GAP: a Genome Annotation Pipeline

Grüning B*, **Erxleben A, Flemming S, Senger C, Günther S**

Department of Pharmaceutical Bioinformatics, Institute for Pharmaceutical Sciences, University of Freiburg, Germany;
**e-mail: bjoern.gruening@pharmazie.uni-freiburg.de*

Introduction

Processing the exploding number of new sequence data requires efficient implementation of several specialised tools on a powerful hardware infrastructure. Employed methods include genome assembly, genome annotation, pathway reconstruction, data visualisation, and pathway modelling. Galaxy[1] is a workflow management system for data processing ideally suited to the combination of various software tools. Furthermore, it ensures data and process reproducibility in terms of repeatability and traceability. Accessibility via web interface facilitates the integration of Galaxy into genome annotation projects. Based on Galaxy, we have developed the comprehensive Genome Annotation Pipeline (GAP) focused on processing newly sequenced bacterial genomes. GAP will soon be publicly available for genome annotation and systems biology for the scientific community.

Methods

The pipeline provides several tools (publicly-available and developed in-house) for statistical analyses of genome sequences, functional gene annotations, and the visualisation of complete genomes. With a genomic sequence as input, GAP uses a workflow (Fig.1) which includes:

- ORF prediction (*Glimmer3*),
- similarity searches in databases (e.g. *NCBI databases*, *UniProt*),
- functional annotation of gene products (*InterProScan*),
- prediction of protein localisation (*SignalP*),
- tRNA and tmRNA detection within the genome (*Aragorn*)

and many more.
GAP facilitates the mapping of genes and proteins to associated metabolic pathways. For example, the linkage of sequences to Gene Ontology terms, PubMed abstracts, or potentially interacting compounds via *CIL*[2] allows for a broad range of applications after first annotation.

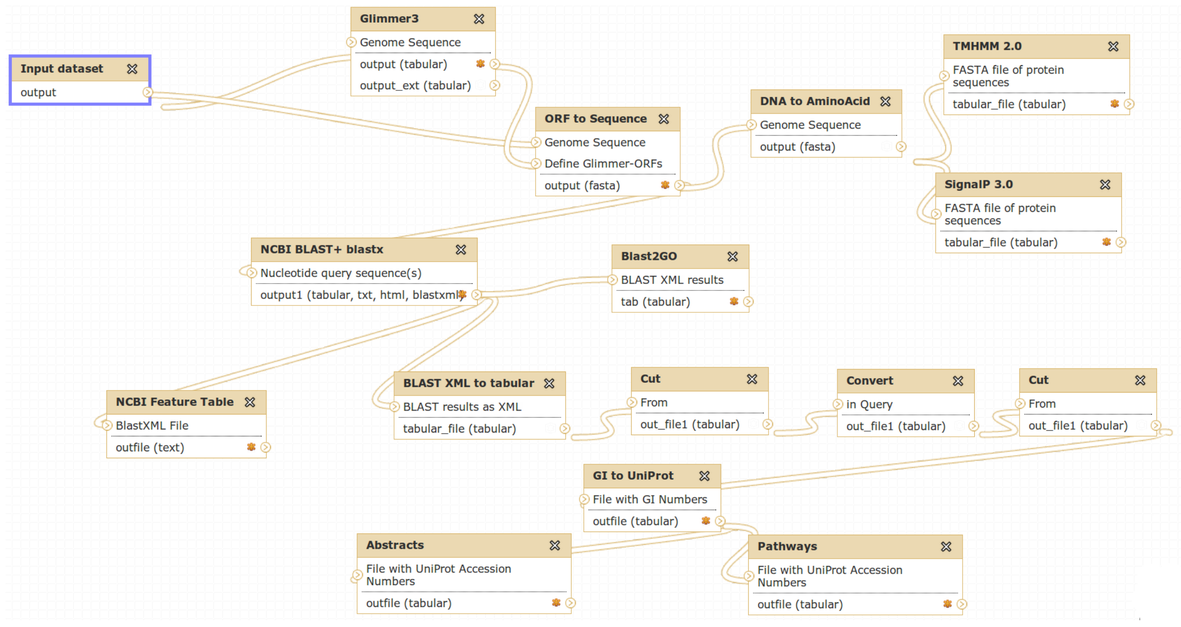

Fig. 1: Genome annotation workflow, including gene prediction and functional annotations

For visualisation of functionally annotated genomes, the genome browser *GBrowse* (Fig.2) and the multiple alignment tool *Mauve* were integrated. Sequence submission to GenBank/NCBI is facilitated by a tool to generate the required feature table.

