

MAGICIAN 養成講座

Materials Genome/Informatics and Chemo-Informatics Activate Networks

第4a回ポリマー・インフォマティクス用のデータベース

2018.9.4 横浜国大非常勤講師 山本博志

情報化学を用いて、データを解析していると、どんどんデータが増えてくる。また、実験化学者との協業などが増えてくると、商品名しかわからない原料やそのカタログ値などもどんどん増えてくる。昔の様にコンピュータの HD が Mbyte クラスであればどうにかなったものが、Gbyte クラスになると、どうにも整理がつかなくなってくる。扱った記憶はあっても、様々な有用な式が、どこにあったかを思い出せなくなり、自分で自分の HP、pirika.com にアクセスして調べる始末になる。本来データをきちんと整理して持っていたいなら、本格的なデータベースを導入して管理すべきであろう。しかし、それを本格的にやろうとすると、それ自体が仕事になるくらい大変なことで、片手間にできる様なことでは無い。MI をしやすい様に構造化されたデータを用意しようという動きもあるが、例えば、ガラスとポリマーを同列に構造化データにすることはできない。化学者が専門的知識を学ぶことなく、お手軽にデータベースを使いMIを利用して研究を加速する。そのやり方を説明しておこう。

表計算のテーブルと DB の違い

前回、フッ素ゴムの物性について検討を行った。その際に次の様なテーブルを示した。

Run	Tg	VDF	HFP	TFE
1	227.5	100	0	0
2	234	95	5	0
3	240.5	89	11	0
4	244	83.7	16.3	0
5	248	80.6	19.4	0
6	250	77.4	22.6	0
7	253.5	74.2	25.8	0
8	255.5	72.2	27.8	0
9	261.3	66.2	33.8	0
10	267.8	59.3	40.7	0
11	271.5	35.6	32.8	31.6
12	264.5	49.4	18.4	32.2
13	260.5	55.4	13	31.6
14	259.5	60.3	24.1	15.6
15	255.4	61.2	16.5	22.3
16	258	63.8	27.8	8.3
17	253.5	64.3	19.1	16.6
18	252	71.8	20.7	7.5

元々は表計算ソフトで作ったテーブルである。テーブルというのは、行と列からできている。この場合、一行目はタイトルで、二行目からは様々なフッ素ポリマーのデータになる。列には2列目にはガラス転移温度(Tg)、その後の列にはモノマーの比率が示してある。1列目には、Run 番号として 1 から順番に数字を割り当てているが、実際には Yamamoto-20180903-1 などと合成結果、分析結果と紐づけられる名称を入力する事が多い。

これはテーブルであって、データベースでは無い。などということはない。これも立派なデータベースである。行の事をレコードと呼ぶ。そして、列の事をフィールドと呼ぶ。そこでこの場合は 5 フィールド、18 レコードのデータベ

ースと言え。データベースをテーブル形式で表現しただけである。

さらに、合成結果のテーブルには、日付、実験のスケール、開始剤の種類、量、重合方法、重合率、反応の後処理などが記載されているかもしれない。さらにできあがったポリマーを分析し、Tg 点やそれ以外の物性値、NMR や元素分析から決定した導入モノマーの比率のテーブルがあるかもしれない。

さらに、このフッ素ポリマーを使い、カーボンブラックや受酸剤などを添加してゴムを作成し、その物性値まで含めたテーブルがある。

FKV種類			カーボンブラック				受酸剤				
GT1	GT4	GT6	BT	BT	SP	SP	SP	PA	PA	PA	PA
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

ポリマーを構成するモノマーの物性値のテーブルもあるだろう。一つの素材を MI するために、テーブルの数はどんどん増えてくる。

見通しが悪くなるので一つのテーブルにまとめようとする。最初のテーブルに合成法や分析結果、ゴムとして処方検討した結果を列としてどんどん付け加えることになる。テーブルはどんどん巨大化していくが、中身はスカスカなテーブルができてくる。コンピュータにとっては何の問題もないことだが、人間にとっては理解不能のテーブルになる。そこで、この問題を回避するために、リレーショナルデータベース(RDB)が採用される。

これは例えば、**Yamamoto-20180903-1** という一つのポリマーを表す“キー”を他のテーブルでも共通に用いる。つまり、合成テーブル、原料テーブル、分析テーブル、処方テーブル中のレコードが **Yamamoto-20180903-1** というキーで横串が刺さった状態にしてあげる。横串は1本である必要はない。モノマーはモノマーで、反応性を示す **Q** 値、**e** 値、沸点、密度などの熱力学的性質、毒性、環境上の指標などと横串を指す。モノマー程度の低分子であれば、キーとしては **CAS** 番号を使い、横串を通す事が多い。そこでデータベースというのは、テーブルの集合体で、テーブル間の相対関係（リレーション）を扱いやすくしたものがリレーショナル・データベースだと思えば良い。マイクロソフトの製品では、**Excel** はテーブルを扱うソフトであるが、**Access** はテーブルを集合体で扱うデータベース・ソフトである。

筆者は、もう **30** 年以上マック・ユーザーなので、残念ながら **Access** は使った事が無い。クラリスワークス、アップルワークスなど、非常に優れたカード型データベースがシステムに付属していたので、それを使っていた。

現在は、それらの後継製品である **Filemaker Pro** を利用している。**SQL** 言語を覚える事なしに手軽に使い、**Excel** などとも親和性が高く、使えるコンピュータの種類も多い。手軽にスタートして、必要であればサーバー版など本格運用にも耐えられるシステムが提供されている。値段もリーズナブルで、個人で買える程度である。試用版がダウンロードできるので今回のサンプルを試してみて、気に入ったら導入を考えてもらえればと思う。(何のコネもないので、紹介割引などは受けられない。)

