

Similar Subsequence Retrieval Of Two Time-series Data Using Homology Search

西井琢真

1 はじめに

本論文では、時系列データの再量子化と同源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。大量の時系列データを素早く探索するためには、並列化アルゴリズムを用いることが有効である。しかし、同源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、つまり文字列化が必要となる。本論文では、SAX と等間隔領域分割という 2 つの手法を用いて再量子化を行い、それらの手法の性質について検討した。検証の結果、時系列データは 4 つのタイプに分かれることがわかった。

2 Introduction

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によって、人間の脳機能が徐々に解明されていくことが期待される。脳機能の解明が進めば、将来的に頭の中で考えたことを可視化し、それを様々な装置に応用することができると考えられる。

光トポグラフィは人間の脳の血流の増減の測定により、脳の活性化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といったことを知ることができる。しかし、光トポグラフィは一回の実験で約 300 個の時系列データを発生させることがあり、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わ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り、光トポグラフィのデータはある程度の時間差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。まず、2 つの時系列データから類似部分を探し出すことを考えると、関連研究は以下のものが挙げられる。

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)³⁾ 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、参照区間自由時系列アクティブ探索法

(Reference Interval-Free Time-Series Active Search)²⁾ や、Reference Interval-Free 連続 DP (RIFCDP)¹⁾ 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。しかし計算量が膨大になるため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⁹⁾。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が⁷⁾、2 本以上の時系列データには対応できないといえる。フーリエ変換によるスペクトルや、相互相関関数による類似部分は時間差を考慮していない。以上から、複数の時系列データから時間差を許容して類似部分を探し出す手法はあまりないといえる。また、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。

大量の時系列データを素早く処理するためには、並列化アルゴリズムを用いることが有効である。さらに並列化アルゴリズムのモデルが既に豊富な分野の手法を用いることができればよい。そこで、本論文では並列化アルゴリズムのモデルが豊富である同源性検索を用いて、複数の時系列データから類似部分を抽出するアプローチを提案する。しかし同源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、つまり文字列化が必要となる。再量子化の手法としては、SAX (Symbolic Aggregation approximation) と等間隔領域分割が挙げられる。本論文では、時系列データを再量子化、つまり文字列に変換し、同源性検索を用いて類似部分文字列を抽出することで、抽出した類似部分文字列から時系列データの類似部分を抽出する手法を提案する。また、時系列データの再量子化を 2 つの手法で行い、それらの手法の比較と評価を行った。

3 2 つの時系列データの類似部分の抽出方法

3.1 概要

本論文では、時系列データを再量子化する手法と、同源性検索による類似部分文字列の抽出の組み合わせによる時系列データからの類似部分抽出法を提案する。Fig.1 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、Fig.1 の時系列データ T1 は「BAABCCB」に T2 は「ABCCBAA」に変換される。再量子化の手法として

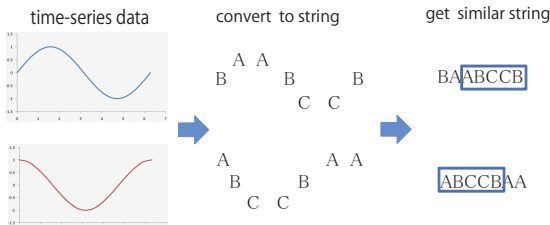


Fig.1 symbolization of sine wave

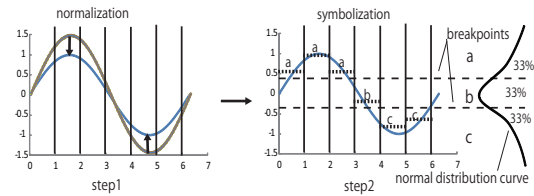


Fig.2 SAXによる文字列変換の流れ

は、SAX と等間隔領域分割がある。それぞれについては後述する。そして、時系列データを文字列に変換することによって同源性検索が適用できる。同源性検索は2つの文字列から最適なアラインメントを探す手法である。この手法により「BAABCCB」と「ABCCBAA」から「ABCCB」を類似部分として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同源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。また、SW法のgapパラメータを変化させることで時間伸縮を考慮した一致度探索が可能である。時系列データを文字列に変換して文字列検索アルゴリズムを適用する手法としては荒木らの手法⁷⁾が挙げられる。しかし、2つ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適用した例はない。

