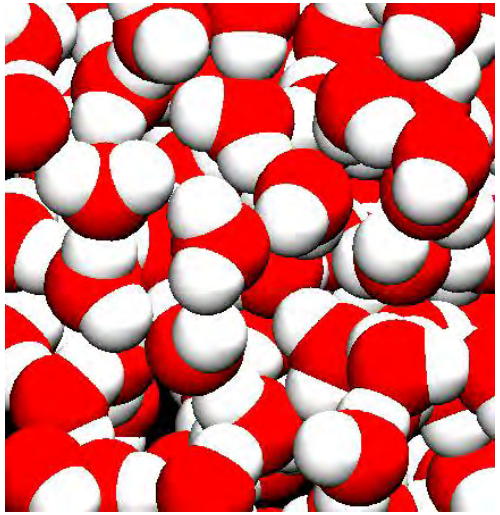
- 得られたビックデータの活用
- 社会の様々なところで使われ始めている
  - 画像認識、自然言語処理、金融工学、etc…

- 応用は様々…中身はどうなっているのだろう?

内容

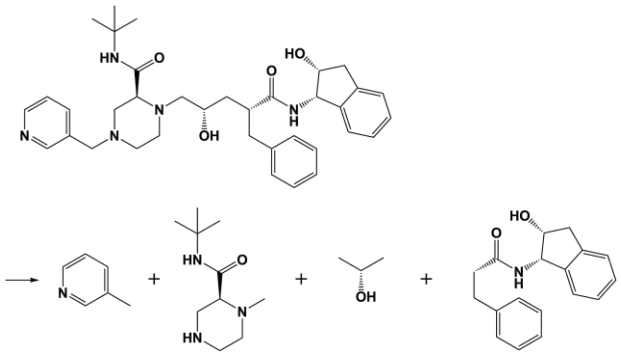
各々の興味に応じたテーマに対してプログラミングを行い結果を発表することで知識を共有した

産総研
手書教育漢字データベースより



- 機械学習
  - 手書き文字認識
- サンプリング
  - 分子動力学(MD)シミュレーション
- 応用(組み合わせ)
  - MDによって得られたサンプルを用いたデータ同化解析

