BayeHem: Bayesian Optimisation of Genome Assembly

DCSI 2018

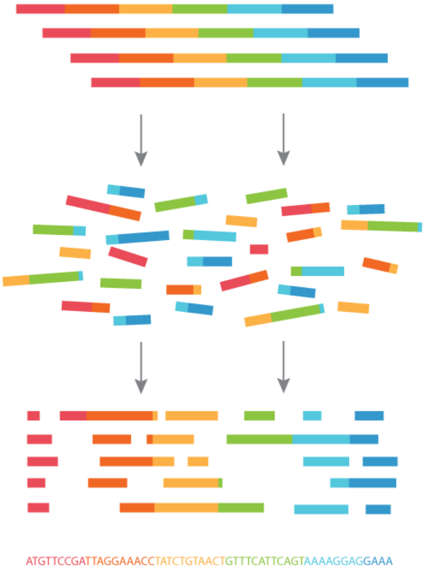
Finlay Maguire

Beiko Lab, FCS, Dalhousie University

Table of contents

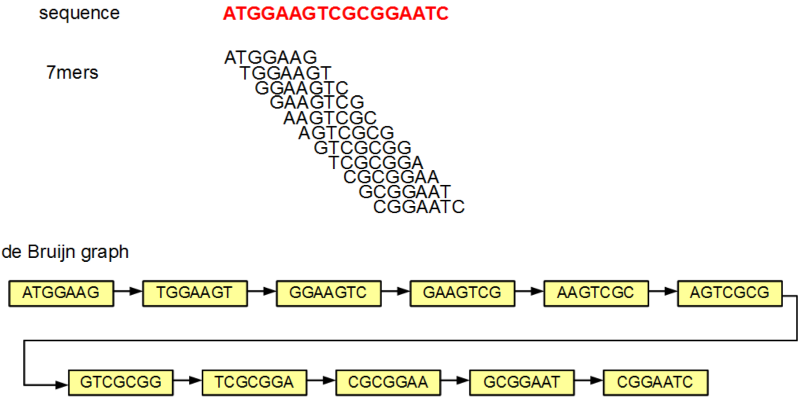
Genome Assembly

2nd Generation Genome Sequencing



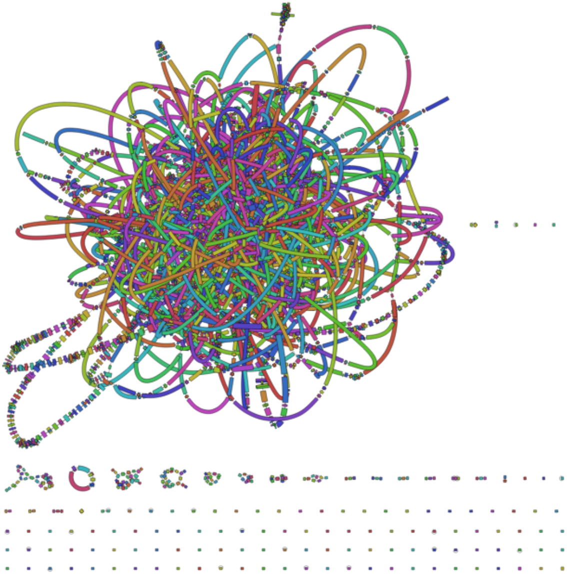
https://www.abmgood.com/marketing/knowledge_base/next_generation_sequencing_data_analysis.php

De Bruijn Graph Assembly



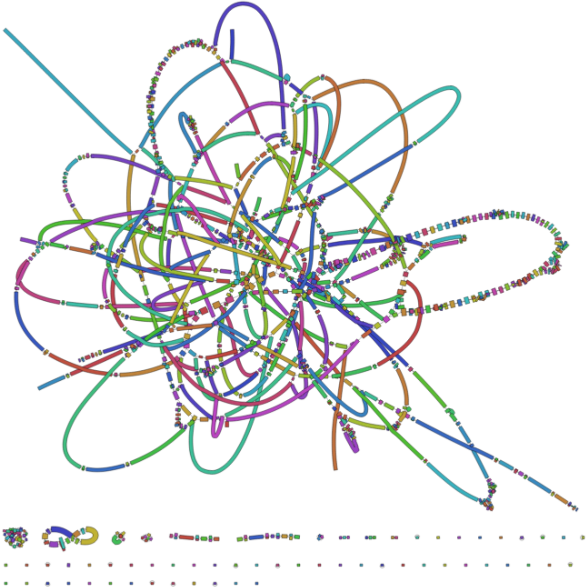
http://www.homolog.us/Tutorials/index.php?p=2.1&s=1

