

Class 1: xPore: An AI-Powered App for Bioinformaticians

คอร์สนี้จะพูดถึง Application ที่มี AI และถูกพัฒนาขึ้นสำหรับนักชีววิทยาสารสนเทศ

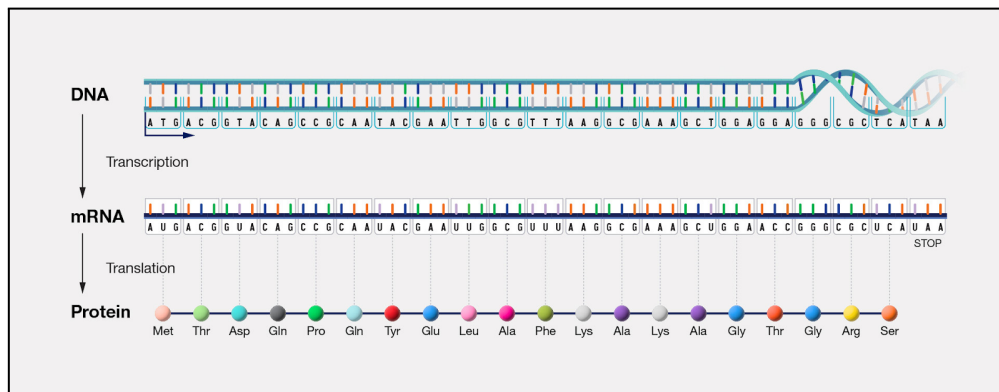
→ xPore เป็นแอปสำหรับการหาตำแหน่ง RNA modification

→ Data ที่ใช้ในการพัฒนาแอปนี้มาจาก nanocore sequencing เป็นเครื่องมือที่สามารถลำดับตำแหน่งของ RNA ให้ output ออกมาเป็น สัญญาณไฟฟ้า และนำ machine learning เข้ามาใช้ในการระบุในแต่ละเซลล์มีสัญญาณไฟฟ้าแตกต่างกันยังไง

Problem Statement

พัฒนา software เพื่อตอบโจทย์อะไร การศึกษา domain หรือความรู้ในงานวิจัยมีความรู้ที่เกี่ยวข้องอย่างไร

☐ Central Dogma >> ลำดับเบส

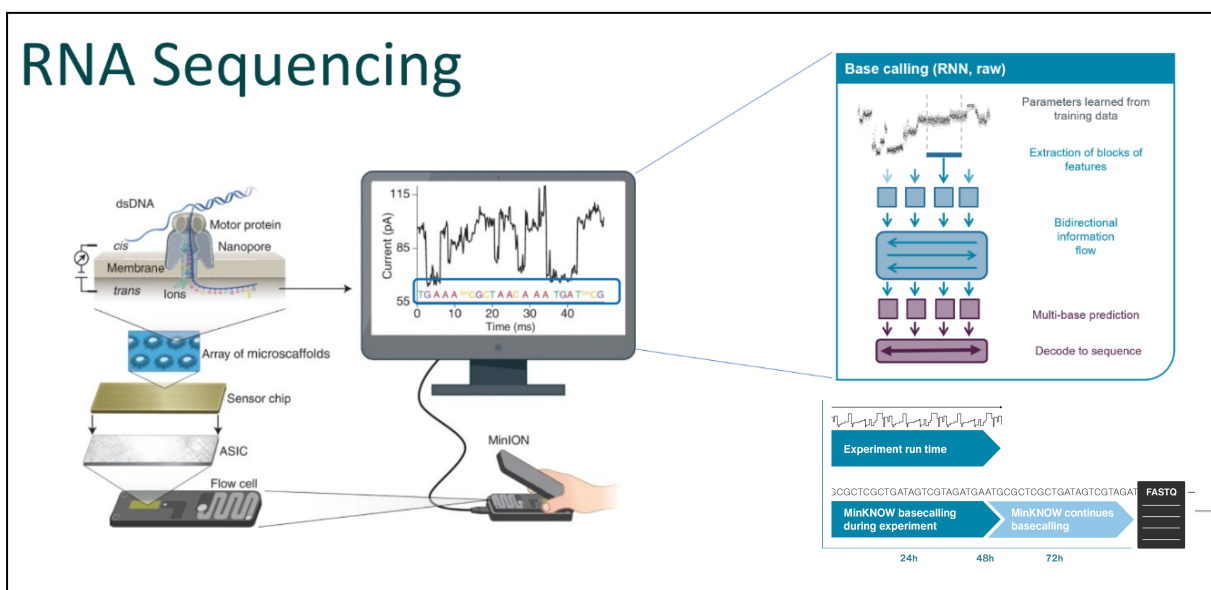


จากสาย DNA เวลาทำงานใน cell ถูกแปลง (transcription) มาเป็น mRNA จากนั้น mRNA ถูกถอดรหัส (translation) มาเป็น Protein ที่ทำหน้าที่ต่าง ๆ ในร่างกาย เช่น การเกิด / รักษา โรคต่างๆ ก็ต้องมาดู mRNA พวกนี้ว่ามันทำงานผิดปกติไปอย่างไร งานวิจัยนี้สนใจ mRNA ซึ่งก็คือสนใจทั้งสาย DNA แต่หากแบ่งเป็นช่วง ๆ ึ่งใน DNA → Gene หรือดู all Gene ซึ่งในเชิงของข้อมูล ก็คือสนใจที่ **การแสดงออกของ gene (Gene expression)** เป็นการ transcribe ตัวเอง มากน้อยแล้วแต่ gene ซึ่งตรงนี้จะบ่งบอก จำนวน Reads เราจะโฟกัสที่ Reads แสดงออกมากหรือน้อย จะมีลำดับเบส (sequence) ที่ปกติหรือไม่ ? ก็คือ เอา mRNA เข้า machine เพื่อดูว่าสาย mRNA ทั้งสาย ในแต่ละ gene ที่มีมากน้อยแตกต่างกันไป สายไหนปกติ สายไหนผิดปกติ...