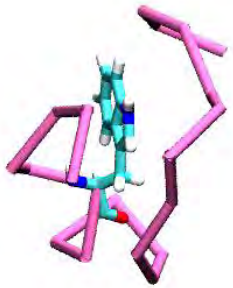
セミナー



- 畑中実穂氏（奈良先端大学）

「自動反応経路探索を用いる触媒反応の機構解明と機械学習による効率的解析」

✓ k-means法による化学構造の分割と解析



- 松永康佑氏（理研、JST さきがけ）

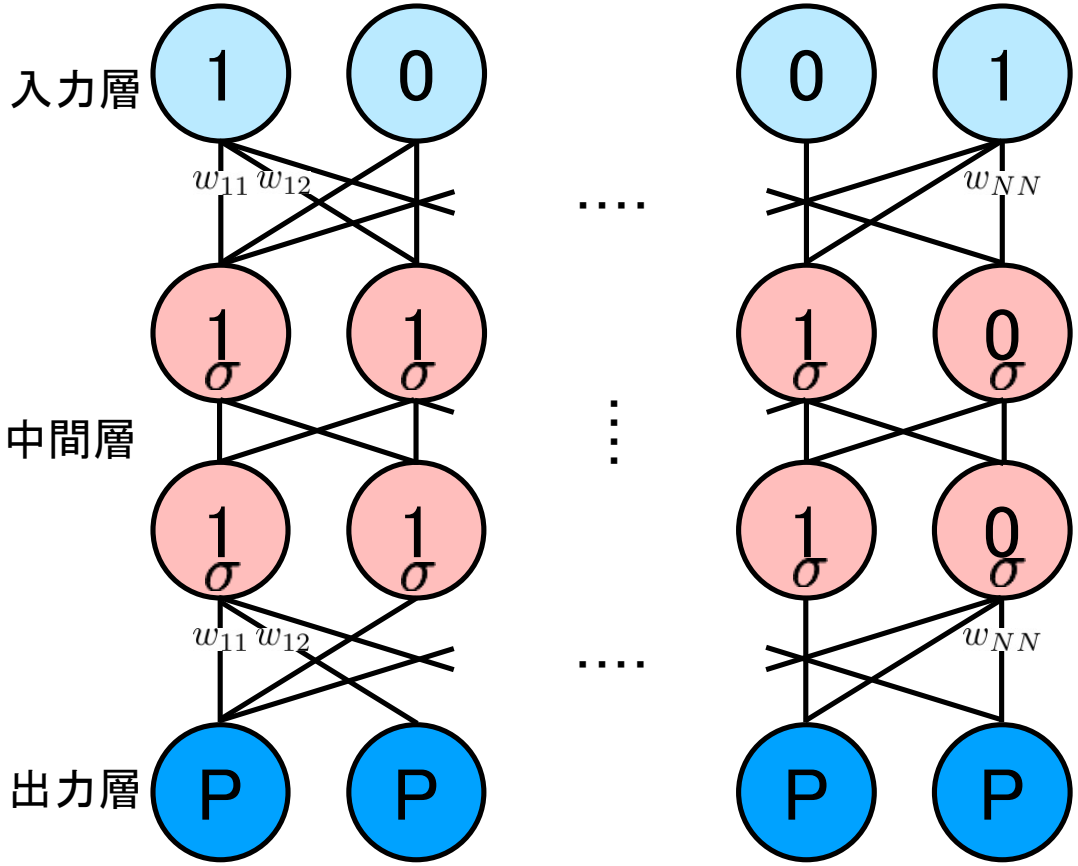
「機械学習を用いた計測とシミュレーションの統合によるタンパク質動態解析」

✓ MD シミュレーションと機械学習によるタンパク質折り畳み過程の解析とデータ同化

機械学習

村田　隆
（M1, 生物物理）

Deep belief network



Forward Propagation

$$\sigma = \frac{1}{1+\exp(\mathrm{input})}$$

$$\mathrm{input_j} = \sum_i w_{ij} \bigcirc_i$$

勾配降下法

$$\mathbf{W} = \mathbf{W} - \epsilon \frac{\partial J}{\partial \mathbf{W}}$$

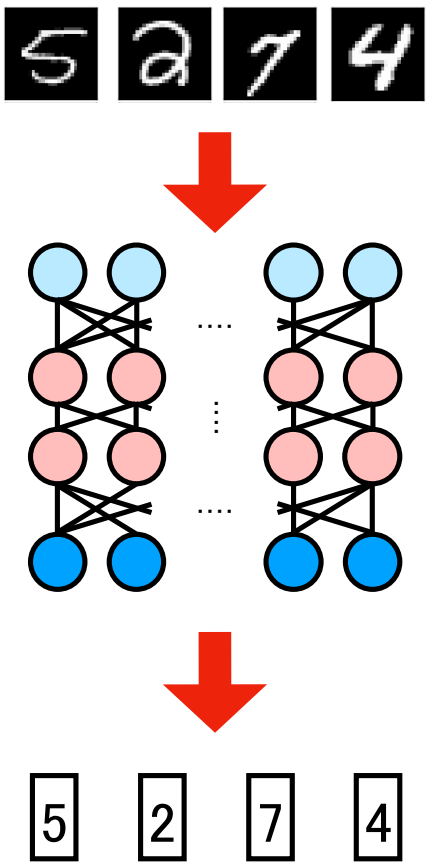
Back Propagation

$$\frac{\partial J}{\partial W_{ij}^{l}} = \frac{\partial J}{\partial \mathrm{input_j^l}} \mathrm{output_i^{l-1}}$$

$$\frac{\partial J}{\partial \mathrm{input_j^l}} = \frac{\partial J}{\partial \mathbf{input^{l+1}}} \cdot \frac{\partial \mathbf{input^{l+1}}}{\partial \mathrm{input_j^l}}$$

手書き文字認識

28x28 pixel



Property

実装:python
(numpy)