3.2 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り、そのために分割線を定める。例えば、時系列データ $T = \{1.0, 1.5, -0.5\}$ は「BAC」に再量子化される。2つの手法は異なる分割線を持つ。分割線の定め方としては、波形が正規分布になることを仮定し区分域を決めるSAXと、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る等間隔領域分割がある。また、1つの波形を何種類の文字で表現するかを定める必要があり、ある時系列データには最適な分割数があることが考えられる。分割数は任意で定めることができるが、本論文では予備実験により5分割にした。

3.2.1 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⁸⁾。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は正規分布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 $\{x(t)\} = \{x(1), x(2), x(3), \dots, x(M)\}$ を標準化すると $\{x(t)\} = \{(x(1) - \text{average}(x(t)))/\text{std}, \dots, (x(M) - \text{average}(x(t)))/\text{std}\}$ となる。averageは平均値、stdは標準偏差を意味する。SAXによる文字列変換の流れをFig.2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。標準化された時系列デー

タは正規分布する可能性が高く、正規分布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 $B = (\beta_1, \dots, \beta_n)$ 、分割記号数を α として Table.1 に示す。Fig.2において、最も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、最も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。

Table1 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	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15
β_2		0.43	0	-0.25	-0.43	-0.57	-0.67
β_3			0.67	0.25	0	-0.18	-0.32
β_4				0.84	0.43	0.18	0
β_5					0.97	0.57	0.32
β_6						1.07	0.67
β_7							1.15

3.2.2 Equal Intervals Area Division

等間隔領域分割は、波形が一様分布に従うと仮定して、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。この手法は、SAXとは異なる領域線を定め、データの標準化をする必要がない。等間隔領域分割による文字列変換の流れをFig.3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つの分割点を定める。よって最大値=1、最小値=-1であれば、領域幅は0.68となり、2つの分割点は(0.34, -0.34)となる。領域幅は(2)のように決定できる。MAXは時系列データの最大値、MINは時系列データの最小値、Wは領域の間隔、NUMは分割文字数

$$w = |MAX - MIN| \div NUM \quad (1)$$

$$\text{breakpoints} = MAX - W \times N \quad (1 \leq n \leq NUM) \quad (2)$$

4 同源性検索とSW法

同源性検索は、生物の同源性を調べるための文字列検索アルゴリズムであり、DNAやタンパク質の遺伝子構

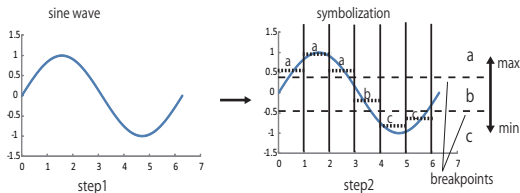


Fig.3 the concept of Equal Intervals Area Division

造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時に、ヒトがその配列と類似した遺伝子を持つかどうか」を調べることにより用いられる。しかし、DNA やタンパク質の遺伝子は大量の文字列で表現されるため、探索には長い計算時間が必要となる場合が多い。そこで並列化アルゴリズムを用いることによる計算時間の短縮化が有用だと考えられる。しかし、並列化アルゴリズムを用いるには GPU を使用するための専用プログラムを作成する必要があり非常に手間がかかる。バイオインフォマティクス分野では大量の文字列処理が求められるため、相同性検索のライブラリには、既に GPU を用いた並列プログラムが豊富に用意されている。このような相同性検索のライブラリを時系列データの探索にも適用することができれば、非常に高速な探索が可能になるのではないかと考えられる。

相同性検索には、精度を重視する Smith Waterman 法 (以下 SW 法)、速度を重視する FASTA, BLAST がある。アルゴリズムの速度では、BLAST や FASTA のほうが高速であるが、本論文では精度検証に用いた時系列データが限られていたため、SW 法を選択した。

SW 法は文字列の局所的アラインメント、つまり類似部分の抽出を行うアルゴリズムである。このアルゴリズムは、動的計画法の 1 種であり、全ての可能な長さのセグメントの比較を行うことで部分文字列の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。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 の文字列をアラインメントするとき、アルゴリズムのオーダーは $O(mn)$ である。また、SW 法の gap パラメータを変化させることで時間伸縮を考慮した一一致度探索が可能である。

