System Introduction and Evaluation of Extracting Similar Subsequences from fNIRS Data

Takuma NISHII* , Tomoyuki HIROYASU** , Masato YOSHIMI*** ,
Mitsunori MIKI*** and Hisatake YOKOUCHI**

(Received October 25, 2011)

One of non-invasive brain functional mapping equipments fNIRS (functional Near-Infrared Spectroscopy) is known for its practicality. One of the characteristics of fNIRS is that fNIRS derives enormous time series data in each experiment, so that it is hard to analyze these data effectively. In this paper, we introduce the novel algorithm which can extract similar subsequence of fNIRS data. In the proposed algorithm, the conventional homology search and Smith Waterman method are applied. Since there are several software libraries of these algorithms, the proposed algorithm is not only useful for getting satisfactory results but also effective for drawing these results quickly. In this paper, the fNIRS data analysis system is illustrated, where the proposed algorithm has been implemented. Using the proposed system, the effectiveness of the proposed algorithm is discussed, and the response time of the system is estimated and illustrated.

Key words ： functional Near-Infrared Spectroscopy, Smith Waterman algorithm, homology search, similar subsequence

キーワード ： fNIRS，Smith Waterman 法，相同性検索，類似部分

Smith Waterman 法を利用した fNIRS データの類似部分抽出システムの提案と評価

西井琢真・廣安知之・吉見真聡・三木光範・横内久猛

1. はじめに

近年，脳血流の変化を測定することにより非侵襲的に脳機能マッピングを行う fNIRS(functional Near-Infrared Spectroscopy) や fMRI(functional Magnetic Resonance Imaging) が注目を集めている [1)]．これらの装置は，脳機能の解明に役立ち，種々の病理の判定や生体信号によるコンピュータ操作などに利用されている [2)]．しかし，これらの装置の性能が向上し出力される時系列データ量が増大すると解析者が効率的にデータを解析できないという問題も生じる．効率的にデータを処理するための課題はいくつか存在するが，その 1 つに解析者がどのデータのどの部分に着目すれば良いのかという問題がある．

* Department of Knowledge Engineering and Computer Sciences, Doshisha University, Kyoto
Telephone:+81-774-65-6130, Fax:+81-774-65-6780, E-mail:dtk0748@mail4.doshisha.ac.jp
** Department of Faculty of Life and Medical Sciences, Doshisha University, KyotoTelephone:
+81-774-65-6932,Fax:+81-774-65-6780,E-mail:thiroyas@mail.doshisha.ac.jp ,hyokouch@mail.doshisha.ac.jp
*** Department of Knowledge Engineering and Computer Sciences,Doshisha University, Kyoto Telephone:
+81-774-65-6930,Fax:+81-774-65-6796,E-mail:mmiki@mail.doshisha.ac.jp ,myoshimi@mikilab.doshisha.ac.jp

本研究ではこの問題を解決するために複数の時系列データの中から特徴的な部分を抽出するアルゴリズムを提案し，システムを構築することで解析者の負担を軽減する．

具体的には，Smith Waterman 法 [3)4)] を利用した fNIRS データの類似部分抽出アルゴリズムの提案し，それを実装したシステムの評価を行う．Smith Waterman 法を利用した fNIRS データの類似部分抽出アルゴリズムは，fNIRS の出力データする時系列データに対して，時系列データの再量子化を行ない，Smith Waterman 法を適用することで，時系列データからの類似部分を抽出する．fNIRS は出力データが大量であるため，解析に用いられるアルゴリズムは高速である必要がある．そのため，時系列データからの類似部分の抽出については，バイオインフォマティクスの分野で盛んに扱われる Smith Waterman 法に注目した．Smith Waterman 法は，アルゴリズムの並列性が高く，CPU マルチスレッドプログラミングや GPU を用いた高速な探索を期待できる．

また，そのアルゴリズムを適用したシステムのインターフェース紹介を第 5 章で行う．インターフェースは，ユーザが注目すべき部分を分かりやすく表示することを目的としている．これにより解析者は低負担で注目部位を特定することが可能である．解析者がシステムを快適に使うためには，高速なレスポンスが必要とされる．第 6 章では実装システムに必要な機能を実現するための処理時間を計測した．第 7 章では，解析システムを用いて実際の fNIRS の出力データを解析した結果を表示した．