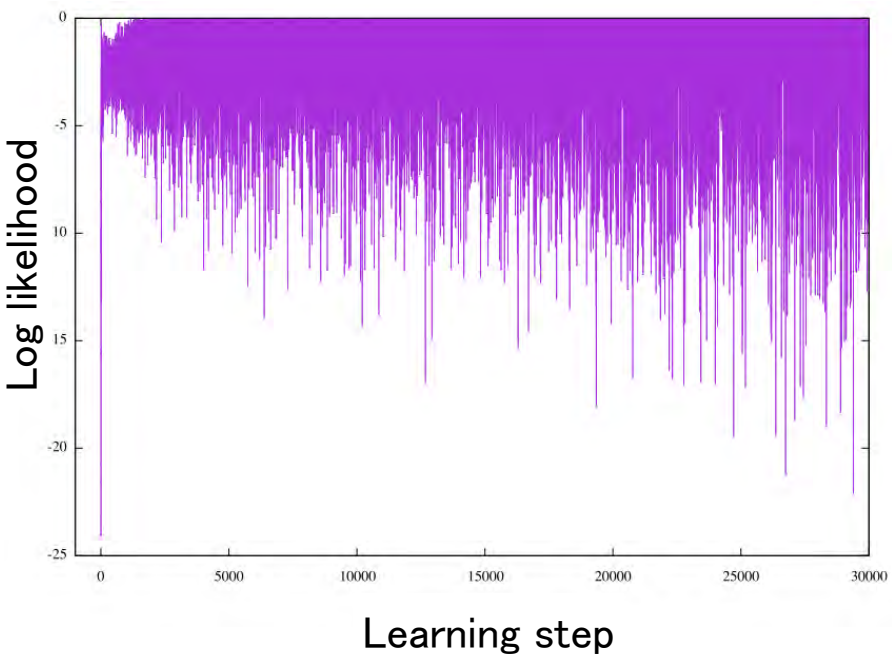
ノード数:784

中間層数:2

備考:
momentum

Result

学習:50000step
正答率:77%

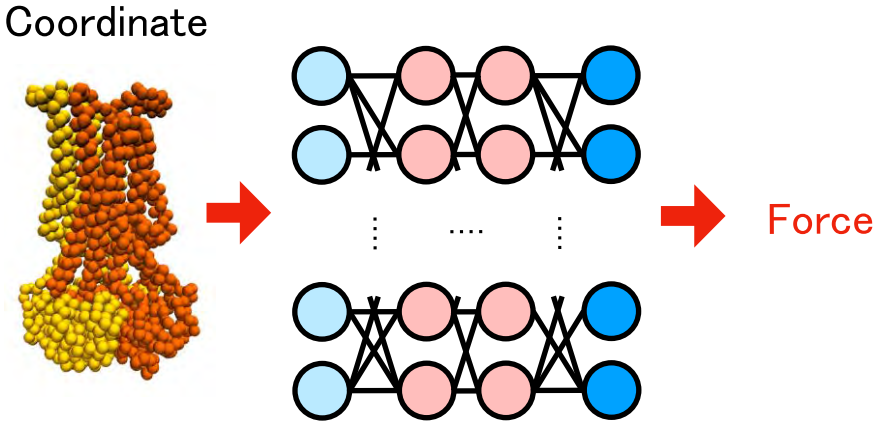


学習を重ねるほどエラーが大きくなっている⇨過学習の兆候

ノード数の過多

生物物理への応用

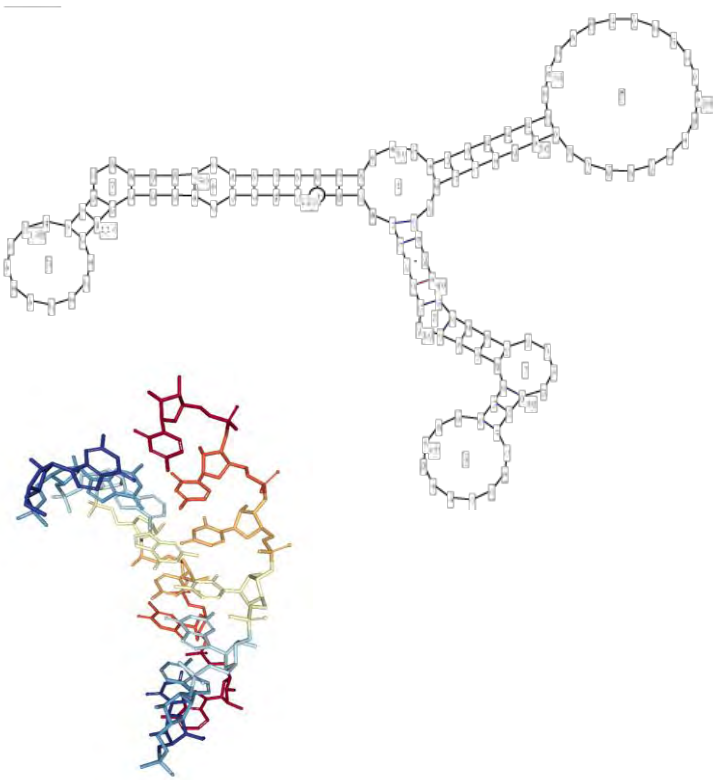
DBNによる力場



既存MD力場パラメータのリファインメント

$$
\begin{aligned}
V_{AICG2}(\mathbf{R}|\mathbf{R}_0) &= \sum_{ibd} K_{b,ibd}(b_{ibd} - b_{ibd,0})^2 + V_{loc}^{flp} \\
&+ \sum_{i+2 \le j \le i+3} \varepsilon_{loc,ij} \exp\left(-\frac{(r_{ij} - r_{ij0})^2}{2W_{ij}^2}\right) \\
