高分子科学最近の進歩（本文 1 頁：24 字×39 行×2 段＝1872 字）_5 頁依頼

PoLyInfo と機械学習

石井真史

物質・材料研究機構
統合型材料開発・情報基盤部門 材料データプラットフォームセンター
[305-0044]　つくば市並木 1-1
グループリーダー，博士（工学）．
専門はデータベース，材料開発・計測，セマンティックウェブ技術．
ISHII.Masashi@nims.go.jp
https://www.nims.go.jp/research/group/materials-database.html

1．序

近年、高分子分野においても、マテリアルズ・インフォマティクスや機械学習（ML）への関心が高まっている。一方で、高分子の多様性を反映して、系統的なデータの少なさや、それをカバーするための形式を一にしたデータの統合が困難であることが問題になっている。物質・材料研究機構では、長期にわたって論文から高分子データを収集し、「PoLyInfo」というデータベースとして公開している[1]。ここでは、世界的な高分子データベースを俯瞰した上での PoLyInfo の位置づけ、インフォマティクスの観点で見た特徴、関連の成果等をまとめる。

2．高分子データベース

ポリマーに特化した世界規模のデータベースはあまり多くはない。有機物や生体のデータベースの中で高分子を部分的に扱っているケースは、今回の趣旨とは別の利用価値があると考えて割愛し、インフォマティクスに適用できる可能性がある比較的大きなデータベースをいくつか紹介しておく。

多くの高分子を収容しているという点では、MatWeb, LLC によって運用されている MatWeb[2]が知られている。エンジニアリング材料にフォーカスし、エンジニア、設計者、加工業者のために作られたデータベースであり、現在、155,000 を超えるデータシートを保有している。このデータベースには、金属、セラミック、半導体、繊維など、高分子以外の材料も含まれる。多くのサプライヤーからの製品情報を集めて利用しやすいように編纂しており、素材の比較や購入のためのカタログとして使うことができる。高分子に限ってみると、2022 年 8 月現在、97,635 materials が登録されている。その内容はいくつかのカテゴリに分かれているが、例えば最も数が多い Thermoplastic の内訳の上位 5 つは**表 1** のとおりである。このうち Nylon で最も多いのは Nylon 66 の 5,778 である。製品より素材に近い例である"Overview of materials for Nylon 66, PTFE Filled"では、物理特性（7）、機械特性（22）、電気特性（6）、熱特性（8）、加工特性（8）といった基本的な物性情報が一覧になっている。ここで括弧内は物性項目数を表す。特に、単位変換がされており他の高分子との比較がしやすい。ただし工業製品であることから、この例のみならず掲載データ全般で重合情報や添加物の情報、構造情報が少なく、高分子の素性は必ずしも明確ではない。公知情報の一つとして、カタログのインフォマティクス利用は考えられる。安定供給可能な工業材料から新たな知識や製品開発を生み出せるか検討することは興味深い。

このほかの産業用高分子データベースとしては、CAMPUS (Computer Aided Material Preselection by Uniform Standards)[3]が知られている。このデータベースの、1988 年のテキストベースの情報提供から始まる歴史は、英語版の wikipedia[4]に詳しく記載してあり、計測データの標準化およびコンピュータの発展と並走してきた過程を概観できる。データ数は明確にはわからないが、例えば、Nylon66 を検索すると 1,457 件ヒットすることから、上記の MatWeb よりも小規模と思われる。データは保護された PDF でダウンロード可能である。特性図とそのテキストデータも取得可能とされている。

これらのデータベースは、基本的に出来上がった製品の特性が提示されているが、重合情報など一次構造のデータは多くない。インフォマティクスにおいて高

表 1 MatWeb の Thermoplastic の内容の上位 5 項目

Material category	Number of data
Expand Nylon (Polyamide PA)	14,409
Nylon (Polyamide PA)	14,409
Polypropylene (PP)	10,481
Expand Elastomer, TPE	7,683
Elastomer, TPE	7,683

分子構造を記述子にして特性を予測するためには、製品名から推し量れる以上の構造特徴量を補うことが求められる。また、良く知られた高分子のデータに限られることから、物性から構造を逆解析して新規高分子を開発するには、多くの外挿が伴うと予想される。勿論、プロセスインフォマティクスなど、より即効性があり応用性に富んだ手法への展開は可能であり、開発目的に応じたデータベースの選択と利用が重要であろう。