☐ RNA >> Nanopore sequencing machine >> ช่วยในการหาลำดับเบสบน RNA มีอะไรผิดปกติหรือไม่ หรือก็คือ ไม่ได้เป็นเบสทั่วไป AGCU เหมือน RNA ทั่วไป คืออาจจะโมเลกุลเปลี่ยนไป >> RNA modification

☐ หลักการทำงานของ Nanopore sequencer machine

RNA Sequencing



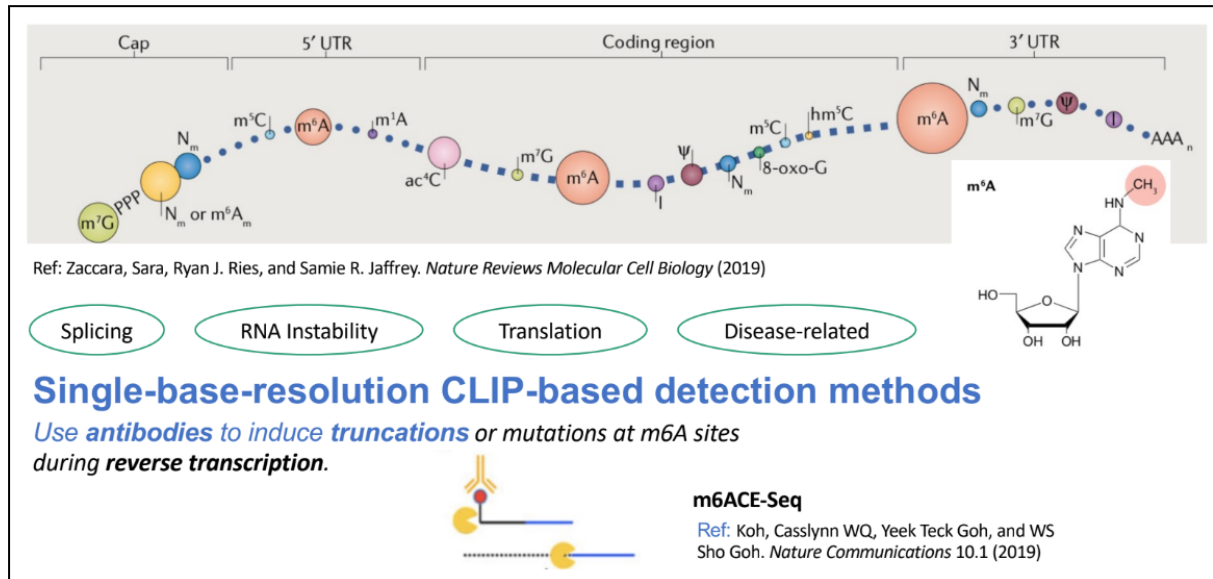
⇒ หยอดสาร หรือ ตัวอย่างเซลล์มนุษย์หรือไวรัสลงไปใน Nanopore จะมี Motor Protein ที่ทำหน้าที่ดึง data (ลำดับเบส)

ซึ่งโปรตีนนั้นจะมีความต่างศักย์ไฟฟ้าอยู่ หากมีโมเลกุลก็就会有ความต้านทานไฟฟ้าที่แตกต่างกันออกไปในแต่ละ stage ซึ่งก็จะส่งผลกับค่าของไฟฟ้า (Current) บนจอ ซึ่ง Intensity ก็จะแตกต่างกันไป

⇒ โมเลกุล มีความต้านเยอะ ค่าไฟฟ้าที่ผ่านระหว่าง 2 ขั้ว น้อยลง ลำดับเบสในแต่ละตัว (A G C U) ก็จะมีขนาดโมเลกุลแตกต่างกันไป ใช้ Pattern ของขนาดไฟฟ้าของแต่ละโมเลกุลเพื่อแยกเบสแต่ละตัวคืออะไร

⇒ ใช้ RNA sequencer == Nanopore ร่วมกับตัวหนึ่งชื่อ Base calling ใช้ในการดูลำดับเบส แปลงสัญญาณไฟฟ้าไปเป็นลำดับเบส ข้อดีคือ 1. ทำ Direct RNA sequencing 2. ขนาดเล็ก 3. ทำ real-time ได้ ในขณะที่เครื่องนี้ sample sequencing อยู่ เราก็ใช้ machine learning ทำ base calling ได้

☐ RNA modifications



⇒ RNA ปกติ. เบส 4 อย่าง A G C U ในแบบต่างๆ คือโมเลกุล แต่ตัวโมเลกุลอาจแตกต่างกัน อาจมีการดัดแปลง (modify)

เพื่อให้การทำงานของเซลล์ในร่างกายทำงานได้ปกติขึ้น เช่น m6A ถูก modify ด้วย CH มาเกาะที่เบส A → common RNA modifications เกิดขึ้นเยอะ มีส่วนสำคัญเกี่ยวกับทำงานของเซลล์ในร่างกาย

⇒ มีงานวิจัย M6ACE-Seq ที่สามารถ หาตำแหน่งไหนบนสาย RNA ว่ามี M6A

⇒ Output Table ⇒ เรา Input genomic position แล้วหา modification rate เพื่อมาดูผลตอนท้ายว่า เช่น คนเป็นมะเร็ง มี % การ RNA modification อยู่ที่เท่าไร ซึ่งก็เทียบหลายๆ sample แต่ละ sample มี gene เป็นหมื่น ๆ set

☐ Research Objective คือ

เราจะเอา สาย RNA เข้าไป ⇒ Nanopore Sequencing ⇒ ได้ Pattern ของสัญญาณไฟฟ้า ⇒ เราจะทราบ ว่า โมเลกุลที่ต่างกันที่ตัวตรงกลางตัวเดียว (จาก A เป็น M6A) สัญญาณไฟฟ้าจะไม่เท่ากัน

- Locate modified position ⇒ M6A เกิดบนตำแหน่งไหนของสาย RNA
- Quantify fraction of modified reads - modified rate ⇒ ต่างเล็กน้อย เพื่อไปวิเคราะห์ต่อว่า การ modify ของ RNA มีผลต่อการเป็นโรคต่าง ๆ หรือการทำงานของร่างกาย อย่างไร

☐ Methods or Softwares package ที่สามารถ detect ตำแหน่งของ M6A ได้

- EpiNano / MINES ⇒ Supervised Learning (requires training data)
- Tombo / Nanocompare ⇒ Unsupervised Learning detect all modification not only M6A ⇒ xPore

Data Collection and Preparation

- Data collection

- ☐ FAST5 ⇒ HDF5 format (binary), storing large and complex data

ได้มาจาก Nanopore Sequencer เป็นไฟล์เก็บ Raw signal หรือ Current Intensity level (pA) ที่เราต้องการนำมาสร้างโมเดลนั่นเอง ⇒ 1 FAST5 file ⇒ 1 Reads ของ RNA ⇒ ความยาวของ Gene

