

分子類似度評価ソフトウェア

COSMOlogic社新製品 COSMOsim3D

2012年リリースを目指して開発中のCOSMOlogic社新製品COSMOsim3Dについて紹介します。COSMOsim3Dは、COSMOthermの物性推算に使用する分子の表面電荷分布をMolecular Field Analysis (MFA) のためのフィールド情報Local Sigma Profile (LSP) に変換する機能と、LSPに基づき分子の類似度を評価する機能を搭載したソフトウェアです。本稿では、COSMOsim3Dの概要とLSPの評価例、ならびに類似度の利用例について紹介します。

■COSMOsim3Dの概要

COSMOthermは、量子化学計算で得られる分子の表面電荷分布を用いて物性推算を行う熱力学物性推算ソフトウェアです。そして、既存の物性推算法と比べて、その予測精度が高いことが評価されています¹⁾。

COSMOlogic社は、分子の表面電荷分布が分子情報として優れていることに着目し、表面電荷分布をMolecular Field Analysis (MFA) に用いるフィールド情報に変換すると共に、フィールド情報に基づき分子の類似性を評価するソフトウェアCOSMOsim3Dの開発を行っています²⁾。

COSMOsim3Dでは、表面電荷分布を分子場の各グリッドに割り当てることで、フィールド情報を作成します。具体的には、表面電荷分布をその近傍のグリッドに割り当て、各グリッドの周囲の電荷分布を表すプロファイルを算出します(図1)。このプロファイルをLocal Sigma Profile (LSP) と呼び、各グリッドの極性や水素結合への寄与を表します。あわせて、LSPの特徴を抽出して数値化したLocal Sigma Moment (LSM) も考案しました。LSMは、LSPの特徴を6個の数値で表す記述子です。LSMはLSPに比べて情報量が少なく、データの取り扱いが容易です(図1参照)。

COSMOsim3Dの分子の類似度は、各グリッドのLSPの類似性に基づき算出されます。まず、各グリッドのLSPの類似性をsigma-match similarity (SMS) で評価します³⁾。

SMSは、表面積および電荷分布の類似性が高いほど、つまり、LSPの重なりが大きいほど、大きな類似度を与えます。全グリッドのSMSを算出し、その合計値より、分子の類似度SMS3Dを算出します。分子の三次元的な表面電荷分布に基づく類似度SMS3Dが、分子の識別・分類の検討などに活用されることが期待されます。

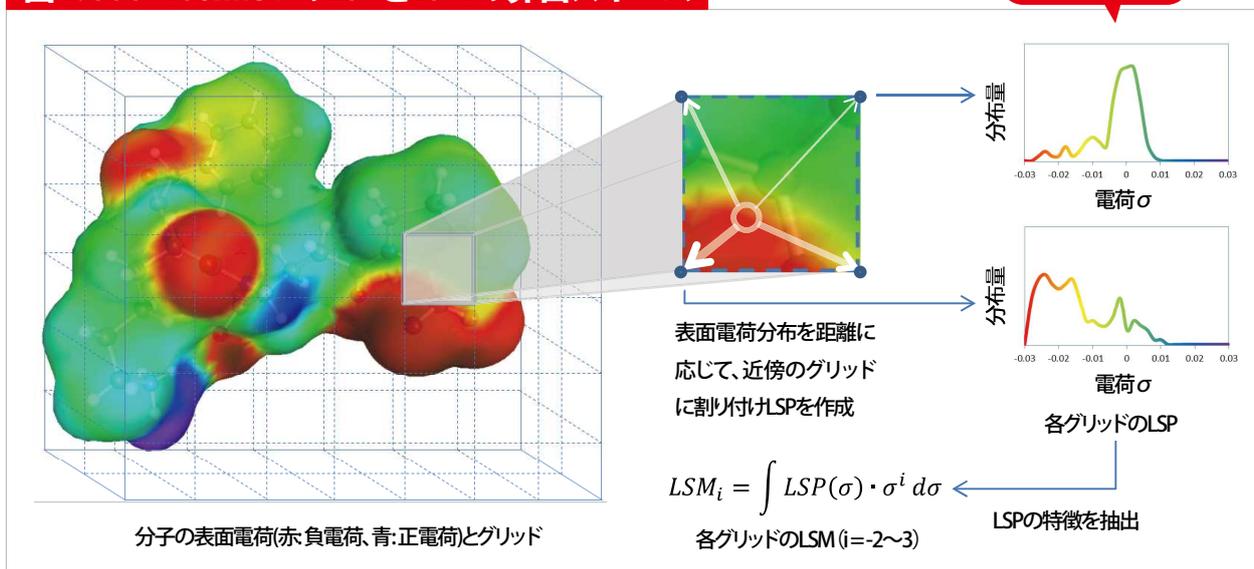
■Local Sigma Profile (LSP) の評価

COSMOsim3Dで得られるフィールド情報LSPの3次元記述子としての有用性を評価しました。評価は、Weaverらが行った8つの生理活性値データセットを用いた記述子の有用性比較と同様の方法を用いました⁴⁾。

まず、各データセットの分子の量子化学計算(汎関数: BP、基底関数TZVP)を実施し、分子の表面電荷情報を得ました。なお、アライメントされた分子構造は、Weaverらを使用した分子構造情報を用いています。次に、COSMOsim3Dを用いて、分子の表面電荷情報をLSPに変換しました。なお、グリッド間隔は1.5Åとしました。

続いて、統合計算化学システムMOEを用いて、LSPを記述子とした活性予測モデルを構築しました。モデル構築には、Partial Least Square (PLS) 回帰を用い、モデルの精度評価には、トレーニングセットの相関係数(R²)、平均二

図1: COSMOsim3DのLSPとLSMの算出スキーム



PICK UP!

