水の計算物理学とデータマイニング

松本 正和 (岡山大学大学院自然科学研究科分子科学専攻理論化学研究室)

1. データマイニング概説

1. 正確さと複雑さ

シミュレーションとは、現実のできごとを模倣することである。なぜ、現実に起こることを、わざわざコンピューターの中で再現する必要があるのだろう。

計算機シミュレーションの本来の目的は、解析的な計算ができない、複雑な積分 (配置積分、多変量の微分方程式の求解) をすることである。計算機は、当初は弾道計算のために開発され、その後高級言語Fortranが開発されるとほぼ同時に分子シミュレーション (はじめはモンテカルロ計算、のちに分子動力学や量子計算) にも使われるようになった。どれも計算内容は数値積分である。これにより、分子の運動を再現できるようになり、化学・物理の目的に使えるようになった。

現在の分子シミュレーションの目的は、実験結果を再現しつつ、実験では見えないものを観察することにある。では、実験で見えないものとは何だろう。著者の考えでは、それは次のようなものである。

(1) 観測装置の時間・空間的解像力を越える情報

(2) 実験では実現不可能な、あるいは非常に難しい条件下の現象

(3) 多自由度系に特有の (いや、むしろ例外的にシンプルな系を除いて普遍的に存在する) 、複雑性。

(1) は、計算の精密・正確さが重要になる。(2) は、どうやって検証するかという問題は残るが、実験に先立って、道標を立てることに意義がある。(3) は、正確に再現することはともかく、いかに法則性を見付けだし、解釈するかが重要になる。今回の講義の対象は (3) である。

2. 理論駆動と実験 (データ) 駆動

新しい、あるいは面白い物理を探しあてるにはどうしたらいいだろう。既存の理論を組みあわせ、新しい物理を探す、という理論駆動 (頭駆動) の方法と、とりあえず実験やシミュレーションをやってみて、結果の中から面白いものがないかを探す、実験駆動 (データ駆動、あるいは手駆動) なアプローチがあるのではないかと思う。誤解をしては困るのだが、後者には理論が要らないわけではない。理論をよく知った上でないと、むやみに実験しても、得られたデータが面白いのかどうか、つぎにどういう手を打てばいいのかがわからないからだ。どちらが良いということもないが、複雑系の現象は、理論駆動だけでは全貌を把握できないだろう[1]。

3. モデル化とスコープ

実験に比べると、シミュレーションでは、膨大な情報を得られ、短時間、近距離の現象はとても正確に近似できる。しかし、膨大な情報はそのままでは理解できないので、情報を減らして、見通しを良くする必要がある。

---

[1] 「(複雑系の研究では) 新しい法則の発見なしに新しい現象の発見がありうる」 (田口善弘他、「複雑系のキーワード」 p. 14、共立出版 (2000).

この、情報量を削る作業をモデル化という。多数の自由度のなかで、なにを重要視し、なにを捨てるか、ということ、言い換えるなら、スコープを定める作業である。実験の場合には、情報の取捨選択の自由度はそれほど広くないが、モデルをあてはめ、そのスコープでものを見るという点に違いはない。

単一のスコープでの観察だけでは、「群盲象を撫づ」になってしまうが、スコープの選び方にはいろいろあり、スコープを変えると違うように見えてくる。一つの現象をいろんな人が異なるスコープで観察し、コンシステントな全体像を作りあげていく。同じ問題でも、スコープが違えば答も違う。「鳥はなぜ飛べるか」という問いに対し、鳥の筋肉の性能に注目する者もあれば、翼の流体力学を解こうとする者もいるだろう。多様性があるからこそ研究は面白い。

ただし、スコープは全く自由に選べるわけではない。研究対象に対し、不適切なスコープで観測すると、大事なことを見落したり、複雑で理解できなくなる。あいまいな問いは、さまざまな研究者をインスパイアするという点では良い問題だが、すぐに解答にとりかかる前に、最善なスコープを選んで、問題を再定義しないと、泥沼に陥いる危険性もある。

モデル化とは近似であり、そこには人の主観が入っている。モデル化により、複雑な系を限られた変数で簡単に表現することで、モデルの適用範囲をせばめ、正確さを失うのとひきかえに、人間にとってのわかりやすさ、予測能力をもたらす。複雑なシステム、複雑なこの現実世界は、たくさんのモデルを集積することでしか理解できないのではないか、というのが、現代の複雑系研究者の共通の認識である。これを、ある人は「21世紀の科学は博物学に戻る」と表現し、ある人は「現代版の百科全書」と表現する。

4. 問題設定は直感が頼り

ある人は、タンパク質が数千分子の水にひたった状態を計算機内に再現し、超長時間のシミュレーションを行い、そこからタンパク質のゆっくりした運動をとりだして、実験で知られている事実と照合する。その時彼は、タンパク質1つの、マイクロ秒程度の動きにスコープをあわせ、ほかの自由度、速い運動、もっとミクロな情報には注意を払わない。人によっては、水を連続体で近似し、溶媒の分子運動を全く観察しないで捨ててしまう。

一方、水を研究している者から見ると、数千分子のマイクロ秒の運動のシミュレーションは、水のダイナミクスの得難い情報である。これをむざむざ捨ててしまうのはなんともももったいない。また、タンパク質の機能やフォールディングに水が非常に重要な役割を担っているのなら、タンパク質にスコープをあわせてしまうと大事な情報をとりこぼす。

彼がなぜタンパク質の遅い運動に限定して情報を見ようと思ったかといえば、遅い運動がタンパク質の機能に関連する、という仮説を置き、それを解明することを目標に設定したからであろう。研究者は、過去の研究と照らし、手持ちの研究資源 (時間、装置、人員など) を勘案し、あるいは次年度の科研費の申請書にどんな話を書くかを考えて、いくつかの仮説の中から、時間内に解けそうで、信憑性のある仮説を選んで、研究資源を投入する。

しかし、スコープの設定が良くなければ、良い解析結果はえられない[2]。いろんなスコープを試せばいいじゃないかと言うかもしれないが、選択肢があまりに多いと、まぐれ当たりを期待するしか方法がなくなる。生体分子ほど複雑な対象になってくると、どういうスコープを選ぶのが、コストと成果に照らして最良

[2] タンパク質の遅い運動にスコープを向けるのが適切かどうか、筆者はそれを批判できるだけの知識がありません。単なる例え話です。

かという、一段メタなレベルの問題に何らかの方法で解を見付ける必要がでてくる。今は計算機の性能も大したことないし、この問題を実質的に解く方法もないので、まぐれを狙ってえいやっとテーマを選ぶのが最良解かもしれない。計算機の速度が年々指数関数的に[3]向上しているなかでは、研究者がスコープを選び、それを計算機で検証する、というサイクルで、ボトルネックになるのは、実は研究者の時間になりかねない。それでも、単一の仮説を選んで、そこにリソースを集中させる方法がうまくいくのは、仮説の数が少なく、同じ問題に異なる視点からとりくむ研究者 (とそのもとで研究する学生たち) の数がそこそこいるから、アンサンブルとしてはいろんな仮説が検討され、検証されるからだろう。

しかし、研究の世界にも流行はあるので、同じような仮説に多数の研究者がしがみつくこともある。また、研究テーマの数が増え、一つのテーマに関与する研究者が少なくなる傾向もある。仮説の数よりも研究者の数が圧倒的に少なく、手が回らない分野もある。

5. 問題探索ツールとしてのデータマイニング

仮説を立てる部分を、研究者のひらめきだけに頼るのではなく、コンピュータに支援させられるのではないか? 例えば、タンパク質の問題で言えば、遅い運動が機能に関連するはず、という前提を捨て、オープンマインドで、全原子の全時系列情報を、様々なスコープで眺めれば (人間には、6N次元の位相空間の情報を「眺める」のは無理だが、コンピュータなら可能だろう)、もっと重要な情報をとりだせるのではないか?

