2006－BIO－6（7）
2006／9／15

モルフォロジーを用いた蛋白質表面の様々な深さのポケット形状の計算

川端 猛 *

概 要

蛋白質分子の構造から凹形状部分（ポケット）を抽出することは、 低分子の結合部位の推定に重要な示唆を与える。ポケットを「小さな球は入れるが大きな球は入れない空間」だと考え、モルフォロジー (mathematical morphology) による画像の集合演算としてポケットを抽出する方法が提案されている。しかし、最適なポケットの深さ（大きな球の半径）は、結合する分子の種類や結合の性質によって様々である。そこで、本研究では、様々な大きさの分子表面およびポケットを同時に計算する方法を提案する。この手法により、ポケットの位置、深さ、大きさは同時に計算され、蛋白質の表面構造の特徴をより広い視点から捉えることができる。また、ポケットの深さ（ある点に到達できない球の半径の最小値）を、近似的に求めることも可能であり、結合分子の環境評価を新たな視点で行うことも可能となる。

Multi-scale Pocket Detection on Protein Surface Using 3D Image Processing Technique

Takeshi Kawabata

Abstract

Detecting pocket regions on protein surfaces is important for prediction of putative binding sites of small ligands. Masuya and Doi (1995) defines a pocket regions using the theory of mathematical morphology. We extend their concept and introduce multi-scale molecular surface and pocket using several sizes of probe spheres, and invent an efficient algorithm to calculate them at the same time. Multi-scale pocket provides more information to characterize protein surface. Multi-scale molecular surface enable us to calculate new depth measure for pockets and binding ligands.

1　はじめに

蛋白質の立体構造データから、それに結合する分子の種類、結合部位、結合の強さを推定することは、機能未知蛋白質の機能推定や、ドラッグの設計において大変重要である。分子の結合のエネルギーは分子間の接触表面積と関係がある。よって、結合に関する予測には、蛋白質表面が他分子に対してどれだけ露出しているかの定量化が必要である。他分子のモデルとして球形のプローブを蛋白質表面に転がす操作を考える方法が 1970 年代に提案され、現在でも広く使われている。その一つが Richards らによって提案された accessible surface (露出面積) であり [1]、プローブ球を蛋白質 VdW 表面に転がしたときの球の中心の軌跡面として定義される（図１）。水分子を想定した 1.4$\AA$ のプローブ球を用いた露出面積は、solvent accessible surface(溶媒露出面積) と呼ばれ、蛋白質に水和する水分子の数との相関が期待されることから、水和自由エネルギーの見積もりによく使われている。一方、同様にプローブ球を転がしたときに、プローブ球が接触できない空間との境界のことを分子表面 (molecular surface)、接触表面

*奈良先端科学技術大学院大学 情報科学研究科 GraduateSchool of Information Science, Nara Institute of Science and Techonology

(contact surface) あるいは提案者の名前をとってコノリー表面 (Connolly surface) と呼ぶ [2,3]。これらは、結合分子にとっての蛋白質の表面であると考えられるため、二つの分子の相補的な結合を求めるドッキング計算によく使われる（図１）。また、低分子の結合サイトは一般にポケット形状をしていることが多いため、ポケット形状を認識してそれを結合部位と候補とする方法が多く提案されている。Kawabata and Go はポケットを「接触する小さな球の集合のうち、大きな球は触れることができない部分」と定義し、３原子に接触するプローブ球の集合を発生させることで、ポケットを近似的に計算する手法を提案した（図１）[4]。後述するように、このポケットの定義はコノリー表面と深い関係がある。