材料科学、情報科学の統合を目指した MGI（Material Genome Initiative）などの取り組みを背景として、インフォマティクスに特化したデータベースとして興味深いものに、NanoMine がある [5]。NanoMine は高分子ナノコンポジットを扱うリポジトリであり、無限の組合せが考えられる複合材を系統的に扱う挑戦的な試みである。データ数は多くはないが、様々なデータ科学的なアイディアが込められており、紙数の都合で十分紹介できないことが惜しまれる。一般論に戻れば、インフォマティクスを目的としたデータの収集において、Processing-Structure-Property（p-s-p）を包括的に扱い、上述の MatWeb などで不足していた構造情報を補うことが重要である。この p-s-p の考えは、主に冶金の分野で扱われていたが高分子にも拡張されてきており [7]、NanoMine でも取り入れられている。PoLyInfo は構造情報に注力しており、p-s-p の考えを内在している。

3. PoLyInfo

3.1 PoLyInfo の ID 体系

PoLyInfo は、学術論文をデータ源としており、MatWeb や CAMPUS のようなカタログデータは収録していない。工業分野での応用よりは、重合等の基礎化学的な情報を収集している。実用性は別にして、合成可能な高分子をできるだけ網羅的にまとめることが基本的なポリシーとなっている。ここで収録条件として、「高分子の構造が確定できること」があり、実際に構造ごとに ID を附番し（PID と称する）、添加物など付加的な条件で細分化したサンプルを PID に紐づけることで、PID ごとに発現し得る機能と特性を俯瞰している。**図 1** に PoLyInfo の ID 体系を示しておく。文献ごとにサンプルを選り分け、PID とは別にサンプル ID を付けている。すなわち、文献に ID（この図では 72103）を附番し、そこに出て来る高分子に枝番を付ける（例えば 72103-1）。ここで重合や分子量など一次構造が異なれば、更に末尾に枝番を加える（例えば 72103-1-1）。更に、成形条件等高次構造が異なれば、末尾に三つ目の枝番を加える（例えば 72103-1-1-1）。基本的に、こうして一意に決められた枝番が三つ付いたサンプル ID に対して測定条件を含めた物性値が紐づけられる。明らかに構造が提供されている PID とサンプル ID の間には二階層の差がある。このギャップを埋める重合情報や成形条件などの補完情報が PoLyInfo に収録されている。つまりデータベースの体系に沿って p-s-p と同様の書き方をするのであれば、ある PID の構造 s が確定した後に、プロセス p が加わって、特性 p が与えられることから s-p-p となる。あるいは、PoLyInfo では高次構造で分類される結晶情報を以て s とすれば、従来のコンセプトと同様に p-s-p となり得る。高分子ミクロからマクロまで様々な形体があることを考えると、その順番は問題ではなく、これらの三要素をデータセットとして扱うことが重要であり、PoLyInfo のサンプル ID の三つの枝番がそれに対応している。

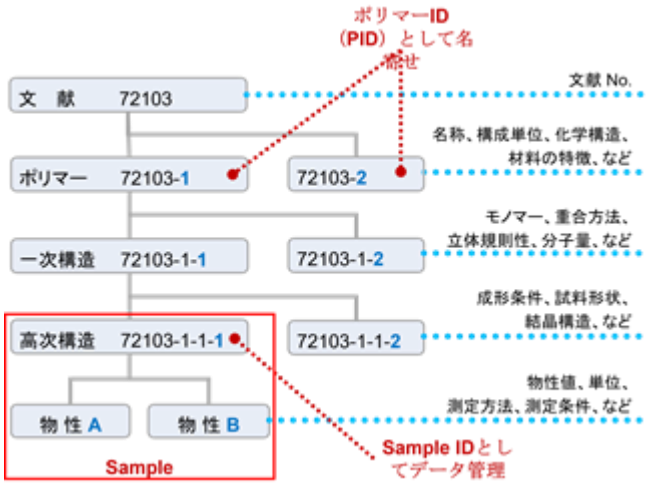

図 1 PoLyInfo の ID 体系

3.2 PoLyInfo の構造情報

PoLyInfo における s-p-p の s、すなわち構造の扱いについて紹介する。前述の通り PoLyInfo では、サンプルは PID に分類され、名寄せされている。2022 年 3 月現在、PID がついている高分子は、ホモポリマーが 18,526 種、コポリマーが 7,442 種ある。これらの構造は MOL ファイルで提供しており、SMILES 等にも変換可能である。しかし、**表 2** に示すような MOL ファイル形式では正確に表現できないものについては公開していない。データベース上は、PID に表にある suffix を付けて表現できない部分を原子団として代替表現して管理している。この三つのカテゴリの代表的なものを**図 2** に示しておく。

