

SVM を用いたタンパク質 2 次構造予測

麻 生 川 稔[†]

タンパク質は 200 個程度の 20 種のアミノ酸の一次配列であり、配列が折りたたまることによって、特異な立体構造を形成する。立体構造は、3 種類の 2 次構造 (螺旋形状の α ヘリックス、シート形状の β シート、どちらにも分類できないコイル) の集合とみなすことができる。配列からの 2 次構造予測に対して 70% 以上の予測率を示すシステムとして、PHD¹⁾ と PsiPred²⁾ が広く用いられている。これらは、配列の連続 7 ~ 21 残基の情報を 2 段重ねた階層型ニューラルネットワークで処理することにより、2 次構造を予測している。本研究では、階層型ニューラルネットワークの代わりに、Support Vector Machine (SVM) を用い、種々の条件で 2 次構造の予測実験を行なった。

Protein secondary structure prediction using SVM

MINORU ASOGAWA[†]

Proteins are consists of more than hundreds of amino residues. Each protein forms particular tertiary structures. The tertiary structure can be classified into three substructures, named secondary structures. Several systems predict these secondary structures by using neural networks based on the information of consecutive residues ranging from seven to twenty one. In this paper, we utilized support vector machines (SVM) instead of neural networks. Several experimental results are shown.

1. Support Vector Machine

Support Vector Machine は、学習データのエラーと VC 次元で表現されたモデルの複雑さの和を、PAC 学習の枠組で最小にするものである³⁾。この和は、未学習データに対するエラーの上限の推定値を与える。従って、**上記の和を最小にするモデルを構築する**ということは、**未学習データに対するエラーの上限値を最小にするモデルを構築する**ということになる。これは、PAC 学習の枠組なので、このことは $1 - \eta$ の確率で成立する (η は小さな正の数)。SVM では、モデルのパラメータ推定を 2 次計画問題として定式化し、カーネル関数を用いることにより、階層型ニューラルネットワークや Radial Basis Function 等のモデルが表現できる。

SVM でのパラメータ推定方法とバックワード・プロパゲーション法 (BP 法) と比較してみる。BP 法では、重み (モデルのパラメータ) の推定には、学習データを逐次的に与え重みを更新する。学習の途中で、適当な学習回数で停めないと過学習の問題が起こることは、良く知られている問題である (これを避けるために、逐次的な

重みの更新に於いて、制限をつける等の方法がある⁷⁾)。また、中間層のセル数を多くすると過学習の問題が顕著となるため、試行錯誤により適当な中間層の数を決定する必要がある。

これに対して、SVM のパラメータ推定では、学習データを用いて 2 次計画問題を解くことによって得られるので、過学習の問題が無い。さらに、パラメータ数も、未学習データに対するエラー最小となるという観点から、適切に決めることができる。これらの長所より、近年多くの場面で SVM が適用され、従来方法と同等かより高い認識率が報告されている⁹⁾。

2. タンパク質 2 次構造予測システム

本研究では、連続 7 ~ 21 残基の情報を SVM で処理することにより、2 次構造を予測するシステムを作成した。データは、PDB (version 89) のうち 2.5 Å よりも精度の高い X 線で決定されたものと NMR で決定されたタンパク質を用いた。これらのタンパク質の配列を、HSSP データベース⁶⁾ の 2 次構造分類 (α ヘリックス、ストランドまたはそれ以外) に従って分類した。予測実験では、連続 7 残基と 21 残基のウィンドウ幅を用い、その中心残基の 2 次構造を予測するものとした。HSSP データベースは立体構造の類似性に基づいてアライメン

[†] 日本電気 (株) 基礎研究所
Basic Research Laboratories, NEC Corporation

トした結果をまとめたデータベースで、実験ではウインドウに含まれる各位置での残基の分布を入力データとした。これは、1つの配列の残基だけではなく、“2次構造を構成する時に望ましい残基の分布”に基づいた学習を目指したためである。構造未知のタンパク質についての推定では、そもそも立体構造が解っていないので、このような入力データが得られないが、SVMでは学習データがモデルに反映されるのでこのような入力データを用いた。

2.1 カーネル関数としてガウス関数を用い、連続7残基を用いた実験

連続7残基を用いた実験では、カーネル関数としてガウス関数($\sigma^2 = 1.0$)を用いた。学習データとしては(ヘリックス vs スtrand)、(ヘリックス vs コイル)、(スtrand vs コイル)の3種データ(各3万)を用いた。

学習データに対する結果		Q_2 99.72%	
	Helix	Coil	← svm の予測
Helix	13,411(44.70%)	45(0.15%)	13,456(44.85%)
Coil	38(0.13%)	16,506(55.02%)	16,544(55.15%)
↑正解	13,449(44.83%)	16,551(55.17%)	30,000(100.00%)