これらのプローブ球を使った形状は、すべて球に関する幾何学的な集合演算であるため、その体積・面積を解析的に得ることができるはずだ。しかしながら、そのアルゴリズムは一般に複雑であり、特にコノリー表面を計算するアルゴリズムは大変混み入った手続きが必要となる [2,3]。この煩雑さから逃れるために、球集合による空間表現をあきらめ、空間を格子に分割して、蛋白質形状を３次元の離散的な二値画像として扱う戦略がある。離散化による誤差を生じるものの、そのアルゴリズムは一般に平易であり、より複雑な演算を行うことも可能となるからだ。画像処理の分野の中で「モルフォロジー」(mathematical morphology) という手法 [5,6] は、構造化要素 (structuring element) というプローブ形状を用いて、対象形状とプローブ形状の集合演算により、様々な特徴抽出を行う。本来、この手法は、鉱石の画像解析のために開発されたものだが、集合論を用いて厳密にかつ一般的に理論が構築されているため、様々な分野の画像処理に対して、数学的に整理された見通しのよい表記法を提供する。。Masuya and Doi (1995) は、前述の露出面積、分子表面などの形状は、モルフォロジーの記法により簡潔に記述できることを示し、また、モルフォロジーによりポケット領域を定義する方法を提案した [7]。本研究では、これらの研究の発展として、複数の大きさのプローブ球を用いた場合の分子表面やポケットを同時に計算する方法を提案する。この方法により、深さの異なる分子表面やポケットを等値面に描画することが可能であり、ポケット領域や結合分子が、どのくらいの深さのポケットに位置するかを定量化することができる。

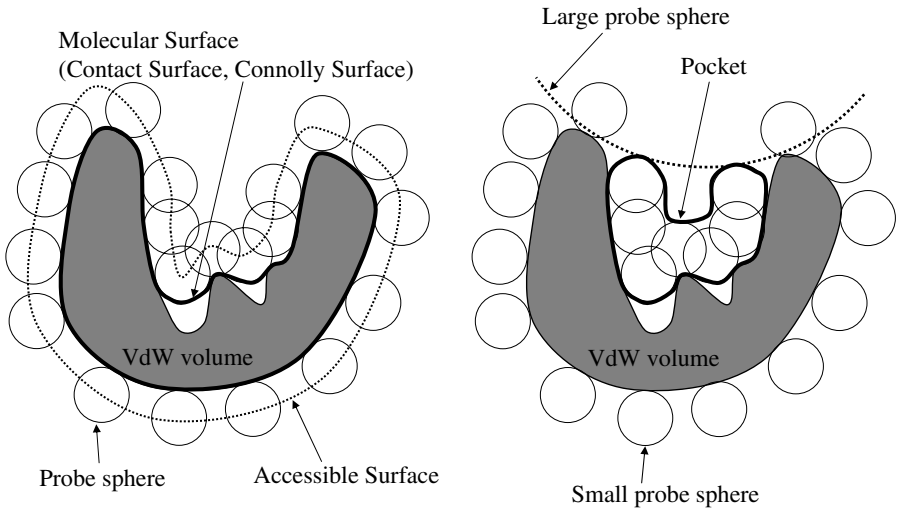

図 1: 高分子の立体構造解析で使用されるプローブ球を用いた表面と体積（左）および Kawabata and Go によるポケットの定義。

2　モルフォロジーの基本操作

最初に、モルフォロジー (mathematical morphology) の基本となる概念を手短かに説明する [5,6]。ここで、X と P をそれぞれ三次元位置ベクトル $x = (x_0, x_1, x_2)$, $p = (p_0, p_1, p_2)$ を要素として持つ集合と

する。X は解析対象となる形状データの二値画像であり、P はプローブ (あるいは構造化要素) 形状の二値画像である。ここで、P は原点に対して対称であるとする。二つの集合操作 dilation ($X \oplus P$) と erosion ($X \ominus P$) を以下のように定義する。

$$X \oplus P = \bigcup_{p \in P} X_p \tag{1}$$

$$X \ominus P = \bigcap_{p \in P} X_p \tag{2}$$

ここで、X_p は、X の全ての要素をベクトル p によって平行移動した集合である。図 2 に dilation と erosion の例を示した。この定義は次のように解釈されてプログラム内で実装される。dilation は、P の中心が X の要素 x になるように並進させたときの、P 全体が通る軌跡の集合である。eroson は、同様に、P の中心が X の要素 x になるように並進させたとき、P の全てが X に含まれる場合だけを選んだ P の中心だけの集合である。図 2 からわかるように、dilation は対称の図形をプローブの半径だけ拡張（膨張）させる操作であり、erosion は逆に図形をプローブの半径だけ収縮（浸食）させる操作である。

