Proc. 2009 IEEE/ICME Int'l Conf.on Complex Medical Engineering (CME'09), 1-4, 2009

Extracting Sequential Episodes Representing Replacements of Bacteria from Bacterial Culture Data

Takashi Katoh*, Kouichi Hirata†, Hiroki Arimura*, Shigeki Yokoyama‡, and Kimiko Matsuoka§

***Abstract*—A *sequential episode*, which is the simplest form of serial episodes, is an episode of the form $A \to B$. This sequential episode means that an event type A is precedent to an event type B. In this paper, we extract the sequential episodes *representing the replacements of bacteria* from Osaka Prefecture General Medical Center in years from 2000 to 2005.**

I. Introduction

In Complex Medical Engineering, *data mining* is one of the key techniques to analyze medical data from medical viewpoint [10]. In order to prevent hospital-acquired infection, we have applied several data mining techniques to bacterial culture data in our previous works [1], [2], [3], [4], [5], [6], [7], [9], [11]. In particular, in order to extract the time-related rules representing *replacements of bacteria* and *changes for drug resistance* as the factors of hospital-acquired infection, we have paid our attention to *episode mining* from the bacterial culture data in the works [4], [5], [6], [7].

The *episode mining*, introduced by Mannila *et al.* [8], is known as one of the methods to discover frequent patterns from time-related data. The purpose of episode mining is to discover *frequent episodes* that are a collection of event types occurring frequently together in event sequences. In episode mining, the frequency is formulated as the number of occurrences of episodes in every *window* that is a subsequence of event sequences under a fixed time span called the *width* of windows.

The above bacterial culture data were proposed from Osaka Prefectural General Medical Center in years from 1995 to 1998 consisting of 15784 records. Unfortunately, since the number of samples in the data is just 8, the extracted rules in [4], [5], [6], [7] were not enough to characterize *replacements of bacteria* and *changes for drug resistance* in the data accurately.

Hence, in this paper, we adopt a new and improved version of the bacterial culture data from Osaka Prefectural General Medical Center in years from 2000 to 2005 consisting of 35827 records and, in particular, of which number of samples is 30. As the simplest form of *serial episode* [8], we introduce a *sequential episode* of the form $A \to B$ where A and B are event types. Then, we extract a sequential episode $A \to B$ *representing the replacements of bacteria*, where A and B are event types of the occurrences of bacteria, from the new data.

*Graduate School of Information Science and Technology, Hokkaido University, Kita 14-jo Nishi 9-chome, Sapporo 060-0814, Japan. Email: {t-katou,arim}@ist.hokudai.ac.jp

†Department of Artificial Intelligence, Kyushu Institute of Technology, Kawazu 680-4, Iizuka 820-8502, Japan. Email: hirata@ai.kyutech.ac.jp

‡Koden Industrial Co., Ltd, Ohmori Nishi 2-22-9, Ohta, Tokyo 143-0015, Japan. Email: shigey@st.rim.or.jp

§Osaka Prefectural General Medical Center, Bandai Higashi 3-1-56, Sumiyoshi, Osaka 558-8558, Japan. Email: cby41060@pop01.odn.ne.jp

II. Sequential Episode

Let $\mathcal{E}$ be the set of *event types*. Then, a pair (e, t) is called an *event*, where $e \in \mathcal{E}$ and t is a natural number which is the *occurrence time* of e. An *event sequence* $\mathcal{S}$ on $\mathcal{E}$ is the set of events, where we deal with bacterial culture data as an event sequence.

A *sequential episode* is of the form $A \to B$, where A and B are event types. We say that a sequential episode $A \to B$ *occurs* in an event sequence $\mathcal{S}$ if both A and B occur in $\mathcal{S}$ and A is precedent to B.