2. 関連研究

脳機能イメージング装置の解析ソフトとしては，日立メディコ社製の ETG-7100 における Wave Analysis ソフト [5)] やスペクトラテック社製の Spectratech OEG-16 の解析ソフトなど装置に付属するソフトが挙げられる．これらのソフトを用いることで生データのグラフ化や簡単な処理は可能である．また，Matlab を用いた解析ソフト「NIRS-SPM」[6)] や Source Signal Imaging 社製の脳波解析プログラム「EMSE」[7)] を用いることでより高度な波形処理や脳の 3D モデリングが可能である．しかしながら，解析すべきデータのうち，どこに着目するべきかを提示する事を目的としたソフトはほとんどないといえる．

また，本論文では出力データを高速に処理するメソッドとして Smith Waterman 法を用いて，そのパラメータ調整を行った．SW 法はアルゴリズムの並列性が高く，GPU を用いた高速実行のための様々な実装が試みられている [8)9)10)11)]．

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．

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部分を最適化する．類似度は文字列テーブルのスコアによって評価される．Fig.1 に SW 法で抽出された類似部分文字列の例を示す．Fig.2 に文字列テーブルを示す．文字列テーブルでは文字列 X のそれぞれの文字が行に，文字列 Y 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．

Fig. 1. String sequence alignment using SW.

x | y	_	C	B	C
_				
B				
B				
C				

Fig. 2. Searching matrix of SW.

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．ここではパラメータを $match = 1, mismatch = -1, gap = -1$ とする．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．$mismatch$ 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．

3.2 アルゴリズム

SW 法のアルゴリズムの流れを以下に示す．

Step1: 文字列テーブルを作成し，それぞれの文字列を列と行に割り当て初期化を行う．(Fig.3(a))

Step2: それぞれのセルにおけるスコアを文字の一致や不一致及び式 (1),(2) に基づき計算する．(Fig.3(b))

Step3: テーブルの終了までスコアを計算する．(Fig.3(c))

Step4: 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．(トレースバック)(Fig.3(d))

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

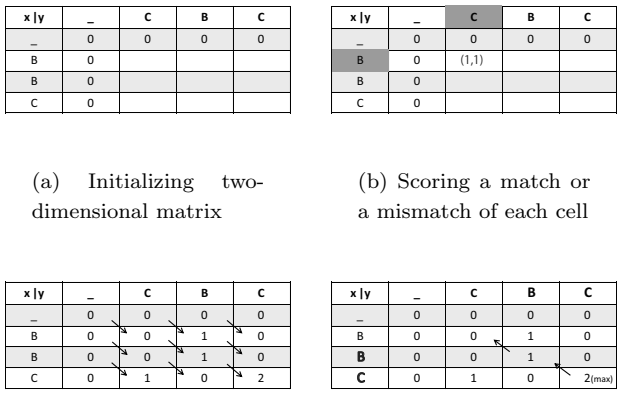
$$SW(y,x) = max \begin{cases} SW(y-1, x-1) + mismatch \\ SW(y-1, x) + gap \\ SW(y, x-1) + gap \\ 0 \end{cases} \tag{2}$$

Fig.3(a)〜Fig.3(d) は，文字列“BBC”と“CBC”の類似部分を SW 法で求めた時の過程を，式 (1)，(2) はスコアの計算式を表している．例えば，Fig.3(a) において最初に計算されるセル (1，1) のスコアは，セルの上の文字が”C”，左の文字が”B”と不一致なので，式 (2) が適用され，式 (3) となる．

$$SW(1,1) = max\{-1, -1, -1, 0\} = 0 \tag{3}$$

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，式 (1) が適用され，式 (4) となる．

$$SW(1,2) = max\{1, -1, -1, 0\} = 1 \tag{4}$$



(a) Initializing two-dimensional matrix

(b) Scoring a match or a mismatch of each cell

(c) Scoring until the end of the matrix

(d) Backtracing from maximum cell score

Fig. 3. The flow of SW algorithm(example of BBC and CBC).

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 になる．Fig.3(c) の状態から類似部分を得るために，”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

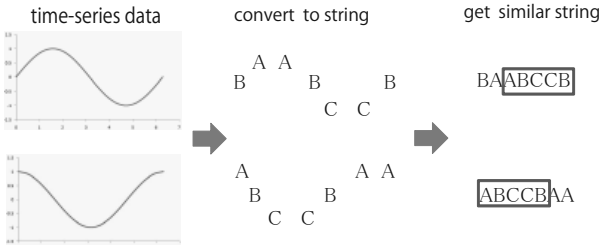


Fig. 4. The outline of proposed method.