- ☐ FASTQ ⇒ Basecalled sequence ⇒ txt format

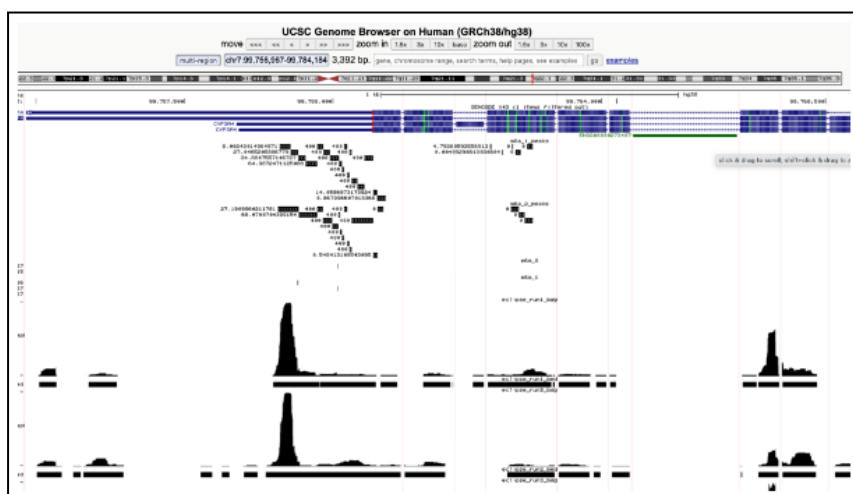
ไฟล์เก็บลำดับเบส ที่ถูกแปลงมาจากสัญญาณไฟฟ้า ผลลัพธ์ที่ได้จากการ basecalled

- ☐ FASTA ⇒ reference sequence ⇒ txt format

ไฟล์ที่บอก RNA ที่ถูก convert เป็น DNA แล้ว

- ☐ BAM (binary) / SAM (txt) ⇒ Alignment result (FASTQ aligned with FASTA)

จะเก็บผลลัพธ์ของการ Alignment ระหว่าง FASTQ กับ FASTA ดูว่า reads เบอร์ไหน เข้ากับ gene ตัวไหน ส่วนนี้จะบอกข้อมูลว่า ตรงไหนที่มัน aligned ได้ ตรงไม่ตรง เว้นวรรคอย่างไร



เครื่องมือ Genome Browser

⇒ ใช้ check aligned result

⇒ สามารถ visualize RNA

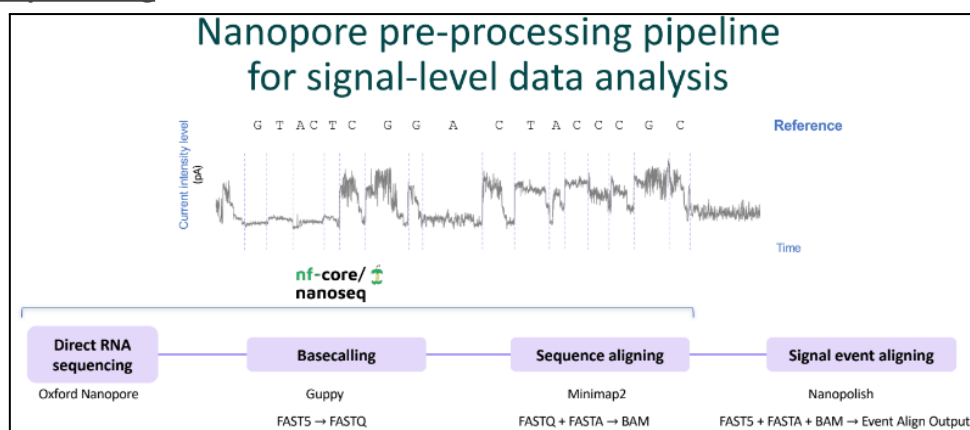
ตัวนี้ มี reads หรือลำดับเบส

แต่ละตัวอย่างไร

⇒ hist บอกสิ่งที่เรา segment

ได้ มีจำนวนมากน้อยขนาดไหน

- Preprocessing



⇒ เริ่มจาก การทำ Direct RNA sequencing ได้สัญญาณไฟฟ้าออกมา เป็น Current Intensity level

⇒ ทำ Base Calling ดูว่าสัญญาณไฟฟ้าแต่ละตรง เทียบแพทเทิร์นแล้ว ตัวไหนเป็นตัวตรงกลาง GTAC โดยใช้ software ที่ชื่อว่า Guppy ⇒ FAST5 → FASTQ

⇒ Sequence aligning โดยใช้ software ที่ชื่อ Minimap2 ⇒ FASTQ+FASTA → BAM alignment result ว่า reads ไหน แมพได้กะ Gene หรือ RNA ตัวไหน ที่ตรงกับ Reference

⇒ signal event aligning ⇒ align ระหว่างสัญญาณไฟฟ้า ว่ามันตรงกับ เบสตัวไหน โดยใช้ software ที่ชื่อ Nanopolish

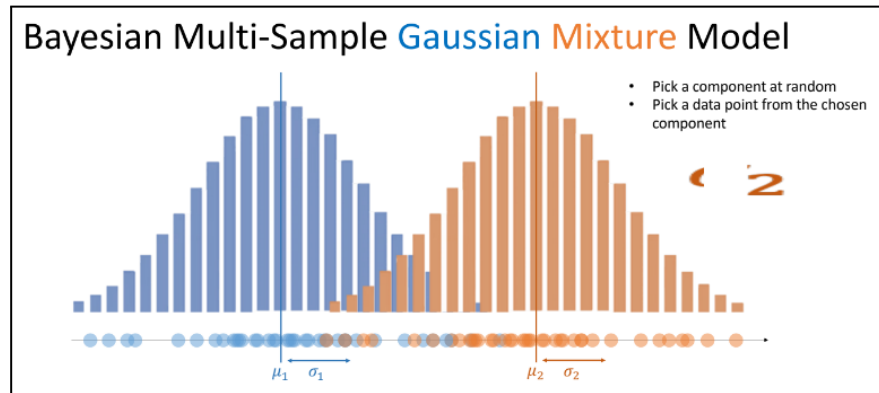
⇒ FAST5+FASTA+BAM → event align Output

Bayesian [Multi-Sample] Gaussian Mixture Modeling เทคนิคหลักที่ใช้ใน xPore

☐ What is gaussian?

Gaussian คือ การแจกแจงรูปแบบหนึ่ง ที่แกน x เป็นค่าที่เราสนใจ (ค่าสัญญาณไฟฟ้า) แกน y เป็น frequency probability.