全原子の全時系列情報という鉱脈の中から、平凡なデータを捨て、金の鉱脈を見付けだす作業を、機械化してしまおう、というのが、分子系のデータマイニングの発想である[4]。

電動アシスト自転車が人の行動範囲を拡げるのと同様、データマイニングは、人間の「気付き」をアシストし、知的探索の範囲を拡げる。ただし、自転車に乗るのが、目的地へ速く楽にたどりつくためであるのと同じように、データマイニングは、新しい物理を見付けだすための手段にすぎない。本当に「わかる」、つまり、背後にあるからくりを表現する部分は、人間にしかできないことに注意してほしい。データマイニングは、答を見付けるための道具というよりは、**問題を見付けるための道具**というべきだろう[5]。

6. 発見科学

データマイニングは発見科学とも呼ばれる。発見科学と書くと聞こえは良いが、英語ではHeuristics = 経験則という意味もある。つまり、規則性は見付かっているが、理論的な裏付けは弱いことを指す。

良い問題 (テーマ) の条件は、まだ誰もその問題に気付いておらず、答に広い応用可能性があることだが、もう一つ、「一定期間内に、なんとかして答に到達できる」という経済的条件も、現実にはとても重要である。発見的な研究手法 (データ駆動的な研究方法) だと、所定の期間内に答を (というより問題を) 得られるかどうか見当がつかない。これは修士課程・博士課程の学生にとっては深刻で、結果として、こじんまりし

[3] 15年で1000倍ぐらい速くなります。

[4] 論文に、「こういう条件でこんな観測をしたら、こんな結果が得られた」と書くよりも「すべての条件で探索した結果、こんな一般則が得られた」と書くほうがインパクトがあるのは言うまでもない。問題発見を自動化することで、人間の仕事は、良い結果が得られそうな特定の条件に目星を付けることから、条件空間全体を均一に探索できる、探索手法を開発する、という一段階メタなレベルに変わる。

[5] 実際、数学的予想を自動生成するシステムが開発されています (S. Fajtlowicz, On Conjectures of Graffiti, Discrete Mathematics, 72 (1988) 113–118) 。

た (安全な) テーマに流れてしまいがちである。データマイニングを身につけることで、冒険的なテーマ選びができる (といいなあ) 。

一番良いのは、データの中に誰もが見落していた関係性が潜んでいて、それに気付けば、答はデータを見れば一目瞭然、というケースである[6]。これも、データマイニングの技術が手助けしてくれるはずである。

7. 複雑系とデータマイニング

どんな対象にでもデータマイニングが使えるわけではないし、使うとより良い結果が得られるわけでもない。水溶液内の化学反応の反応式を例に考えてみよう。水が反応に関与していないなら、いかに水分子が多数存在していても、反応物だけに注目すれば反応機構はおおよそ理解できる。地 (客体) と図 (主体) が明確に分かれているので、スコープは図のほうに向ければいいのである。

これに対し、水の中のプロトン移動反応や、水の相転移、あるいは生体分子の機能発現では、客体と主体の区別はあいまいで、多数の水分子が反応に関与する。ある場所で起こっているイベントが、いつどこに影響を及ぼしているかを、あらかじめ予測することができないし、実際にシミュレーションで可視化しても、何が起こっているのかが読みとれない。こういう問題には、データマイニングが役立つかもしれない[7]。

8. 反証可能性とシミュレーション、精密化と粗視化

Wikipediaによれば、ある仮説が反証可能性を持つとは、その仮説が何らかの実験や観測によって反証される可能性があることを意味する。平易な意味では「どのような手段によっても間違っている事を示す方法が無い仮説は科学ではない」（科学が覆されるのは科学のみ）と説明される。

シミュレーションで得られる情報の多くは、実験で確認することができない。また、人為的なモデルを使った計算結果は、対応する実験ができない。従って、シミュレーションの結果から導かれる仮説は、実験的に反証できないものになるかもしれない[8]。

では、シミュレーション屋は、できるだけ正確で精密な計算を心掛け、実験観測で反証可能であることが明らかな仮説しか述べてはいけないか?筆者はそれは狭量すぎる考えだと思う。

第一に、現在は反証する手段がないからといって、将来も不可能であるとは限らない。

第二に、直接的に実験で確認できなくても、過去の様々なシミュレーションと実験と理論に照らして、仮説が妥当かどうかは、ある程度推論できる。

第三に、現時点では反証できない仮説を礎として、さらに枝葉の仮説を導くこともできる。この場合、最初の仮説が誤っていれば、その先の仮説も共倒れになるが、枝葉が増えることで、反証する手段も増えてくる。

---

[6] 「データ自身が明確に語ることを尊重すれば、自ずと意味が浮かび上がるということだ。」 (田口善弘、数理科学2006-9「ランダムネス」、サイエンス社)

[7] 「このような大規模なシミュレーションではもはや結果を予測することは難しい。むしろ予測できるような結果を出すのでは面白くないのである。」 (木村英紀、数理科学1998-9「モデルとモデリング」、サイエンス社)

[8] シミュレーションの論文を書いて、実験家がレフェリーになると、しばしばこのような批判を受けます。

実験値に合わせるために、精緻なシミュレーションを行うことも時には必要だが、もっとラフなモデルを使って、シミュレーションから仮説を導き、そこにさらに枝葉を発展させることも重要である。主張が正しいかどうかは、歴史が裁定してくれる。

複雑系では、一つの法則があてはまる範囲は狭いので、いくつもの法則をつぎあてて全体像を描かざるをえない。だから、モデルをどんどん単純化し、あるいはパラメータをいじって、どこで法則性が破綻するかを知ることは、モデルを精密化する以上に重要であると筆者は考えている。

9. 情報の構造を捉える

シミュレーション結果からとりだせる情報には、いろんな階層がある。ある情報は個々の分子の属性であり、ある情報は、いくつかの分子の集団で定義される。またある情報は、スペクトルの形で得られる。

情報の関連性を吟味するためには、これらの情報が、できるだけ明確な構造 (定常分布、連続値か離散値か、上限下限、次元、データ値の偏在、などなど) をもっていたほうが取り扱いやすい。特に、情報が多次元量の場合は、前処理が必須である。

例えば、人の顔かたちが、ほかの情報 (年齢、既往症、など) とどう関連しているかを検討するなら、顔写真のピクセルデータを使うよりは、前処理で目鼻口の特徴ごとに**分類**しておいたほうが、有意な相関を得られるし、あとの処理も少なくてすみ、解析結果に意味付けしやすくなる。

前処理によって単純化されたデータの間に、**関連性**があるかどうかは、相関係数や相互情報量といった距離の指標で測れる。そして、それらの量にもとづいて、類似なデータを集めたクラスターを作り、さらにそれらの大分類を作る。それぞれの階層は独立であれば言うことはないが、実際には人間が階層を切っているわけだから、階層の間にも関連性があり、階層の切り分け方には任意性があるし、階層に分けないほうがいい場合もある。ともあれ、複雑で大量のデータを解釈する過程で、**類似度 (関連性、距離) の計量**と、それにもとづく**分類**という作業が繰り返し発生する[9]。前者については、極めて強力で一般的な尺度が最近提案されたので、次節で簡単に紹介する。後者は、対象となる情報の構造に依存し、かつ研究者の主観・直感に依る面が大きい。

10. 関連性を測る

2つの変数の関連性とは、片方の値を知ることで、もう片方の値をどれだけ正確に予測できるかと言い換えられる。関連性がないなら、それらは独立であるという。関連性の尺度として、一番身近で手軽な方法は、相関係数を求めることであろう。→Wikipedia: 相関係数

相関係数は2つの変数の間に比例的関係があるかどうかを計量する。相関係数 (相関関数) は物理で広く使われているが、それはあらかじめ近似的に線形関係がなりたつのがわかっているから使えるのであって、比例関係がない情報にこれを適用しても、関連性は捉えられない。また、実用的な問題として、例えば線形目盛の場合と対数目盛の場合でも結果が変わってきてしまうし、数値でないもの (塩基やアミノ酸の記号列など) の関連性も測れない。

