Microbial Cell Factories

Commentary

Open Access

Open access to sequence: Browsing the *Pichia pastoris* genome

Diethard Mattanovich*[1,2], Nico Callewaert[3,4], Pierre Rouzé[5,6], Yao-Cheng Lin[5,6], Alexandra Graf[1,2], Andreas Redl[1,2], Petra Tiels[3,4], Brigitte Gasser[1] and Kristof De Schutter[3,7]

Address: [1]Department of Biotechnology, University of Natural Resources and Applied Life Sciences, Vienna, Austria, [2]School of Bioengineering, University of Applied Sciences FH-Campus Wien, Vienna, Austria, [3]Unit for Molecular Glycobiology, Department for Molecular Biomedical Research, VIB, Ghent-Zwijnaarde, Belgium, [4]Unit for Molecular Glycobiology, L-ProBE, Department of Biochemistry and Microbiology, Ghent University, Ghent-Zwijnaarde, Belgium, [5]Department of Plant Systems Biology, VIB, Ghent-Zwijnaarde, Belgium, [6]Department of Plant Biotechnology and Genetics, Ghent University, Ghent, Belgium and [7]Department for Biomedical Molecular Biology, Ghent University, Ghent-Zwijnaarde, Belgium

Email: Diethard Mattanovich* - diethard.mattanovich@boku.ac.at; Nico Callewaert - Nico.Callewaert@dmbr.vib-ugent.be; Pierre Rouzé - pierre.rouze@psb.vib-ugent.be; Yao-Cheng Lin - yao-cheng.lin@psb.vib-ugent.be; Alexandra Graf - alexandra.graf@boku.ac.at; Andreas Redl - andreas.redl@boku.ac.at; Petra Tiels - petra.tiels@dmbr.vib-ugent.be; Brigitte Gasser - brigitte.gasser@boku.ac.at; Kristof De Schutter - kristof.deschutter@dmbr.vib-ugent.be

* Corresponding author

Published: 16 October 2009

Microbial Cell Factories 2009, **8**:53 doi:10.1186/1475-2859-8-53

Received: 25 December 2008
Accepted: 16 October 2009

This article is available from: http://www.microbialcellfactories.com/content/8/1/53

Abstract

The first genome sequences of the important yeast protein production host *Pichia pastoris* have been released into the public domain this spring. In order to provide the scientific community easy and versatile access to the sequence, two web-sites have been installed as a resource for genomic sequence, gene and protein information for *P. pastoris*: A GBrowse based genome browser was set up at http://www.pichiagenome.org and a genome portal with gene annotation and browsing functionality at http://bioinformatics.psb.ugent.be/webtools/bogas. Both websites are offering information on gene annotation and function, regulation and structure.

In addition, a WiKi based platform allows all users to create additional information on genes, proteins, physiology and other items of *P. pastoris* research, so that the *Pichia* community can benefit from exchange of knowledge, data and materials.

Commentary

Modern biological research requires genome sequence information of the organisms of interest for numerous applications: the development of transcriptomic methods like DNA microarrays relies on genome data, proteomics needs a genome sequence for efficient identification of proteins, metabolic modelling and flux analysis is based on the knowledge of ideally all enzymatic reactions encoded in the genome of an organism. Systems biology, as the synthesis of the above mentioned techniques [1], relies on comprehensive genome sequence data. Systems biology is most advanced for a few model organisms, for which genome sequencing has been an international challenge funded with public support. Systems biotechnology, the application of these approaches to biotechnological strain and process development, faces the same needs [2]. However, genome sequencing of biotechnologically relevant organisms has mainly been pur-

sued with corporate support, and the results were kept confidential over years for commercial exploitation. A major disadvantage of this strategy is the delay of basic research related to these organisms, negatively affecting the knowledge of organisms with the highest relevance for industry.

One such example is the yeast *Pichia pastoris*, widely used for heterologous protein production (reviewed in [3,4]), but also for the production of metabolites [5,6]. The major research areas towards implementing *P. pastoris* as a production host for heterologous proteins are engineering of glycosylation [7-9] and protein folding and secretion (reviewed in [10]). A draft genome sequence has been available commercially since appr. 5 years and omics methods have been developed based on this sequence (transcriptomics [11,12]; proteomics [13]; metabolic flux analysis ()[14,15]), but the strict obligation to keep sequence information confidential has hampered publication of relevant data and collaborations, so that the community could not benefit from exchange of knowledge, data and materials.

To bridge this gap we have published the genome sequences of two *P. pastoris* strains, DSMZ 70382 [16] and GS115 [17], obtained with next generation sequencing technologies. Versatile access to genome sequences is a prerequisite for efficient utilisation of the information. Therefore a genome browser was set up at http://www.pichiagenome.org[18] with a main focus on *P. pastoris* DSMZ 70382 and a genome portal with the gene annotation and browsing functionality for *P. pastoris* GS115 at http://bioinformatics.psb.ugent.be/webtools/bogas[19].