&+ \sum_{i<j-3}^{nat\ contact} \varepsilon_{go,ij} \left[5\left(\frac{r_{ij0}}{r_{ij}}\right)^{12} - 6\left(\frac{r_{ij0}}{r_{ij}}\right)^{10}\right] \\
&+ \sum_{i<j-3}^{non-native} \varepsilon_{ev} \left(\frac{d}{r_{ij}}\right)^{12}.
\end{aligned}
$$

核酸構造予測



分子動力学
シミュレーション

西尾　宗一郎
（B4, 化学）

単原子分子のシミュレーション

- Lennard-Jones potential

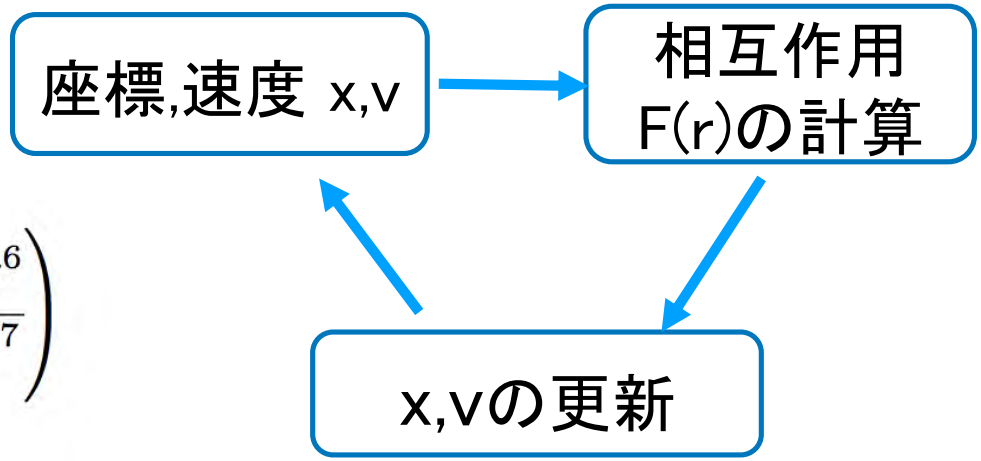
$$U(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

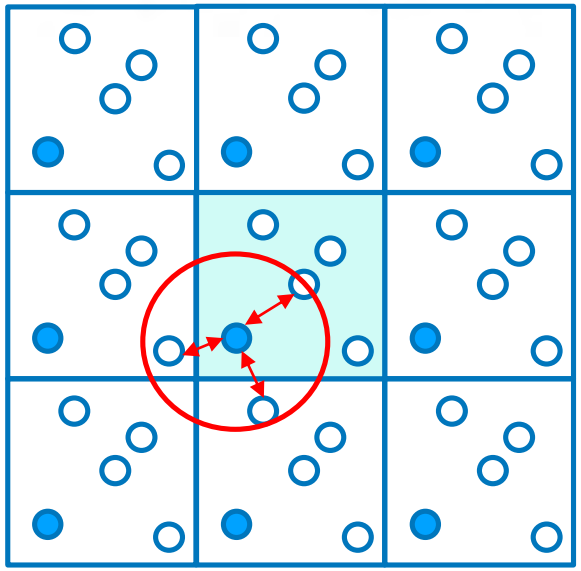
$$F(r) = -\frac{d}{dr}U(r) = 4\epsilon\left(12\frac{\sigma^{12}}{r^{13}} - 6\frac{\sigma^{6}}{r^{7}}\right)$$