dilation と erosion を組み合わせて、closing($X \bullet P$) および openinig($X \circ P$) が定義される。

$$X \bullet P = (X \oplus P) \ominus P \tag{3}$$

$$X \circ P = (X \ominus P) \oplus P \tag{4}$$

closing は、P の中心が X の要素 x になるように並進させたときに、P の全てが X に含まれる場合の P 全体の通る軌跡の集合であるともいえる。opening は、同様の closing の操作を X の補集合について行った結果の補集合である ($X \bullet P = [X^c \circ P]^c$)。図 2 に closing と opening の例を示した。closing は、凹部を埋め、opening は凸部を削る効果がある。

3　モルフォロジーによる分子表面およびポケットの定義

次に、これらのモルフォロジーの集合操作を用いて、生体高分子の分子体積やポケットを定義する方法を説明する [7]。まず、対象となる生体高分子の形状は、Van der Waals 体積 X であるとする。

$$X = \bigcup_i A_i \tag{5}$$

ここで、A_i は生体高分子を構成する i 番目の原子の VdW 球の形状である。次に、プローブ球 P を定義する。プローブ球の半径は、水分子に相当する $1.4\mathring{A}$ を用いることが多い。まず、分子の Solvent accessible 体積 X_S は以下のように X の P による dilation で表される。

$$X_S = X \oplus P \tag{6}$$

solvent accessible 体積 X_S の表面だけを抽出した部分が、水和自由エネルギーの見積もりに広く使われている solvent accessible surface(溶媒露出表面積) である。次に、プローブ球 P による分子体積 X_M は、X の P による closing で定義される。

$$X_M = X \bullet P = (X \oplus P) \ominus P \tag{7}$$

分子体積 X_M の表面だけを抽出した部分が、分子表面 (コノリー表面 (Connolly surface)) である。

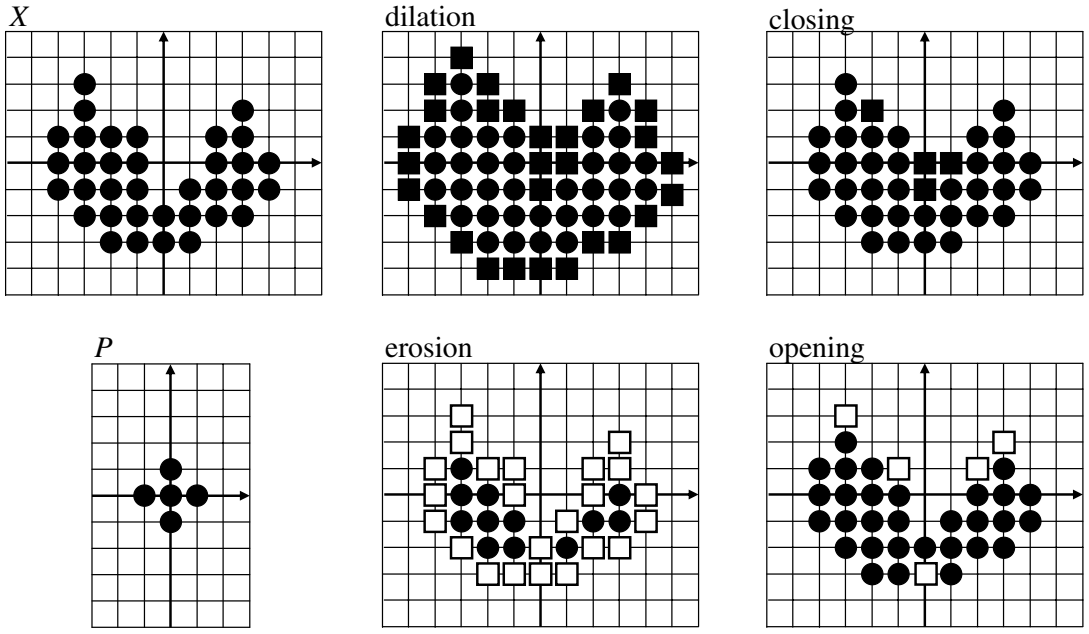

図 2: モルフォロジーの基本操作の二次元の二値画像に対する計算例。左端に対象形状 X とプローブ形状 P を示した。中の二つは dilation $X \oplus P$ と erosion $X \ominus P$ である。右端は closing $X \bullet P$ と opening $X \circ P$ である。対象の形状 X の要素を●、操作によって、X に追加された要素を■、X から削除された要素を□で示している。