4.1 parameters of score

SW 法には *match*, *mismatch*, *gap* の 3 つのパラメータがある。gap はスペースの発生を意味する。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。例えば、 $match=1, mismatch=-1, gap=-1$ は、標準的な設定であるが、これを $match=1, mismatch=-1, gap=-0.5$ とすることで、よりスペースが入りやすい文字列を抽出することができる。また、 $match=1, mismatch=-2, gap=-2$ とすることで、長さは短い但其の一致度の高い文字列を抽出できる。どのパラメータが最適であるかは、元デー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。本論文では $match=1, mismatch=-1, gap=-1$ を採用した。

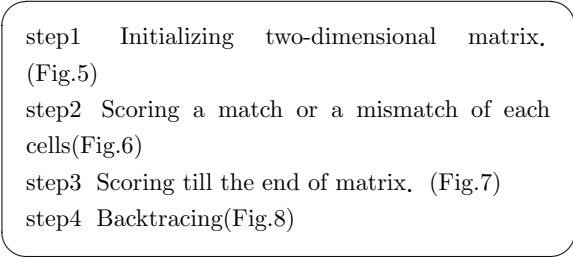


Fig.4 SW 法のアルゴリズム

4.2 algorithm

SW 法における局所的アラインメントの過程を以下に示す。まず、文字列テーブルを作成し、それぞれのセルにおける文字の一致や不一致に基づきスコアを計算する。そして最後のセルまで計算を行い、最も高いスコアを持つセルからトレースバックを行う。トレースバックとは最も高いスコアを持つセルからスコアが 0 のセルまで経路をたどることにより、文字列を取り出すことを意味する。Fig.4 にアルゴリズムの流れを示す。(3),(4) はスコアの計算式を表している。

Fig.5, Fig.6, Fig.7, Fig.8 は、文字列「BBC」と「CBC」の類似部分を SW 法で求めた時の様子を示している。まず、(3),(3) によって文字列テーブルの各セルにおけるスコアを計算する。各セルのスコアは、各セルの文字が一致か不一致か及び、左セル、左上セル、上セルのスコアによって定められる。例えば最初に計算されるセル (1,1) においてスコアは、セルの上の文字が「C」、左の文字が「B」と不一致なので、不一致の計算式が適用され、 $SW(1,1) = \max\{SW(0,1)-1, SW(0,0) - 1, SW(1,0) - 1, 0\} = 0$ となる。また、セル (1,2) においてスコアは、セルの上の文字が「B」、左の文字が「B」と一致なので、 $SW(1,2) = \max\{SW(0,2) - 1, SW(0,1) + 1, SW(1,1) - 1, 0\} = 1$ となる。このように最後のセルまで計算を行う。途中でスコアがマイナスになった場合は、そのセルのスコアを 0 にする。

類似部分を得るためには、「最大のスコア」のセルからスコアが 0 のセルまでトレースバックを行う。そのため、スコアの計算時にそれぞれのセルに対してどのセルから辿ってきたか目印をつける必要がある。もし、左セル、左上セル、上セルでスコアが重なっていれば、左上セル、左セル、上セルの順に優先して矢印をつけることとする。例えば、Fig.8 において、最大値が (3,3) に当たるのでトレースバックは (3,3), (2,2), (1,1) という経路をたどる。0 に辿りつけばトレースバックは終了し、辿ってきたセルの上と左の文字から類似部分を抽出する。(3,3) の上の文字は「C」、左の文字は「C」であり、(2,2) の上の文字は「B」、左の文字は「B」である。これにより「BBC」から「BC」が「CBC」から「BC」の部分文字列が抽出される。Fig.9 は SW 法のアラインメントの例を示す。

$$SW(y, x) = \max \begin{cases} SW(y-1, x-1) + match \\ SW(y-1, x) + gap \\ SW(y, x-1) + gap \\ 0 \end{cases} \quad (3)$$

$$SW(y, x) = \max \begin{cases} SW(y-1, x-1) + mismatch \\ SW(y-1, x) + gap \\ SW(y, x-1) + gap \\ 0 \end{cases} \quad (4)$$

x \ y	-	C	B	C
-	0	0	0	0
B	0			
B	0			
C	0			

Fig.5 initializing two-dimensional matrix

x \ y	-	C	B	C
-	0	0	0	0
B	0	(1,1)		
B	0			
C	0			

Fig.6 scoring a match or a mismatch of each cells

x \ y	-	C	B	C
-	0	0	0	0
B	0	0	1	0
B	0	0	1	0
C	0	1	0	2

Fig.7 scoring till the end of matrix

x \ y	-	C	B	C
-	0	0	0	0
B	0	0	1	0
B	0	0	1	0
C	0	1	0	2(max)