コポリマーの場合は、必要に応じて接合単位の MOL ファイルを加えている。PoLyInfo では、この接合部についても一意の ID を付けて管理している。

表2 MOLファイル形式で正確に表現できない高分子

Category	Polymer example	PID	Suffix
Borane	poly[1,10-octacarborane-1,10-diyl(dimethylsilanediyl)oxy(dimethylsilanediyl)]	P170058	1
	poly{[2,4-bis(chlorodimethylsilyl)-2,4-dicarba-closo-heptaborane-7]-alt-[2,4-bis(methoxydimethylsilyl)-2,4-dicarba-closo-heptaborane-7]}	P470001	2
	poly[(benzene-1,4-diol)-alt-(m-carborane-1,7-dicarbonyl dichloride)]	P390137	3
	poly{[bis(p-phenoxyphenyl)-p-carborane]-alt-(4,4'-oxydibenzoic acid)}	P079002	6
	poly{[bis(p-phenoxyphenyl)-o-carborane]-alt-(biphenyl-4,4'-dicarboxylic acid)}	P079001	7
Ferrocene	poly(vinylferrocene)	P039034	4
Phthal	poly[hydroquinone-alt-(terephthaloyl dichloride)]	P099041	5

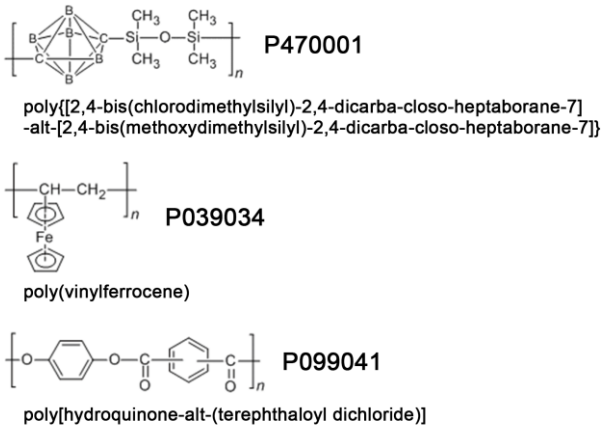

図2 MOLファイル形式で正確に表現できない高分子例

3.3　PoLyInfoにおけるプロセス情報

PoLyInfoにおけるs-p-pの最初のp、すなわちプロセスの扱いについて紹介する。PoLyInfoで扱っているプロセスは、大きく分けて重合を中心にした一次構造に関するものと、成形などの高次構造に関する二過程がある。重合については、PoLyInfoのMonomer searchで概観することができる。しかし、PoLyInfoは高分子を中心に扱っているために、低分子の情報は十分とは言い難い。2022年3月時点で19,136の低分子が登録されているが、PubChem[8]などの大きなデータベースでは、登録・公開数が既に1億種を越えている[9]。勿論、これらの低分子の全てが重合に寄与するわけではないが、この圧倒的な差を縮めることは困難であろう。現在、PoLyInfoは、高分子での独自性は維持しつつ、低分子では協調を重視し、外部の低分子データベース（日化辞[10]及びPubChem）との連携を進めている。そのスキームの中では、PubChemのCompound IDや日化辞番号から特定された低分子で重合できるPoLyInfoの高分子が特定できるようになる[11]。そもそもバイオロジーを基盤とする両データベースはデータベースの統合を積極的に進めており、RDF（Resource Description Framework）[12]といった国際的規格に基づく連携技術を10年近く前に既に導入している。PoLyInfoもRDF化を進めており、PubChemで公開している23種の記述子を使って、PoLyInfoの物性をターゲットとした統合的なインフォマティクスが可能になるであろう。一方で、PoLyInfoでは原料となる低分子情報に加え、重合反応分類（付加重合、開環重合など）、重合機構（ラジカル重合、酸化重合など）、重合様式（塊状重合、溶液重合など）、触媒、溶液、開始剤、停止剤、添加剤等が、重合条件と共に登録されている。これまでPoLyInfoでは、PIDレベルで重合情報の対応付けを行ってきたが、それを拡大する形で、他の物性データと同様にサンプルIDレベルでの対応付けを行うデータ収集ポリシーへ移行している。新ポリシーでのデータは2023年に公開される予定である。

