MAGICIAN 養成講座

MAterials Genome/Informatics and Chemo-Informatics Activate Networks

第 6b 回 再構築学習法を取り入れた

ニューラルネットワーク法とドラッグ・デザイン

2018.9.22　横浜国大非常勤講師　山本博志

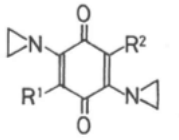
Materials Genome を使って医薬品を設計することを考えてみよう。かと言って、医薬品のリード化合物を設計するのは実に難しいことである。そこで、既にある医薬品の基本骨格を利用して、置換基を様々に変えた時に薬理活性がどう変化したか？を解析し、薬理活性を上げる置換基とその位置を特定する方法を学ぼう。製薬会社は高度な計算機科学を駆使して行っている研究である。これまでの研究蓄積と、高速なコンピュータを使って高精度な分子軌道計算などから“分子の遺伝子“を特定して、設計を行っている。そのような方法は授業で教えられるものでも無いので、もっと簡単な Materials Genome を使ったやり方を解説しよう。

準備

まず、青山智夫先生の「ニューラルネットワークによる薬物の構造活性相関に関する研究」という学位論文をネットからダウンロードしよう。実は、筆者が使っているニューラルネットワーク法のプログラムは、青山先生が開発された再構築学習法を取り入れたものである。共立出版の「階層型ニューラルネットワーク　非線形問題解析への応用」市川紘著(1993)に Fortran のプログラム(PSDD: Perceptron Simulator for Drug Design)が記載されている。昔は JCPE(日本化学プログラム交換機構：Japan Chemistry Program Exchange)からダウンロードできたが、今は手に入らないようだ。化学系の学生が化学系のプログラムのソースコードを読む機会が失われたのは残念なことではある。（ソースコードを読んで見たい学生はこっそり連絡をいただきたい。JCPE で配布されたコードは手元にある。）とあれ、25 年以上前に開発されたニューラルネットワーク・シミュレータで Drug Design をどのように行っていたのか、青山先生の学位論文に記載の例題を実際に解析してみよう。ニューラルネットワーク・シミュレータ(NNS)自体は、最近、様々なものがネットからダウンロードできる。興味のある学生は、データセットだけ昔のものを使って、最新のシミュレータの実力を調べてみよう。

例題解析

カルボキノン誘導体は、制ガン抗生物質マイトマイシン C に基づいて合成された化合物群で実際に医療に用いられている。



置換基、R1, R2 を変えた時の生物活性値が求められている。これらの化合物の活性値を予測するために、説明変数を準備する。（この部分が、こうした研究を行ったことがない初学の研究者やコンピュータしか知らないデータサイエンティストには難しいところになる。）

ペースト　フォント　配置　数値　セルのスタイル

E12　2.28

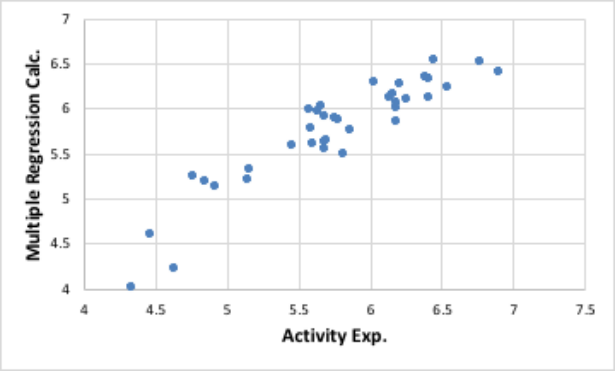
	A	B	C	D
1	NO	R1	R2	Activity
2	1	CH3	COCH3	
3	2	C6H5	C6H5	4.33
4	3	CH3	CH2CH2CH2Ph	4.47
5	4	C5H11	C5H11	4.63
6	5	CH(CH3)2	CH(CH3)2	4.77
7	6	CH3	CH2Ph	4.85
8	7	C3H7	C3H7	4.92
9	8	CH3	CH2OPh	5.15
10	9	CH2CH2OCON(CH3)2	CH2CH2OCON(CH3)2	5.16
11	10	C2H5	C2H5	5.46
12	11	CH3	CH2CH2OCH3	5.57
13	12	OCH3	OCH3	5.59
14	13	CH3	CH(CH3)2	5.6

青山先生たちは、分子屈折定数(MR)、疎水性定数(π)、置換基定数（F および R）、立体効果と疎水定数を推定するための $MR_{1,2}$、$\pi_{1,2}$ の 6 種類の説明変数(識別子:Descriptor)を用いている。これらの識別子の意味は置いておいて、まずは解析を行うためにデータを打ち込んでみよう。（近い将来、このような打ち込み作業は AI がやってくれるかもしれない。しかし現状ではまだまだ誤認識が多い。汚いデータが入ってくると、修正する手間は、経験上、自分で打ち込む倍以上の労力がかかる。）

データ打ち込みが終わったら、まずは、重回帰計算を行ってみる。ソフトウエアーは授業で配布した YSB を用いるか、Excel に搭載されている分析ツールを使ってもよい。

