Towards a Model of Provenance and User Views in Scientific Workflows

Shirley Cohen

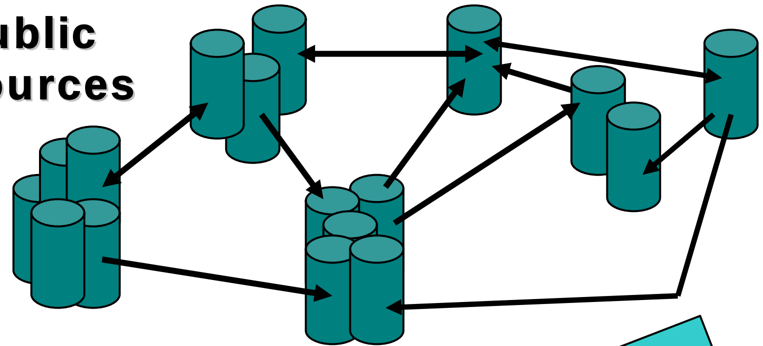
Sarah Cohen-Boulakia

Susan Davidson

University of Pennsylvania

Need for provenance!

Public sources



Bioinformatics protocols

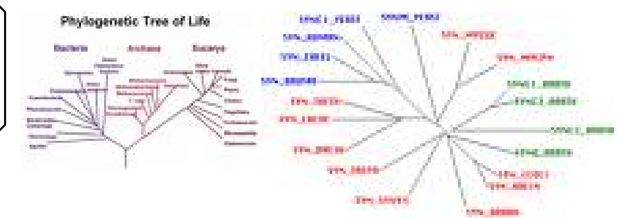
ATGGCCGTGTG
GCTAAATGTCT
GTGCCTAACTA
ACTAA...

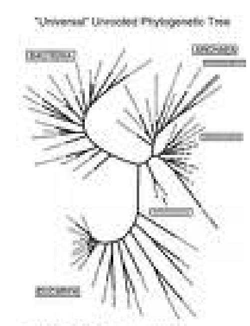
Alignments
ClustalW
PAUPS
Phillips
Bootstrap
...

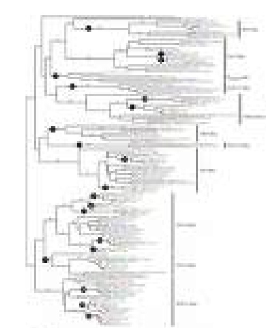
How this tree has been generated?

Which sequences have been used to produce this tree?

Can I throw away some of these data? Which ones are really important to keep?







Biologist's workspace

CIPRES project

Cyberinfrastructure for Phylogenetic RESearch

Scientific Analysis

- Explosion of biological data, must be analyzed to create knowledge
- Scientific analysis is complex
- Reproducing, interpreting results depends on the **provenance** of the data (how, where, who…)
- Workflow systems
  - Support scientists in their analysis
  - **Trace** the data used / generated at each step
  - Are **heterogeneous**
    - Different graph-based models
    - Different technologies

➔ Need a generic **model of provenance**

Provenance

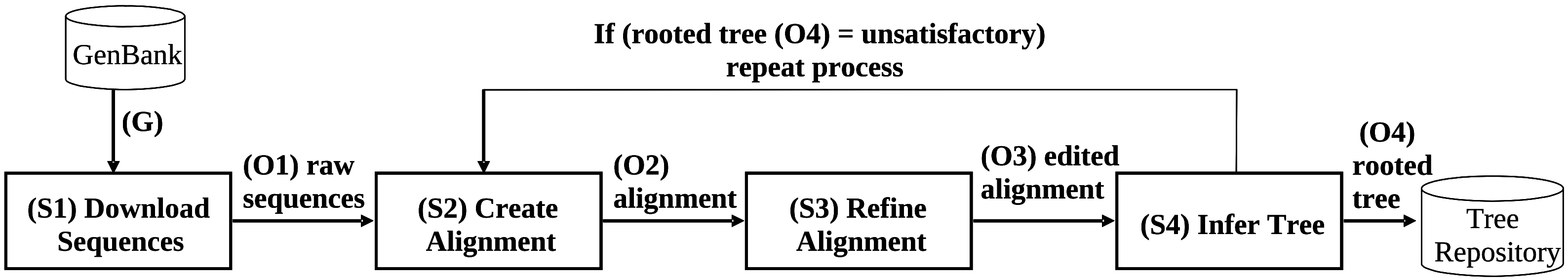
- Provenance is an increasingly important topic
  - specialized workshops, survey papers…
- Models for data provenance exist in the database community
  - E.g. [Buneman *et al.*,01], [Bhagwat *et al.*,04], [Widom *et al.*,06]
- However, several features of scientific workflows are not addressed
  - Data are derived by **chaining** and **composing** analytical tools
  - Steps are **black boxes**
  - Different **views** of a given workflow (sub-steps) may be considered

