Vol.2010-MPS-80 No.24
2010/9/29

相同性検索を用いた2つの時系列データからの類似部分抽出手法とDTWによる類似部分の評価

廣安知之[†1]　西井琢真[†2]　吉見真聡[†3]
三木光範[†3]　横内久猛[†1]

本研究では，2本以上の時系列データに対して類似部分の抽出手法を提案している．提案手法は，時系列データを再量子化し，文字列検索アルゴリズムを用いて類似部分を抽出する方法である．文字列検索アルゴリズムには，相同性検索を利用する．相同性検索を利用することで，将来，既存の並列アルゴリズムを利用することで高速に処理が可能である．数値実験を通じて，再量子化手法の違いによって抽出される類似部分にどのような差異が生まれるか検討した．また，既存の時系列データの距離測定手法であるDTWとどの程度一致する類似部分を抽出するのかについて検討した．実データへの適用として，fNIRSを使った脳機能実験を行い，提案手法による時系列データの抽出を試みた．

Homology Search Retrieval from Two Time Series Data and Evaluation by Dynamic Time Warping

Tomo Hiroyasu [†2] Takuma Nishii [†1] Masato Yoshimi [†3]
Mitsunori Miki [†3] and Hisatake Yokouchi [†1]

Here, we proposed a method for extracting the most similar subsequences from two time series data. In the proposed method, time series data is quantized and similar subsequences are extracted by string search algorithms. As a string search algorithm, homology search algorithm is utilized. Using homology search, strong parallel libraries can be used in the future. Through the numerical examples, the differences of results between two methods of quantization were discussed. At the same time, the results of the proposed method and those of the conventional distance measurement method, the Dynamic Time Warping (DTW) method, were discussed. For the real world problems, the proposed method were applied to time series date of brain function experiments using fNIRS and extracted the most similar subsequences.

1. はじめに

近年，医療分野では光トポグラフィや脳波計などの生体情報診断機器が普及している．これらの診断機器から得られたデータの解析が進むことで，人間の脳機能が解明されていくことが期待される．これにより，将来的に頭の中で考えたことを可視化し，様々な装置に応用することができると考えられる．光トポグラフィは，脳血流の増減を測定することで脳の活性度を把握する装置であり，“ テレビを観賞している時”や“ 歌っている時”に活性化する脳の部位を調べることができる．また，“ 楽しく音楽を聴いている時”や“ 楽しく食事している時”の実験データを解析すれば，“ 楽しい”という感情で共通して活性化する部位，つまり類似した活動をする部位があるのかなどを知ることができる．しかし，光トポグラフィは1回の実験で数百個の時系列データを発生させるため，複数の実験結果を比較する際に，大量のデータのどこに着目すればよいのかが分からないという問題がある．また，脳の血流の増減速度には個人差があるという問題が存在する．血流の急激な増減や緩やかな増減に対応するため，光トポグラフィのデータはある程度の時間伸縮を許容する必要がある．この問題に対して，複数の時系列データの中から時間伸縮を許容して類似している部分を自動的に探し出す手法があれば，解析者の負担を軽減できると考えられる．

そこで本研究では類似している部分を探し出すため，2つの時系列データからの類似部分の抽出を試みた．提案手法は，時系列データを再量子化し，相同性検索を用いて類似部分を抽出する方法である．多チャンネルの時系列データは大規模であり，その計算コストが問題となる．本論文では検討に至っていないが，相同性検索を用いることで，特にバイオインフォマティクスの分野で開発されている並列ライブラリを利用することで，高速に処理することが期待される．提案アルゴリズムは，まず人工的な時系列データによりその探索を検討する．つづいて，実際に光トポグラフィを利用した脳機能実験から得られた時系列データに対して提案手法を適用し，類似部分の抽出を試みる．

†1 同志社大学生命医科学部
Faculty of department of life and medical science, Doshisha University
†2 同志社大学大学院工学研究科
Graduate School of Engineering, Doshisha University
†3 同志社大学理工学部
Faculty of Science and Engineering, Doshisha University