Both of these *Pichia* sites serve as a resource for genomic sequence data and gene and protein information for *P. pastoris*. The genome browser (GBrowse for DSMZ 70382 and AnnoJ [20] for GS115) allows users to view and navigate genomic sequences including non-translated regions of the genome. BLAST searches for comparing any query sequence against the *P. pastoris* dataset, full text searches and gene/sequence resources (Get Sequence) serve to retrieve, display and analyze a gene or sequence in many ways, such as protein translation. In the near future, a comparison of the genome of different strains will be added to both genome browsers.

The genome browser of *P. pastoris* DSMZ 70382 is based on the Generic Genome Browser (GBrowse) which consists of a web interface and a database backend. The system was developed by the Generic Model Organism Database project [21,22] for the purpose of exploring genomic sequences together with annotated data. GBrowse has already been used successfully in various genome database projects like SGD, FlyBase or WormBase and its functionality will therefore be familiar to many researchers. The browser simultaneously provides a bird's eye view and detailed views of the genome and facilitates easy navigation through the genome using its zoom capacity. A flexible display of a variety of features, including genes, proteins, RNAs, GC content and restriction sites, on separated customizable tracks permits the user to adapt the browser to his or her needs. The visualization of Microarray probe locations allow for the direct access to specific probe sequence and location of published microarray designs [12]. The *Pichia* Genome Browser further allows locating DNA or protein sequence patterns, to design sequencing and PCR primers and to display restriction maps for a sequence. Several search functions are implemented, including a full text search of the gene annotation. Each gene has a details page where further information about the gene such as its annotation or assigned Gene Ontology (GO) terms [23] is displayed. Apart from the DNA, the coding and the translated sequence of a gene, an up- or downstream region can be specified to be displayed on this page. At the bottom of each details page, links allow users to directly send the specific sequence to other analysis tools such as BLAST. Furthermore, the results of a precalculated InterProScan pattern search [24] are displayed for each annotated protein and can be accessed through the respective link. A comments section enables researchers to add information to their genes of choice. Data downloads are available either in the format of decorated FASTA files or gff files which include gene annotation. Future work on the genome browser of *P. pastoris* DSMZ 70382 will include a genome snapshot which will summarize the status of annotation and the distribution of gene products among functional groups. Batch download processes and an extension of the tools section are planned as well as a platform for the community to share experiences and knowledge in order to promote collaboration. Tutorials for GBrowse are available at [25] or [26].

Except the basic genome browsing and search function, the genome portal of GS115 strain also provides a comprehensive protein-coding gene annotation by the BOGAS (Bioinformatics Gent Online Genome Annotation System). The BOGAS is a gene centric concept, which means the information is provided based on the information related to the gene. Each gene has it's own annotation page which provides an overview of the gene information including the annotator, gene function, gene ontology, protein domain, protein homologs, gene structure, CDS and protein. The annotator information tells who and when annotated this gene and the history log to go back to previous version. Gene function field is filled by anno-

tators with the full gene function and a dictionary to provide a standardized gene nomenclature (short name). The BOGAS system automatically updates the protein information to provide the gene ontology and protein domain by InterProScan, the protein homologs and the multiple alignment by BLASTP and MUSCLE [27] when the user updates the gene structure.

The most important feature of BOGAS system is that it allows the registered users to update the information. Users can correct existing gene structure or create new genes by the annotation software (Artemis [28] or GenomeView [29]) and contribute their expert biological domain in the gene function field. Since the BOGAS provides the history log function, other experts can update the information and people in the community can trace these changes in few clicks. The full text search function in BOGAS can search across locus id, protein domain, genomic location and annotator information. The BLAST function also provides bidirectional link between the query sequence and the possible gene or genomic region. After running the sequence similarity search to fish out the candidate gene or genomic sequence, the user will be linked between the BLAST search result and the corresponding gene region.

As it has been adopted already to a large extent, we suggest that *P. pastoris* gene names should follow the format established for *S. cerevisiae* gene names. A detailed guide to *S. cerevisiae* nomenclature has been published in Trends in Genetics [30]. The gene name should consist of three letters followed by an Arabic number (e.g. *TPI1*). Where *P. pastoris* and *S. cerevisiae* genes appear to be orthologous, they should share the same gene name. The use of prefixes adds clarity to papers discussing genes from different species that share a name (e.g., Pp*URA3* vs. Sc*URA3*), but the gene names themselves do not include the prefix.

These two *Pichia pastoris* genome sites have been developed as a service for the scientific community. The remote annotations can be added either by informing the authors or through the BOGAS system. The WiKi based platform will allow to create additional information on genes, proteins, physiology and other items of *P. pastoris* research. We invite the *P. pastoris* community to join our efforts by providing new information on gene annotation, function, regulation and structure.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors contributed equally to this manuscript.