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，任意で優先順位を定める必要がある．ここでは，左上セル，左セル，上セルの順に矢印をつけることとする．Fig.3(d) 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．

4. SW 法を用いた 2 つの時系列データの類似部分の抽出方法

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用いた [12]．Fig.4 に抽出手法の流れを示す．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.4 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(Symbolic Aggregation approXimation) や等間隔領域分割がある [13]．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．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．本論文では，本手法を用いて抽出された部分を類似部分として定義した．fNIRS の出力データに対して提案手法を適用した例を Fig.5(a)，Fig.6(b) に示す．

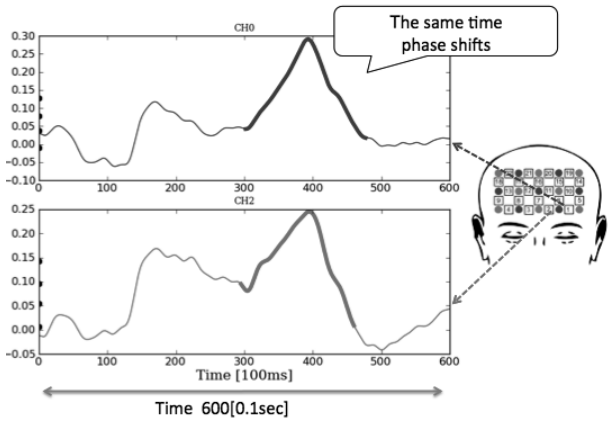


Fig. 5. Sample extracted data using the proposed algorithm (Time phase shifts in the two data are the same).

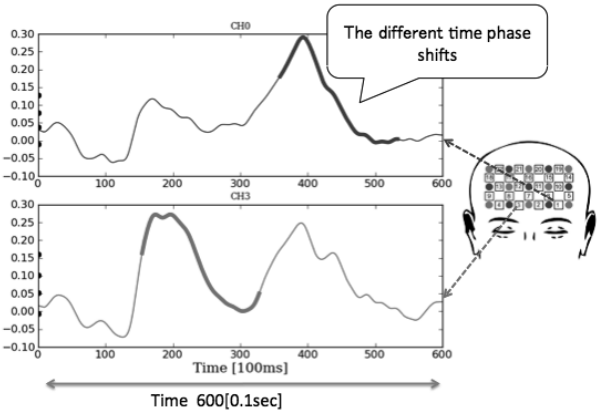


Fig. 6. Sample extracted data using the proposed algorithm (Time Phase shifts in the two data are the different).

5. 解析システムの提案

5.1 概要

このシステムの目的は，データ解析において従来は注目されなかった脳部位データのどれに着目すべきかを提示することで解析者に新たな知見を与えることである．現在，fNIRS の解析者は注目した脳部位のデータに対してのみデータ処理を行っており，その方法では注目している脳部位以外に重要な要素がある場合にそれを見落としてしまう可能性がある．これを解決するためには，ある特徴的な波形に類似した波形が他の脳部位のどこにあるのかを特定することが有効なのではないかといえる．これにより，解析者は低負担で着目するべき部位を見つ

けることが可能である．

5.2 機能

ここでは作成するインターフェースの機能について説明する．インターフェースに求められる機能は以下の 3 つである．

機能 1：fNIRS のファイルを読み込み，CH データ一覧を表示

機能 2：CH を 1 つ選択し，その CH データに類似するデータを表示

機能 3：ある CH データの範囲を指定し，それと類似するものを他の CH データ群から表示

6. SW 法の処理速度性能

6.1 概要

第 5 章で述べた解析システムの機能を実現するためには，提案手法を何度も繰り返す必要がある．この手法において，特に処理に時間がかかるのは SW 法の部分であると予想される．そのため，インターフェースを実現するための処理時間を計測し，複数のスレッドを利用して高速化することを検討する．

また，解析者が快適に作業するためには，クリックなどのアクションをしてから 1 秒以内で反応が返ってくることが必要であると考えた．よって解析システムにおいてそれを実現する必要がある．

6.2 実験目的

ここでは，マルチコア CPU 環境においてマルチスレッドを利用した SW 法の速度計測を行った．マルチスレッドの SW 法を実行するためには，スレッド数と部分ブロックのサイズをパラメータとして指定する必要がある．パラメータの設定によっては高速化がうまくいかない場合も考えられるため，事前に最適なパラメータを把握する必要があった．