➔ Model of provenance for scientific workflows must incorporate these features

Outline

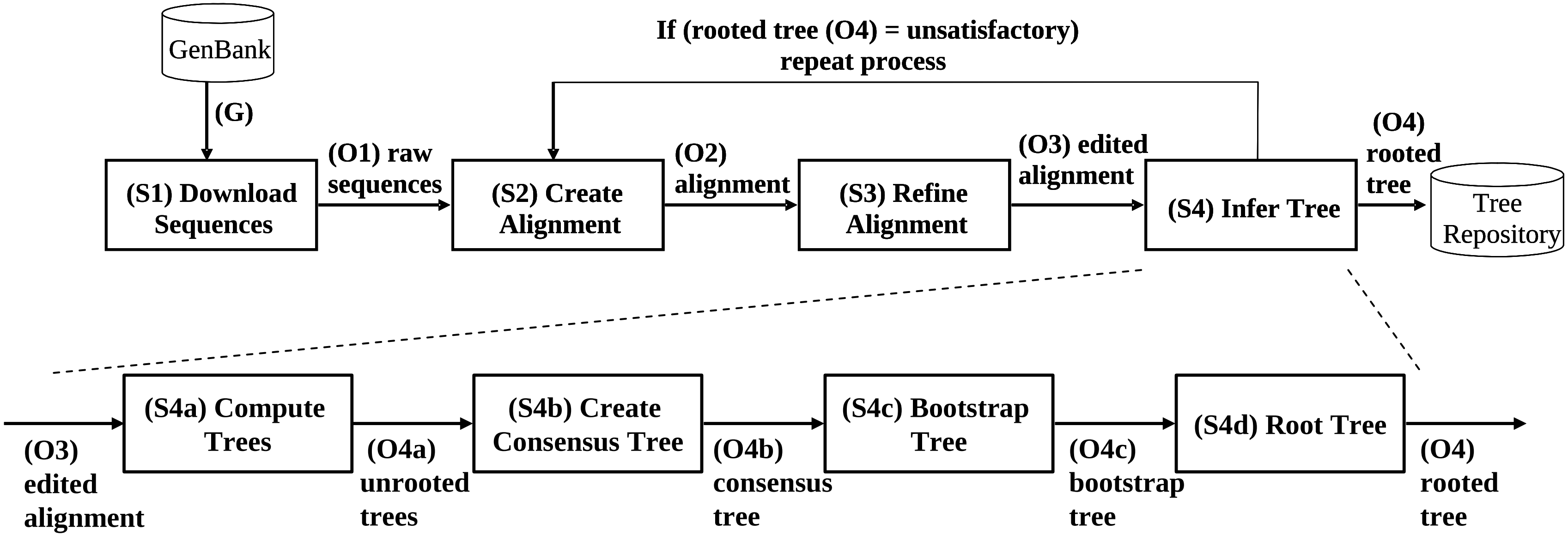
- Motivation
- **Case study: Tree Inference**
- Model for provenance and user views
- Querying provenance
- Conclusion