---

[9] 「分かる」は「分かれる」の古語です。情報の分野では、分類することを学習と呼ぶことからもわかるように、分類し名前をつけることは、物事を「分かる」一番基本的なやりかたです。

より一般的な方法として、相互情報量がある。これは、2つの確率変数の間の従属性を測る指標である[10]。
→Wikipedia: 相互情報量

相互情報量$I(X, Y)$は、2つの確率変数の間に何らかの関連性がある (独立でない) なら、線形な関係でなくても0よりも大きい値になるという点で、相関係数よりも汎用的な指標である。また、記号列にも適用できるという強みがある。一方で、連続量同士の相互情報量を計算する場合には次の問題が生じる。連続値から情報量を計算する場合には、必ず区間分けをして離散値に変換する必要があるが、区間分けのやり方には特に決まりがないにもかかわらず、区間分けのしかたによって、情報量が変化してしまうのである。

これを解決したのがReshefらによる最大情報係数 (MIC) [11]である [1] 。彼らは、区間分けに依存しない形で相互情報量を定義する方法を考えた。可能なすべての区間分けの方法を網羅的に試し、その中で最も相互情報量が大きくなるような区間分けを探すのだ[12] 。新しい論文なのでまだ応用事例はないが、相互情報量の弱点を克服したうまい手法なので、今後いろいろな分野のデータマイニングに利用されるだろう。

相関係数を汎化して、相互情報量や最大情報係数が導かれたのと同じように、時間差のある相関係数 (時間相関関数) を汎化したものが輸送エントロピー[13]である [2] 。輸送エントロピー$I(X_{i+1}, Y_i | X_i)$は3つの情報がからんでいるのでわかりにくいが、次のような意味の量である。x_iを時刻iでの物理量Xの値、x_{i+1}はその次の時刻 (時刻は離散化されているものとする) の値であるとする。また、y_iは時刻iでの物理量Yの値としよう。一般に、x_iとx_{i+1}は関連性があるので、x_iを知ればx_{i+1}をより正確に知ることができる。これを相互情報量で表現するなら、$I(X_{i+1}, X_i) > 0$ と書ける。ここで、もし別の物理量y_iとx_{i+1}の間にも関連性があるなら、y_iを知ればより正確にx_{i+1}を予測できる。この、y_iを知ることで、x_iしか知らない場合に比べてどれだけx_{i+1}を余剰に正確に知ることができるかを定量化したものが輸送エントロピー$I(X_{i+1}, Y_i | X_i)$である。

Y_iからX_{i+1}への輸送エントロピー$I(X_{i+1}, Y_i | X_i)$と、逆にX_iからY_{i+1}への輸送エントロピー$I(Y_{i+1}, X_i | Y_i)$は全く別の量である。Yを知れば未来のXがより正確に予測できるからといって、Xを知れば未来のYが同じように正確に予測できるとは限らない。

輸送エントロピーも条件付きの相互情報量の一種なので、相互情報量と同じ弱点を持つ。MICの発想を予想エントロピーに適用すれば、時間のずれのある関連性を見付けだす、汎用性のある解析手法になるだろう。

11. 因果関係

時間のずれを含む相関やそれに類するもの (e. g. 輸送エントロピー) は、因果関係 (「AならばBである」) をさがすのに役立ちそうに思える。しかし、観測データから因果関係に言及する場合は、かなり慎重になるべきである。

事象Aの結果として、事象Bと事象Cが起こるとしよう。この場合、AとB、AとCの間には因果関係がある。一方、観測にひっかかる量は、BとCだけで、Aは観測できないものとする。観測できないAが起こると、BとCが多少の時間のずれをともなって起こり、これらの間の時間相関や輸送エントロピーは大きな値になる

[10] 3変数以上の場合にも、一般的に相互情報量を定義できます。また、確率と同じように、同時相互情報量、事後相互情報量も定義できます。

[11] Maximal Information Coefficient

[12] 相互情報量は、区間の幅のとり方だけでなく、個数にも依存するので、区間数から求められる最大相互情報量であらかじめ規格化した上で、区間の幅を調節します。

[13] Transfer Entropy

(Bが起こったことを知れば、すこしあとにCが起こることを予測できる)。だからといって、BとCの間には因果関係はない[14]。同じように、もしAが観測できたとして、AからBへの輸送エントロピーが非常に大きいとわかったからといって、Aの背後に、真の原因があって、AとBはどちらもその結果かもしれない。結局、観測量だけでは、因果関係があるかどうかはわからないのである。

おそらく、外的要因 (明らかに外部に由来する原因、例えば、人間が観測系に干渉する、など) があってはじめて、因果関係を見付けだせる。つまり、因果関係に言及したい場合は、積極的に実験対象に干渉する必要がある。

論文を書くときに、観測データにもとづいて、「Aの結果Bが起こる」、というストーリーを書きたくなった時は、少し立ちどまって、本当に因果関係であることを示せているかを考えよう。

2. 水

与えられた時間でデータマイニングの一般手法を網羅的に語るのは不可能だし、具体的な対象なしに、抽象的な説明をしてもわかりにくいので、水のこれまでの研究で、どのような手法を利用し、あるいは開発してきたかを紹介する。

1. 水の異常性、多形、polyamorphismの起源

水の百科事典であるMartin Chaplinのサイト[15]によれば、水には60を越える異常性がリストされていて、そのリストは今も伸び続けている。しかし、それらすべてが独立というわけではなく、同じ異常を異なる物性で観測しているものも多数ある。これらを大雑把に次の数種類に分類してみた。

1. 相互作用 (水素結合) の相対的な強さに起因する、表面張力、高い沸点、高い融点などの物性。
2. プロトンが軽く、解離できることに起因する、プロトン移動に関連する物性。
3. 4配位のネットワーク構造を作る性質に起因する、結晶多形の多様性など。
4. 融点付近〜過冷却の水の異常性。
5. その他 (ムペンバ現象[16]など)

1の、気液界面の界面エネルギーが強く、負圧でも容易にキャビテーションを起こさない性質は、樹木が水を吸いあげるために不可欠である。2は、燃料電池や生体分子でのエネルギー・信号伝達で重要である。3は宇宙の様々な環境にある水の性質を理解する上で重要な性質であり、5は未解明なものもいくつか残されている。そして、4が、常温以下の水の性質を理解する上で最も重要であり、実際、Chaplinのリストの最多項目である。

以下では、主に4について説明するが、その前に3について述べておく。

---

[14] 地震の直前にナマズが暴れるからといって、ナマズが地震を起こしているわけではないが、ナマズを観察していれば、より正確に地震が予測できる。

[15] http://www.lsbu.ac.uk/water/

[16] 特定の状況下では高温の水がより低温の水よりも短時間で凍ることがあるという物理学上の主張。必ず短時間で凍るわけではないとされる (Wikipediaより引用)。

水には10種類以上の異なる結晶構造 (多形) が知られている。準安定相も含めると、さらにその種類は増えると思われる。凝縮相 (氷、液体) では、水分子同士は水素結合で結びついている。我々が日常目にする氷は氷Ih (六方晶氷I) と呼ばれており、それぞれの水分子は周囲の4つの水分子と、4本の水素結合を形成している (図1)。そのうち、2本は中心の水分子が供与する水素結合、残り2本は周囲の水分子から受容する水素結合である。これをBernal-Fowlerのアイスルールと呼ぶ。氷Ihはプロトンディスオーダ氷とも呼ばれ、酸素原子位置はウルツ鉱構造の格子点にある一方、水素結合の向きはおおよそランダムになっていると考えられる。氷Ihを絶対零度付近まで冷却すると、水素の配向のランダムネスを反映して$k_B \ln(3/2)$ の 残余エントロピーが観測される。