Effect of K-mer Size: 51-mer



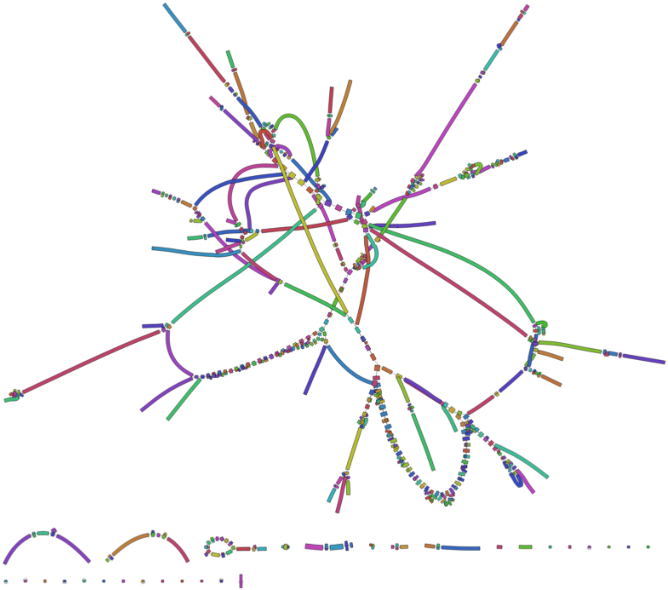
https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size

Effect of K-mer Size: 61-mer



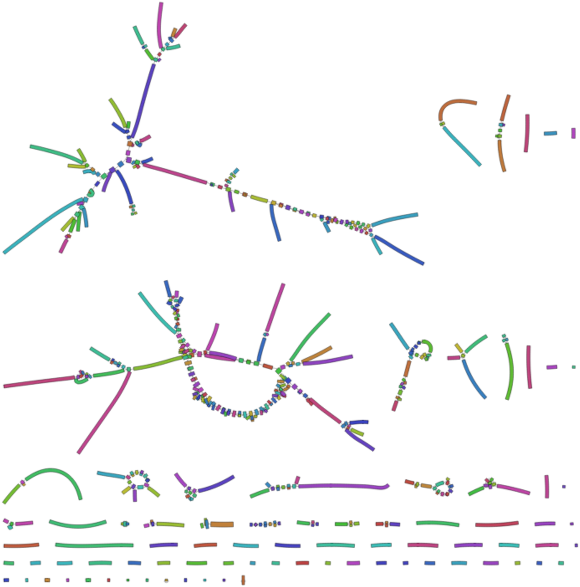
https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size

Effect of K-mer Size: 71-mer



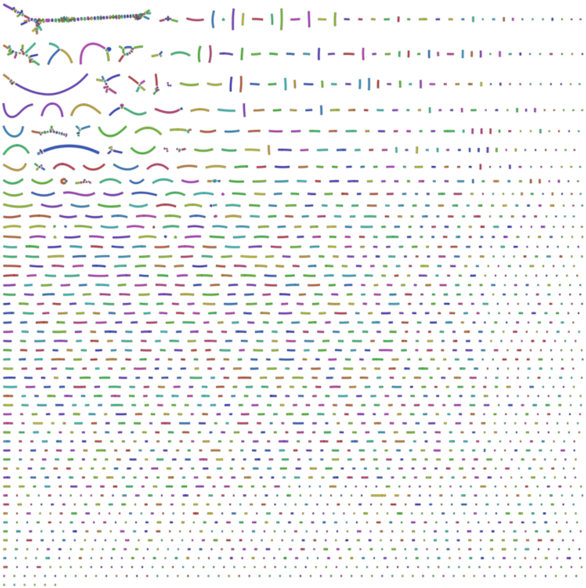
https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size

