Proc. 2010 IEEE/ICME Int'l Conf.on Complex Medical Engineering (CME'10), 193-197, 2010

Aligned Bipartite Episodes between the Genera of Bacteria

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

***Abstract*—An *aligned bipartite episode between the genera of bacteria* is of the form $\mathcal{A} \to \mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ are the sets of genera of bacteria. This episode means that the earliest occurrence of each genus in $\mathcal{A}$ is precedent to the earliest occurrence of one in $\mathcal{B}$. In this paper, we extract such aligned bipartite episodes from bacterial culture data provided from Osaka Prefectural General Medical Center in years from 1999 to 2007.**

I. Introduction

In Complex Medical Engineering, *data mining* is one of the key techniques to analyze medical data from medical viewpoint [9]. In order to prevent hospital-acquired infection, in our previous works, we have extracted the time-related rules representing *replacements of bacteria* and *changes for drug resistance* as the factors of hospital-acquired infection, by paying our attention to *episode mining* from the bacterial culture data [1], [2], [3], [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.

In the previous work [4], as the simplest form of *serial episode* [8], we have introduced a *sequential episode* of the form $A \to B$ where A and B are event types of the occurrences of bacteria, and then extracted a sequential episode $A \to B$ *representing the replacements of bacteria* from bacterial culture data provided from Osaka Prefectural General Medical Center in years from 2000 to 2005.

On the other hand, the frequency of the sequential episodes representing the replacements of bacteria is too small to find some tendencies between the *genera* of bacteria in the previous work [4]. Hence, in order to find them, in this paper, we introduce an *aligned bipartite episode between the genera of bacteria* of the form of $\mathcal{A} \to \mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ are sets of event types representing the occurrences of the genera of bacteria. We can regard that such an aligned bipartite episode represents the replacement of the set of genera of bacteria from $\mathcal{A}$ to $\mathcal{B}$. Here, the word "aligned" means that the earliest time when each genus in $\mathcal{A}$ occurs is same and the earliest time when each genus in $\mathcal{B}$ occurs is also same. Then, we extract such episodes from bacterial culture data provided from Osaka Prefectural General Medical Center in years from 1999 to 2007.

*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

‡KD-ICONS Co., Ltd, Ohmori Minami 4-6-15, Ota, Tokyo 143-0013, 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. Aligned Bipartite Episodes between the Genera of Bacteria

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. In this paper, we regard bacterial culture data as an event sequence, and a detected bacterium as an event type.

In particular, we collect the *set $\mathcal{A}$ of genera of bacteria* such that every genus in $\mathcal{A}$ occurs at the same time t. Then, we regard $\mathcal{A}$ as an event type and $(\mathcal{A}, t)$ as an event.

An *aligned bipartite episode between the genera of bacteria* is of the form $\mathcal{A} \to \mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ are the sets of event types representing the occurrences of the genera of bacteria. In particular, we call an aligned bipartite episode $\mathcal{A} \to \mathcal{B}$ such that $\mathcal{A} \neq \mathcal{B}$ *proper*.

The word "aligned" in the aligned bipartite episode comes from the following *earliest occurrence*. Let $\mathcal{S}$ be an event sequence, s and t time stamps, w the width of windows and $\mathcal{A} \to \mathcal{B}$ an aligned bipartite episode between the genera of bacteria. Then, we say that $\mathcal{A} \to \mathcal{B}$ *occurs at* (s, t) *in* $\mathcal{S}$ *within* w if $(\mathcal{A}, s), (\mathcal{B}, t) \in \mathcal{S}$ and $0 < t - s \leq w$. Furthermore, we say that a pair (s, t) is the *earliest occurrence* of $\mathcal{A} \to \mathcal{B}$ if $\mathcal{A} \to \mathcal{B}$ occurs at (s, t) in $\mathcal{S}$ within w such that s and t are minimum. In this case, we say that an aligned bipartite episode $\mathcal{A} \to \mathcal{B}$ occurs *through the earliest occurrence*.

By using the following naïve algorithm, we extract the aligned bipartite episodes of the from $\mathcal{A} \to \mathcal{B}$ representing the replacements of genera of bacteria that every genus in $\mathcal{A}$ is precedent to one in $\mathcal{B}$ from the bacterial culture data.

