生化学分野における Video&Language データセットの構築

迫田航次郎 [1]　西村太一 [1]　森信介 [2]　小野富三人 [3]　田中奈津子 [3]
[1] 京都大学大学院情報学研究科 [2] 京都大学学術情報メディアセンター
[3] 大阪医科薬科大学
{sakoda.kojiro.48z,nishimura.taichi.43x}@st.kyoto-u.ac.jp
forest@i.kyoto-u.ac.jp

概要

本研究では生化学分野における Video&Language データセット（Biochemical Video-and-Language: BioVL）を構築した．このデータセットは生化学実験動画，実験プロトコル，動画に映る動作とプロトコルに記述された手順との対応関係を示すアノテーションで構成される．動画は 4 種類の実験ごとに各 4 回ずつの 16 動画，合計 1.6 時間分を収録した．本研究の実験では構築した BioVL データセットを用いて，手順と動作との対応関係を推定し，アライメントを獲得するタスクを行った．

1　はじめに

科学の世界には「再現性の危機」と呼ばれる問題がある．これは，研究者が自分自身が過去に行った実験や，他の研究者が行った実験を再現できないという問題である．[1] によると，生化学分野や生命科学分野の 80%以上の研究者が他の研究者が行った実験を再現できなかった経験があると回答している．また，再現性の危機が生じる原因の一つとして，研究者同士で実験方法や実験中の詳細な情報が共有できていないことが示されている．実験の実施に必要な，実験の手順や機器の使用法，試薬の調整法などの情報はプロトコルに記述され，研究者はプロトコルを参照することで実験を再現できる．プロトコルは実験に必要な手順をテキスト化しているため，多くの人が簡単に実験を実施できるようになるが，プロトコルの記述だけではそれぞれの手順が実際にどのような動作で行われているかなどの情報が伝わりづらい．そこで，実験の再現性を高める手段の一つとして，研究者が行う実験の様子を収録し，収録した動画とプロトコルとを共に参照することが挙げられる．

本研究では生化学分野を対象に，Video&Language の技術を用いて，実験方法や実験中の詳細な情報の共有を支援することで再現性の危機の解消を目指す．具体的には，研究者が行なっている実験を収録した動画と実験の手順や条件が記述されたプロトコルとの対応関係を推定し，手順と動作のアライメントを獲得する．アライメントを獲得することで，プロトコルに記述された手順から得られる定量的な情報と動画に映る動作から得られる定性的な情報の両方を参照しながら実験を行えるため，再現実験が容易になると考えられる．

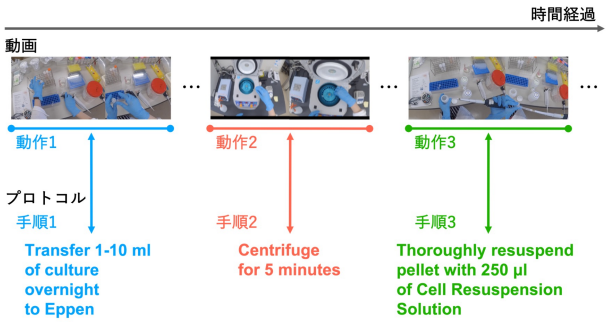
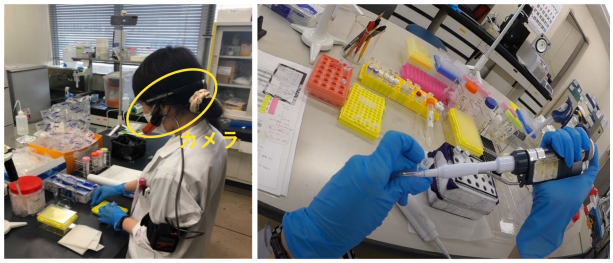


図 1　BioVL データセットの例．生化学実験動画，実験プロトコル，動画に映る動作とプロトコルに記述された手順との対応関係を示すアノテーションで構成される．

この課題を達成するための取り組みとして，まず初めに生化学分野における新しい Video&Language データセット (Biochemical Video-and-Language: BioVL) を構築する（図 1）．このデータセットは生化学実験動画，実験プロトコル，動画に映る動作とプロトコルに記述された手順との対応関係を示すアノテーションで構成される．次に，構築した BioVL データセットを用いてプロトコルに記述された手順と動画に映る動作とのアライメントを獲得するタスクを行った．本研究の実験結果は，最先端の事前学習済みモデルを用いても大幅な改善の余地があることを示している．また，BioVL データセットは研究目的でのみオンラインで利用する