未学習データに対する結果		Q_2 79.87%	
	Helix	Coil	← svm の予測
Helix	8,760(30.52%)	3,759(13.10%)	12,519(43.62%)
Coil	2,018(7.03%)	14,164(49.35%)	16,182(56.38%)
↑正解	10,778(37.55%)	17,923(62.45%)	28,701(100.00%)

図1 ヘリックス vs コイルに対する結果

		Q_2 81.61%	
	Helix	Strand	← svm の予測
Helix	11,571(56.12%)	568(2.76%)	12,139(58.88%)
Strand	3,224(15.64%)	5,254(25.48%)	8,478(41.12%)
↑正解	14,795(71.76%)	5,822(28.24%)	20,617(100.00%)

図2 ヘリックス vs スtrandの未学習データに対する結果

		Q_2 78.66%	
	Strand	Coil	← svm の予測
Strand	4,224(17.19%)	4,468(18.18%)	8,692(35.37%)
Coil	775(3.15%)	15,107(61.48%)	15,882(64.63%)
↑正解	4,999(20.34%)	19,575(79.66%)	24,574(100.00%)

図3 スtrand vs コイルの未学習データに対する結果

これらの結果から分かるように、学習データに関しては、高い認識率を達成している。モデルでは約2,300個のSV(Support Vector)が用いられており、大半の入力サンプルを覚えていないことに起因すると考えられる。この状態は、過学習の様に思えるが、未学習データに対する結果では、78～81%と高い認識率を持っているので一概に過学習とはいえない。

実験で得られた3つのSVMを組み合わせ、3状

態の認識を行なった。認識では、(ヘリックス vs スtrand)のSVMのスコア、(ヘリックス vs コイル)のSVMのスコア、(スtrand vs コイル)のSVMのスコアの各々の各カテゴリに関する値を総計した。例えば、スtrandに関するスコアは、(ヘリックス vs スtrand)のSVMのスコアに-1を乗じたものと(スtrand vs コイル)のSVMのスコアとの和を用いた。この結果、 Q_3 (=(各3状態を正しく認識した残基数)/(全残基数))が70.26%であった(図4)。PHD(ニューラルネットワークを用いた2次構造予測プログラム)では、 Q_3 が約72%であることを考慮すると、SVMの結果は比較的良好な値であると言える。しかしながら、この結果はタンパク質の配列全体の予測ではなく、切り出した一部の解析なので、単純には比較できない(連続残基の結果をまとめて、配列の解析すると、予測率は向上する等の効果があり得る)。また、使用しているデータセットが全く違うので(本実験の方がより緩い条件で作成した)、安直な比較はできない。

Q ₃ 70.26%				
	Helix	Strand	Coil	← svm の予測
Helix	8281	556	2737	11574
Strand	1231	4678	2534	8443
Coil	2300	1187	11952	15439
↑正解	11812	6421	17223	35456

	Helix	Strand	Coil	← svm の予測
Helix	23.36%	1.57%	7.72%	32.64%
Strand	3.47%	13.19%	7.15%	23.81%
Coil	6.49%	3.35%	33.71%	43.54%
↑正解	33.31%	18.11%	48.58%	100.00%

図4 3状態の未学習データに対する結果

2.2 大きな誤りを学習データから除き再学習

SVMはノイズに起因するエラーに敏感といわれているので、 λ の値の大きなデータを学習データから除き、再度SVMで学習した。そのSVMに対する、未学習データの結果を図5に示す。 Q_3 は、わずかながら改善されている。

2.3 ガウス関数の σ^2 を2.0とした実験

前記の実験では、入力データ数(3万)に比べて、SV数が約2,300個と多かったため、ガウス関数の σ^2 を2.0として、より多くのデータをカバーし、SV数を減少させようとした。その結果を図6に示す。 Q_3 はわずかながら改善された。残念ながら、SVMの枠組では、 σ^2 の学習はできないので、いくつかの実験を繰り返す必要がある。

実験結果より、SVMがスtrandと予測したものは4,721で、正解が3,877(82.12%)と高い認識精度である。しかしながら、スtrandがコイルと予測された

ものが、8,433 中 4,721(55.98%) あり、ストランドのコイルの区別が難しいことが分かる。

2.4 連続 21 残基を用いた実験

入力ウィンドウ幅として連続 21 残基を用いた時の SVM の結果を図 7 に示す。今までの結果と比較して、コイル状態の予測率が向上している (37% → 42%) が、その他の予測率が向上していないため、全体として Q_3 は低い。

