Prediction of Microbial Characters by a Computerized Learning Machine

Part II. Application to Pseudomonads

By Eitarō MASUO and Toshio NAKAGAWA

*Shionogi Research Laboratory, Shionogi & Co. Ltd.,
Fukushima-ku, Osaka*

Received May 19, 1971

A computerized learning machine was applied to pseudomonads to predict the presence or absence of the following four features; the production of phenazine compounds, the utilizations of D-sorbitol, of *n*-hexadecane, and of histamine. On all of the three measures of evaluation proposed in Part I, the results obtained for all of the above four features were far better than those expected in the case of random guess.

The object of this series of papers is to see if a computerized learning machine can be used effectually to predict a microbial character, or more strictly speaking, to conjecture the presence or absence of a certain feature in microorganisms under examination. The principle and calculation process of a learning machine have been described in the preceding paper. In the present paper are reported the results of application to pseudomonads. Predictions were made for the following four features, (a) the production of phenazine compounds, (b) the utilization of D-sorbitol, (c) the utilization of *n*-hexadecane, and (d) the utilization of histamine. The first three features were taken up on account of the authors' interest in (a) the antibiotic and anti-cancer potency of phenazine compounds, (b) the fermentational production of ascorbic acid, and (c) the industrial fermentation of petroleum. The last feature, utilization of histamine, was examined because the prediction of this feature seemed very difficult to the authors.

On all of the three measures of evaluation proposed in Part I, the results obtained for all of the above four features were far better than those expected in the case of random

guess.

MATERIALS AND METHODS

Microorganisms and features. R. Y. Stanier *et al.* have published a paper which described more than 100 features for 267 strains of genus *Pseudomonas*.¹⁾ Because of missing data and ambiguous descriptions, however, the present authors picked up only 214 strains and 94 features, in which the presence or absence of every feature is known for every strain. These 214 strains cover more than 17 species in terms of current taxonomy (Table I), and are maintained in the American Type Culture Collection (ATCC). They were divided into two groups, A and B, both of which contained 107 strains as shown in Table II. Table III illustrates the selected 94 features, four of which (the underlined ones) were subjected to the computer prediction. For both groups was prepared a character table that listed up the presence (denoted by 1) or absence (denoted by 0) of every feature for every strain.

A number of phenazine derivatives, such as pyocyanine, 1-hydroxyphenazine, iodinin, chlororaphine, aeruginosin, griseolutein, and several related compounds are known to be produced by some strains of

1) R. Y. Stanier, N. J. Palleroni and M. Doudoroff, *J. Gen. Microbiol.*, **43**, 159 (1966).

TABLE I. MICROORGANISMS EXAMINED

Species	No. of strains
(I) <i>P. acidovorans</i>	12
(II) <i>P. aeruginosa</i>	29
(III) <i>P. angulata</i>	1
(IV) <i>P. aureofaciens</i>	7
(V) <i>P. chlororaphis</i>	13
(VI) <i>P. denitrificans</i>	1
(VII) <i>P. desmolytica</i>	5
(VIII) <i>P. fluorescens</i>	65
(IX) <i>P. genticulata</i>	3
(X) <i>P. indoloxidans</i>	1
(XI) <i>P. lemonnieri</i>	1
(XII) <i>P. marginalis</i>	2
(XIII) <i>P. marginata</i>	1
(XIV) <i>P. multivorans</i>	18
(XV) <i>P. putida</i>	38
(XVI) <i>P. schuylkilliensis</i>	1
(XVII) <i>P. testosteroni</i>	5
(XVIII) <i>Pseudomonas</i> sp.	11

TABLE II. GROUPING OF TEST MICROORGANISMS

A-group				B-group			
(I)	5	(X)	0	(I)	7	(X)	1
(II)	14	(XI)	1	(II)	15	(XI)	0
(III)	1	(XII)	1	(III)	0	(XII)	1
(IV)	3	(XIII)	0	(IV)	4	(XIII)	1
(V)	7	(XIV)	9	(V)	6	(XIV)	9
(VI)	1	(XV)	19	(VI)	0	(XV)	19
(VII)	4	(XVI)	0	(VII)	1	(XVI)	1
(VIII)	32	(XVII)	2	(VIII)	33	(XVII)	3
(IX)	2	(XVIII)	6	(IX)	1	(XVIII)	5
Total 107				Total 107			

Pseudomonas. In the present work, these compounds having a phenazine skeleton were lumped together under the term of "phenazine compounds." D-Sorbitol is utilized by pseudomonads through either of two courses, *via* L-sorbose or *via* D-fructose. However, the discrimination of the two path-ways was not attempted in the present work, because insufficient information was available.