すべての氷の多形はアイスルールをみたす。つまり、どんなに高圧であっても、氷の作るネットワークは次数4である。次数4という制約のもとで (いや、制約があるからこそ)、氷は様々な結晶構造に変化する。

たまたま安定相となるような、結晶構造 (観測できる構造) だけでも多数ある一方、局所的には安定だが周期性のない構造や、準安定結晶構造も含めると、局所的安定構造はさらに多様であると推察される。このことを念頭に、低温の水の性質を見ていこう。

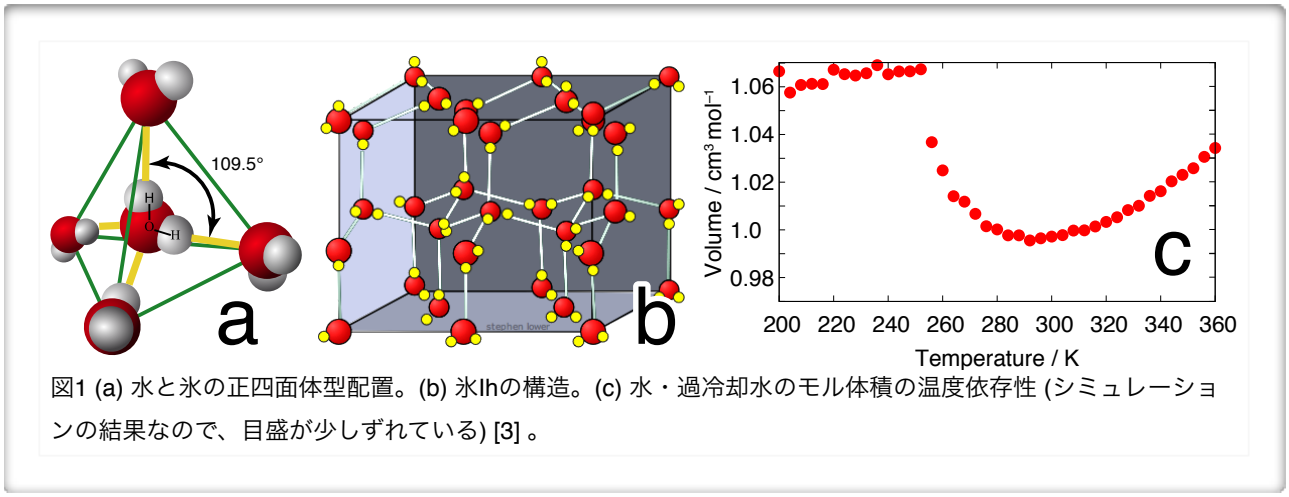


図1 (a) 水と氷の正四面体型配置。(b) 氷Ihの構造。(c) 水・過冷却水のモル体積の温度依存性 (シミュレーションの結果なので、目盛が少しずれている) [3]。

通常の物質は、温度を下げると熱振動が抑制されるため、体積が小さくなる。水も同じように冷やすにつれて収縮するが、4°Cで密度が極大となり、それ以下では膨張しはじめる。この膨張は、低温で水の構造 (平均的な分子配置) の変化が起こっていることを示唆している。0°Cでの凍結を回避し、過冷却すると、水は氷点下でも膨張し続ける。シミュレーションにより、過冷却での温度密度曲線を描くと、融点より40Kぐらい下まで急激に膨張し続ける (図1)。

この膨張の原因は、温度が低くなるにつれ、水分子が形成する水素結合ネットワークの結合角 (隣接する3分子の水が作る角度) が、氷の109.5度に近づくとともに、水1分子あたりの水素結合の本数も氷と同じ4本に漸近することにあると考えられている。つまり、過冷却水の構造は、常温の水の熱揺らぎを単純に小さくしたものではなく、ネットワークの構造自体が、より密度の低い別の構造に変化したものとみなせる [4]。もし、結晶化を回避して過冷却を続けると、水は低密度なアモルファス氷 (LDA) になると推定される。実験では、バルク水をゆっくりと過冷却してアモルファス化させることには成功していないが、急冷など別のルートで、LDAを得られる。

一方、より高圧 (3000気圧以上) で同じように液体の水を過冷却すると、温度を下げても膨張することができず、常温の液体に近い構造を保ったまま、水の体積は収縮し続け、ついにはアモルファス化すると推定される。こうして高圧下で得られるアモルファス氷は、低圧で膨張をともなって形成される低密度アモルファ

スと区別するため、高密度アモルファス (HDA) と呼ばれる。なお、どちらのアモルファス状態も、準安定であり、同じ温度圧力での最安定相は、通常の氷Ihである。

高圧でできるアモルファスが高密度で、低圧でできるアモルファスが低密度であっても特に不思議はないのだが、水の場合、低温でこれら2つ準安定相の間に一次相転移が起こることが、三島らにより発見された [5] 。つまり、水を過冷却した上で、圧力をすこしずつ加えると、アモルファスの密度は連続的に変化するのではなく、ある圧力で不連続に収縮するというのである。準安定相であるアモルファス相が2種類あって、そのあいだに共存線が存在し、さらに共存線の末端には準安定臨界点があるのだ、その臨界点が生みだす揺らぎのせいで、臨界点に比較的近い、常温の水でも上掲のようないろんな特異な性質が生まれるのだ、だから、準安定の水の性質を徹底的に調べ、様々な特異な性質を統一的に説明しよう、という考えが、ここ20年ほどの、水の研究で最もホットなトピックだった [6] 。

高密度アモルファス氷では、高い圧力のために、水素結合ネットワークは歪み、あるいは切断されて、次数が4ではない (4配位ではない) 水分子が多数生じていると考えられている。一方、低密度アモルファス氷では、ほぼすべての水分子が、氷と同じく次数4になっているが、長距離秩序は失われた、Continuous Random Network (CRN) [17] と呼ばれるネットワーク構造になっていると考えられる [7] 。

では、高密度アモルファス氷の構造の乱れ方と、低密度アモルファス氷の構造の乱れ方は、どう違うのだろう。乱れの大きさは、エントロピーで測れる。熱測定により、確かに低密度アモルファス氷のほうが秩序が大きいことは確認できる。しかし、ここでいうアモルファスの「秩序」とは何を意味するのか。

水と油を混ぜた場合には、水は水同士、油は油同士引きあうことで、相分離が起こる。この場合、相分離は個々の分子の性質 (相互作用) によってひきおこされ、分子配置に反映される。シミュレーションで分子スケールで観察しても界面は明確だし、分離する理由を明快に説明できる。しかし、水と水が相分離する場合にはこの考えは通用しない。相分離を、個々の分子の性質に帰着することはできないのである。4配位の水と、4配位でない水が存在するとしても、そのことが直接相分離に結びつくことはない。常温の水の中にも、4配位の水とそうでない水は多数混在するが、それらはほぼ均等に混ざっている。一方、過冷却水の中では、4配位の分子同士が凝集する傾向はシミュレーションで確認されている。なぜ低温になると、4配位の水は4配位の水同士、そうでないものはそうでないもの同士集まってしまうのだろう[18]。

これを説明するには、配位数というローカルな情報だけでは不十分なのは明らかである。アモルファスの中に潜む、中距離構造の乱れ方を理解しなければいけない。

前述の通り、水は水素結合でつながりあってネットワークを形成している。水の構造を、ネットワーク = グラフで表現することが妥当であれば、構造を比較する場合にも、座標のような連続量を使って比較するよりもはるかに容易である。実際、氷の結晶構造は、水素結合ネットワークのトポロジーの違いで明確に区別できる。液体やアモルファスでも、温度が低い限りグラフとして取り扱って構わない [8] 。