2. 関 連 研 究

時系列探索法は，音声や映像の時系列データを扱う分野において，ある音や映像の信号が長大な時系列データ内のどこに存在するかを探索する手法である．時系列探索法には，一致検索を目的とした時系列アクティブ探索法 (Time-Series Active Search:TAS)[1] や，時間伸縮を許容する検索を目的とした DTW(Dynamic Time Warping) がある．また，2 つの時系列データの類似部分を求める手法として，RIFTAS(Reference Interval-Free Time-Series Active Search)[2] や，RIFCDP(Reference Interval-Free Continuous DP)[3] がある．これらの手法はウィンドウ幅を決め，部分時系列データを作成し，アクティブ探索や DTW を繰り返すものである．TAS を用いて任意個数の時系列データに含まれる類似部分の探索法として RDDS-n 法[4] が提案されている．しかしながら時系列データ数の増加に伴い，処理時間が爆発的に増加するという問題がある．これらの手法は計算回数の軽減が試みられている[5] が，計算量は膨大であり実データへの適用には向いていないと考えられる．

豊田らは，DTW を用いて 2 つのデータストリームから類似する部分シーケンスペアを検出する手法を提案している[6]．この手法は計算量，精度ともに優れているが適切な類似の閾値をパラメータとして定める必要がある．Keogh は 1 本の時系列データ中の頻出パターンを抽出する手法を提案している[7] が，2 本以上の時系列データには対応できない．

上記以外には，相互相関関数により 2 つの時系列データの相関値の最大を求める相互相関分析[8][9] を用いることで，時系列データ同士が最も類似する部分を抽出する手法がある．しかし，この手法は類似部分の時間伸縮を考慮していないという問題がある．同様にフーリエ変換を用いた手法も類似部分の時間伸縮を考慮していないと言える．

時系列データを再量子化し，パターンを抽出する試みは既に行われている[10]．しかし，2 本以上の時系列データに対して，文字列検索アルゴリズムを用いて類似部分を抽出した例はない．また，複数の時系列データから時間伸縮を許容して類似部分を探し出す手法は少なく，複数の時系列データの処理が大きな計算量を必要とすることが問題となる．

3. 相同性検索と SW(Smith Waterman) 法

相同性検索は，バイオインフォマティクスの分野で広く用いられている文字列検索アルゴリズムであり，DNA や塩基配列の類似度測定や類似部分の抽出が可能である．例えば，ハツカネズミの未知の遺伝子を発見した際に，ヒトがその配列と類似した遺伝子を持つかどうかを調べる場合などに用いられる．相同性検索には，精度を重視する SW(Smith Waterman) 法，速度を重視する FASTA(FAST-All)，BLAST(Basic Local Alignment Search Tool) がある．本論文では精度を評価するため SW 法を選択した．SW 法は動的計画法の 1 種であり，全ての部分文字列の比較を行うことで類似度を最適化する．類似度は文字列テーブルのスコアによって評価される．図 1 に文字列テーブルを示す．文字列テーブルでは文字列 A のそれぞれの文字が行に，文字列 B のそれぞれの文字が列に割り当てられる．長さが m と n の文字列から類似部分を抽出する場合，アルゴリズムのオーダーは $O(mn)$ である．図 2 に SW 法で抽出された類似部分文字列の例を示す．

x | y	_	C	B	C
_				
B				
B				
C				

図 1 文字列テーブル
Fig. 1 The matrix

“ PELICAN” → “ ELICAN”
“ PAWHEAE” → “ AW_HE”

“ COELACANTH” → “ ELACAN”
“ HEAGAWGHEE” → “ AWGHE”

図 2 SW 法で抽出された類似部分文字列の例
Fig. 2 Example of SW algorithm’s alignment

3.1 スコアパラメータ

SW 法には $match, mismatch, gap$ の 3 つのパラメータがある．$match$ は文字列の一致に，$mismatch$ は文字列の不一致に，gap はスペース発生に関わるパラメータである．標準的な設定は $match = 1, mismatch = -1, gap = -1$ である．これらのパラメータが変化すれば抽出される文字列も変化する．$mismatch$ が $match$ より低ければ，類似部分の長さは短くなるがその分，一致度の高い文字列を抽出することができる．また，類似部分の比較を行う際，スペースが入ることで類似度が高くなる部分文字列も存在する．gap は 0 に近いほど，スペースが入りやすい文字列を抽出することができる．gap によるスペースの発生は，時系列データにおける時間的な伸縮を意味することになる．どのパラメータが最適であるかは，元デー