データベースの作成

今回はガスバリアー性のポリマーをターゲットに、“MI”を行う為に便利なデータベース構築”を学ぼう。

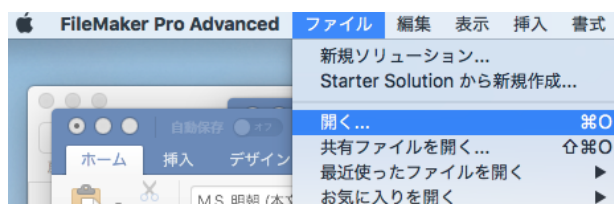
Pcode	Name	logPO2
50000	Polyethylene	
50004	Polypropylene	
50005	Poly(vinyl alcohol)	
50006	Polyoxyethylene	
50007	Poly(vinyl fluoride)	
50008	Polyacrylonitrile	
50010	Poly(cis-1,4-butadiene)	
50024	Poly(vinyl chloride)	
50027	Polymethacrylonitrile	
50081	Poly(vinyl acetate)	
50082	Poly(methyl acrylate)	
50092	Polychloroprene	
50106	Poly(vinylidene chloride)	
50112	Polytetrafluoroethylene	
50115	Poly(methyl methacrylate)	
50138	Polystyrene	

まず、今回はビニル系のポリマーの酸素透過係数について解析を行おう。**Pcode** というのはハンセンの溶解度パラメータ(**HSP**)を取り扱うソフトウェア、**HSPiP** でつけられているポリマー用のコードである。

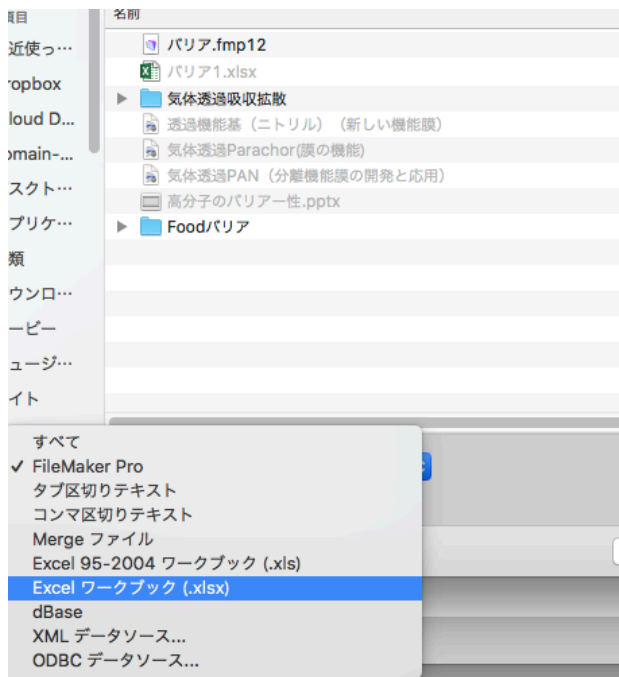
そして、**Pcode** に該当するポリマーの酸素透過係数を収集する。テーブルを作り、相当するデータを、ハンドブック、文献、データベースから収集する事は良く行う事だろう。これを、テーブルで扱わないで、データベースとして扱うとどんな良い事があるのだろうか？

テーブルの読み込み

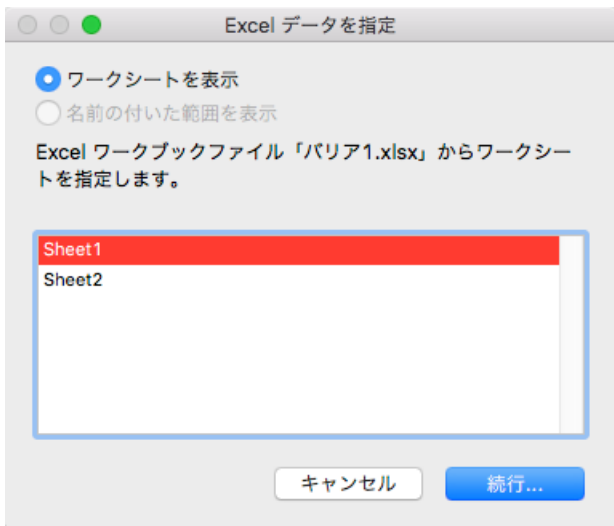
データベースをスクラッチから作成するのも良いが、まずは、慣れ親しんだテーブルをデータベース化してみよう。先ほどの **Excel** のテーブルをセーブしておく。



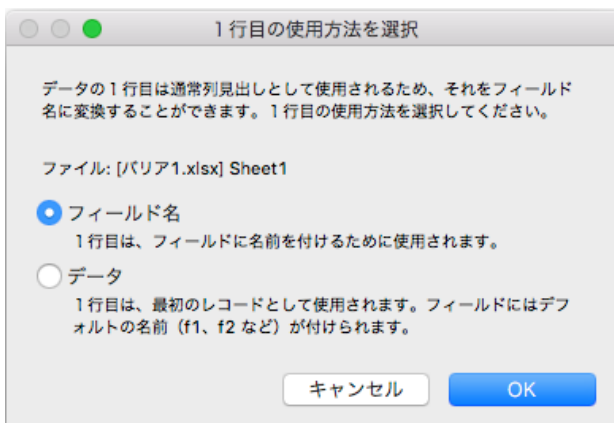
ファイルメーカー・プロ(**FMP**)を立ち上げて、ファイルを開く、を選択する。(Excel のテーブルを直接 **FMP** にドロップしても良い)



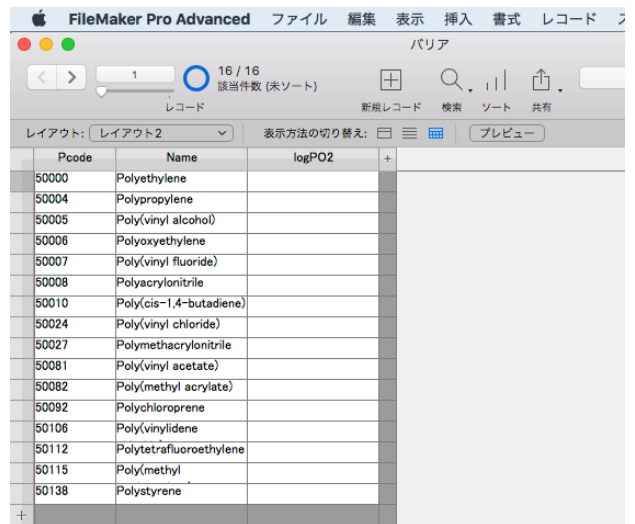
ファイルの種類を選択 (Excel ワークブック(.xlsx)して、ファイルを開く。