---

[17] シミュレーションが行われるよりはるか昔に考案された、アモルファスシリコンの構造モデル。

[18] 「傾向・法則性を見付ける」ことと、「わかる」ことの違いを如実に示す良い例。4配位の水が凝集する傾向を見付けても、なぜそうなるかを説明できなければ、わかったことにはならない。データマイニングは、法則性を見付けだすのには役立つが、わかるには人間の解釈が必要である。

2. グラフと3次元構造

グラフは頂点を辺で結んだもので、つながり方 = トポロジーを表現する道具である。つながり方さえ変えなければ、辺を曲げたり伸ばしたりしても構わないし、2次元平面 (紙) の上に射影しても3次元で表しても相同だとみなす。一方、水のネットワーク構造は明らかに立体的な情報を含んでいて、結合の長さや角度といった幾何学的な制約を受ける。舟型六員環と椅子型六員環は時には別のものとして扱わなければいけない。

また、構造をグラフ/ネットワークとして表現する場合に、ネットワーク全体を取り扱うのは不可能なので、部分を切り出し、部分ネットワークの多様性を統計的に扱うのが現実的である。この時、どのように部分を切りだすかは任意性がある (例えば、ある分子からネットワークに沿ってn歩でたどりつける部分ネットワークを切り出す、リングをさがす、など)。うまい方法を見付ければ、結晶の種類を見分け、液体の中の局所構造の違いも識別できるようになるが、さもないと、構造の種類が多すぎて手におえなくなったり、逆に多様性が小さすぎると、結晶多形の間の構造の違いすら判別できなくなったりする。

このように、水素結合ネットワークの構造をグラフとみなす場合には、水素結合ネットワークの幾何学的情報が、グラフのトポロジーにうまく畳みこまれるように配慮が必要である。

例えば、球形分子の構造の問題を扱うなら、四面体パッキングと八面体パッキング (どちらも最密充填構造の要素) に対応する、四面体グラフと八面体グラフの組み合わせで、より高次の構造を表現すると良いだろう (Appendix A クラスレートハイドレートの構造解析)。

水の場合には、フラグメント (vitrite) と呼ぶ多面体的部分構造を使うと、水の局所構造の特徴を的確にとらえられる。ほかの物質の場合には、それぞれその系の特徴を捉える、うまい部分構造の定義を個別に考案する必要がある (Appendix B 過冷却水の構造解析)。

3. グラフの相同性と分類

技術的には、数万〜数億の異なる構造を識別し、数えあげることは、現在のコンピュータの性能であればそれほど困難ではない。特に、水のように、構造をグラフで表現できる場合には容易である。

あるグラフFを、頂点のラベルをてきとうにつけかえることで、別のグラフGと同一にできる時、FとGはグラフ同型 (graph isomorphism) と呼ばれる[19]。Fから定義されるある量$f(F)$ (例えばFの頂点の数、辺の数など) が、$f(G)$ に等しい時 (つまり、ラベルの順序をいれかえても変化しない時)、この量をグラフ不変量と呼ぶ。FとGが同型かどうかを直接比較するのは計算量を要するが、それぞれのグラフについて、あらかじめいくつかの不変量を計算しておき、先にそれらを比較すれば、無駄な計算をしなくてすむ。さらに、FとGが同型な場合に限り、$f(F) = f(G)$ となるような不変量を、グラフの正準ラベル (canonical label) と呼ぶ。正準ラベルはグラフのIDそのものなので、グラフデータベースのキーに使える[20]。

---

[19] graph isomorphismはNP完全な、組み合わせ最適化問題である。

[20] もし正準ラベルを多項式時間で計算できるなら、graph isomorphismも多項式時間で解けることになり、P = NP?問題が解決してしまう。逆に言えば、正準ラベルの計算もまた、NP完全だと思われる。実用的には「ほぼ」正準なラベルで十分である。水の部分ネットワーク構造を表すグラフの種類は、同じサイズ(ノード数)の可能なグラフの総数に比べればほんのわずかであり、その範囲内で異なる構造に異なるIDを振ることができれば (あるいは、同じIDを異なるグラフに振ってしまう確率が十分小さければ)、構造分類には事足りる。

分子動力学計算を実施し、時々刻々出現する構造の正準ラベルを計算し、データベースに照会して、新しい構造ならデータベースに追加することで、出現する構造の種類と出現頻度を網羅的に調べあげる。一旦データベースができてしまえば、異なる相では部分構造 (グラフ) の出現頻度が異なるので、部分構造の出現頻度を、構造を見分ける「指紋」として利用できる。このような、グラフを使ったデータマイニング手法をグラフマイニングと呼ぶ。

Appendix Cでは、溶融過冷却シリコンにおいて、結晶とは異なる局所安定構造が形成され、結晶核生成と競合する現象をグラフマイニングで見付けた例を紹介する。

4. 配置空間の構造

数分子からなるクラスタ系を考える。ある温度で長時間のシミュレーションを行った結果、このクラスタが、ほぼエネルギーの等しい3つの異なる構造の間をうつりかわっていることがわかったなら、あなたは、このクラスタには3つの典型的な (準) 安定構造がある、と描写することに躊躇しないだろう。

では、1000分子からなるバルクの水について、同じように超長時間の (平衡状態での) シミュレーションを行った結果、ほぼエネルギーの等しい10^{10}種類の構造の間を遷移していることがわかったなら、あなたは同じように「このシステムには10^{10}種の典型的な構造がある」と述べて良いのか?

もちろんそう表現しても問題ないはずである。しかし、シミュレーション以外の方法では、瞬時にうつりかわるこれらの構造を見分けることは不可能なので、これまでの慣例に従うなら、構造の多様性を数えて、「この系は1分子あたり $(\log 10^{10})/1000$のエントロピーを持つ液体 (ガラス、アモルファス) である」と表現するだろう。

液体の構造の多様性を、エントロピーにすべて押しこんでしまって良いのか、10^{10}種類を完全に分類すべきなのか、それともももうちょっと粗視化することで10^{2}種類ぐらいに分類したほうが良いのか、はたまた低密度液体と高密度液体の2種類に分けるべきなのか、と考えていくと、液体という概念がゆらぐ不安を覚えはじめる。これまで、液体を単一の状態と考えていたのは、「見た目にそう見える」「その方が取扱いに便利」以上の根拠があるのか? 2種類の液相を持つ水のような物質があるのなら、3種類の液体相がある物質だってあってもいいのではないか?

それ以外にも、分子数が少ない系や、非平衡系や、分子のダイナミクスに注目する時には、個々の構造を見分けることが非常に重要な意味を持つ場合がある (たとえば、均一核生成における最初の核の構造、水溶液における溶質周囲の水の配置、など)。例えば、水20分子からなるクラスタがとりうる “液体” 構造は、水素結合の方向性を無視して無向グラフとみなせば、おおよそ100万種類程度の多様性があると推定される [9]。100万種類程度の構造の間の遷移確率や反応経路を網羅することもできるだろう。相互に (単一のバリアを経て) 到達できる構造同士をつないで、反応面 (ポテンシャルエネルギー面) 全体をグラフで表現し、その連結性から、互いに到達しやすい構造の集団 (ベイスン) や、その中の副ベイスンといった階層的なポテンシャルエネルギー面地形を描くこともできる。この地形の特徴を、客観的かつ簡潔に表現する方法はないか? 水の地形と、球形分子の地形と、タンパク質の地形はどう違うと言えるだろう。その違いは、どのような性質を生みだしているだろう。

5. 構造の類似度と距離

ある構造が別の構造と似ているか、という、形の類似度の尺度は、形が運動に関連しているなら、重要な意味を持つ。例えば、水20分子のクラスターのうち、ある (比較的安定な) 構造は、ほかの多くの安定構造に「似ていない」ため、その構造に遷移する経路が少なく、めったに出現しない。[10] また、一旦その構造が実現されれば、非常に長い時間構造が維持される。ポテンシャルエネルギー面の地形で表現するなら、その構造に対応するエネルギー極小は、深く、孤立していると言ってよいだろう。しかし、比喩的にそう述べるのはともかく、本当に孤立しているということ (ほかの構造は孤立していないこと) をどうやって示せば良いか。言い換えれば、形が似ている、とはどういうことを意味するのか。