高次構造のプロセスについては、体系的に整理できるいくつかの項目が登録されている。成形方法（射出成形、圧縮成形、キャスティングなど）、成形条件（自由記述）、処理方法（エイジング、アニーリングなど）、処理条件（自由記述）などがあるが、おそらく、これらがMLに用いられるケースは限られるであろう。これらは実験の再現性や信頼性を示すものであり、インフォマティクス応用ではなく、個々のデータを参照して研究開発の出発点とする従来の方法において、引き続き活用されると考えられる。現在のMLにおいて記述子となり得るものは、最終的な形状や画像から得られる特徴量と思われる[5,6]。これらの特徴量に相当するものはPoLyInfoの中では多くはない。例えば試料形状（シート、フィルムなど）及び形状情報（自由記述）があるが、MLという観点では粗い粒度であろう。他に特徴量になり得る項目として、結晶構造情報がある。結晶系、格子定数、対称性は特徴量となり得るが、添加剤や充填剤など他の登録情報と合わせて使うことになると考えられる。

3.4　PoLyInfoにおける特性情報

PoLyInfoにおけるs-p-pの最後のp、すなわち特性について紹介する。PoLyInfoでは、凡そ100種の物性を収録している。2022年3月時点での物性情報数は、492,645である。物理的性質、光学的性質、熱的性質、電気的性質、他機械的特性などに分類され、物性値、単位の他、測定方法、測定条件、規格などが上述のSample IDごとにまとまっている。こうした情報はインフォマティクスへの直接的な活用を思わせるが、いくつかの注意点がある。

測定値は測定方法や測定条件に依存する。また論文では、これらの依存を積極的に記載し、更にこの依存性に論文のオリジナリティが現れる場合が多い。PoLyInfoではサンプルIDに物性値を紐づけているが、データ収集元の論文の趣旨を考えると、IDと物性値

の一対一の紐づけで情報を完結させることは正しくない。例えば、サンプル ID 01317-1-1-1 を見てみる。このサンプル、polyetheretherketone （P070466）融点 335C、結晶化温度 180C に対して、様々な DSC の設定温度範囲に対するアブラミ式の成長速度 k、核生成速度 n がまとめて掲載されている（**表 3**）。
これらの表現形式は、論文ごとに異なり、またこの表の中ですら Measurement condition の書き方は様々である。物性値についても k で 7 桁に及ぶ差がある中で、どれを 01317-1-1-1 のターゲット物性値とするかは、目的に応じた検討が必要である。おそらく PoLyInfo を多くのデータの源と考えるだけではなく、少数データを精査して抜き出し、転移学習などに用いることが活用幅を広げるであろう。

表 3 PoLyInfo サンプル ID 01317-1-1-1 に関する、様々な成長速度 k と核生成速度 n、及びそれらの測定条件

k	n	Measurement conditions
1.80E-07	5.1	Heating rate;50C/min from 80C to 280C
2.50E-09	5.4	Cooling rate;20C/min from the melt
0.00008	3.3	315[C]
0.0017	3	312[C]
0.01	3.05	308[C]
0.031	3	164[C]
0.00067	2.8	160[C]

4. データ活用例

最近は、PoLyInfo をはじめとする、高分子データを使った ML 例は少なくない。最近の s-p-p の最後の p である特性予測の例をいくつか挙げてみると、機械特性であれば引張試験における破断強度のモデル化の報告例がある [13]。電気特性であれば、データライブラリー、文献、ハイスループット計算など様々なデータリソースからデータを収集して高分子誘電率を予測した文献 [14]はナノコンポジットの特性予測として興味深い。逆浸透膜の設計では、未知低分子と作成条件の組合せのデータ空間を作り、ベイズ最適化によって水の透過性と塩阻止性の最適条件が特定され、実際に現在の特性を上回る膜の作成に成功している [15]。また、s-p-p の二番目の p であるプロセスに関しては、様々なデータ源から抽出した数百の合成テンプレートを学習させ、更に類似性を指標にして、新しいターゲット高分子の合成テンプレートを提示する事も行われている [16]。これらの事例は、確かに ML の成功例として重要であるが、一般化や実用化に向けた学究的な立場に立って全体の動きを眺めた時、客観的な指針を示す報告が現れ始めている点を強調しておく。