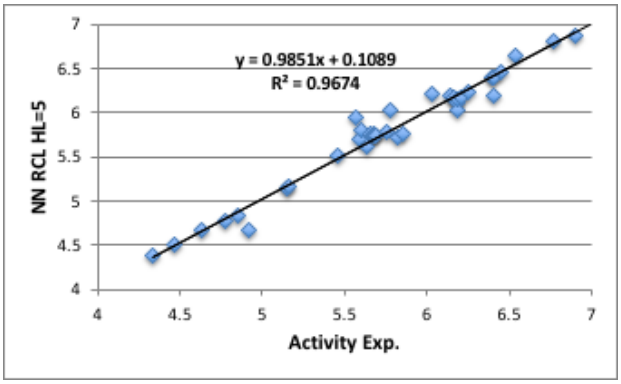
$$Activity = -0.02169*MR_{1,2} - 0.09473*\pi_{1,2} - 0.4152480*\pi_2 - 0.274901*MR_1 - 1.771798*F - 0.749842*R + 6.174982$$

容易に重回帰式が求まり、それをグラフ表示すると下図のようになる。大きく外れるものは、打ち込み間違いのこともあるのでデータを見直す。

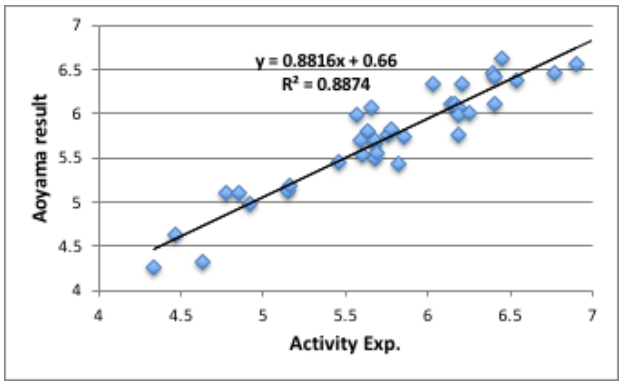


何故、NN 法で推算する前に必ず重回帰法で計算するのかというと、NN 法は非常にフィティング性能が高いので、多少の打ち込みエラーがあってもフィティングしてしまい、どのデータが誤りかがわからなくなるからだ。（例年いくら言っても端折る学生はいるが、“学んで”もらうしかない。）

次に同じデータを用いて、ニューラルネットワーク法で学習させてみる。



学位論文では、入力層=7、中間層=12、出力層=1 で計算しているが、筆者の NNS は流石に 20 年以上改良を重ねているので、中間層は 5 でも十分高い精度で収束している事がわかる。比較として学位論文中のデータをプロットすると下図のようになる。



通常はここまで出来上がると、Leave-n-out(L-n-o)解析を行う。これは、学習データの中から、n 個を外してモデル式を作成し、そのモデル式を用いて、外した n 個を予測して予測性能を検証する。しかし、自分は L-n-o はあまり意味が無いように思っている。例えば今回作成したモデルで、一番エラーの大きな 5 つを除いてモデルを作成し、その 5 つを予測してみれば、かえって誤差は大きくなる。L-n-o が成立するのは、もっとビッグデータが存在して、その大多数が線に近いもので、少しの外れるものに使った場合だけではないかと思う。それよりは、外れるものの特徴をよく考え、新たな識別子を導入するなどが実際には大事であると考えている。（例えば、フェニル基が入ると常に外れるなど。）

学位論文によれば、L-n-o の結果は良好だということで、ここまでできれば、次は置換基を片っ端から色々なものに変えて、活性が最大になる構造を探索するだけになる。そこで問題なるのが、識別子の問題である。青山先生たちは、分子屈折定数(MR)、疎水性定数(π)、置換基定数（F および R）、立体効果と疎水定数を推定するための $MR_{1,2}$、$\pi_{1,2}$ の 6 種類の識別子を用いている。それでは、新しい置換基を考えたときに、その置換基の 6 つの識別子をどうやって手に入れたら良いのだろうか？分子屈折は屈折率、疎水定数は log P(オクタノール/水分配比率)などから導き出される定数であるが、置換基定数はどうやって求めたら良いのだろうか？

また、これらの定数は置換基が 2 つある時の定数で、どう取り扱ったら良いのかわからない。論文などを遡る必要がある。この分野に明るければ、これらの定数は Hansch 法のパラメータであることがわかる。

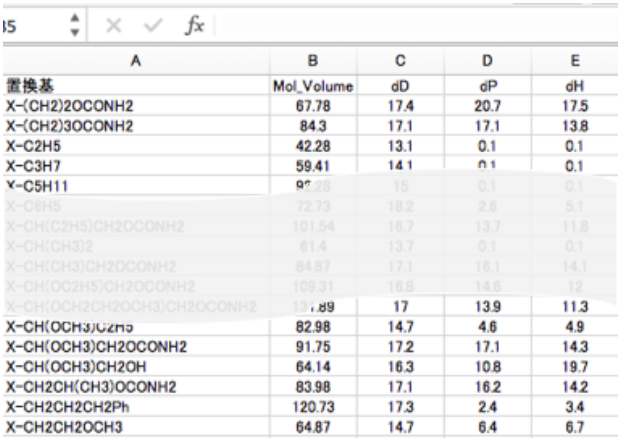
「定量的構造活性相関　Hansch 法の基礎と応用」C. Hansch / A. Leo 江崎俊之訳　地人書館

