2I3-OS-10a-3

PU Classification による ナノデバイス出力信号からの DNA 塩基パルスの抽出

Extraction of DNA base pulses from nano device output signals by PU Classification

吉田 剛 Takeshi Yoshida　大城 敬人 Takahito Ooshiro　鷹合 孝之 Takayuki Takaai　谷口 正輝 Masateru Taniguchi　鷲尾 隆 Takashi Washio

大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research Osaka University

Needs of single molecule measurement technology are increasing. However, the measurement always involves a problem of noise contamination, which is very significant in the nanoscale measurement. We attempted to remove noise from measured data by using PU Classification (Classification from Positive and Unlabeled Examples). We applied this technique to a single molecule DNA identification. We report the significant improvement in the accuracy of monomer identification using this noise removal principle.

1. はじめに

ナノセンシング，微量計測，量子計測など先端センシングデバイス開発分野では，微細・微量な対象を計測するためのデバイスが次々と開発されつつあり [Rosenstein 12]，この分野で日本は世界をリードする立場にある．しかしながら，これら多くのデバイスは，計測系や計測対象が微小であるが故に，対象の部分的な情報のみを出力し，かつ出力が熱雑音や量子ノイズなどの影響を受けることが多い．

そのために，実用的に高精度な計測結果を得るには，出力情報に基づき時空間や特徴空間上で計測値の適切な推定を行い，不要なノイズを除去し有用な情報のみ取り出すということが不可欠である．特に 1 分子計測技術を用いた次々世代 DNA シークエンサーでは，このようなノイズ除去により塩基識別精度を向上させるニーズは顕著である．次々世代シークエンサーでは DNA の 1 塩基分子を電流パルスとして計測するが，計測されたパルスには塩基分子由来のものだけではなく電極表面の金属原子のゆらぎや不純物による電流パルスも含まれている．これらのノイズパルスのために，本来は塩基由来であるパルスを見逃したり，逆にノイズパルスであったのに塩基分子パルスが計測されたと誤判定する可能性が起こり，その結果，DNA 塩基分子の識別が困難になる．

そこで，計測されたパルスの集合から適切にノイズパルスを除去し，高精度に塩基種類の識別を行うことを本研究の目的とする．そのために我々は「PU Classification (Classification from Positive and Unlabeled Examples)」と呼ばれる手法を用いて対象とする計測パルス集合からノイズパルスの除去を行い，その後に対象計測パルスの分類実験を行った．この結果，PU Classification でノイズ除去をしなかった場合と比べて大幅に識別精度が向上した．この手法と結果について報告する．

2. ナノ電極ギャップと測定実験

次々世代 DNA シークエンサーとして期待されるデバイスであるナノギャップ電極 [Tsutsui 10] は，機械的破断接合と呼ばれる手法を用いて作成されたごく微細な隙間をもつ電極ギャップである．この電極ギャップに一定の電圧をかけると，ギャップ付近を物質が通過する際に量子力学的トンネル効果による電流（トンネル電流）が流れる．このトンネル電流が，物質が通過した瞬間のパルス電流として計測される．このナノギャップ電極によるトンネル電流パルスを計測することにより，DNA 塩基分子の種類を 1 分子単位で識別する研究が進んでおり [Tsutsui 11]，既存技術では困難であったペプチドのアミノ酸配列や，疾病マーカーとなる修飾アミノ分子の識別などが可能になってきている [Ohshiro 12, Ohshiro 14]．

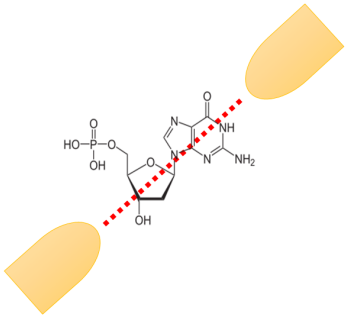

図 1: ナノギャップ電極の概念図

我々は，このような約 1nm の電極ギャップをもつナノギャップ電極を用いて，電極付近を通り過ぎる 1 分子に流れるトンネル電流パルスを計測した．計測分子として，人工核酸塩基であるジチオフェンウラシル誘導体（以下では BithioU），TTF ウラシル誘導体（TTF）を用いた．これらの分子は，識別を容易にするためにエピジェネティック部位（DNA メチル化などが起こる後天修飾部位）を化学的に修飾したものである．