Effect of K-mer Size: 81-mer



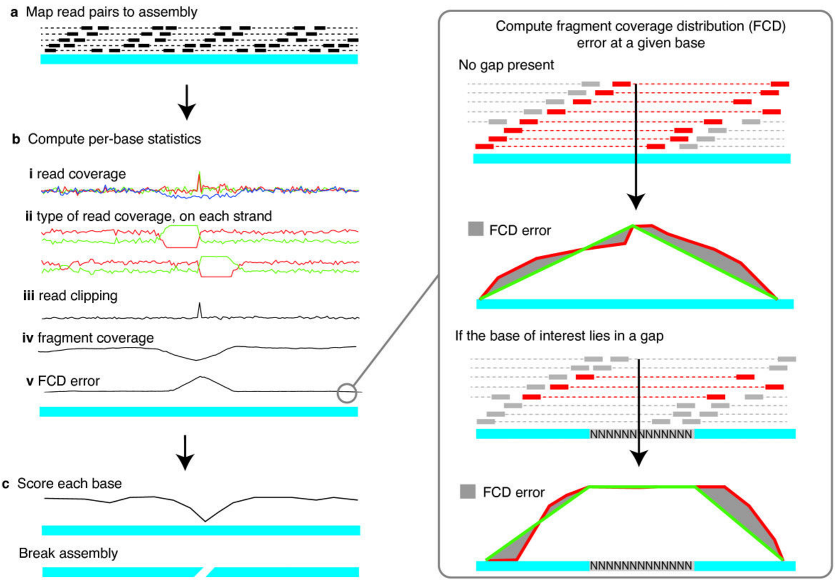
https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size

Effect of K-mer Size: 91-mer



https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size

Assessing Assemblies



[2]

Bayesian Optimisation

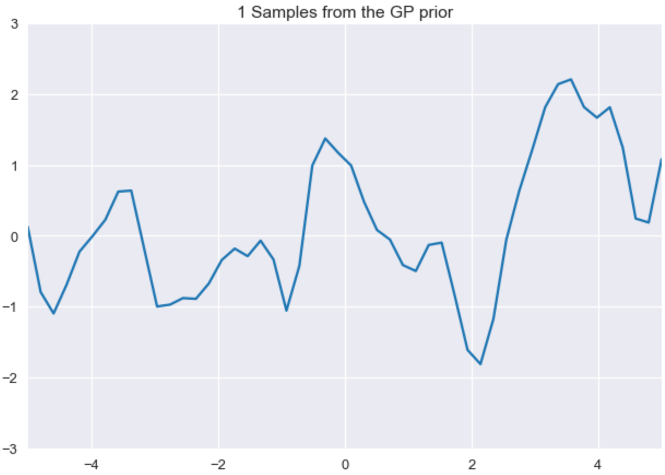
Gaussian Processes

- Form of functional regression.
- Powerful base for Sequential Model Based Optimisation [6].
- Every draw is a multivariate Gaussian random variable.

