アルゴリズム 82－5
（2002．1．24）

位置依存スコア行列の例からの学習の計算複雑度について

阿久津 達也[1,2]　坂内 英夫[3]
宮野 悟[3,1]　オット ザーシャ[3]

[1] 京都大学 化学研究所 バイオインフォマティクスセンター
[2] 京都大学大学院 情報学研究科 知能情報学専攻
[3] 東京大学 医科学研究所 ヒトゲノム解析センター

位置依存スコア行列（PSSM）はバイオインフォマティクスにおいて幅広く応用されている。例えば、アミノ酸配列群からの共通パターン（モチーフ）検出や、遺伝子制御領域推定、タンパク質の二次構造予測などに利用されている。PSSMは通常は学習データから推定されるが、本稿ではその学習問題について計算論的観点から研究を行なった。具体的には、学習データとして与えた正負の例から、正負の例を正しく分類するPSSMがあるかないかを判定する問題を考え、その問題が一般にはNP困難であるが、PSSMのサイズが定数で押さえられる場合には多項式時間で解けることがわかった。また、PSSMの混合分布の推定などの関連する問題についてもいくつかの結果を得た。

On the Complexity of Deriving Position Specific Score Matrices from Examples

Tatsuya Akutsu[1,2]　Hideo Bannai[3]
Satoru Miyano[3,1]　Sascha Ott[3]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto-Fu 611-0011, Japan.
[2]Dept. Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshidahoncho, Sakyo-ku, Kyoto 606-8501, Japan.
[3]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.

takutsu@kuicr.kyoto-u.ac.jp　{bannai,miyano,ott}@ims.u-tokyo.ac.jp

PSSMs (Position-Specific Score Matrices) have been applied to various problems in Bioinformatics such as protein structure prediction, identification of DNA regulatory regions and detection of motifs. We studied the following problem: given positive examples (sequences) and negative examples (sequences), find a PSSM which correctly discriminates between positive and negative examples. We prove that this problem is solved in polynomial time if the size of the PSSM is bounded by a constant. On the other hand, we prove that this problem is NP-hard if the size is not bounded. Several related results are also presented.

1 Introduction

Position-Specific Score Matrices (PSSMs) have been applied to various problems in *Bioinfomatics* such as detection of remote homology, identification of DNA regulatory regions and detection of motifs [2]. Usually, PSSMs are derived from training data. Therefore, how to derive a good PSSM from training data (examples) has been a key issue in these application. Various methods have been proposed for this purpose. Among them, simple statistical methods based on residue frequencies and local search algorithms (such as Expectation Maximization algorithms) have

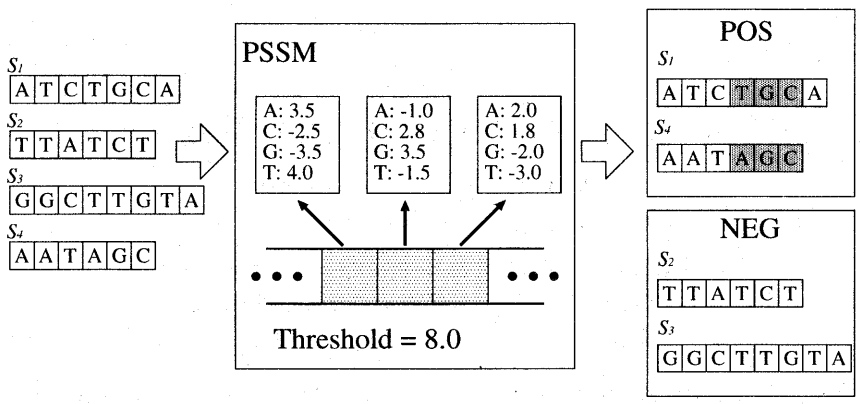

Figure 1: An example of PSSM. S_1 (resp. S_4) is classified as a positive sequence because $f(\mathrm{TGC}) = 9.3 \geq 8.0$ (resp. $f(\mathrm{AGC}) = 8.8 \geq 8.0$).