測定により得られたパルス波形の一例を図 2 に示す．このようなパルス波形について解析を行い，ノイズパルスを除去した後に 2 種塩基 BithioU，TTF のパルスの分類を試みた．これらの人工塩基の識別が可能となれば，DNA 記憶媒体の情報圧縮技術や，人工塩基対を用いた医薬品創薬などへの応用が期待できる．

3. パルス波形を用いた属性ベクトル

機械学習分類器によって塩基種類を識別するためには次元のそろった属性ベクトルを入力として用いなければならないが，抽出したパルスは波長も波高もまちまちであるため，以下に述

連絡先: tyoshida@ar.sanken.osaka-u.ac.jp

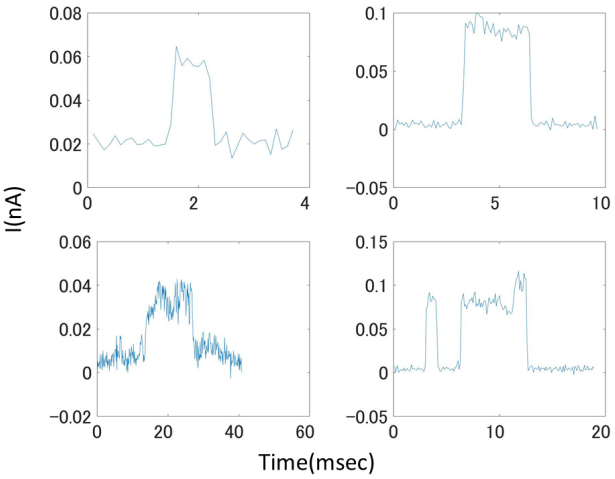

図 2: ナノ電極により計測されたパルス波形の例

べるような，一種の粗視化を施し，パルス波形を反映した属性ベクトルを作成する前処理を行う．

3.1 波高ベクトル

図 3 左図のような計測パルス波形について，波長方向に d_h 分割し各分割区分ごとに計測電流値の平均値を計算する．これを d_h 次元の属性ベクトルとする．この属性ベクトルは，波高方向に規格化したもの，しないものの 2 種類を作成する．

3.2 波長方向時間ベクトル

図 3 右図のように，パルスのピーク前後で計測電流値を 2 つのグループに分けた上で，波高方向に d_w 分割する．するとパルスの計測電流値は $2d_w$ のグループに分割されるが，この分割区分ごとに，パルス開始時点からのステップ数の平均値を算出し，これらの値を成分としてもつ $2d_w$ 次元の波長方向時間ベクトルを作成する．また，パルス開始時点から終了時点までの時間を 1 とする規格化を施した規格化波長方向時間ベクトルも作成する．

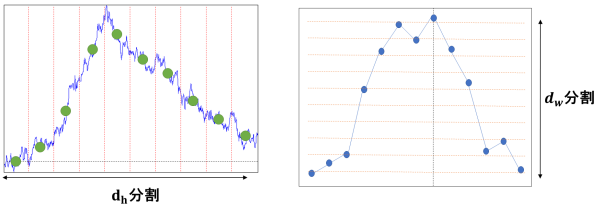

図 3: 波高ベクトル・波長方向時間ベクトル

3.3 波高ベクトルと波長方向時間ベクトルの組合せ

以上の波高ベクトルと波長方向時間ベクトルに加え，これらを単に連結した属性ベクトルも作成する．

1. パルスピーク値を 1 に規格化した波高ベクトル
2. 規格化しない波高ベクトル
3. パルス波長時間を 1 に規格化した波長方向時間ベクトル
4. 規格化しない波長方向時間ベクトル
5. 1 と 2 を連結した $d_h + 2d_w$ 次元ベクトル
6. 1 と 4 を連結した $d_h + 2d_w$ 次元ベクトル
7. 2 と 3 を連結した $d_h + 2d_w$ 次元ベクトル
8. 2 と 4 を連結した $d_h + 2d_w$ 次元ベクトル