ในทุกๆ x จะมี probability ที่จะถูกแรนด้อมในทุก sample \Rightarrow การสร้างข้อมูลจาก 1 Gaussian Distribution



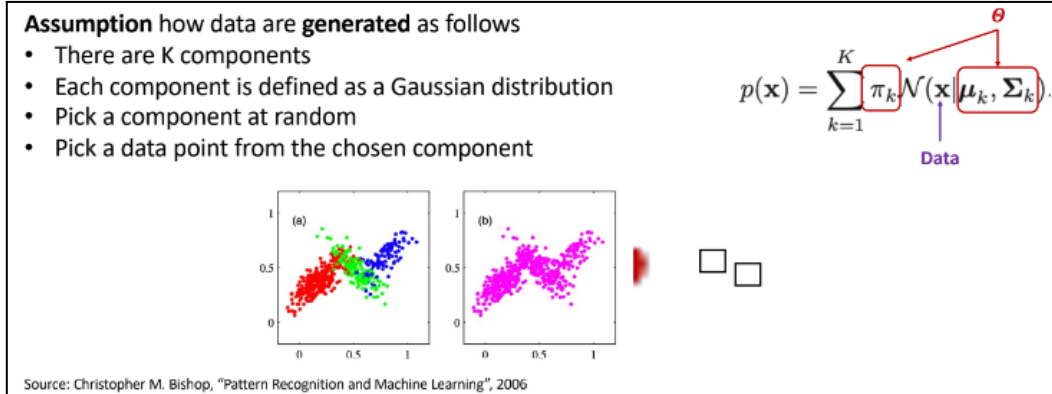
หากมีหลาย Gaussian \Rightarrow Gaussian Mixture ก็คือมีหลาย Distribution เช่น การกระจายตัวที่บ่งบอกโมเลกุลของ RNA

\Rightarrow เลือกก่อนว่าข้อมูลนั้นจะ random มาจาก component ไດ

\Rightarrow แล้วค่อยเลือก data point

ในทุก ๆ x จะมี probability ที่จะถูกแรนด้อมแล้วเลือก component ที่เท่าไร ในทุก sample \Rightarrow การสร้างข้อมูลจาก Gaussian Mixture

☐ What is a gaussian mixture model (GMM)?



Source: Christopher M. Bishop, "Pattern Recognition and Machine Learning", 2006

รูปซ้ายเป็นเพียงสมมติฐานของการที่ data ถูก generated มาโดยใช้ GMM แต่ในความเป็นจริงคือรูปขวา เราไม่สามารถรู้ได้เลยว่า K จริงๆแล้วมีกี่ component , เราไม่รู้ค่า μ และ σ ของแต่ละอัน , และไม่รู้ค่า π ของแต่ละ component ด้วย แต่เรา assume ได้จาก assumption \rightarrow สุดท้ายเราก็ไป learn และหาค่าออกมาว่าจริงๆแล้ว แต่ละ data point นั้นมาจาก component ไດ จุดนี้จริงๆคืออะไร โดยดูจาก probability ของแต่ละ component และ แต่ละจุด

☐ gaussian mixture model (GMM) Inference

\Rightarrow เริ่มจาก random μ และ σ ของแต่ละ component

\Rightarrow วิธีการ learn ของ gaussian ก็คือจะพยายาม assign ว่า จุดที่เราแรนด้อม มันน่าจะอยู่ใน component ไหนของ distribution ที่มี ก็เลย assign ว่า data point มี probability ที่ใกล้เคียงกับ component ไหน ก็จะ update μ และ σ ใหม่ไปเรื่อยๆ \Rightarrow iterative algorithms

\Rightarrow จนสุดท้ายเราจะเจอ μ และ σ ที่ fit กับ data ของเรามากที่สุด

- **Discriminative AI** \Rightarrow model ที่สามารถจัดกลุ่ม data ให้เราได้ บอก class ของ data ได้ เช่น clustering
- **Generative AI** \Rightarrow model ที่สามารถสร้าง data ให้เราได้ เช่น image processing \Rightarrow GMM

☐ Bayesian or Posterior

⇒ เป็น learning algorithm for making inference on the latent variables วิธีการประมาณค่าตัวพารามิเตอร์

⇒ $P(\text{Data} \mid \theta) \times P(\theta) \Rightarrow$ ดู Prior ด้วย คือ ดู distribution ของพารามิเตอร์ตัวนี้ บอกความไม่แน่นอนของการประมาณค่า

☐ Frequentist or Point estimate

⇒ $\text{argmax } P(\text{Data} \mid \theta) \Rightarrow$ พารามิเตอร์ตัวไหน ค่าเท่าไร ที่จะทำให้ probability ของ data ทั้งหมด สูงที่สุด

☐ Bayesian Multi-Sample Gaussian Mixture Model

⇒ เริ่มมา เราเอา RNA ของคนไข้ 3 คน หา RNA sequencing ได้สัญญาณไฟฟ้าออกมา แล้วเทียบกัน จุดที่มันปกติ สัญญาณก็จะซ้อนทับกันไม่มีปัญหาในขณะนั้น

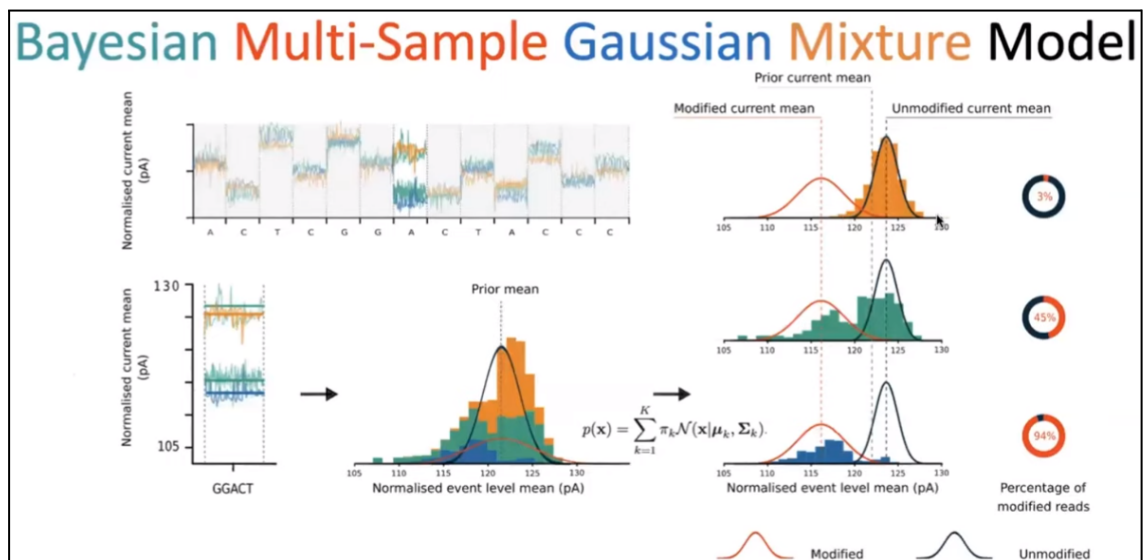
⇒ จุดที่ไม่ซ้อนทับกัน จุดที่มันเกิด RNA modified 1 เส้นสัญญาณ == 1 reads สามารถ plt hist เพื่อแยกคนไข้