- 速度ベルレ法

$$x_{t+\Delta t} = x_t + v_t \cdot \Delta t + \tfrac{1}{2}F(x_t, t)\cdot \Delta t^2$$

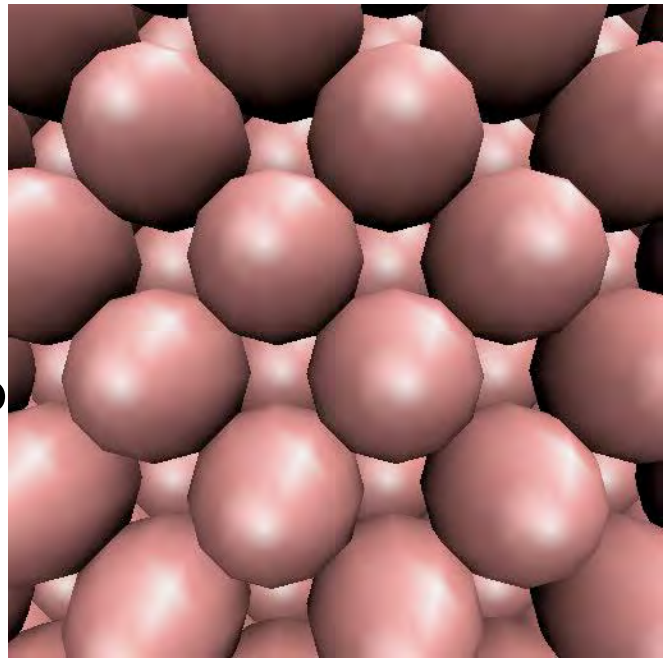
$$v_{t+\Delta t} = v_t + \tfrac{1}{2}\Delta t\left(F(x_t, t) + F(t+\Delta t, x_{t+\Delta t})\right)$$



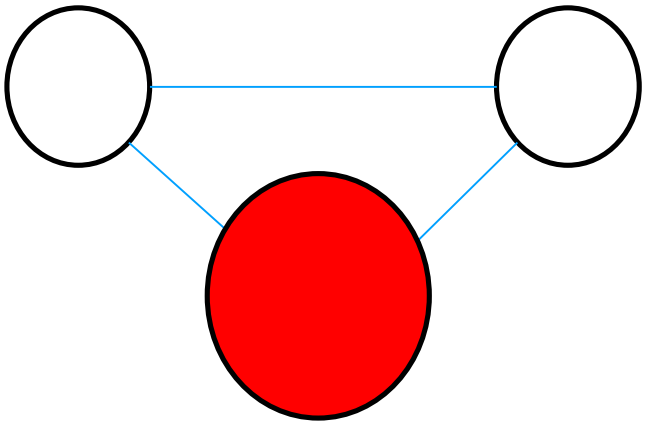


●: 注目する分子に対し、

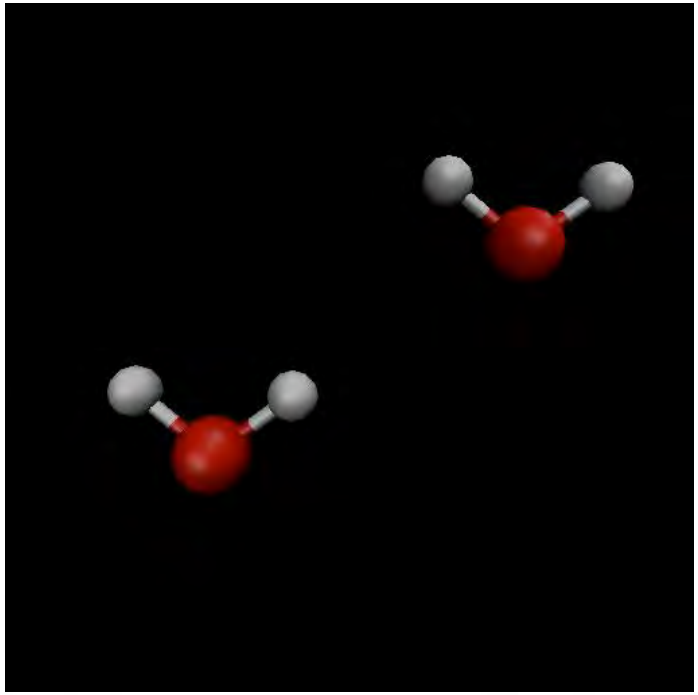
○ の範囲内にある分子との相互作用を考慮する



H_2O分子のシミュレーション



・O−H, H−H間の結合長の固定と速度の補正（RATTLE法）



- クーロン相互作用

$$U(r) = \frac{1}{2}\sum_{\mathbf{n}}\sum_{i} q_i \sum_{j} \frac{q_j}{r_{ij\mathbf{n}}}$$