$$f \sim GP(0, K)$$

$$K \sim k(x_i, x_j) = exp(-\frac{1}{2}d(x_i/l, x_j/l)^2)$$

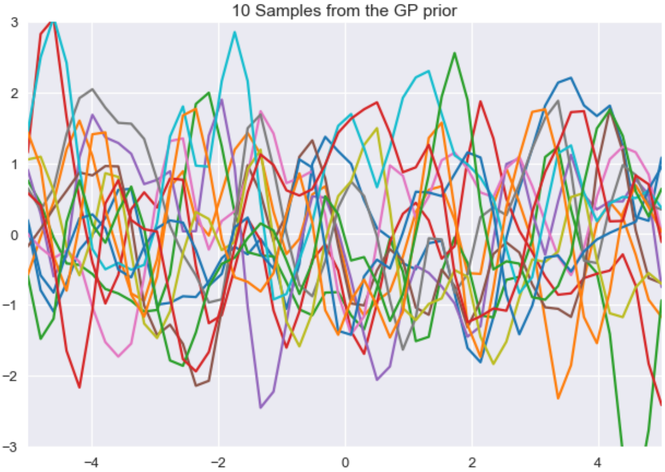
Gaussian Process Prior



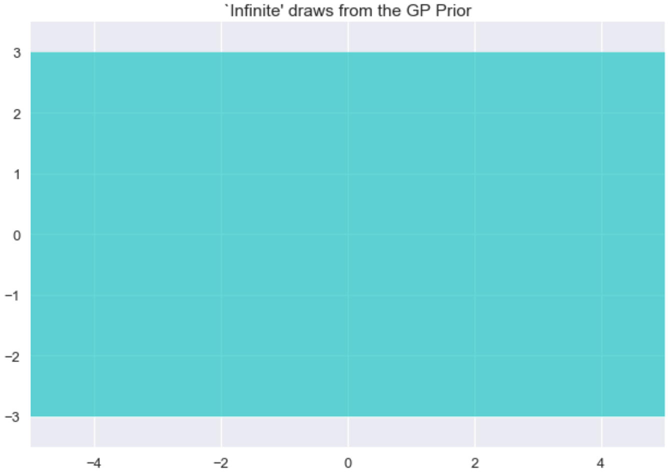
Visualisation code modified from `http://katbailey.github.io/post/gaussian-processes-for-dummies`