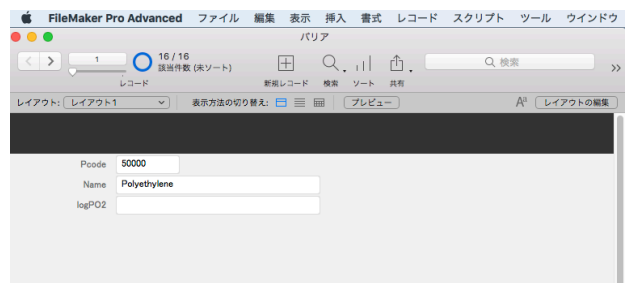
ワークシートが複数ある場合には、データベースに取り込むワークシートを指定する。



一行目はフィールド名に指定する。(フィールド名に使えない文字もある。)



各行を取り込んだ、テーブル・レイアウトのデータベースが作成される。

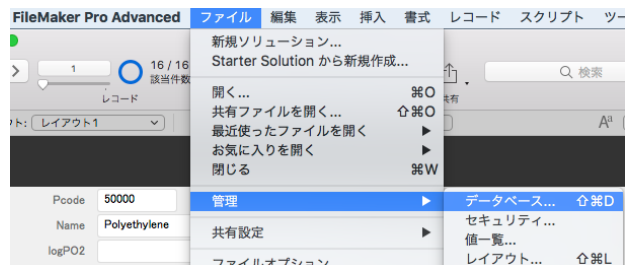


カード型のレイアウト表示。

非常に簡単に、データベース化できる事が判るだろう。

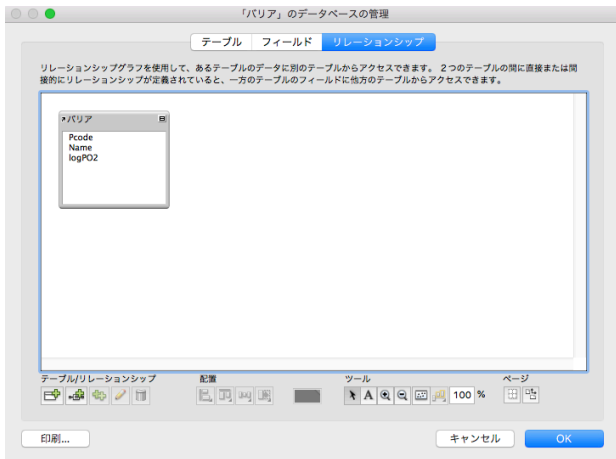
データベースを用いたデータ収集

我々、HSPの研究Grは様々なデータベース、ハンドブック、書籍、文献などからテーブルを作成し、同じようにデータベース化してある。(PDFからAdobeのAcrobat proを使って Excel のテーブルを作り、FMP のファイルにしてある) それらの DB と、今作成した DB を、リレーション機能を使って接続する。



