第11回バイオインフォマティクス研究会

Dynamic Time Warping (DTW)を用いた時系列データからのパターン抽出法

2014/04/26

バイオインフォマティクス解析室

中林潤

トピック

- 時系列データを扱う際の一般的な問題点
- 類似パターン抽出法

  Dynamic Time Warping　(DTW)
- 実例

  遺伝子発現の時間発展パターンの抽出

時系列データ

- 複数のタイムポイントで取得されたデータの系列

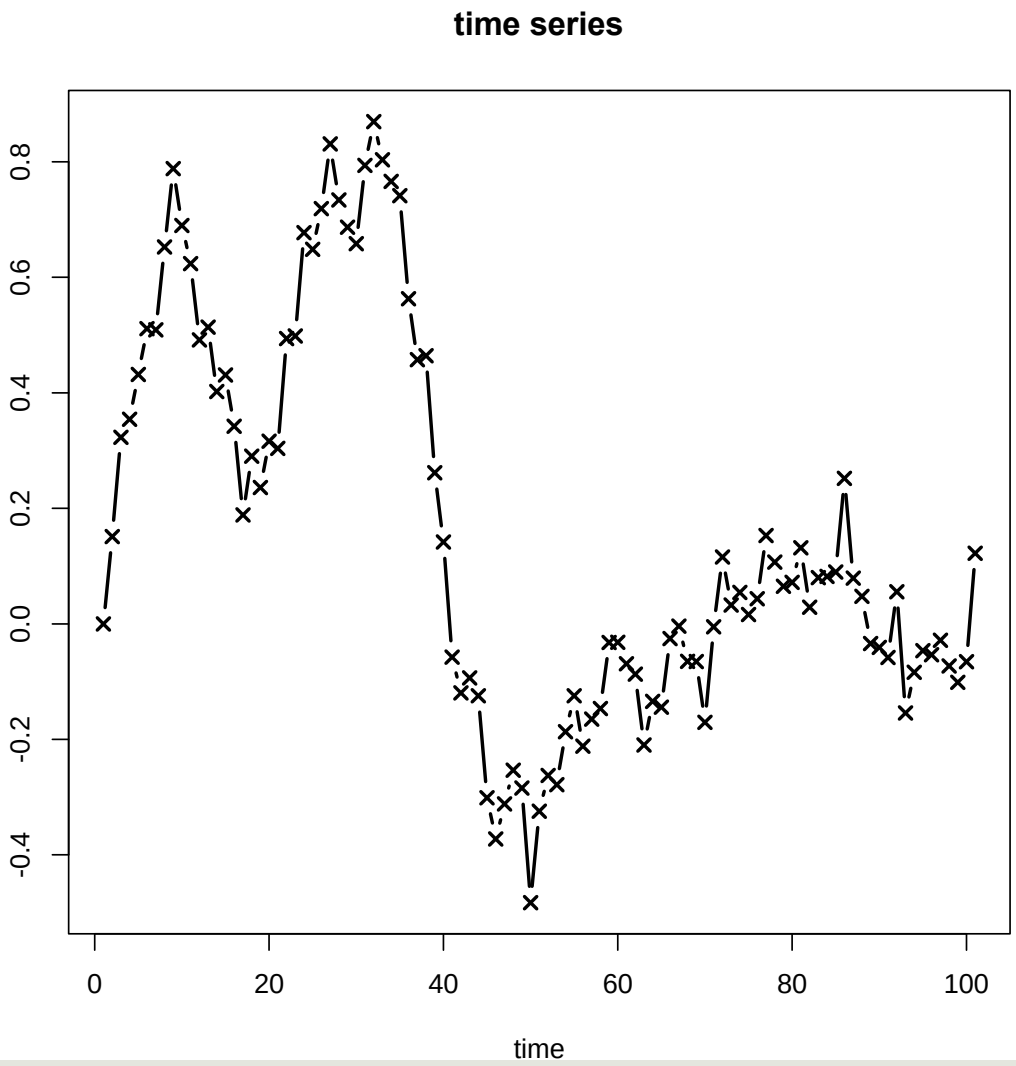
- 時系列データの問題点

  偽相関

  見せかけの回帰

  …

時系列データの例



偽相関

- 媒介変数を介して本来は無関係な2変数の間に相関があるかのように見える

- 地域ごとの小学校数と歯科医数

  住民数が媒介変数

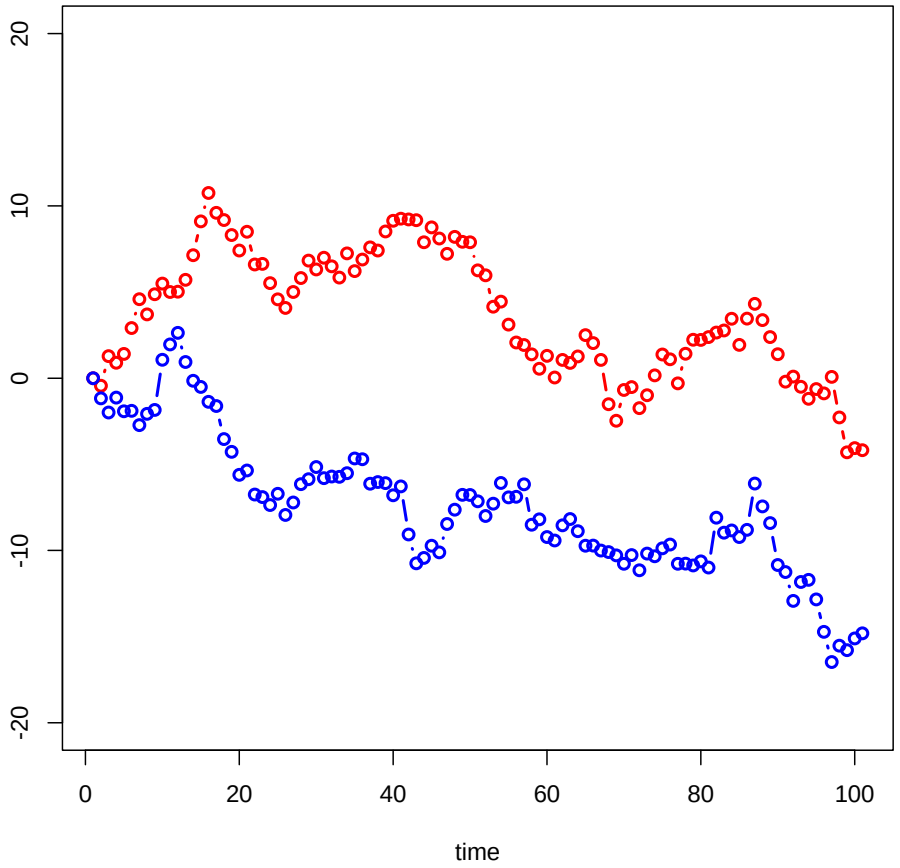
見せかけの回帰