been widely used [2]. However, from the algorithmic viewpoint, almost no theoretical studies have been done on the derivation of PSSM's. Therefore, we study the following fundamental version of the problem: *given positive examples (sequences) and negative examples (sequences), find a PSSM which completely discriminates between positive and negative examples.* We prove that this problem is NP-hard in general but can be solved in polynomial time if the size of a PSSM is bounded. It follows from the latter result that derivation of position *non-specific* score matrices can be derived from examples in polynomial time, where hydropathic indices [5] are well-known examples of position *non-specific* score matrices.

We also consider derivation of *a mixture of PSSMs.* Mixture of PSSMs are widely used in Bioinformatics since a single PSSM is not always sufficient for characterizing sequences having common biological properties. We consider a special case in which the regions to be identified are already known. Derivation of a single PSSM can be solved in polynomial time by a naive algorithm which uses linear programming. We show that derivation of a mixture of two PSSMs is NP-hard even for this retricted case.

2 Deriving a PSSM from Examples

Let Σ be an alphabet. Let $POS = \{P^1, P^2, \ldots\}$ and $NEG = \{N^1, N^2, \ldots\}$ be sets of strings on Σ, where POS and NEG mean a set of positive examples and a set of negative examples respectively. For string S, $S[i]$ denotes the i-th letter of S and $S_{i,j}$ denotes the substring $S[i]S[i+1]\ldots S[j]$ of S. For strings S_1 and S_2, $S_1 \cdot S_2$ denotes the concatenation of S_1 and S_2. Let L be a positive integer indicating the length of a motif region to be detected.

Definition 1 (PSSM)
A PSSM is a function $f_k(a)$ from $[1, \ldots, L] \times \Sigma$ to the set of real numbers, where $k \in [1, \ldots, L]$ and $a \in \Sigma$ (see Fig. 1).

For string S of length L, we define $f(S)$ by $f(S) = \sum_{i=1}^{L} f_i(S[i])$.

Problem 1 (Derivation of a PSSM from examples)
Given Σ, POS, NEG and L, find a PSSM and a threshold Θ which satisfy the following

conditions:

- For all $P^h \in POS$, $f(P^h_{j,j+L-1}) \geq \Theta$ holds for some $j \in [1,\ldots,|P^h| - L + 1]$,
- For all $N^h \in NEG$ and for all $j = 1,\ldots,|N^h| - L + 1$, $f(N^h_{j,j+L-1}) < \Theta$.

Theorem 1. Problem 1 is NP-hard.

(Proof) We use a polynomial time reduction from 3SAT.

Let $C = \{c_1,\ldots,c_m\}$ be a set of clauses over a set of boolean variables $X = \{x_1,\ldots,x_n\}$, where each clause consists of three literals.

From this example, we construct an instance of the PSSM derivation problem (see Fig. 2). Let $\Sigma = \{0,1\}$ and $L = 4n$. Let $S(i_1,i_2,\ldots)$ denote the string of length $4n$ such that $S[i] = 1$ for $i = i_1,i_2,\ldots$, otherwise $S[i] = 0$. Then, NEG is defined by

$$\begin{aligned} NEG \;=\; & \{\, S(), S(4n) \,\} \cup \{\, S(i,j) \mid 1 \leq i < j < 4n \,\} \cup \\ & \{\, S(i) \mid i = 1,\ldots,4n \,\} \cup \{\, S(2i-1, 2i, 4n) \mid i = 1,\ldots,n \,\}. \end{aligned}$$

It should be noted that the $(2i-1)$-th position and the $(2i)$-th position $(i = 1,\ldots,n)$ of each string correspond to literals x_i and $\overline{x_i}$ respectively.