References

1. Ideker T, Galitski T, Hood L: **A new approach to decoding life: systems biology.** *Annu Rev Genomics Hum Genet* 2001, **2:**343-372.
2. Lee S, Lee D, Kim T: **Systems biotechnology for strain improvement.** *Trends Biotechnol* 2005, **23(7):**349-358.
3. Cereghino JL, Cregg JM: **Heterologous protein expression in the methylotrophic yeast Pichia pastoris.** *FEMS Microbiol Rev* 2000, **24(1):**45-66.
4. Macauley-Patrick S, Fazenda ML, McNeil B, Harvey LM: **Heterologous protein production using the Pichia pastoris expression system.** *Yeast* 2005, **22(4):**249-270.
5. Marx H, Mattanovich D, Sauer M: **Overexpression of the riboflavin biosynthetic pathway in Pichia pastoris.** *Microb Cell Fact* 2008, **7:**23.
6. Hu H, Qian J, Chu J, Wang Y, Zhuang Y, Zhang S: **DNA shuffling of methionine adenosyltransferase gene leads to improved S-adenosyl-L-methionine production in Pichia pastoris.** *J Biotechnol* 2009, **141(3-4):**97-103.
7. Hamilton S, Davidson R, Sethuraman N, Nett J, Jiang Y, Rios S, Bobrowicz P, Stadheim T, Li H, Choi B, *et al.*: **Humanization of yeast to produce complex terminally sialylated glycoproteins.** *Science* 2006, **313(5792):**1441-1443.
8. Hamilton S, Gerngross T: **Glycosylation engineering in yeast: the advent of fully humanized yeast.** *Curr Opin Biotechnol* 2007, **18(5):**387-392.
9. Jacobs P, Geysens S, Vervecken W, Contreras R, Callewaert N: **Engineering complex-type N-glycosylation in Pichia pastoris using GlycoSwitch technology.** *Nat Protoc* 2009, **4(1):**58-70.
10. Gasser B, Saloheimo M, Rinas U, Dragosits M, Rodríguez-Carmona E, Baumann K, Giuliani M, Parrilli E, Branduardi P, Lang C, *et al.*: **Protein folding and conformational stress in microbial cells producing recombinant proteins: a host comparative overview.** *Microb Cell Fact* 2008, **7:**11.
11. Gasser B, Maurer M, Rautio J, Sauer M, Bhattacharyya A, Saloheimo M, Penttilä M, Mattanovich D: **Monitoring of transcriptional regulation in Pichia pastoris under protein production conditions.** *BMC Genomics* 2007, **8:**179.
12. Graf A, Gasser B, Dragosits M, Sauer M, Leparc G, Tuechler T, Kreil D, Mattanovich D: **Novel insights into the unfolded protein response using Pichia pastoris specific DNA microarrays.** *BMC Genomics* 2008, **9(1):**390.
13. Dragosits M, Stadlmann J, Albiol J, Baumann K, Maurer M, Gasser B, Sauer M, Altmann F, Ferrer P, Mattanovich D: **The effect of temperature on the proteome of recombinant Pichia pastoris.** *J Proteome Res* 2009:1380-92.
14. Solà A, Maaheimo H, Ylönen K, Ferrer P, Szyperski T: **Amino acid biosynthesis and metabolic flux profiling of Pichia pastoris.** *Eur J Biochem* 2004, **271(12):**2462-2470.
15. Solà A, Jouhten P, Maaheimo H, Sánchez-Ferrando F, Szyperski T, Ferrer P: **Metabolic flux profiling of Pichia pastoris grown on glycerol/methanol mixtures in chemostat cultures at low and high dilution rates.** *Microbiology* 2007, **153(Pt 1):**281-290.
16. Mattanovich D, Graf A, Stadlmann J, Dragosits M, Redl A, Maurer M, Kleinheinz M, Sauer M, Altmann F, Gasser B: **Genome, secretome and glucose transport highlight unique features of the protein production host Pichia pastoris.** *Microb Cell Fact* 2009, **8:**29.
17. De Schutter K, Lin YC, Tiels P, Van Hecke A, Glinka S, Weber-Lehmann J, Rouze P, Peer Y Van de, Callewaert N: **Genome sequence of the recombinant protein production host Pichia pastoris.** *Nat Biotechnol* 2009, **27(6):**561-566.
18. **Pichia Genome browser** [http://www.pichiagenome.org]
19. **BOGAS** [http://bioinformatics.psb.ugent.be/webtools/bogas]
20. **Anno-J** [http://www.annoj.org/]
21. **GMOD** [http://www.gmod.org/]
22. Stein L, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich J, Harris T, Arva A, *et al.*: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12(10):**1599-1610.
23. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25(1):**25-29.

24. Zdobnov EM, Apweiler R: **InterProScan--an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17(9):**847-848.
25. **GBrowse Tutorial** [http://www.openhelix.com/gbrowse]
26. **GBrowse Tutorial** [http://gmod.org/wiki/Gbrowse]
27. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32(5):**1792-1797.
28. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16(10):**944-945.
29. **GenomeView** [http://genomeview.sourceforge.net]
30. Cherry JM: **Genetic nomenclature guide. Saccharomyces cerevisiae.** *Trends Genet* 1995:11-12.