For an event sequence $\mathcal{S}$, a subset $\{(e, t) \in \mathcal{S} \mid t_s \leq t < t_e\}$ of $\mathcal{S}$ is called a *window* in $\mathcal{S}$ with *width* $t_e - t_s$. We denote the set of all windows in $\mathcal{S}$ with width k by $W_{\mathcal{S}}(k)$, and the set of all windows in $\mathcal{S}$ with width k in which a sequential episode $A \to B$ occurs by $W_{\mathcal{S}}(k, A \to B)$.

Let $\mathcal{S}$ be an event sequence k an integer. Then, we formulate the *support* $supp_{\mathcal{S},k}(A \to B)$ of a sequential episode $A \to B$ as $|W_{\mathcal{S}}(k, A \to B)|/|W_{\mathcal{S}}(k)|$. In the following, the subscripts $\mathcal{S}$ and k are omitted if they are clear by the context. For the *minimum support* σ $(0 \leq \sigma < 1)$, we say that a sequential episode $A \to B$ is *frequent* if $supp(A \to B) \geq \sigma$.

By using the following naïve algorithm, we extract the frequent sequential episodes of the from $A \to B$ *representing the replacements of bacteria* that the occurrence of the bacterium A is followed by one of B from the bacterial culture data.

1) For each of 30 samples, construct an event sequence by fixing the sample and connecting data of every patient with the span of 15 days.
2) Apply the procedure from 3) to 5) to the event sequence constructed by 1).
3) Compute the number of windows in an event sequence and the number of windows that a bacterium occurs.
4) For every pair (A, B) of bacteria, compute the number of windows that a sequential episode $A \to B$ representing the replacement of bacteria and then compute $supp(A \to B)$.
5) Output the frequent sequential episodes and their information of occurrences.

III. Experimental Results

In this section, we give the experimental results for extracting frequent sequential episodes representing the replacements of bacteria. In Section III-A, we summary the number of windows and all of the sequential episodes under the minimum support 0%. In Section III-B, we give the frequent sequential episodes for the samples of which number of windows is greater than 10000. In Section III-C, we give the frequent sequential episodes under the minimum support 1%.

A. Data

The data are provided from Osaka Prefectural General Medical Center in years from 2000 to 2005 consisting of 35827 records. Figure 1 describes the number of windows (#windows) and the extracted sequential episodes under the minimum support 0% (#episodes) for each sample sorted by #windows.

sample	#windows	#episodes
sputum	185745	1627
pus	109103	1700
respiratory mucus	94689	273
urine	58699	265
blood	19117	169
stool	18820	85
indwelling vascular catheter	9772	70
tissue	7111	87
drain	5993	390
pertoneal fluid	4604	483
internal organ	4240	0
bile	3413	152
pleural effusion	2782	164
puncture fluid	1893	106
perfusate	1654	7
spinal fluid	1033	1
secretion (prostate gland, urethra)	980	2
amnion liquid	840	0
catheter	460	10
joint fluid	425	10
gastric fluid	245	2
heart pacer	171	1
secretion (cornea, conjunctiva)	120	0
intestinal fluid	120	0
attachment	77	0
fluid of bone marrow	57	0
lymph node	45	0
duodenal fluid	45	0
breast milk	30	0
nasal cavity	0	–

Fig. 1. The number of windows and extracted sequential episodes under the minimum support 0% for each of 30 samples.

In the remainder of this section, we refer both the sample and the number of windows of the sample to the form of "sample (#windows)."

B. The most 5 frequent sequential episodes

In this subsection, we describe the most frequent 5 sequential episodes extracted from 6 samples, blood, respiratory mucus, pus, stool, urine and sputum, of which number of windows is more than 10000. Here, #patients is the number of different patients included by the sequential episode.