- 原系列が非定常過程、差分が定常過程（タイムポイント間の変動は時間に非依存）の2変数間で回帰分析を行ったとき、実際よりも高い説明力が生じる現象

  原系列は非定常過程：実測されたデータは時間と伴に変化する

  差分は定常過程　　：データ間の差は時間に非依存

見せかけの回帰



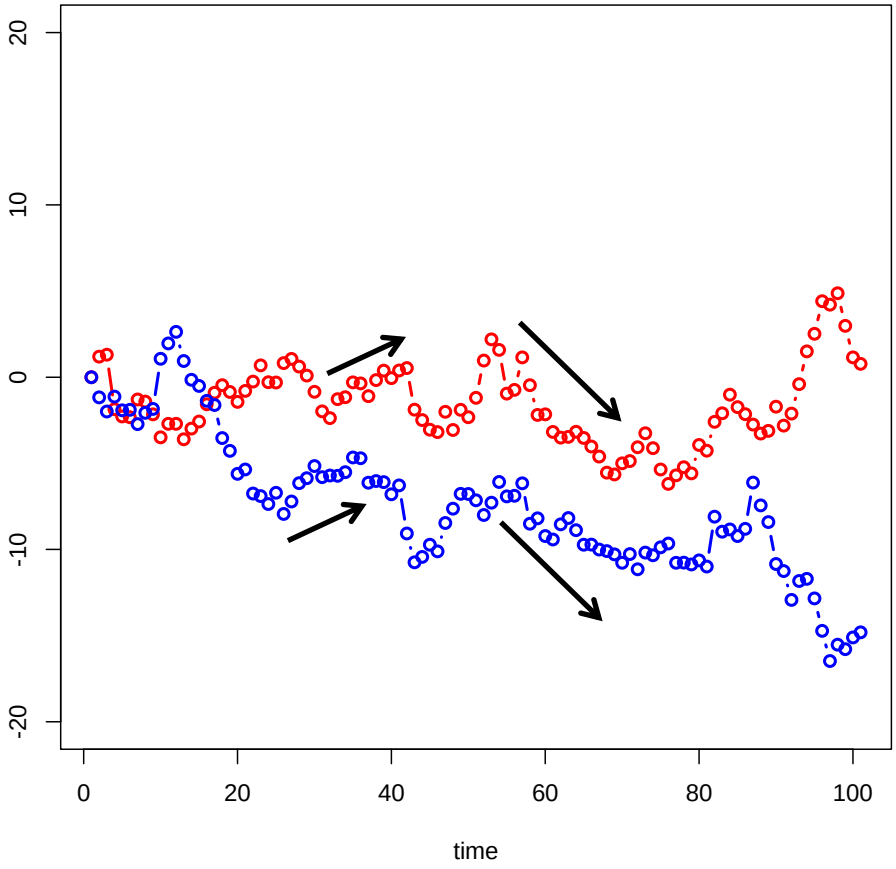
ある時刻における値に乱数を加えて時系列データを作成

$$x(t) - x(t-1) \rightarrow \text{random}$$

$$y(t) - y(t-1) \rightarrow \text{random}$$

見せかけの回帰

直観的理解



ある時刻における値に乱数を加えて時系列データを作成

$$x(t) - x(t-1) \rightarrow \text{random}$$

$$y(t) - y(t-1) \rightarrow \text{random}$$

時系列データからのパターン抽出

- 時系列データにおける類似度（距離）