また，解析システムにおいて実際に機能 2，3 がどれくらいの処理時間になるかを計測し，高速なレスポンスを実現できるかどうかを把握する必要があった．

6.3 実験結果

文字列の長さを 256，512，1024，2048，4096 と変化させ SW 法の処理時間を計測した．Fig.10 から Fig.13 にその結果を示す．それぞれの実行計測は 3 回ずつ行い，実験結果にはその中央値を用いた．使用したマシンの環境を Table.3 に示す．それぞれの文字列サイズにおける最適パラメータを Table.2 に示す．インターフェースの機能 1〜3 に必要な SW 法の繰り返し回数と，文字列の長さが 600 のときの処理時間を Table.1 示す．fNIRS のサンプリングレートは 0.1 秒で有ることが多いため，1 秒あたり 10 文字となる計算である．なお，文字列の内容によって処理時間が変化することはない．

Table 1. The number of times of executing SW algorithm for functions (1) - (3).

Function name	The number of execution of SW algorithm	Calculation time[sec]
F1	0	0
F2	24	0.165
F3	24	0.165

Table 2. Optimum parameters for each table size.

Table size	The number of threads	Sub-block size
256	4	32
512	4	32
1024	4	128
2048	8	32
4096	8	128

6.4 評価

Fig.10 から Fig.13 より，パラメータの設定によっては処理時間が数倍異なることが分かった．このことから処理の高速化のためには，パラメータ調整も必要であるということがいえる．しかし，Table.2 にある最適パラメータはマシン環境により変化することが考えられる．

また，Table.1 より，SW 法の処理時間は機能 2，3 を実行すると 0.165 秒となり，1 秒以内の高速なレスポンスが可能であることが分かった．

Table 3. Operating environment.

OS	Ubuntu10.10 64bit
CPU	Intel Core i7-2600 4cores 3.40GHz
RAM	8GB

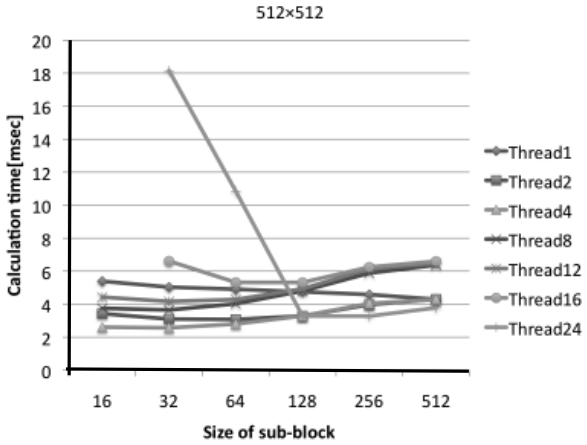


Fig. 10. Calculation time with along to the number of threads (table size 512 × 512).

7. 構築したシステムによる fNIRS データの解析

解析システムを用いて，実際の fNIRS 出力データを解析した．用いた実験データは，タスクが「ストループ効果に関するタスク」，実験時間は 60sec，サンプリングレートは 0.1sec である．この場合，文字列の長さが 600 になることから，Table.2 より，スレッド数を 4，サブブロックサイズを 32 にすることが適切だと考えた．SW 法の処理時間は，Table.1 と同様となった．5 章で説明したシステムの機能を Fig.7〜Fig.9 に示す．ある 1 つのプローブにおける解析システムの処理結果を Fig.14 に示す．Fig.14 を見ると，CH1〜24 において CH2 を基準としたときの類似部分が現れており，注目部位の選択に役立つことがわかる．

8. まとめと今後の課題

本稿では，大量の実験データを出力する fNIRS のデータ解析のためのアルゴリズムとシステムの提案と評価を行った．相同性検索を用いて fNIRS データから類似部分を抽出するアルゴリズムを紹介し，それを実現したシステムの機能とレスポンスについて述べた．その結果，現

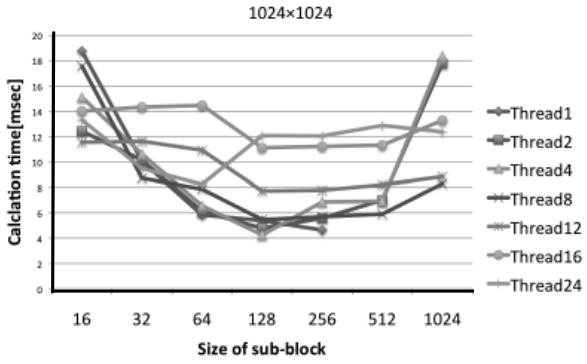


Fig. 11. Calculation time with along to the number of threads (table size 1024 × 1024).