⇒ assume distribution ให้กับ hist ของสัญญาณไฟฟ้าของคนไข้ 2 อัน แล้วดูว่า จุดข้อมูลของคนไข้จะไปอยู่ใน component ไหน \Rightarrow fit gaussian \Rightarrow use bayesian หา best parameter \Rightarrow สุดท้าย iterative algorithm จะแยกว่าอันไหนจะ fit กับ data มากที่สุด

⇒ สุดท้ายก็จะได้ modification rate เพื่อเปรียบเทียบ differential

Why do we use this model?

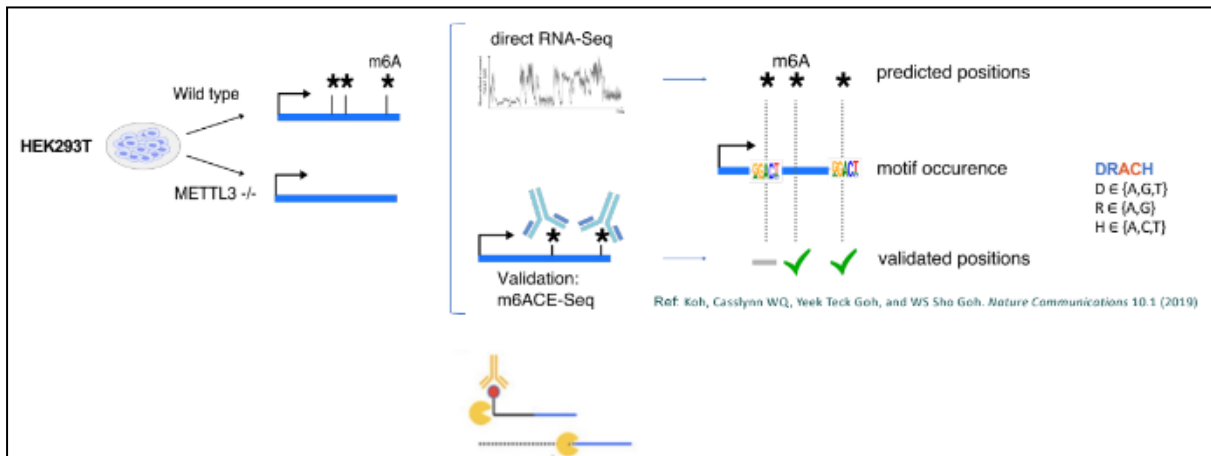
- Each site has 2 distribution at maximum (2K \Rightarrow 2 component)
 - Unmodified
 - Modified only one type \Rightarrow M6A
- To accommodate comparison across many samples
- Nanopolish Event Align assume a Gaussian distribution for each signal event (k-mer)
- μ and σ of unmodified k-mer is estimated by the eventalign algorithms in prior
- Fast \Rightarrow Parallelisation



Evaluation วิธีการประเมินค.ถูกต้องของ xPore

ถ้าได้ result จากการรันโมเดลมาแล้ว ว่าตำแหน่งไหนมี M6A จะมีวิธีการ ประเมินว่าsoftware ของเราน่าเชื่อถือ เพื่อให้คนเอาไปใช้

☐ Experiment setup ⇒ ออกแบบการทดลองว่าตอบโจทย์ research objective หรือไม่



⇒ โดยเริ่มจาก ตัวอย่างเอา cell ที่ถูก cut มาจากไตของมนุษย์ มาทำ

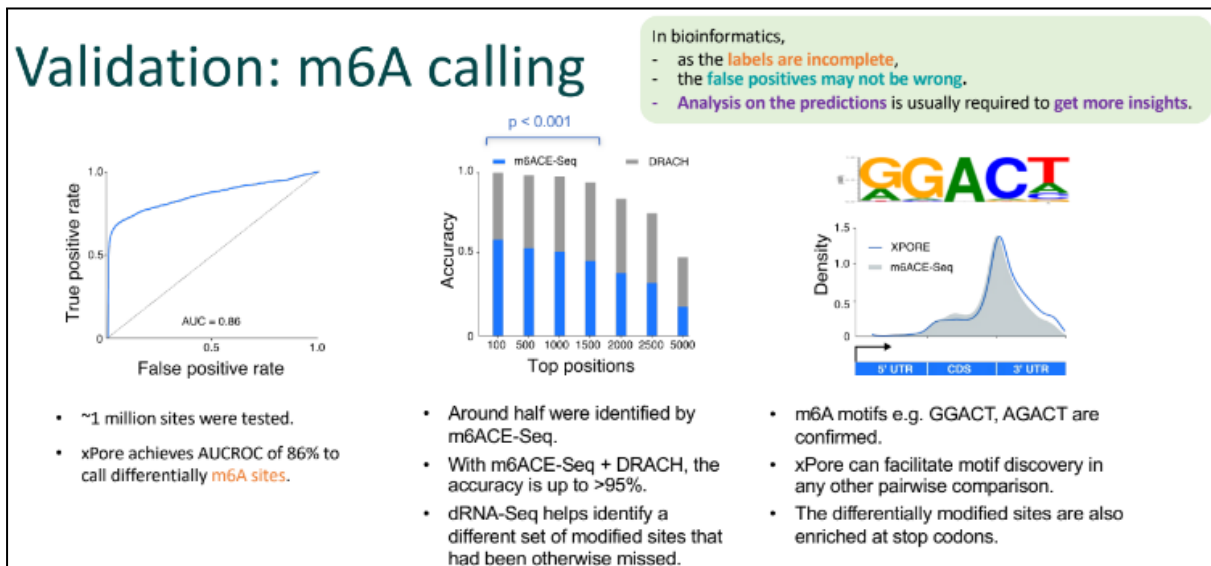
- wild type มี M6A position
- knock out → gene ที่ไปสร้าง M6A ถูกทำให้มันหายไป