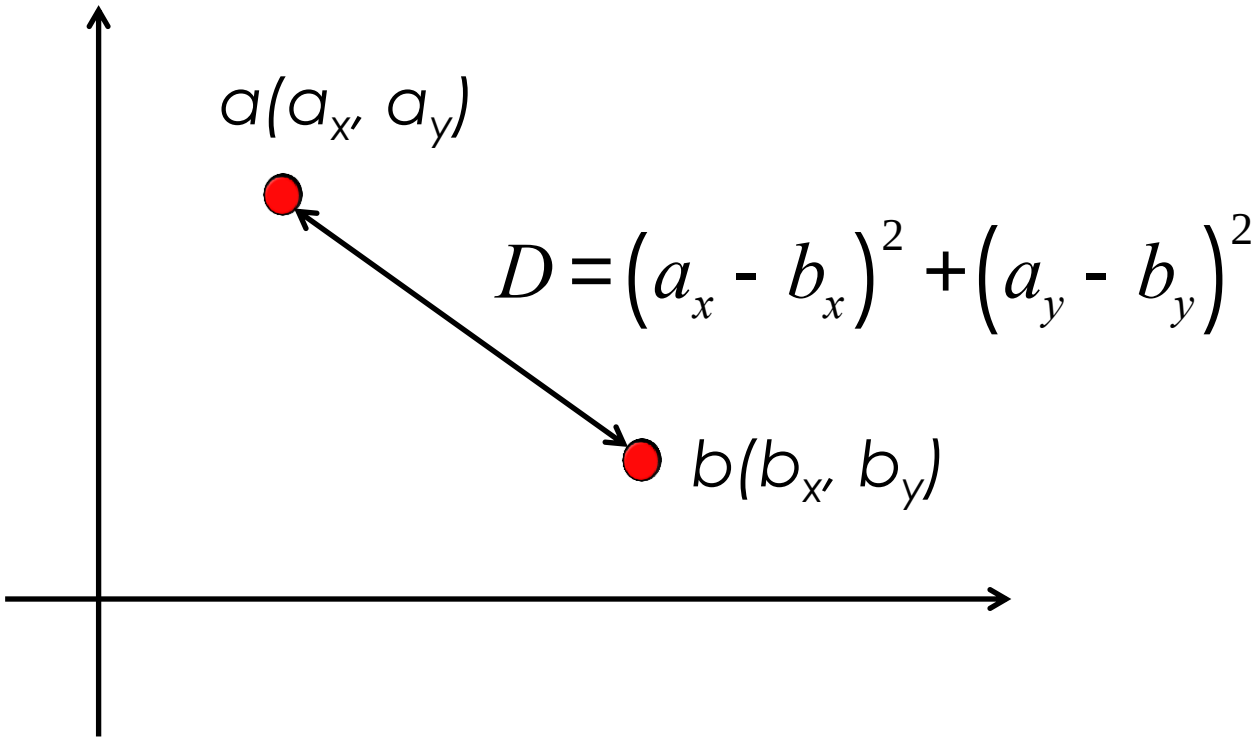
  ユークリッド距離ではパターン抽出

  が困難

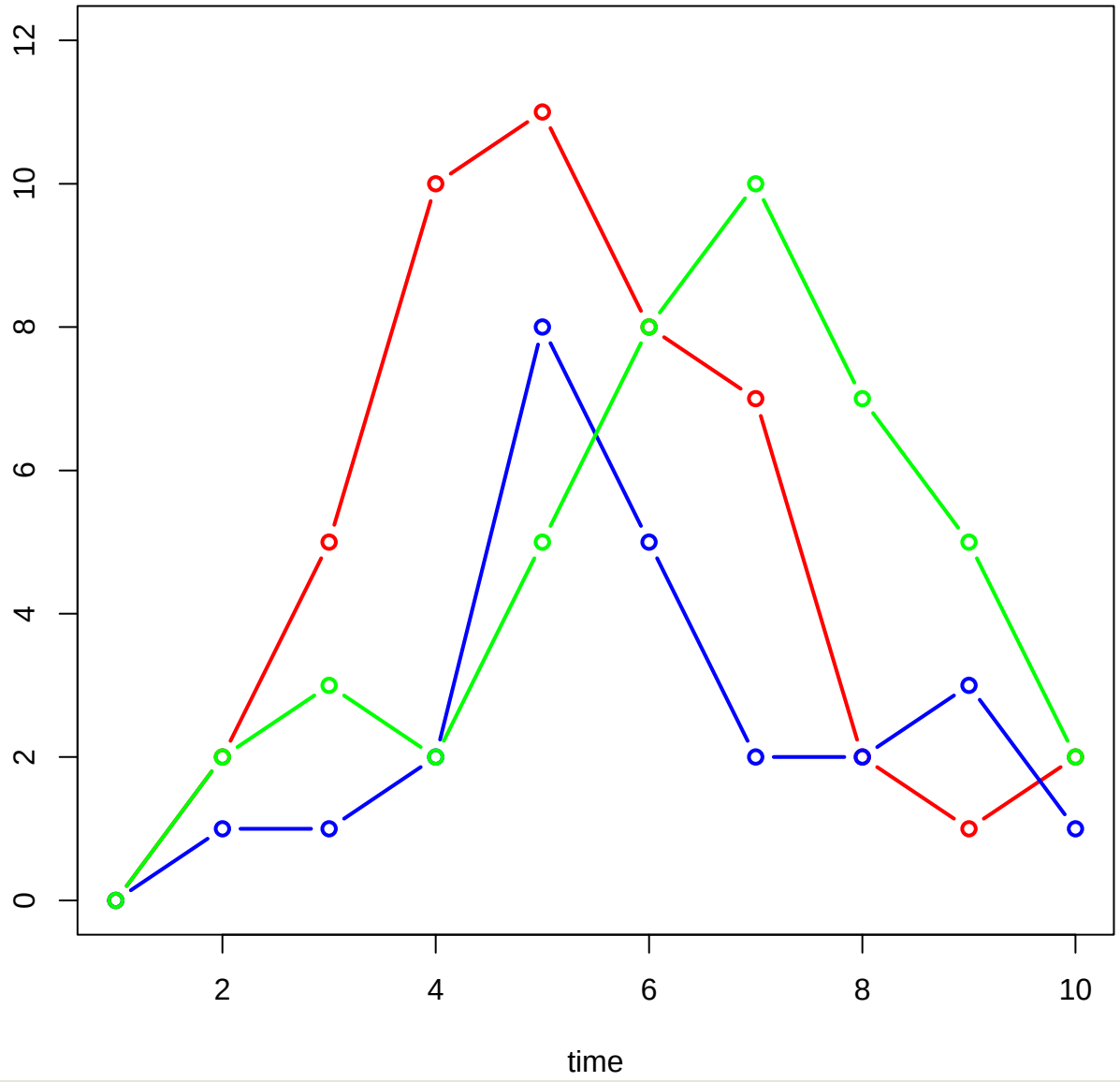
  データ長が異なる

  位相がずれている

類似度（距離）



類似度（距離）



時系列サンプル間の類似度

シーケンス長の異なるサンプル間の比較

位相のずれの影響を受ける

Dynamic Time Warping (DTW)

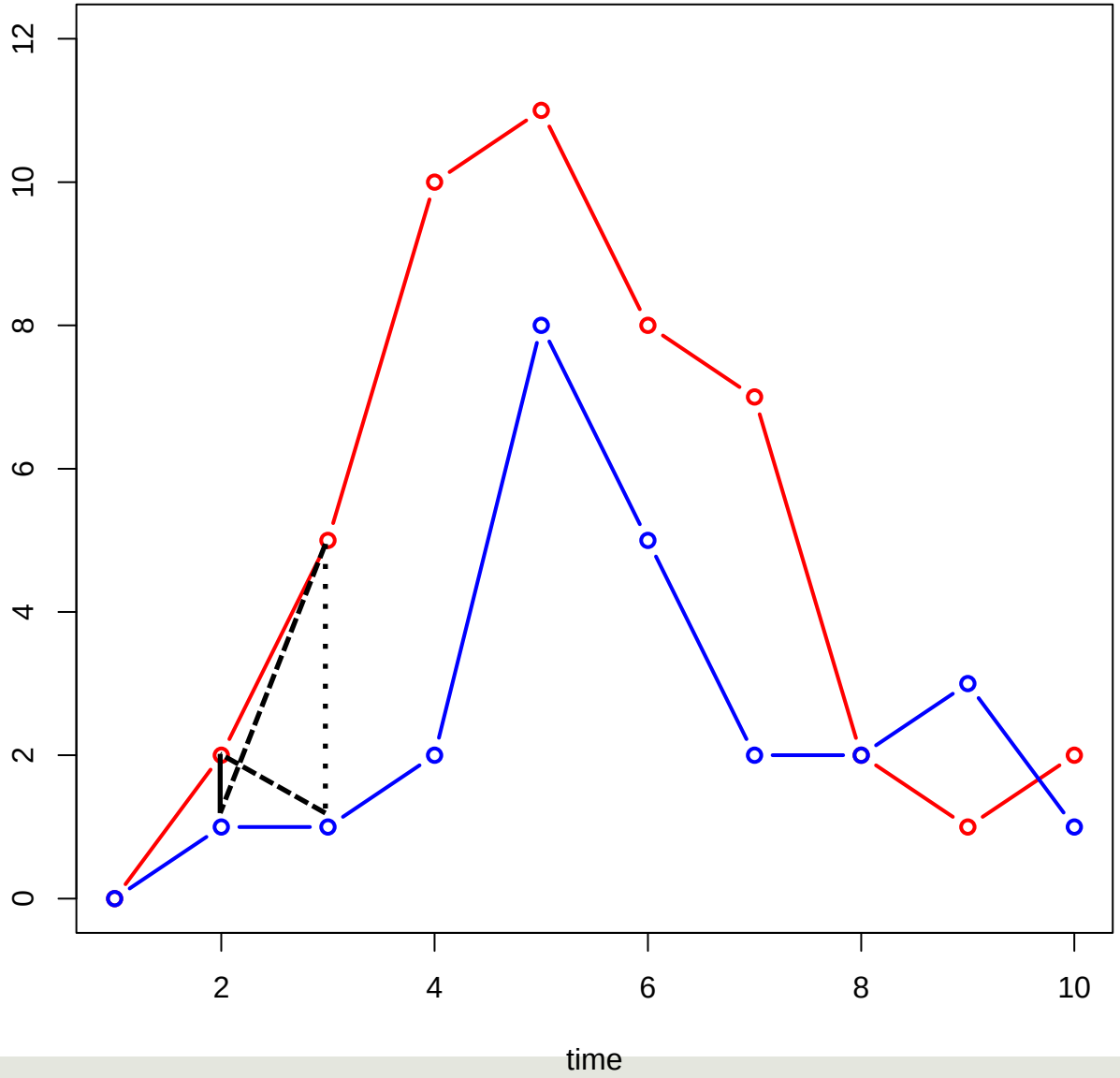
時間スケールを考慮した距離尺度の一つ。シーケンス間の距離を最小化するよう時間軸方向にサンプリングスケールを調整する。
シーケンス長の異なるサンプル間の類似度を計算できる

$$d(i,j) = \left\| x_i - y_j \right\| + \min \begin{cases} d(i, j-1) \\ d(i-1, j) \\ d(i-1, j-1) \end{cases}$$