Tree Inference Workflow



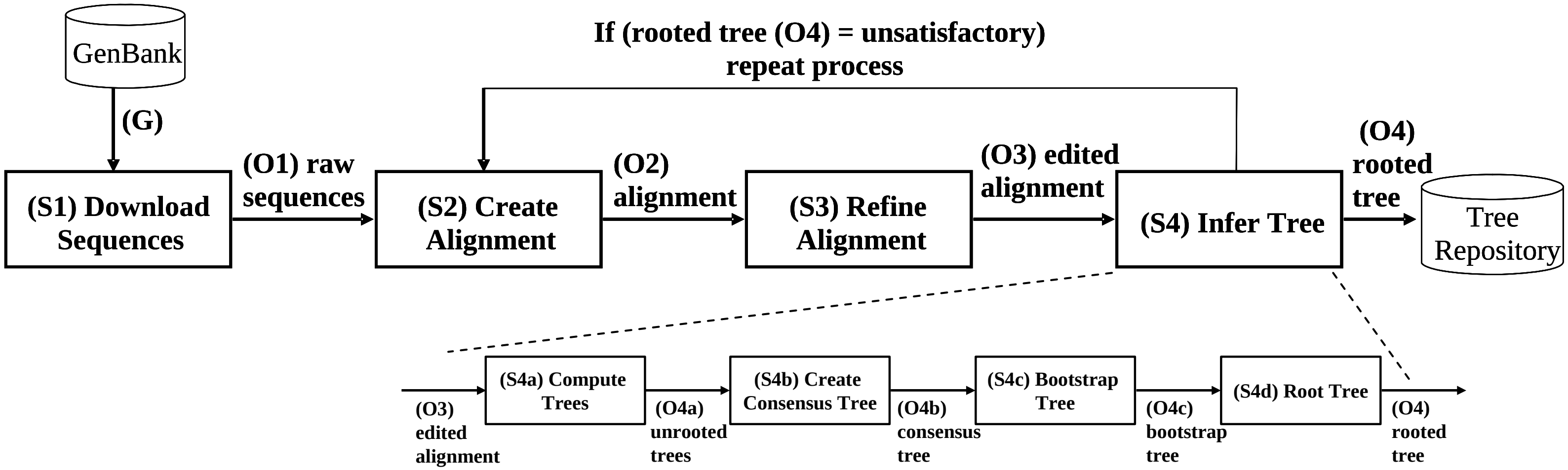
- Designed in the context of the *CIPRES* project
- Represents how phylogeneticists analyze data
- Terminology
  - Nodes are **step-classes** (static)
  - Edges capture the **flow of data** between step-classes
    - Loops are possible
  - An **execution** of a workflow generates a partial order of steps (dynamic)
    - Instances of step classes
  - Each step has **input** and **output** data

Tree Inference Workflow, cont.



- A step-class may itself be a workflow
- Users may zoom-in to the boxes
  - Kepler, myGrid…
- Different **user views** can be considered
  - Am I allowed to zoom in S4?

Querying Provenance



- From what **immediate data products** did this tree originate?
- What are **all the data products** which have been used to produce this tree?
- What **step** produced this tree?
- What **sequence of steps** produced this tree?

➔Data vs step provenance

➔Immediate vs deep provenance

Outline

- Motivation
- Case study: Tree Inference
- **Model for provenance and user views**
- Querying provenance
- Conclusion

Model of Provenance: Logs

- A log is a sequence of entries
  - Input(sid,iid,ts) *sid takes iid as input at time ts*
  - Output(sid,did,ts) *sid produces did at time ts*
- Immediate provenance
  - All the data and steps directly used to produce did

ImmProv(did,sid,iid):- Input(sid,iid,tsi) $\wedge$ Output(sid,did,tso) $\wedge$ tsi $\leq$ tso

ImmDProv and ImmSProv are also defined

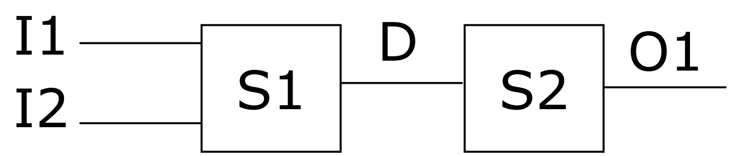
Input

SID	IID	TSI
S1	I1	1
S1	I2	1
S2	D	3

Output

SID	DID	TSO
S1	D	2
S2	O1	4

Each input/output data is stored!