⇒ จากนั้น เอาไปทำ Direct RNA seq ทั้ง 2 samples

⇒ แล้วก็ เอาไป predict โดยใช้ GMM บอกว่าตำแหน่งไหนที่มี M6A บ้าง ซึ่ง จาก M6ACE-Seq (เปเปอร์เดิม) มีตำแหน่ง reference อยู่แล้ว ถูกนำมาใช้เป็นตัว validation: ว่าสิ่งที่เราทำนายมา ถูกหรือไม่

⇒ เพิ่มเดิม M6A ตำแหน่งนั้นจะต้องมี motif (5-mer) เป็น DRACH เท่านั้น ถ้าเป็นอย่างอื่น จะไม่ถูก modified by M6A → จากองค์ความรู้เดิม

☐ Validation: M6A calling

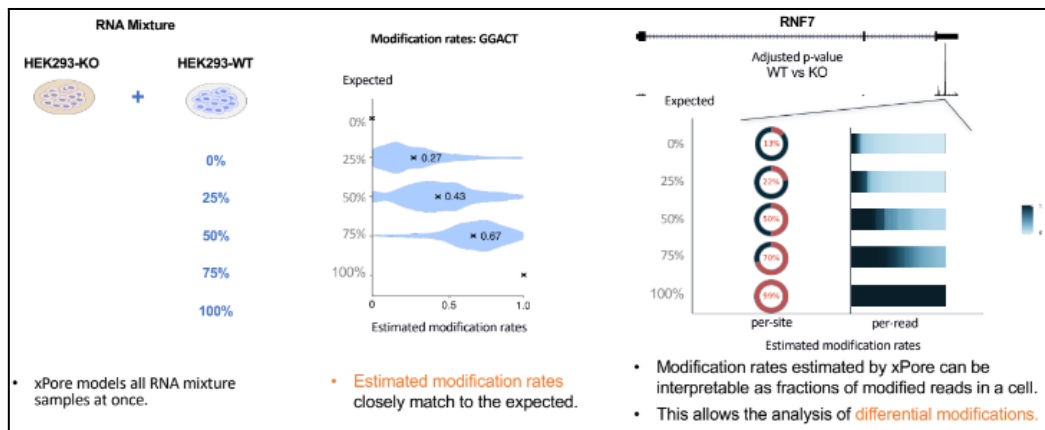


⇒ มีลิสต์ของตำแหน่ง RNA แล้วก็เอาไปเทียบกับตำแหน่งของ M6A ว่าตรงไหนทำนายถูก/ผิด → AUCROC

⇒ xPore ถูก confirm ว่าทำนายได้ดี เพราะสามารถบอก top position มี DRACH และยังบอกว่า ตำแหน่งหางของ RNA มักเกิด M6A ซึ่งตรงกับความรู้เก่า

⇒ ตอบโจทย์วัตถุประสงค์ว่า XPORE สามารถทำนายได้ว่า M6A อยู่ที่ตำแหน่งไหนบน RNA ได้อย่างความแม่นยำและยังค่อนข้างเพิ่มความรู้อีก

☐ Validation: M6A stoichiometry quantification

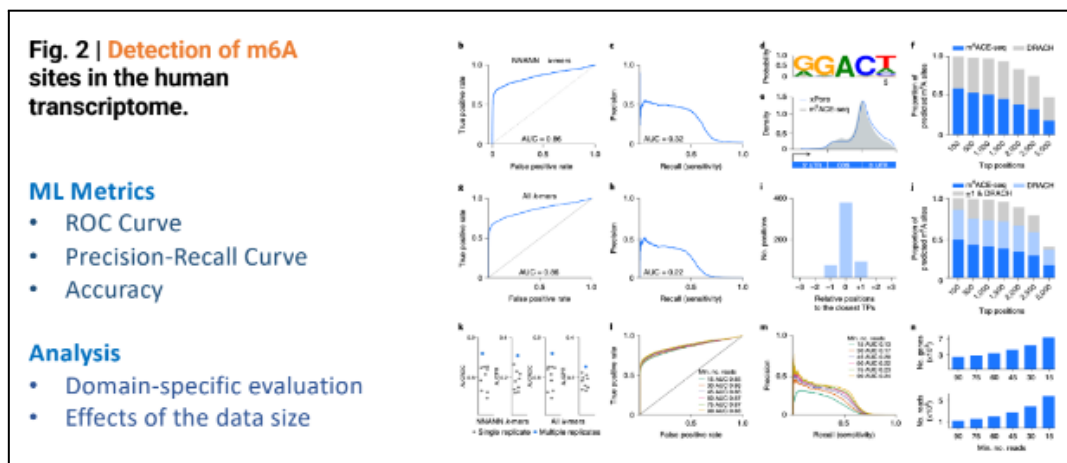


⇒ เราจะต้องหาให้ได้ด้วยว่า ตำแหน่งที่เกิดการ modified มีกี่ read

⇒ ลองใช้ xPore ดูว่า ตำแหน่งนั้นมี modification rate เท่าไหร่

⇒ เมื่อเทียบ ณ ตำแหน่งเดิม ถ้าเราไม่ใส่ wt modification rate \Rightarrow 0% และหากเพิ่มไป xPore ก็สามารถ estimated modification rate ได้ใกล้เคียงกับ เปอร์เซ็นการ mixture sample

☐ Validation: ML Metrics & Result Analysis



⇒ ภาพรวมของการ validation ก็ใช้ standard ML Metrics เลย

⇒ มีการ Analysis เพิ่มเติม

- Domain-specific evaluation เช่น
 - ⇒ ดูว่า motif มันตรงไหม
 - ⇒ ตำแหน่งที่ xPore predict ว่ามี M6A มันอยู่ที่ท้ายๆของ mRNA จริงไหม
 - ⇒ ดูถึง false positive ที่อาจจะไม่ได้ผิดซะทีเดียวเพราะมันคือตรงกับ DRACH motif หลายๆ %
- Effects of the data size
 - ⇒ ถ้า data size น้อย accuracy เป็นอย่างไร \Rightarrow method เราเวิร์คไหม

