Large Scale Sequence Difference Visualization

Michael DiBernardo

December 19th, 2005

Outline

- Review of problem
- Implementation
- Challenges
- Results

Specific Problem

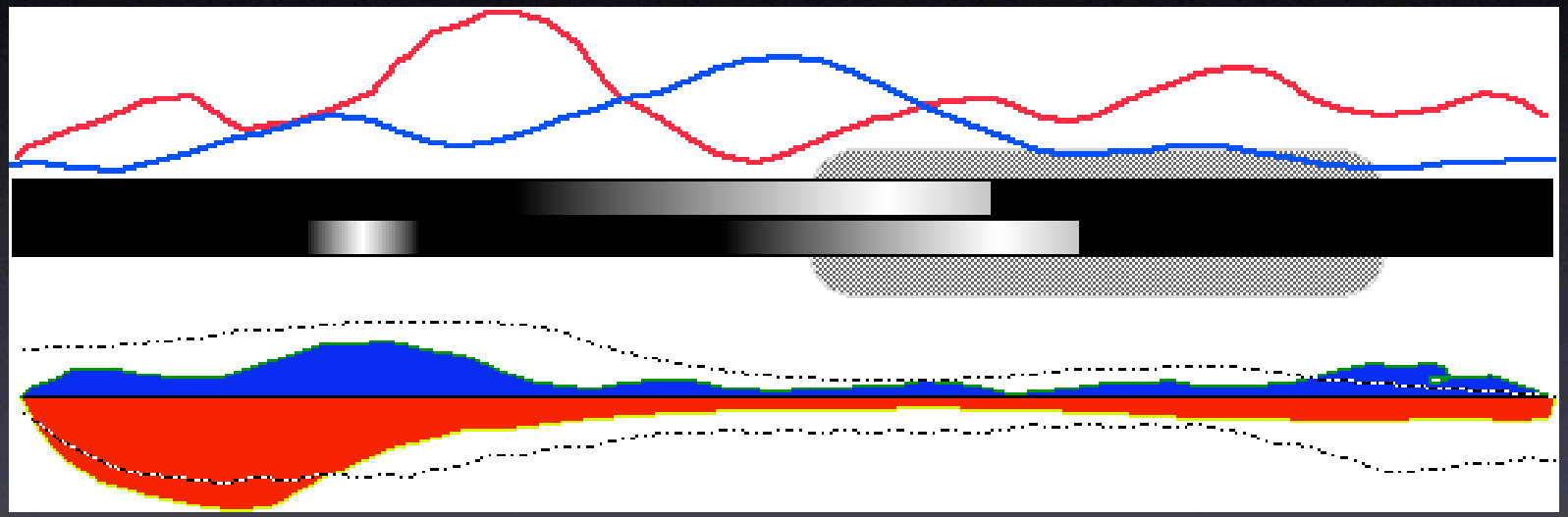
- Two populations
  - One susceptible to HIV
  - One immune to HIV
- HIV sequences extracted from both cohorts
- How do the extracted viruses differ from the canonical HIV?

Objective

- Design a static overview that can:
  - Succinctly describe results in a journal figure etc.
  - Serve as a linked overview in an exploration tool