Weight determination and prediction. For each of the four features to be predicted, 94 weights (w_1 , w_2 , ..., w_{93} , and w_0) were determined by processing the character table of A-group with the calculation

method detailed in Part I. The weights obtained were used to predict the presence or absence of the feature in question for the pseudomonads in B-group. In the original character table of A-group, 107 strains were arranged in the order shown in the left half of Table II (classification order). By randomizing the order of arrangement, a new character table was prepared (random order). The weights obtained by processing this new table gave similar but different predictions for B-group. In the same way, predictions were made for A-group by exchanging the roles of A- and B-groups. Thus, four sets of predictions were worked out with respect to each of the four features (Table IV).

An IBM-1620 computer was used for calculation.

RESULTS AND DISCUSSION

Perfect recognition was attained both for A- and B-groups in the cases of production of phenazine compounds, utilizations of D-sorbitol, and of *n*-hexadecane. In the case of utilization of histamine, the "Per cent recognition" was 99.07% for A-group arranged in the classification order, 96.26% for A-group in random order, 97.20% for B-group in classification order, and 94.39% for B-group in random order.*

The results of predictions are summarized in Table IV. For example, the first row shows the result with respect to the production of phenazine compounds, predictions being made for B-group by using the weights obtained from A-group arranged in the classification order. The last three columns indicate the three measures (predictive ability, EAF, and EPF) evaluating prediction results, about which detailed explanation has been given in Part I.

Recalling that the predictive ability is statistically equal to 0.5 for the random guess and equal to 1 for the perfect prediction, one may conclude that the prediction by a learning machine was much superior to the random

* The meaning of "recognition" has been described in the Principle in Part I.

TABLE III. SELECTED FEATURES

1 Possession of multitrichous flagella	48 <i>n</i> -Butanol
2 Production of phenazine compounds	49 Isobutanol
3 Production of fluorescent pigments	50 Geraniol
4 Growth at 4°C	51 Benzoylformate
5 Growth at 41°C	52 Benzoate
6 Accumulation of poly- β -OH-butyrate	53 <i>m</i> -Hydroxy-benzoate
7 Denitrification ($\text{NO}_3 \rightarrow \text{NO}_2$)	54 <i>o</i> -Hydroxy-benzoate
8 Gelatin-liquefaction	55 Phenylacetate
9 Lipase production	56 Phenol
10 Egg-yolk reaction ^{a)}	57 Testosterone
11 Possession of arginine dehydrase	58 Glycine
12 Diphenol decomposition ^{b)}	59 D- α -Alanine
Utilization of ^{c)}	60 β -Alanine
13 D-Xylose	61 L-Serine
14 L-Arabinose	62 L-Threonine
15 D-Glucose	63 L-Leucine
16 D-Galactose	64 L-Isoleucine
17 D-Fructose	65 L-Norleucine
18 Sucrose	66 L-Valine
19 Trehalose	67 L-Lysine
20 2-Keto-gluconic acid	68 L-Arginine
21 Saccharic acid	69 DL-Ornithine
22 Butyric acid	70 L-Citrulline
23 Valerate	71 γ -Aminobutyrate
24 Caprate	72 δ -Aminovalerate
25 Caprylate	73 L-Phenylalanine
26 Pelargonate	74 L-Tryptophan
27 Malonate	75 L-Kynurenine
28 Maleate	76 Kynurenate
29 Adipate	77 Anthranilate
30 Azelate	78 Ethanolamine
31 L-Malate	79 Benzylamine
32 L-(+)-Tartrate	80 Putrescine
33 Glycollate	81 Histamine
34 Hydroxymethylglutarate	82 Butylamine
35 Laevulinate	83 α -Amylamine
36 Citraconate	84 Creatine
37 Itaconate	85 Urea
38 Mesaconate	86 Acetamide
39 D-Mannitol	87 Trigonelline
40 D-Sorbitol	88 <i>n</i> -Hexadecane
41 Mesoinositol	89 D-Mannose
42 Adonitol	90 Betaine
43 Glycerol	91 D-Fucose
44 Propyleneglycol	92 Cellobiose
45 2,3-Butyleneglycol	93 Mucate
46 Ethanol	94 Pimelate
47 <i>n</i> -Propanol	