ある形と、別の形が似ているかどうかは、2つの形を重ね、ずれの大きさを調べれば計量できる。タンパク質の構造の研究では、RMSD (根平均二乗変位) という量が、構造の照合によく用いられる。これは、2つのタンパク質の、対応する原子の位置の間の距離の二乗の総和 (Qとしよう) が、最も小さくなるように、片方のタンパク質を平行移動 + 回転して得られる、Qの最小値のことである。

タンパク質の場合にはうまく使えるRMSDも、水クラスターのように、同一な分子の集合体の場合には、ひと工夫必要になる。水クラスターの場合、分子のラベルを入れかえても、幾何学形状は全く変化しないので、2つの構造を照合する場合には、ラベルの入れ替えをすべて試したうえで、RMSDを最小にしなければならない。これは、典型的な組み合わせ最適化問題なので、分子数が多くなると、厳密解を求めるのが困難になるが、近似解でもほとんど支障はない。

水の場合には、構造をグラフで表現できるので、もっと簡便な照合法が考えられる。片方の構造のグラフから何本辺をとりさり、何本追加すればもう一方のグラフに変形できるか (これを編集操作と呼ぶ) を数え、上と同じように、ラベルの入れ替えをすべて試した上で、最小限必要な編集操作の回数を求める。この回数のことを編集距離と呼ぶ。編集距離あるいは (ラベル入れかえを許した) RMSDにより、任意の構造の間の類似度を数値化できる。

2つの構造が似ているなら、その間の構造変化は比較的容易だと予想されるが、本当だろうか。形の類似性と、それらをつなぐ反応経路のバリアの高さと、構造変化の頻度はきちんと整合性があるだろうか。

Appendix Dでは、編集距離を利用した解析の例として、氷の融解の研究を紹介する。

6. おわりに

著者は、自分の必要に応じて、いろんな知識をかじっては利用している利用者の一人にすぎないので、総体的かつ観念的な話に終始してしまった。具体的な解析方法を、網羅的に紹介するには時間が足りないし、著者の知識も不十分であることをお詫びしたい。著者の知るかぎり、このタイプのデータにはこの解析をやれば良い、という決定的な方法はまだない。ただ、情報理論に基いたデータマイニングの論文を読んでいると、まだこんなことが見付かっていなかったのかと思うことは多い。輸送エントロピーが提案されたのは2000年、最大情報係数が提案されたのは2011年である。前者は実際には条件付き相互情報量という、昔から知られていた量に新しい意味付けを与えたものであり、後者は区間幅の選び方を最適化することで、もともとは絶対値の意味があいまいだった相互情報量を、他のデータでの値と直接比較しうる形に書きなおしたものとみることができる。どちらのアイデアも、わかってみれば、なあんだと思う人は多いだろう。データマイニングを使う前に、情報関連性を抽出する技術をまずマイニングすれば、未発見の有用な指標がまだまだ見付かるのではないか。

実際に発見的な方法を使った例ということで、後半では水のいろんな物性の解析を紹介した。筆者は、究極的にはシミュレーションでえられた分子運動のムービーを見ながら、実況解説ができればいいなあと思っている。講義や講演で分子の反応や構造変化を説明する際、模式図や、模式化された動画が使われることは多い。しかし、平均的な像の説明を聞きながら、生の分子運動の映像を見ても、納得できることはまずないのではないかと思うのだ。生の映像と模式化された像が異なって見えるのは、映像に騙されているだけかもしれないが、逆に統計処理の結果大事なものを消し去っている可能性もある。これらのギャップを埋めるには、より実情に近い情報を平均に埋もれさせない解析手法と、それを的確に表現する用語が要ると思う。

水に関連する部分は、大峰巌氏 (分子科学研究所) 、望月建爾氏 (総合研究大学院大学) 、田中秀樹氏 (岡山大学理学部) ほかとの共同研究による。

7. 参考書

1. 森下真一、宮野悟編、「発見科学とデータマイニング」、共立出版 (2000).
2. C. M. ビショップ、「パターン認識と機械学習」、シュプリンガージャパン (2008).
3. R. O. Duda他、「パターン識別」、新技術コミュニケーションズ (2001).
4. J. Gasteiger他編、「ケモインフォマティックス」、丸善 (2005).

8. 参考文献

[1] D. N. Reshef et al., Science 334, 1518 (2011).
[2] T. Schreiber, Phys. Rev. Lett. 85 (2), 461 (2000).
[3] http://www.chem1.com/acad/webtext/states/water.html
[4] M. Matsumoto, Phys. Rev. Lett. 103, 017801 (2009).
[5] O. Mishima, L. D. Calvert, and E. Whalley, Nature 310, 393 (1984).
[6] O. Mishima and H. E. Stanley, Nature 396, 329 (1998).
[7] W. H. Zachariasen, J. Am. Chem. Soc. 54, 3841 (1932).
[8] M. Matsumoto, J. Chem. Phys. 126, 054503 (2007).
[9] M. Matsumoto, unpublished data; C. L. Brooks, J. N. Onuchic, and D. J. Wales, Science 293, 612 (2001).
[10] D. J. Wales and I. Ohmine, J. Chem. Phys. 98, 7245 (1993) ; K. Nishio and M. Mikami, J. Chem. Phys. 130, 4302 (2009).
[11] M. Matsumoto, A. Baba, and I. Ohmine, J. Chem. Phys. 127, 134504 (2007).
[12] S. Sastry and A. Angell, Nature material 2, 739 (2003).
[13] J. -F. Sadoc and R. Mosseri, “Geometrical Frustration”, Cambridge University Press (1999).
[14] K. Mochizuki, M. Matsumoto, and I. Ohmine, unpublished data (2012).

9. Appendix

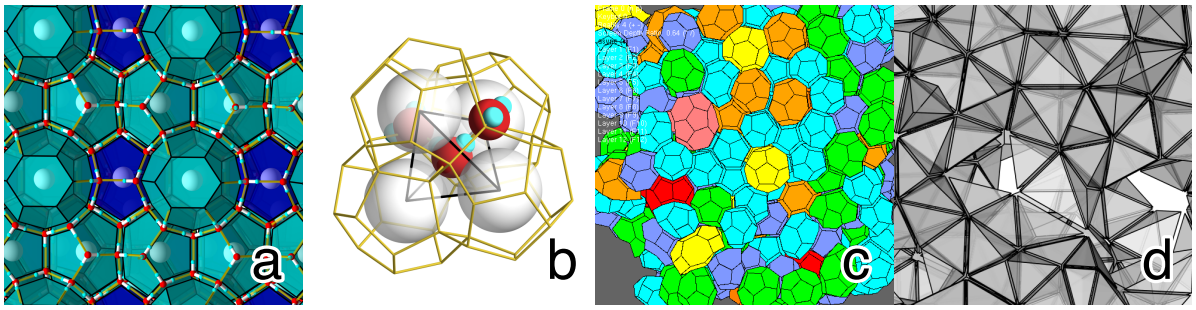
A.クラスレートハイドレートの構造解析

クラスレートハイドレートは、水が作るカゴ状の格子の中に、気体などの小分子 (ゲスト分子) が大量に捉えられた固溶体である。メタンを内包したものはメタンハイドレートと呼ばれ、昨今ではエネルギー資源として注目されている。クラスレートハイドレートでは、1つのゲスト分子を、20〜28分子の水が作るカゴがとりかこんでいると見る一方で、視点を変えてみると、1つの水分子を、4つのゲスト分子が囲む構造ともみなせる (図A (a, b)) 。クラスレートハイドレートの中には、気体分子が膨大に含まれているため、どの水分子も気体分子と隣接せざるをえないが、水と気体分子は弱い相互作用 (主にファンデルワールス相互作用) しかできないので、水はあくまで水同士で水素結合を形成したほうが安定になれる。前述の通り、水はすべての氷の多形の中で四面体ネットワークを形成しようとするので、気体分子と共存しつつ四面体ネットワークを