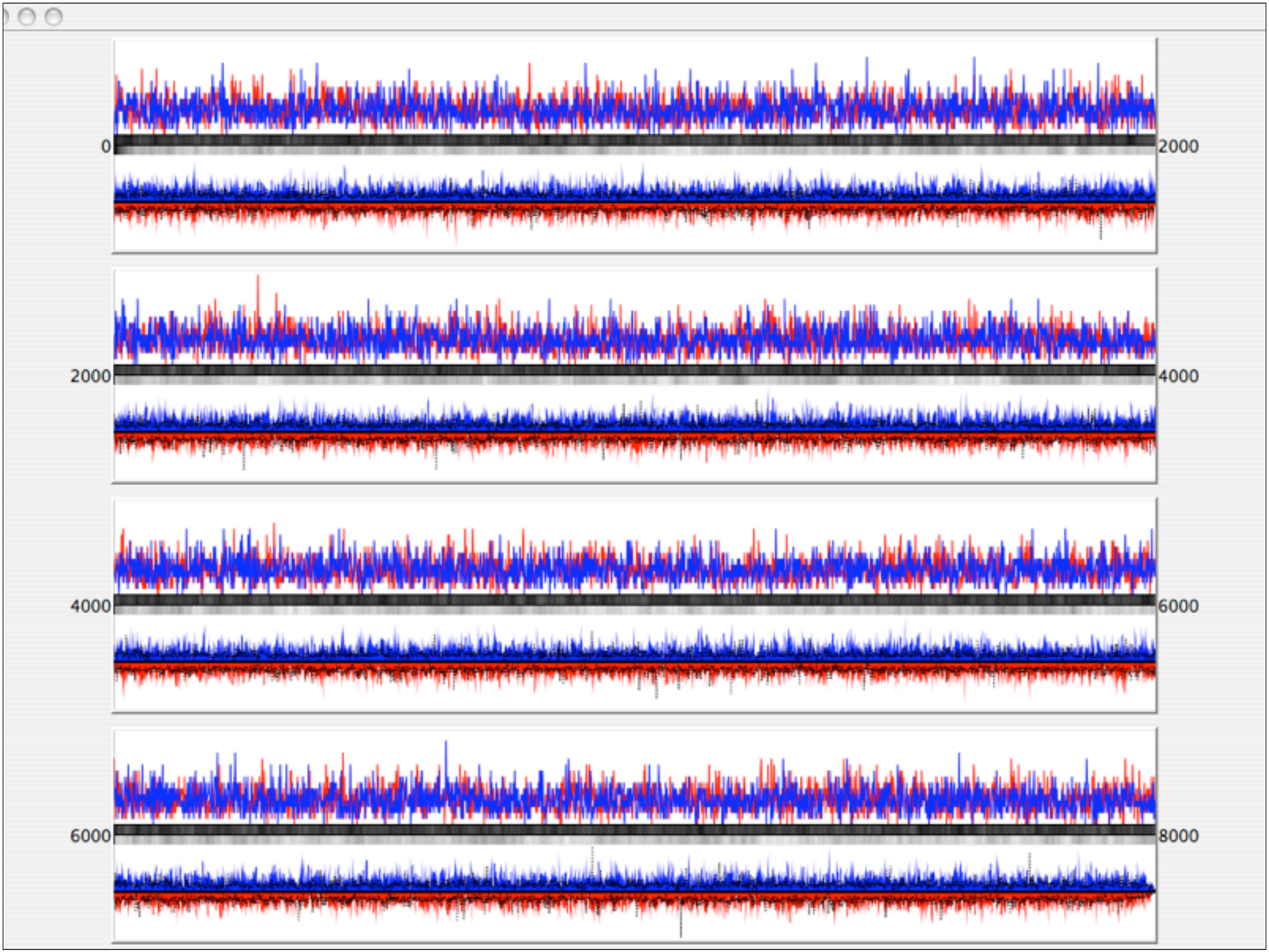
Design goals

- Immediate familiarity to biologists
- Allow for sufficient detail
  - SNPs
  - Small indels



Implementation

- JFreeChart used for substitution, insertion views
- Custom Java2D component used for gradients