(a) (b)

図 2 (a)：カメラを装着した状態で実験を行なっている様子．(b)：一人称視点カメラで収録した実験の様子．

ことができ，これは生化学分野の Video&Language データセットとしては初めての試みである．

2 BioVL データセットの構築と統計情報

BioVL データセットは生化学実験動画，実験プロトコル，動画に映る動作とプロトコルに記述された手順との対応関係を示すアノテーションで構成される．以下に BioVL データセットの動画収録方法，アノテーション方法，データセットの統計情報について示す．

2.1 実験動画の収録

2.1.1 収録した実験

本研究では DNA 抽出，PCR，アガロースゲル作成，ミニプレップの，生化学分野の研究室で一般的によく行われている 4 種類の実験を収録対象とした．収録は実験ごとに 4 回ずつ実施し，合計 16 動画，1.6 時間分を収録した．また，DNA 抽出にはフェノールクロロホルム抽出とエタノール沈殿という 2 つの方法があり，DNA 抽出に関してはこれらを 2 回ずつ収録した．

2.1.2 実験収録の様子

本研究では大阪医科薬科大学の研究者に，カメラを装着した状態で実験を実施してもらうよう依頼した．動画収録の負担を最小限にし，将来的なデータセットの拡張のために，未編集な一人称視点 [2] のカメラを採用した．また，収録のために追加のカメラやセンサー [3, 4] は用いない．カメラは Panasonic の HX-A500 を使用し，図 2 の (a) のようにヘッドセットに固定されている．また，実験は普段と同じ環境で 1 人の研究者がプロトコルを参照しながら実施した．

表 1 PCR のアノテーション例．

手順	開始時間 (s)	終了時間 (s)
add sterile distilled water	30	45
add primer1	64	99
add primer2	106	130
add template	149	173
add primeSTAR	190	238
set in DNA engine	260	266

2.2 アノテーション方法

本研究では [5] に則り，非専門家によるアノテーションと，専門家による検証を順に行うことでアノテーションを付与した．生化学分野の非専門家のアノテーターは，収録した実験動画を見ながら，プロトコルの各手順と動画に映る動作の開始時間と終了時間を対応付けた．このアノテーションの段階で，専門家のアノテーターが非専門家のアノテーターの付けたアノテーションを検証し，誤りがあれば訂正した．BioVL データセットとしては専門家が検証した後のアノテーションのみを保存したため，非専門家と専門家のアノテーションの一致率は計算できない．将来的には，他の専門家にさらにアノテーションの検証を依頼することでアノテーションの品質を評価する予定である．表 1 に PCR のアノテーションの例を示す．

2.3 統計情報

2.3.1 プロトコル側の統計情報

前処理として，プロトコルに記述された手順を動詞ごとに手作業で分割した．例として，"Invert 4 times to mix and add 10μl of Alkaline Protease Solution." という記述があった場合，"Invert 4 times to mix." と"Add 10μl of Alkaline Protease Solution."の 2 つの手順に分割する．この前処理を行なった後のプロトコル側の統計情報を表 2 に示す．表 2 では実験によって手順の数が大幅に異なることが示されている．手順数が最大の実験はミニプレップであり，最小の実験はフェノールクロロホルム抽出である．また，同じ実験であっても手順数が異なり，標準偏差が生じているのは，一部の実験で手順の簡略化を行なったためである．手順ごとの平均単語数が最大の実験はミニプレップであり，最小の実験は PCR である．

表 2　実験ごとの手順数と手順ごとの平均単語数.

実験名		手順数	単語数/手順数
DNA 抽出	フェノールクロロホルム抽出	4.0(± 0.0)	6.0(± 1.9)
	エタノール沈殿	9.0(± 0.0)	4.9(± 2.9)
PCR		6.0(± 0.0)	3.0(± 1.0)
アガロースゲル作成		10.3(± 0.4)	4.7(± 2.4)
ミニプレップ		28.2(± 0.4)	6.4(± 2.5)

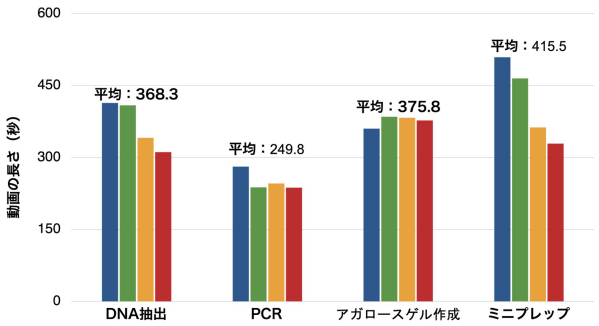


図 3　実験ごとの動画の長さ.