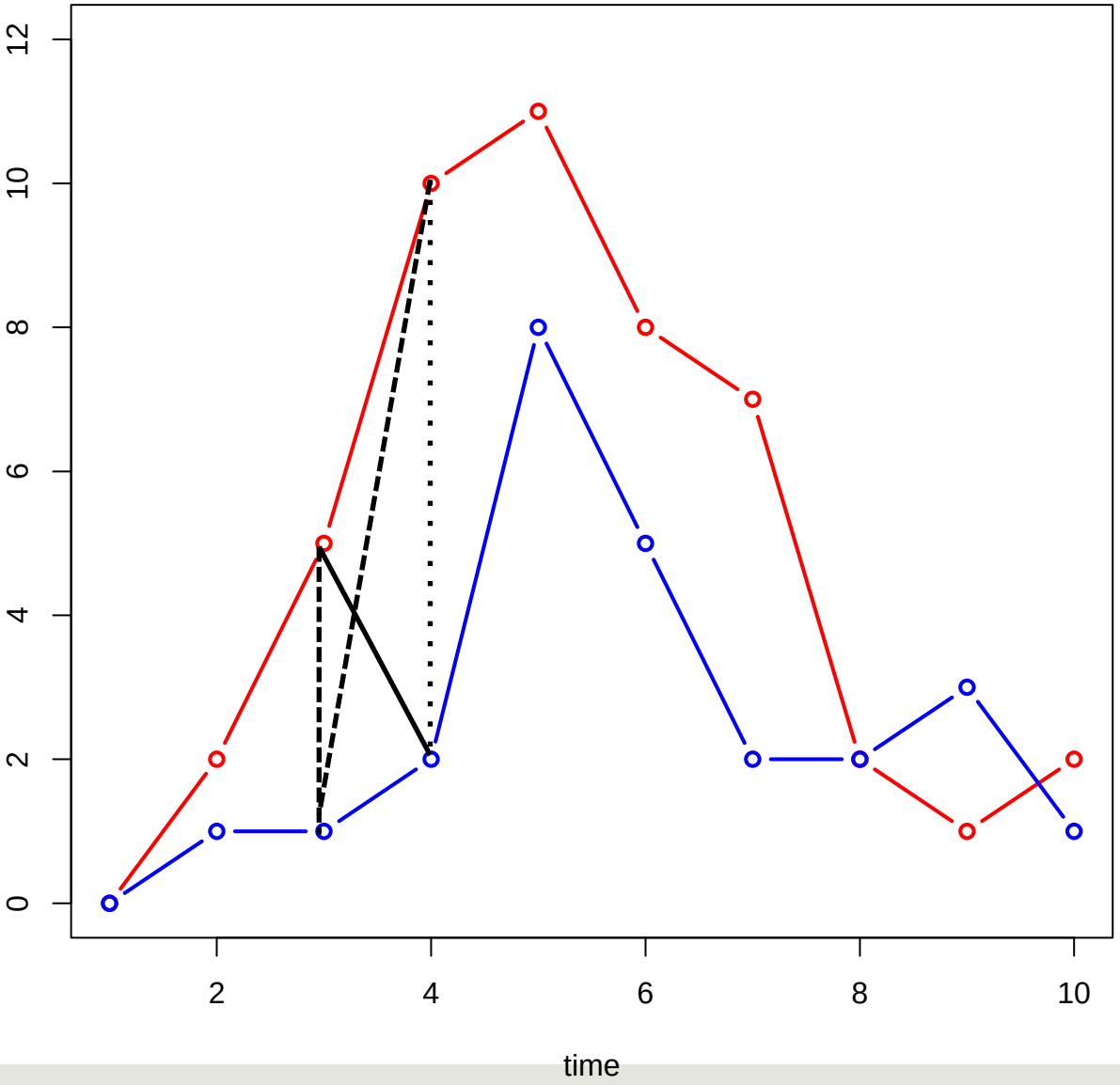
$$d(0,0) = 0, d(i,0) = d(0, j) = \infty$$

