肝炎データからのB-GBI法による時間変化を重視したパターン抽出

- 肝炎データからのパターン抽出
  - 前処理
  - 4回のサイクル
- 今後の計画
- 議論

大阪大学産業科学研究所
松田　喬，吉田　哲也，元田　浩，鷲尾　隆

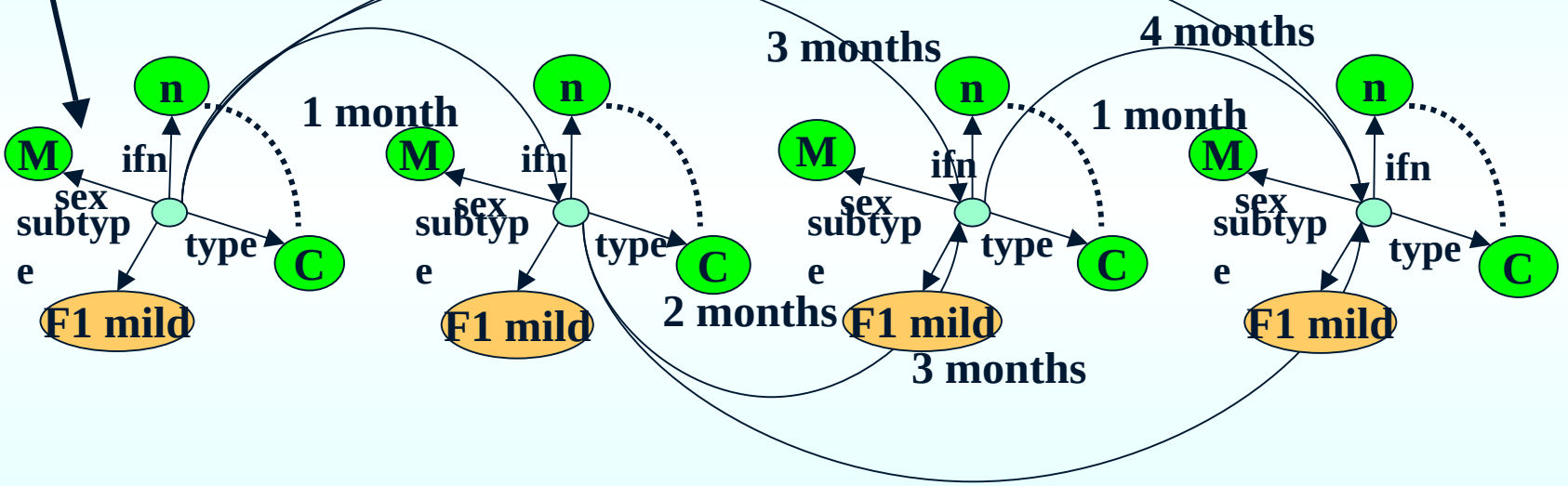
肝炎データからパターン抽出

- 前処理
  - データ洗浄
  - 表形式への変換
  - 離散化
  - グラフ構造データへの変換
- ４回のマイニングサイクル
  - 繊維化，活動性
  - **B**型ウィルスの活動性
  - 時間変化の影響の範囲
  - 時間変化の重み付け

グラフ構造データへの変換

mid,	date,	sex,	ifn,	A2PI,	……,	type,	subtype,	activity,
1,	19810428,	M,	n,	?,	……,	C,	F1 mild,	A1
1,	19820525,	M,	n,	?,	……,	C,	F1 mild,	A1
1,	19810722,	M,	n,	?,	……,	C,	F1 mild,	A1
1,	19811025,	M,	n,	?,	……,	C,	F1 mild,	A1
2,	19900324,	F,	n,	?,	……,	B,	CAH2B,	A2
2,	19900425,	F,	n,	?,	……,	B,	CAH2B,	A2

一定期間の平均値
(e.g., 1ヶ月)



- 各患者ごとに1つのグラフを作成
  - リンクラベル：各検査項目
  - ノードラベル：リンクラベルの検査項目に対する検査結果
  - ダミーノード：一定期間の検査結果
  - 時系列リンク：ダミーノードを

パターン抽出における評価関数

- チャンキングへの評価関数
  - 頻度（閾値 0.3）
    - 閾値以上の数のグラフに含まれ，最も頻度の多いペアから
- パターンの評価関数
  - 規格化した確率（閾値 0.3）

$$Max.\left\{ \frac{\frac{n_{C_1}}{N_{C_1}}}{\sum_{k=1}^{k=K} \frac{n_{C_k}}{N_{C_k}}}, \frac{\frac{n_{C_2}}{N_{C_2}}}{\sum_{k=1}^{k=K} \frac{n_{C_k}}{N_{C_k}}}, \cdots, \frac{\frac{n_{C_k}}{N_{C_k}}}{\sum_{k=1}^{k=K} \frac{n_{C_k}}{N_{C_k}}}, \right\}$$