2.5 カーネル関数として 2 次多項式を用いた実験

カーネル関数 $K(\mathbf{x}, \mathbf{y})$ として、2 次多項式 $(\mathbf{x} \cdot \mathbf{y} + 1)^2$ を用いた実験を行なった。このカーネル関数を用いることにより、ウィンドウ中の 2 残基がどのような組合せで出現するかに対する情報に基づき処理することができる。これは、タンパク質の立体構造上、 α ヘリックスでは、注目残基と 3 もしくは 4 残基離れた残基、 β ストランドでは、注目残基と 2 あるいは 4 残基離れた残基と物理的に相互作用をしていると考えられるため、ウィンドウ中の 2 残基の組合せ情報を用いることによって、高い認識精度が得られると期待したためである。しかしながら、 Q_3 は低かった (60.33%)(図 8 参照)。

3. タンパク質単位の 2 次構造予測実験

上記の実験では、基本的な性能を調べるために、テスト用データは学習データとは異なるものを用いたが、相同性のことは考慮していなかった (高い相同性のものも含まれていた)。ここで行なったタンパク質単位の 2 次構造予測実験では、学習データとの相同性が、高くても 30% 未満のタンパク質配列 (相互には 25% 以下の相同性) を選び、実験を行なった。25% 以下の相同性は、PDB_select 25 を参考にした⁴⁾。この時の予測率の分布を図 9 に示す。各タンパク質配列の予測率の平均は 63.6% で、標準偏差が 13.0% であった。

テストしたタンパク質中で、HSSP データベースで立体構造が類似の配列が 10 以上ある残基の予測率は、これよりも精度が高い (平均の予測率は 66.2% で標準偏差が 12.9% であった)。また、類似の配列が 40 以上ある残基に対しては、平均の予測率は 70.1% で標準偏差が 13.0% であった (図 10 参照)。

以上から分かるように、HSSP データベースで類似の配列が多いほど予測率が高くなる。図 11 に、残基単位での予測率と HSSP で類似の配列の本数との関係を示した。図より分かるように、HSSP の類似配列が多くなれば、予測率が向上する。しかしながら、全体の中では、類似配列の少ない部分も多く、40 本以上の相同性配列が揃っている残基は、全体の 22% と少ない。

しかしながら、構造未知のタンパク質についての 2 次

構造予測では、そもそも立体構造が解っていないので、このような入力データが得られない。そのために、構造未知配列に対しては、配列の類似性に基づいて整列したデータに基づき入力する等の方法が必要がある。この辺りについては、今後の研究課題である。

ま と め

本研究では、タンパク質のアミノ酸配列の連続 7 ~ 21 残基の情報に基づき、SVM を用いて 2 次構造を予測した。種々の条件で実験したところ、階層型ニューラルネットワークにはやや劣るがほぼ同程度の認識精度が得られた。また、SVM のモデルのパラメータに依存性があり、高い 2 次構造認識精度を得るためには、種々のパラメータに対する実験が必要であることが解った。また、立体構造の類似性に基づいてアライメントした結果である HSSP データベースを用いて実験を行なったところ、類似の立体構造が多いデータに基づいた予測では、高い認識精度が得られた。

本研究の一部は、農林水産省プロジェクト **食品製造におけるニューロ制御技術の開発**の一環として行なわれ。

参 考 文 献

- 1) <http://www.embl-heidelberg.de/predictprotein/predictprotein.html>
- 2) <http://bioinf.cs.ucl.ac.uk/psipred/>
- 3) Cortes C. and Vapnik V., "Support vector networks", *Machine Learning*, vol. 20, pp. 1-25, (1995).
- 4) ftp://ftp.embl-heidelberg.de/pub/databases/pdb_select/recent.pdb_select
- 5) Rost B. and Sander C., "Prediction of Protein Secondary Structure at Better than 70% Accuracy", *J. Mol. Biol.*, vol. 232, pp. 584-599, (1993).
- 6) Sander C. and Schneider R., "Database of homology-derived protein structures", *Proteins*, vol. 9, pp. 56-68, (1991).
- 7) Ishikawa, M., "A structural learning algorithm with forgetting of link weights", *Proceedings of IJCNN '89 at Washington DC.*, Vol. 2 (1989).
- 8) B. Hassibi and D.G. Stork, "Second order derivatives for network pruning: Optimal brain surgeon", *Advances in Neural Information Processing Systems*, vol. 5, pp. 164-172, (1993).
- 9) Schölkopf, B., Burges, C., and Smola, A. (Eds), *Advances in Kernel Methods*, Mit Press, (1999).