1 つのパルス波形から以上の 8 通りの属性ベクトルを作成し，これらの識別精度の比較を行う．属性ベクトル作成時の分割数は予備解析を行った上で，一律に $d_h = 10$，$d_w = 5$ とした．

4. PU Classification （PUC）

PU Classification[Elkan 08] は，正例とラベルなしデータから学習し，正例/負例の 2 値分類をするための半教師あり学習アルゴリズムの一種である．以下で簡単にこのアルゴリズムについて述べる．

4.1 手法 1

事例 x（入力）を，3 節で述べたあるパルス波形に関する属性ベクトルとし，$y \in \{0,1\}$ をそのクラスラベル，事例にクラスラベルが付けられているか否かを示すフラグを $s \in \{0,1\}$ とする．入力事例の集合のなかで，正例（$y = 1$）の一部のみがラベルされており（$s = 1$），他の正例と全ての負例（$y = 0$）はラベルされていない（$s = 0$）．すなわち，サンプルが負例であるならばラベルされている確率はゼロであり，$p(s = 1|x, y = 0) = 0$ である．

このような事例集合を通常の 2 値分類器の学習アルゴリズムの入力とし，サンプルがラベルされている確率 $g(x) = p(s = 1|x)$ を求める．しかし，本来求めたいものは $g(x)$ ではなく $p(y = 1|x)$ である．そこでさらに，以下の補正を加える．全事例集合において，サンプルがラベルされている確率は

$$\begin{aligned} g(x) &= p(s = 1|x) \\ &= p(y = 1 \wedge s = 1|x) \\ &= p(y = 1|x)\, p(s = 1|y = 1, x) \\ &= p(y = 1|x)\, p(s = 1|y = 1) \end{aligned}$$

であるから，$c = p(s = 1|y = 1)$ とすると，サンプルが正例である確率は $p(y = 1|x) = g(x)/c$ と与えられる．ただし，正事例集合中でラベル付けされる確率が一様ランダム，すなわち x によらず $p(s = 1|y = 1, x) = p(s = 1|y = 1) = c$ は一定値であると仮定している．

ここで，c は以下のようにして推定する．正事例集合中で一様ランダムにラベル付けされているならば，$g(x)$ は x が正例である場合には正例中に含まれるラベル付き事例集合の割合に一致し，$g(x) = p(s = 1|y = 1) = c$ となる．そこで通常の 2 値分類器で求めた $g(x)$ を用いて，正事例であるラベル付き事例集合 L 中の平均 $\sum_{x \in L} g(x)$ として c を推定する．以上を「手法 1」と呼ぶことにする．

4.2 手法 2

ラベル付き事例は全て正例であるが，ラベルなし事例は正例，負例のいずれの可能性もある．ラベルなし事例が正例である確率を $w(x)$ とすると，その負例である確率は $1 - w(x)$ である．そこで，ラベルなし事例をすべて 2 倍に複製し，一方を正例として扱い，もう一方を負例として扱う．正例として扱うラベルなし事例 x には 重み $w(x)$ を与え，負例として扱うラベルなし事例 x には重み $1 - w(x)$ を与える．ラベル付き事例はすべて正例であるから，重み 1 で正例として扱う．これら重み付き事例集合を学習データとして分類器を作成する．