1) Sputum: The most 5 frequent sequential episodes for the sample of sputum (185745) are described as follows.

replacement of bacteria	frequency (%) #patients
Staphylococcus aureus → Pseudomonas aeruginosa	3010 (1.621%) 179
Pseudomonas aeruginosa → Staphylococcus aureus	2662 (1.433%) 168
yeast → Staphylococcus aureus	2609 (1.405%) 192
Staphylococcus aureus → yeast	2566 (1.381%) 201
yeast → Pseudomonas aeruginosa	1189 (0.640%) 101

2) Pus: The most 5 frequent sequential episodes for the sample of pus (109103) are described as follows.

replacement of bacteria	frequency (%) #patients
Staphylococcus aureus → Pseudomonas aeruginosa	384 (0.352%) 36
Staphylococcus aureus → Enterococcus faecalis (Group D)	377 (0.346%) 42
Pseudomonas aeruginosa → Staphylococcus aureus	376 (0.345%) 29
Enterococcus faecalis (Group D) → Staphylococcus aureus	373 (0.342%) 38
Staphylococcus aureus → Escherichia coli	252 (0.231%) 21

3) Respiratory mucus: The most 5 frequent sequential episodes for the sample of respiratory mucus (94689) are described as follows.

replacement of bacteria	frequency (%) #patients
Staphylococcus aureus → yeast	298 (0.315%) 29
Staphylococcus aureus → Pseudomonas aeruginosa	215 (0.227%) 18
Staphylococcus aureus → Staphylococcus coagulase (-)	185 (0.195%) 25
yeast → Staphylococcus aureus	166 (0.175%) 17
Pseudomonas aeruginosa → Staphylococcus aureus	151 (0.159%) 13

4) Urine: The most 5 frequent sequential episodes for the sample of urine (58699) are described as follows.

replacement of bacteria	frequency (%) #patients
Escherichia coli → yeast	124 (0.211%) 12
Escherichia coli → Enterococcus faecalis (Group D)	119 (0.203%) 12
Staphylococcus aureus → Pseudomonas aeruginosa	115 (0.196%) 17
Enterococcus faecalis (Group D) → Staphylococcus aureus	109 (0.186%) 14
Enterococcus faecalis (Group D) → Pseudomonas aeruginosa	104 (0.177%) 12

5) Blood: The most 5 frequent sequential episodes for the sample of blood (19117) are described as follows.

replacement of bacteria	frequency (%) #patients
Staphylococcus aureus → Streptococcus constellatus	117 (0.612%) 9
Staphylococcus aureus → Streptococcus intermedius	112 (0.586%) 8
Streptococcus intermedius → Staphylococcus aureus	111 (0.581%) 9
Streptococcus constellatus → Staphylococcus aureus	96 (0.502%) 7
Klebsiella pneumoniae → Bacteroides distasonis	50 (0.262%) 4

6) Stool: The most 5 frequent sequential episodes for the sample of stool (18820) are described as follows.

replacement of bacteria	frequency (%) #patients
Staphylococcus aureus → yeast	335 (1.780%) 29
yeast → Staphylococcus aureus	206 (1.094%) 17
Pseudomonas aeruginosa → yeast	122 (0.648%) 9
Pseudomonas aeruginosa → Staphylococcus aureus	109 (0.579%) 10
yeast → Pseudomonas aeruginosa	71 (0.377%) 6

7) Summary: We can observe the replacements of bacteria between Staphylococcus aureus and (1) Streptococcus constellatus or Streptococcus intermedius for the sample of blood and (2) Pseudomonas aeruginosa for the samples except blood. Also the occurrences of bacteria of Enterococcus faecalis (Group D) and Escherichia coli are characterized as the samples of pus and urine. Furthermore, we can observe the replacement from/to yeast for the samples of respiratory mucus, stool, urine and sputum.

On the other hand, by focusing the number of different patients, we can observe the sequential episodes with more than 100 different patients for the sample of sputum, and ones with more than 10 different patients for the 3 samples of respiratory mucus, pus and urine.

C. The frequent sequential episodes under the minimum support 1%

In this subsection, we describe all of the frequent sequential episodes under the minimum support 1%. We can extract them for the following 10 samples; joint fluid, pleural effusion, sputum, bile, pertoneal fluid, catheter, stool, heart pacer, gastric fluid and secretion (prostate gland, urethra).