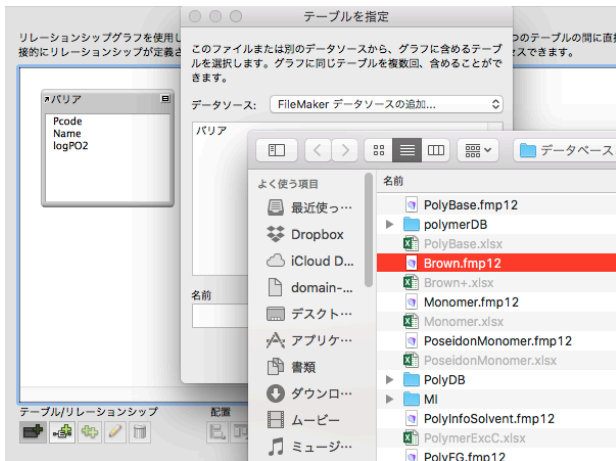
ファイル/管理/データベースを選択する。

データベースの管理画面で、リレーションシップのタグを選ぶ。

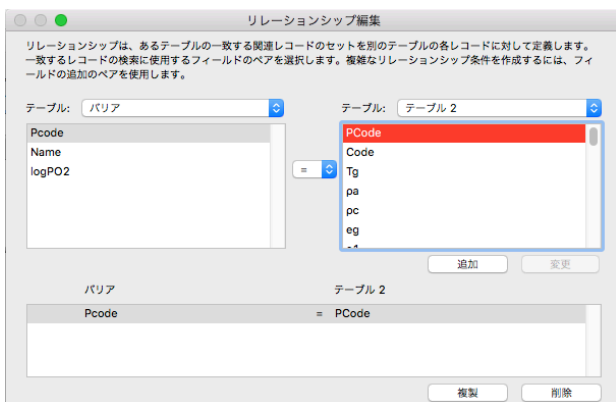


左下のアイコンからテーブルを指定する。

Filemaker データソースの追加を行い、データベースを選択する(Brown.fmp12)。



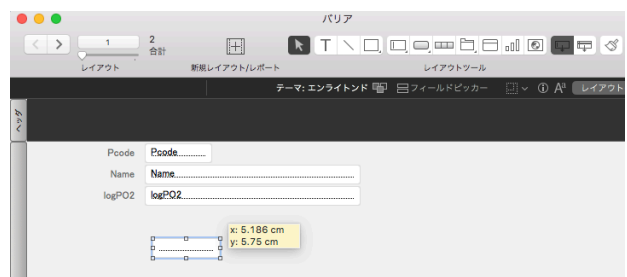
リレーションシップ編集を行い、テーブル中の共通キーを指定する。



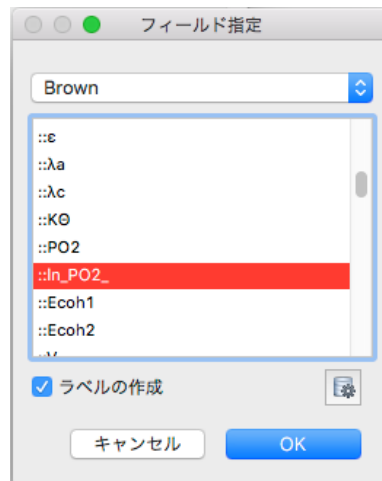
他のデータベースとも、Pcode を用いて接続する。



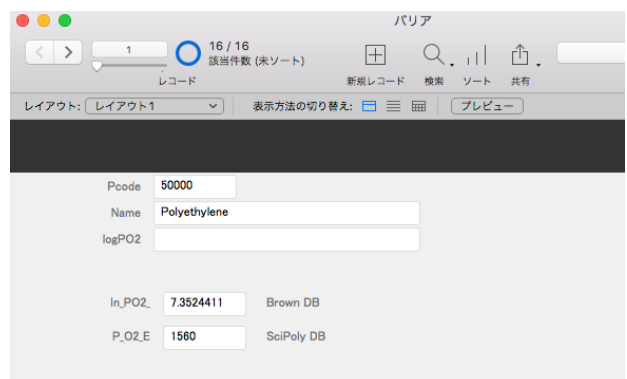
こうしたリレーショナル化が終了すると、バリアのデータベースに対して、他の DB に登録されている値を表示する事ができるようになる。



レイアウト・モードで右上のアイコンからフィールドをメイン画面にドロップする。



フィールドを指定する。

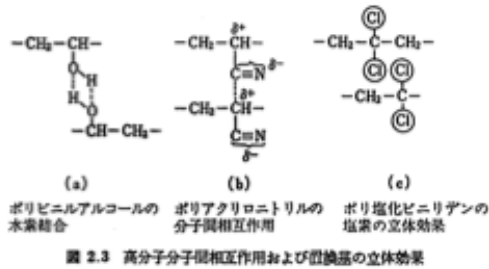
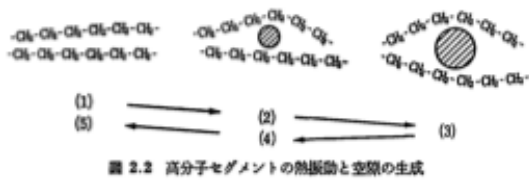


最終的に各 Pcode に対して、BrownDB と SciPolymerDB から酸素透過係数を表示できるカード型のデータベースが作成される。

ここまでであるなら、実はテーブルに対して、各データベースを叩いて、値を持って来た方が早い。データベース化する時のメリットは別にある。

データベースを用いた解析

ガスバリアー性のポリマーは、例えば食品包装用の素材として非常に有望な市場を持っている。例えば、食品用のラップ、鯉節などの匂いのきつい食材の包装材、太陽電池用のバックシート、塗料などもバリアー性が重要である。例えば、講談社サイエンティフィク「新しい機能膜」には膜に対する透過現象に関して、次のような記載がある。



高分子の分子間相互作用が強くなると、セグメントの熱振動がしにくくなり、膜の透過係数が小さくなる。水素結合、電荷相互作用、立体障害がその原因である。

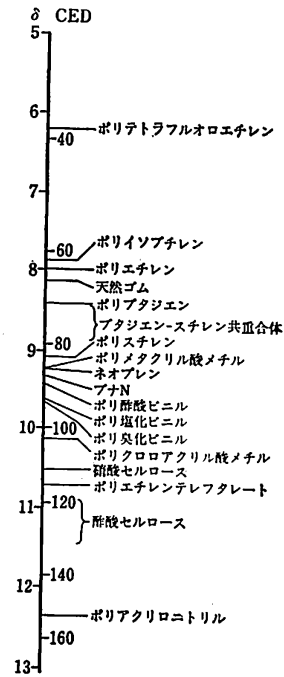
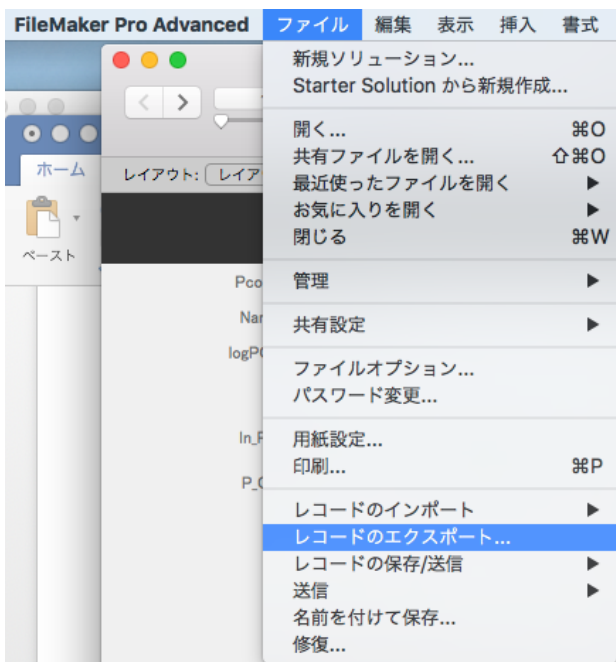


図 2.4 高分子の分子間凝集エネルギー密度 (CED) および溶解度パラメータ (δ) [P. A. Small, *J. Appl. Chem.*, 3(2), 71(1953)]