Dynamic Time Warping

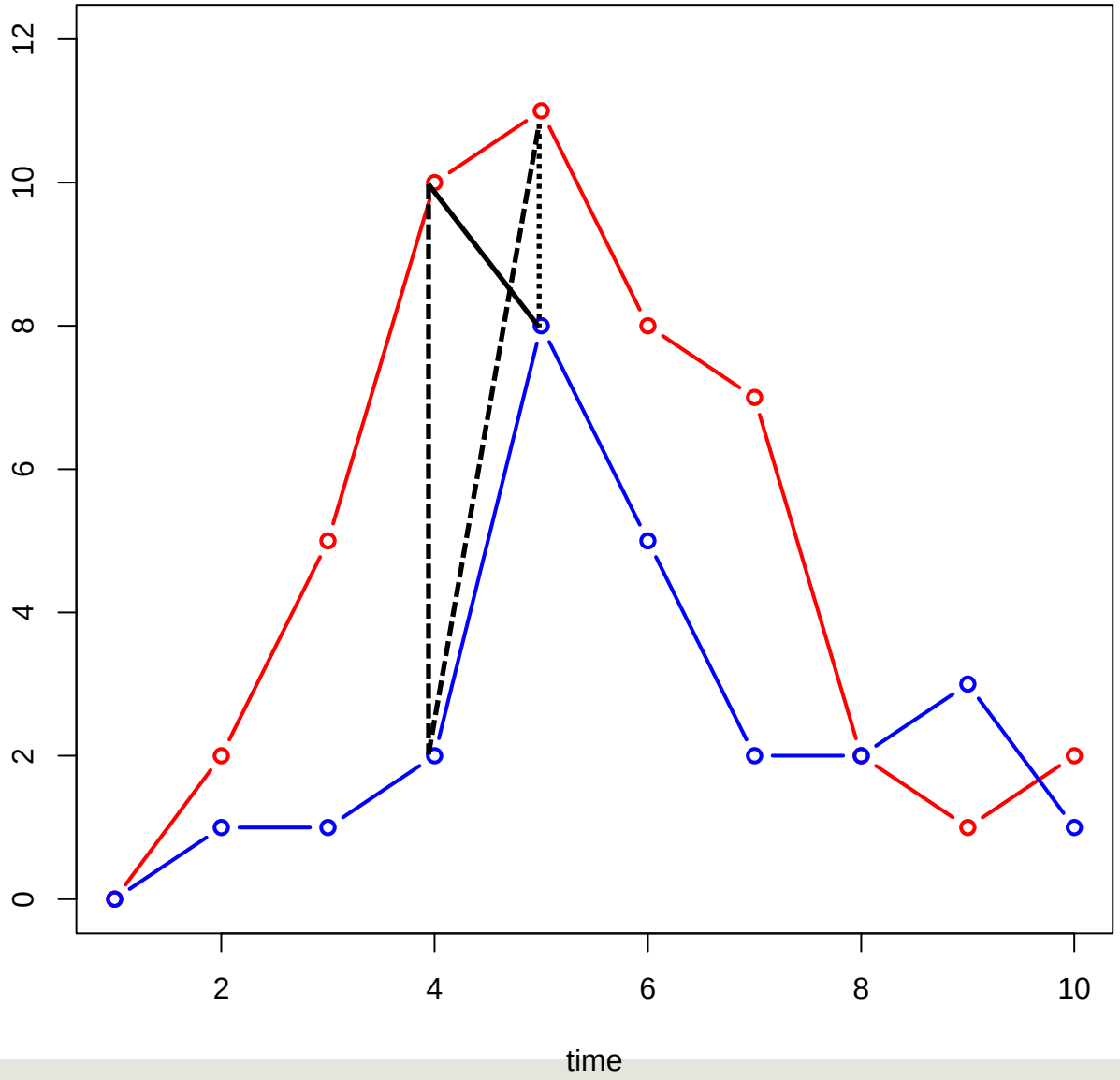
i=3
j=3



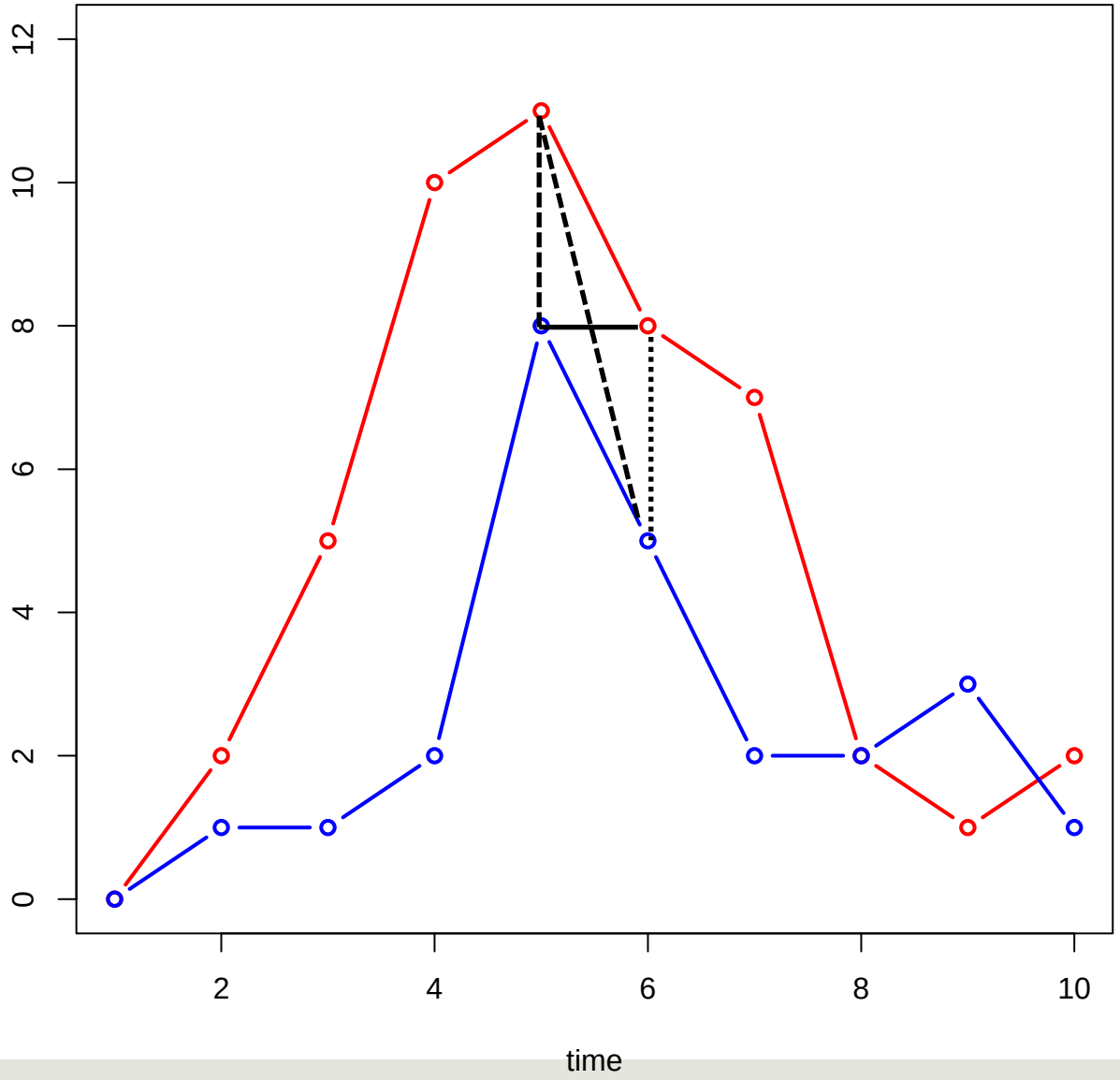
Dynamic Time Warping

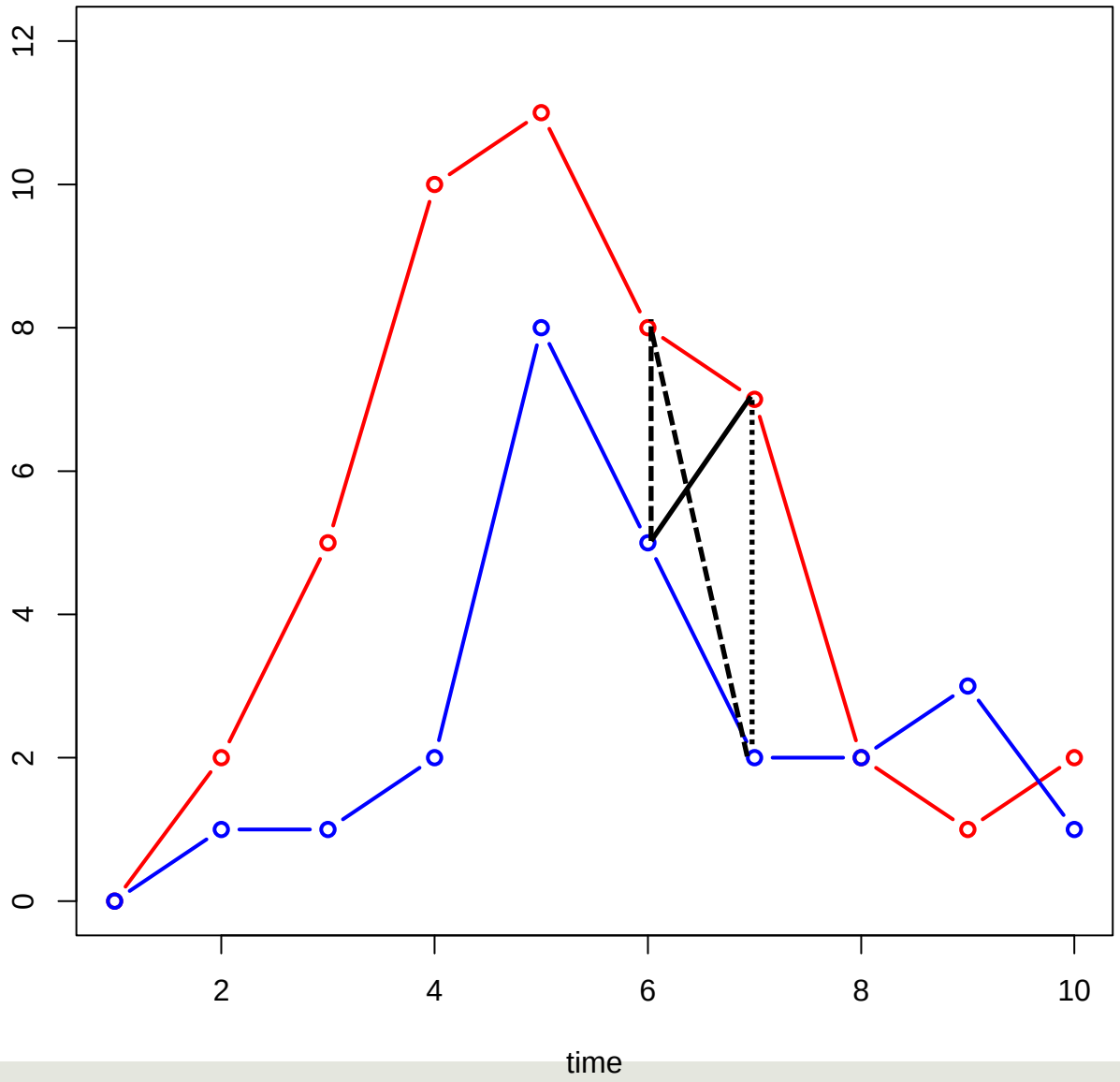


Dynamic Time Warping

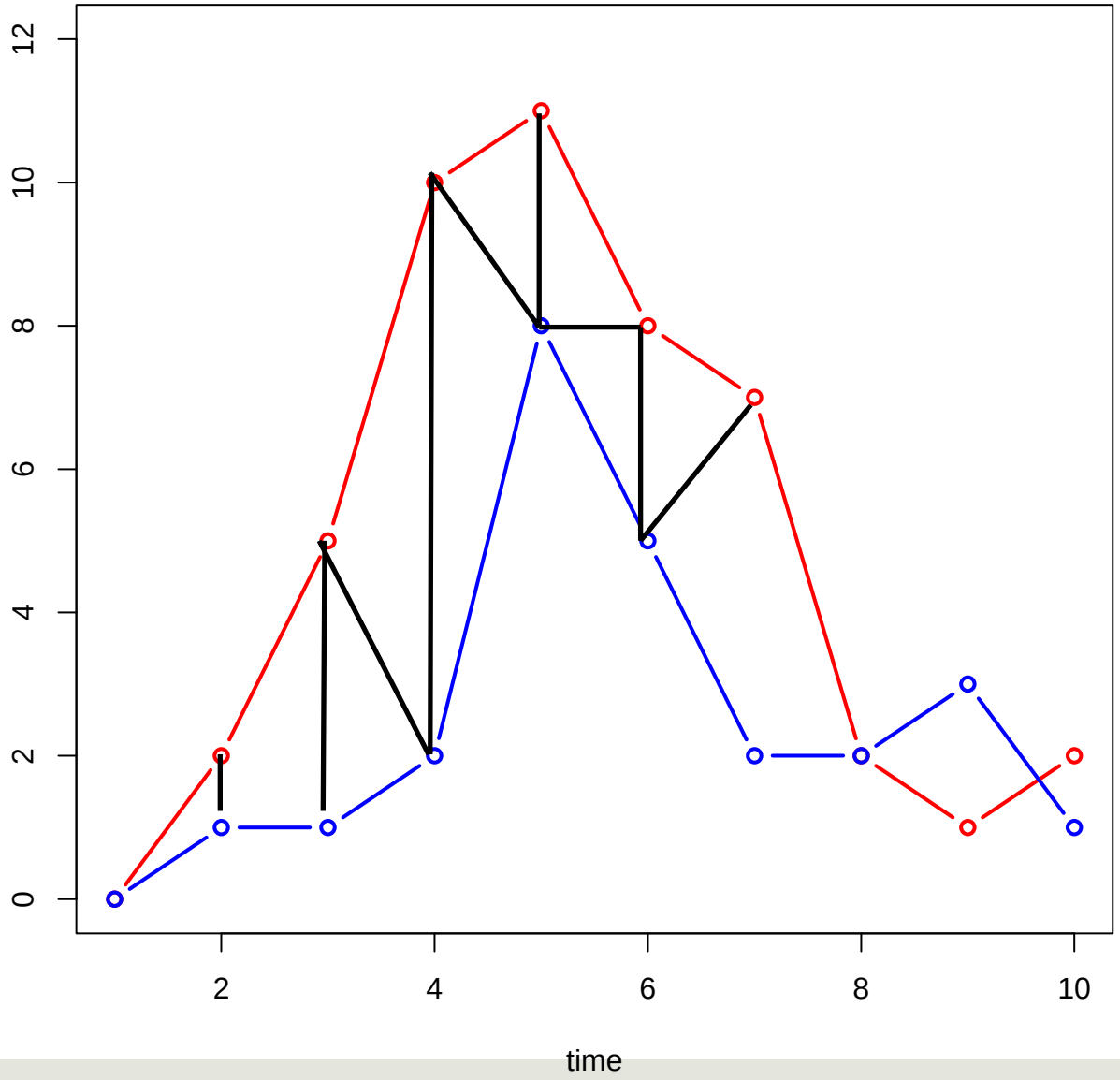


