Visual Analysis of Next-Generation Sequencing Data to Detect Overlapping Genes in Bacterial Genomes

Svenja Simon*
University of Konstanz, Germany

Daniela Oelke†
University of Konstanz, Germany

Richard Landstorfer‡
Technical University of Munich, Germany

Klaus Neuhaus§
Technical University of Munich, Germany

Daniel A. Keim¶
University of Konstanz, Germany

Abstract

Next generation sequencing (NGS) technologies are about to revolutionize biological research. Being able to sequence large amounts of DNA or, indirectly, RNA sequences in a short time period opens numerous new possibilities. However, analyzing the large amounts of data generated in NGS is a serious challenge, which requires novel data analysis and visualization methods to allow the biological experimenter to understand the results.

In this paper, we describe a novel system to deal with the flood of data generated by transcriptome sequencing (RNA-seq) using NGS. Our system allows the analyzer to get a quick overview of the data and interactively explore interesting regions based on the three important parameters coverage, transcription, and fit. In particular, our system supports the NGS analysis in the following respects: (1) Representation of the coverage sequence in a way that no artifacts are introduced. (2) Easy determination of a fit of an open reading frame (ORF) to a transcript by mapping the coverage sequence directly into the ORF representation. (3) Providing automatic support for finding interesting regions to address the problems that the overwhelming volume of data comes with. (4) Providing an overview representation that allows parameter tuning and enables quick access to interesting areas of the genome.
We show the usefulness of our system by a case study in the area of overlapping gene detection in a bacterial genome.

Index Terms:
J.3 [Computer Applications]: Life and Medical Science—Biology and genetics; H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces;

1 Introduction

DNA contains the heredity information of an organism consisting of four chemical letters (nucleotides G, A, T and C). Substrings of this DNA are translated into proteins, and - together with their regulatory parts - are called a "gene". Each protein coding DNA-substring is defined by a start and stop codon but not all such sections which are called open reading frames (ORFs) actually encode a protein.

The building blocks of a protein are amino acids. A single amino acid is encoded by a nucleotide triplet, named codon, within an ORF. This feature causes three different reading frames to exist, depending on whether the amino acid chain commences at the first, second or third base to continue in triplets. Since DNA is double stranded another three reading frames are found on the antisense strand, thus, there is a total of six reading frames at the same DNA locus. Consequently, this enables the presence of two or more ORFs at the same site. Existence of such overlapping genes (OLGs) in which one ORF is embedded in another, is accepted for viral genomes, but for bacterial genomes, only a minority of about 30 OLGs are acknowledged to date (e.g. [17]). However, recent studies suggest that their number is largely underestimated [8, 16]. The aim of our project is to support the discovery of new OLGs in bacteria so that their total number and biological role can be explored (see Figure 1 for an illustration of these basic terms).

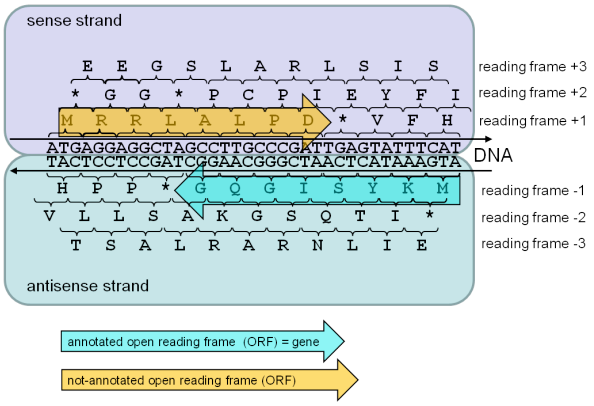


Figure 1: Illustration of the DNA double strand with its six reading frames. An annotated gene (in blue) overlaps with an ORF (in orange).

When a gene is translated into a protein, the ORF-containing part of the DNA is transcribed into several RNA-copies, depending on how much protein is needed. These copies are translated into the final protein. Next generation sequencing (NGS) technologies enable the strand specific identification of the RNA-copies (RNA-seq) and any transcribed genome region can be suspected to contain protein-coding ORFs.