Q_3 70.83%

	Helix	Strand	Coil	← svm の予測
Helix	8089	226	3266	11581
Strand	1131	3877	3435	8443
Coil	1667	618	13154	15439
↑正解	10887	4721	19855	35463

	Helix	Strand	Coil	← svm の予測
Helix	22.81%	0.64%	9.21%	32.66%
Strand	3.19%	10.93%	9.69%	23.81%
Coil	4.70%	1.74%	37.09%	43.54%
↑正解	30.71%	13.31%	55.99%	100.00%

図5 大きな誤りを学習データから除いた時の未学習データに対する結果

Q_3 70.87%

	Helix	Strand	Coil	← svm の予測
Helix	8089	226	3259	11574
Strand	1122	3874	3447	8443
Coil	1660	613	13166	15439
↑正解	10871	4713	19872	35456

	Helix	Strand	Coil	← svm の予測
Helix	22.81%	0.64%	9.19%	32.64%
Strand	3.16%	10.93%	9.72%	23.81%
Coil	4.68%	1.73%	37.13%	43.54%
↑正解	30.66%	13.29%	56.05%	100.00%

図6 カーネルの σ^2 を 2.0 とした 3 状態の未学習データに対する結果

Q_3 56.58%

	Helix	Strand	Coil	← svm の予測
Helix	4506	0	14084	18590
Strand	39	3639	9580	13258
Coil	346	214	23471	24031
↑正解	4891	3853	47135	55879

	Helix	Strand	Coil	← svm の予測
Helix	8.06%	0.00%	25.20%	33.27%
Strand	0.07%	6.51%	17.14%	23.73%
Coil	0.62%	0.38%	42.00%	43.01%
↑正解	8.75%	6.90%	84.35%	100.00%

図7 連続 21 残基を用いた 3 状態の未学習データに対する結果

Q_3 60.33%

	Helix	Strand	Coil	← svm 予測
Helix	6994	962	3618	11574
Strand	1969	3201	3273	8443
Coil	2823	1420	11196	15439
↑正解	11786	5583	18087	35456

	Helix	Strand	Coil	← svm の予測
Helix	19.73%	2.71%	10.20%	32.64%
Strand	5.55%	9.03%	9.23%	23.81%
Coil	7.96%	4.00%	31.58%	43.54%
↑正解	33.24%	15.75%	51.01%	100.00%

図8 2 次の多項式をカーネル関数として用いた時の 3 状態の未学習データに対する結果

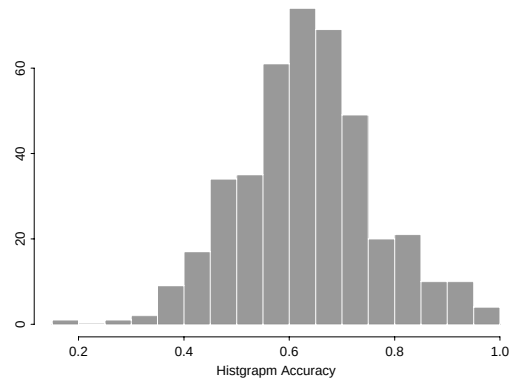


図9 タンパク質単位での予測率

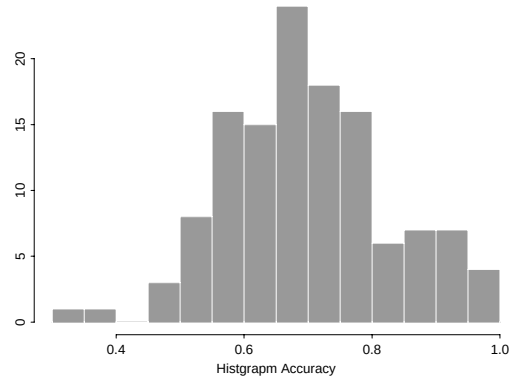


図10 タンパク質単位での予測率 (類似の配列が 40 以上)

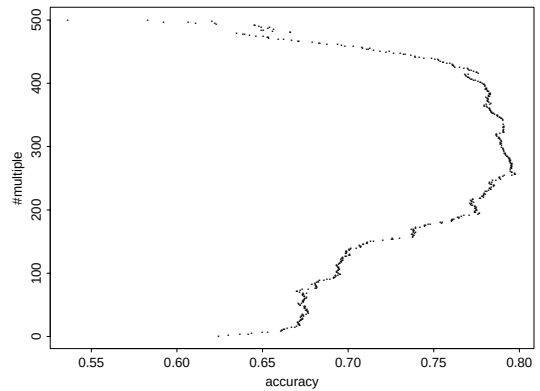


図11 予測率と HSSP で類似の配列の本数