その端緒として、Lee F.L.らの論文 [17]によれば、ごく基本的な物性である、Tg、融点、密度の予測に現時点で障壁はないが、どのモデルやフィンガープリントも実際のポリアミドの引張弾性率 E を正確に予測できなかったことを明らかにしている。単純なフィンガープリントでは捉えられない潜在的な情報が必要になると結論付けているが、上記の PoLyInfo の PID とサンプル ID の二段階の差を埋める記述子の必要性に対応すると考えられる。Sattari K.は最近の Review[18]の中で、今後、高分子の ML が取り組むべき課題として

・ホモポリマーからより複雑な高分子への拡張
・高分子アーキテクチャーへの拡大

を挙げている。ここでのアーキテクチャーとは、ハイパーブランチや、スター、ブラシなどの構造的特徴を意味し、溶解度やガラス転移温度などの物理的性質にも大きな影響を与える要素を指している。また、DNA 等の配列が重要なバイオポリマーへの拡張も含まれている。いずれにしても最も単純なホモポリマーから実際の複雑な系への展開が必要であることは間違いない。Yang Y.らも、高分子の特性は多くの複雑な構造的要因に影響されるため、繰り返し単位だけでは、生産のためのガイドラインとなる予測は示せないことを明言している [19]。その上で、相状態を判定した上で、そのプロセス温度依存性を予測するモデルの構築を行っている。こうした、s-p-p 全体にわたって複雑な系をモデル化することが、高分子のマテリアルズ・インフォマティクスの今後の進むべき方向の一つと言えよう。
図 3 は、PoLyInfo に登録されているホモポリマーについて、横軸に繰り返し単位内の炭素（C）の数、縦軸にその数以上の C を含むホモポリマー数（累積数）を示す。おおむね、C の数が 40 以上で指数関数に良くフィットすることが分かる。高分子は繰

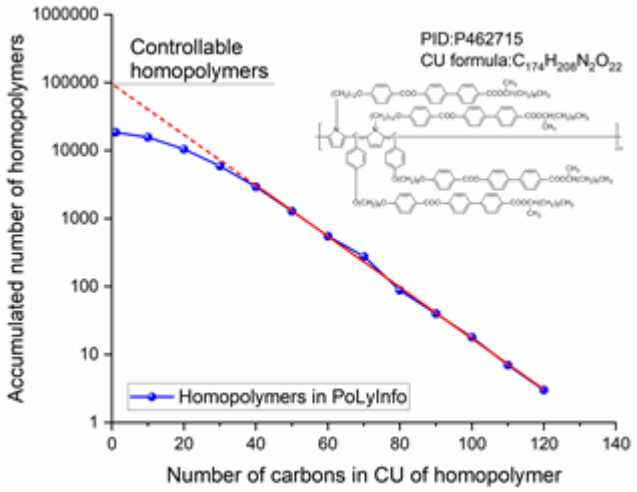

図 3 PoLyInfo に登録されているホモポリマーの繰り返し単位内の C の数に対する、その数以上の C を含むホモポリマー数（累積数）

り返し単位を際限もなく大きくすれば無限の組合せがあり得るが、実際はそのような高分子は機能的にも作成技術に照らしても現実的ではない。この応用上の限界がこの指数関数に現れていると考えられる。挿入図は、最近PoLyInfoに登録された最も繰り返し単位中のCの数が多い高分子P462715であるが、Cは高々174である。いささか乱暴な方法だが、この指数関数をCの数が1まで外挿すると、現実的に扱えそうなホモポリマー数は10万と見積もられる。勿論PoLyInfoに収録されていない高分子もあるので、この数の正確さは低い。それにしても、人にとって少なくないこの数も、機械にとっては多いとは言えない。それでも高分子のインフォマティクスが難しいのは、ここに高次構造の複雑さが数桁以上のオーダーで掛け算されるためであろう。いずれにしてもそこに向かうならば、検索ではなく複雑な系のMLや統計処理が材料開発上、真に必要になると思われる。