Fig.8 backtracing from the maximum value cell

"PELICAN" → "ELICAN"
"PAWHEAE" → "AW_HE"
"COELACANTH" → "ELACAN"
"HEAGAWGHEE" → "AWGHE"

Fig.9 Example of SW algorithm's alignment

5 数値実験による精度検証

5.1 目的

ここでは、数値実験により提案手法が有効に働くかどうかを検証する。実験の目的は、ある時系列データにたい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⁴⁾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。Fig.10~Fig.17には2本の時系列データがあり、類似部分が黒色の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。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 $match = 1$, $mismatch = -1$ and $gap = -1$ 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。抽出された類似部分が長くて類似度が低い場合と類似部分は短いが高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評価を行った。評価は抽出された類似部分の長さとして DTW 距離を用いて行う。

DTW は、時間伸縮を許容した時系列データの距離測定手法である。例えば、ある人はゆっくり話し、もう1人はより速く話している場合でも、スピーチ中の発話パターンを類似を検出することができる。DTW は、2つの時系列データの時間軸を伸縮して距離を算出するため、2人のスピーチ音声を類似として判断できる。DTW 距離は以下の様に定義される。

$$DTW(P, Q) = f(n_p, n_q) \quad (5)$$

$$f(i, j) = |p_i - q_j| + \min \begin{cases} f(i, j-1) \\ f(i-1, j) \\ f(i-1, j-1) \end{cases}$$

$$f(0, 0) = 0, f(i, 0) = f(0, j) = \infty \\ (i = 1, \dots, n_p, j = 1, \dots, n_q)$$

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。この誤差が低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よい類似部分を抽出しているのか判断できる。Table.2 にそれぞれの図における類似部分間の DTW 距離の結果を示す。

Table2 DTW distance of similar subsequences in each figure

Figure Name	Distance	Length	Dis/Len
Fig.10	41.6	114	0.36
Fig.11	53.3	118	0.45
Fig.12	32.2	112	0.29
Fig.13	16.3	67	0.24
Fig.14	10.6	63	0.17
Fig.15	23.2	222	0.10
Fig.16	1.1	1	1.09
Fig.17	42.6	4	10.65

5.4 結果

ここでは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。等間隔領域分割とSAXの両方で類似部分が抽出された時系列データをFig.10, Fig.11に示す。また、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時系列データをFig.12, Fig.13に示す。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時系列データをFig.14, Fig.15に示す。SAXでも等間隔領域分割でも抽出されなかった時系列データをFig.16, Fig.17に示す。

Fig.10, Fig.11を見ると、上下のデータのうち位相がずれた部分がうまく抽出されていることがわかる。このことから提案手法は位相のずれたデータに対して適用できることがわかる。Fig.12, Fig.13を見ると、saxではデータが中途半端に抽出されたことが分かるが、この原因として分割線が適切に定められなかったことが考えられる。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と考えられる。Fig.14, Fig.15を見ると等間隔領域分割の分割線とSAXの分割線が大きく異なっていて、等間隔領域分割は類似部分を抽出できていないことがわかる。Fig.16は外れ値があるため等間隔領域分割では抽出できず、Fig.17は標準化によりスケールが変化し分割線が異なったために適切な抽出が行えなかったことがわかる。

また、Table.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大体同じになることがわかった。つまり抽出された類似部分の誤差の程度は同じであり、できるだけ長く類似部分が抽出されたほうがよいではないかといえる。

これらの結果から、SAXと等間隔領域分割が適用できる時系列データには向き・不向きがあるといえる。データに外れ値がない場合は等間隔のほうが分割の幅を広くとることができるため、より適切な分割ができると考え