Dynamic Time Warping

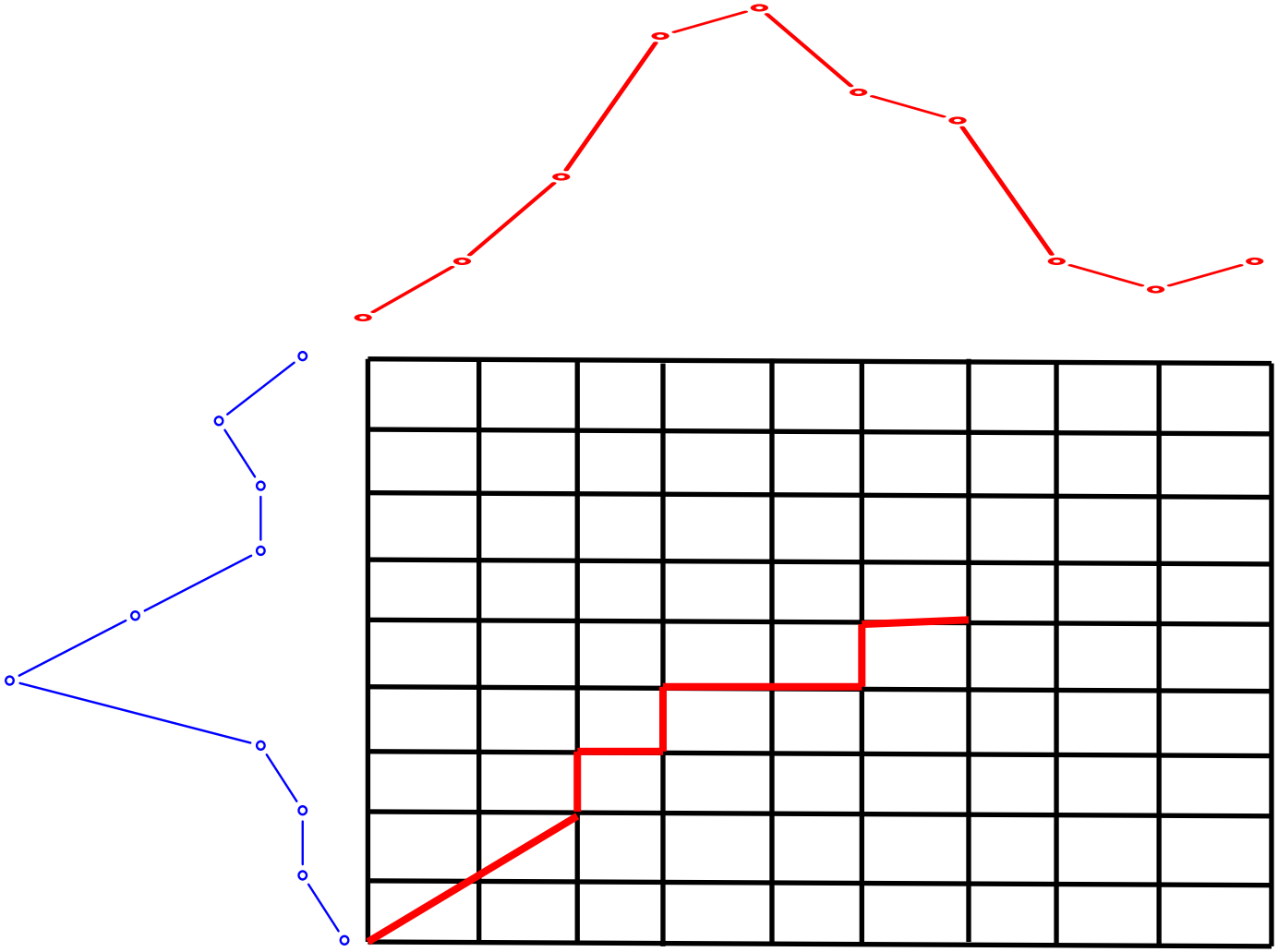



0
2
4
6
8
10
12
2
4
6
8
10
time

Dynamic Time Warping



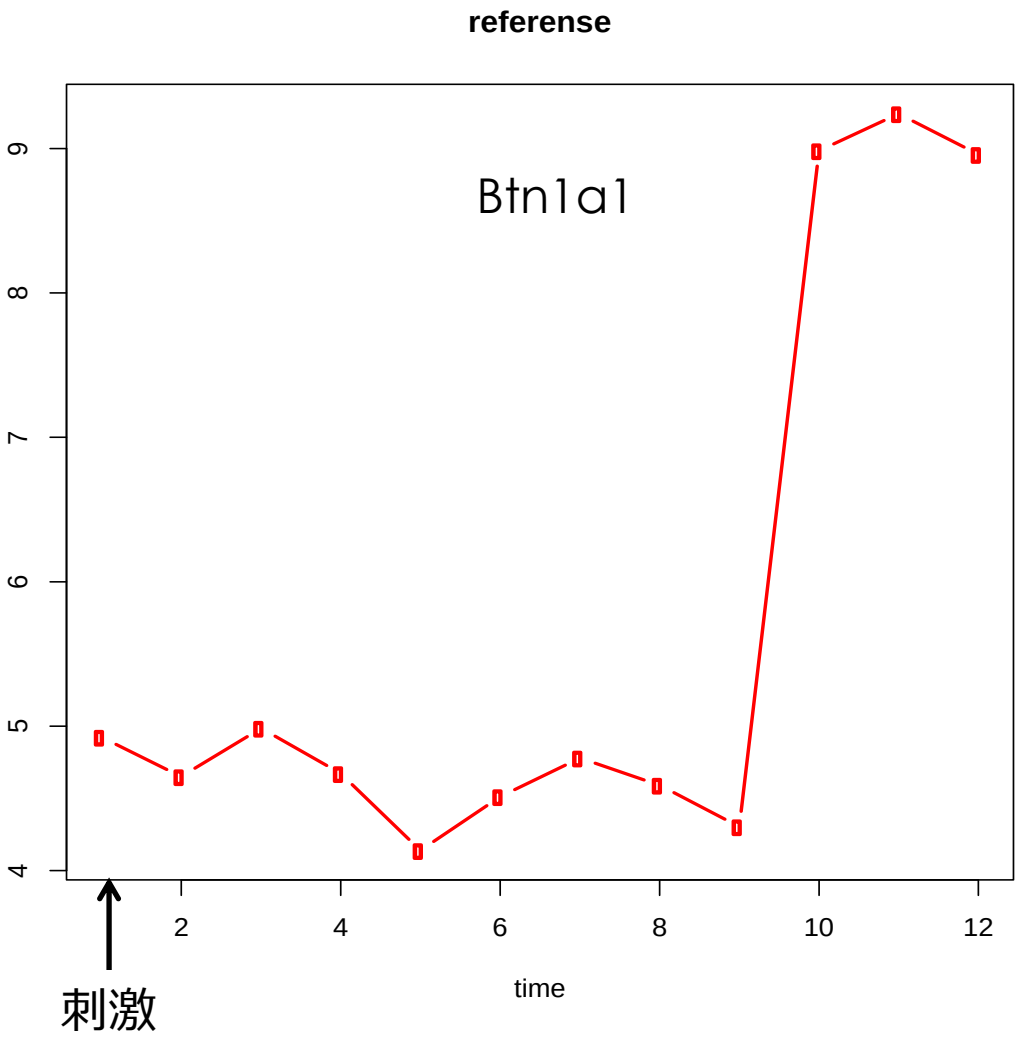
Warping Path



例

- GEO database から遺伝子発現データを取得
- 薬物刺激→12 time pointでマイクロアレイ