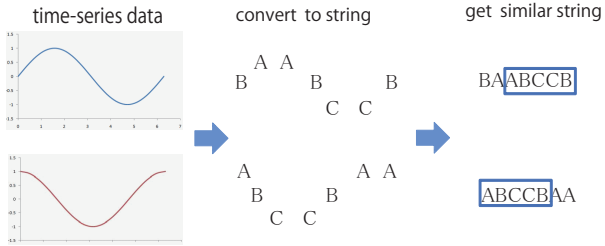


図 3 提案手法の流れ
Fig. 3 Symbolization of sine wave

タやどのような類似文字列を抽出するかによって異なる．

4. 2 つの時系列データの類似部分の抽出方法

4.1 概　　要

本論文では時系列データの再量子化と相同性検索の組み合わせによる，2 つの時系列データからの類似部分の抽出手法を提案する．図 3 に提案手法の流れを示す．まず，時系列データを再量子化する．本論文では再量子化は文字列化を意味する．例えば，図 3 の上部の時系列データは“ BAABCCB”に，下部の時系列データは“ ABCCBAA”に変換される．再量子化の手法としては，SAX と等間隔領域分割がある．時系列データを文字列に変換することによって，相同性検索を適用することが可能となる．相同性検索は 2 つの文字列から最適な類似部分を探索する手法である．この手法により“ BAABCCB”と“ ABCCBAA”から類似部分として“ ABCCB”を取り出すことができる．このように時系列データの再量子化と相同性検索による部分文字列の抽出によって時系列データの類似部分の抽出が可能になる．

4.2 時系列データの再量子化

時系列データを再量子化する手法には，SAX と等間隔領域分割が挙げられる．再量子化するためには，時系列データの数値と文字の対応関係が必要である．よって分割線を定める必要がある．例えば，分割線を $\{-0.6, 0.6\}$ に設定すると，時系列データ $T = \{0.5, 1.5, -0.8\}$ は“ BAC”に再量子化される．2 つの手法は分割線の定め方が異なっている．SAX は波形が正規分布になることを仮定し区分域を決める．等間隔領域分割は一様分布になることを仮定し，最小値と最大値を等間隔で区切り分割線を定める．再量子化では，1 つの波形を何種類の文字で表現するか定める必要がある．ある時系列データの分割数は任意に定めることができる．

4.2.1 SAX(Symbolic Aggregation approXimation)

SAX は，Keogh らによって提案された時系列データの表現手法である[13]．この手法は時系列データが正規分布することを仮定し，データを文字列に変換する．SAX は波形が正規分布することを仮定しているため，時系列データを文字列に変換する前に時系列データの標準化が必要となる．標準化とは，平均が 0，分散が 1 となるようにデータを変換することである．例えば，時系列データ

$$X(t) = \{x(1), x(2), x(3), \ldots, x(M)\}$$

を標準化すると

$$X(t) = \{\frac{x(1)-\mu}{\sigma}, \ldots, \frac{x(M)-\mu}{\sigma}\} (\mu : \text{X の平均値}, \sigma : \text{標準偏差})$$

となる．SAX による文字列変換の流れを図 4 に示す．step1 によって標準化がなされ，step2 によってデータが文字列に変換される．これにより標準化された時系列データを等領域に分割する分割線を定めることができる．これらの分割線は標準正規分布表に基づいている．分割線を $B = (\beta_1, \ldots, \beta_n)$，分割記号数を α として表 1 に示す．図 4 において，下の分割線より下にあるデータは“ c”に変換され，上の分割線と下の分割線の間にあるデータは“ b”に，上の分割線より上にあるデータは“ a”に変換される．結果的に時系列データは“ aaabcc”に変換される．

表 1 分割記号数が 2 から 7 のときの正規分布を等領域に分割する分割線
Table 1 The breakpoints that divides a normal distribution in an arbitrary number from 2 to 7 of equiprobable regions

α	2	3	4	5	6	7	8
β_1	0.00	-0.33	-0.67	-0.84	-0.97	-1.07	-1.15
β_2		0.33	0.00	-0.25	-0.33	-0.57	-0.67
β_3			0.67	0.25	0.00	-0.18	-0.32
β_4				0.84	0.33	0.18	0.00
β_5					0.97	0.57	0.32
β_6						1.07	0.67
β_7							1.15

4.2.2 等間隔領域分割

等間隔領域分割は，波形が一様分布に従うと仮定し，時系列データ中の最大値と最小値の間を等分割する分割線を定める手法である．等間隔領域分割による文字列変換の流れを図 5 に示す．例えば，波形を 3 つの文字に変換する場合，波形の最小値と最大値を等分割する 2 つの分割線を定める．最大値 = 1，最小値 = −1 であれば，領域幅は 0.68 となり，2 つの分割線は (0.34, −0.34) となる．