0
2000
2000
4000
4000
6000
6000
8000

Challenges

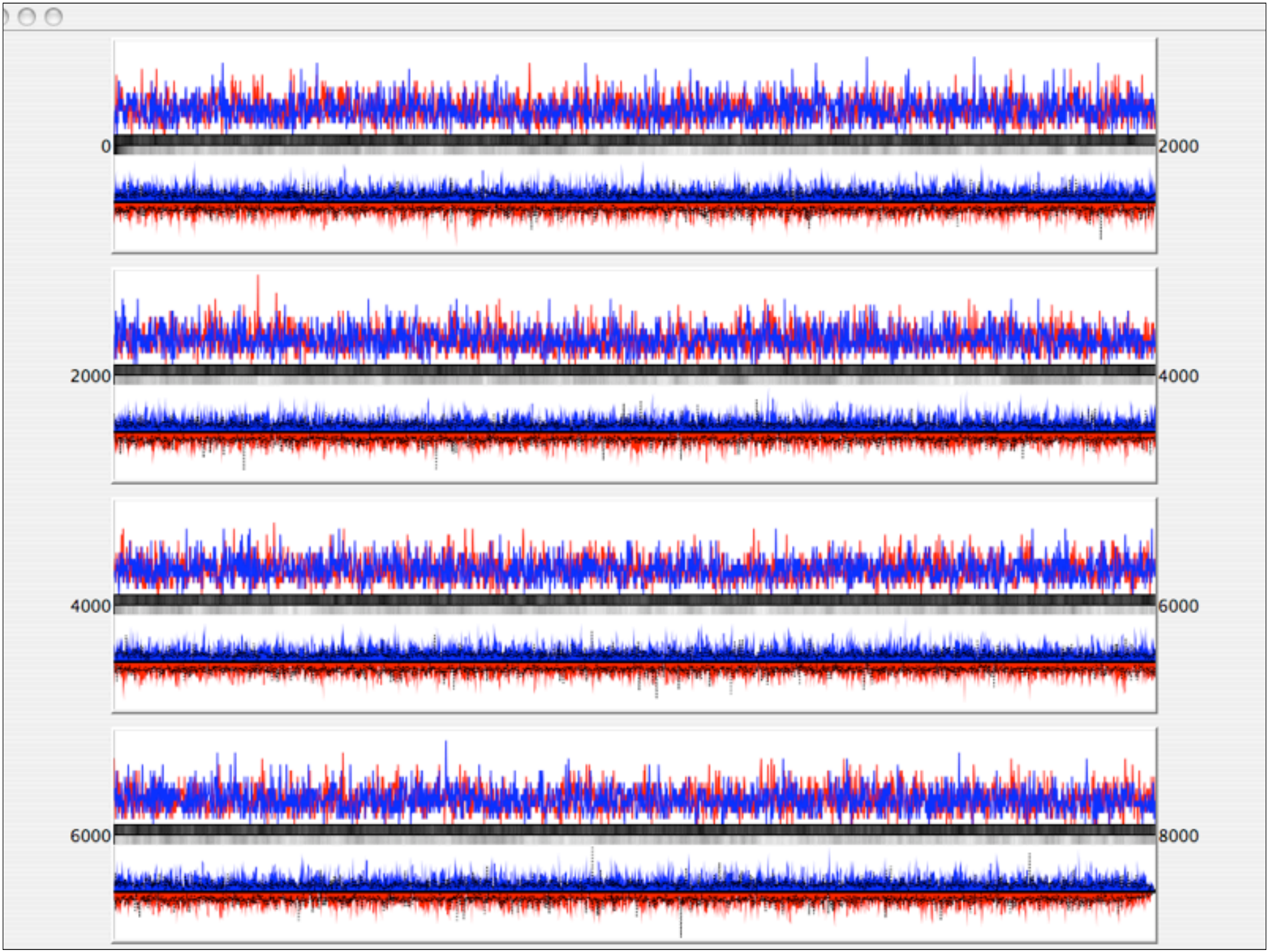
- Lack of critical feedback
- Solutions:
  - Cognitive walkthroughs
  - Input from previous colleagues in molecular biology