Gaussian Process Prior

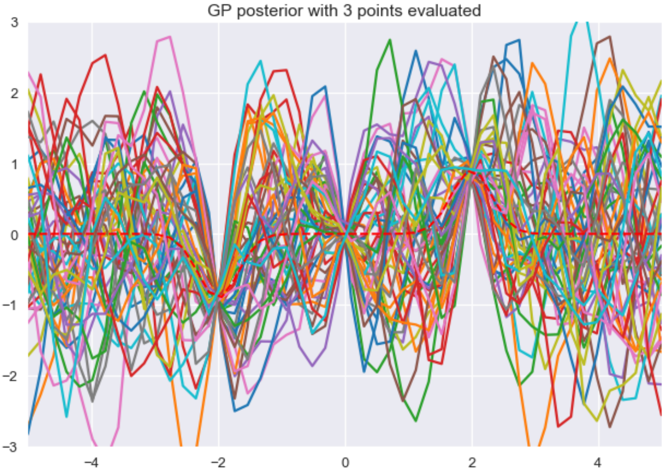
Gaussian Process Prior



Gaussian Process Prior

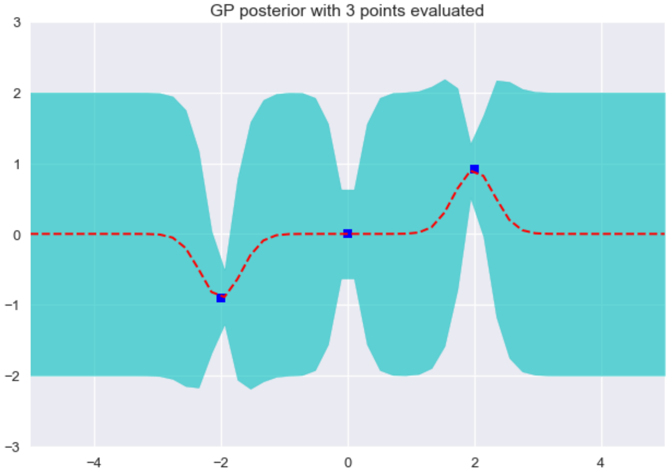


Gaussian Process Posterior

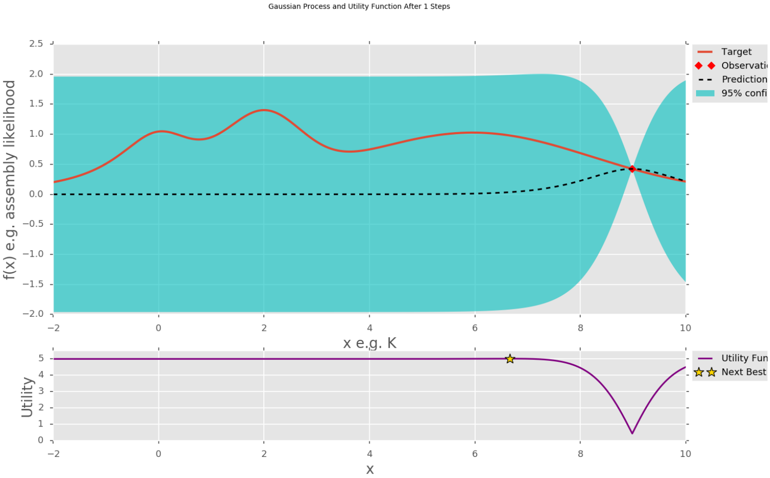


Gaussian Process Posterior

Gaussian Process Posterior

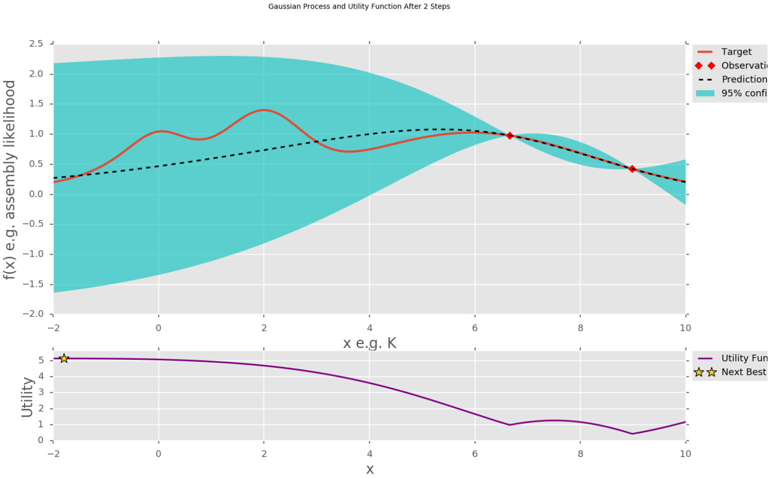


Acquistion Function

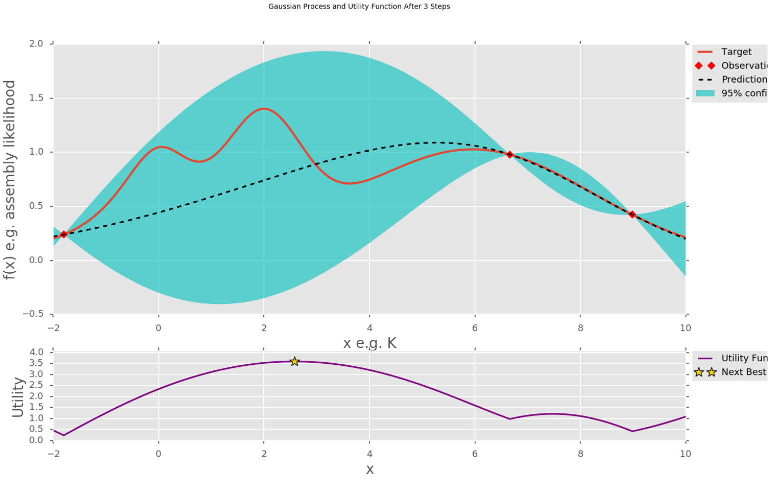


Adapted from code found here: `https://github.com/fmfn/BayesianOptimization`

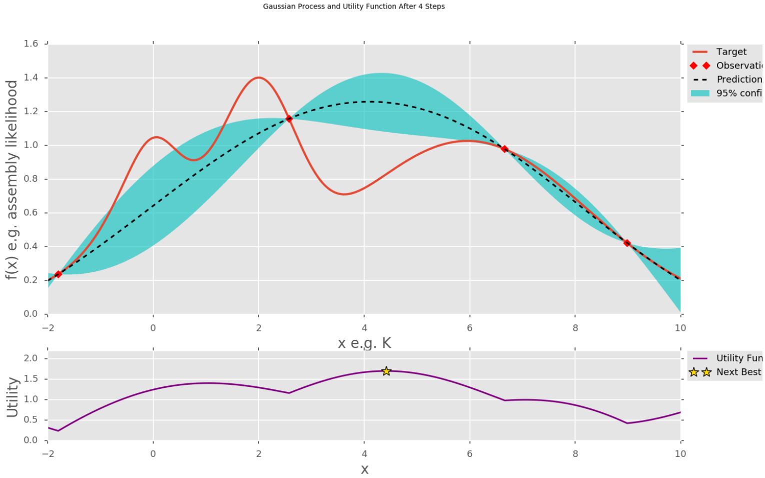
Acquistion Function



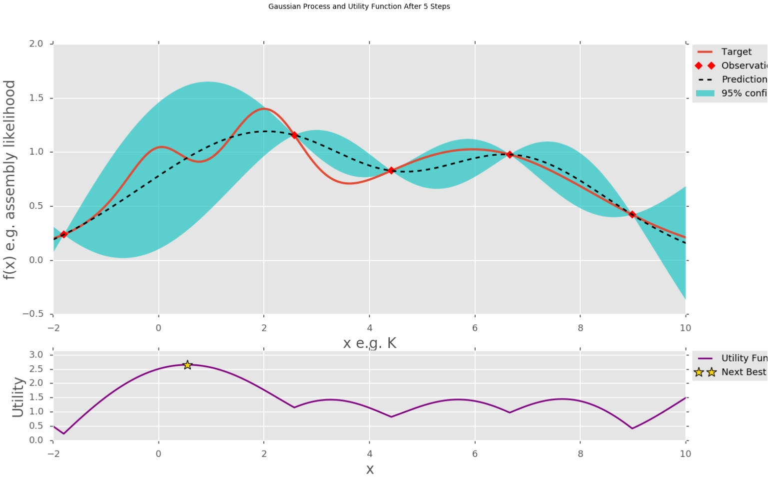
Acquistion Function



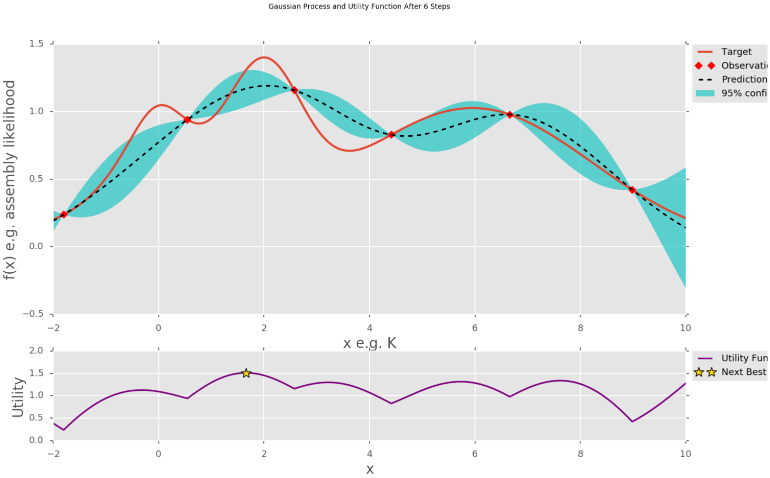
Acquistion Function



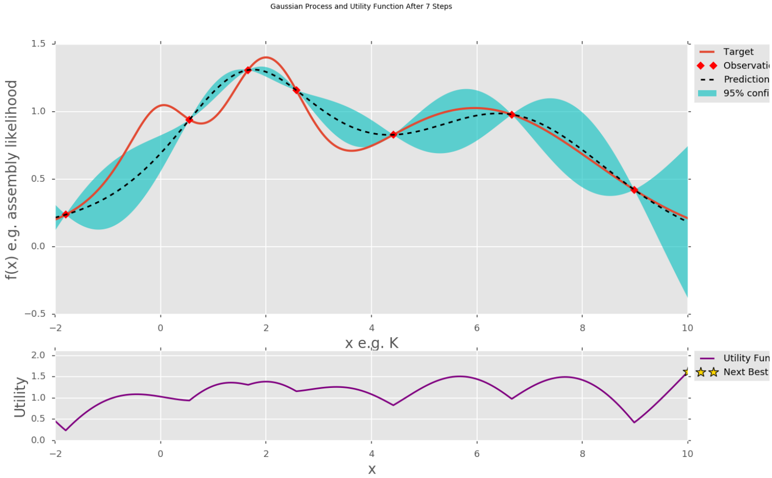
Acquistion Function



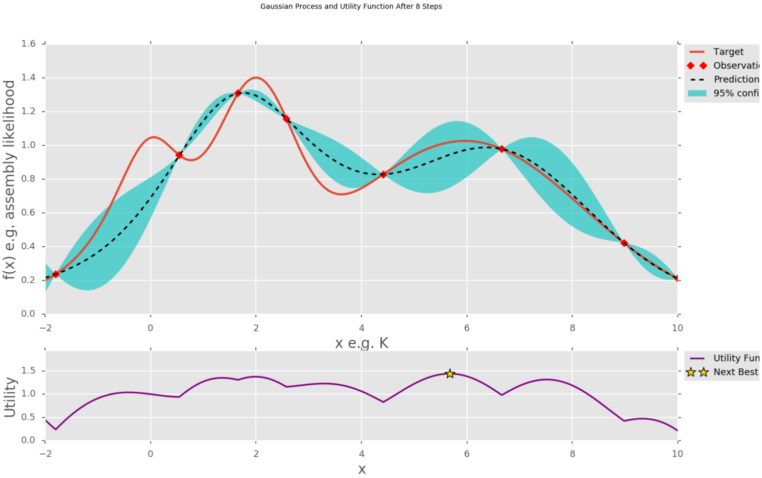
Acquistion Function



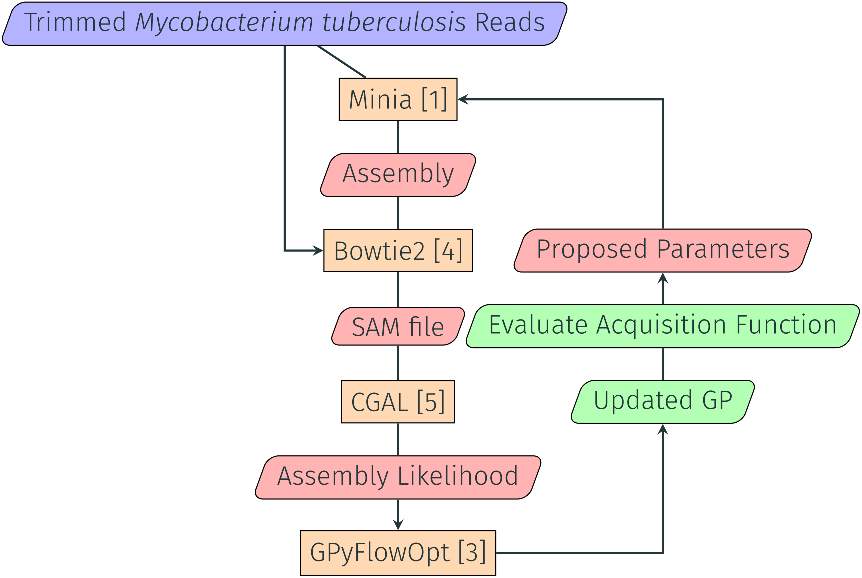
Acquistion Function



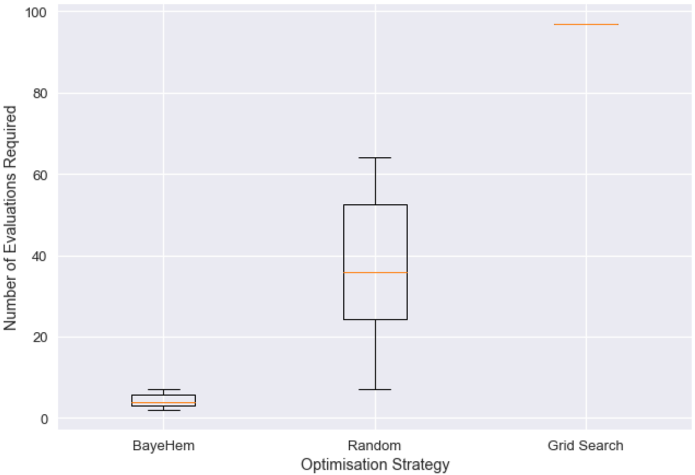