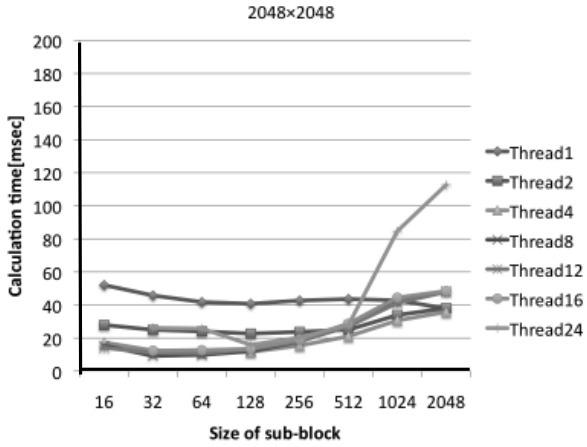


Fig. 12. Calculation time with along to the number of threads (table size 2048 × 2048).

在想定している機能についてシステムのレスポンスは十分高速である事がわかった．また，実際の fNIRS の出力データに対してシステムを適用し，類似部分がどのように表示され，解析者の負担軽減に繋がるかについて述べた．

今回作成したインターフェースは，ユーザが基準となる CH を選択し，それに類似したデータを表示するというものであった．今後は，基準となる CH の自動選択機能も検討したい．

また，今回は提案手法を用いて抽出された部分を類似部分として定義した．この手法を用いて抽出される類似部分は，時系列データの再量子化や SW 法を用いて抽出された部分文字列に基づいているため，ユークリッド距離や相関値のように，2 つの部分時系列データの類似度を数学的に定義することはできない．類似部分は SW 法の

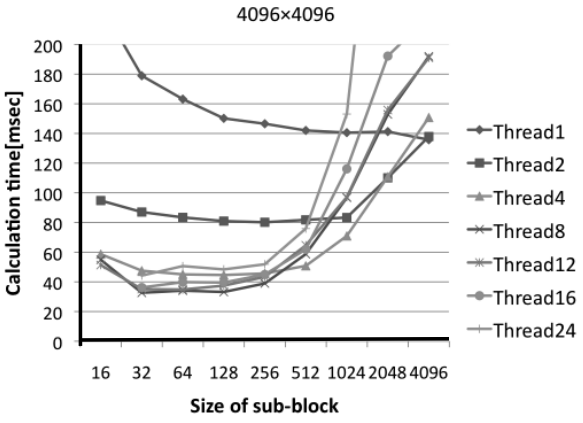


Fig. 13. Calculation time with along to the number of threads (table size 4096 × 4096).

スコアパラメータや再量子化の手法によっても変化するため，これらのパラメータを変化させることで提示される類似部分も変化していく．将来的には，fNIRS のデータの特性に合わせた適切なパラメータを発見したいと考えている．

本研究は，2010 年度同志社大学理工学研究所研究助成金によって行った．ここに記して謝意を表する．

参 考 文 献

1) James C. Eliassen; Erin L. Boespflug; Martine Lamy; Jane Allendorfer; Wen-Jang Chu; Jerzy P. Szaflarski. Brain-Mapping Techniques for Evaluating Poststroke Recovery and Rehabilitation. Neuroplasticity: Changing Minds and Changing Function. 2008, vol.15, no.5, p.427-450.

2) Xu Cuia; Signe Braya; Daniel M. Bryanta; Gary H. Gloverc; Allan L. Reissa. A quantitative comparison of NIRS and fMRI across multiple cognitive tasks. NeuroImage. 2011, vol.54, no.4, p.2808-2821.

3) T.F.Smith; M.S.Waterman. Identification of Common Molecular Subsequences. J.Mol.Bwl.1981, vol.147, p.195-197.

4) Peiheng Zhang; Guangming Tan; Guang R. Gao. Implementation of the Smith-Waterman algorithm on a reconfigurable supercomputing platform. HPRCTA '07 Proceedings of the 1st international workshop on High-performance reconfigurable computing technology and applications. doi:10.1145/1328554.1328565.