^{a)} Cf. Manual of Microbiological Methods (1957).

^{b)} Cf. reference 1. *ortho*-Cleavage was recorded with 1.

^{c)} The exact meaning of "utilization of something" is "utilization of something as carbon and/or energy source(s)".

FIG. 1. Distribution of Mispredictions in the Case of Production of Phenazine Compounds.

Predictions being made for A-group by using the weights obtained from B-group arranged in random order.

- ×: Correctly predicted strain (hit).
- : Incorrectly predicted strain (miss).

guess in all of the examined cases. The same conclusion is derived from EAF and EPF, both of which are statistically equal to 1 for the random guess and equal to 2 for the perfect prediction.

As mentioned in Part I, the prediction, *i.e.* the decision making between the presence and absence of a feature in question, is done solely based on the sign of S . However, in reconsidering the principle of a learning machine detailed in Part I, one perhaps conceives that the probability of misprediction may be the larger for an organism possessing the smaller absolute value of S . Figures 1, 2 and 3 reveal that this conception is generally true. Solid circles in the lower halves of Figs. 1 and 2, each of which indicates a misprediction, dis-

FIG. 2. Distribution of Mispredictions in the Case of Utilization of Histamine.

Predictions being made for B-group by using the weights obtained from A-group arranged in classification order.

- ×: Correctly predicted strain (hit).
- : Incorrectly predicted strain (miss).

tribute densely in the proximity of $S=0$. The curve depicted in the upper half illustrates the mode of the distribution. The height of curve at $S=a$ represents the ratio of mispredictions (the number of mispredictions divided by the number of predictions) in a range from $S=0$ to $S=a$.* Figure 3 reproduces all of the sixteen distribution curves

* Near $S=0$, only a small number of predictions are available to calculate the height of curve, and therefore, the shape of curve cannot be accurately determined. The curves in Figs. 1, 2 and 3 have been drawn by assuming tentatively that their height at $S=0$ is 0.5.

TABLE IV. PREDICTION RESULTS

Feature to be predicted	Group to be predicted	Order of arrangement	Predictive ability	EAF	EPF
Production of phenazine compounds	B	class.	0.93	1.84	1.74
	B	random	0.93	1.84	1.74
	A	class.	0.92	1.72	1.83
	A	random	0.92	1.72	1.83
Utilization of D-sorbitol	B	class.	0.95	1.89	1.90
	B	random	0.94	1.88	1.88
	A	class.	0.92	1.82	1.81
	A	random	0.93	1.85	1.86
Utilization of <i>n</i> -hexadecane	B	class.	0.93	1.70	1.61
	B	random	0.91	1.68	1.53
	A	class.	0.93	1.71	1.71
	A	random	0.87	1.33	1.50
Utilization of histamine	B	class.	0.83	1.67	1.65
	B	random	0.82	1.67	1.67
	A	class.	0.81	1.63	1.64
	A	random	0.80	1.61	1.61

Underlined are the largest and smallest values.

FIG. 3. The Ratio of Mispredictions as a Function of S .

A: The production of phenazine compounds.

B: The utilization of D-sorbitol.

C: The utilization of *n*-hexadecane.

D: The utilization of histamine.

Prediction for	by using the weights obtained from	arranged in
— B-group	A-group	class. order
- - - B-group	A-group	random order
- · - A-group	B-group	class. order
···· A-group	B-group	random order

obtained.

When compared with the random guess, this true when compared with predictions given by trained biologists? The authors' the superiority of predictions given by a efforts are now directed to answering this learning machine seems to be established. Is question.