MakeElevator

Tyler Weirick, PhD
What is MakeElevator?

• A blockchain-based protocol for communicating computational workflows.
• Increases the speed, reproducibility, and accessibility of computational workflows.
• Intended for use with “workflow managers”.
• Consists of:
  • A set of rules for formatting and organizing input and output data.
  • A set of rules for controlling write access to the blockchain.
  • The InterPlanetary File System (IPFS)
  • BigchainDB
• Name comes from Snakemake & space elevators.
• Functions by associating hashed key-value pairs.
  • Ex: "QmamX8N8Gpm3yjxvYbQVkJL" : "QmR56UJmAaZLXlDT1ALrE"
• Designed for bioinformatics, could also be interesting for publishers.
What is Bioinformatics?

• An interdisciplinary field based on using data to study biology.

• Examples:
  • The human genome contains ~3 billion basepairs, we need computer science to work with data on this scale.
  • Is a variant of gene is associated with a disease? We need statistics to make inferences.

• Heavily utilizes software and large amounts of data.
  • I feel bioinformatics doesn’t quite fit the traditional publishing model.
  • MakeElevator is intended to improve our ability to do bioinformatics research, but cloud also address some of these publishing related problems.
Evolutionary Landscapes

• **Evolutionary landscapes**: a very abstract metaphor used to conceptualize and visualize evolution.

• Ex: A gene impacting muscle building speed.
  - x & y-axis = a high dimensional space representing all possible nucleotide sequences for a gene & environmental factors.
  - z-axis = fitness (chance to reproduce)

• Blockchains also have evolutionary landscapes i.e. “motivational landscapes”.

• Blockchains are useful tools for modifying motivational landscapes via previously unavailable game theory applications.

• What is the motivational landscape for MakeElevator?
  - First we need to introduce a different kind of landscape...

Depending on the environment building muscle faster or slower (during a famine) might be preferred.

https://whattomine.com/
Perception Landscapes & Intelligent Agents

- **Agents** - anything that perceives its environment via sensors and acts upon its environment via effectors.
  - Humans, bacteria, thermostats, etc.

- Agents have perception and action “surfaces”.
  - Ex: Robot Vacuum
    - **Actions**: \{Move-Left, Move-Right, Vacuum, Wait\}
    - **Precepts**: \{Dirty, Clean, Left-Square, Right-Square\}

- Humans also have a limited perception landscape.

- Bioinformatics is focused on our perception landscape.

Artificial Intelligence: A Modern Approach

We can only observe things within a radius of earth. What’s outside?
Perception Landscapes & Science/Technology

- **We can increase our perception landscape with science.**
  - Ex: Cosmic Distance Ladder - In addition to seeing stars we can now know their distance from earth.

- Not limited to large things, we can see small things too, or even things that happened long ago.

- Expanding our perception is important as it enables us to better interact with our environment.
Complexity another factor of perception landscapes

• Biotechnology where is my cool stuff?
  • Smartphones - pocket size computers more powerful than those in spaceships that went to the moon. It can even tell me where I am thanks to satellites and the special and general theories of relativity.
  • We can detect planets in other solar systems, yet we are biological entities.
  • It’s not a matter of size, we can see viruses.

• Complexity!
  • Even when the answers are in front of us, we need logic to see them.
  • This is the kind of perception landscape we want to expand via bioinformatics.
  • I’m creating MakeElevator to improve this perception landscape.

Mr. Bill, Chris, Sam, and Tim each have a son. One of the sons and fathers each work as a baker, carpenter, smith, or tailor.

1. No one's job and name share the same first letter.
2. No son has the same job as his father.
3. Bill has the same job as Chris's son.
4. Sam's son is a baker.
Actually, there is plenty of cool stuff happening in biotech.

• Predicting faces from genomes.

• Cancer vaccines

• Cardiovascular regeneration!
  • The field I work in!
  • Other animals like fish can regenerate their heart if damaged. Why can’t humans?


More Specifically: IncRNA and Epitranscriptomics in Cardiovascular Research

- RNA-seq – similar to whole genome sequencing, except measuring the RNAs present within a sample.

- Long non-coding RNA
  - Large RNAs that do not code for proteins.
  - Complex controllers of gene expression.
  - Numbers increase in higher organisms.
  - Highly tissue specific expression.

- We study IncRNAs to uncover the gene expression “program” needed for regeneration.

- RNA-seq data is made freely-available online after being published in a research paper.

- RNA-seq has high re-use value as it captures all RNAs in a sample.

- We can study things not considered in the original study!

- The more samples we have, the better our perception landscape!

• RNA Editing
  • Post-transcriptional (after the RNA is created) modifications.
  • There are actually more than four bases (A, C, G, T/U)
  • Some types of editing detectable within standard RNA-seq data!
    • Even more utility from re-analysis!