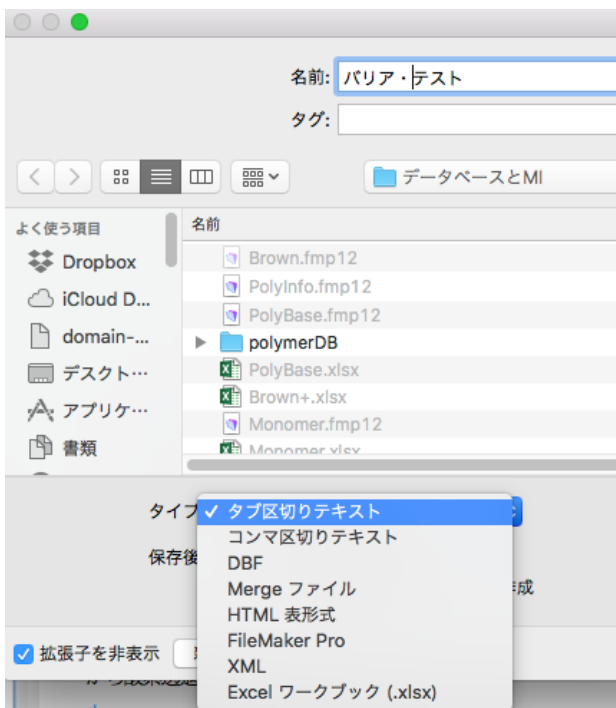
そして、「その分子間相互作用を定量化するのに、凝集エネルギー密度 (CED) や SP 値が有効である」とある。例えばポリアクリロニトリルは凝集エネルギー密度が非常に高く、熱をかけても溶融しない。そして蒸し焼きにすれば炭素繊維という非常に魅力的な素材になる。セルロース系の素材も同じように溶融しない。

そこで、MI 的な考え方に立てば、酸素の透過係数と例えば SP 値を比較してみたい。その時に先ほどのテーブルに SP 値のデータを継ぎ足していく。Tg 点や Tm 点のデータを継ぎ足していく。ポリマーを構成する原子団の情報を継ぎ足していく。と言う作業を繰り返すか、RDB を使ってその作業を効率化するかの選択になる。

Filemaker Pro には、データベースの値を Excel の形式で書き出す機能がある。

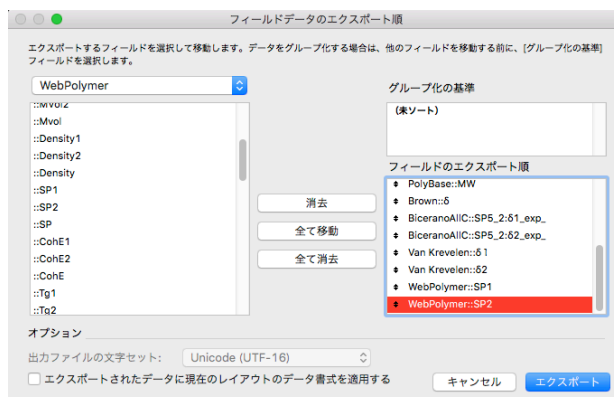
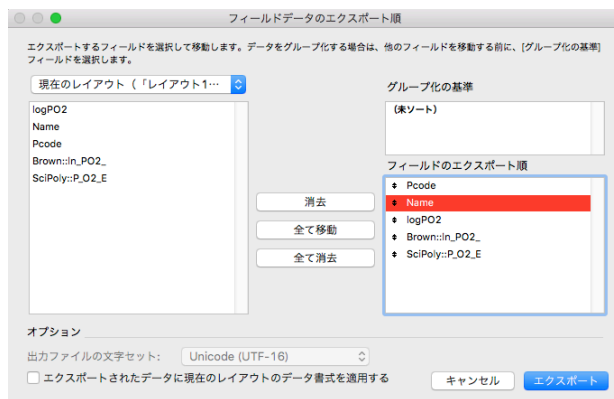


ファイル/レコードのエクスポートを選択する。

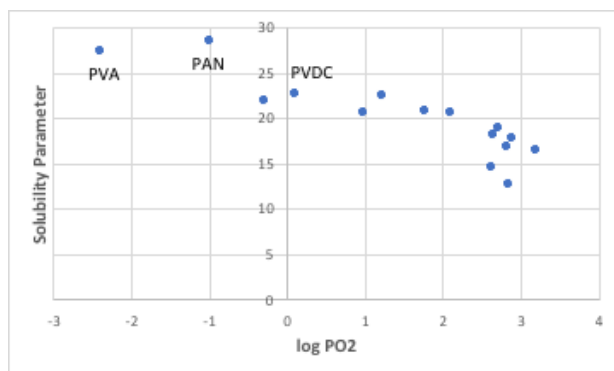


ファイル名とファイル形式を指定する。Excel ワークブックを選択する。

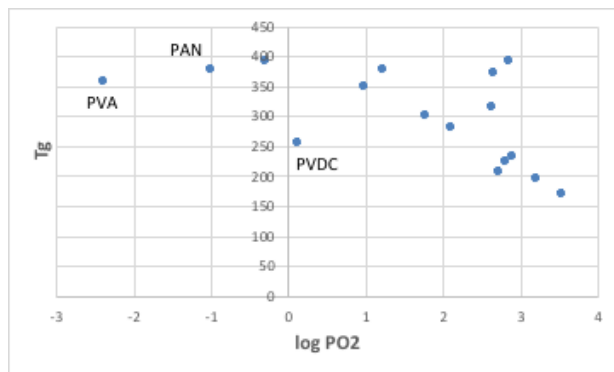
エクスポートするフィールドを指定する。その際に、RDBで指定しているDBのフィールドも指定できるので、例えば各DBに登録されているSP値も含めてエクスポートすることができる。TgやTmを含める事も簡単にできる。これが、RDBを使って、データをまとめておくことの最大のメリットであろう。