***nCi*:評価するペアが含まれているクラスC*i*のグラフ数**
***NCi*:全グラフにおけるクラスC*i*のグラフ数**

- ビーム幅：3

第1回

- 前処理
  - 1ヶ月平均
  - 時系列リンク
    10年まで
- クラス
  - 活動性
  - 繊維化
- パターン抽出
  - 大量のパターン
  - 計算時間： **16時間**

活動性

クラス	A1	A2	A3	All
グラフ数	51	54	11	116
ノード数の平均値	1975	1758	1776	1855
ノード数の最大値	6723	6688	4572	6723
ノード数の最小値	63	53	214	53

繊維化

クラス	F0	F1	F2	F3	F3-4	F4
グラフ数	1	49	33	26	1	12
ノード数の平均値	908	1637	2306	1469	1155	2677
ノード数の最大値	908	6907	6880	3895	1155	5511
ノード数の最小値	908	86	61	54	1155	397

活動性

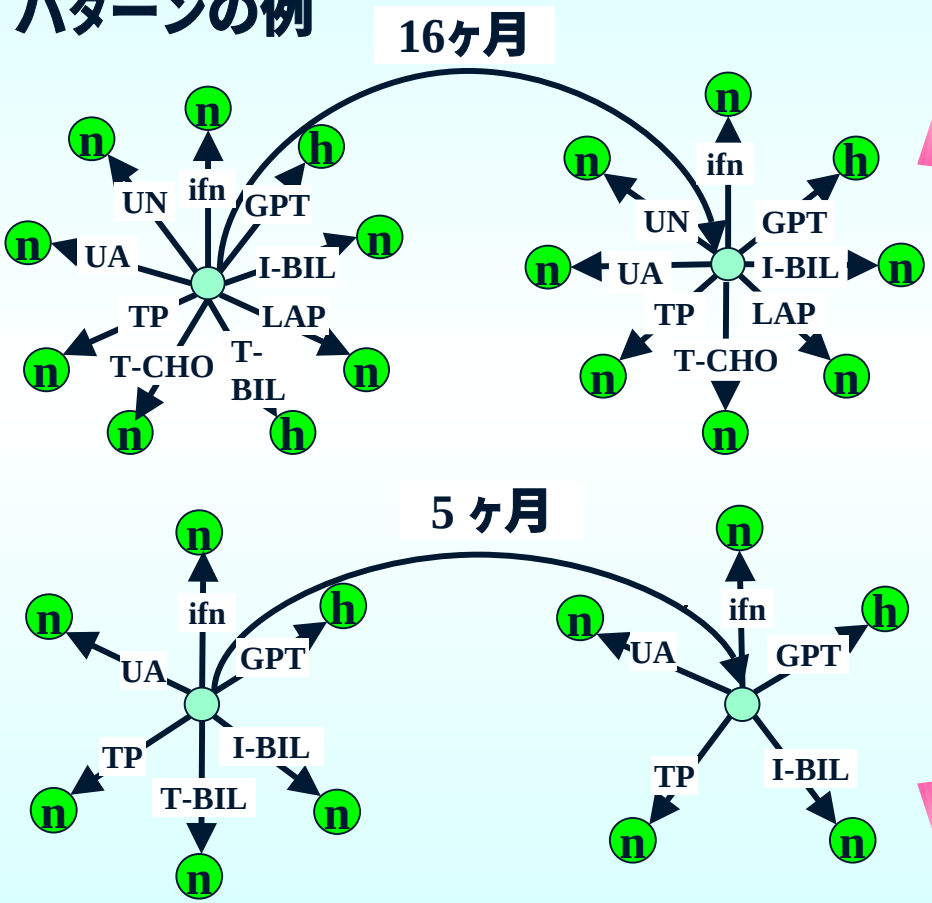
閾値	パターン数
0.4	192826
0.5	172437
0.6	140991
0.7	117628
0.8	105426