ラベルなし事例が正例である確率 $w(x)$ は以下のように求める．c と $g(x) = p(s = 1|x)$ は手法 1 により得られているとする．

$$
\begin{aligned}
w(x) &= p(y = 1|x, s = 0) \\
&= \frac{p(s = 0|x, y = 1)\, p(y = 1|x)}{p(s = 0|x)} \\
&= \frac{[1 - p(s = 1|x, y = 1)]p(y = 1|x)}{1 - p(s = 1|x)} \\
&= \frac{(1 - c)p(y = 1|x)}{1 - p(s = 1|x)} \\
&= \frac{(1 - c)p(s = 1|x)/c}{1 - p(s = 1|x)} \\
&= \frac{1 - c}{c} \frac{p(s = 1|x)}{1 - p(s = 1|x)} \\
&= \frac{1 - c}{c} \frac{g(x)}{1 - g(x)}
\end{aligned}
$$

5. 識別実験と結果

我々は実験的に得た計測パルス集合から，1) まず前処理として PUC を用いてノイズ由来のパルスを除去して塩基由来のパルスのみの集合を取得し，2) そのようにして得た塩基由来のパルス集合に対して塩基種別の識別精度を評価した．以下に手順と結果を述べる．

5.1 PUC によるノイズ除去

まず，あらかじめ塩基（BithioU，TTF）を含んでいない溶媒のみに対して，ナノギャップ電極により計測したトンネル電流パルスを取得しておく．このパルス集合は塩基とは関係のないノイズ由来のパルスであり,「ノイズパルス集合」と呼ぶことにする．

次に，溶媒に塩基 BithioU を混入したものについて計測した電流パルスを取得する．TTF についても同様に取得する．このパルス集合には，塩基由来の「塩基パルス」とノイズパルスの双方が含まれている．そこでこれを「塩基+ノイズパルス集合」と呼ぶことにする．

ノイズパルス集合中のパルスは必ずノイズパルスであるので，それを正事例集合とみなし，塩基＋ノイズパルス集合中のパルスはいずれのパルスであるか不明なので，それをラベルなし事例集合とみなせば，PUC によりノイズパルス（正例）と塩基パルス（負例）の識別ができ，正例であるノイズを除去することで，ほぼ塩基パルスのみからなる集合（塩基パルス集合）を得ることができる．

PUC はこのノイズパルスと塩基パルスの正負例分類のために 1 度使用するだけであり過学習による問題は起こらないので，全パルス集合を学習用データとして用いて PUC 分類器を作成し，それにより全パルス集合をノイズパルスと塩基パルスに分離した．このようにして，BithioU の塩基＋ノイズパルス集合から PUC により BithioU の塩基パルス集合を取得，TTF の塩基＋ノイズパルス集合から PUC により TTF の塩基パルス集合を取得した．

5.2 塩基パルスと塩基＋ノイズパルスの識別精度評価

このようにノイズから分離した BithioU と TTF の塩基パルス集合に対し，通常の 2 値分類器による塩基種類の識別実験を行った．2 種塩基の塩基パルス数のいずれかが 10 に満たない場合は，学習用事例が少なすぎるために実験対象から除外した．識別精度の指標には F-measure を用い，10-Fold crossvalidation（10CV）により精度評価を行った．10CV の際には，BithioU と TTF の塩基パルス数は同数とした．すなわち，PUC により得た BithioU と TTF の塩基パルス数がそれぞれ N_B，N_T であるとき，10CV に用いる塩基パルス数を BithioU，TTF ともに $N = min(N_B, N_T)$ に揃えた．パルス数が N より大きい塩基パルス集合については N 個の塩基パルスをランダム抽出した．

また，PUC によるノイズ除去の効果を見るために，PUC ノイズ除去を施す以前の塩基＋ノイズパルス集合に対しても，BithioU と TTF の識別実験を行った．BithioU，TTF それぞれの塩基＋ノイズパルス集合から N 個ずつランダム抽出して得たパルス集合に対し，塩基パルス同様に 10CV で精度評価を行った．

5.3 実験条件

実験では，機械学習プラットフォームフリーウェア Weka[WEK] を用い，種々の分類器を含めさまざまな条件下で識別精度を調べた．各条件についていかに述べる．

5.3.1 パルス抽出パラメータ

パルス抽出の際には，計測電流値のベースラインからどれだけ外れたらパルス開始と判定するかという波高閾値 α と，何ステップ以上波高閾値を超えたらパルスであると判定するかという波長閾 k 値の 2 つのパラメータを用いている．これらのパラメータを様々試し,「波高閾値 α について 4 通り×波長閾値 k について 4 通り=16 通り」に対して実験を行った．