フーリエ変換

$$U = U_{real} + U_{self} + U_{wave}$$

$$= \frac{1}{2}\sum_{\mathbf{n}}\sum_{i} q_i \sum_{j} \frac{q_j}{r_{ij\mathbf{n}}} \mathrm{erfc}\left(\frac{r_{ij\mathbf{n}}}{\eta}\right) - \frac{1}{\pi^{1/2}}\sum_{i} q_i^2$$

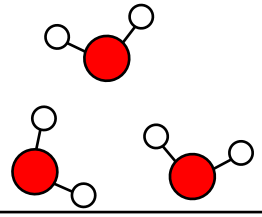
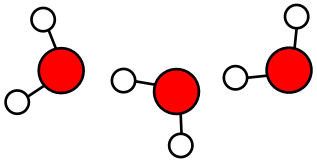
$$+ \frac{1}{2}\sum_{\mathbf{K}} \frac{\exp(-\pi^2\eta^2\mathbf{K}^2)}{\mathbf{K}^2}\left\{\left(\sum_{i} q_i \cos 2\pi\mathbf{K}\cdot\mathbf{r}_i\right)^2 + \left(\sum_{i} q_i \sin 2\pi\mathbf{K}\cdot\mathbf{r}_i\right)^2\right\}$$

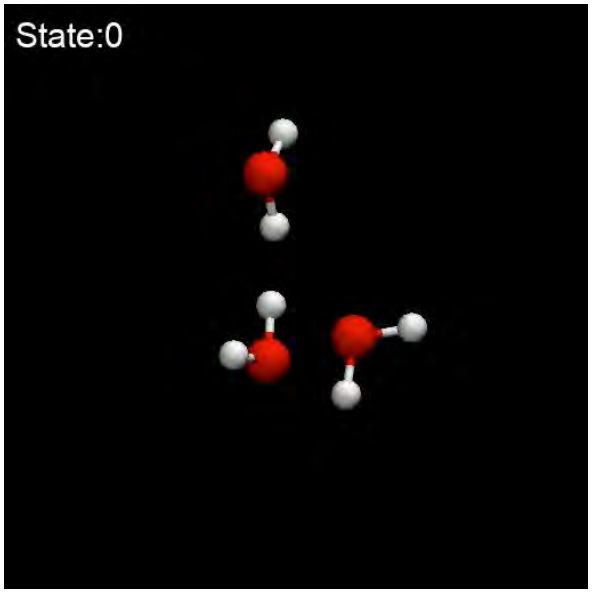
隠れマルコフモデルによるデータ同化

松澤　優太
（B4, 化学）

水分子クラスターの構造変化

- 水3分子の配置を4状態に分類

	状態	配置
	0	三角形
	1	直線状(水分子1が中心)
	2	直線状(水分子2が中心)
	3	直線状(水分子3が中心)

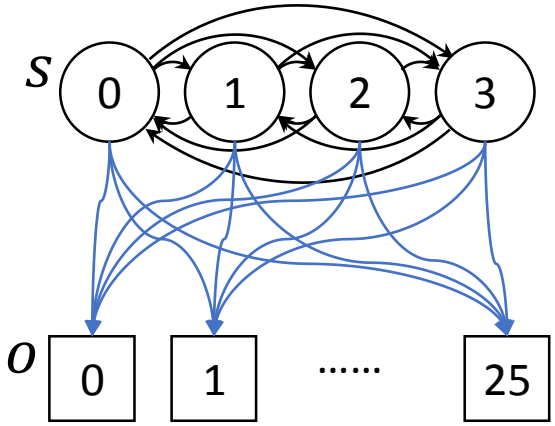


- 状態を隠れ変数、重み付きダイポールモーメントを観測値とする隠れマルコフモデルで系を表現

$$\begin{bmatrix} P(s_{k+1}=0) \\ P(s_{k+1}=1) \\ P(s_{k+1}=2) \\ P(s_{k+1}=3) \end{bmatrix} = \begin{bmatrix} .876 & .127 & .127 & .128 \\ .041 & .873 & 0 & 0 \\ .041 & 0 & .873 & 0 \\ .041 & 0 & 0 & .873 \end{bmatrix} \begin{bmatrix} P(s_k=0) \\ P(s_k=1) \\ P(s_k=2) \\ P(s_k=3) \end{bmatrix}$$