In brief, for a RNA-seq-experiment, all RNAs of an organism are isolated, fragmented, converted to DNA and these fragments are sequenced. The short sequencing reads are mapped to the genome. This allows discovering any active site on the genome. The dynamic range of NGS is astounding; i. e., transcripts present only in a subset of the examined biological sample ($<$ 1 transcript / cell) or strongly up regulated genes with more than 100,000 transcripts per cell can be quantified accurately within one experiment [4]. Furthermore, NGS allows to probe all regions of a bacterial genome

*e-mail: simon@dbvis.de
†e-mail: oelke@dbvis.de
‡e-mail: richard.landstorfer@wzw.tum.de
§e-mail: neuhaus@wzw.tum.de
¶e-mail: keim@dbvis.de

with the same probability, a task which is otherwise only approachable using "tiling microarrays", but with a lesser dynamic range.

Therefore, NGS is the method of choice to identify new OLGs as some of these genes are supposed to be transcribed lowly and their genomic position is not known beforehand. Since OLGs are different in size, location, expression, and signals are mixed with strongly transcribed neighboring genes or non-coding RNA transcripts, analyzing NGS data is challenging. Besides, the data sets are quite large. A single bacterium may contain up to 10,000 annotated genes in genomes of up to 13 million nucleotides and a single NGS-experiment generates data in the range of 2 to 200 Gbp (billion base pairs). Thus, there is a clear need to support the task with visual analytics methods.

2 What is needed to support the task?

To display NGS-data, some visualization tools exist but all of them have shortcomings. Available programs are either designed to examine the expression of annotated genes only (RNA-seq) or for complete genome sequencing. Some of the most cited tools are, e. g., Genome Studio [2], the UCSC browser [3] or Artemis [13, 1]. These programs have in common that the raw data, which are the sequence reads of a length between 36 to 400 base pairs, are only visualized as stacks. Therefore, artificial gaps and peaks emerge as artifacts between stacks. Thus, boundaries between genes cannot be determined accurately (see top of Figure 2 for an example from the Artemis tool). Genome Studio and the UCSC browser are able to show the direction of the reads (sense and antisense), but only color coded and not separated for each strand. Artemis has a strand specific view, but the problem with the stack view remains. Further, the read stacks in Artemis are neither shown to a fixed scale, nor is a true stack height presented since identical reads are merged to a single read (green, Fig. 2), there are cases where it is reasonable to merge identical reads but for our propose we want to represent all reads).

Caused by the NGS technique, sampling of the sequenced reads depends largely on chance causing any transcript coverage to appear as a rugged transcript pattern above an active site. This in turn prevents the use of fixed values to determine if a site is indeed transcriptionally active or only background. The absolute values depend on the specific experiment and its sequencing depth. Furthermore, the distribution of expression peaks or valleys, fit of the transcript to an ORF, gene length and distance of the new OLG to existing and transcribed up- or downstream annotated ORFs, number of gaps and gap-width are important parameters which need careful inspection. The necessary synopsis of all of these parameters hampers the definition of fixed thresholds to identify OLGs.

Since in the existing tools only raw data are shown without any further focus, detection of OLGs is time demanding. The amount of not-annotated ORFs is much higher than the number of annotated ORFs (genes) and every one might be a potential candidate if covered by transcription. However, most of the genome will not show overlapping transcripts since at each experimental condition only a subset of those genes will be active. Thus, a focus on interesting regions highly accelerates the search.

Therefore, we designed a new, improved visualization tool in which the following aspects are specifically taken into account:

1. Transcription coverage is visualized color coded in each frame, which also shows all annotated and not-annotated ORF. This allows easy determination if an ORF fits to a transcript.

2. Interesting sites are automatically determined according to a user-defined interestingness function to help analyst to deal with the large amount of data.

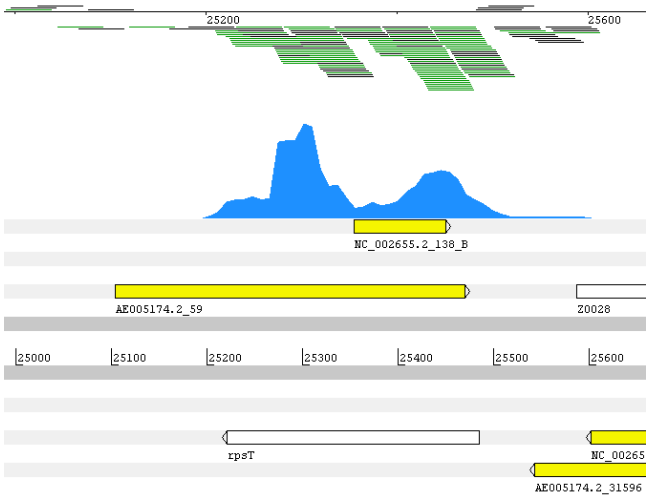


Figure 2: Stack view in the Artemis tool [1]. An annotated gene (rpsT) is irregularly covered by read stacks. Identical reads are merged to a single green read. Due to the stack representation artificial gaps emerge as artifacts between stacks. The corresponding coverage is displayed by the blue graph, but not strand specific. Overlapping ORFs of ≥ 93 bp are displayed by yellow bars.