には、229 種類の官能基の分子屈折定数(MR)、疎水性定数(π)、置換基定数（F、R）が記載されている。

それであっても、この学位論文の置換基を全ては網羅して

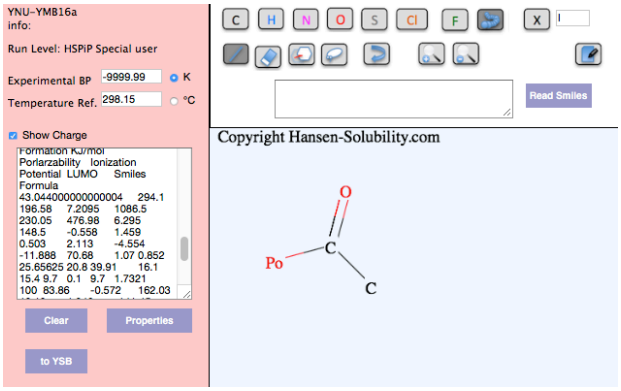
いない。そこで新しい官能基を考えた場合、自分で6種類の識別子を作り出さなくてはならない。置換基のペアになった時の識別子の混合ルールを覚えなくてはならない。
そこで、最近の風潮としては、DFT 使った高精度の分子軌道計算やって、HOMO, LUMO、ダイポールモーメント、分子の正負に荷電した表面積など計算結果を識別子にしてしまえ、という乱暴な方法に頼ることになる。すると、スーパーワークステーションと高価なソフトウエアーが必要になり、一般人から遠い世界の研究開発になる。それは、それで専門家に任せて、みんなが楽しめるマテリアルズ　インフォマティクス(MI)で医薬品を設計してみよう。MI の大元は、Chemo-Informatics(化学情報学)である。化合物の情報は分子構造の中に存在する。疎水性定数の意味合いは明瞭で、水酸基やアミド基など水溶性の高い官能基は値が小さくマイナスになる。疎水的な官能基はプラスになる。分子屈折定数（MR１）は重原子の数が増えるにつれて大きくなる値なので、置換基の大きさを表している。F（場誘起効果）R（共鳴効果）はハメット定数に由来する定数である。ハメット式は1935年ごろ、安息香酸類のイオン化に及ぼす置換基の効果として定数化された。昔のやり方は、ある置換基を m 位や p 位に導入した時の、安息香酸エステルの加水（加溶媒）分解速度から定数化するものだった。当然ながら、複数の置換基が入り込んでくると、加成性はないのでぐちゃぐちゃになる。
近年は分子軌道計算などの結果と置き換わってきているが、実際の設計段階では、置換基定数として与えられているF、R 定数の方が逆設計上便利である。

m 位や p 位の特異性は検討できないが、溶解度パラメータを置換基の識別子として使う簡便な方法を紹介しておこう。ハンセンの溶解度パラメータ（HSP）は、蒸発のエネルギーを、分散項(dD)、分極項(dP)、水素結合項(dH)に分割した3次元の溶解度パラメータとして知られている。似たベクトルの溶媒は似たベクトルの溶質を溶解するので、医薬品の吸収などの解析に用いられている。様々な利用方法が筆者の HP(www.pirika.com)に記載されている。この溶解度パラメータと分子体積を各原子団に対して決定する。



A	B	C	D	E
置換基	Mol_Volume	dD	dP	dH
X-(CH2)2OCONH2	67.78	17.4	20.7	17.5
X-(CH2)3OCONH2	84.3	17.1	17.1	13.8
X-C2H5	42.28	13.1	0.1	0.1
X-C3H7	59.41	14.1	0.1	0.1
X-C5H11	93.28	15	0.1	0.1
X-C6H5	72.73	18.2	2.6	5.1
X-CH(C2H5)CH2OCONH2	101.54	16.7	13.7	11.8
X-CH(CH3)2	61.4	13.7	0.1	0.1
X-CH(CH3)CH2OCONH2	84.87	17.1	16.1	14.1
X-CH(OC2H5)CH2OCONH2	109.31	16.8	14.6	12
X-CH(OCH2CH2OCH3)CH2OCONH2	133.89	17	13.9	11.3
X-CH(OCH3)C2H5	82.98	14.7	4.6	4.9
X-CH(OCH3)CH2OCONH2	91.75	17.2	17.1	14.3
X-CH(OCH3)CH2OH	64.14	16.3	10.8	19.7
X-CH2CH(CH3)OCONH2	83.98	17.1	16.2	14.2
X-CH2CH2CH2Ph	120.73	17.3	2.4	3.4
X-CH2CH2OCH3	64.87	14.7	6.4	6.7

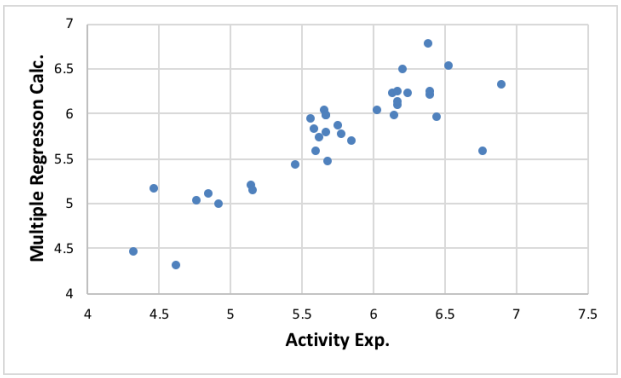
パラメータの決定には、授業で配布している YMB を用いる。分子のお絵かきをして計算結果を取り出すだけなので非常に簡単に行える。注意点としては、官能基の先頭が炭素以外の酸素、窒素だった場合には、ダミーアトム（下図で Po と表記）は使えない。C（炭素）を付加して、計算結果からメチル基分を差し引く。このようにして、自分の用いたい置換基の HSP と置換基体積の一覧表を作成しておく。