Imm. Provenance of O1

ImmDProv: D

ImmSProv: S2

Deep Provenance

- **Recursive** definition
  - **Deep Data** provenance (D):

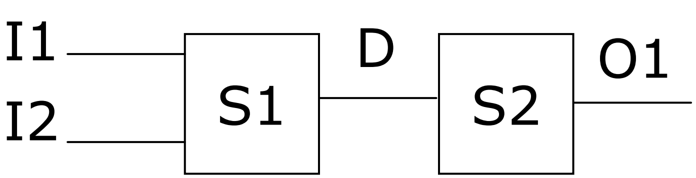
    DProv(did, iid):- ImmProv(did,_, iid)

    DProv(did, iid):- ImmProv(did,_, x) $\wedge$ DProv(x, iid)

  - **Deep Step** provenance (S):

    SProv(did, sid):- ImmProv(did,sid,_)

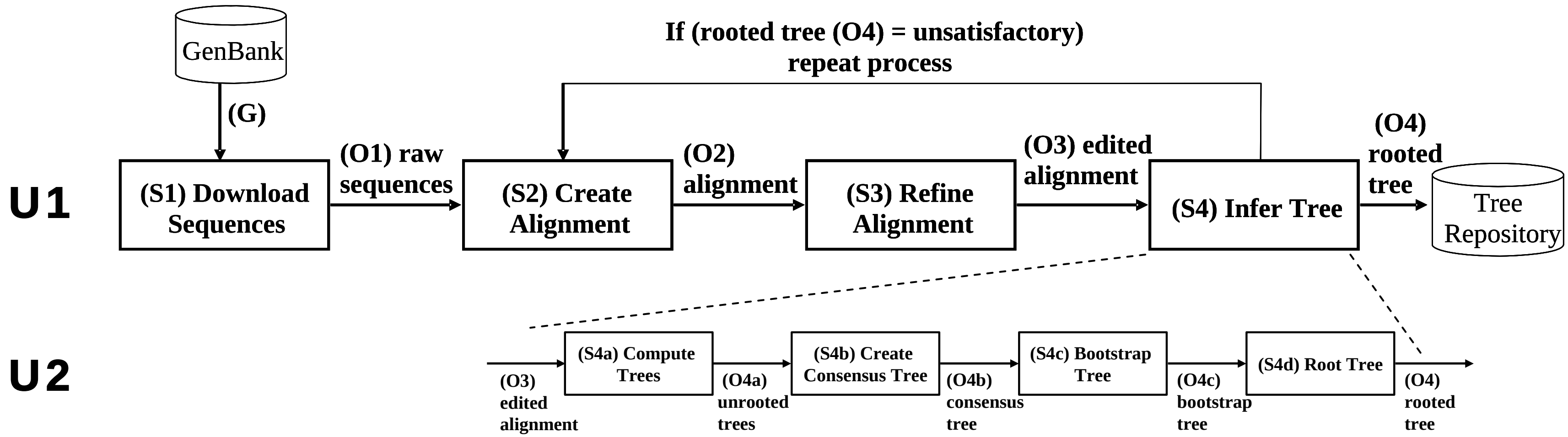
    SProv(did, sid):- ImmProv(did,_, x) $\wedge$ Sprov(x,sid)



DProv for O1: [{D}, {I1, I2}]

SProv for O1: [{S2}, {S1}]

Composition and User Views



- What is the immediate data provenance of O4?
  - If I can zoom into S4 ➔ O4c
  - Otherwise ➔ O3
- **UserView(U):** set of the lowest level step classes that U is entitled to see.
- **Ordering on user views: U2 $>_u$ U1**

  U2 is finer than U1 (sees provenance in more detail)

User Views

- **What** are User views?
  - Level of **detail** the user wishes to track
  - **Permissions** given to the user
  - **Ability** of the user to see / know the sub-steps (distributed computation)
  - Similar to **checkpoints** in logs