$$P(o_k = n) \equiv P(5n \le \mu < 5(n+1)) = h(o_k|s_k)P(s_k)$$

重み付きダイポール $\mu = \sum_i^N q_i \boldsymbol{r_i} \exp(|\boldsymbol{r}_\mathrm{i}|^2/30)$



- 電荷1.5倍の“変な水分子”1個を含む系を現実系とし重み付きダイポールモーメントの時系列データからその遷移行列を求める

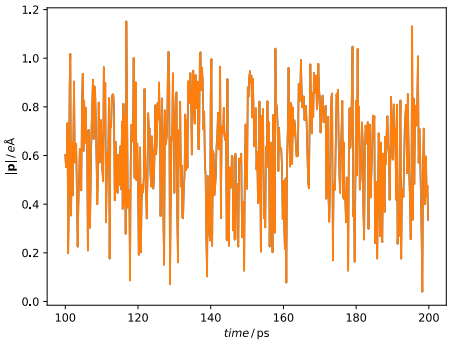
MDシミュレーション系
普通の水3分子

遷移行列

$$\begin{bmatrix} .876 & .127 & .127 & .128 \\ .041 & .873 & 0 & 0 \\ .041 & 0 & .873 & 0 \\ .041 & 0 & 0 & .873 \end{bmatrix}$$

$h(o|s)$

重み付き
ダイポールモーメント



現実系のダイポールモーメントを
再現するように
遷移行列を最適化
（データ同化）

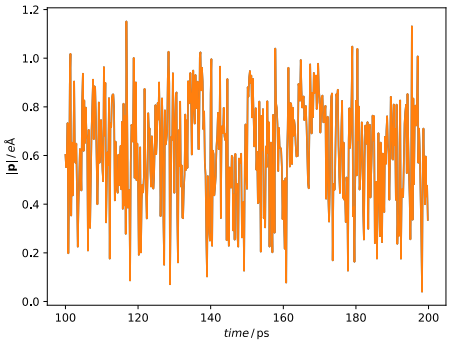
現実系
変な水を含む水3分子

遷移行列

$$\begin{bmatrix} & & & \\ & ? & & \\ & & & \end{bmatrix}$$

$h(o|s)$

重み付き
ダイポールモーメント

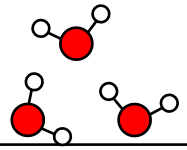
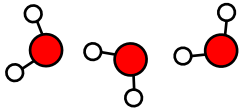


結果

右式を最大化する

$$P(o_1, \dots, o_N) = \sum_{s_1 \dots s_N} P(s_1, \dots, s_N, o_1, \dots, o_N)$$

$$= \sum_{s_1 \dots s_N} \underbrace{h(o_N|s_N)}_{\text{観測量}} \underbrace{P(s_N|s_{N-1})}_{\text{遷移確率}} \cdots \underbrace{h(o_1|s_1)}_{\text{観測量}} \underbrace{P(s1)}_{\text{遷移確率}}$$

	状態	配置
	0	三角形
	1	直線状(水分子1が中心)
	2	直線状(水分子2が中心)
	3	直線状(水分子3が中心)

- 0→1, 2→0, 3→0遷移の確率が増加
- 0→2, 0→3遷移の確率が減少

最適化前（MD系）

	0	1	2	3
0	.876	.127	.127	.128
1	.041	.873	.0	.0
2	.041	.0	.873	.0
3	.041	.0	.0	.873

最適化後

	0	1	2	3
0	.756	.129	.202	.202
1	.171	.867	.016	.016
2	.037	.002	.783	.0
3	.037	.002	.0	.783

現実系

	0	1	2	3
0	.802	.158	.180	.183
1	.164	.842	.001	.001
2	.018	.0	.819	.0
3	.018	.0	.0	.816

現実系の遷移行列の特徴を取り込むことができた