繊維化

閾値	パターン数
0.4	488172
0.5	433011
0.6	344322
0.7	301911
0.8	276073

活動性 (A_1, A_2, A_3)に特徴的なパターンの例

ウィルスの活動性（A_3）に特徴的パターンの例



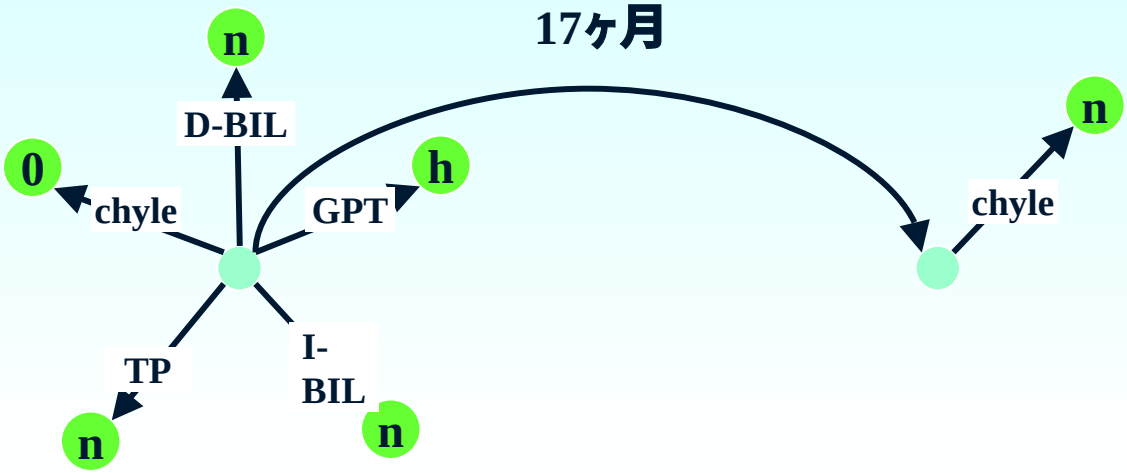
Evaluation = 0.905
A1 1
A2 1
A3 4

専門家のコメント

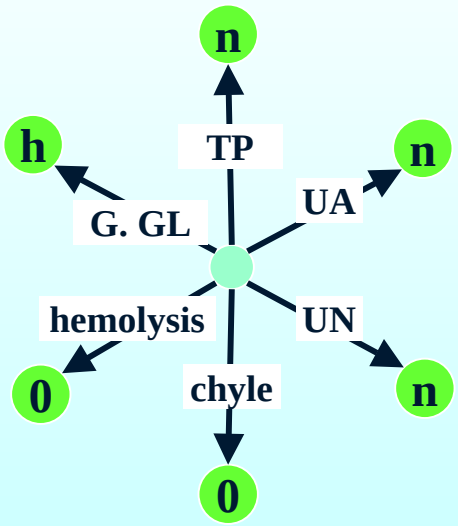
UN, UAといったデータが正常値を示すということは肝疾患と腎疾患には強い関連性がないという，既知の医学知識に合致する．また，肝硬変に至っていない患者ではTPが正常を示すことも知られている．

Evaluation = 0.827
A1 2
A2 2
A3 4

繊維化　(F_1, F_2, F_3, F_4)に特徴的なパターン



Evaluation = 0.706
F1　9
F2　3
F3　1
F4　9



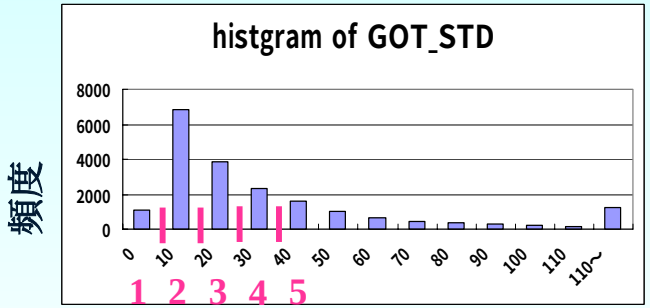
Evaluation = 0.730
F1　2
F2　3
F3　3
F4　8

第2回

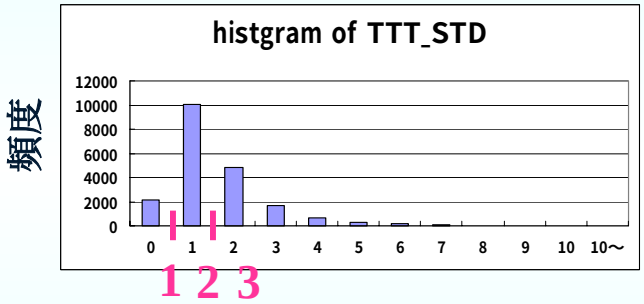
- 前回の成果
  - 妥当なパターンの抽出
- 課題
  - 既存，当たり前?
  - 計算時間
    - インタラクティブな解析が困難
- 専門家との共同作業
  - 6月，7月
  - 属性選択
  - 属性構築
- グラフサイズの抑制
  - 属性選択
  - 平均化処理