3. Furthermore, interesting sites are shown in a "genome overview bar" combining several, freely adjustable parameters in color codes, which allows to see the most interesting cases immediately.

3 Design of the tool

One of our design criteria was to create a tool that resembles as closely as possible the state-of-the-art tools for next generation sequencing data that biologists are already acquainted to. This led to our basic setup in which we depict the genome as a linear sequence. The three reading frames of the sense and antisense strand are shown above and below the sequence, respectively. Open reading frames are depicted as boxes and positioned in the corresponding reading frame. In contrast to most sequence viewers, our task requires to depict *all* ORFs (in our example with at least 93bp in length) and not just already annotated genes. To distinguish annotated from not-annotated, the former are shown with a red frame.

3.1 Visualization of sequence coverage

In typical tools for next generation sequence analysis, two basic methods to represent sequence coverage can be found. Either a bar chart is used in which the number of observed reads at a specific position is mapped to the height of a bar or the reads are visualized as stacks. An advantage of the latter one is that not only the coverage at a specific position is visible but also the position of a specific read sequence.

Our goal is to visualize the sequence coverage in a way that the following criteria are met:

1. Showing the sequence coverage without introducing artifacts that are a problem of common stack view representations (see Fig. 2).

2. Making the positions of the read sequences visible.

3. Reducing the mental load of determining an ORF-fit to a transcript.

Therefore, sequence coverage was integrated in two ways: With an extended bar chart visualization and by mapping the values directly to the ORF representations. In both cases we distinguish between the sense strand and the antisense strand.

The bar chart represents complete genome coverage and the coverage values are logarithmically scaled to allow a consistent scale for the whole genome. Further, the information where reads begin might be important. To not lose this information we plot in orange the start positions of the reads in the bar chart. Highlighting a read end is not necessary in this case, since the example data are obtained with the SOLiD 4.0 system from ABI and each read has the same length.

Next, we map the transcription level to the ORF representation. By directly showing the read coverage color coded within the ORF rectangles, we are able to reduce the mental effort to determine if an ORF fits to the transcribed region or not. Since, we are specifically interested in overlapping genes (not annotated so far), it is not only the question if an ORF fits to a transcribed region but to which one of a few ORFs fits best. Therefore, correct data interpretation is critical. Because the transcription is rugged and uneven, inspection by an expert is mandatory.

From the perspective of visualization the challenge was to find a representation that permits to fit the transcription graphs directly into the shallow rectangles that represent the ORFs. Coloring of the rectangles using a mean value included by each ORF would cause loss of necessary information, such as if the complete ORF is transcribed or not. Standard line charts do not work as well because space in the y-direction of the graph is too limited to depict the fluctuations truthfully.

The solution to this is two-tone coloring [14, 5], in which each value is represented by two discrete colors (see colorscale in Fig. 3 D). Using this technique, values can be read quite precisely even if not much space is available for drawing.

3.2 Providing an interestingness function

NGS experiments provide vast amounts of data, which complicates data analysis. Due to technical (random sampling of reads) and biological reasons (background transcription, multiple promoter sites, untight termination regions, etc.) transcripts always appear in rugged course of values. For genes with low expression and rare transcripts, it is difficult to tell whether a gene is or is not transcribed. To decide, the following criteria are taken into account by experts: [1]

- **Coverage:** Percentage of bases of an ORF with a count of at least one. A low value means either coverage by only a few reads (background transcription) or overlap with a transcription signal from the untranslated regions (UTR) of an adjacent ORF. A high coverage value indicates a good fit of a certain ORF to a transcriptional signal.

- **Transcription:** Average number of counts of the bases of an ORF. To make sure that the numbers for different experimental conditions are comparable, this value has to be normalized. The higher the transcription value the more likely it is that the ORF was indeed transcribed.