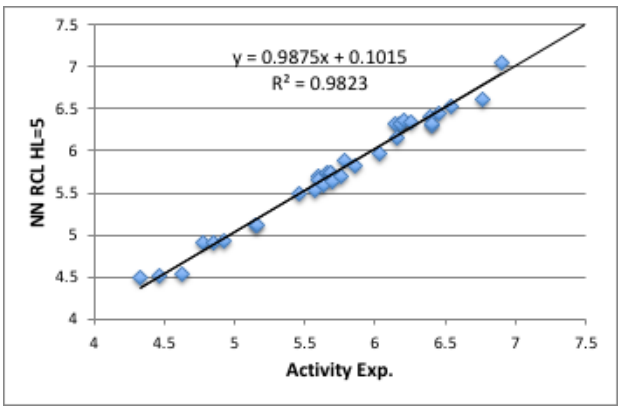
分子を描いて Properties ボタンを押すと HSP も含め各種計算結果が出力される。

A	B	C	D	E	F	G	H	I	J
NO	Activity	Mol_Volume1	dD-1	dP-1	dH-1	Mol_Volume-2	dD-2	dP-2	dH-2
2	4.33	72.73	18.2	2.6	5.1	72.73	18.2	2.6	5.1
3	4.47	22.7	11.2	0.1	0.1	120.73	17.3	2.4	3.4
4	4.63	92.28	15	0.1	0.1	92.28	15	0.1	0.1
5	4.77	61.4	13.7	0.1	0.1	61.4	13.7	0.1	0.1
6	4.85	22.7	11.2	0.1	0.1	87.48	17.8	2.3	4.6
7	4.92	59.41	14.1	0.1	0.1	59.41	14.1	0.1	0.1
8	5.15	22.7	11.2	0.1	0.1	93.68	18.3	4.9	6.4
9	5.16	101.75	17.6	14.1	7.6	101.75	17.6	14.1	7.6
10	5.46	42.28	13.1	0.1	0.1	42.28	13.1	0.1	0.1
11	5.57	22.7	11.2	0.1	0.1	64.87	14.7	6.4	6.7
12	5.59	43.2	16.4	8.6	8.8	43.2	16.4	8.6	8.8

そして、先ほどの6つの識別子の代わりに、2組の [体積、dD, dP, dH]を識別子として使ってテーブルを作成する。分散項(dD)は分子の屈折率と相関があるパラメータであり、HSP の組みが疎水性を表現するので、誘起や共鳴は表現していないが、医薬品の溶解性に関しては必要な情報は揃っている。このテーブルを、まずはお決まりの重回帰計算を行う。数点大きく外れるケースもあるが、大まかには悪く無い識別子であることがわかる。

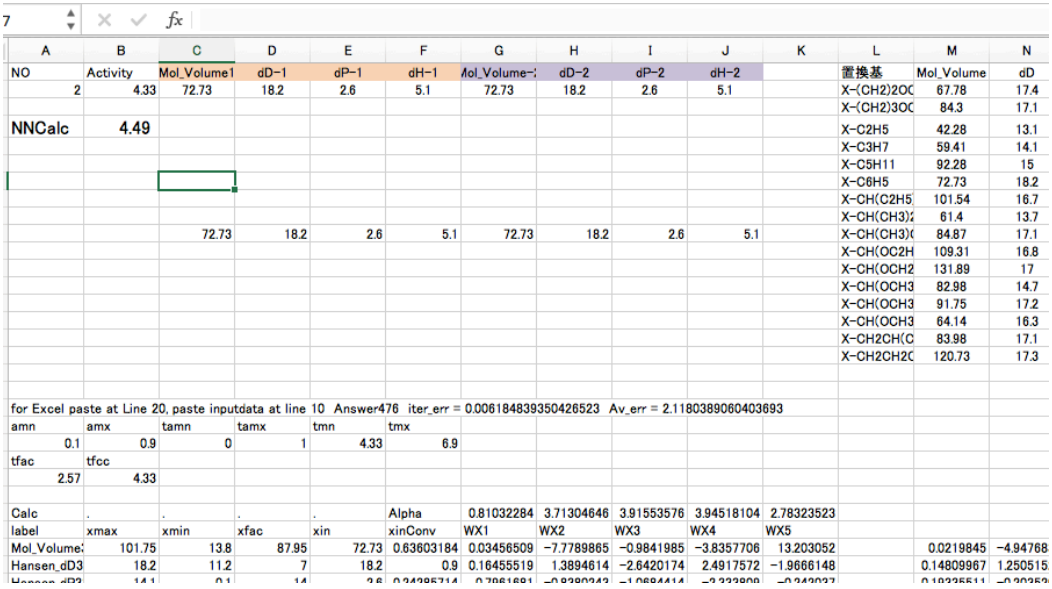


次に、このテーブルをニューラルネットワークに、再構築学習の頻度は1/500, 中間層=5 で学習をさせる。結果としては、下図に示すように良好に活性を表現できていることが分かる。この HSP で収束したことは、次のことを示している。この化合物の活性は、溶解性が支配的で、置換基が電子を押したり、引いたり、共鳴したりはあまり関係ないと言える。



この予測式が成立すると、置換基の識別子として[体積、dD, dP, dH]が得られれば、活性を予測でき、さらに活性の高いカルボキノン誘導体を分子設計できることが可能になる。HSP は最近、dH をドナー性、アクセプター性に分割する試みもなされている。次世代バージョンでは dH_{ED}, dH_{EA} などの指標も使えるようになる。そうした新しい識別子も含めて置換基のデータベースを作成し、新たな医薬品設計を行うことが、普通の PC と授業で使っているソフトウエアーだけでできてしまう。それが、Materials Genome の面白さであり、醍醐味である。逆に言えば、高価なソフトと高価なハードウエアーが医薬品開発の参入障壁であった時代は終わったということだ。