1) For each of samples, construct an event sequence by fixing the sample and connecting data of every patient with the span of 15 days. Also collect the genera of bacteria stated as above.
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 the genera of bacteria

occur.

4) For every pair $(\mathcal{A}, \mathcal{B})$ of the genera of bacteria, compute the number of windows that an aligned bipartite episode $\mathcal{A} \to \mathcal{B}$ occurs through the earliest occurrence, and then compute the frequency of it.
5) Output the aligned bipartite episodes and their information of occurrences.

III. Experimental Results

In our previous work [4], we have adopted the bacterial culture data provided from Osaka Prefectural General Medical Center in years from 2000 to 2005, of which number of records is 35827. In this paper, we also adopt the updated data in years from 1999 to 2007, of which number of records is 164713.

Furthermore, in this paper, we use two kinds of data, one is called *type-1* data that a detected bacterium always exists, another is called *type-2* data that contains "none of detected bacteria," where we denote it by $\emptyset$. The number of type-1 and 2 data 54260 and 75137, respectively.

Figure 1 describes the number of different patient (in the left column) and the possible maximum frequency of episodes (in the right column) in type-1 and type-2 data for every sample.

sample	type-1		type-2	
catheter	44	6	56	8
pleural effusion	163	58	1091	302
blood	1328	322	5995	2305
indwelling vessel catheter	729	149	3115	1514
respiratory mucus	3796	631	7095	1328
internal organ	82	2	147	6
tissue	455	76	727	121
drain	621	148	1725	433
pus	3973	1000	6953	1663
heart pacer	15	1	239	13
lymph node	3	0	19	1
gastric fluid	27	3	59	5
joint fluid	42	12	320	58
breast milk	2	0	2	0
duodenal fluid	5	0	5	0
spinal fluid	78	15	1096	186
puncture fluid	123	25	295	55
stool	1140	248	3934	937
bile	161	56	233	71
intestinal fluid	12	0	20	1
urine	2174	620	4463	1407
attachment	6	2	10	2
peritoneal fluid	415	82	904	202
secretion	78	2	94	3
amnion liquid	113	0	155	0
sputum	4342	1996	7283	3300
perfusate	20	12	38	24
nasal cavity	212	10	286	15

Fig. 1. The number of different patients and the possible maximum frequency of episodes for every sample.

A. The number of extracted aligned bipartite episodes between the genera of bacteria

In the remainder of this section, we denote the number of aligned bipartite episodes between the genera of bacteria by #abe and the number of proper ones by #pabe. Then, Figure 2 describes #abe and #pabe for every sample.

sample	type-1		type-2	
	#abe	#pabe	#abe	#pabe
catheter	11	9	12	9
pleural effusion	139	125	229	215
blood	292	270	576	553
indwelling vessel catheter	119	109	262	249
respiratory mucus	562	523	881	740
internal organ	2	1	5	3
tissue	133	123	183	171
drain	284	271	405	390
pus	3214	3138	3917	3841
heart pacer	1	0	4	2
lymph node	0	0	1	0
gastric fluid	3	3	5	4
joint fluid	12	7	32	27
spinal fluid	8	5	31	27
puncture fluid	52	46	78	71
stool	163	148	310	293
bile	193	180	215	201
intestinal fluid	0	0	1	1
urine	1163	1109	1499	1444
attachment	2	1	2	1
peritoneal fluid	203	195	283	274
secretion	4	4	8	7
sputum	7636	7474	9543	9375
perfusate	30	24	61	54
nasal cavity	10	8	15	12

Fig. 2. The number of aligned bipartite and proper aligned bipartite episodes between the genera of bacteria.

B. The most top 5 aligned bipartite episodes between the genera of bacteria

In this subsection, we describe the extracted aligned bipartite episodes between genera of bacteria for the samples of which number of episodes is more than 100 (from type-1 data) in Figure 2.