2.3.2　動画側の統計情報

いくつかの実験では研究者が特定の手順を完了する間（試薬を遠心分離するなど）に待ち時間が生じる．待機している間，研究者にはカメラをヘッドセットに固定したまま，ヘッドセットを外してもらった．このような待機時間は実験手順とは無関係なため，手動でその間のシーンをカットした．プロトコル側と同様に，動画の前処理を行なった後の動画側の統計情報を図 3 に示す．図 3 は実験ごとの動画の長さを示している．DNA 抽出について，左側の 2 つがエタノール沈殿，右側の 2 つがフェノールクロロホルム抽出を示している．また，ミニプレップの動画の長さが左側の 2 つと右側の 2 つで大きく異なるのはミニプレップの 3 回目，4 回目の収録で試料の反応時間や手順を簡略化したためである．動画の平均時間について，ミニプレップが最長であり，PCR が最短である．

3　実験と評価

本研究では BioVL データセットを用いてプロトコルに記述された手順と動画に映る動作とのアライメントを獲得するタスクを行なった．このタスクを行うにあたって，BioVL データセットの大きさが限られているため，機械学習モデルを学習させることが困難だった．そのため，大規模なデータセットで事前学習したモデルである Video and Language Embeddimg model (VLE)[6] と Contrastive Language–Image Pre-training (CLIP)[7] をそれぞれ，モデルの重みを再学習することなく BioVL データセットを用いたタスクに適用した．以下に，実験に用いた事前学習済みモデル，行ったタスクの詳細と結果について示す．

3.1　実験に用いた事前学習済みモデル

3.1.1　VLE

VLE は動画と言語の対応関係を推定するモデルである．HowTo100M データセット [6] を用いて，対応する動画と言語のペアを共有潜在空間上に埋め込み，それらの表現ベクトルの cos 類似度が大きくなるように事前学習が行われている．VLE は事前学習に用いていない様々な Video&Language データセットに対する動画検索タスクで高い精度を達成している．

3.1.2　CLIP

CLIP は言語から画像表現を学習する画像分類モデルである．英語版 Wikipedia から，4 億枚の画像とその画像を説明する言語のペアを収集することで構築したデータセットである，WebImageText(WIT) を用いて事前学習されている．事前学習は共有潜在空間上で，対応する言語と画像との表現ベクトルの cos 類似度を最大化する一方で，対応しない言語と画像との表現ベクトルの cos 類似度を最小化することで行われている．CLIP は VLE と同様に，事前学習に用いていない様々なデータセットに対する画像分類タスクで高い精度を達成している．また，CLIP は画像を入出力するモデルであるため，本研究では [7] と同様に，モデルに入力する各動画区間の中央のフレームを入力画像として用いた．