ちなみに筆者の作成した NNS は出力として、JAVA(C#) のプログラムを出力するようになっているので、プログラムが書ければ、置換基を順番に入れ替えて全ての組み合わせを計算して一番高い活性の構造を取り出すことも可能である。また、Excel の計算式としても出力される。



NO	Activity	Mol_Volume1	dD-1	dP-1	dH-1	Mol_Volume-2	dD-2	dP-2	dH-2
2	4.33	72.73	18.2	2.6	5.1	72.73	18.2	2.6	5.1
NNCalc	4.49								
		72.73	18.2	2.6	5.1	72.73	18.2	2.6	5.1

置換基	Mol_Volume	dD
X-(CH2)2OC	67.78	17.4
X-(CH2)3OC	84.3	17.1
X-C2H5	42.28	13.1
X-C3H7	59.41	14.1
X-C5H11	92.28	15
X-C6H5	72.73	18.2
X-CH(C2H5	101.54	16.7
X-CH(CH3)2	61.4	13.7
X-CH(CH3)C	84.87	17.1
X-CH(OC2H	109.31	16.8
X-CH(OCH2	131.89	17
X-CH(OCH3	82.98	14.7
X-CH(OCH3	91.75	17.2
X-CH(OCH3	64.14	16.3
X-CH2CH(C	83.98	17.1
X-CH2CH2C	120.73	17.3

for Excel paste at Line 20, paste inputdata at line 10 Answer476 iter_err = 0.006184839350426523 Av_err = 2.1180389060403693

amn	amx	tamn	tamx	tmn	tmx
0.1	0.9	0	1	4.33	6.9
tfac	tfcc				
2.57	4.33				

Calc	.	.	.	.	Alpha	0.81032284	3.71304646	3.91553576	3.94518104	2.78323523
label	xmax	xmin	xfac	xin	xinConv	WX1	WX2	WX3	WX4	WX5
Mol_Volume:	101.75	13.8	87.95	72.73	0.63603184	0.03456509	-7.7789865	-0.9841985	-3.8357706	13.203052
Hansen_dD3	18.2	11.2	7	18.2	0.9	0.16455519	1.3894614	-2.6420174	2.4917572	-1.9666148

0.0219845	-4.94768	
0.14809967	1.250515	

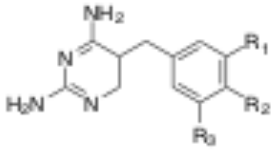
Excel のマクロが記載できるなら、置換基データを順に入れ替えながら、計算結果を取り出すことも可能である。大事な点は、もし新しい実験を行って結果が出たら、すぐにニューラルネットワークを組み直せるかどうかであろう。使えば使うほど、性能が高くなっていく。論文を書いたら終わりの大学の研究とは違い、企業の研究は競争に勝つために、開発スピードをどうあげていくか？　が焦点になる。そのために、Materials genome や Materials Informatics を

利用したいということを忘れてはならない。また、そうした再計算、再アプリケーション化が、その道の専門家に頼らなくてはできないのだとすると、スピードアップは期待できない。自分でやるしか無いし、AI の教育係としては、そちらの方が適している。

ともあれ、25 年前にできていた Drug Design のシステムでも有効に働くことが再確認できたし、識別子としても HSP（これも 50 年前のパラメータではあるが）が有効であることがわかった。それではもう少し複雑な系でも確認しておこう。

応用練習

Nonlinear Quantitative Structure-Activity Relationship for the Inhibition of Dihydrofolate Reductase by Pyrimidines
J. Med. Chem. 1996, 39, 3526-3532



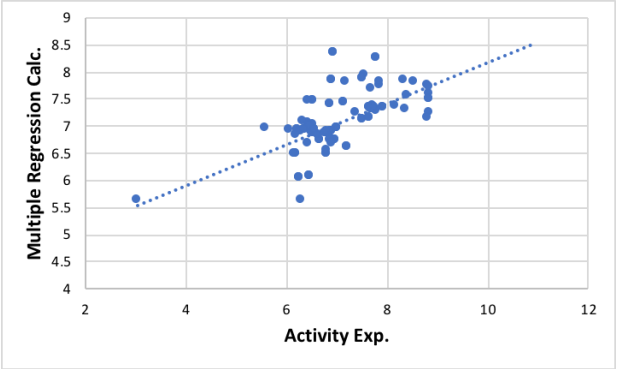
この論文に Pirimidine 誘導体の薬理活性のデータが記載されているので、これを解析してみよう。R1, R2, R3 の置換基の種類と活性が記載されているので、置換基の[体積、dD, dP, dH]を全て計算しておこう。

no.	substituents	activity meas	sd
			Set 1
11	4-F	6.35	
31	4-NHCOCH$_3$	6.89	
34	3-Br	6.96	
42	3,5-(OCH$_3$)$_2$, 4-(CH$_2$)$_2$OCH$_3$	8.35	
20	3-OCH$_2$CONH$_2$	6.57	
24	3-CH$_3$	6.70	
30	4-O(CH$_2$)$_4$CH$_3$	6.89	
23	3-Cl	6.65	
37	3-CF$_3$	7.02	
08	3-CH$_2$OH	6.28	
39	3-I	7.23	
			Set 2
16	3-OH	6.47	
54	3,5-Cl$_2$, 4-NH$_2$	8.87	
43	3,5-(OCH$_3$)$_2$	8.38	