1) Pleural effusion: For the sample of pleural effusion, #abe and #pabe for type-1 data are 139 and 125, and ones for type-2 data are 229 and 215, respectively. Then, the most 5 frequent aligned bipartite episodes between the genera of bacteria for type-1 and type-2 data are described as follows.

freq.	type-1 data
5	Staphylococcus → Staphylococcus Streptococcus
4	Streptococcus → Staphylococcus
3	Staphylococcus Streptococcus → Staphylococcus
3	Staphylococcus → Pseudomonas
2	Staphylococcus → Peptostreptococcus Staphylococcus
2	Peptostreptococcus Staphylococcus → Staphylococcus

freq.	type-2 data
25	∅ → Staphylococcus
18	Staphylococcus → ∅
9	Streptococcus → ∅
5	Staphylococcus → Staphylococcus Streptococcus
5	∅ → Staphylococcus Streptococcus

2) Blood: For the sample of blood, #abe and #pabe for type-1 data are 292 and 270, and ones for type-2 data are 576 and 553, respectively. Then, the most 5 frequent aligned bipartite episodes between the genera of bacteria for type-1 and type-2 data are described as follows.

freq.	type-1 data
14	Staphylococcus → Staphylococcus Streptococcus
12	Staphylococcus Streptococcus → Staphylococcus
10	Escherichia → Staphylococcus
8	Peptostreptococcus Staphylococcus → Staphylococcus
6	Staphylococcus → Escherichia
6	Enterococcus → Staphylococcus

freq.	type-2 data
274	Staphylococcus → ∅
261	∅ → Staphylococcus
49	Escherichia → ∅
36	Streptococcus → ∅
36	∅ → Escherichia

3) Indwelling vessel catheter: For the sample of indwelling vessel catheter, #abe and #pabe for type-1 data are 119 and 109, and ones for type-2 data are 262 and 249, respectively. Then, the most 5 frequent aligned bipartite episodes between the genera of bacteria for type-1 and type-2 data are described as follows.

freq.	type-1 data
8	Enterococcus → Staphylococcus
7	Staphylococcus → Enterococcus Staphylococcus
7	Staphylococcus → Enterococcus
7	Candida → Staphylococcus
5	Staphylococcus Streptococcus → Staphylococcus
5	Staphylococcus → Pseudomonas
5	Staphylococcus → Candida

freq.	type-2 data
182	∅ → Staphylococcus
162	Staphylococcus → ∅
24	Enterococcus → ∅
22	∅ → Candida
20	∅ → Enterococcus

4) Respiratory mucus: For the sample of respiratory mucus, #abe and #pabe for type-1 data are 562 and 523, and ones for type-2 data are 881 and 740, respectively. Then, the most 5 frequent aligned bipartite episodes between the genera of bacteria for type-1 and type-2 data are described as follows.

freq.	type-1 data
93	MRSA screening → Staphylococcus
23	Staphylococcus → MRSA screening
21	Staphylococcus → Staphylococcus yeast
18	Staphylococcus → Pseudomonas Staphylococcus
15	Staphylococcus → yeast

freq.	type-2 data
243	Staphylococcus → ∅
162	∅ → Staphylococcus
148	MRSA screening → ∅
98	∅ → ∅ Staphylococcus
74	MRSA screening → Staphylococcus

The reason why the frequency of the aligned bipartite episode "MRSA screening → Staphylococcus" between type-1 data and type-2 data changes is the occurrences of the following episodes containing ∅ in type-2 data; The part of the following episodes are extracted as the episodes from type-1 data.

freq.	type-2 data
26	MRSA screening → ∅ Staphylococcus

5) Tissue: For the sample of tissue, #abe and #pabe for type-1 data are 133 and 123, and ones for type-2 data are 262 and 249, respectively. Then, the most 5 frequent aligned bipartite episodes between the genera of bacteria for type-1 and type-2 data are described as follows.

freq.	type-1 data
4	Staphylococcus → Enterococcus Staphylococcus
3	Enterococcus Staphylococcus → Staphylococcus
2	Staphylococcus Streptococcus → Staphylococcus
2	Pseudomonas Staphylococcus Streptococcus → Staphylococcus
2	Pseudomonas Staphylococcus → Staphylococcus
2	Escherichia Staphylococcus → Staphylococcus
2	Enterococcus Staphylococcus Streptococcus → Staphylococcus

freq.	type-2 data
18	Staphylococcus → ∅
10	∅ → Staphylococcus
5	Candida → ∅
3	Staphylococcus → Enterococcus Staphylococcus
3	∅ → Candida