5) 株式会社日立メディコ. "Optical Topography ETG-7100". http://www.hitachi-medical.co.jp/product/opt/etg/func.html

6) BiSPL(Bio Imaging Signal Processing Lab). "NIRS-SPM".http://bisp.kaist.ac.kr/NIRS-SPM.html

7) Source Signal Imaging Inc. "EMSE 脳波解析プログラム". http://www.miyuki-net.co.jp/jp/product/emse.htm

8) Keisuke Dohi; Khaled Benkrid; Cheng Ling. Highly efficient mapping of the Smith-Waterman algorithm on CUDA- compatible GPUs. ASAP'10. 2010, vol.36, p29-36.

9) Yongchao Liu; Douglas L Maskell; Bertil Schmidt. CUDASW++: optimizing Smith-Waterman sequence database searches for CUDA-enabled graphics processing units. BMC Research Notes. 2009. doi:10.1186/1756-0500-2-73.

10) Svetlin A Manavski; Giorgio Valle. CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment. BMC Bioinformatics. 2008. doi:10.1186/1471-2105-9-S2-S10

11) nVIDIA. "Tesla BIO Workingbench - 実用化する新科学" http://www.nvidia.co.jp/object/tesla_bio_workbench_jp.html

12) Takuma Nishii; Tomoyuki Hiroyasu; Masato Yoshimi; Mitsunori Miki; Hisatake Yokouchi. Similar subsequence retrieval from two time series data using homology search. Systems Man and Cybernetics.2010, p.1062-1067. doi:10.1109/ICSMC.2010.5641809

13) 廣安 知之，西井 琢真，吉見 真聡，三木 光範，横内 久猛. 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, 数理モデル化と問題解決（MPS）. 2010, vol.80, no.24

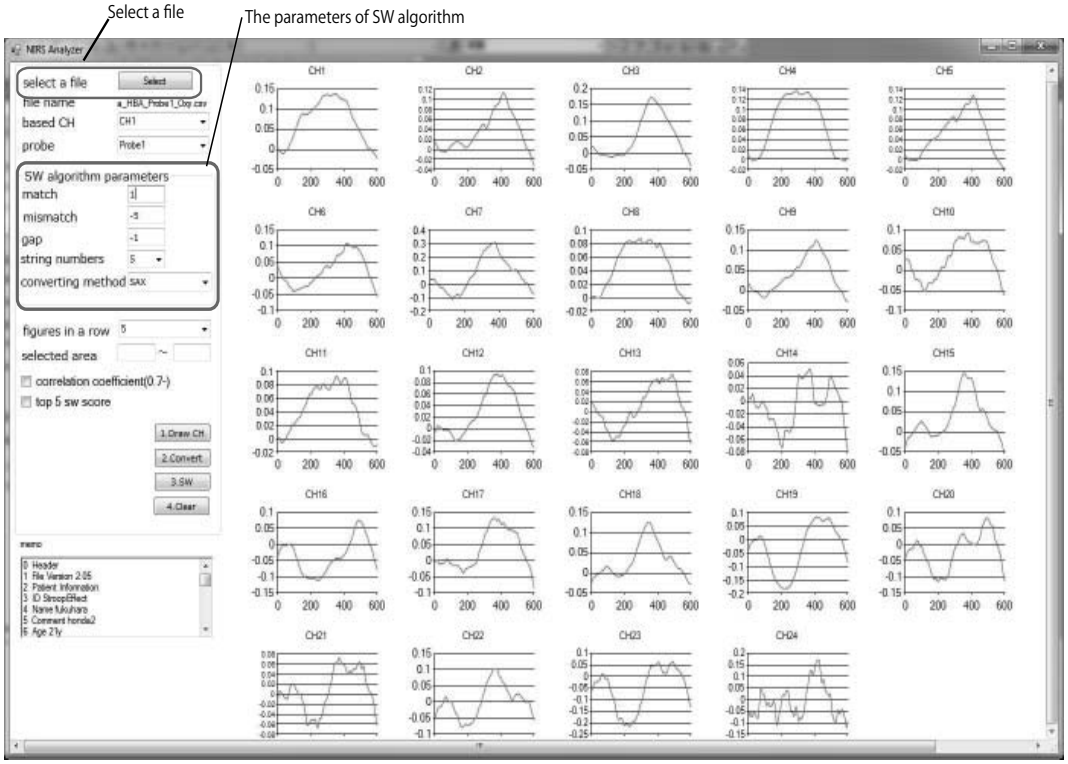


Fig. 7. Reading fNIRS output data and showing all the CHs data.