短期変動への新指標（属性）の追加

- 短期変動が重要なGOT, GPT, TTT, ZTT
  - 6ヶ月間の標準偏差
- ヒストグラムに基づき離散化
  - GOT, GPT: 5値
  - TTT,ZTT ：3値



GOT_STD



TTT_STD

mid,	date,	……,	GOT	GPT	……,	GOT_STD	GPT_STD	……,
1,	19810428,	……,	54	108	……,			……,
1,	19810722,	……,	63	112	……,	7.97	15.71	……,
1,	19811025,	……,	71	318	……,	25.45	26.96	……,
1,	19820125,	……,	97	144	……,	37.91	30.5	……,
…….	……,	……,	……	……	……,	……,	……,	……,
2,	19900324,	……,	33	65	……,			……,
2,	19900425,	……,	72	80	……,	5.66	13.88	……,
2,	19900921,	……,	47	118	……,	8.58	22.63	……,

B型肝炎ウィルスの状態に特徴的なパターン抽出

- 既存の4属性の組合わせとして専門家が定義
  - HBV : active, inactive, cured
- 属性選択 → 23属性に絞込み
- 6ヶ月平均

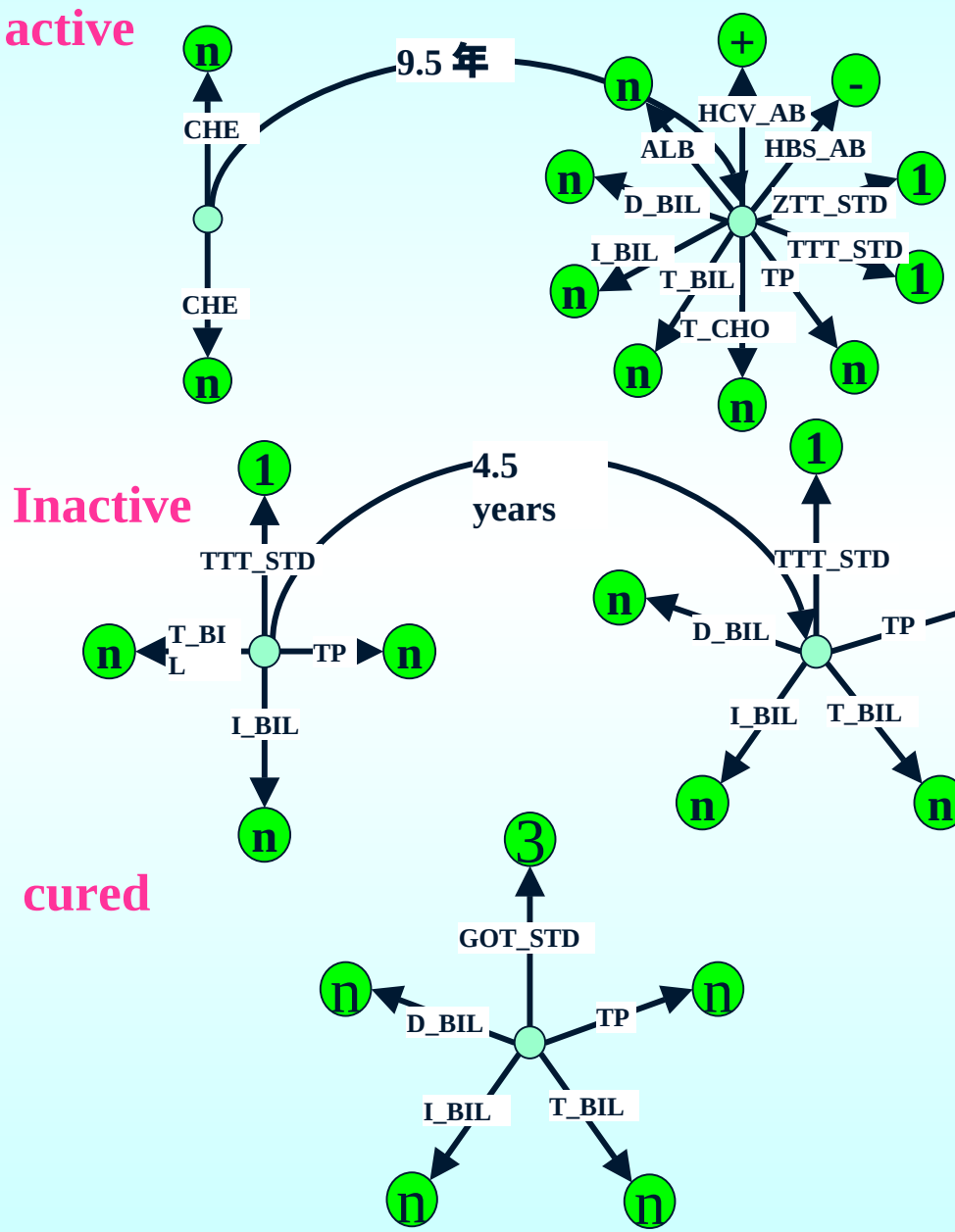
- 時系列リンク:10年まで
- クラス：HBV

B型肝炎ウィルスの状態に対する判定ルー

HBS-AG	HBE-AG	HBE-AB	HBS-AB	HBV
+	+	-	-	active
+	-	+	-	inactive