The reason why the frequency of the aligned bipartite episode "Staphylococcus → Enterococcus, Staphylococcus" between type-1 data and type-2 data changes is the occurrences of the following episodes containing ∅ in type-2 data; The following episode is extracted as the episode from type-1 data.

freq.	type-2 data
1	Staphylococcus → ∅ Enterococcus Staphylococcus

6) Drain: For the sample of drain, #abe and #pabe for type-1 data are 284 and 271, and ones for type-2 data are 405 and 390, respectively. Then, the most 5 frequent aligned bipartite episodes between the genera of bacteria for type-1 and type-2 data are described as follows.

freq.	type-1 data
3	Pseudomonas → Staphylococcus
3	Enterococcus Staphylococcus → Staphylococcus
3	Candida Staphylococcus → Candida
3	Candida → Staphylococcus
2	Staphylococcus → Staphylococcus Streptococcus
2	Staphylococcus → Pseudomonas
2	Staphylococcus → Prevotella Staphylococcus
2	Peptostreptococcus Staphylococcus → Staphylococcus
2	Enterococcus → Staphylococcus
2	Enterococcus → Enterococcus Staphylococcus
2	Enterococcus → Candida
2	Enterobacter → Candida Enterobacter
2	Candida Staphylococcus Streptococcus → Staphylococcus
2	Candida Staphylococcus → Staphylococcus
2	Candida → Candida Enterococcus

freq.	type-2 data
54	∅ → Staphylococcus
31	Staphylococcus → ∅
6	∅ Staphylococcus → ∅
6	∅ → Streptococcus
6	∅ → Enterococcus

7) Pus: For the sample of pus, #abe and #pabe for type-1 data are 3214 and 3138, and ones for type-2 data are 3917 and 3841, respectively. Then, the most 5 frequent aligned bipartite episodes between the genera of bacteria for type-1 and type-2 data are described as follows.

freq.	type-1 data
33	Staphylococcus → Pseudomonas Staphylococcus
33	Pseudomonas Staphylococcus → Staphylococcus
32	Enterococcus → Staphylococcus
31	Staphylococcus → Pseudomonas
27	Pseudomonas → Staphylococcus

freq.	type-2 data
220	Staphylococcus → ∅
162	∅ → Staphylococcus
52	Streptococcus → ∅
40	∅ → Streptococcus
35	Pseudomonas → ∅

8) Stool: For the sample of stool #abe and #pabe for type-1 data are 163 and 148, and ones for type-2 data are 310 and 293, respectively. Then, the most 5 frequent aligned bipartite episodes between the genera of bacteria for type-1 and type-2 data are described as follows.

freq.	type-1 data
19	Staphylococcus → yeast
14	Staphylococcus → yeast yeast
13	Staphylococcus → Staphylococcus yeast
12	yeast → Staphylococcus
11	Staphylococcus → Staphylococcus yeast

freq.	type-2 data
138	Staphylococcus → ∅
95	∅ → Staphylococcus
47	yeast → ∅
42	Salmonella → ∅
38	∅ → yeast

9) Bile: For the sample of bile, #abe and #pabe for type-1 data are 193 and 180, and ones for type-2 data are 215 and 201, respectively. Then, the most frequent aligned bipartite episodes between the genera of bacteria for type-1 and type-2 data are described as follows. Here, the frequency of the other aligned bipartite episode is just 1.

freq.	type-1 data
2	Escherichia → Enterococcus Escherichia

freq.	type-2 data
2	Klebsiella → ∅
2	Escherichia → Enterococcus Escherichia
2	∅ → Enterococcus

10) Urine: For the sample of urine, #abe and #pabe for type-1 data are 1163 and 1109, and ones for type-2 data are 1499 and 1444, respectively. Then, the most 5 frequent aligned bipartite episodes between the genera of bacteria for type-1 and type-2 data are described as follows.

freq.	type-1 data
20	yeast → Candida
13	Candida → yeast
13	Candida → Escherichia
12	Staphylococcus → Pseudomonas
10	Escherichia → yeast

freq.	type-2 data
91	∅ → Escherichia
83	∅ → Candida
80	Escherichia → ∅
76	Staphylococcus → ∅
71	∅ → Staphylococcus