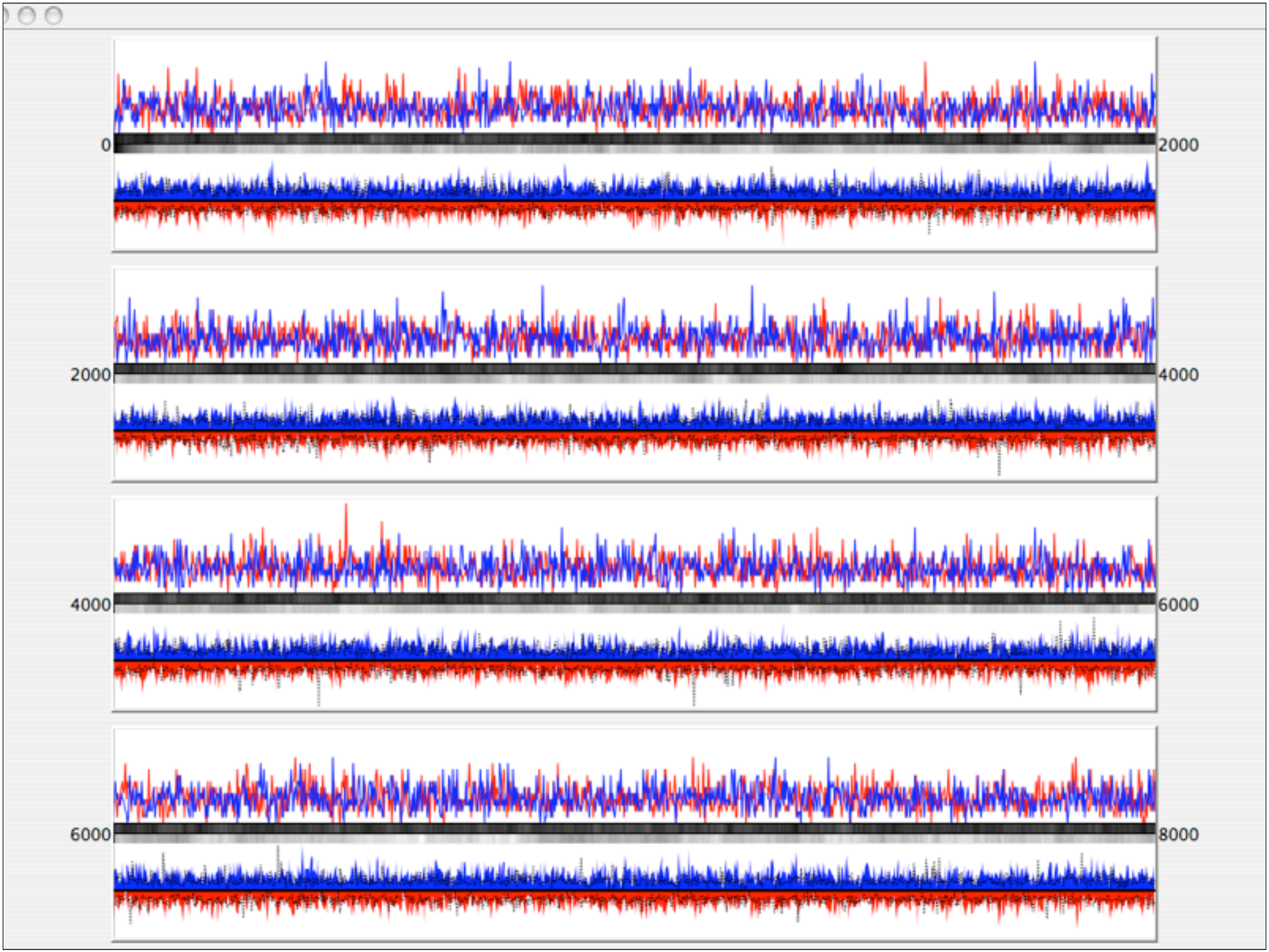
Challenges

- Lack of real data
- Solutions:
  - Generated data from probabilistic models of evolution
  - Invented a ‘toy’ experiment based on existing, publicly available data

Challenges

- Data density
  - Even with stacking, 4+ bases per pixel
  - JFreeChart behaviour undefined
- ‘Solution’:
  - Window averaging