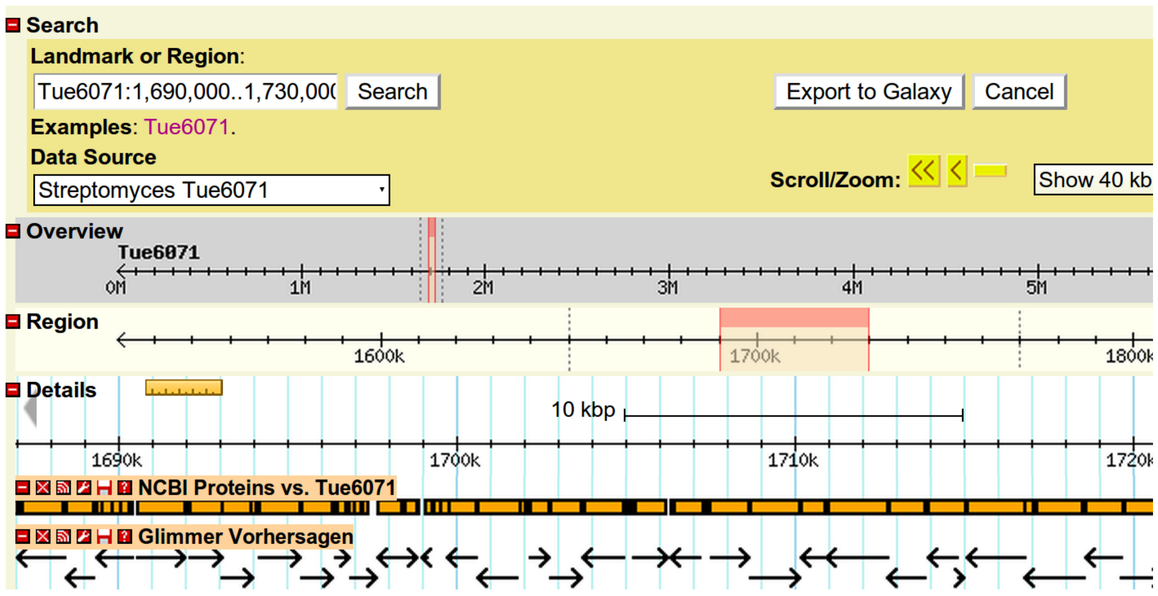

Fig. 2: Genome and corresponding annotation visualisation using GBrowse

Case study

Streptomyces Tü6071 is a bacterium which has a highly-active isoprenoid biosynthesis and produces the industrial important terpene Phenalinolactone which has anti-bacterial activity against several Gram-positive bacteria. The identification of biosynthetic gene clusters involved in metabolic pathways of secondary metabolites allows for the understanding of the biosynthesis of pharmaceuticals and will enhance rational genome modification and metabolic engineering.

We have successfully applied GAP for gene and protein annotation of the genome sequence of *Streptomyces Tü6071*[3].

The genome of *S. Tü6071* consists of one linear chromosome composed of 7,359,294 bp with 73.1% G+C as well as one linear plasmid (147,347 bp, 70.9% G+C).

Analysis of the *Tü6071* genome revealed that its chromosome contains 6466 protein-coding genes with at least 4887 proteins with assigned functions. On the plasmid, from 176 predicted ORFs, 73 have been functionally annotated. Furthermore, at least 6 rRNA operons and 74 tRNAs on the linear chromosome were predicted. The Phenalinolactone biosynthetic gene cluster examined by Dürr *et al.*[4] could be identified on the linear chromosome at position 1.69-1.73 Mb.

The annotated genome sequence of *Streptomyces Tü6071* is now available under the NCBI accession number AFHJ00000000.

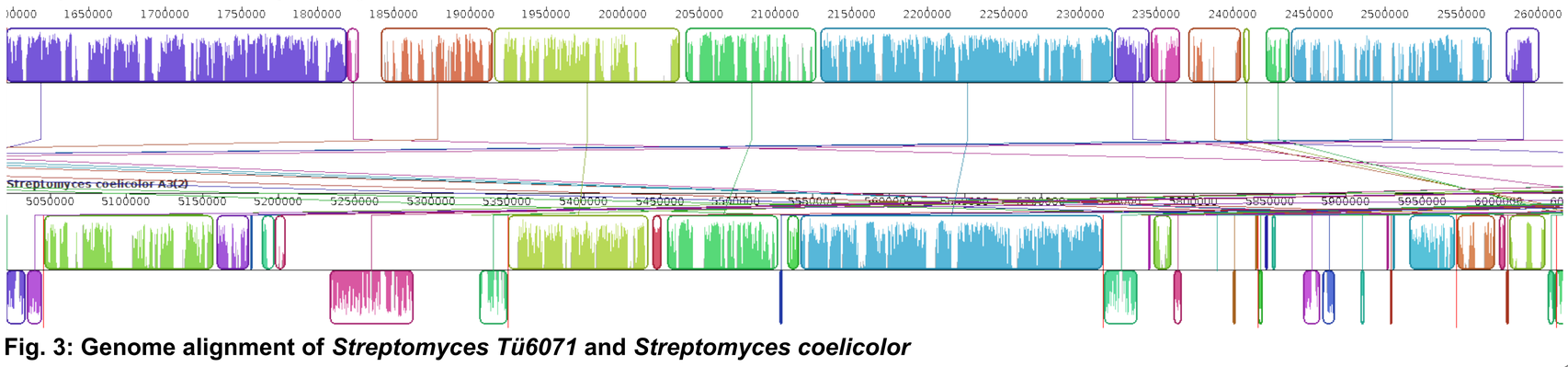

Fig. 3: Genome alignment of *Streptomyces Tü6071* and *Streptomyces coelicolor*

Albert-Ludwigs-Universität Freiburg

Black Forest Grid

Bundesministerium für Bildung und Forschung

References

[1] Goecks *et al.* Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010. 25:R86
[2] Grüning *et al.* Compounds In Literature (CIL): screening for compounds and relatives in PubMed. Bioinformatics. 2011. 27:1341-1342
[3] Erxleben *et al.* Genome sequence of *Streptomyces sp. Tü6071*. J. Bact. 2011. *accepted*
[4] Dürr *et al*. Biosynthesis of the terpene phenalinolactone in *Streptomyces sp. Tü6071*: analysis of the gene cluster and generation of derivatives. Chem Biol. 2006. 13:365-77

UNI FREIBURG