• Relation to Cardiovascular Research
  • A-to-I editing increases in atherosclerotic plaques.
  • Increased editing in MED13 during congenital heart disease.
  • Nucleocytoplasmic shuttling of editing enzyme in newt heart during regeneration.

One Last Problem: Agents and Error

• Remember Vacuum World? Here is a slightly more complicated example.
  • Maze traversal with simulated sensor and actuator error.

• Error rapidly reduces perception and action surfaces.
  • High sensing error – mistake position in maze
  • High action error – run into walls

• Closer to doing science than you might think.
How can we expand our perception landscape in bioinformatics?

• Things the community is already doing:
  • Sharing Data.
    • Unprocessed high-throughput data stored by NCBI.
    • Scripts used to run analyses.
  • Encourage use of open source software.
    • Much easier to find bugs.

• Things we could do better.
  • Reducing self-hosting – programs and data missing when sites no longer maintained.
  • Even if analysis scripts are shared, it is difficult to determine if the programs ran correctly.
    • **Ex: Too many observations for humans to check.**
    • **Workflow managers are a helpful solution.**

Self-hosting: Even I have tools which are no longer available.
What are workflow managers? Example: Snakemake

• Instead writing explicitly, share a file with rules for each step and a configuration file describing the desired output.

A Bash Script

```bash
# Pipeline

# rule I:
input: "{id}.x"
output: "{id}.A"
shell: "cmd1 {input} > {output}"

# rule II:
input: "{id}.y"
output: "{id}.B"
shell: "cmd2 {input} > {output}"

# rule III:
input: A="{id}.A", B="{id}.B"
output: temp("{id}.C")
shell: "cmd3 {input.A} > {input.B} {output}"

# rule IV:
input: "{id}.C"
output: "{id}.D"
shell: "cmd4 {input} > {output}"
```


```bash
# rule I:
input: "{id}.x"
output: "{id}.A"
shell: "cmd1 {input} > {output}"

# rule II:
input: "{id}.y"
output: "{id}.B"
shell: "cmd2 {input} > {output}"

# rule III:
input: A="{id}.A", B="{id}.B"
output: temp("{id}.C")
shell: "cmd3 {input.A} > {input.B} {output}"

# rule IV:
input: "{id}.C"
output: "{id}.D"
shell: "cmd4 {input} > {output}"
```

Are all these commands correct?
Features of Workflow Managers

• Easier to read.
• Easy re-use for new analyses, just modify the file describing which samples to run.
• **Built in error detection.**
• Integration with HPC and cloud environments.
• Automatically detects parts of the pipeline that can run in parallel.
• Optionally delete intermediate files.

Unfortunately workflow managers still aren’t very popular 😞
Can’t we just trust scientists like in a biological paper?

• Many observations in bioinformatics aren’t as concrete as you might think.

• The locations and even existence of many genes has changed over time.
  • Ensembl protein-coding genes.
  • The number of human protein-coding genes has changed in every GENCODE version!

• Dead mice tend to stay dead.
Not just over time, also between reference annotations.

• Guided alignments – Use a reference genome and list of genomic features to improve accuracy and speed.

• Gene structure – The start, stop, exon, and intron locations.

• If the size of a gene changes this can affect the detected expression level.

• Most genes do have matching locations and most research papers would require other validation. So not dire, but still concerning.
What can we do to combat these problems?

• Make it trivial to report exactly what was done.
• Remove the need for self-hosting.
• Make re-analysis easy enough that automated systems can perform them.

MakeElevator!
What is the IPFS?

• A protocol and network designed to create a content-addressable, peer-to-peer method of storing and sharing in a distributed file system.
  • **No need for centralized DNS!**

• Functions via hashing files and addressing contents via a Merkel tree.
  • **If you have the hash you can download the file!**
  • **Simplifies file sharing without a need for centralized servers!**
  • The technology behind Filecoin.

• **Problem: We can share files but we can’t connect inputs to outputs.**

Example IPFS Hash:
QmU23breoah19eqYv6mymNrfgUmarxqrBJnKLHvBxd8Q2h
What is BigchainDB?

• A scalable blockchain database.
  • **Merges the best of two worlds: the “traditional” distributed database world and the “traditional” blockchain world.**

• Distributed Database Characteristics
  • Scale (high-throughput, capacity, low latency)
  •Queryable

• Blockchain Characteristics:
  • Decentralized (no single entity owns or controls it)
  • Immutable (tamper-resistance)
  • Assets (yours if you own the private key)
With the combination of IPFS and BigchainDB

• Basically, we just want to store and retrieve key-value pairs!
  • Ex: { Input: "Qma...", Output: "QmR5...", Metadata: [ ... ] }

• Key value pairs are comprised of the input data and arguments used to run the program
  • Ex: Consider the file “S1.x” and rule “I”

1. Generate the command.
   cmd1 S1.x > S1.A

2. Hash the initial components of the command. Output files get special reserved word hashes.
   Cmd1 = “Qm1111”
   S1.x = “Qm2222”
   S1.A = “QmOut1”

3. Rebuild command with hashes
   Qm1111 Qm2222 > QmOut1

4. Hash the command built from hashes.
   Qm1111 Qm2222 > QmOut1 = “QmInput”

5. Search for the output hash.
   If found: download.
   Else: run original command, and report hash of resulting output.