-	-	+	+	cured

クラス	active	inactive	cured
グラフ数	24	49	25
ノード数の平均値	351	474	409
ノード数の最大値	703	788	717
ノード数の最小値	17	32	135

B型肝炎ウィルスの状態に特徴的なパター



専門家のコメント

9.5年も離れた期間で共起性があるということは専門知識からは**解釈不能**.

ウィルスの活動性が低い場合にはTTTの変化が少ない, という既知**医学知識に合致**する.

GOTの変化が大きくてもウィルスの動性が低い場合がある, ということを示唆するため**意外性**がある.

第3回

- 前回の成果
  - 妥当なパターン
  - 解釈不能なパターン
  - 意外なパターン
- 課題
  - 長期間離れた場合での共起性　は解釈が困難
- 時系列の影響範囲を制限
  - 時系列リンクの張り方：2年までに

時系列リンクを2年までに制限した場合

10年ま

閾値	パターン数
0.4	67239
0.5	58868
0.6	48849
0.7	39579
0.8	37743

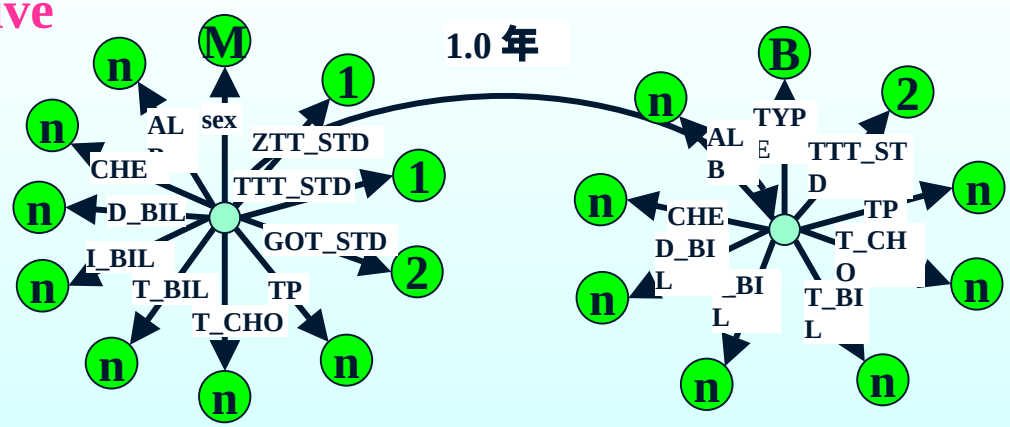
2年ま

閾値	パターン数
0.4	17505
0.5	14641
0.6	11567
0.7	9016
0.8	8512

専門家のコメント

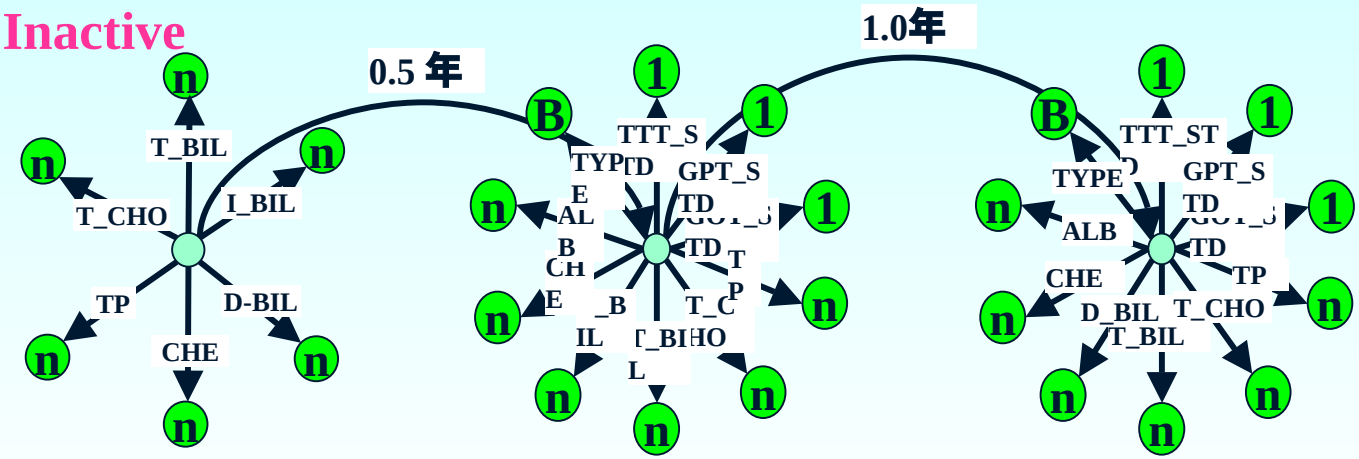
GOTとTTTのSTDがGrade2ということで、かなり臨床的には、うなずける結果である。この形で精度を高めれば、いわゆるエキスパートシステム等に応用できるルールになりうると思われる