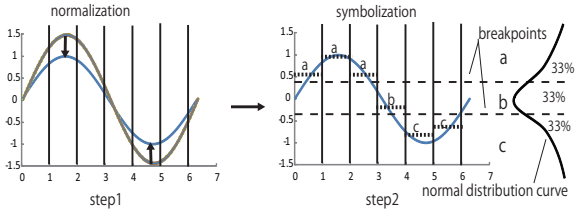


図 4 SAX による文字列変換の流れ
Fig. 4 The concept of Symbolic Aggregation approXimation

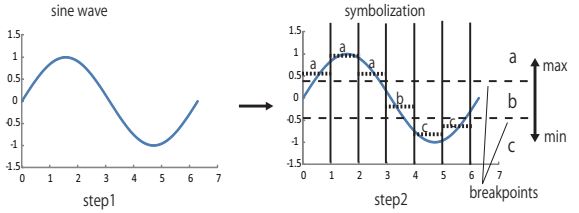


図 5 等間隔領域分割による文字列変換の流れ
Fig. 5 The concept of Equal Intervals Area Division

例えば，時系列データ

$$X(t) = \{x(1), x(2), x(3), \ldots, x(M)\}$$

の領域幅は (1), 分割線は (2) のように決定できる．X は時系列データ，w は領域の間隔，num は分割文字数を表す．

$$w = \mid max(X) - min(X) \mid / num \tag{1}$$

$$breakpoints(n) = max(X) - w * num \ (1 \leq n \leq num - 1) \tag{2}$$

5. 数値実験による精度検証

5.1 目　　的

数値実験により提案手法が有効に働くかどうか検証する．実験の目的は，ある時系列データに対して提案手法を適用したとき，SAX と等間隔領域分割で抽出される類似部分がどのように異なるかを比較することである．また，それらの 2 つの手法に適したデータセット，適さないデータセットがあるかどうかを確認することである．

5.2 データセット

使用したデータセットは時系列クラスタリングデータセット[14] である．このデータセットにはあるタイプの時系列データ (実数) が複数のクラス別に含まれている．同じクラス内に属する時系列データは互いに類似していることが考えられる．ここでは同じクラス内に属する 2 つの時系列データの全体を類似部分として抽出できるかどうかを検証した．なお，SAX のグラフは元の時系列データを標準化したものであるため，等間隔領域分割のグラフとは差異がある．分割文字数は予備実験より 5 文字とした．SW 法のパラメータは $match = 1, mismatch = -1, gap = -1$ とした．

5.3 評 価 方 法

SAX と等間隔領域分割では異なる部分が類似部分として抽出される．今回は，類似する 2 つの時系列データを用いたため，データ全体を類似部分として抽出ことが望ましい．抽出された類似部分が長く類似度が低い場合と，類似部分は短いが類似度が高い場合に，2 つの手法のどちらが適しているのかを判断するため，抽出された類似部分の長さと DTW 距離を用いて評価を行った．

DTW 距離を抽出された類似部分の長さで割ることにより，長さ 1 あたりの DTW 距離を誤差として求めることができる．この誤差が小さいほど抽出された類似部分は似ていることになり，SAX と等間隔領域分割でどちらがより良い類似部分を抽出しているのか判断できる．表 2 にそれぞれの図におけるデータの長さ，抽出された類似部分の長さ，類似部分の抽出率，2 つの時系列データの類似部分間の DTW 距離，長さ 1 あたりの DTW 距離の結果を示す．

5.4 結　　果

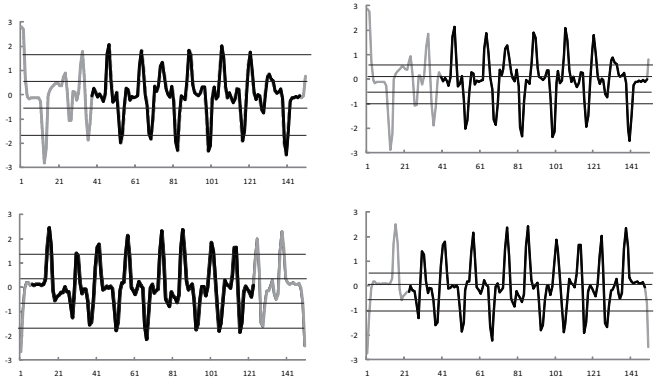
4 つのデータに対して SAX と等間隔領域分割を適用し，それぞれの特徴を調べた．図 6〜図 9 には 2 本の時系列データがあり，類似部分が黒色の太線でそれ以外の部分が灰色の線で表されている．簡単のため，変換で得られた文字列は標記しない．