**rule I:**
input: "{id}.x"
output: "{id}.A"
shell: "cmd1 {input} > {output}"
Benefits

• Faster analyses.
  • Download instead of running.
  • Hash skipping – only download final file.

• Data permanently preserved.

• Share analyses with a hash.

• Permanent non-cheatable record of analyses.

• Allows random testing of large projects.
Demo: Simple Snakemake Pipeline

• Simple RNA-seq assembly pipeline.
  • Simulated files and programs.
  • Programs sleep for 1.5 seconds then calculate md5 sum.

```python
samples = ["sample1", "sample2"]

FASTQ_FILE = "pseudo/initial_files/{sample}.fastq"
TRIMMED_FASTQ = "pseudo/generated_files/{sample}.trimmed.fastq"
SAM_FILE = "pseudo/generated_files/{sample}.sam"
GENE_EXP_FILE = "pseudo/generated_files/{sample}.gene_exp.tsv"

rule all:
  input:
    expand(GENE_EXP_FILE, sample=samples)

rule Trim_Fastqs:
  input: FASTQ_FILE,
  output: TRIMMED_FASTQ
  shell: """"./pseudo/programs/trim {input} > {output} """

rule Align_Reads:
  input: TRIMMED_FASTQ,
  output: SAM_FILE
  shell: """"./pseudo/programs/hisat2 {input} > {output} """

rule Assemble_Reads:
  input: SAM_FILE
  output: GENE_EXP_FILE
  shell: """"./pseudo/programs/stringtie {input} > {output} """
```
ls -l pseudo/generated_files/

rm pseudo/generated_files/*
cat first_md5sums.txt
Is there really enough re-analysis happening to make this valuable?

• Some processing steps are extremely common.
  • Reference files – genome sequences, gene locations, variant information.
  • Indexes
    • Data structure built from reference files.
    • Help speed up data access for aligner.
    • Building indexes is time consuming.
    • Can require expensive computing resources (Ex: 500GB RAM)

• Re-analysis for studying new phenomena.
  • IncRNAs
  • RNA Editing

---

Further improving the motivational landscape.

• We can create a citation-like system via creative use of the metadata in BigchainDB assets.
  • Capable of “citing” a variety of things.
    • Raw data.
    • Processed data – both original submissions & replications.
    • Seeding (downloading) data.
    • Software.

• The “citation” score can also be used as a “replication score” we can use to decide if we want to re-run or download a process.

• Ultimate goal is enabling users to earn money by participating.
  • A motivation landscape which improves the perception landscape!
  • Some sources of inspiration – Steemit, storage coins (Filecoin, Storj), Decred
  • Academia & red tape.
  • Would love to hear others thoughts.
How is this useful for journals?

• Journals add value through ensuring quality observations and corresponding causal statements.
  • Until recently, journals added value by warehousing data, but it is increasingly easy to host large amounts of data online.

• Improving the quality of computational analyses in papers.
  • Exact details and files for each step easily and instantly available.
    • Hashes of whole analyses can be described within a journal article.
  • Re-analysis to increase reliability.
  • “Living” review articles
    • Bioinformatics software bakeoffs
      • Which program is the best?
      • This can change with a single update.
Summary

- MakeElevator is intended to create a motivational landscape that will promote increasing our perception landscape in computational sciences.
  - A massive amount of easily accessible data will help with the complexity problem.
  - A “Cosmic Distance Ladder” for biology.
  - RNA-seq is like a movie compressed averaged into a single frame.
  - There is a surprising amount of deductive logic we can to with RNA-seq data.
    - A-to-I RNA editing only occurs in double stranded RNA – Could we predict structure or even age of RNAs?
    - IncRNAs are highly tissue specific – Could we predict the cellular composition of a tissue sample?

- For scientists:
  - Faster computational pipelines.
  - Another way to get attribution,
    - Especially for currently untracked, but useful contributions.
  - Possibly monetizable in the future.

- For journals:
  - Helps increase the quality of bioinformatic analyses.
  - Improves reproducibility.
  - Reduce file hosting cost.
Thanks for listening, Questions?

• Thanks to:
  • Shizuka Uchida, PhD
  • UofL Institute of Molecular Cardiology

• Like my presentation? I’m looking for work.
  • I am currently a post-doc, but my lab moved during my PhD and wife became pregnant shortly before defending.
  • Now we have a healthy baby so I am looking for new opportunities.