Acquistion Function



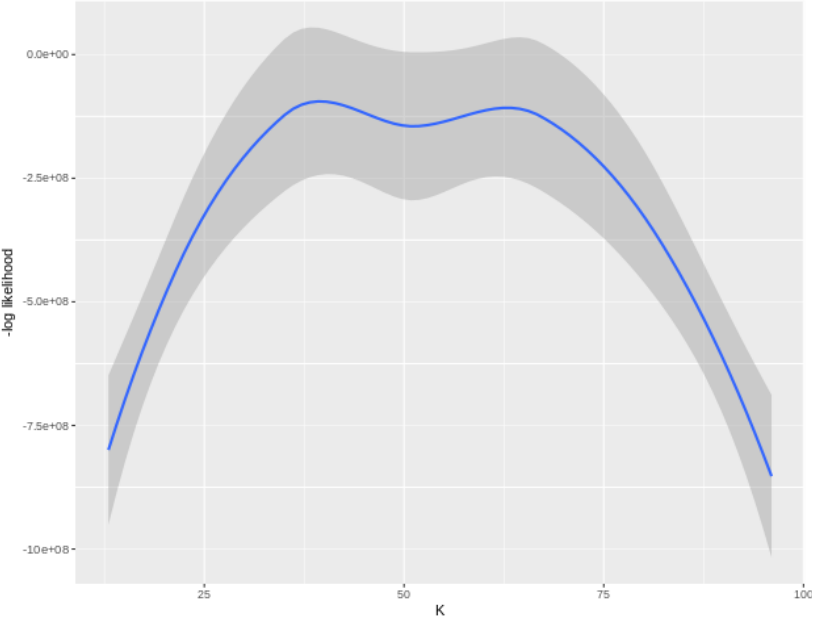
BayeHem


Trimmed *Mycobacterium tuberculosis* Reads
Minia [1]
Assembly
Bowtie2 [4]
SAM file
CGAL [5]
Assembly Likelihood
GPyFlowOpt [3]
Updated GP
Evaluate Acquisition Function
Proposed Parameters

BayeHem Proves Very Efficient



K Likelihood Surface



Limitations and Future Work

- Alternative GP covariance kernels
- Tuning acquisition (and parametrisation)
- Expand to other parameters in assembly pipelines
- Potentially flawed objective function.
- Multi-objective optimisation possible solution.

Conclusion

Summary

- Proof of concept for effectiveness of BayeHem.
- Assemblies are difficult to evaluate by a single metric.
- Large scope for improvement and development of this approach.

Questions?