1) Joint fluid: For the sample of joint fluid (425), we can extract 10 frequent sequential episodes under the minimum support 1% as follows.

replacement of bacteria	frequency (%) #patients
Streptococcus agalactiae (Group B) → Streptococcus constellatus	12 (2.824%) 1
Streptococcus constellatus → Staphylococcus aureus	10 (2.353%) 1
Bacteroides fragilis → Streptococcus constellatus	7 (1.647%) 1
Bacteroides fragilis → Prevotella loescheii/denticola	7 (1.647%) 1
Bacteroides uniformis → Prevotella oralis	7 (1.647%) 1
Bacteroides uniformis → Streptococcus constellatus	7 (1.647%) 1
Bacteroides fragilis → Prevotella oralis	7 (1.647%) 1
Bacteroides uniformis → Prevotella loescheii/denticola	7 (1.647%) 1
Prevotella oralis → Prevotella loescheii/denticola	7 (1.647%) 1
Prevotella oralis → Streptococcus constellatus	7 (1.647%) 1

2) Pleural effusion: For the sample of pleural effusion (2782), we can extract 6 frequent sequential episodes under the minimum support 1% as follows.

replacement of bacteria	frequency (%) #patients
Escherichia coli → Enterococcus faecalis (Group D)	37 (1.330%) 2
Bacteroides distasonis → Escherichia coli	33 (1.186%) 2
Bacteroides distasonis → Enterococcus faecalis (Group D)	33 (1.186%) 2
Escherichia coli → Fusobacterium varium	30 (1.078%) 2
Candida albicans → Enterococcus faecalis (Group D)	29 (1.042%) 2
Enterococcus faecalis (Group D) → Escherichia coli	29 (1.042%) 2

3) Sputum: For the sample of sputum (185745), we can extract 4 frequent sequential episodes under the minimum support 1% as follows.

replacement of bacteria	frequency (%) #patients
Staphylococcus aureus → Pseudomonas aeruginosa	3010 (1.621%) 179
Pseudomonas aeruginosa → Staphylococcus aureus	2662 (1.433%) 168
yeast → Staphylococcus aureus	2609 (1.405%) 192
Staphylococcus aureus → yeast	2566 (1.381%) 201

4) Bile: For the sample of bile (3413), we can extract 3 frequent sequential episodes under the minimum support 1% as follows.

replacement of bacteria	frequency (%) #patients
Staphylococcus aureus → Candida albicans	46 (1.348%) 4
Candida albicans → Staphylococcus aureus	40 (1.172%) 3
Escherichia coli → Enterococcus faecalis (Group D)	39 (1.143%) 1

5) Pertoneal fluid: For the sample of pertoneal fluid (4604), we can extract 3 frequent sequential episodes under the minimum support 1% as follows.

replacement of bacteria	frequency (%) #patients
Enterococcus faecalis (Group D) → Escherichia coli	54 (1.173%) 6
Escherichia coli → Pseudomonas aeruginosa	53 (1.151%) 5
Escherichia coli → Fusobacterium varium	48 (1.043%) 6

6) Catheter: For the sample of catheter (460), we can extract 2 frequent sequential episodes under the minimum support 1% as follows.

replacement of bacteria	frequency (%) #patients
Enterobacter cloacae → Staphylococcus coagulase (-)	6 (1.304%) 1
Enterococcus faecium (Group D) → Staphylococcus coagulase (-)	6 (1.304%) 1

7) Stool: For the sample of stool (18820), we can extract 2 frequent sequential episodes under the minimum support 1% as follows.

replacement of bacteria	frequency (%) #patients
Staphylococcus aureus → yeast	335 (1.780%) 29
yeast → Staphylococcus aureus	206 (1.094%) 17

8) Heart pacer, gastric fluid and secretion (prostate gland, urethra): For the samples of heart pacer (171), gastric fluid (245) and secretion (prostate gland, urethra) (120), we can extract just one frequent sequential episodes under the minimum support 1% as follows under this order of samples.