このようにSP値付きで酸素の透過係数が書き出されると、すぐに下図のようなグラフを書く事ができる。



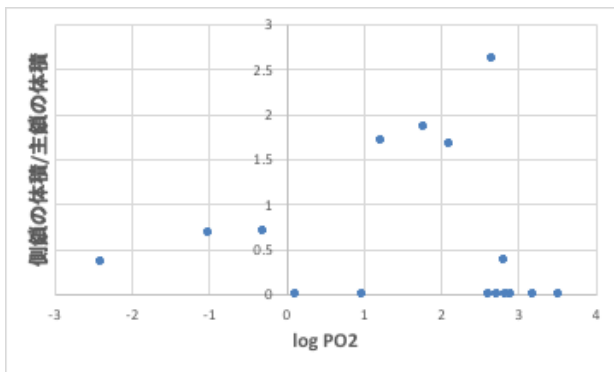
酸素をバリアするには、SP値が高い事が重要である。



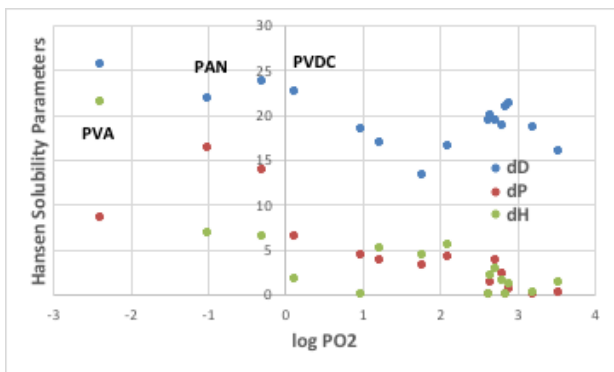
縦軸をガラス転移温度(Tg)に変える事も容易である。Tg点が高いポリマーはポリマー鎖が動きにくいポリマーである。そこで、酸素が透過する空隙が発生しにくい。PVDCはTg点が高いにもかかわらず、透過性が低い特殊なポリマ

一である事が解る。

そうした空隙は、側鎖の大きさにも依存する。



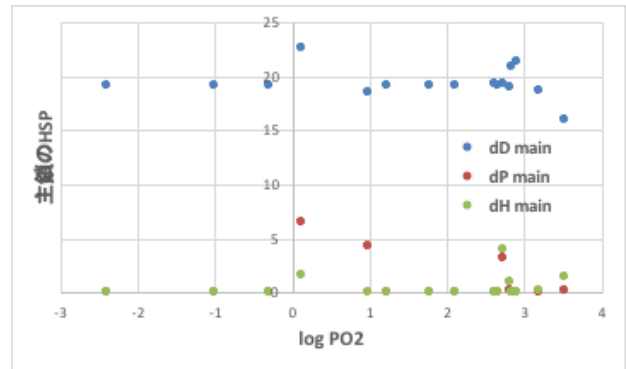
側鎖の体積/主鎖の体積の値が大きなポリマーは、酸素の透過性が高い。(ただし、主鎖と側鎖の定義の仕方にもよる) PVDCは-CCl2-を主鎖にとりまわっているため、側鎖の体積はゼロになっている。そこで、側鎖の大きさによる酸素の透過性は、SP 値や Tg 点から考えるよりは小さくなる。(分子鎖のパッキングが高くなる。)



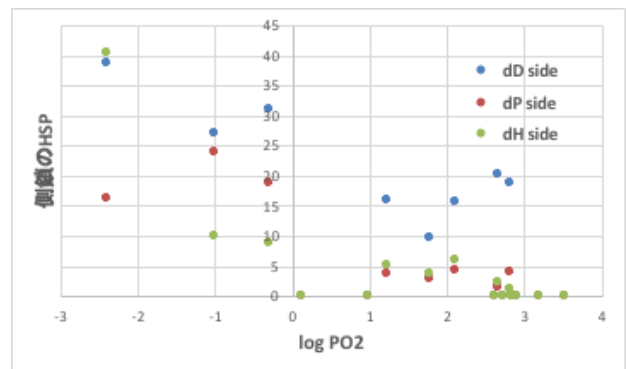
ハンセンの溶解度パラメータ (HSP) は溶解のエネルギーを dD:分散項、dP:分極項、dH:水素結合項に分割したものである。ポリビニルアルコール(PVA)は水素結合項が大きく、ポリアクリロニトリル(PAN)は分極項が大きく、ポリ塩化ビニリデン(PVDC)は分散項が大きい事が明確にわかり、「新しい機能膜」に記載のガスバリア性の機構を定量化するのに一番適している事がわかる。

以上のように、ポリマーの様々な情報がリレーショナル・データベースで連携されていると、必要な情報にアクセスするのが非常に容易になる。ポリマーの、ある機能を考察しようとした場合に、どのような情報が必要か?はケースバイケースで、データベース中に必要な情報が存在しない場合もある。例えば、あるポリマーの SP 値がデータベース中に存在しない場合を考えてみよう。新規なポリマーや、新しい共重合系のポリマーを作った場合には、SP 値がない事も多い。すると、ガスバリア性を改良しようとしたポリマーであっても、性能が上がるかどうかは実験して

見るまでわからないと言うことになる。SP 値がわかればバリアー性を予測できる。SP 値を知るためには、実験をしなければならぬ。それでは、MI による研究の加速には繋がらない。無い情報はリレーショナル化しても出てこない。例えば、ポリマーの主鎖の体積、側鎖の体積は HSP の研究者が独自に開発したパラメータなので、他のデータベースからは入手が不可能だ。しかしこの情報を持っていれば、主鎖の HSP 値、側鎖の HSP 値と言う情報が得られ、その情報とガスバリアー性を関連づける事ができる。



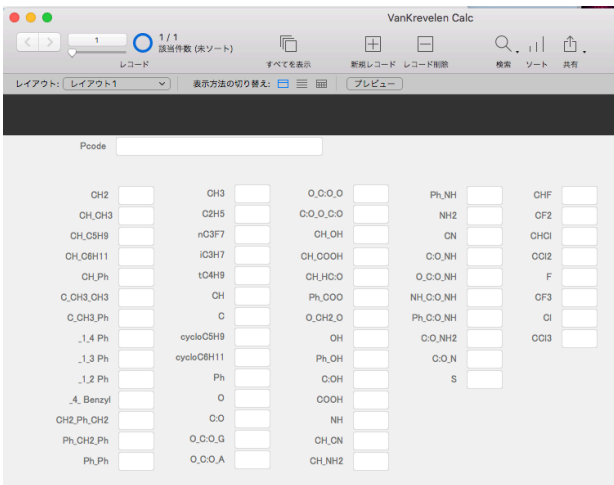
すると、ビニル系のポリマーの主鎖に関してはどのポリマーの HSP もほぼ同じである事が解る。