Let $c_i = l_{i_1} \vee l_{i_2} \vee l_{i_3}$, where l_{i_k} is either x_{i_k} or $\overline{x_{i_k}}$. We define $g(i_k)$ by $g(i_k) = 2i_k - 1$ if $l_{i_k} = x_{i_k}$, otherwise $g(i_k) = 2i_k$. Then, P^i is defined by $P^i = S(g(i_1),4n) \cdot S() \cdots S(g(i_2),4n) \cdot S() \cdots S(g(i_3),4n)$. POS consists of $P^1, P^2, \ldots, P^n$.

First we show that if C is satisfiable then there exists a PSSM f_k satisfying the condition of Problem. 1 for $\Theta = 3$. From the truth assignment to X satisfying all the clauses, we construct f_k by:

- for $i = 1,\ldots,n$,
  $f_{2i-1}(0) = 0$, $f_{2i-1}(1) = 1$, $f_{2i}(0) = 1$ and $f_{2i}(1) = 0$ if x_i is true,
  $f_{2i-1}(0) = 1$, $f_{2i-1}(1) = 0$, $f_{2i}(0) = 0$ and $f_{2i}(1) = 1$ otherwise,
- for $i = 2L+1,\ldots,4n-1$, $f_i(0) = f_i(1) = 0$,
- $f_{4n}(0) = 0$ and $f_{4n}(1) = 1$.

Then, it is easy to see that $f(N^h) < 3$ holds for all $N^h \in N$, and either $f(P^h_{1,4n}) = 3$, $f(P^h_{8n+1,12n}) = 3$ or $f(P^h_{16n+1,20n}) = 3$ holds for all $P^h \in POS$.

Next we show that if there is a PSSM satisfying the condition of Problem 1, there exists a truth assignment satisfying all the clauses in C. Let $\hat{x_i} = f_{2i-1}(1) - f_{2i-1}(0)$, $\hat{\overline{x_i}} = f_{2i}(1) - f_{2i}(0)$ and $\hat{a} = f_{4n}(1) - f_{4n}(0)$. Let $f(S()) = z$. Then, it is easy to check that for all $P^h \in POS$, all substrings of length L except $P^h_{1,4n}$, $P^h_{8n+1,12n}$ and $P^h_{16n+1,20n}$ appear in NEG. Therefore, the following relations hold if Problem 1 has a solution:

- $z < \Theta$, $z + \hat{a} < \Theta$
- $z + \hat{l_i} + \hat{l_j} < \Theta$ for all $l_i \neq l_j$,
- $z + \hat{l_i} < \Theta$ for all l_i,
- $z + \hat{x_i} + \hat{\overline{x_i}} + \hat{a} < \Theta$ for all x_i,
- $z + \hat{l_{i_k}} + \hat{a} \geq \Theta$ holds for some $k \in [1,2,3]$ for all $c = l_{i_1} \vee l_{i_2} \vee l_{i_3}$.

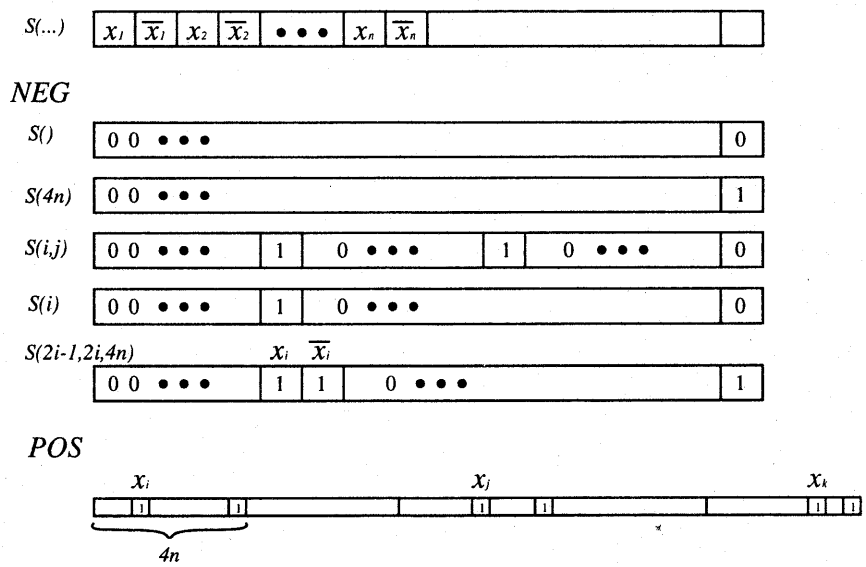

Figure 2: Construction of POS and NEG in Theorem 1.