次にポケット領域の定義を導入する。Masuya and Doi(1995) は、ポケット領域を「小さな球は入れるが大きな球は入れない空間」とし、以下のような操作で表現した (図 3)[7]。最初に、二つの異なるサイズのプローブ球、P_L、P_S を導入する。次に、大プローブ球 P_L を用いて分子体積 X_{ML} を求め、これと X との間の空間 X_C を求める。この X_C が、大きな球が入れない空間である。X_C の中で、小プローブ球 P_S が入れる空間がポケット X_P であり、これは、X_C の P_S による opening の処理によって得られる (図 3 を参照)。この一連の処理は以下の三段階の集合操作にまとめられる。

$$X_{ML} = X \bullet P_L \tag{8}$$

$$X_C = X_{ML} \bigcap [X]^c \tag{9}$$

$$X_P = X_C \circ P_S \tag{10}$$

4 複数の大きさのプローブを用いた場合の分子体積とポケット

前節まで、モルフォロジーを用いた分子体積およびポケットの定義を説明した。ポケットを抽出するには、二つのプローブ球の半径を適用前に決定しておく必要がある。小プローブ球には、原子や官能基を想定して、1.4 から 2.0$\AA$ の半径を使うのが一般的である。一方、大プローブ球の大きさは、結合分子の埋もれ度に相当するため、結合する分子の特徴や結合の強さによって、最適な値が異なる（Kawabata and Go,submitted）[4]。よって、結合する分子が未知の場合はいくつかの大きさの大プローブ球を試すことが推奨される。本研究では、複数の大きさのプローブ球を用いた場合の分子体積とポケットの計算を

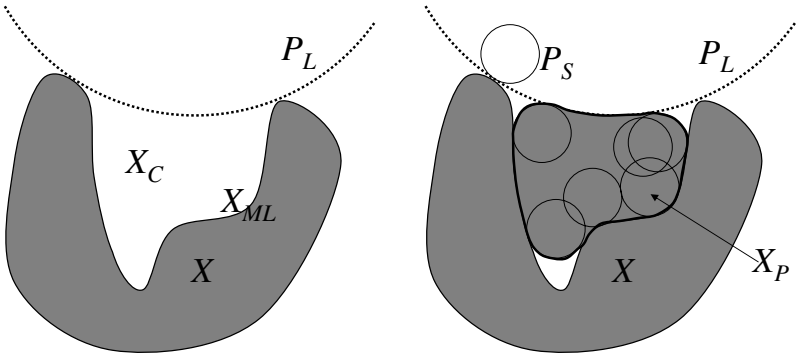

図 3: Masuya and Doi によるモルフォロジー操作によるポケットの定義

できるだけ効率的に計算する方法を導出したい。そのために、2 種の大きさのプローブ球を用いたとき、モルフォロジーの操作の結果の包含関係を考える。まず、半径 R の大きさのプローブ球の形状を $P(R)$ とする。ここで、二通りの半径 r と R を考え、$r < R$ であるとすれば、当然、$P(r) \subset P(R)$ が成り立つ。よって、任意の形状 X における、dilation, erosion, closing, opening について、$X \oplus P(r) \subset X \oplus P(R)$, $X \ominus P(r) \supset X \ominus P(R)$, $X \bullet P(r) \subset X \bullet P(R)$, $X \circ P(r) \supset X \circ P(R)$ の関係が成り立つことが知られている。次に、分子体積とポケットについてもプローブ球の大きさによる包含関係を考える。半径 R のプローブ球を用いた場合の分子体積を $X_M(R)$、同様に半径 R の球を大プローブ球として用いた場合のポケットを $X_P(R)$ と定義する。すると、$R > r$ であれば、分子体積とポケットについて以下の関係があることがわかる。