11) Peritoneal fluid: For the sample of peritoneal fluid, #abe and #pabe for type-1 data are 203 and 195, and ones for type-2 data are 283 and 274, respectively. Then, the most 5 frequent aligned bipartite episodes between the genera of bacteria for type-1 and type-2 data are described as follows. Here, the frequency of the other aligned bipartite episode for type-1 data is just 1.

freq.	type-1 data
2	Candida → Streptococcus

freq.	type-2 data
15	∅ → Staphylococcus
9	Staphylococcus → ∅
5	Candida → ∅
5	∅ → Candida
4	Streptococcus → ∅

12) Sputum: For the sample of sputum, #abe and #pabe for type-1 data are 7636 and 7474, and ones for type-2 data are 9543 and 9375, respectively. Then, the most 5 frequent aligned bipartite episodes between the genera of bacteria for type-1 and type-2 data are described as follows.

freq.	type-1 data
108	Staphylococcus → Pseudomonas Staphylococcus
103	Staphylococcus → Staphylococcus yeast
92	Staphylococcus → Staphylococcus yeast
92	Staphylococcus → Pseudomonas
90	Pseudomonas → Staphylococcus

freq.	type-2 data
396	∅ → Staphylococcus
325	Staphylococcus → ∅
169	∅ → Pseudomonas
136	Pseudomonas → ∅
128	∅ → Pseudomonas Staphylococcus

13) Summary: For the samples except bile and urine, the episodes of both "Staphylococcus → ∅" and "∅ →Staphylococcus" are the most 2 frequent episodes than others for type-2 data. On the other hand, the episodes of "Klebsiella → ∅," "Escherichia → Enterococcus, Escherichia" and "∅ → Enterococcus" are the most frequent for the sample of bile and the episodes of "∅ → Escherichia," "∅ → Candida" and "Escherichia → ∅" are the most 3 frequent for the sample of urine.

By pay our attention to the genera of bacteria, the genus of Pseudomonas occurs in the episodes for the samples except blood, stool, bile and peritoneal fluid. Also the genera of (1) Streptococcus, (2) Enterococcus, (3) Escherichia and (4) Peptostreptococcus occur in the episodes for the samples of (1) pleural effusion, blood, indwelling vessel catheter, tissue and drain, (2) indwelling vessel catheter, tissue, drain, pus and bile, (3) blood, tissue, bile and urine, and (4) pleural effusion and blood, respectively.

On the other hand, the yeast occurs in the episodes for the samples of stool, urine and sputum. In particular, for the sample of stool, the most 5 frequent episodes for type-1 data always contain the yeast.

Also Candida occurs in the episodes for the samples of indwelling vessel catheter, drain, urine and peritoneal fluid. In particular, for the sample of urine, the episodes both "yeast → Candida" and "Candida → yeast" occur uniquely.

Finally, the episode containing MRSA screening occurs uniquely for the sample for respiratory mucus, the episode of "Salmonella→ ∅" occurs uniquely for the sample of stool,

IV. Conclusion

In this paper, we have extracted the *aligned bipartite episodes between the genera of bacteria* from bacterial culture data provided from Osaka Prefecture General Medical Center in years from 1999 to 2007.

It is an important future work to analyze such episodes in more detail from medical viewpoint, by accessing other information in the original data in which the extracted episodes occur, together with antibiotics information. It is also a future work to extend aligned bipartite episodes to bipartite episodes [2] between the genera of bacteria, by extending the earliest occurrence of the genera to an arbitrary one.

References

[1] T. Katoh, H. Arimura, K. Hirata: *An efficient depth-first search algorithm for extracting frequent diamond episodes from event sequences*, IPSJ Transactions on Databases (to appear).

[2] T. Katoh, H. Arimura, K. Hirata: *Mining frequent bipartite episodes from event sequences*, Proc. DS'09, LNAI **5808**, 136-151, 2009.

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

[4] T. Katoh, K. Hirata, H. Arimura, S. Yokoyama, K. Matsuoka: *Extracting sequential episodes representing replacements of bacteria from bacterial culture data*, Proc. CME'09, 4 pages, 2009.

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