From f and Θ satisfying the above relations, we construct a truth assignment to X as follows: x_i is true if $z + \hat{x_i} + \hat{a} \geq \Theta$, x_i is false if $z + \hat{\overline{x_i}} + \hat{a} \geq \Theta$, otherwise x_i is arbitrary.

It is sufficient to show that either $z + \hat{x_i} + \hat{a} < \Theta$ or $z + \hat{\overline{x_i}} + \hat{a} < \Theta$ holds. Suppose that $z + \hat{x_i} + \hat{a} \geq \Theta$ held. Then, $\hat{x_i} > 0$ would hold from this inequality and $z + \hat{a} < \Theta$. From $\hat{x_i} > 0$ and $z + \hat{x_i} + \hat{\overline{x_i}} + a < \Theta$, $z + \hat{\overline{x_i}} + a < \Theta$ would hold. □

Theorem 2. Problem 1 can be solved in polynomial time if Σ and L are fixed.
(Proof) We construct an *arrangement* of hyperplanes, where the arrangement is a well-known concept in computational geometry [3]. We construct the arrangement in the $(|\Sigma|L + 1)$-dimensional Euclidean space for the following hyperplanes:

- $f(P^h_{j,j+L-1}) - \Theta = 0$ for $j = 1, \ldots, |P^h| - L + 1$ and for all $P^h \in POS$,
- $f(N^h_{j,j+L-1}) - \Theta = 0$ for $j = 1, \ldots, |N^h| - L + 1$ and for all $N^h \in NEG$.

Then, we pick an arbitrary point (which corresponds to a pair of f and Θ) from each cell and check whether or not the condition of Problem 1 holds. Since the sign of each function (i.e., $f(P^h_{j,j+L-1}) - \Theta$, $f(N^h_{j,j+L-1}) - \Theta$) does not change within a cell, this algorithm correctly solves Problem 1.

Since the arrangement of hyperplanes in fixed dimensions can be constructed in polynomial time and the combinatorial complexity of the arrangement is also polynomially bounded [3], the algorithm works in polynomial time. □

Hydropathic indices have been used for the identification of transmembrane domains of membrane proteins. Usually, hydropathic indices are not position-specific, i.e., $f_i[a] = f_j[a]$ for

all $i \neq j$. The algorithm above can also be applied to this case. In this case, the arrangement in the $(|\Sigma|+1)$-dimensional Euclidean space is constructed. Since $|\Sigma|$ is 4 or 20, we have:

Corollary 1. Hydropathic indices satisfying the condition of Problem 1 can be derived from examples in polynomial time.

In most cases in deriving hydropathic indices, positive examples given as training data contain information about the positions of the transmembrane domains. Therefore, various learning algorithms have been applied to derivation of hydropathic indices. But, the above theorem suggests that hydropathic indices can be derived even if the positions of the transmembrane domains are not known. Of course, the time complexity of the algorithm is still too high even for $|\Sigma| = 4$. Thus, an improved algorithm should be developed.

3 Deriving a mixture of PSSMs from Examples

In this section, we consider the following problem.