0
2000
2000
4000
4000
6000
6000
8000


0
2000
2000
4000
4000
6000
6000
8000

Performance

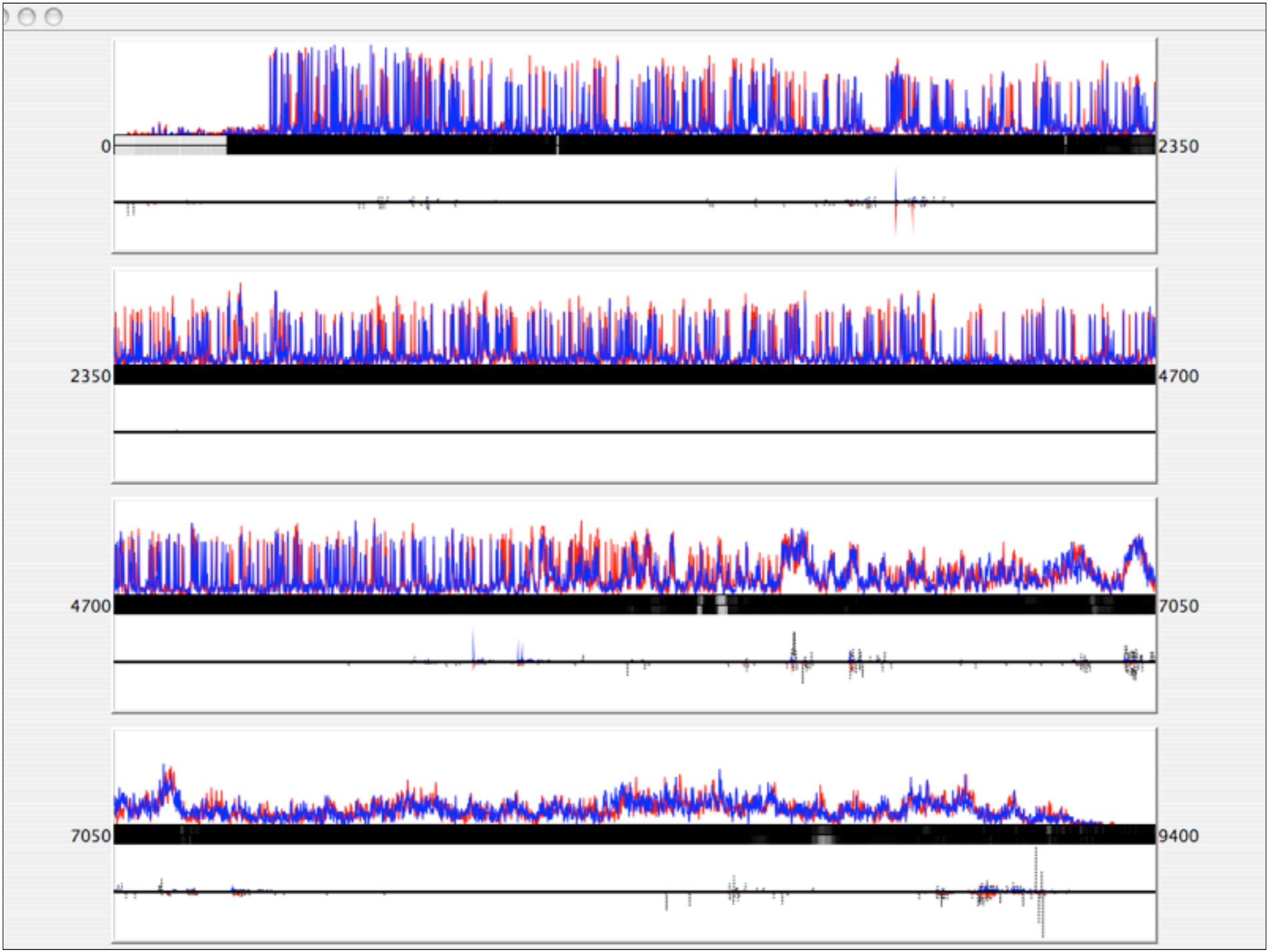
- For 200k data points first prototype took:
  - 2 minutes to launch
  - 45 seconds to render
- Current version:
  - 10 seconds to launch
  - 8 seconds to render (still too slow)