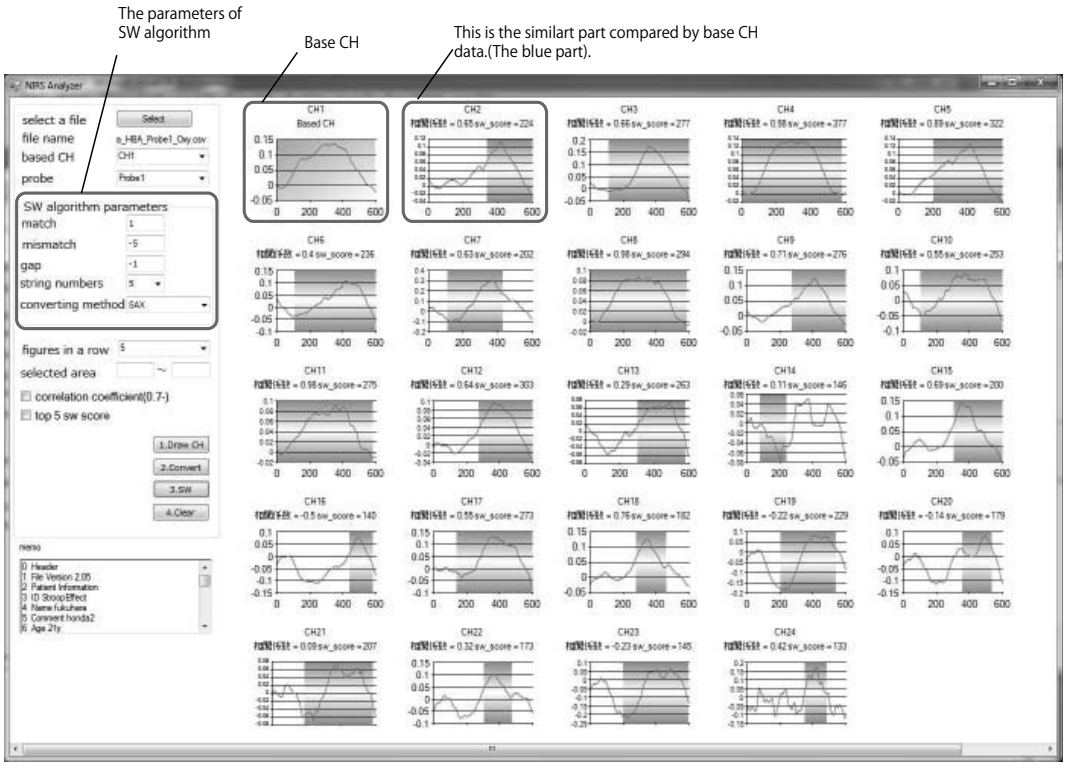


Fig. 8. Selecting a CH and then showing the related data.