R. Chikhi, G. Rizk, R. Idury, M. Waterman, M. Grabherr, Y. Peng, H. Leung, S. Yiu, F. Chin, P. Peterlongo, N. Schnel, N. Pisanti, M. Sagot, V. Lacroix, Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, G. McVean, G. Sacomoto, J. Kielbassa, R. Chikhi, R. Uricaru, P. Antoniou, M. Sagot, P. Peterlongo, V. Lacroix, R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, J. Simpson, K. Wong, S. Jackman, J. Schein, S. Jones, I. Birol, T. Conway, A. Bromage, R. Warren, R. Holt, P. Peterlongo, R. Chikhi, C. Ye, Z. Ma, C. Cannon, M. Pop, D. Yu, J. Pell, A. Hintze, R. Canino-Koning, A. Howe, J. Tiedje, C. Brown, A. Kirsch, M. Mitzenmacher, J. Miller, S. Koren, G. Sutton, R. Chikhi, D. Lavenier, C. Kingsford, M. Schatz, M. Pop, G. Marçais, C. Kingsford, G. Rizk, D. Lavenier, R. Chikhi, G. Rizk, D. Lavenier, S. Salzberg, A. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren,

T. Treangen, M. Schatz, A. Delcher, M. Roberts, G. Marçais, M. Pop, J. Yorke, B. Chazelle, J. Kilian, R. Rubinfeld, A. Tal, A. Bowe, T. Onodera, K. Sadakane, and T. Shibuya.
Space-efficient and exact de Bruijn graph representation based on a Bloom filter.
Algorithms for Molecular Biology, 8(1):22, 2013.

M. Hunt, T. Kikuchi, M. Sanders, C. Newbold, M. Berriman, and T. D. Otto.
REAPR: A universal tool for genome assembly evaluation.
Genome Biology, 14(5), 2013.

N. Knudde, J. van der Herten, T. Dhaene, and I. Couckuyt.
GPflowOpt: A Bayesian Optimization Library using TensorFlow.
pages 0–1, 2017.

B. Langmead and S. L. Salzberg.
Fast gapped-read alignment with Bowtie 2.
Nature Methods, 9(4):357–9, apr 2012.

A. Rahman and L. Pachter.
CGAL: computing genome assembly likelihoods.
Genome Biol, 14:R8, 2013.

J. Snoek, H. Larochelle, and R. P. Adams.
Practical Bayesian Optimization of Machine Learning Algorithms.
In *Advances in Neural Information Processing Systems*, volume 25, pages 2951–2959, 2012.