Problem 2. (Derivation of mixture of PSSMs from examples)
Given Σ, POS, NEG, L and N where N denotes the number of PSSMs, find a set of PSSMs with cardinality N and a threshold Θ which satisfy the following conditions:

- For all $P^i \in POS$, $f^k(P^i_{j,j+L-1}) \geq \Theta$ holds for some $j \in [1,\ldots,|P_i|-L+1]$ and for some $k \in [1,\ldots,N]$,
- For all $N^i \in NEG$, for all $j = 1,\ldots,|N^i|-L+1$ and for all $k \in [1,\ldots,N]$, $f^k(N^i_{j,j+L-1}) < \Theta$,

where f^k denotes the score given by the k-th PSSM.

Clearly, Problem 2 is NP-hard from Theorem 1. We are interested in the case where all of strings are of the same size L (**Problem 3**).

Proposition 1. Problem 3 can be solved in polynomial time if $N = 1$ or $N \geq |POS|$.
(Proof) The case of $N = 1$ is trivial and well-known. We simply construct the linear inequalities: $f^1(P^h) \geq \Theta$ for all $P^h \in POS$, and $f^1(N^h) < \Theta$ for all $N^h \in NEG$. Then, we can obtain PSSM and Θ by applying any polynomial time algorithm for linear programming.
In the case of $N = |POS|$, we solve the following inequalities: $f^h(P^h) \geq \Theta$ for all $P^h \in POS$, and $f^k(N^h) < \Theta$ for all $N^h \in NEG$ and for all $k \in [1 \ldots N]$. □

Theorem 3. Problem 3 is NP-hard even for $N = 2$.
(Proof) We reduce NOT-ALL-EQUAL 3SAT (LO3 in [4]) to Problem 3.
Let U be a set of variables and C be a set of clauses over U such that each clause $c \in C$ has $|c| = 3$. Given the instance (U, C) for NOT-ALL-EQUAL 3SAT, we define an instance $I(U, C)$ for Problem 3 as follows.
Let L denote the set of literals over U and let $p : L \to \{1,\ldots,2|U|\}$ be a bijection such that for each $u \in U$ $p(\overline{u}) = p(u) + 1$ holds. We define POS as the set

$$POS = \{0^i 10^{2|U|-i-1} \mid i = 0,\ldots,2|U|-1\}$$

and NEG as

$$\begin{aligned} NEG \;=\; & \{\, 0^i 110^{2|U|-i-2} \mid i = 0,2,\ldots,2|U|-2 \,\} \cup \{\, 0^{2|U|} \,\} \cup \\ & \{\, 0^{p(x)-1}10^{p(y)-p(x)-1}10^{p(z)-p(y)-1}10^{2|U|-p(z)} \mid \{\, x,y,z \,\} \in C,\ p(x) < p(y) < p(z) \,\}. \end{aligned}$$

We denote the Problem 3-instance consisting of POS and NEG as $I(U, C)$. We say a string in POS corresponds to a literal in L, iff the 1 appears at the $p(L)$-th position in the string. For a literal $l \in L$, we denote the string corresponding to l with w_l. In the same way, we denote strings $0^i 110^{2|U|-i-2} \in NEG$ as $w_{u,\overline{u}}$ for the variable $u \in U$ with $p(u) = i + 1$.

We have to show that there is a not-all-equal truth assignment for (U, C), iff there is a solution for $I(U, C)$ with 2 matrices.

Let σ be a not-all-equal truth assignment. Call the set of all strings of POS, for which the corresponding literal is satisfied by σ, POS_A, and the set of all other strings of POS POS_B. Let A denote the PSSM, which assigns the value 1 to all occurrences of the character 1 at positions corresponding to satisfied literals, the value -2 to all other occurrences of 1, and the value 0 to all occurrences of 0. In the same way, let B denote the PSSM assigning 1 to the occurrences of 1 at positions corresponding to literals not satisfied by σ and -2 to the other occurrences.

For $\Theta = 1$, A accepts all strings in POS_A and B accepts all strings in POS_B. Furthermore, for every clause $c \in C$, there is a literal in c satisfied by σ and a literal not satisfied by σ. Therefore, for every string $s \in NEG$, $A(s) \leq 0$ and $B(s) \leq 0$, which shows that (A, B) is a solution for $I(U, C)$.