- **Fit:** Absolute value of the difference of the transcript length and the ORF length. In case of annotated ORFs (genes), a high fit value indicates that this gene is part of an operon; in case of not-annotated ORFs which overlap a gene on the same strand this is an indication that the coverage is only due to the untranslated regions (UTR) of the gene (which in turn, would decrease coverage).

[1]Note that in the following definitions *count* is defined as the number reads covering a single base pair in sense or antisense. *Total number of counts* is defined as the sum of counts over all bases of the complete DNA of a cell less rRNA reads, because rRNA has been depleted in the course of the experiments.

All three criteria are selectable and thresholds can be defined (Fig. 3 B). Combining these three criteria, we set up an interestingness function to highlight interesting ORFs. Any other ORF, not meeting the thresholds is faded out. Alternatively, we may also restrict the analysis to regions with annotated or not-annotated ORFs. Further, we distinguish between the two reading directions because separate transcript values are available for the two strands, enabling an immediate overview of regions of transcription.

3.3 Genome Overview Bar

Having a flexible and expressive interestingness function is beneficial as the automatic fade-out puts focus on interesting areas that need deeper inspection. Still, scrolling through the whole genome sequence would be necessary. Displaying only regions of interest would reduce the amount of data but with the expense of losing context information. Now, an overview representation of the genome gives for all ORFs insight into all values of the interestingness function. Figure 3 A shows the visualization. In this tabular view each horizontal line represents a single parameter of the interestingness function (coverage, transcription, and fit). Each column represents an interesting ORF (only annotated ORFs (genes), only not-annotated ORFs or both) of the genome. A cell is colored in a range of green to purple. The more saturated the green is, the more the threshold was passed. Similarly, intensifying shades of purple encode increased fail of a threshold.

A summary line below further reduces this information to a single cell to enable an extremely quick overview. Again, sections failing one or more are shown in shades of purple. Regions passing all thresholds are drawn in yellow. Additionally, the bars are distorted so that columns which meet more criteria get more space. Because the length of the different ORFs in each column varies significantly and longer ORFs are often considered as more interesting, the length of each region is encoded in grey values below. Alternatively, length could also be used as the variable that determines the distortion factor of a column.

By clicking on a column in the overview bar the corresponding ORF is centered in the genome view. Thus, the overiew bar can be used to browse the genome faster and in a more focused and task specific way.

4 Working with the Tool

4.1 Tuning parameters

Because random sampling is involved in the sequencing process, the resulting sequence coverage is inevitably rugged. For the same reason also gaps have to be expected, even in regions of clearly transcribed genes. The probability of gaps depends on the sequencing depth in general and the transcript level in particular. The latter depends on the experimental conditions. The same holds for the average coverage of an ORF which is measured by the criterion 'transcription'. Consequently, it is not possible to specify default threshold values for filtering.

Thus, every analysis process starts with the challenge of choosing meaningful parameter values. This is an interactive process that imperatively needs an analyst with expert knowledge and some experience in evaluating next generation sequencing data. Our genome overview representation supports the task.

Figure 4(a) shows part of the resulting display when the expected coverage is set to 100% and the threshold for the transcription value is 10. Furthermore, the threshold value for the fit is set to 45 nucleotides. Overall these are quite stringent settings. Consequently, only few regions in the genome are able to pass all three thresholds.