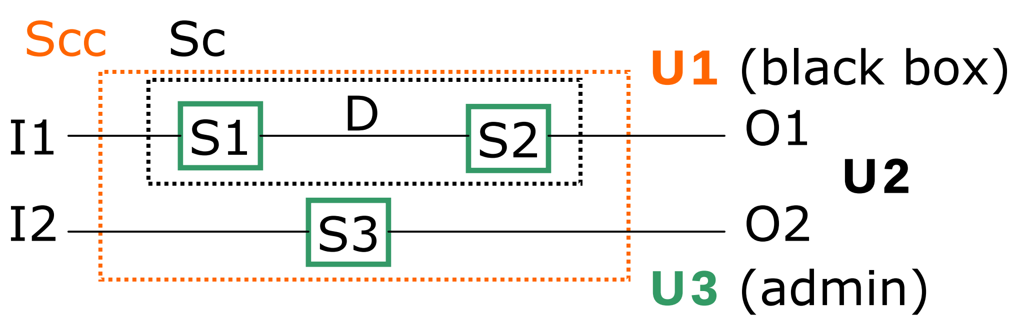
- **Why** use User Views?
  - **Throw away** unimportant intermediate results
  - **Reduce** the amount of work to be redone

➔ Storage **efficiency**

Reasoning with User Views

- Logging occurs at lowest level steps
- Reasoning uses information from
  - Workflow: Step-classes containment and user views
  - Cinput(sid,idid,tsi), Coutput(sid,idid,tso) calculated from log
- Immediate user-provenance
  - ImmUserProv(u,did,sid,idid):- Cinput(sid,idid,tsi) $\wedge$ Coutput(sid,did,tso) $\wedge$ tsi$\leq$ tso $\wedge$ **userView(u,sid)**

➔ ImmUser**D**Prov
➔ ImmUser**S**Prov



CInput

SID	IDID	TSI
S1	I1	1
Sc	**I1**	**1**
Scc	I1	1
S3	I2	1
Scc	I2	1
S2	**D**	**3**

COutput

SID	DID	TSO
S1	D	2
S2	**O1**	**4**
Sc	**O1**	**4**
Scc	O1	4
S3	O2	5
Scc	O2	5

ImmUserDProv for O1 viewed by U2: {I1}

ImmUserDProv for O1 viewed by U3: {D}

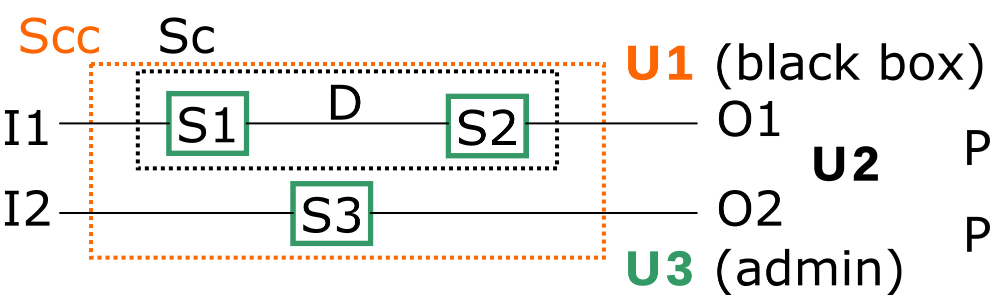
- User Deep provenance is analogously defined

Reasoning with User Views (cont.)