られる。逆に、外れ値がある場合はSAXを用いたほうがよいと思われるが、Fig.17より実際にはSAXを用いても必ずしも外れ値に対応できるわけではないことがわかった。Fig.12は等間隔領域分割が対応できているが、この結果はたまたま分割線がうまく設定されただけであり、SAXでも等間隔領域分割でも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と考えられる。2つの手法のどちらも位相がずれただけのデータに対してはうまく適用できると考えられる。これらからいえることは、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ということである。前処理を行うことで提案手法は時系列データの位相がずれた部分を類似部分として抽出できるといえる。

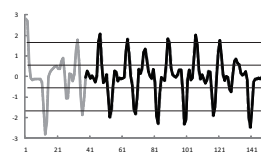


Fig.10 Leaf_all dataset:EIAD

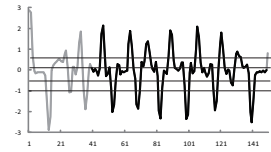


Fig.11 Leaf_all dataset:SAX

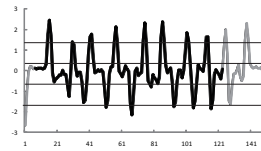


Fig.12 CBF dataset:EIAD

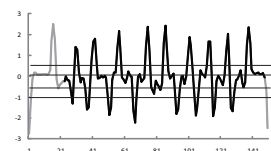


Fig.13 CBF dataset:SAX

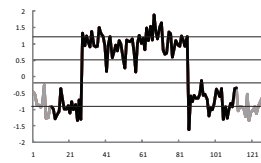


Fig.14 Trace dataset:EIAD

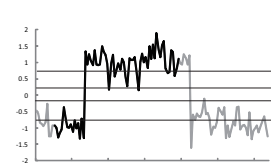
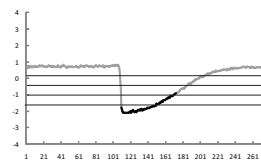
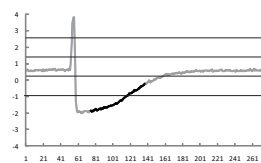


Fig.15 Trace dataset:SAX



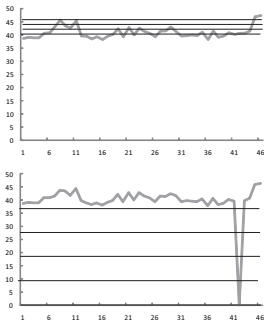


Fig.16 Q8Humid
dataset:EIAD

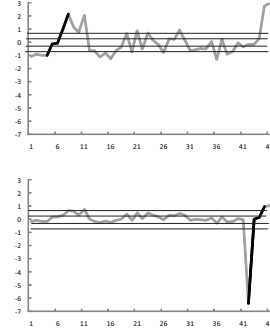


Fig.17 Q8Humid
dataset:SAX

6 結論と今後の課題

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。大量の時系列データを素早く探索するためには、並列化アルゴリズムを用いることが有効である。相同性検索のライブラリにはGPUを用いた並列プログラムが豊富に用意されているため、これを時系列データの探索に適用することを考えた。さらにSW法のgapパラメータを設定することで時間差を許容した類似部分の探索が可能になる。しかし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、つまり文字列化が必要となる。本論文では、SAXや等間隔領域分割を用いて再量子化を行い、それらの手法の精度検証を行った。精度検証の結果、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わかった。また、外れ値がない場合に等間隔領域分割はより適切な分割ができることがわかった。逆に、外れ値がある場合はSAXを用いたほうがよいと思われるが、実際にはSAXを用いても必ずしも外れ値に対応できるわけではないことがわかった。さらに、SAXでも等間隔領域分割でも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わかった。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といえる。

今後の課題としては最適な分割文字数やSW法のパラメータの設定、外れ値の除去や平滑化等の時系列データの前処理、2つ以上の時系列データの類似部分の抽出が挙げられる。

参考文献

- 1) 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
- 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) 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
- 4) 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/
- 5) "Tesla BIO Workingbench", http://www.nvidia.co.jp/object/tesla_bio_workbench_jp.html
- 6) 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
- 7) Abdullah Mueen, Eamonn Keogh, Qiang Zhu, Sydney Cash, Brandon Westover, "Exact Discovery of time-series Motifs", SDM 2009: 473-484
- 8) 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.
- 9) TOYODA Machiko, SAKURAI Yasushi, ICHIKAWA Toshikazu "Stream Matching based on Dynamic Programming"