

2P13 分子構造重ね合わせのための Hopfield Neural Network のエネルギー関数の検討

○ 山田吉朗、荒川正幹、船津 公人
豊橋技術科学大学

1. はじめに

CoMFA(Comparative Molecular Field Analysis)法による 3D-QSAR(Quantitative Structure-Activity Relationship)では、解析に用いる化合物をあらかじめ適切に重ね合わせておくことが要求される。重ね合わせ手法は、様々なものが提案されており、当研究室においても Hopfield Neural Network(HNN)を用いた分子構造重ね合わせ手法を提案している[1]。この手法では、はじめに重ね合わせを行う二つの分子上に Hydrophobic group(HY), Hydrogen-bonding acceptor(HA)などの化学プロパティを設定し、HNN によりそれらに対応付け、重ね合わせを行う。その時に用いるエネルギー関数には各項の重要度を調整する係数があり、この係数の値が重ね合わせ結果に影響を与える。

本研究では、X 線結晶解析により正解の分かっている構造に対して重ね合わせ実験を行うことにより、係数が結果に与える影響を調べた。その結果、多くの化合物に対して安定して最適なプロパティ対応付けを得ることができる係数の割合を推定することができた。

2. 手法

HNN は Hopfield らにより提案された人工ニューラルネットワークの一種である。相互結合型のニューラルネットワークである HNN は、巡回セールスマン問題などの NP 完全問題において効率よく近似解を与えることが証明されている。次に、HNN で用いているエネルギー関数を示す。

$$\text{energy} = \sum_i T_i \quad (1)$$

$$T_1 + T_2 = \sum_{iI} \sum_{jJ} S_{iI} S_{jJ} [A \delta_{iI} (1 - \delta_{jJ}) + B \delta_{jJ} (1 - \delta_{iI})] \quad (2)$$

$$T_3 = C \left[\min(n, m) - \sum_{iI} S_{iI} \right] \quad (3)$$

$$T_4 = D \left[\sum_{iI} \sum_{jJ} S_{iI} S_{jJ} (1 - \delta_{iI}) (1 - \delta_{jJ}) + |d_A(ij) - d_B(jI)| \right] \quad (4)$$

$$T_5 = E \left[\sum_{iI} S_{iI} P(i, I) \right] \quad (5)$$

ここで m, n は各分子上に定義されているプロパティの数、 S_{iI} は i 行 I 列のニューロンの値を示す。また σ_{ij} は i と j が等しいときには 1 を異なるときには 0 を返す関数である。P は同じプロパティ同士が対応する場合は負の値、異なるときは無限大を示す行列である。

エネルギー関数値は式(1)により、 T_1 から T_5 までの 5 つの項の線形和で計算される。式(2)は各行、各列の 1 の数を 1 つ以下にし、プロパティを 1 対 1 で対応させる。式(3)はプロパティの対応の数を最大にするための項である。式(4)は対応するプロパティ間の距離の差の最小化、式(5)は同種のプロパティ同士の対応を目的とした項である。

3. 結果

文献[1]で用いられている酵素阻害剤について重ね合わせ実験を行った。図 1 はそのうちの一組である

Carboxypeptidase-A の阻害剤、FVF と AGF の化学構造と設定されたプロパティである。右図は PDB より得られている重ね合わせであり、これと同様な結果が得られた場合を正解とした。

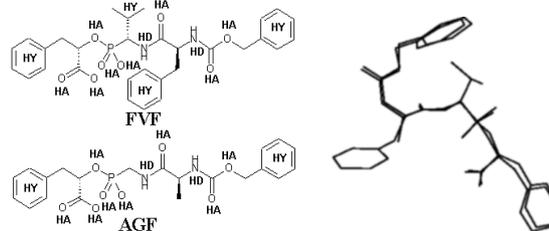


図 1.FVF と AGF の化学構造とその重ね合わせ図

係数 A と B は式(2)に示すように対称であり、等しい値を設定するのが適切である。そこで本研究では、A と B の値を 100 に固定した。そして、他の係数 C, D, E の値を一定間隔で変化させ、それぞれ 100 回ずつ HNN による重ね合わせを行った。そのときの正解率のグラフを図 2 に示す。色が濃い部分ほど多くの試行が成功したことを示している。

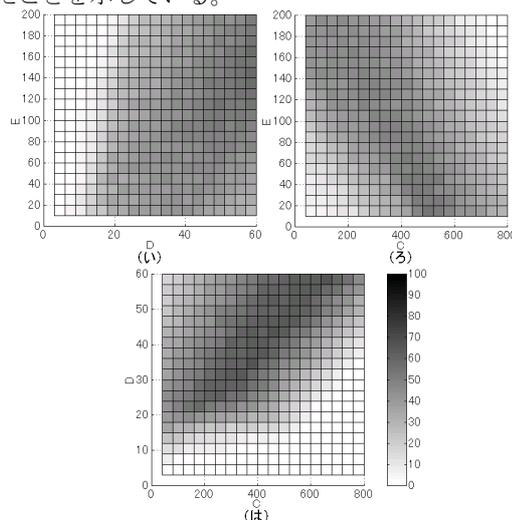


図 2.係数の比率による正解率の違い

4. 考察

図 2.(い)と図 2.(は)においては正解率の高い領域が右上がりに分布している。つまり、係数 D の値を大きくした場合、係数 E, C の値も大きくする必要がある。しかし、図 2.(ろ)では正解率の高い領域が右下がりに分布している。これは係数 C と係数 E のどちらかを高くするならば、どちらかを低くする必要があるということを示している。上記以外の二組についても同様の傾向がみられた。D については 10~20 以下の値では、ほとんど正解が得られなかった。C については比較的大きな値を持つ場合に良好な結果を与えることがわかった。E については、はっきりとした傾向がみられず、ある程度自由な値を用いることができるという結果が得られた。

5. 参考文献

- [1] 荒川正幹, 他, J. Comput. Aided Chem., 2, 29-36 (2001).