$$X_M(r) \subset X_M(R) \tag{11}$$

$$X_P(r) \subset X_P(R) \tag{12}$$

さらに N 通りの大きさの大プローブ球を考え、その半径を $R_1, R_2, ..., R_N$ とし、$R_1 < R_2 < ... < R_N$ の関係が成り立っているとする。すると、これら N 通りのプローブ球を用いて得られた分子体積とポケットの間には以下の関係が成り立つ。

$$X_M(R_1) \subset X_M(R_2) \subset ... \subset X_M(R_N) \tag{13}$$

$$X_P(R_1) \subset X_P(R_2) \subset ... \subset X_P(R_N) \tag{14}$$

つまり、分子体積 $X_M(R_i)$、ポケット $X_P(R_i)$ は、大きな半径による集合が小さな半径による集合を含むような関係になっている（図４）。これらの結果から、複数の大きさのプローブ球を用いた計算アルゴズムに関して以下の二つの知見が得られる。

- 各格子点ごとに N 通りの非負の整数を格納するような 3 次元データを一つ用意し、各格子点ごとに、それを要素とする最小の $X_M(R_i)$ を格納することにすれば、一つの 3 次元データで、同時に N 通りの分子体積（あるいはポケット）を表現することができる。

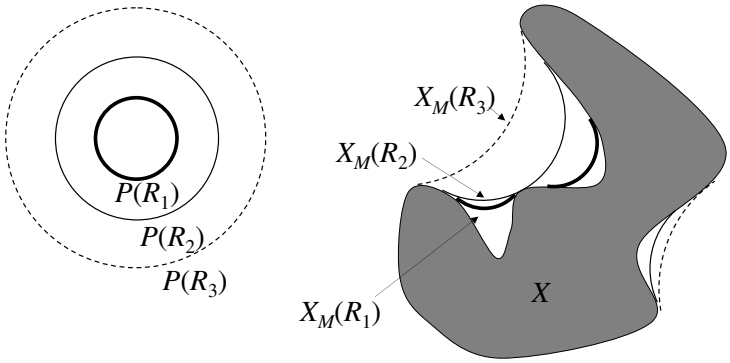

図 4: 複数の大きさのプローブ球 $P(R_1)$,$P(R_2)$,$P(R_3)$ を用いたときの分子体積 $X_M(R_1)$, $X_M(R_2)$, $X_M(R_3)$

- ある点 x が、ある半径 R_i の分子体積に属するなら、それ以上の半径 $R_j(i < j \leq N)$ の分子体積にも必ず属する $(x \in X_M(R_i) \rightarrow x \in X_M(R_j))$。ポケットについても同様である。この性質をうまく利用すれば、計算量を減らせる可能性がある。

これらから、N 通りの分子体積、およびポケットを少ない計算コストで計算することが可能である。アルゴリズムの詳細は、別誌で発表予定であるが、最大の半径 R_N の分子体積の計算と同様の計算時間で、それ以下の半径の分子体積も同時に計算することが可能である。図 5 に 2 種類の大プローブ球を用いたときのの分子体積とポケットを示す。

5 適用例と応用可能性

本研究で提案された複数の大プローブ球を用いた分子表面、ポケットの計算には、大きく二つの用途がある。一つは、認識された様々な深さのポケット領域を、その深さごとに色を分けて、表示することで、ポケットの位置、大きさ、深さを同時に視覚的に把握することだ。本研究の手法を用いて、生体高分子の構造の異なる深さのポケットを計算した例を図 6 、 7 に示す。図 6 は典型的な DNA の二重らせん構造のポケットである。主溝は白く（浅く）、副溝は黒い（深い）ことが明瞭に理解できる。図 7 はプロテアーゼに対する計算結果である。分子中央に非常に深く大きなポケットがあり、それが浅い入り口に連なっていること、他に深いが小さなポケットが複数あることが観察される。また、結合分子の種類によって、結合するポケットの深さが異なることは知られており [4]、結合する分子が未知である蛋白質に対しては、こうした多様な深さのポケットを一度に計算できる手法の有効性は高いと思われる。