☐ Applicability

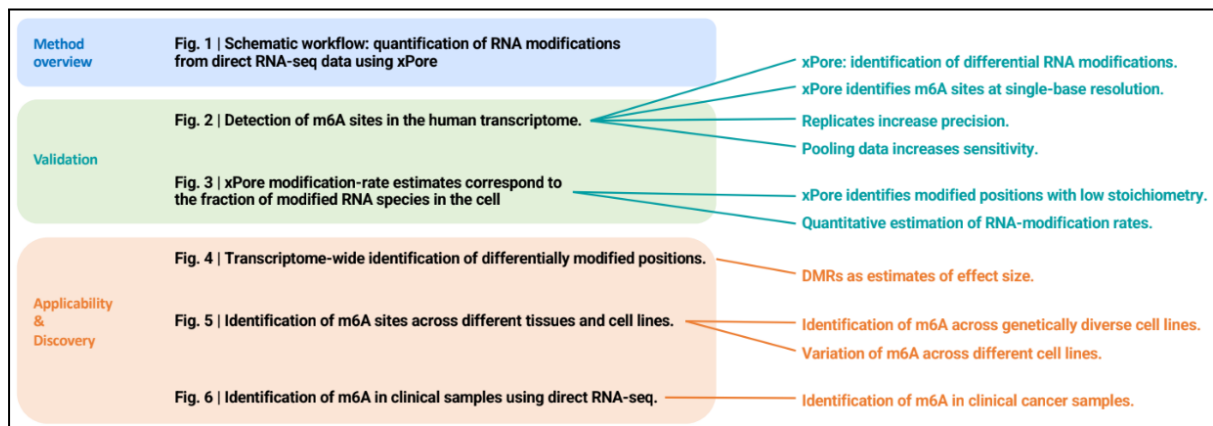
⇒ ถ้าเอา xPore ไปใช้กับ full dataset เค้าจะสามารถพบ gene ที่มี M6A อยู่ตรงไหน เป็นประโยชน์

⇒ add scenario compare other datasets \Rightarrow identification of M6A sites across different tissues and cell lines

⇒ เนื้อเยื่อ ดับ ไต ปอด หัวใจ ตรงไหนมี M6A

⇒ Clinical Dataset \Rightarrow using patient clinical samples หว่าตำแหน่งไหนบ้างที่ modified และไม่เหมือนคนอื่น

Visualization & Presentation



□ Storylining ⇒ ใช้ Figure เป็นหลักในการเล่าเรื่อง

⇒ เขียน outline เอาไว้ก่อนว่าในแต่ละ section เราจะพูดถึงอะไรบ้าง และงานนี้เขียน paper ไปพร้อมกับการทำ experiment เขียน idea + storyline ออกแบบเป็นสเตปมา ทำให้ไม่หลงทางง่าย ๆ

⇒ เน้นไปที่ method overview ที่อธิบายเกี่ยวกับภาพรวมของวิธีการทำงานวิจัยใช้ เป็นรูปหน้าปกของเปเปอร์ เมื่อคนมาดูรูป + อ่านคำบรรยายเล็กน้อย ก็สามารถเข้าใจว่าเราทำอะไร

⇒ ตอบโจทย์วัตถุประสงค์งานวิจัย ก็เลยโชว์ไปที่ Fig2. and Fig3. หา topic ของเรื่องเพื่อแบ่งประเด็นออกมา เพื่อ represent ในแต่ละ section

⇒ ส่วนสุดท้ายเป็นการ Applicability & Discovery เอาไปใช้จริงจะเกิดอะไรขึ้น ได้ผลอย่างไรบ้าง

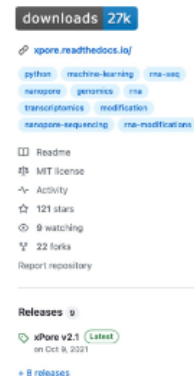
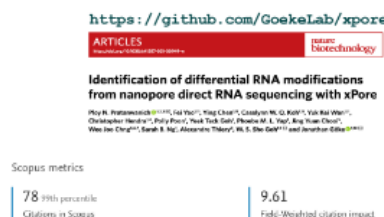
☐ Choosing the right plots

⇒ 6 รูป เล่าเรื่องเกี่ยวกับงานวิจัยเราทั้งหมด สำหรับคนที่สแกนอ่านวิจัยเราเร็วๆ ดูรูป + คำบรรยาย แล้วเข้าใจได้เลย ดึงความสนใจได้ง่ายกว่า text ใน 1 รูปมีหลายรูปย่อย และ เป็นเรื่องราวเดียวกัน

⇒ using another equipment สำหรับจัด layouts ของภาพ ให้มันมีความหมายเข้าใจง่าย แล้วค่อยสวยงาม

3 Key Success to Develop AI-Powered Apps


1. Alignment with the actual needs
2. Sufficient generalization and evaluation
3. Simple deployment and serving
 - Online documentation
 - Easy installation
 - Source code
 - Data availability
 - Lightweight
 - Fast

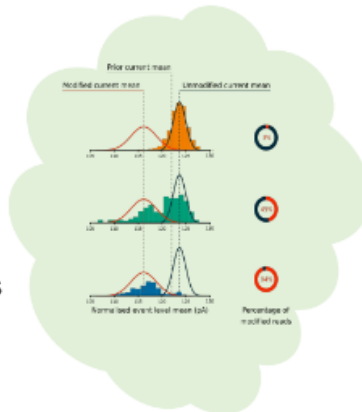


Future Work

Future Work

Domain-Oriented

- m6anet • 
- <Gaussian> mixture model
- Interpretability
 - Modification or basecalled errors
- End-to-end
 - Why?
 - Nanopolish eventalign / Guppy basecaller are subject to change



Method-Oriented

- Deep autoencoder + GMM
- CNN + GMM
- Other models + GMM

1. ระบุ limitation ของงานวิจัยนี้ให้ได้
2. พิจารณา changes in the future ที่เกี่ยวข้องกับ software เพื่อเตรียมแก้ไขปัญหา

☐ Domain-Oriented ต่อยอดเกี่ยวกับ การหา Nanopore หรือ identified RNA modification ให้ดียิ่งขึ้น ทำอย่างไร?