形成する方法として、4つの気体分子に囲まれたすきまに居場所を見付けるのである。こうすれば、水は気体分子のすきまから隣の水分子に水素結合の手を4本延ばせる。

実際、すべてのクラスレートハイドレート結晶構造では、水は隣接水分子と4本の水素結合を作りつつ、4つの気体分子 (あるいは空隙) が四面体型に配置している。気体分子だけに注目すれば、すべての気体分子は四面体パッキングになる。これは、気体分子だけが密集した場合にはとりえない構造である。球形に近い形状の分子は、密度が高くなるといずれも最密充填構造に近付くが、これは正四面体配置と正八面体配置が2: 1の比率で混ざった構造である。対して、クラスレートハイドレートの中の気体分子の配置は (水を無視すれば) 四面体配置しか存在しない (図A (c, d)) 。

クラスレートハイドレートの結晶化過程をシミュレーションすると、ゲスト分子が水分子1つを囲んで四面体配置でパッキングすると、ゲスト分子とその周囲の水分子の拡散が著しく遅くなるのがわかる。また、ゲスト分子が水を囲んだ八面体パッキングは全く出現しないこともわかる。このようにクラスレートハイドレートでは、局所構造と運動が密接に関係している。



図A (a) クラスレートハイドレートの構造。水の水素結合が作る多面体型のカゴの中に疎水性分子が捉えられる。(b) 4つのケージが1つの水分子を共有する。4つのガス分子が1つの水分子をとりかこむ。(c) 均一核生成の分子動力学計算により実際に形成されたメタンハイドレートの構造。多面体の面数ごとに色分けした。(d) 隣接する多面体の中心同士を線でつなぐと、四面体でうめつくされた構造が見える。

B.過冷却水の構造解析

LDA以外で、至るところ4配位で、かつランダムなネットワークを形成するもので、最も身近なものといえば、泡が挙げられる。理想的な泡は多面体でうめつくされた構造を持ち、多面体のすべての頂点では、4つの辺が互いに109. 5度で交わっている (図B (a)) 。

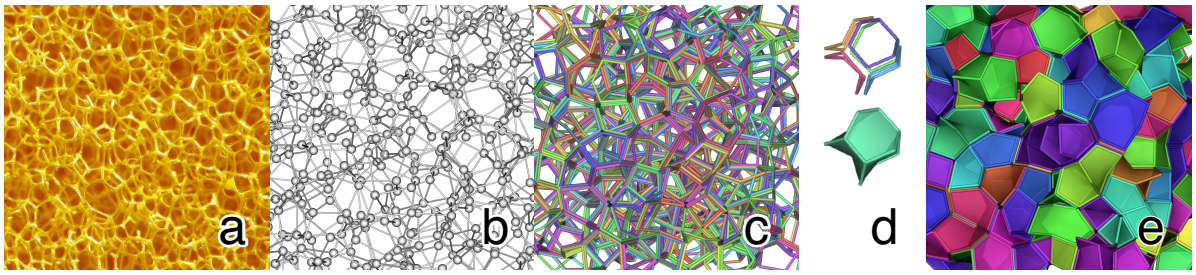
ただし、LDAのネットワークと泡では、制約条件が違う。泡の構造では、多面体を構成する面の総面積がエネルギーに比例するため、多面体の各面はほぼ平面になる一方、辺の長さはまちまちである。これに対し、LDAのネットワークでは、辺の長さに比較的強い制約があり、ランダムネットワーク構造であっても、辺の長さはほぼ等しくなっている一方、辺が作る環は、椅子型/舟型六員環にみられるように、平面ではない場合が多い。

このような制約の違いはあるものの、やはりこれらのネットワーク構造は似ている。そう思って見れば、LDAの構造は、多面体の集合体に見える。そこで、実際にLDAの構造を多面体に分割することを試みた [11] 。

まず、水素結合ネットワークをグラフで表現する。次に、グラフの中に存在する、8員環以下の環をすべて列挙する[21]。得られた環を組みあわせ、閉包を作る。この時、水のネットワークらしさを反映するために、次の3つの制約を課す: (1) 1つの辺は2つの環で共有される。(2) 1つの頂点は2つまたは3つの環で共有される。(3) Eulerの式 $f-e+v=2$ を満たす。こうして得られた閉包をフラグメントまたはvitriteと呼ぶことにする (図B (b-e))。

実際に、LDAのネットワークからフラグメントを構成すると、互いに重なることなく、フラグメントがほぼ全空間を埋めつくす (タイリングできる) ことがわかる[22]。つまり、LDAの構造は、フラグメントを構成単位とする泡のような構造を実際に持っていたのである。

同じフラグメントの定義を使って、クラスレートハイドレートや氷の結晶構造 (低圧) もフラグメントでタイリングできる。低温低圧の水は、液体も氷も泡のような構造を持っているということになる。クラスレートハイドレートがその泡の中央にゲスト分子を捉えられるのと同じように、高圧下では、氷Iや氷IIの泡の中にゲストが入った、filled iceというものもできる。このことから類推すれば、LDAの泡の中にゲストが入った、filled LDAも高圧では作りうる (ただし、おそらく準安定相) と思われる。



図B (a) キッチンスポンジの拡大像。スポンジは泡から膜をとりさった構造で、膜と膜が交わる辺だけが残っている。4つの辺が1点で連結する、次数4のネットワークである。(b) 過冷却水の水素結合ネットワークの構造。こちらも4つの辺が1点で連結する。(c) 水素結合ネットワークの作る環をすべて探し出し、違う色で彩色した。(d) 環を組みあわせ、Eulerの式をみたす、多面体的構造 (フラグメント) を組み立てる。(e) すべての多面体的構造を違う色で彩色した。

C.過冷却液体に生じる局所安定構造

水を過冷却すると、徐々にLDAに変化していき、その構造はアモルファスシリコンのモデル “CRN”に似ている、と述べた。シリコンもまた過冷却すると、水と同じようにHDA-LDA転移 (あるいは高密度液体と低密度アモルファスの転移) を起こすと言われている [12] 。このように、シリコンと水は、過冷却状態で互いにいろいろ似たところがあるため、水では探りにくい物性をシリコンで代わりに探ったり、あるいはその逆を行ったりすることは多い。

---

[21] 環の定義の仕方は、reducible (可約) とirreducible (既約) におおまかに分けられる。前者では、例えば六員環の対角頂点を結ぶ辺が存在する場合にも、4員環2つと6員環1つを数えるが、後者では六員環は数えない。ここでは既約な環のみを数えるが、それでも環の数え方の流儀はいくつもあり、数え方によって結果が違ってきてしまう。既約な環のみを数える場合、9員環よりも大きい環は、水のネットワークにはほとんど出現しない。

[22] 多少の例外はある。環を数える際、環の交差が生じる場合がある。また、LDAのネットワークといっても、ごく少数の3配位や5配位分子を含むため、その周辺では上の方法でうまくフラグメントを作れない場合がある。