It remains to show, that there is a not-all-equal truth assignment for C, if there is a solution for $I(U, C)$. Let A, B be PSSMs solving $I(U, C)$. Let σ denote the truth assignment satisfying all literals, for which the corresponding string in POS is accepted by A, and unsatisfying all other literals. If there were a variable $u \in U$ with $\sigma(u) = \sigma(\overline{u})$, then $A(w_u) + A(w_{\overline{u}}) \geq 2\Theta$ or $B(w_u) + B(w_{\overline{u}}) \geq 2\Theta$ would hold. Therefore, since $A(w_u) + A(w_{\overline{u}}) = A(0^{2|U|}) + A(w_{u,\overline{u}})$ holds (and analogously for B), $0^{2|U|}$ or $w_{u,\overline{u}}$ would be accepted by either A or B, a contradiction. Thus, σ is well-defined.

To see that σ has the not-all-equal property, assume that there is a clause $\{x, y, z\} \in C$ with $\sigma(x) = \sigma(y) = \sigma(z)$. If $\sigma(x) = 1$, then we have

$$A(w_x) + A(w_y) + A(w_z) \geq 3\Theta.$$

Moreover,

$$A(w_x) + A(w_y) + A(w_z) = 2A(0^{2|U|}) + A(w_{x,y,z})$$

holds, where $w_{x,y,z}$ denotes the string of NEG corresponding to clause $\{x, y, z\}$. Therefore, since $A(w_{x,y,z}) < \Theta$ holds, we have $A(0^{2|U|}) > \Theta$, contradicting $0^{2|U|} \in NEG$. If $\sigma(x) = 0$, the same contradiction follows for B.

□

4 Concluding Remarks

In this paper, we have shown that derivation of a PSSM is NP-hard in general but is polynomial time solvable if the size of the PSSM is bounded by a constant. We also showed that derivation of mixture of two PSSM's is NP-hard even if the regions to be identified are known. Development of approximation algorithms for the NP-hard problems and development of faster algorithms for derivation of bounded-size PSSMs are important future work.

Relating to derivation of PSSMs, Akutsu and Yagiura studied the following problem [1]: given correct examples and incorrect examples, find a score function with which the scores of the correct examples are optimal and the scores of the incorrect examples are not optimal. They proved that this problem is computationally hard for protein threading with contact potentials and for multiple alignment with SP-scoring. On the other hand, they showed that this problem is solvable in polynomial time for pairwise alignment and RNA secondary structure prediction with simple pseudo-energy functions, by using reductions to linear programming. We recently

consider the problem of deriving score matrices for pairwise alignment under the condition that each (positive or negative) example consists of a pair of sequences (i.e., alignment results are not given). We proved that this problem is NP-hard for general Σ. This result is interesting because this general case can be solved in polynomial time if alignment results are given [1]. It is not yet known whether this problem is NP-hard for fixed Σ.

Acknowledgement

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas (C) for "Genome Information Science" from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan. Tatsuya Akutsu was also partially supported by HITOCC (Hyper Information Technology Oriented Corporation Club) and Grant-in-Aid #13680394 from MEXT, Japan.

References

[1] Akutsu, T. and Yagiura, M., On the complexity of deriving score functions from examples for problems in molecular biology, *Proc. ICALP'98 (LNCS 1443)*, 832-843, 1998.

[2] Durbin, R., Eddy, S., Krogh, A. and Mitchison, G., *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998.

[3] Edelsbrunner, H., *Algorithms in Combinatorial Geometry*, Springer-Verlag, 1987.

[4] Garey, M. R. and Johnson, D. S., *Computers and Intractability*, Freeman, 1979.

[5] Kyte, J. and Doolittle, R. F, A simple method for displaying the hydropathic character of a protein, *J. Molecular Biology*, 157, 105-132, 1982.