active



リンクを2年までに制限した場

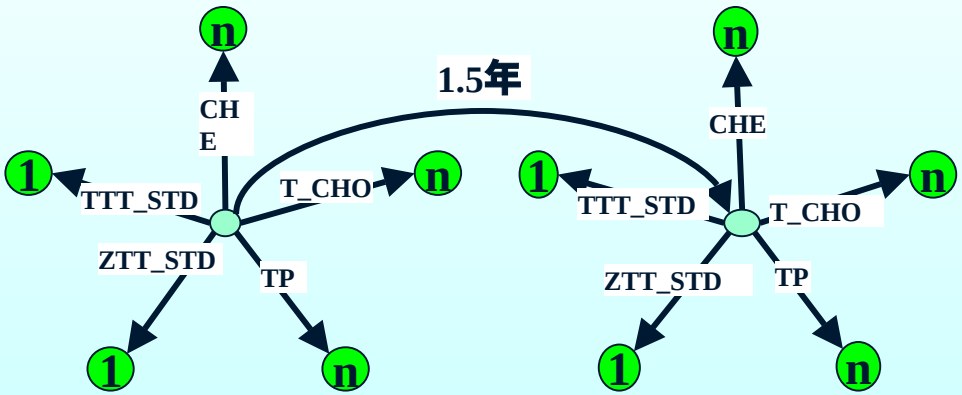
Inactive



専門家のコメント

一年の間隔について TTT,GOT,GPTのSTDがgrade1であるというルールは非常にリーズナブルで臨床家の持っているイメージに合致し，**妥当な結果**である．

cured



TTT,ZTTについてSTDがgrade1ということで治癒後のデータであるとすれば**妥当**である．

第4回

- 前回の成果
  - 時系列の影響範囲を考慮したパターン
- 課題
  - 妥当なパターンが多い
  - 時系列的な推移が少ない
    - 2、3ステップに留まる
- 時系列パターンへのバイアス
  - チャンクの重み付けを変更
- クラス：繊維化
  - **HBV(2,3回目のクラス): 属性の1つとして使用**