5. まとめ

本稿では、最近の高分子のMLのデータ源として、PoLyInfoおよびその他の国際的なデータベースを使うことを念頭に、データベースの特徴、プロセス-構造-特性（p-s-p）の考えに基づいた具体的内容と課題、最近のMLの動向から見られる方向付けを述べた。多くの検討要素がある中で、MLですべきことは、単純な系を使って短期的な成果を追及することだけではなく、高分子の展開と課題解決を見越した未知のデータ空間の探査と考えている。

文　献

1) 物質・材料研究機構, "高分子データベース(PoLyInfo) - DICE :: 国立研究開発法人物質・材料研究機構", 高分子データベース (PoLyInfo), https://polymer.nims.go.jp/, 2022年9月6日.
2) MatWeb, LLC, "Online Materials Information Resource - MatWeb", MatWeb, https://www.matweb.com/index.aspx, 2022年9月6日.
3) CWFG mbH, "CAMPUSplastics", CAMPUS® - a material information system for the plastics industry, https://www.campusplastics.com/, 2022年9月6日.
4) Wikipedia, "CAMPUS (database) - Wikipedia", CAMPUS (database), https://en.wikipedia.org/wiki/CAMPUS_(database), 2022年9月6日.
5) He Zhao, Yixing Wang, Anqi Lin, Bingyin Hu, Rui Yan, James McCusker, Wei Chen, Deborah L. McGuinness, Linda Schadler, and L. Catherine Brinson, "*Applied Physics Letter Mater.*", **6**, 111108, (2018).
6) Duke Univeristy, "NanoMine Nanocomposites Data Resource", NMaterialsMine, https://materialsmine.org/nm#/, 2022年9月6日.
7) Chung D.D.L., "*Materials Science and Engineering*", **R 113**, 1, (2017).
8) National Institutes of Health (NIH) National Center for Biotechnology Information, "PubChem", PubChem, https://pubchem.ncbi.nlm.nih.gov/, 2022年9月6日.
9) National Institutes of Health (NIH) National Center for Biotechnology Information, "PubChem Statistic", PubChem Data Count, https://pubchemdocs.ncbi.nlm.nih.gov/statistics, 2022年9月6日.
10) 国立研究開発法人科学技術振興機構, "J-GLOBAL 科学技術総合リンクセンター", J-GLOBAL, https://jglobal.jst.go.jp/#%7B%22category%22%3A%227%22%7D , 2022年9月6日.
11) Masashi Ishii, Taro Takemura, and Mikiko Tanifuji, Proceedings of the ISWC 2019 Posters & Demonstrations, Industry, and Outrageous Ideas Tracks co-located with 18th International Semantic Web Conference (ISWC 2019), http://ceur-ws.org/Vol-2456/paper18.pdf, 2022年9月6日
12) World Wide Web Consortium (W3C), "RDF - Semantic Web Standards", Resource Description Framework (RDF), https://www.w3.org/RDF/, 2022年9月6日.
13) Cravero F., Díaz M.F., Ponzoni I., "*Journal of Chemical Physics*", **156**, 204903, (2022).
14) Zhu M.-X., Deng T., Dong L., Chen J.-M., Dang Z.-M., "*IET Nanodielectrics*", **5**, 24, (2022).
15) Gao H., Zhong S., Zhang W., Igou T., Berger E., Reid E., Zhao Y., Lambeth D., Gan L., Afolabi M.A., Tong Z., Lan G., Chen Y., "*Environmental Science and Technology*", **56**, 2572-2581, (2022).
16) Chen L., Kern J., Lightstone J.P., Ramprasad R., "*Applied Physics Reviews*", **8**, 31405, (2021).
17) Lee F.L., Park J., Goyal S., Qaroush Y., Wang S., Yoon H., Rammohan A., Shim Y., "*Polymers*", **13**, 3653, (2021).
18) Sattari K., Xie Y., Lin J., "*Soft Matter*", **17**, 7607, (2021).
19) Yang Y., Ye H., Zhu W., Zou X., Dong H., "*Industrial and Engineering Chemistry Research*", **60**, 12068, (2021).