複数の大プローブ球を用いた分子表面のもう一つの用途は、蛋白質周辺の空間に対する深さ・浅さの定量化である。Kawabata and Go は、ポケットの浅さの指標、「最小非接触半径 (minimum radius of inaccessible sphere)」を提案している。これは、ある点に到達できない球の最小半径であり、本研究の方法により、同様の量がより高速に計算できる。最小非接触半径は、結合分子の埋もれ度合いの新しい定量化の指標となり、既知の結合分子の環境評価や、ドッキング計算によって得られた結合分子の信頼性の評価などに適用できるのではないかと考えている。またモルフォロジーによる画像処理の分野では、多数の大きさのプローブの opening の結果をパターンスペクトルとしてまとめ、形状の特徴のプロファイルとして用いる研究が進んでいる [8]。生体高分子においても、ある種のパターンスペクトル（各深さのポケットの体積の分布など）が、その蛋白質の機能的特徴を表している可能性があり、今後、調査を

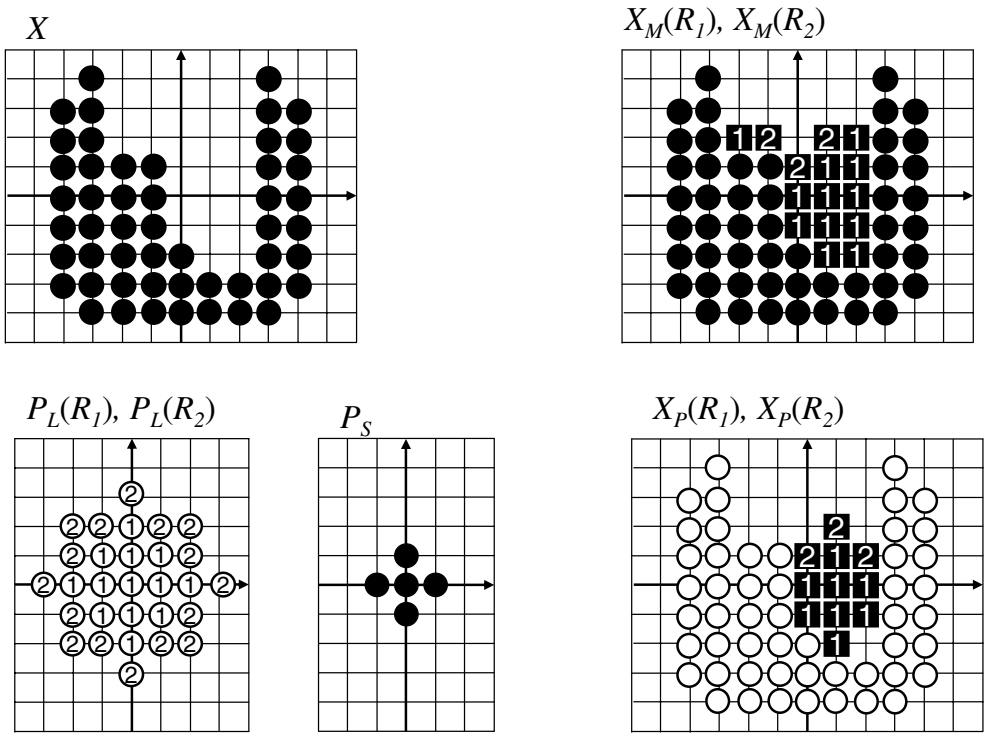

図 5: 複数の大プローブ球を用いたときの分子体積とポケットの二次元の二値画像に対する計算例。解析対象の形状 X、二つの大きさの大プローブ球 $P_L(R_1)$,$P_L(R_2)$ と小プローブ球 P_S を左に示した。これらのプローブ球を用いた場合の 2 種の分子体積 $X_M(R_1)$, $X_M(R_2)$, と 2 種のポケット $X_P(R_1)$, $X_P(R_2)$ を右に示した。

進めていきたいと考えている。

6 参考文献

[1] Richards, FM. Areas, Volume, Packing and Protein Structure. Ann.Rev.Biophys.Bioeng., 1977, 6,151-176.

[2] Connolly ML. Analytical molecular surface calculation. J Apl Cyst 1983;16:548-558.

[3] Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983: 221:709-713.

[4] Kawabata T and Go N. “A novel definition of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites”. Submitted.