5.3.2 属性ベクトル

前記の属性ベクトル 1〜8 の 8 種類について試した．

5.3.3 分類器

分類器としてアンサンブル学習「Rotation Forest」を採用し，その内部で用いるベース分類器として，Weka に実装されているもののうち，入力事例連続値ベクトルの 2 値分類を行える以下の 22 種類の分類器を使用した．

functions.SMO	rules.ZeroR
bayes.DMNBtext	trees.ADTree
bayes.NaiveBayes	trees.BFTree
bayes.NaiveBayesUpdateable	trees.DecisionStump
functions.SimpleLogistic	trees.FT
functions.SPegasos	trees.J48
functions.VotedPerceptron	trees.J48graft
lazy.IBk	trees.LADTree
lazy.LWL	trees.RandomForest
rules.JRip	trees.RandomTree
rules.PART	trees.REPTree

5.3.4 PUC 手法

手法 1，手法 2 のいずれの手法も用いた．

5.4 実験結果

識別実験は，パルス抽出パラメータ 16 通り×属性ベクトル 8 通り× 22 分類器× PUC 手法 2 通りの全ての組合せのうちで，塩基パルス数が 2 塩基とも 10 以上であった 3272 ケースについて行った．ただし，今回は単純化のため，抽出したパルスに対して，1)BithioU のノイズ除去，2)TTF のノイズ除去，3) ノイズ除去後の 2 塩基識別，のこれら 3 者に用いた条件（パルス抽出パラメータ，属性ベクトル，分類器，PUC 手法）は全て共通とした．また 5.2 節後半に記述した通り，同様な条件で PUC によるノイズ除去を用いないで塩基＋ノイズパルス集合に対しても識別実験を行った．

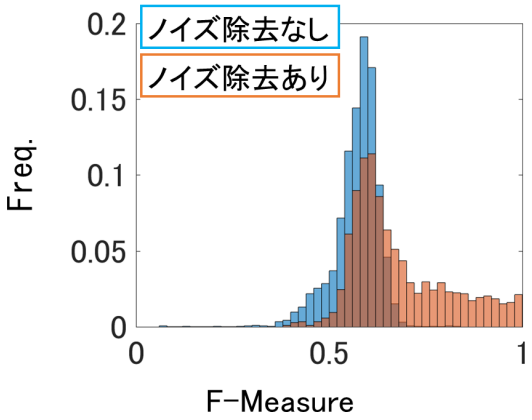

図 4: 識別精度のヒストグラム

これらノイズ除去あり/なし のそれぞれ 3272 ケースについて，F-measure のヒストグラムを図 4 に示す．この図からわかるように，ノイズ除去の効果により多くのパラメータ・条件下において顕著に全体的に精度が向上することが分かった．

また図 5 は，F-measure が 0.93 であった解析条件で使用した計測パルス集合について，塩基と判定したパルス (青)，ノイズと判定したパルス（橙）のパルスピーク波高のヒストグラムを示したものであるが，2 者の分布は非常に重なりが大きい．このようにパルスピーク波高だけでは塩基/ノイズの判定が困難な場合であっても，PUC においてパルス波形特徴をうまくつかんだ属性ベクトルを用いることで適切にノイズパルスを除去し，高い塩基分類精度を得ることができる．