（困ったことに、38. COH2CH2OCH3 とか、62. OCH2O とかいう多分誤植の置換基があるが、これらは除いて考える。）

この解析の場合は、カルボキノン誘導体の場合には無視できた、置換基の位置特異性（同じ置換基でも、m 位、p 位に付く時で活性が異なる）ケースが生じる可能性がある。特に R1 と R3 は対称な関係の同じ m 位置換なので、３、４置換と５、４置換は学習させるデータに工夫が必要である。

まず、お決まりの重回帰計算を行うと下図のようになる。



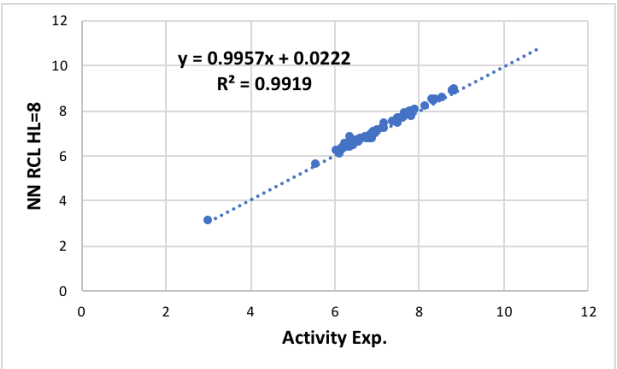
一点大きく外れる化合物は、#10 の 3,5 位に（m 位に）CH2OH が付く化合物である。＃１の 3,5 位に（m 位に）OH が付く化合物は今回検討されている化合物の中で最も活性が低い。CH2OH と OH では識別子の値が非常に似通ってしまうので、このような結果になったと思われる。

Activity=
0.00191*Vol-0.0367*dD+0.08908*dP-0.04885*dH+
0.00197*Vol-0.08230*dD+0.0089*dP+0.024390*dH+
0.00191*Vol-0.0367*dD+0.08908*dP-0.04885*dH
+9.04893

重回帰式から明らかなように、３位と５位の計算値は同じになる。例えば OH が付加する場合、[体積、dD, dP, dH]は、[16.5, 21.1, 21.9, 45.6] で同じであるが、

３位、５位
0.00191*Vol-0.0367*dD+0.08908*dP-0.04885*dH=−1.02
４位
0.00197*Vol-0.08230*dD+0.0089*dP+0.024390*dH=-0.40

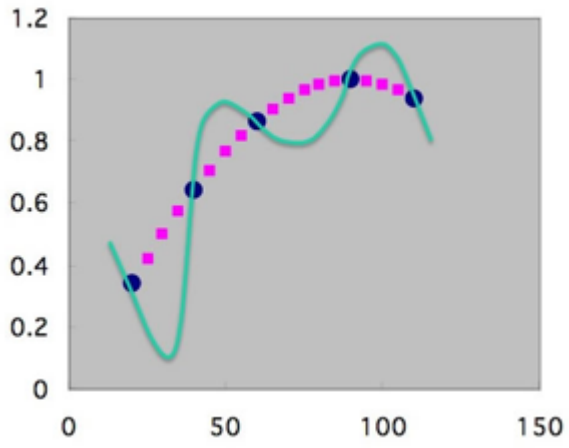
となる。つまり同じ置換基であっても、付加する位置によって活性が異なる事が表現できている。

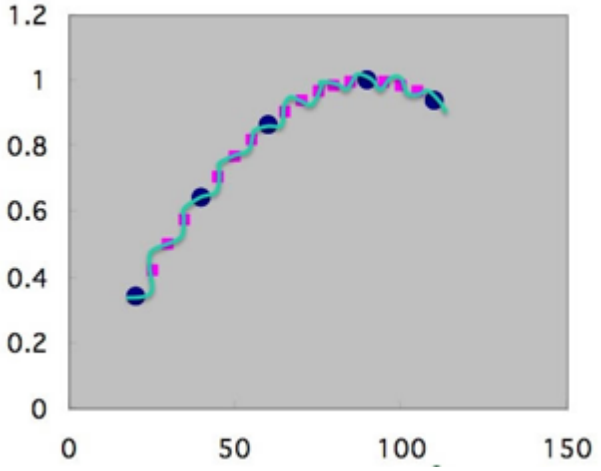


次にニューラルネットワークに、再構築学習の頻度 1/500, 中間層=8 で学習をさせる。結果は上図に示したように綺麗に収束した。このケースでも、化合物の活性は溶解性が支配的で、置換基が電子を押したり、引いたり、共鳴したりはあまり活性に関係ない事が示唆される。

これらの例のように、ニューラルネットワーク法を使った QSAR（定量的構造活性相関）は 2000 年ごろにはブームであった。しかし、ブームは急速に終焉し、サポートベクターマシン（SVM）、ベイズ統計などに移って行った。20 年の時を経て、またディープラーニング法というニューラルネットワーク法がブームになっているが、20 年前は何故ブームが終わり、今回は一時期のブームで終わらないと言えるのか？　筆者の見解も示しておこう。

問題は、QSAR 式の使い方にある。ある置換基の 3 つの組みが決まれば、活性が決まる。置換基は、今回の場合３組の体積と HSP、[体積$_1$、dD$_1$, dP$_1$, dH$_1$], [体積$_2$、dD$_2$, dP$_2$, dH$_2$], [体積$_3$、dD$_3$, dP$_3$, dH$_3$]で表すことができる。そこでニューラルネットワークの学習によって、活性と置換基の持つ識別子（体積と HSP）の関係を正しく表現できているのなら、予測性能も高いと期待してしまう。しかし多くの場合、この予測は裏切られる。前の 6a 回で述べたようにデータ数を増やせば、この過学習はある程度抑えることができる。