乗誤差 (RMSE)、および交差検定 (leave-one-out) の相関係数 (R^2_{CV}) とテストセットの相関係数 (R^2_{TEST}) を用いました。

各モデルの統計値を表1に示します。また、参照値として、3次元記述子として代表的なCoMFA⁵を使用した場合の各値も表1に記載しました。 R^2 と R^2_{CV} から、AchEとCOX2を除く6つのセットで、CoMFAに比べてLSPを使用したモデルの精度が高いことが確認されました。また、 R^2_{TEST} から、CoMFAに比べて、新規化合物に対するモデルの予測精度が高いことが確認できました。

今回の結果によって、既存の3次元記述子に対するLSPの優位性が確認できました。

■類似度SMS3Dの利用例

COSMOsim3Dで得られる類似度SMS3Dを分子のクラスターリングに使用しました。使用したデータセットは、LSPの評価で用いたACEデータセットです。

すべての分子同士の類似度を計算し、主成分分析 (PCA) によって類似度行列の第1~3主成分を抽出し、3次元プロットしました (図2、PCAおよびグラフ表示にはMOEを使用)。

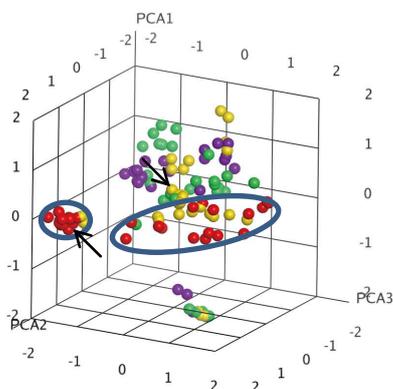


図2: 類似度の主成分によるクラスターリングと生理活性値

赤>黄>緑>紫の順で活性値が減少。
矢印は図3の描画に用いた分子を示す。

表1: LSPを記述子とした活性予測モデルの実測値との相関性および予測精度^a

Data set ^b	Components	RMSE	R^2	RMSE _{CV}	R^2_{CV}	R^2_{TEST}
ACE	3	0.82	0.88 (0.80)	1.30	0.69 (0.68)	0.58 (0.49)
AchE	5	0.50	0.83 (0.88)	0.90	0.47 (0.52)	0.47 (0.47)
BZR	3	0.34	0.74 (0.61)	0.51	0.43 (0.32)	0.24 (0.00)
COX2	5	0.69	0.54 (0.70)	0.81	0.37 (0.49)	0.24 (0.29)
DHFR	5	0.48	0.86 (0.79)	0.72	0.68 (0.65)	0.62 (0.59)
GPB	4	0.18	0.97 (0.84)	0.67	0.60 (0.42)	0.60 (0.42)
THER	4	0.55	0.91 (0.85)	1.19	0.60 (0.52)	0.42 (0.54)
THR	4	0.30	0.90 (0.86)	0.56	0.65 (0.59)	0.57 (0.63)

a. 括弧内はCoMFAを記述子とした場合。b. 生理活性データおよび構造は、参考文献4より得た。

図2より、活性値の高い分子は特定の座標位置に分布する傾向があり、分子の表面電荷分布に共通の特徴があると推察されます。このような結果を利用して、活性値の高い分子の特徴が見い出されることが期待されます。

また、COSMOsim3Dの描画機能を用いて、図2の矢印で示す分子の表面電荷を可視化し、LSPの類似性を比較しました (図3)。図3より、電荷分布の差異を容易に判別できると共に表面電荷と部分構造の関連も確認することができます。

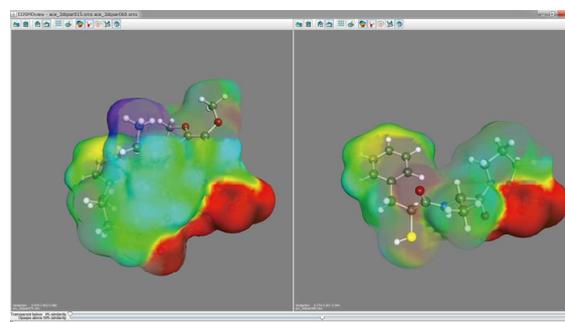


図3: 分子の表面電荷分布と類似性の表示例

ACEデータセットのうち最も活性の高い分子との比較。
左が活性値が高い分子、右が低い分子。
LSPの類似性が50%以下の分子表面を半透明で表す。
半透明部分が大きいほど分子間の類似性が低い。

■リリース時期と関連製品

COSMOlogic社では、COSMOsim3Dの2012年前半のリリースを目指して開発中です。関連ソフトウェアとして、構造活性相関解析ソフトウェアCOSMOsar3Dも同時にリリース予定です。

*1 <http://fluidproperties.org/results-sixth-challenge>

*2 A. Klamt et al., Publication in preparation.

*3 A. Klamt et al., *J. Chem. Inf. Model.*, 46, 1040-1053 (2006).

*4 D. F. Weaver et al., *J. Med. Chem.*, 47, 5541-5554 (2004).

*5 R. D. Cramer et al., *J. Am. Chem. Soc.*, 110, 5959-5967 (1988).