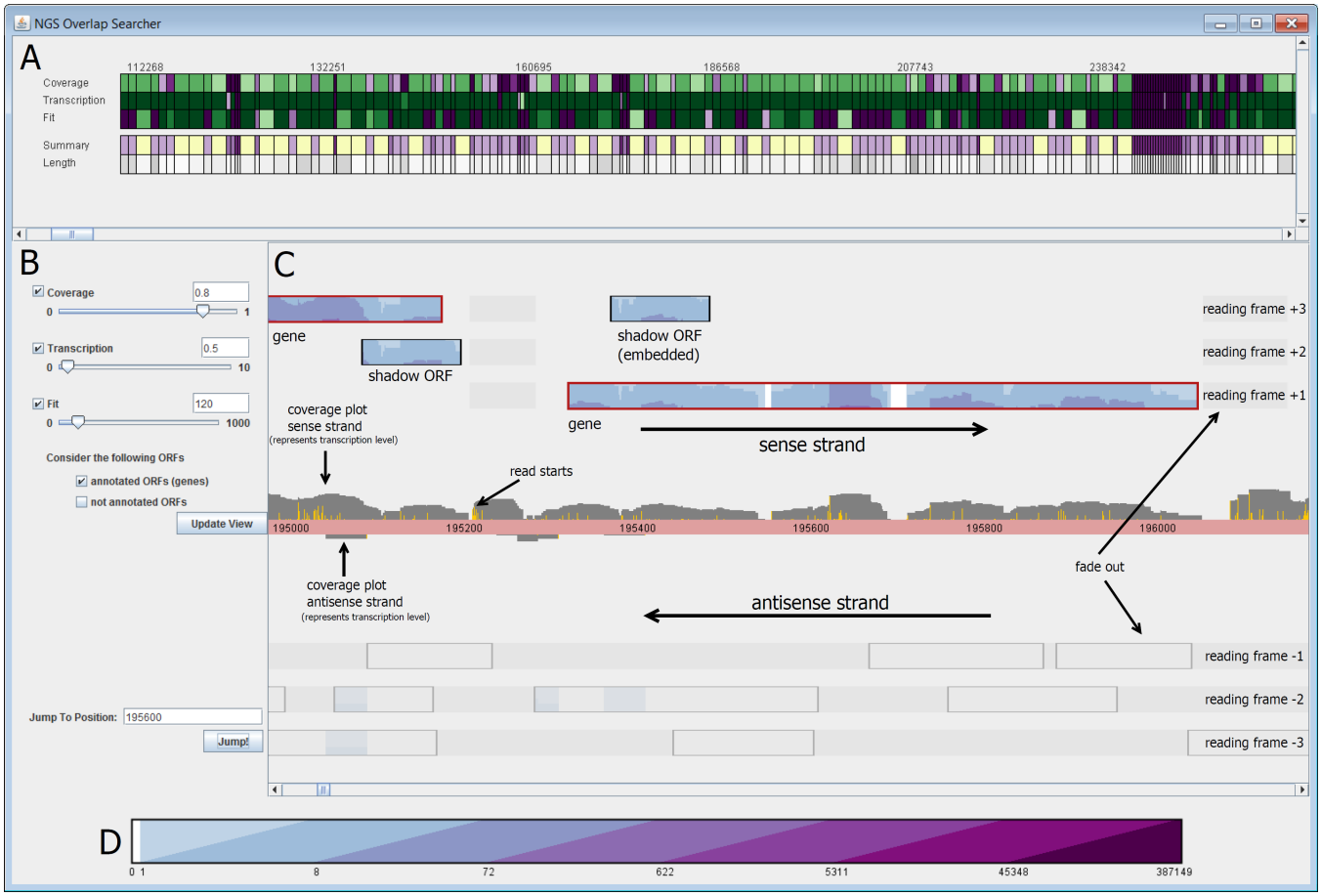
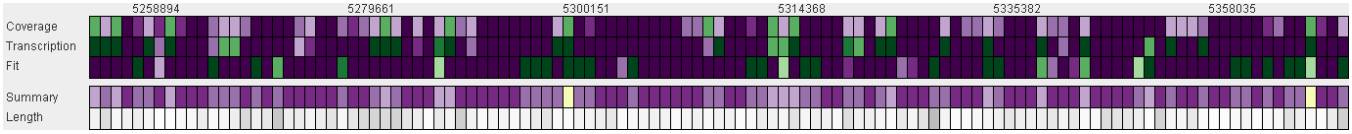
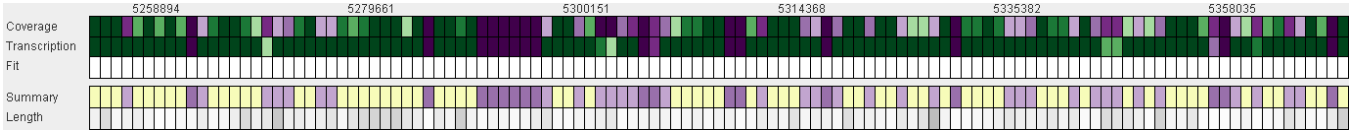


Figure 3: **A** shows the Genome Overview Bar. Every ORF (here only annotated ORFs) is represented by a column. Each line represents one parameter of the interestingness function and the coloring of the cells encodes whether the given threshold was passed or not. Distortion is used to highlight interesting regions. In **B** the interestingness measure can be parameterized. Furthermore, the search may be restricted to genes or not-annotated ORFs only. The genome view **C** consists of a plot for the read coverage in the middle, plus the six reading frames of the sense and antisense strand. ORFs with no transcription or a transcription level that does not allow them to pass the thresholds are faded out. **D** shows the color scale for the coverage plot in the ORFs. In the plot a gene is shown which can be considered as active, since nearly its whole region is covered with reads. There are only two small gaps, which decrease the coverage value (percentage of ORF covered with reads).