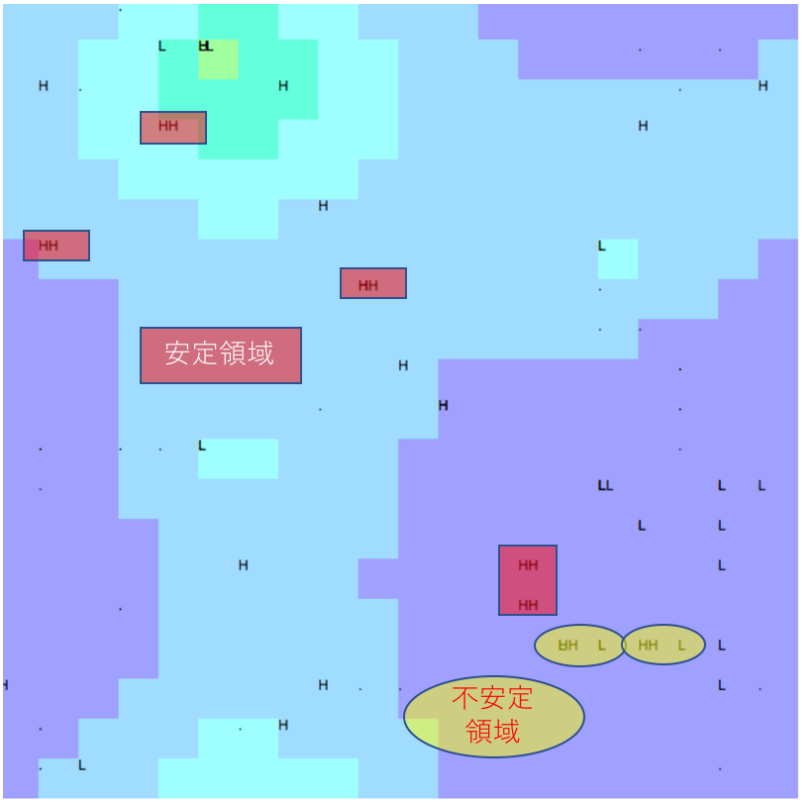


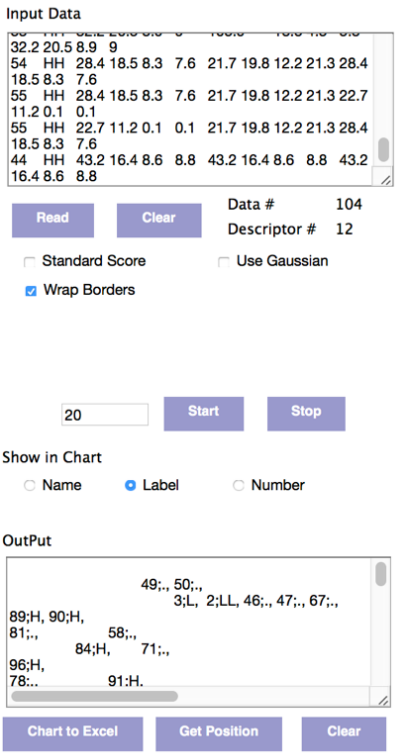


そこで、ビッグデータがあればニューラルネットワーク法の QSAR でも予測性を高くする事は可能である。こうした Sin カーブであれば、自由にデータ数を増やす事は可能であろう。しかし、医薬品の開発では、置換基の HSP は同じで、体積だけ少し小さくというデータは取ることができない。ポリマーの設計の回で、量子化ベルトの話をしたが、飛び飛びの値しか取れない。

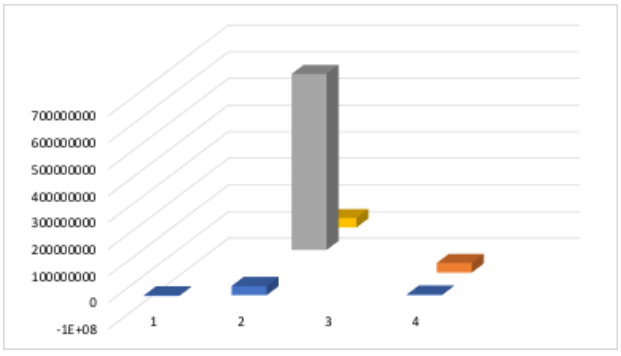
そこで、ある現象を、ある識別子で表現できた（収束した）後には、自己組織化ニューラルネットワーク（SOM）で検証する。SOM: Self-Organizing Map 法は、多次元ベクトルを 2 次元にマッピング方法で、1981 年にコホネンによって開発された。この SOM 法の解説は他の回に譲るとして、簡単に説明すると、似たようなベクトルは 2 次元上の似た位置にマッピングされると言う特徴を持っている。つまり、今回３組の[体積、HSP]は 12 次元ベクトルで、この 12 次元ベクトルが、対称も含め 104 種類存在する。この 104 種類の 12 次元ベクトルを、似たベクトルは似た位置にマッピングする事ができる。口で説明すると難しいので、エクセルのテーブルをコピーし、授業で配布した SOM のアプリケーションにペーストし、プログラムを走らせてみよう。

104 データを全て表示すると見にくくなるので、
活性が 8.6 以上のものを HH,
活性が 7.74 以上 8.6 未満のものを H
6 以下のものを LL,
6.5 以下 6 以上のものに L というラベルをつけた。