- 特徴的な遺伝子発現パターンを抽出

リファレンス発現パターン

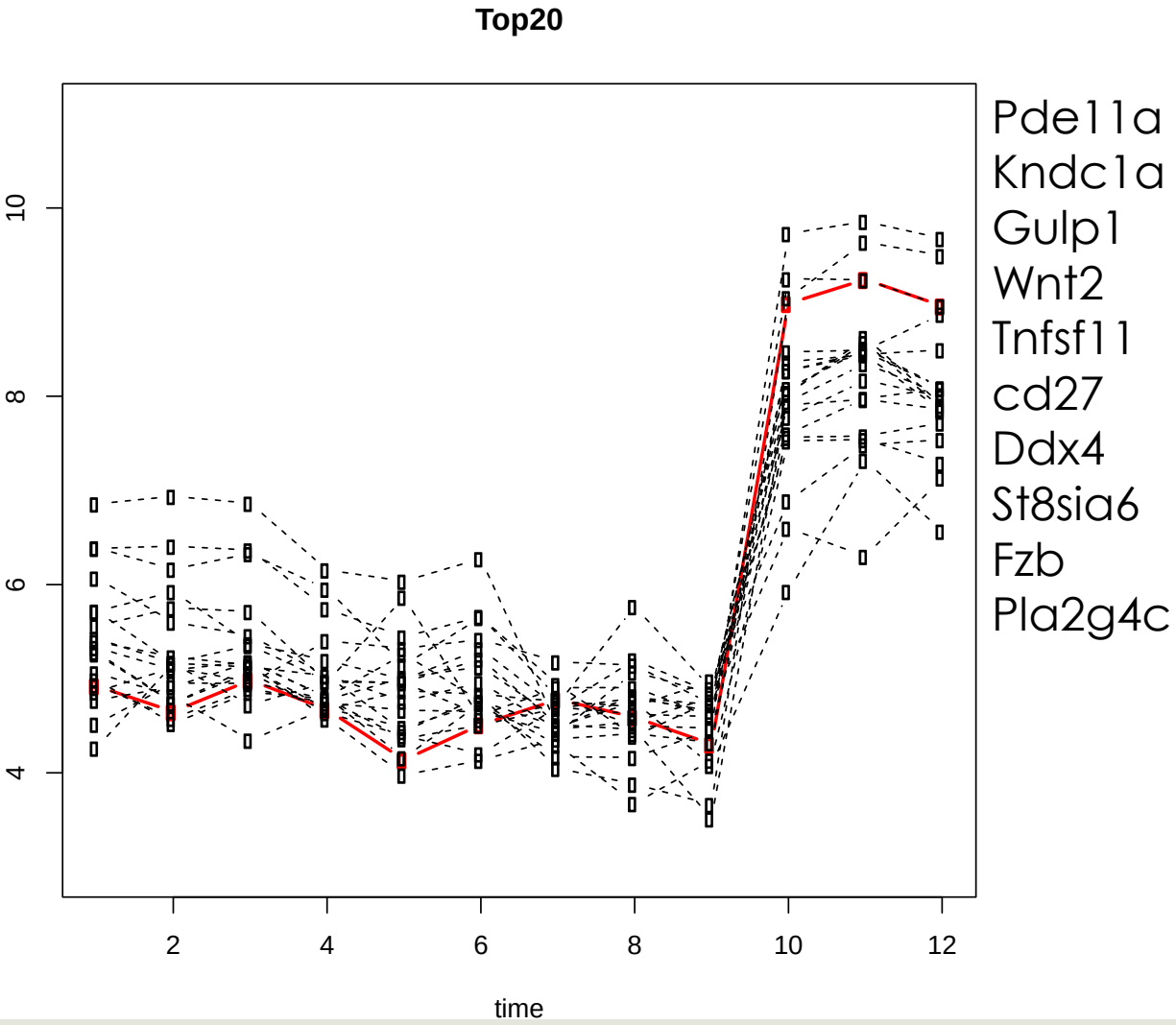


DTW

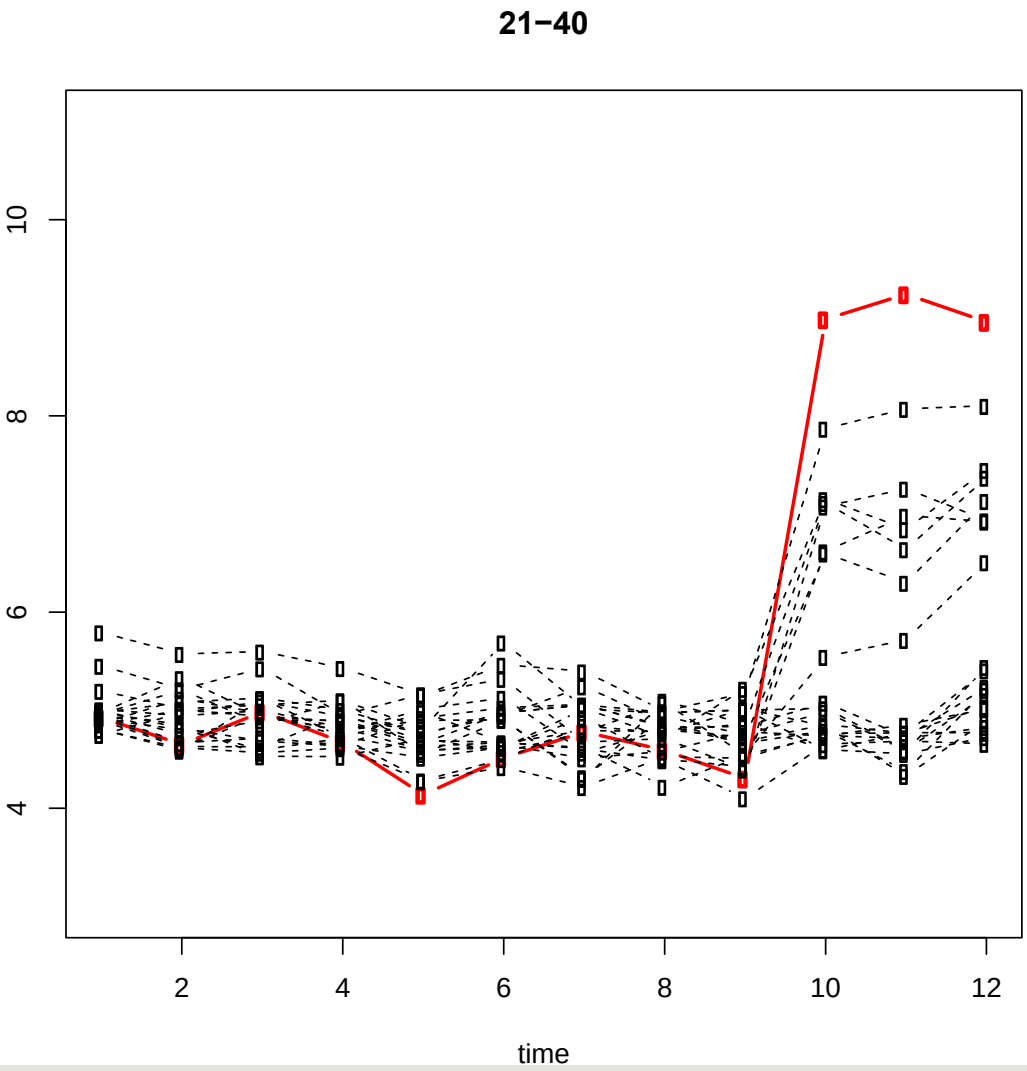
- リファレンス遺伝子ーその他の遺伝子間でDTW距離を計算
  （統計解析ソフトRのパッケージを使用）

- DTW距離の小さい順番に遺伝子をソート

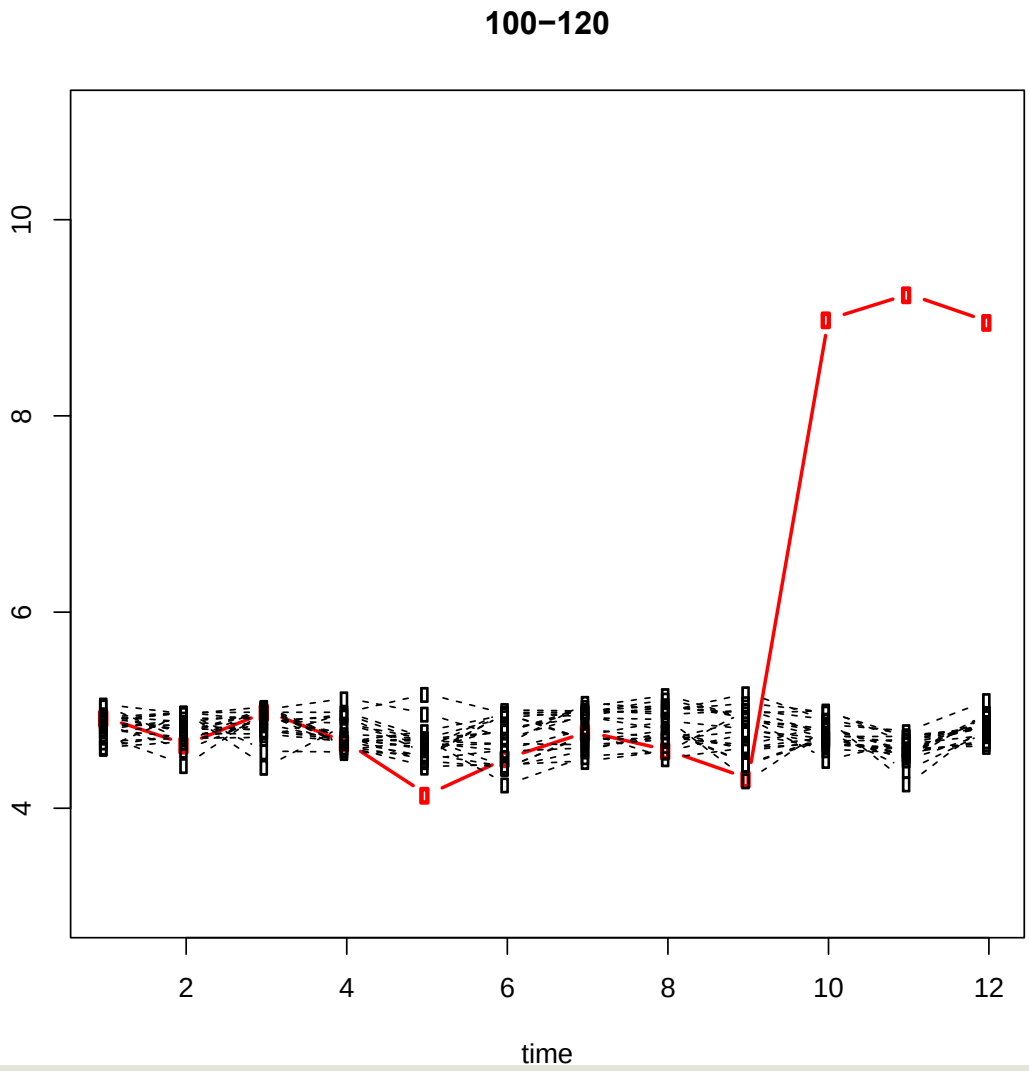
類似度上位20位


Top20
10
8
6
4
2
4
6
8
10
12
time
Pde11a
Kndc1a
Gulp1
Wnt2
Tnfsf11
cd27
Ddx4
St8sia6
Fzb
Pla2g4c

類似度20～40位



類似度100位以下



まとめ

- DTW距離を用いて似た遺伝子発現パターンを持つ遺伝子を抽出することができた。

- データ長が長くなると多くの計算時間が必要となる。

  O(m,n)