- A finer **user view** allows
  - more data and steps to be seen
  - more precise reasoning about data provenance
- **Lemma**

  Given a data object did and two user views u1 and u2 such that u1 $<_u$ u2 and *did* is *visible* in *u1. Then*

  Prov-visible(u1,u1,did) $\supseteq$ Prov-visible(u1,u2,did)



Prov-visible(U1,U3,O1)={I1}

Prov-visible(U1,U1,O1)={I1,I2}

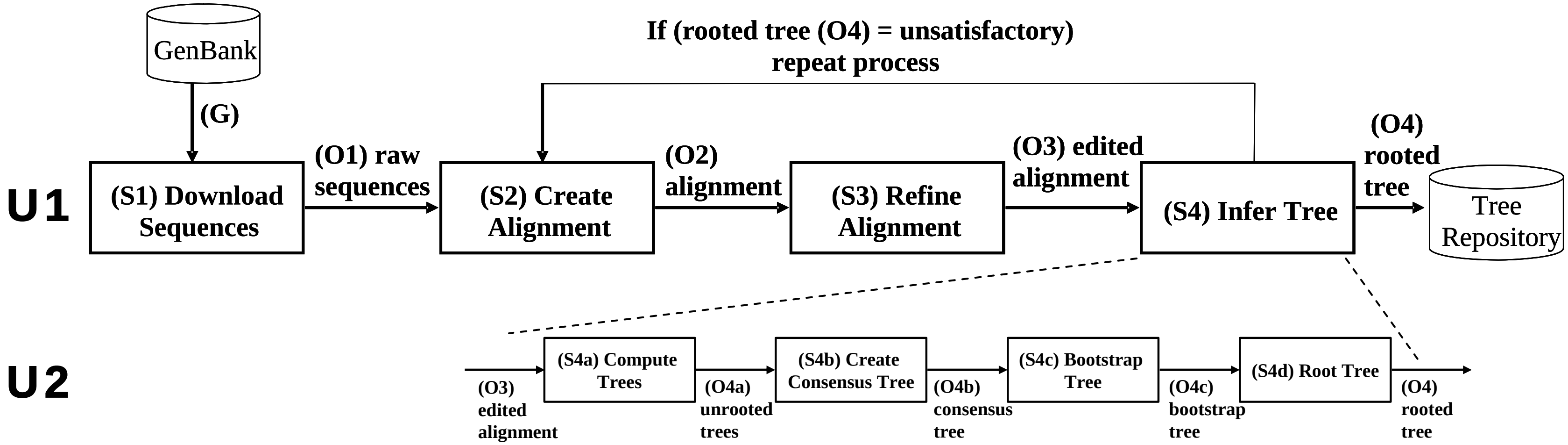
➔ Different **granularity** levels of provenance

➔ Storage **efficiency**

Outline

- Motivation
- Tree Inference use case
- Model for provenance
- **Querying Provenance**
- Conclusion

Querying Provenance



- *From what direct data products did this tree originate?*

  ImmUserDProv (U1,O4): O3

  ImmUserDProv (U2,O4): O4c

- *What are all the data products which have been used to produce this tree?*

  userDProv (U1,O4): O3,O2,O1,G

  userDProv (U2,O4): O4c,O4b,O4a,O3,O2,O1,G

- *What sequence of steps produced this tree?*

  userSProv (U1,O4): S4,S3,S2,S1

  userSProv (U2,O4): S4d,S4c,S4b,S4a,S3,S2,S1

Conclusion

- Model of **provenance**
  - Based on study of user requirements (**Tree Inference Workflow**)
  - Uses **generic** and **minimal** information
    - Based on careful studies of workflow systems (Kepler, MyGrid, Chimera)
- Definitions include
    - **Data** and **Step** provenance
    - **Immediate** and **Deep** provenance
- **User Views**
  - Multi-**granularity** levels of provenance
  - Only *visible and necessary* data are kept

  ➔ **Efficiency** in storage
- Model is rich enough to answer the collected queries

Ongoing Work

- Experiment with the **expressiveness** of the language
  - Queries over concurrent and partial executions
  - Use an object-oriented data model (JDBC/Oracle)

- **Implement** the model (efficiently)
  - Experiment with storage models
  - Collect real scientific logging information
  - Study use within in real workflow system
    - Collaboration with the Kepler group

Acknowledgements

- Kepler Group
  - Shawn Bowers
  - Bertram Ludascher
  - Timothy McPhillips
- Biologists from the *CIPRES* project
- Members from the Database group, University of Pennsylvania
- This work is supported by NSF grants IIS0513778 and IIS0415810