それに対して、側鎖の HSP はバリアー性の高いポリマーは dP, dH が大きい事が解る。主鎖の体積、側鎖の体積の情報が無ければこのような解析はできない。

それでは、必要なのに DB に無い情報を、どうやって入手するのか?を解説しておこう。ここでは、Filemaker Pro の計算機能を利用する。Filemaker Pro にはフィールドの値を用いて計算を行う機能がある。その使い方を説明しよう。

物性推算式のデータベース化



まず、ポリマー中に含まれる原子団の数を定義するデータベースを作成する。

どんな原子団を定義するかは、ポリマーハンドブック（下テーブル参照）やVan Krevelenの書籍などを参照にする。原子団の数などからポリマー物性を予測する式=原子団寄与法が様々開発されている。

TABLE 2. GROUP CONTRIBUTIONS TO COHESIVE ENERGY DENSITY

Group	F [MPa ^{1/2} cm ³ /mol]			ΔE _v [J/mol]
	Small (112)	van Krevelen (121)	Hoy (61)	
2.1. CARBON-CONTAINING GROUPS				
-CH ₃	437	420	303	
>CH ₂	272	280	269	
>CH-	57	140	176	
<C	-190	0	65	
=CH ₂	388	-	259	
=CH-	277	222	249	
=C<	39	82	173	
-CH=(aromatic)	-	-	240	
-C=(aromatic)	495	560	(479)	
-CH(CH ₃)-	685	641	(672)	
-C(CH ₃) ₂ -	685	641	(672)	
-CH=CH-	454	444	497	
-C=CH-	265	304	421	
CH ₃ -C<	(704)	724	(725)	
H-C=C-	583	-	-	
-C=C-	454	-	-	
Cyclopropyl	-	1380	1300	
Cyclohexyl	-	1660	1470	
Phenyl	1500	1520	(1400)	
p-Phenylene	1350	1380	(1440)	
Naphthyl	2340	-	-	
2.2. OXYGEN-CONTAINING GROUPS				
-O-, ether	143	255	235	
-O-, epoxide	-	-	360	
-OH	-	754	462	
-OH, aromatic	-	-	350	
-(C=O)-	562	685	538	
-(C=O)-O-	634	511	688	
-(C=O)-OH	-	651	(998)	
-O-(C=O)-O-	-	767	(904)	
-(C=O)-O-(C=O)-	-	767	1160	

この予測式をデータベース中に載せてしまう。

例えば、繰り返しユニットの分子量を計算で出すのは簡単だろう。

$$MW = 14.03 \times \text{CH}_2 \text{ 個数} + 28.05 \times \text{CH}_2\text{CH}_3 \text{ 個数} \dots$$

これでポリプロピレン・ユニットの分子量は計算できる。

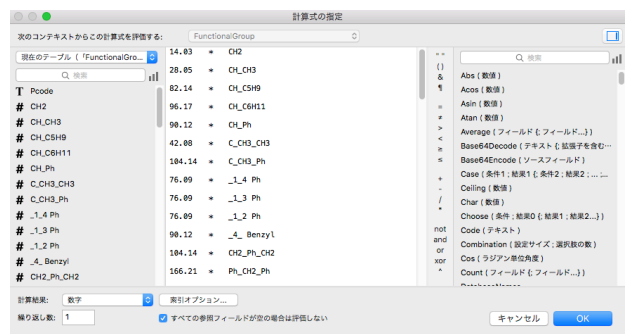
MW	C	D	E	F
14.03 *			CH2	+
28.05 *			CH ₂ CH ₃	+
82.14 *			CH ₂ C ₅ H ₉	+
96.17 *			CH ₂ C ₆ H ₁₁	+
90.12 *			CH ₂ Ph	+
42.08 *			C ₂ CH ₃ CH ₃	+
104.14 *			C ₂ CH ₃ Ph	+
76.09 *			.1,4 Ph	+
76.09 *			.1,3 Ph	+
76.09 *			.1,2 Ph	+
90.12 *			.4_Benzyl	+
104.14 *			CH ₂ Ph ₂	+
166.21 *			Ph ₂ CH ₂ Ph	+
152.18 *			Ph ₂ Ph	+
15.03 *			CH ₃	+
29.06 *			C ₂ H ₅	+
43.09 *			nC ₃ F ₇	+
43.09 *			iC ₃ H ₇	+
57.11 *			tC ₄ H ₉	+
13.02 *			CH	+
12.01 *			C	+

原子団ごとの加算値のテーブルを用意する。

先ほどのデータベースを立ち上げ、管理/データベースから、フィールドのタブを選択する。



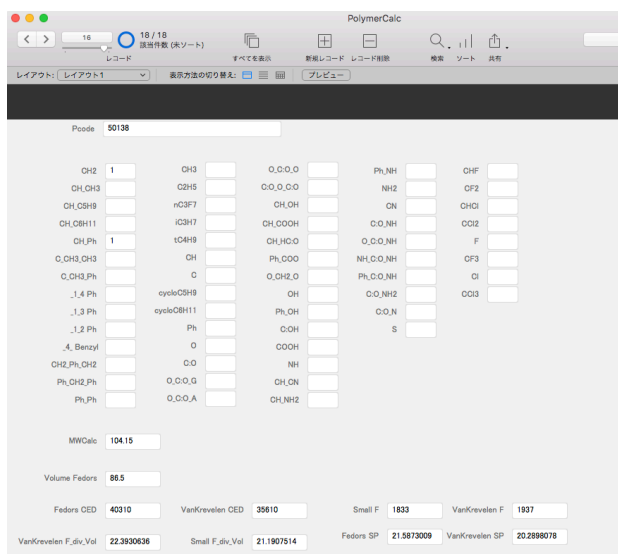
新しいフィールド名、MWCalcを入力して、タイプを“計算”にする。作成ボタンを押すと計算式を入力する画面になる。