チャンキングにおける重み付け

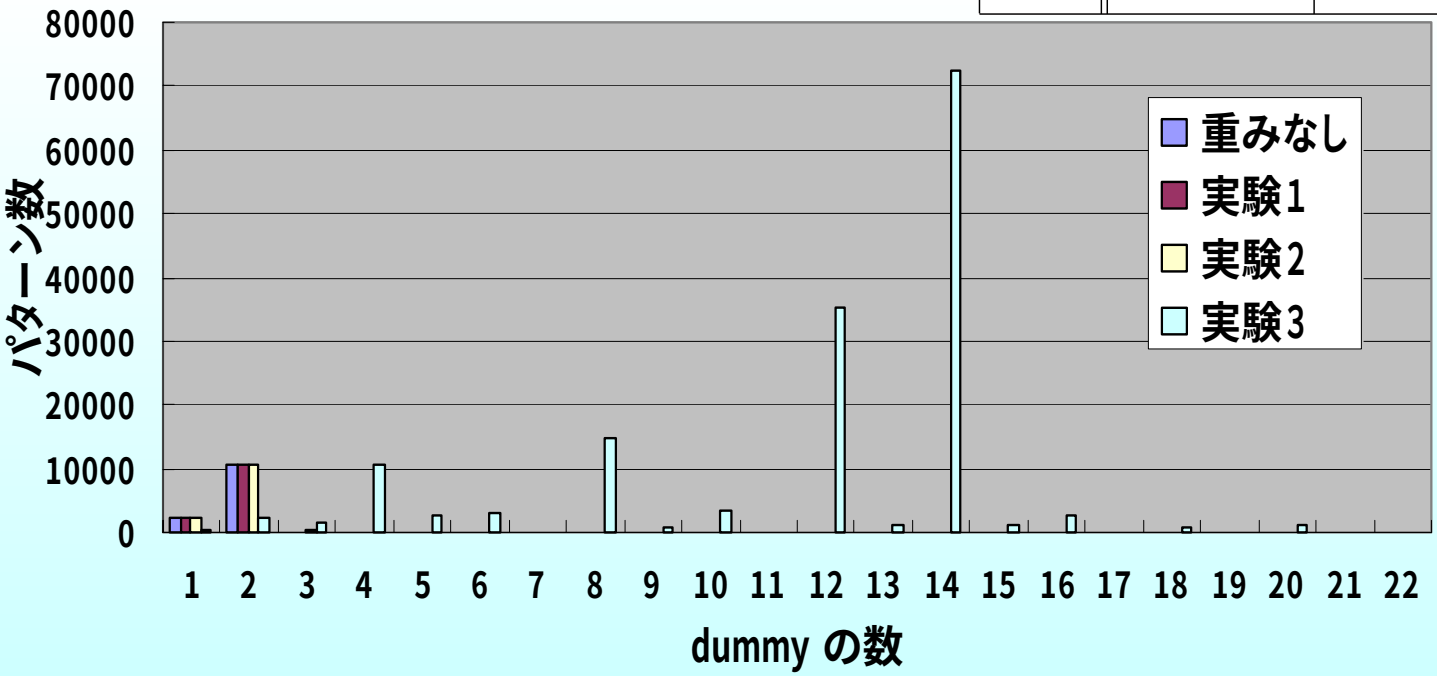
通常のペア　1

時系列リンクを含むペア　1＋(時系列リンク数＊α)

	α
重みなし	0
実験1	+0.05
実験2	+0.1
実験3	+0.2

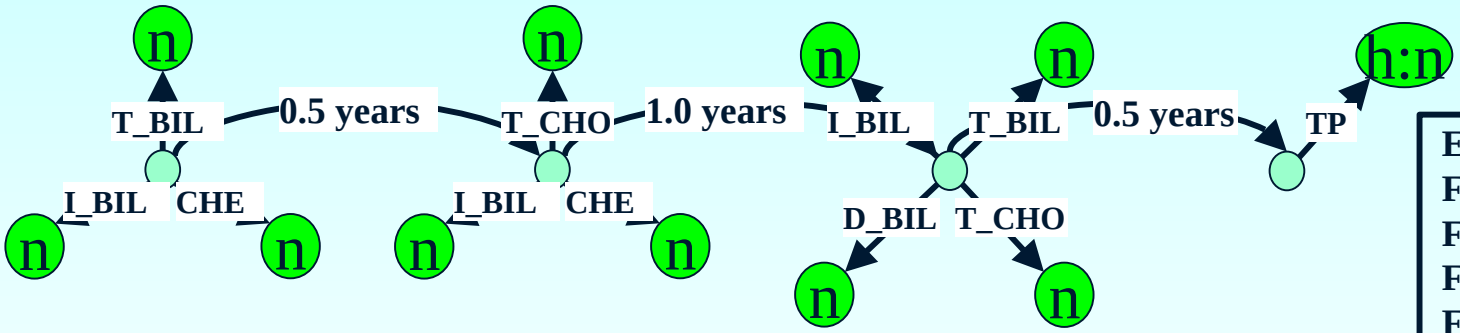
クラス	F0	F1	F2	F3	F4
グラフ数	11	217	107	78	83
ノード数の平均値	322	301	271	262	277
ノード数の最大値	632	751	774	646	788
ノード数の最小値	80	16	16	15	16

閾値	重みなし	実験１	実験２	実験３
0.4	8883	8883	9411	117586
0.5	6328	6328	6580	87388
0.6	4248	4248	4554	62840
0.7	3244	3244	3577	52391
0.8	2566	2566	2780	44769