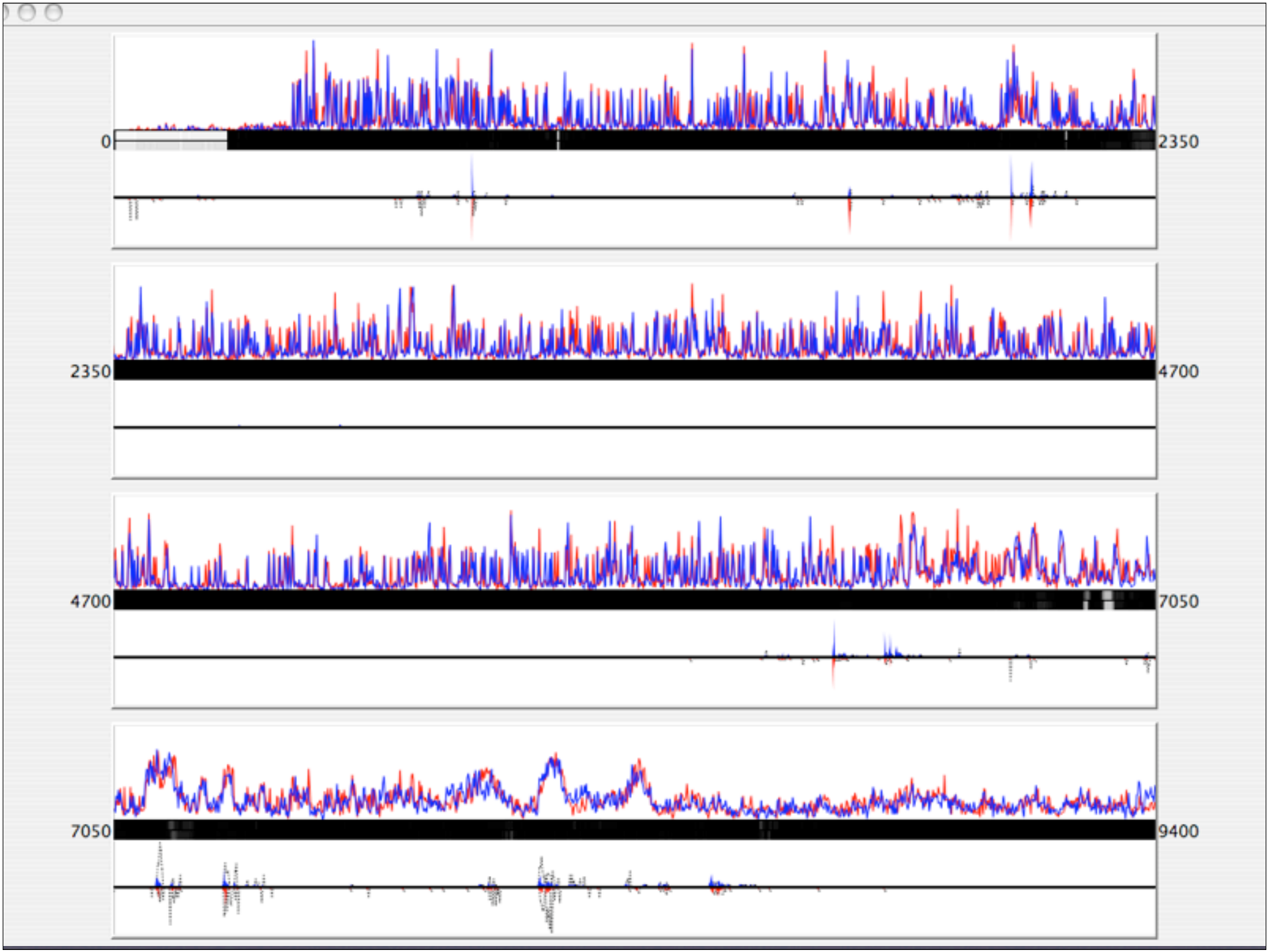
Evaluation

- Problems are easiest to see on real data

Experiment

- 74 viral samples isolated from Kenyan and Ugandan victims
- Compared to a single sample from Ethiopia


0
2350
2350
4700
4700
7050
7050
9400


0
2350
2350
4700
4700
7050
7050
9400

Conclusions

- Averaging is a poor solution
  - Hides some details and ‘lies’ to user
- Insertions are sparse and would be a good candidate for abstraction
- Using a gradient to represent deletions is a decent approach

Conclusions

- I am not smart enough to work alone