(a) Overview for a stringent parameter setting: Only two genes in this subsection of the Genome Overview Bar passed all three thresholds (shown in light yellow in the summary row). Some of the cells in the first line are colored in bright purple. This suggests that these regions are only slightly below the threshold for this criterion. Similarly, for the ones in bright purple in the summary row only a single criterion was below the threshold, hindering them to get through.



(b) Overview for a relaxed parameter setting which does not use the fit criterion. The majority of genes pass these thresholds.

Figure 4: Genome Overview Bar for different parameter settings. Only genes are shown and distortion is not applied.

However, through the coloring (many cells in this line are in light purple) it becomes apparent that slightly lowering the coverage threshold would let quite a couple of more regions pass this threshold. On the other hand, it might be more advisable to obtain a couple of very good hits that can be tested in wet lab experiments, which are time consuming. Thus, depending on the goals, the rate

of false positive or false negative hits can be adjusted. E.g., it might be of interest to get an impression of the amount of not-annotated ORFs which are transcribed. This could be achieved by lowering the thresholds to not miss too many interesting ORFs. To further support this task, not-annotated ORFs can be ignored. This way, adequate parameters can be assessed according to genome regions already better researched. The influence of adjusted parameter settings can be viewed in the genome overview bar. By clicking on a column in the genome overview bar the genome view jumps to the corresponding ORF. In this way promising ORFs can be investigated, but also columns representing ORFs that have not passed the thresholds can be inspected for a better understanding of the parameter settings.

It is also possible to exclude parameters from the search. In Figure 4(b) the minimum coverage was set to 60% and a transcription value of 0.5 was chosen. The parameter 'fit' could be excluded as bacteria can have genes that are located very close to each other. For those cases, the fit value is not meaningful anymore. On the other hand, to find transcribed not-annotated ORFs that overlap with a known gene, the fit value is very helpful. Otherwise, cases in which the transcript of an ORF is actually part of the overlapping gene would lead to many false positive hits.

4.2 Case Study: Detecting new overlapping genes

As said, a possible application of the tool is to search for overlapping genes. Many not-annotated ORFs overlap with a known gene in a different frame. How many of them encode proteins is debated, but surely more than anticipated before [6]. Our goal is to find such incidences by analyzing the transcripts of a genome under different experimental conditions. Thus, we are searching for transcribed ORFs that are overlapping with annotated genes [2].

With the help of the interestingness measure it is easy to locate such regions in the genome. However, false positives appear as well. In most of all cases, where the transcription covers a same strand overlapping ORF the transcription belongs to the gene (annotated ORF) and not the overlapping same strand ORF (see Figure 5 for an example). Thus, same strand overlapping ORFs are excluded.

For the analysis the thresholds were set as follows: Coverage = 80%, Transcription = 0.5, Fit = 120 (see Fig. 6). It can easily be seen in the summary line of the Genome Overview Bar that only few regions in the genome are able to pass all thresholds and meet the specified additional requirement. Next, the highlighted regions can be inspected one by one by clicking on them, to assess if they are indeed meaningful.

Figure 6 shows an example for a region that might contain an overlapping gene. Three not-annotated ORFs that are encoded in the sense strand do meet the specified criteria. One of them, which would then overlap with the already known gene of the antisense strand, could indeed encode a protein. Further evidence can be gained by comparing the specific region for different experimental conditions tested, taking additional meta-data into account (which will be added to future updates of the tool), and finally by testing the assumption in wet lab experiments. Figure 7 shows two further examples for promising findings.

5 Lessons Learned and Future Work

From our experience working with biologists using the tool, we learned the following lessons which we address in our future work.

1. **Comparison between different experimental conditions**
   Different experimental conditions are important to judge if an transcribed ORF is protein-coding or not. An overlapping gene might be found weakly expressed under one, but highly expressed under a different condition. Often multiple RNA-seq experiments are conducted for different conditions. In the future, we would like to ease the comparison of multiple experiments by integrating the RNA-seq data into one single view.