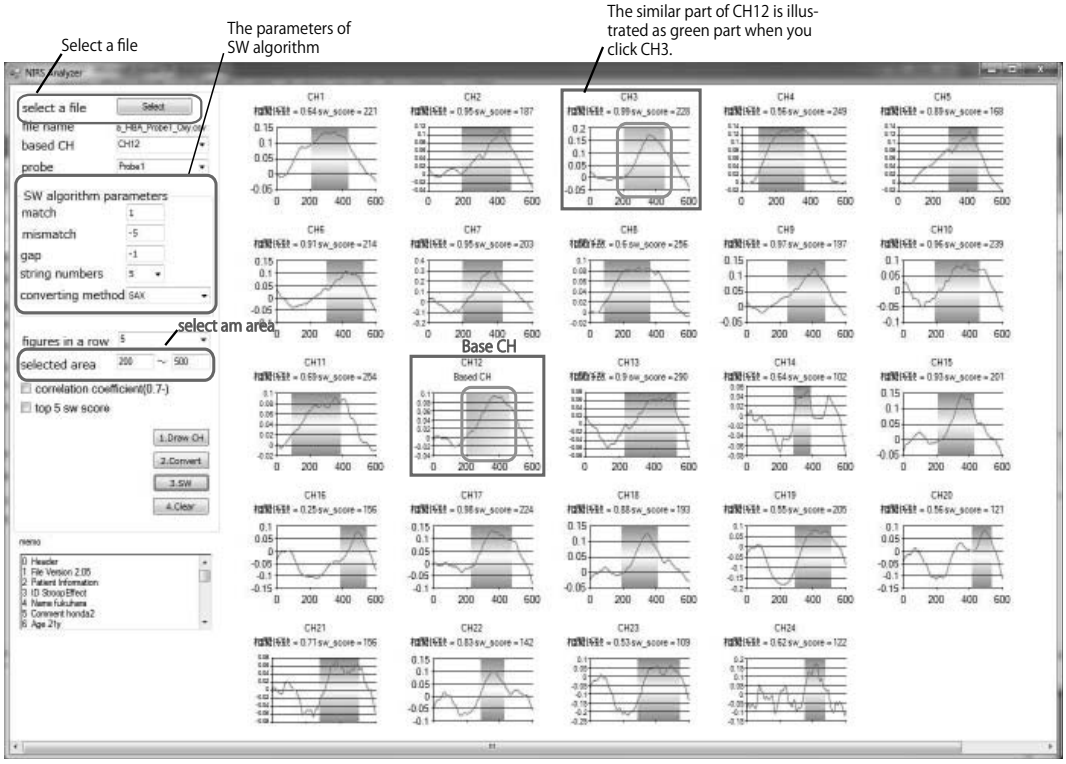


Fig. 9. Selecting a CH and its part and then showing the related data.

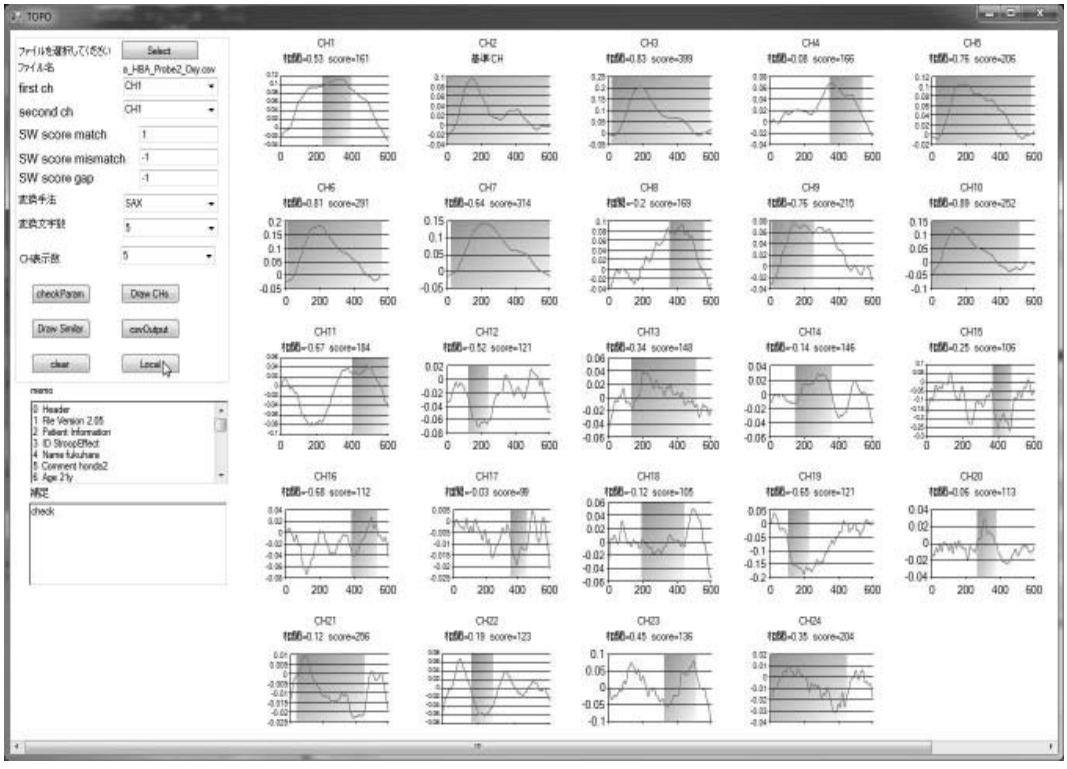


Fig. 14. Output data of fNIRS(CH2 and its similar part).