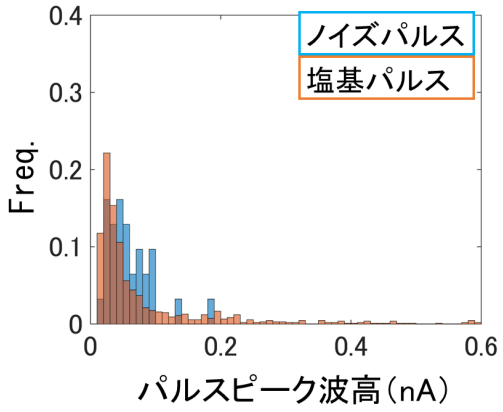

図 5: 塩基/ノイズの波高ヒストグラム

しかし，F-measure > 0.9 となった結果の中には，一方の塩基は PUC によりノイズ除去されたがもう一方の塩基は塩基＋ノイズパルス集合のほとんど全てが塩基パルスと判定された例も多く見られた．つまり一方の塩基だけが PUC によるノイズ除去が施され他方の塩基にはほとんど PUC ノイズ除去が施されずノイズパルスが残ったままであるにもかかわらず，ノイズ除去後の 2 種塩基を高精度で識別できた．これは，一方の塩基＋ノイズパルス集合からノイズ除去できてしまえば，もう一方に塩基パルスとノイズパルスが混在していても，異なる種類の塩基パルスは識別でき，塩基パルスとノイズパルスの識別もできてしまうためであろう．このように，完全にノイズ除去しきれていなくとも見かけ上の塩基の識別精度が高くなってしまうことが起こりうる点は今後の課題である．

6. おわりに

我々は，試料中に含まれるノイズを除去する方法として Elkan and Noto[Elkan 08] の方法を用い，開発中のナノデバイスからの出力パルス波形の識別精度の検証を行った．この結果，ノイズ除去の効果によりさまざまな条件下において大きく識別精度が向上することを確認できた．一方で，以下のような課題も浮かび上がった．

- 一方のデータにのみノイズ除去がなされ他方にはノイズが残留していたとしても，2 塩基の識別精度は高くなってしまう．つまりノイズ除去が不完全であっても見かけ上高精度になってしまう可能性がある．
- ノイズ除去に適切な条件についてさらなる詳細な調査が必要である．

今後は，これらの課題を改善すべく本研究を継続していく予定である．

7. 謝辞

本研究に用いた人工核酸塩基（ジチオフェンウラシル誘導体，TTF ウラシル誘導体）は，東京大学工学部山東研より提供いただいた．この場をお借りして感謝申し上げる．

本研究は，JST，CREST「計測技術と高度情報処理の融合によるインテリジェント計測・解析手法の開発と応用」の支援を受けたものである．

参考文献

[Elkan 08] Elkan, C. and Noto, K.: Learning Classifiers from Only Positive and Unlabeled Data, in *KDD '08 Proceedings of t,he 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213–220, Las Vegas, Nevada, USA (2008), ACM New York, NY, USA

[Ohshiro 12] Ohshiro, T., Matsubara, K., Tsutsui, M., Furuhashi, M., Taniguchi, M., and Kawai, T.: Single-Molecule Electrical Random Resequencing of DNA and RNA, *Scientific Reports*, Vol. 2, No. 501 (2012)

[Ohshiro 14] Ohshiro, T., Tsutsui, K., Yokota, K., Furuhashi, M., Taniguchi, M., and Kawai, T.: *Nature Nanotechnology*, Vol. 9, pp. 835–840 (2014)

[Rosenstein 12] Rosenstein, J. K., Wanunua, M., Merchant, C. A., Drndic, M., and Shepard, K. L.: Integrated nanopore sensing platform with sub-microsecond temporal resolution, *Nature Methods*, pp. 487–492 (2012)

[Tsutsui 10] Tsutsui, M., Taniguchi, M., Yokota, K., and Kawai, T.: Identifying Single Nucleotides by Tunneling Current, *Nature Nanotechnology*, Vol. 5, pp. 286–290 (2010)

[Tsutsui 11] Tsutsui, M., Matsubara, K., Ohshiro, T., Furuhashi, M., Taniguchi, M., and Kawai, T.: Electrical Detection of Single Methylcytosines in a DNA Oligomer, *J. Am. Chem. Soc.*, Vol. 133, (2011)

[WEK] http://www.cs.waikato.ac.nz/ml/weka/