3.2　手順と動作のアライメントを獲得するタスク

このタスクではプロトコルに記述された手順と動画に映る動作との対応関係を推定する．手順と動作

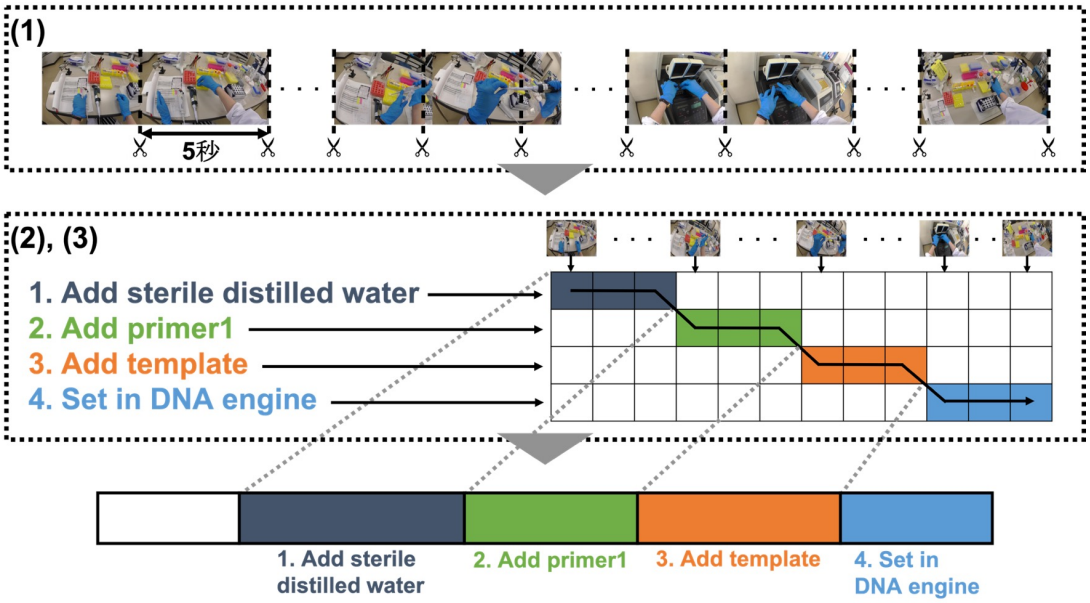


図 4　アライメントモデル.

表 3　アライメントタスクの結果.

	mIoU(%)
ベースライン	28.6
VLE	30.4
CLIP	**33.2**

のアライメントを獲得することで，研究者がプロトコルと動画の両方から実験に必要な情報を得ることができるようになるため，実験の再現性を高めることができる．本研究では３つの処理で構成されるアライメントモデルを提案した（図 4）．(1) で動画を 5 秒間隔のセグメントに区切る．(2) では事前学習済みモデルから出力された手順の表現ベクトルと各セグメントの表現ベクトルの cos 類似度を計算する．最後に，(3) では動的計画法の手法の１つである Needleman-Wunsch アルゴリズム [8] に基づいて最適な経路を探索することでアライメントを獲得する．

3.2.1　評価尺度

アライメントの結果を評価するために，Intersection over Union (IoU) を計算し，その平均 (mIoU) を取る．mIoU は手順に対してアノテーションされた動画に映る動作の時間幅と，入力した手順に対して事前学習済みモデルが予測した動作の時間幅が一致している割合の平均を示す．

3.2.2　結果

表 3 にアライメントタスクの結果を示す．ベースラインは動画を手順数で均一な時間幅に分割した時の結果である．本研究では最先端の事前学習済みモデルを用いたが，結果はベースラインとほとんど同じとなった．このような結果となったのは，事前学習済みモデルの学習に用いたデータセットに生化学分野のデータが含まれておらず，BioVL データセットの動画に対して適切な表現ベクトルが得られていないことが原因であると考えられる．

4　おわりに

本研究では Video&Language 技術を用いて，生化学分野の実験の再現を支援することを目的としている．この課題を達成するための取り組みとして，まず初めに生化学実験動画，実験プロトコル，動作と手順の対応関係を示すアノテーションで構成される BioVL データセットを構築した．また，BioVL データセットを用いて，プロトコルに記述された手順と動画に映る動作とのアライメントを獲得するタスクを実施した．本研究の実験結果は，最先端の事前学習済みモデルを用いても大幅な改善の余地があることを示している．今後の方針としては BioVL データセットの動画数を増やし，機械学習モデルを学習させることや，本研究で用いた事前学習済みモデルを再学習させるために Web 上から生化学実験動画を収集することなどが考えられる．

参考文献

[1] Monya Baker. 1,500 scientists lift the lid on reproducibility. **Nature**, No. 533, pp. 452–454, 2016.

[2] Giovanni Maria Farinella Sanja Fidler Antonino Furnari Evangelos Kazakos Davide Moltisanti Jonathan Munro Toby Perrett Will Price Dima Damen, Hazel Doughty and Michael Wray. Scaling egocentric vision: The epickitchens dataset. In **Proc. ECCV**, pp. 753–771, 2018.

[3] Qiguang Liu Henry Kautz Jiebo Luo Iftekhar Naim, Young Song and Daniel Gildea. Unsupervised alignment of natural language instructions with video segments. In **Proc. AAAI**, pp. 1558–1564, 2014.

[4] Michaela Regneri Sikandar Amin Mykhaylo Andriluka Manfred Pinkal Marcus Rohrbach, Anna Rohrbach and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. **IJCV**, No. 119.

[5] Chenliang Xu Luowei Zhou and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In **Proc. AAAI**, pp. 7590–7598, 2018.

[6] Jean-Baptiste Alayrac Makarand Tapaswi Ivan Laptev Antoine Miech, Dimitri Zhukov and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In **Proc. ICCV**, pp. 2630–2640, 2019.

[7] Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark Gretchen Krueger Alec Radford, Jong Wook Kim and Ilya Sutskever. Learning transferable visual models from natural language supervision. In **Proc. ICML**, 2021.

[8] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Jornal of Molecular Biology**, No. 48, pp. 443–453, 1970.