パターンが含まれるグラフ数の平均

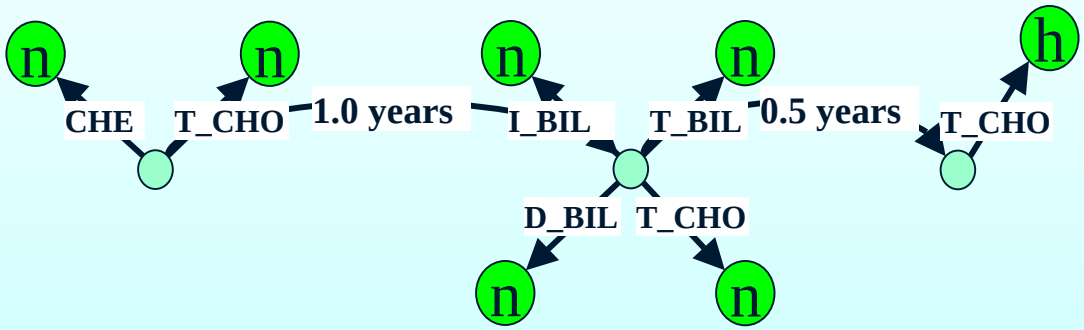
	平均
重みなし	20
実験1	20
実験2	20
実験3	12



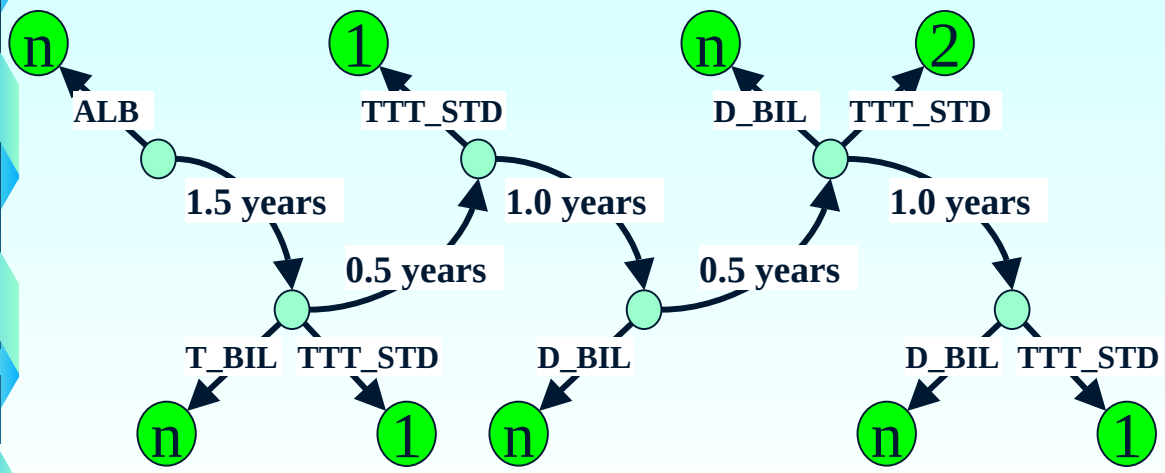
Evaluation = 1.0
F0 0
F1 5
F2 0
F3 0
F4 0

専門家のコメント

時系列最後のパターンがhighに絡んでいるのが意外である．上の方はTP:(h:n)、下の方はT-CHO:hとなっている．これらがそれぞれ，F1,F2の特徴として挙られるのかが一つの論点だと思う．



Evaluation = 0.89
F0 0
F1 1
F2 4
F3 0
F4 0



Evaluation = 1.0
F0 0
F1 0
F2 5
F3 0
F4 0

時系列リンクの重みを変えた場合に対する専門家のコメント：

TTTのパターンが浮き彫りになっているのだと思う．

他のF stageではどのようなTTTのパターンが出るのか興味がある．

仮説：「TTTやGOTなどの動きが大きい方が、悪いステージ(大きいF stage)につながっている」と言えないか．

今後の計画

- グラフ構造の類似性
  - 大量の抽出パターン
  - **TFS等の検討　←高橋先生，岡田先生**
  - 課題：
    - 類似性の定義**?**
      - フーリエ展開のアイデア
- 属性選択　- 前処理として -
  - **GOT, GPT, TTT, ZTT**の絶対値の併用
  - インターフェロン投与の扱い
- 属性構築
  - **GBI**の再帰呼び出し

議論

- 7つの目標のどこを狙うか?
  1. **病理像と血液検査データとの相関性** ← 4
  2. **肝炎の病理像（繊維化の程度）と発ガンまでの期**
  3. **血液データと発ガンまでの期間**
  4. **時系列に関する血液データ積算の有用性**
  5. **B型肝炎とC型肝炎の経過の違い**
  6. **INF治療の有用性**
  7. **GOT,GPTがは「進行速度」の指標か**
- 時系列の共起パターン抽出　が適する課題?
  - グラフ構造への変換方法
  - チャンキングの行い方
    - チャンクの重み付け 以外の方法?

まとめ

- 肝炎データからのパターン抽出
  - 前処理
  - 4回のサイクル
- 今後の計画
  - グラフ構造の類似性
  - 属性選択・属性構築
  - 時系列の共起パターン抽出　が適する課題?