また，表 2 より長さ 1 あたりの DTW 距離は SAX でも等間隔領域分割でもほぼ同じになることが分かった．抽出された類似部分の誤差はほぼ同じであるため，できるだけ長く類似部分が抽出されたほうが良いのではないかと言える．

6. 結論と今後の課題

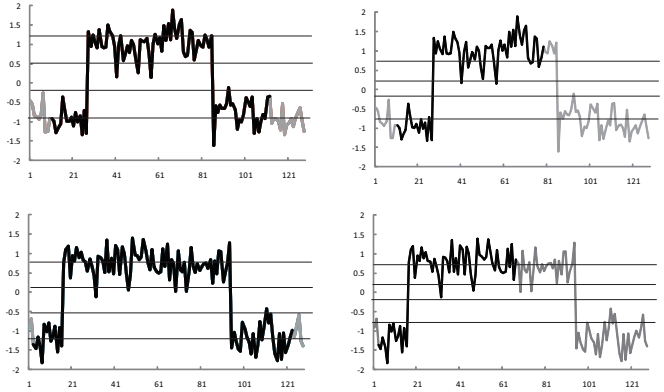
光トポグラフィは 1 回の実験で数百個の時系列データを発生させるため，複数の実験結果

Vol.2010-MPS-80 No.24
2010/9/29



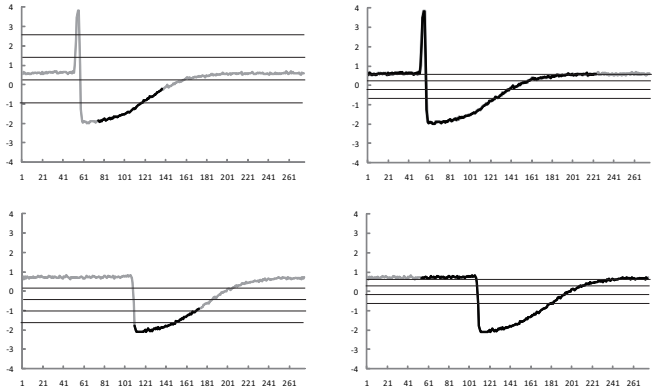
(a) 等間隔領域分割　　(b) SAX

図 6　等間隔領域分割と SAX の両方で類似部分が抽出された例
Fig. 6　The good results of SAX and EIAD



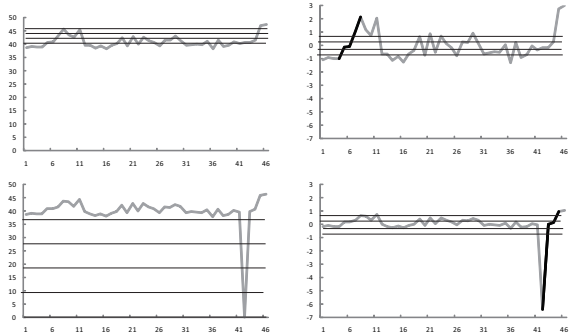
(a) 等間隔領域分割　　(b) SAX

図 7　等間隔領域分割では類似部分が抽出されたが SAX では抽出されなかった例
Fig. 7　The good results of EIAD but poor results of SAX



(a) 等間隔領域分割　　(b) SAX

図 8　SAX では類似部分が抽出されたが等間隔領域分割では抽出されなかった例
Fig. 8　The poor results of EIAD but good results of SAX



(a) 等間隔領域分割　　(b) SAX

図 9　等間隔領域分割と SAX の両方で類似部分が抽出されなかった例
Fig. 9　The poor results of SAX and EIAD

表 2 それぞれの図における類似部分の抽出率，長さ 1 あたりの DTW 距離
Table 2 Extraction rate and DTW distance of similar subsequences in each figure

図の名前	データの長さ	抽出された類似部分の長さ	類似部分の抽出率	2 つの時系列データの類似部分間の DTW 距離	長さ 1 あたりの DTW 距離
図 6(a)	150	114	0.76	41.6	0.36
図 6(b)	150	118	0.79	53.3	0.35
図 7(a)	130	112	0.86	32.2	0.29
図 7(b)	130	67	0.52	16.3	0.24
図 8(a)	270	63	0.23	10.6	0.17
図 8(b)	270	222	0.82	23.2	0.10
図 9(a)	46	1	0.02	1.1	1.09
図 9(b)	46	4	0.09	42.6	10.65