⇒ อีกงานหนึ่ง ที่ชื่อว่า m6anet ⇒ supervised learning ⇒ ใช้ผลลัพธ์ที่ได้จาก xPore ไป train model อีกตัว

⇒ limitation is gaussian ซึ่งอาจจะไม่เหมาะ อาจจะ sensitive กับ long tailed distribution ทำให้ค่า μ ถูก shift ไป และส่งผลให้ model ไม่ค่อยแม่นยำ ⇒ **อาจจะเปลี่ยน distribution**

⇒ interpretability

- การที่ current ไม่เท่ากันถึง 2 อัน มันเป็น modification (type ที่ไม่ใช่ M6A) หรือ basecalled errors ⇒ filter basecalled error ออกไปยิ่งยิ่งได้บ้าง
- ถ้าในอนาคต Software เปลี่ยน เช่น guppy หรือ nanopolish สำหรับ segment เปลี่ยน ⇒ สามารถหาโมเดลที่ไม่ต้อง segment pattern ของไฟฟ้า สร้างโมเดลแบบ end-to-end คือ **ขึ้นกับ software อื่นๆได้น้อยที่สุด** ใช้แต่โมเดลเลย

☐ Method-Oriented ต่อยอดเกี่ยวกับ โมเดล GMM

⇒ งานนี้เราได้เรียนรู้ GMM อย่างลึกซึ้ง ถ้าเราสามารถผนวกความรู้ของ GMM เข้ากับ โมเดลอื่นๆได้ และสามารถนำไปต่อยอดเป็นโมเดลใหม่ที่สามารถแก้ปัญหาก็ไม่เคยถูกแก้มาก่อนได้

WorkShop

Q & A

ทำไมต้องใช้ Gaussian Mixture Model \Rightarrow หลังจากไปศึกษาโดเมนเลยรู้วิธีการ preprocess สัญญาณมันออกมาในรูปแบบ gaussian มันเร็ว ไม่ซับซ้อน ไม่ต้องการ training data จำนวนเยอะ ตอบโจทย์วิจัย ต้องการ modification หลาย samples Data นี้ยังไม่เคยมีใครใช้

Method ก็ต้องเป็น Interpretable เพราะร่วมกับนักชีววิทยา

Literature หลายงานก็ใช้ GMM

เทคนิคในการหาไอเดียที่สามารถสร้าง impact กับสิ่งต่างๆ สนใจ health science / AI วินิจฉัยโรค \Rightarrow การที่เราเห็นผลงานของคนอื่นเยอะ ๆ เราจะได้ไม่ไปทำซ้ำ เราต้องทำได้ดีกว่า สืบจลตลาด การจะทำ product ก็ต้องเน้นที่ pain point ของมัน ใช้ภาพหรือใช้อะไร แล้วเวลาทำ literature review ควรจะลึกลงไปเลย หรือกระโดดข้ามสายไปเลยก็ได้ เช่น การวิเคราะห์ลักษณะบุคคล ธนาคารก็มีการวิเคราะห์ สุ่มอ่านเอา แล้วเรามองเราจะรวบรวมปัจจัยต่างๆเพื่อหาไอเดียให้ตัวเองได้ อ่านทั้งแนวลึกและแนวกว้าง

Output ออกมาเป็นเชิงรูปภาพหรือว่าเป็นคลื่น \Rightarrow ออกมาเป็น list, array พอเอามาพล็อตก็ออกมาเป็น คลื่นสัญญาณไฟฟ้า 1 เส้น พอมาพล็อต hist ดู frequency ทราบ variance คล้าย normal distribution

Data Collection เอามาอย่างไร \Rightarrow เค้าเอามาให้ เป็นไฟล์ FAST5

RNA modification percentage \Rightarrow DNA เป็น gene ถูก transcribe ออกมาได้ RNA หลายสาย ตั้งสมมติฐานว่า ไม่ว่าตำแหน่งไหนที่มัน modify ทุกเส้น modify หมด

Case: เอาสาย RNA ที่ 100% modified + RNA 50% modified ผสมกัน 1:1 \rightarrow 50% modification

การบ้านใช้ synthetic data (สมมติเอา) จะต้องเหมือนกับของตัวอย่างของอาจารย์ไหม? \Rightarrow มันเป็น AI ที่เดาค่า mean กับ variance ไม่ถูกต้อง 100% ดังนั้นมันสลับกันได้ ผิดพลาดได้ แต่ต้องใกล้เคียงกับข้อมูลตัวอย่าง \rightarrow เป็นปัญหาที่เกิดขึ้นสำหรับการ clustering ในการรันแต่ละครั้งมันจะให้ผล output มาไม่เท่ากันอยู่แล้ว

Timeseries \Rightarrow ควรใช้ trend หรือ slope เป็นอีก feature หนึ่ง ทำไมเดอะไรก็ได้ที่สามารถ เรียนรู้ความถี่ได้ แพทเทิร์นของ time series นั้นๆได้ แต่การหา corr ของ time series \Rightarrow ต้องไปหาเอางานอื่นๆ

อ.คิดว่าตัวโมเดลสามารถนำไปใช้ในที่มีข้อจำกัดด้านทรัพยากรคอมพิวเตอร์ไหมคะ เช่น พื้นที่ห่างไกล \Rightarrow ของอาจารย์ใช้ CPU ไม่เยอะ nanopore สามารถใช้ได้โดยไม่ต้องต่ออินเทอร์เน็ต

การใช้ข้อมูล RNA มีข้อแตกต่างกับใช้ protein อย่างไรบ้างคะ เนื่องจากว่าที่หนูเข้าใจคือได้จากการ express ของยีนเหมือนกัน \Rightarrow protien เหมือนเป็นการ express ของ RNA บางโรคมันจะลึกถึงระดับ RNA

การศึกษา field นี้ ประยุกต์ใช้ใน precision medicine ได้ด้วยไหมคะ เนื่องจากใช้ข้อมูล genetic ของแต่ละบุคคล \Rightarrow ถ้าเรารู้ว่ามัน modification เกี่ยวกับโรคนั้น ค่อยเชื่อมโยงไปที่ medicine ได้ แต่มันยากหน่อย เพราะใช้กับคนจริงๆ อาจจะต้องทดลองใน lab \Rightarrow animal model \Rightarrow ผ่าน clinical trial ต่างๆ แต่ xPore ช่วยสกรีนก่อนเฉยๆ

สามารถเอาโปรเจกต์นี้ไปทำจบป.เอกในสาขาที่เกี่ยวข้องกับโปรเจกต์เราได้เลยมั้ยครับหรือว่าต้องเรียนผ่านวิชาต่างๆของโทแล้วก็เอกหรือครับ \Rightarrow ตอนนี้ก็มีหลักสูตรจบตรี แล้วต่อเอกทำวิจัยจบได้เลย แต่ระวัง plagia ต้องมี reseach skill

เทรนสายงานด้าน bioinformatics ในไทยเป็นยังไงบ้างคะ \Rightarrow มี startups tools สำหรับการวิเคราะห์ ใช้ bioinfo เยอะ แต่ในไทยยังน้อย ยังอยู่ในระดับ reseach แต่ในตปท. มีมานานแล้ว \Rightarrow ตอบเชิงไม่ค่อยโต โตช้า