[5] Serra J. “Image analysis and mathematical morphology”. 1982. Academic press.

[6] 小畑秀文 「モルフォロジー」1996 年、コロナ社

[7] Masuya M, Doi J. Detection and geometric modeling of molecular surfaces and cavities using digital mathematical morphological operations. J. Mol. Graphics 1995; 13(6):331-336.

[8] Maragos P. “Pattern spectrum and multiscale shape representation”. IEEE transactions on pattern analysis and machine intelligence, 1989, Vol. 11, 701

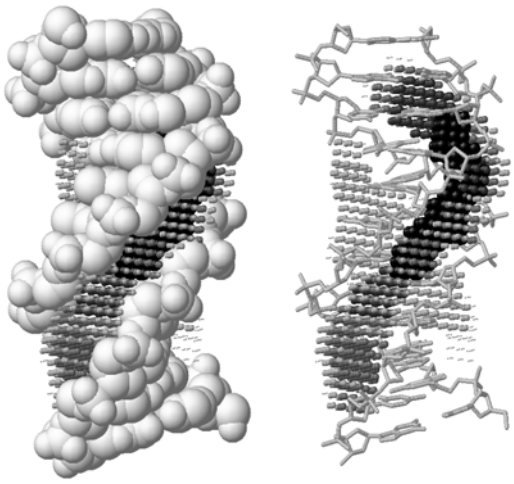

図 6: DNA の二重らせん構造 (PDB ID コード:1bna) に対して本研究手法で認識されたポケット領域。左図は蛋白質を空間充填モデルで表示、右図はワイアーフレームモデルで表示している。グリッドサイズは 1 $\mathring{A}$、小プローブ球は半径 1.87 $\mathring{A}$ を使用。大プローブ球は 3 から 10$\mathring{A}$ まで 1$\mathring{A}$ 刻みで 8 通りを試している。黒から灰色の小さな球の塊が認識されたポケット領域であり、色が黒いほど深いポケット、白いほど浅いポケットを示す。

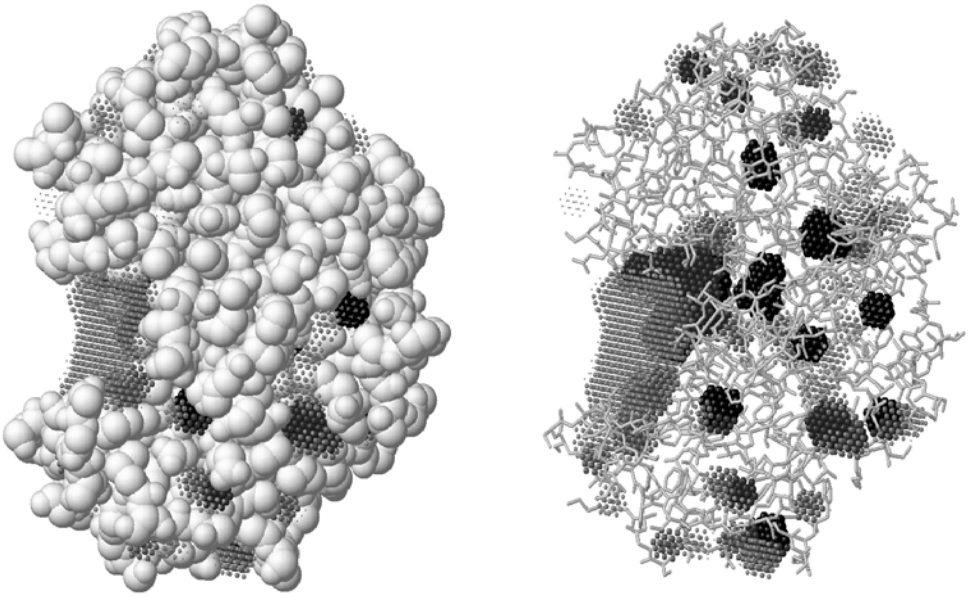

図 7: プロテアーゼ endothiappsin (PDB ID コード:3er5 chain E) 対して本研究手法で認識されたポケット領域。左図は蛋白質を空間充填モデルで表示、右図はワイアーフレームモデルで表示している。計算条件は図 6 と同じ。