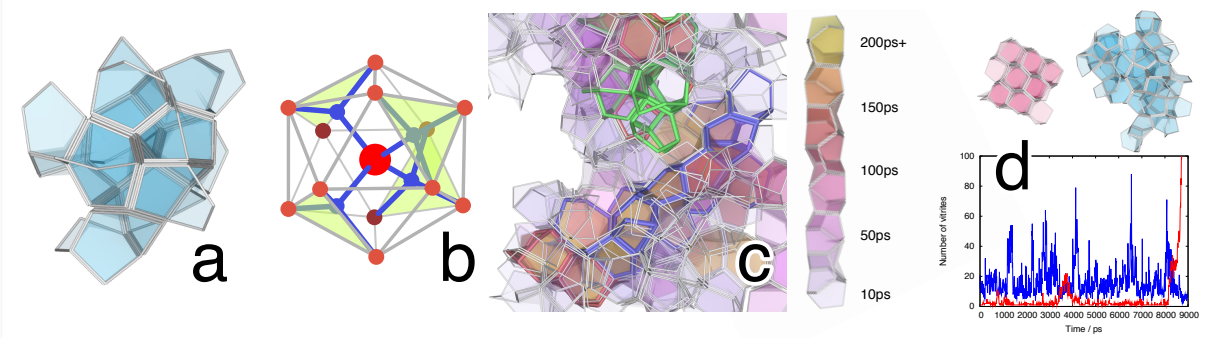
例えば、過冷却シリコンからの結晶の均一核生成は、過冷却水からのそれに比べて、はるかに容易に実現できる。分子動力学法により、液体シリコンを液液共存温度近くまで過冷却すると、数nsの待ち時間のあと結晶化が起こる。結晶化が起こるまでの時間に、液体の中でどんな風に初期の核が形成され、また消失するかを詳しく調べると、面白いものが見えてくる。

前節で紹介したフラグメント解析を用いることで、過冷却液体中に過渡的に形成される泡構造を、もれなく調べることができ、それらを組みあわせた中距離構造がどれぐらいの時間生き残るかもわかる。例えば、氷を構成するフラグメントがいくつか集まった中距離構造を探せば、結晶核がどのタイミングで空間のどこに形成されているか、それがどれぐらい残存するか、どれぐらいの大きさになれば成長が始まるかもわかる。

逆に、ある程度長い時間、構造変化がフリーズする領域を見付けだし、それがどんなフラグメントで構成されているかを調べれば、結晶以外の準安定構造が形成される様子も観察できる。このような構造は、長距離構造を持たず、分子数も少なく、さらに存在している時間も短かいので、実験的にはまず捉えることはできない。

その結果、過冷却シリコンの中には、既知のどんな氷・シリコン結晶構造とも異なる局所安定構造が、かなり長い時間形成されることがわかった。この構造は、六員環だけで構成されており、ネットワークの歪みも非常に小さいものの、正20面体の対称性を持つため、長距離秩序を作ることができない Polytope '240' と呼ばれる構造であることがわかった [13] 。同じ構造は、過冷却の水の中にも見付け出すことができ、やはり長寿命であることが確認できる (図C) 。

水にしてもシリコンにしても、個々の水分子 (原子) は、どんな構造になれば、全体の (自由) エネルギーを低くできるかを知っているわけではないので、周囲の水の配置と同調しながら、局所的に安定な構造を作ろうとする。それが偶然、安定相の結晶構造と同じなら良いが、ほかにも準安定な構造はたくさんあるので、間違った構造を作って安定 (安心) してしまうことも実際には起こっているのだ。このような、核生成の初期に準安定構造ができることで、核生成が遅くなったり、逆に結晶核を作る土台として使われたりする。だが、最終的にできる結晶構造には全く痕跡が残らない。つまり、結果としての結晶構造を見ただけでは、「どうやって結晶化したか」ということはわからないのである。



図C (a) Polytope '240' 構造。氷と同じように六員環だけで形成される局所安定構造。(b) Polytope '240'の構成方法。正二十面体を近似的に20個の正四面体の組みあわせとみなし、そのうち4つの四面体の重心に点を追加する。(c) 過冷却シリコン融液中に形成される様々なフラグメント。面の色はフラグメントの寿命ごとに塗りわけた。緑の辺を持つのは結晶フラグメント、青の辺を持つのがPolytope '240'構造。(d) 核生成前の融液中に見られる、結晶クラスター (赤) とPolytope '240'クラスターの大きさ。どちらのクラスターも間欠的に大きくなるが、最終的に結晶成長に至るのは必ず結晶クラスターである。

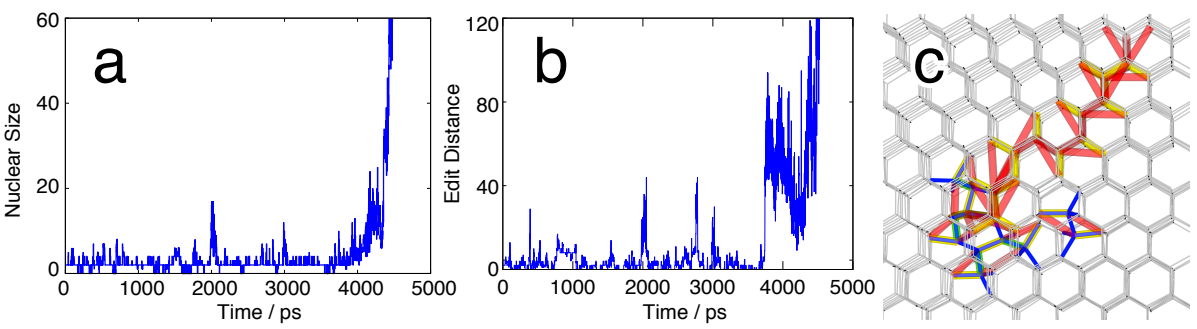
D.氷の融解

氷をあたためると、通常は表面から融解がはじまる。一方、赤外線などにより、氷の内部を直接あたためることにより、融解の均一核生成を起こせる。この場合、氷を融かすには、融点よりも高い温度にする (過熱する) 必要がある。欠陥やゴミや界面などからの融解が起こらない場合、氷はどうやって融けはじめるだろう。

シミュレーションにより氷を過熱すると、氷の格子がところどころで一瞬壊れて、10分子程度の大きさの、構造の乱れた小領域がときどき出現しては、またもとの氷に戻る。そのうち、小領域の一つが、氷の結晶の中を走りまわりはじめるのが見える。この領域は、臨界核サイズよりもずっと小さいので、すぐに融解にはつながらないが、氷の中での構造変化を促進し、結果的に融解までの時間を短くする [14] 。

この小領域は、実は水分子を1つ余計に含んだ高密度領域 (I欠陥) である。氷の構造が熱揺らぎで乱れ、また元に戻るのを繰り返している間に、誤って水分子の密度を均等にしないままに氷の構造に戻ろうとすると、このような、1分子余計に含んだI欠陥と、逆に1分子不足したV欠陥の対ができてしまう。I欠陥は水に近い密度を持ち、構造変化しやすいため、一旦形成されたI欠陥は、ふたたび偶然にV欠陥とであって対消滅するまで、格子の中を走り回るのだ。

I欠陥は、V欠陥とふたたび巡りあわない限り消滅できない。つまり、I-V欠陥対が離れているほど、氷に戻れる可能性は小さくなることが直感的にわかる。その遠さを計量するのに、編集距離が役に立つ。見た目には大きく壊れていないように見える構造でも、I欠陥とV欠陥の分離が起こっていると、完全な氷の構造に戻すためには、たくさんの水素結合を組替える必要がある。つまり、氷からの編集距離が遠い構造と言える。逆に、見た目にはネットワークの乱れが大きくても、氷からの編集距離が近い場合には、いろいろ構造を試行錯誤している間に、容易に完全な氷に戻る。実験的には編集距離は測定できないので、構造の見掛けの乱れの大きさ (融解核の大きさ) などをオーダパラメータとして、融解過程を記述するのが一般的だが、シミュレーションデータを扱う場合には、編集距離のほうが、適切な乱れの尺度と言える。



図D (a) 融液核の大きさ。結晶の格子点からずれた水分子が作るクラスターのうち最大のものに含まれる分子数。(b) 完全な氷の構造からの編集距離。(c) 格子点からずれた水分子が作るクラスター (青線) と、完全な氷の構造に戻すために必要な、最小限の水素結合の再構成 (赤: 結合の追加、黄色: 結合の除去) を示す。