そこに、先ほどの MW を計算するテーブル全体をペース

トする。すると、原子団の数を入力すると自動的にユニットの分子量を計算する。

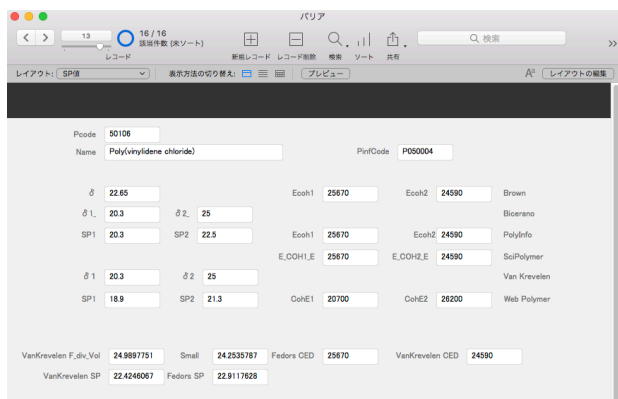
同様に、凝集エネルギー密度、SP 値を計算するフィールドを付け加える。



既存の推算式から、物性値を予測するデータベースが出来上がる。

そして、このデータベースと、バリア材のデータベースをリレーショナル化する。

新たに SP 値のレイアウトを作成し、各種データベースの値と計算値の値を両方表示する事が可能になる。



自分で作成した新たな式とリレーショナル化することも容易である。

様々なデータベースに記載されている SP 値、CED 値はある範囲で与えられている事が多い。また実験値ではなく、Fedors 法などの計算値が収録されている事が、こうした RDB で比較して見ると明らかになる。酸素の透過係数と SP 値、Tg 点などをプロットした時に、直線から外れるポリマーはこうした RDB からデータをチェックし直す必要がある。こうしたことを複数のテーブルから行う事は不可能では無いにしろ、非常に困難である事が解るだろう。PolyInfo のデータベースからは、そのコード番号を含めダ

ウンロードしてあるので、オンラインで簡単に最新のデータにアクセスすることも可能である。この PolyInfo のコードをキーにしてリレーショナル化をするのも一つの選択肢になる。

本来、既存の物性推算式は、作者によって原子団の取り方が異なる。そこで式ごとに計算用の DB を作成しなければならない。ポリマーの構造を見ながら、必要に応じて原子団に分割するのは手作業では大変である。そこで、ポリマー構造から自動で原子団に分割するプログラムを作成してある。ポリマーの両末端に X を付加した、Polymer Smiles の構造式があれば良い。これについては次回説明しよう。

雑感

ポリマーは名称のつけ方が非常に複雑で、ものによっては 10 種類以上の名称を持つものがある。物性値も低分子の熱物性と異なり、合成方法、精製方法、プロセス条件などによって値が大きく異なる。単独成分のポリマーだけではなく、共重合ポリマー、それも成分の入り方がランダムであったり、ブロックであったりする。

そのような材料を扱うときには、どこかで意を決してデータベース化を行う必要があると思う。しかし、データベース化するにしても、様々なツール開発が不可欠で、それがデータベース化を遅らせていると思われる。

ツールが揃ってくると、MI を使った材料開発も容易になってくる。

Excel のテーブルの DB 化、キーを使ったリレーショナル化、計算機能を使った予測用 DB の作り方はきちんと理解しておこう。

ニューラルネットワークというのは、ニューロンとシナプスが複雑に結合して、脳の複雑な機能を発揮する。一説によると脳には 1000 億個のニューロンがあるそうだ。この数は、この銀河の恒星の数と同じだ。そのニューロンが 1 つあたり平均 1 万個のシナプスで結合されている。そして学習によってシナプスの結合強度が変化していく。それを模して開発されたニューラルネットワーク法は情報伝達のスピードは脳を圧倒的に凌駕するが、素子同士の結合や結合強度まで考えると、まだまだ脳にしかできないことは残るだろう。

データに関しても同じで、構造化された超巨大なビッグデータはコンピュータにとっては合理的かもしれないが、人間の脳にとっては意味をなさない。文献一つが 1 つのマ

イクロ・DB でも良いから作り始めるのが大事だろう。それがリレーショナルというシナプスで合理的に結合されたときに大いに研究が加速するのだと思う。優れた人間の研究者は、コンピュータが存在する前からこれができていた。そして、リレーショナル化の際に、興味や本人の資質などでフィルターがかかるので、同じ論文を読んでも、その後が変わってきて“個性”になる。

ニューラルネットワーク法の本質は、誰が、何を、どう教えるのか？ だ。NNへ教える教育者の資質と教育用の資料がAIの能力を決める。そして出来上がったAIは教育者にとって一番“個性”の近いAIになる。そうしたAIに“AIアシスト”してもらいながら研究を進めれば、個性的な研究が加速するのだらう。

汎用のスーパーAIが研究を助けてくれる時代はすぐに来るだろう。論文、特許、ネットを調べて、様々な情報を自動的に収集して構造化してくれるかもしれない。そうした雑用は大いにAIに頼んでしまおう。そのときに、人間は、リレーショナルというシナプスで、どう個性化していくかが問われる時代になっていくのだらう。

Pirika [マテリアル・ゲノム](#) のページ

以下PDF

[第1回 インTRODクシヨN](#) 2018.8.23

[第2回 データ収集と昔ながらのやり方](#) 2018.8.24

[第0回 物性推算と逆設計と呼んでいた時の話](#) 2000.8.28

なんと18年前！

[第3a回 ポリマー設計と3つのMI \(その1\)](#) 2018.9.3

[第3b回 ポリマー設計と3つのMI \(その2\)](#) 2018.9.3

[第4a回 MIに適した簡単なデータベースの利用法](#)
2018.9.4

[第4b回 複雑なポリマーのデータベース化](#) 2018.9.7

プレゼン用：[MIを使う時のデータベース構築法](#)

2018.9.11

プレゼン用：[複雑なポリマーの設計とDB](#) 2018.9.15

[第5回 データのクレンジング](#) 2018.8.28

[第6a回 ニューラルネットワーク法の初歩](#) 2018.9.25

[第6b回 ニューラルネットワーク法を使った Drug Design](#) 2018.9.22

第7回 遺伝的アルゴリズム(GA)を理解しよう