2. **Including additional data sources**
   Additional meta data will help the analysis. If for example some regions were found to carry more overlapping gene transcripts than others, it would be important to see if these regions belong to genome-integrated bacterial viruses (prophages) or "normal" genome regions. Furthermore, ORFs that have a significant BLAST hit are more likely to indeed encode a functional protein. Other protein identifying features, such as Shine-Dalgarno sequence, promoters, terminators, regions with signal peptides or a good secondary structure prediction would also support the hypothesis of a new gene.

3. **Evaluation**
   When working with the biologists, it turned out that it is difficult for them to describe their course of action when deciding on whether an ORF should be considered as transcribed or not. Some important criteria only became clear, when working together with them and discussing their analysis results. It certainly would be valuable to investigate the approach of the experimenters more systematically by conducting a field study. This would also help to understand better how the tool is currently used and what extensions are necessary.

4. **Finding appropriate thresholds**
   Setting the right thresholds in the interestingness function is difficult but critical for the analysis. Few NGS transcriptional studies have been published so far and general thresholds have not been established. Future wet lab confirmations will form a feedback-loop which helps to determine meaningful thresholds.

5. **Scalability**
   Finally, the scalability of the tool is an important issue. In all genome analysis projects, long linear sequences have to be processed which are difficult to display. In our research, we address this problem by an overview representation, which also eases navigation. Future versions of the tool may include a metabolic or regulatory pathway representation instead of linear sequences. Especially, in analyzing a time series of NGS experiments, similar transcriptional response patters of different ORFs can hint towards common regulators.

6 Related Work

NGS is a powerful technique that can be used not only for transcriptome sequencing, but also for *de novo* sequence assembly of a previously unknown genome. The DNA is fragmented and subsequently sequenced to obtain short reads. The challenge of the sequence assembly process is to reconstruct the original sequence from the sequenced pieces. In contrast to this, the task of read alignment, necessary for the transcription studies, is to map the sequenced reads to a given reference sequence. In both cases, visualization can help to inspect and are necessary to correct the result of the automatic algorithms. Popular visualization tools in this domain include Hawkeye [15], the ABySS-Explorer [11], or EagleView [7] (see also table 1 in [10]).

However, the analysis task conducted in this study starts after read alignment has been successfully finished. Given a reference genome and aligned read sequences of RNA transcripts, we aim learning about transcription at specific conditions. Viewing a genome is a classical task for genome browsers of which numerous variations exist (see e. g., [10]). Several genome browsers are

[2]Note that not any transcribed ORF necessarily also encodes a protein, but it provides some evidence that this might be the case.

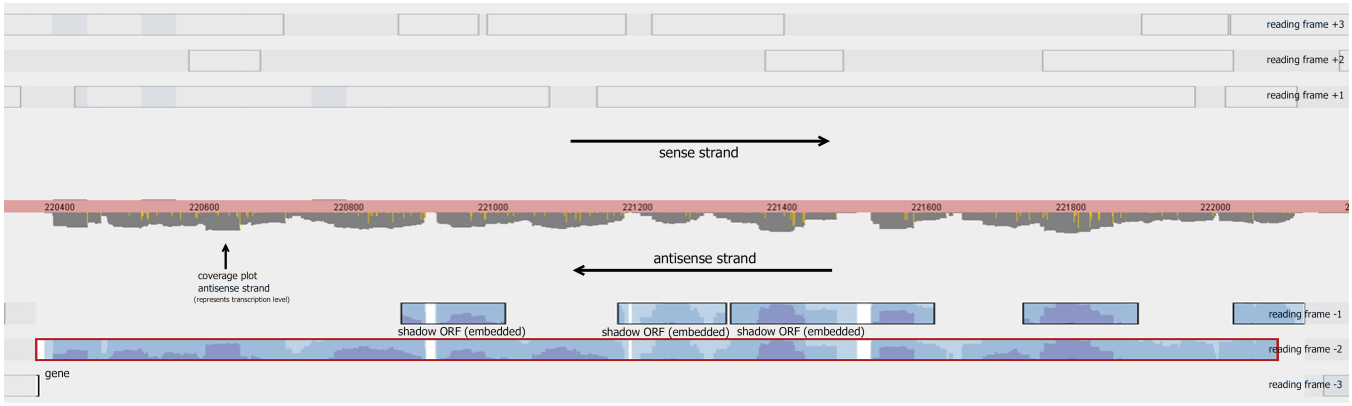


Figure 5: Example in which the transcription on the antisense strand clearly belongs to the gene (shown by the rectangle with the red border in reading frame -2) and not to the overlapping ORFs (in reading frame -1).