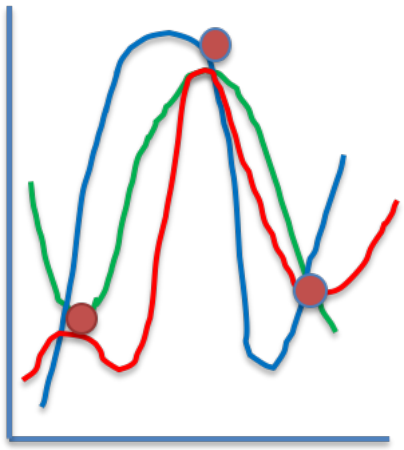




SOM の結果はキャンバスに表示されるが、Chart to Excel ボタンを押し、結果をコピーし、エクセルにペーストする事ができる。細かい検討はエクセルで行った方が便利である。HH の近辺を調べよう。ほとんどの HH はその周辺に、LL, L, “.”、H はいない。ところが、２箇所ほど HH の周辺に L がくる領域がある。その近辺では活性は L から HH に急変する。



それを模式図として表すと下図のようになる。



ニューラルネットワーク法で収束するという事は、活性が大きく変化する点を滑らかにつなぐ曲線をニューラルネットワークが見つけたということになる。しかし、そのニューラルネットワークの中身はブラックボックスで、赤、青、緑のどのような曲線でつないだかはわからない。これらの点の中間ぐらいの修飾基があれば良いのだが、修飾基の「HSP を同じにしたまま体積を減らす」などという事はできない。従って多くの場合は不安定領域では、たとえ NN 法がよく実験値を再現できていても予測性能は出ない。安定領域にいる、HH、H などの修飾基を変化させた時の活性を QSAR で予測して、開発する際の優先順位をつけて行く。つまり、定量的解析と定性的解析を組み合わせて評価していかないと欲しい答えは見つからない。ここで大事なのは、H、M、L のラベリングは研究者の主観でつけ

たものだという事である。つけ方次第で安定領域、不安定領域は変わる。つまり研究者のセンスで考え方が変わる。勉強のしがいがある科学と言える。
ディープラーニングなどの技術では、粗視化層の設計が重要になる。画像解析の場合には、目、鼻、口とかで粗視化層を設計して行くようであるが、構造式の場合はこうした自己組織化ニューラルネットワーク（SOM）で計算したマップを粗視化層に使うのも一つの考え方であろう。

Indian Journal of Advances in Chemical Science 4(4) (2016) 374-385
なども、おもしろい論文なので、時間を見つけて書き足して行こうと思う。

雑記

ニューラルネットワーク・システム(NNS)自体は、今や非常に簡単に入手できる。自分は、化学には化学特有の問題があり、それは数学者には理解できないと思っている。そこで、青山先生が作られ、自分が成長させた NNS が、意味が無いとは思ってはいない。もし、化学系の学生がシステム構築の進むのであれば、最新の NNS に化学特有のオントロジーを組み込んで欲しいと思う。
NNS はそのうち良いものが使えるようになる。プログラミングなどに興味のない大多数の化学系の学生は、NNS に食べさせる識別子をどれだけ合理的に作り出せるか？を自分の売りにして欲しいと思う。
3 つの位置に 100 種類の官能基を順番につける。
全部で 1000000 種類の Pirimidine 誘導体を全部 DFT 計算して一番良いものを選ぶ。
30 分で、100 種類の官能基の HSP 識別子を作り出し、1 時間でニューラルネットワークに学習させ、3 分で 1000000 種類の Pirimidine 誘導体を評価する。
DFT 計算もできるに越したことは無いが、企業が欲しいのはどちらの人材だろうか？　よく考えて欲しい。

最近の AI ブームで、化学系の学生も、計算機化学系を志望する事が多くなってきていると聞いている。ハメット定数など、化学の深淵を覗かなくても簡単に識別子が手に入るのだから、勉強しなくても、楽して稼げる。
そんな、楽なやり方は、AI のもっとも得意とするところだ。近い将来、勝手に DFT 計算しながら材料設計する AI ができても自分は驚かない。その時に AI に仕事を明け渡すか、自分の考えを組み込んだニューラルネットワークに AI アシストしてもらいながら仕事をするか。
人生観が問われる。

ハンセンの溶解度パラメータ(HSP)を識別子に使うことの有効性を今回紹介した。HSP を原子団に割り振るのは、少しテクニックが必要だが慣れれば簡単だ。欧米の医薬品メーカーは全て HSPiP という HSP を扱うソフトウエアーは導入済みである。最近、さらに HSP を Materials Genome, Informatics で扱いやすくした、CLI(Command Line Interface)バージョンの HSP の需要が高まっている。このバージョンは、Pipeline Pilot を使って高速に MI を回すのに適している。日本でもやっと引き合いが来始めたが、最初に購入するのはどこであろうか？
CLI バージョンは個人で買える値段では無いが、学生バージョンの YMB が使いこなせれば、今回紹介した Drug Design ぐらいは、より簡単にできるだろう。
（逆に Pipeline Pilot や Material Studio は個人で買えるようなソフトでは無いので、その中でどう使うのかは、自分は知らない。誰か、これらのソフトを寄付してくれないものだろうか？　とは言っても、自分は Mac なので寄付されても動かないのだが。。。。）

Pirika マテリアル・ゲノム のページ
以下 PDF
第 1 回 イントロダクション 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