replacement of bacteria	frequency (%) #patients
Staphylococcus aureus → Staphylococcus epidermidis	9 (5.263%) 1
Candida glabrata → Pseudomonas aeruginosa	10 (4.082%) 1
Enterobacter cloacae → Staphylococcus coagulase (-)	10 (1.020%) 1

9) Summary: We can first observe that the smaller number of windows tends to give sequential episodes with larger frequency, except for the sample of sputum and stool, of which number of windows is over 10000. Also we can observe that the occurrences of bacteria of Enterococcus faecalis (Group D) and Escherichia coli are characterized as the samples of pleural effusion, bile and pertoneal fluid. Furthermore, we can observe that the occurrences of bacteria of Streptococcus constellatus and Prevotella oralis/loescheii/denticola, Pseudomonas aeruginosa, and Candida albicans are characterized as the samples of joint fluid, sputum, and bile, respectively.

IV. Conclusion

In this paper, we have extracted sequential episodes representing the replacement of bacteria from bacterial culture data from Osaka Prefectural General Medical Center in years from 2000 to 2005.

In this paper, the sequential episodes for the samples of sputum and stool are valuable, because, while the number of windows is large, the extracted sequential episodes are frequent and have many different patients, as described in Section III-B and III-C. Hence, it is an important future work to analyze such episodes in more detail, by accessing other information in the original data in which the extracted episodes occur. It is also a future work to apply our episode mining techniques [4], [5], [6], [7] to the new data.

References

[1] K. Hirata, M. Harao, M. Wada, S. Ozaki, S. Yokoyama, K. Matsuoka: *Attribute selection measures with possibility and their application to classifying MRSA from MSSA*, in [10], 143–151.

[2] K. Hirata, Y. Shima, M. Harao, S. Yokoyama, K. Matsuoka, T. Izumi: *Disjunctive rules extracted from MRSA data with verification*, Proc. CME'05, 326–330, 2005.

[3] Y. Ikenaga, K. Hirata, M. Harao, S. Yokoyama, K. Matusoka: *Risk management system for hospital-acquired infection based on bacterial culture database*, Proc. CME'05, 423–427, 2005.

[4] T. Katoh, K. Hirata: *Mining frequent elliptic episodes from event sequence*, Proc. LLLL'07, 46–52, 2007.

[5] T. Katoh, K. Hirata, M. Harao: *Mining sectorial episodes from event sequences*, Proc. DS'06, LNAI **4265**, 137–145, 2006.

[6] T. Katoh, K. Hirata, M. Harao: *Mining frequent diamond episodes from event sequences*, Proc. MDAI'07, LNAI **4617**, 477–488, 2007.

[7] T. Katoh, K. Hirata, M. Harao, S. Yokoyama, K. Matsuoka: *Extraction of sectorial episodes representing changes for drug resistance and replacements of bacteria*, Proc. CME'07, 304–309, 2007.

[8] H. Mannila, H. Toivonen, A. I. Verkamo: *Discovery of frequent episodes in event sequences*, Data Mining and Knowledge Discovery **1**, 259–289, 1997.

[9] Y. Shima, K. Hirata, M. Harao, S. Yokoyama, K. Matsuoka, T. Izumi: *Extracting disjunctive closed rules from MRSA data*, Proc. CME'05, 321–325, 2005.

[10] J. L. Wu, K. Ito, S. Tobimatsu, T. Nishida, H. Fukuyama (eds.): *Complex medical engineering*, Springer, 2007.

[11] S. Yokoyama, K. Matsuoka, S. Tsumoto, M. Harao, T. Yamakawa, K. Sugahara, C. Nakahama, S. Ichiyama, K.Watanabe: *Study on the association between the patients' clinical background and the Anaerobes by data mining in infectious diseases database*, Biomedical Soft Computing and Human Sciences **7**, 69–75, 2001.