を比較する際に，大量のデータのどこに着目すればよいのかが分からないという問題がある．そこで本論文では，時系列データの再量子化と相同性検索の組み合わせによる 2 つの時系列データからの類似部分の抽出手法を提案した．相同性検索は文字列検索アルゴリズムであるため，時系列データに対して適用するためには再量子化が必要であった．

まず，本論文では 2 種類の再量子化手法ににおける精度検証を行った．その結果，2 つの手法のどちらも位相がずれたデータに対してはうまく適用できることが分かった．しかし，SAX・等間隔領域分割の両手法において，外れ値がある時系列データや分割線を境として増減するデータは類似部分が抽出しにくいことが分かった．よって，提案手法を適用するには時系列データの外れ値の除去や平滑化等の前処理が必要であることが分かった．

次に，既存の時系列データの距離測定手法である DTW とどの程度一致する類似部分を抽出するのかについて検討した．その結果，提案手法による類似部分は，ランダムで類似部分を抽出した場合よりも DTW による類似部分と近くなることが分かった．今回行った実験ではパラメータのチューニングを行っていない．そこで今後は，分割文字数や SW 法のパラメータチューニングを行なう必要がある．

最後に，実データへの適用として，fNIRS を使った脳機能実験を行い，提案手法による時系列データの抽出を試みた．

上記以外に，多チャンネルの時系列データは大規模であり，その計算コストが問題となる．大量の時系列データを素早く探索するには，並列化アルゴリズムを用いることが有効である．相同性検索は，GPU を用いた並列プログラムのライブラリが豊富に用意されているので，今後より高速なアルゴリズムを用いることが必要であると思われる．

参 考 文 献

1) KASHINO Kunio,SMITH Gavin A,MURASE Hiroshi, ”A Quick Search Algorithm for Acoustic Signals Using Histogram Features”, The transactions of the Institute of Electronics, Information and Communication Engineers. D-II pp.1365-1373 19990925
2) NISHIMURA Takuichi,MIZUNO Michinao,OGI Shinobu,SEKIMOTO Nobuhiro,OKA Ryuichi, ”Same Interval Retrieval from Time-Sequence Data Based on Active Search : Reference Interval-Free Time : Series Active Search (RIFAS)”,The transactions of the Institute of Electronics, Information and Communication Engineers. D-II pp.1826-1837 20010801
3) ITOH Yoshiaki,KIYAMA Jiro,KOJIMA Hiroshi,SEKI Susumu,OKA Ryuichi ”Reference Interval-free Continuous Dynamic Programming for Spotting Speech Waves by Arbitrary Parts of a Reference Sequence Pattern”, The Institute of Electronics, Information and Communication Engineers pp.1474-1483 19960925
4) 菅井 康祐, 杉山 雅英 ”任意個数の時系列に含まれる類似部分探索法”, 情報処理学会研究報告. SLP, 音声言語情報処理 2007(129), 265-270, 2007-12-20
5) 杉山 雅英 ”距離に基づく時系列 Active 探索法における l_p 距離の探索効率の比較”, 電子情報通信学会論文誌 情報・システム J91-D(8), 2071-2079, 2008-08-01
6) TOYODA Machiko,SAKURAI Yasushi, ICHIKAWA Toshikazu ”Stream Matching based on Dynamic Programming”
7) Abdullah Mueen, Eamonn Keogh, Qiang Zhu, Sydney Cash, Brandon Westover, ”Exact Discovery of time-series Motifs”, SDM 2009: 473-484
8) KATAYAMA Erika,YAMADA Yoshio,TSUZUKI Shinji ”A Method for Peak Position Estimation of Cross Correlation Functions Using Neural Network”, The Institute of Image Information and Television Engineers pp.21-24 20010302
9) 倉橋 節也, 寺野 隆雄 ”学習分類子システムを用いたプロセス時系列からのデータマイニング,” 情報処理学会研究報告 [知能と複雑系] 2002(45), 1-6, 2002-05-23
10) Yutaka Araki,Daisaku Arita,Rin-ichiro Taniguchi, ”Motif Extraction from Multi-dimensional Motion Information”,IEICE IEE-IIS 2006(19) pp.39-44 20060320
11) Lin, J., Keogh, E., Lonardi, S., Chiu, B. (2003). A Symbolic Representation of time-series, with Implications for Streaming Algorithms. In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery.
12) Eamonn Keogh, Xiaopeng Xi, Li Wei, and Chotirat (Ann) Ratanamahatana, ”Welcome to the UCR time-series Classification/Clustering Page”, http://www.cs.ucr.edu/~eamonn/time_series_data/