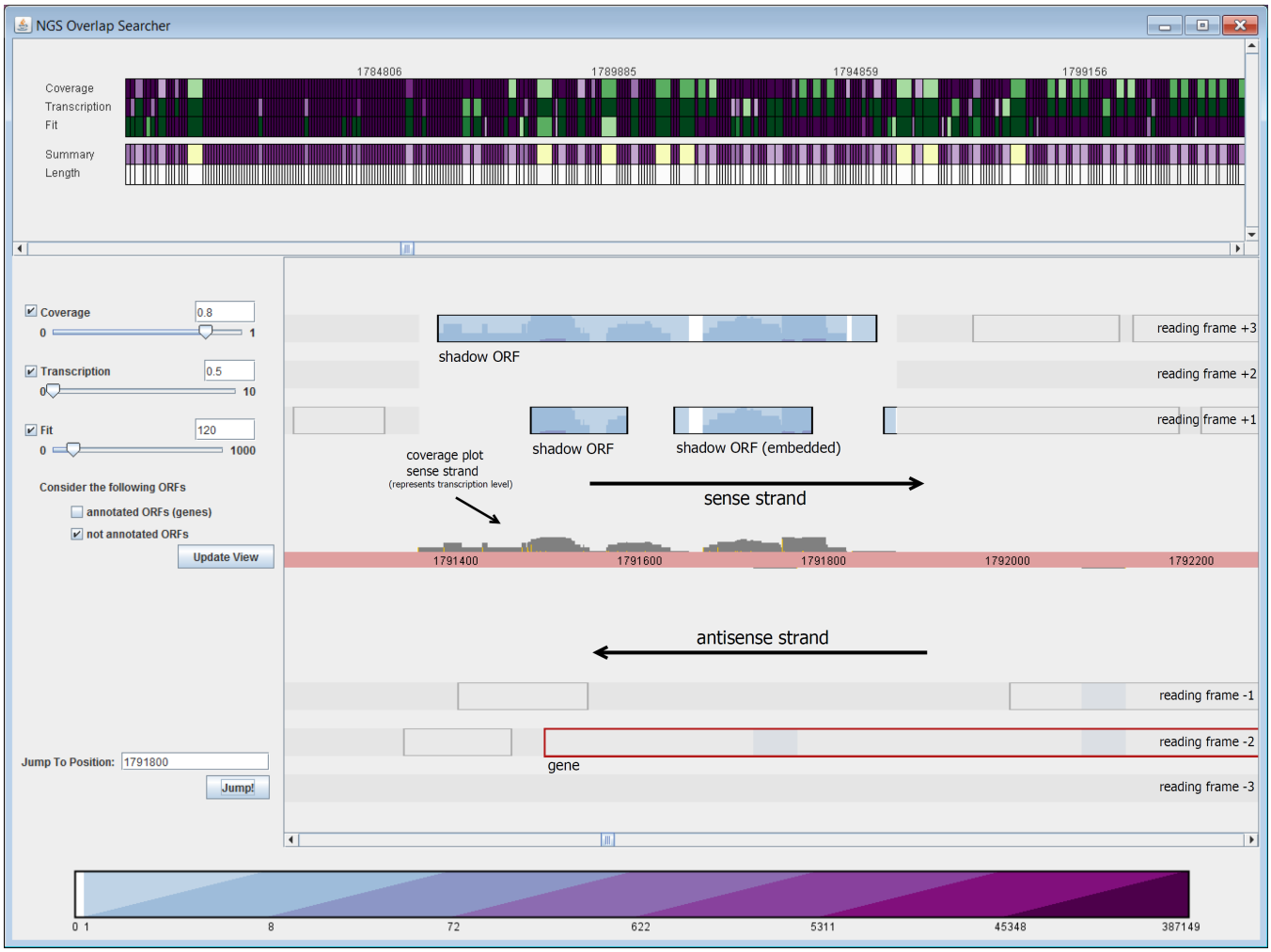


Figure 6: Example in which the region of the depicted ORF in reading frame $+3$ is nearly completely covered with reads. There are no genes on the same strand (sense strand) which could potentially be responsible for this transcription. On the antisense strand we can see a gene (hown by the rectangle with the red border in reading frame -2), which overlaps this ORF. Thus, this ORF would be a good candidate for future examination by wet lab experiments.

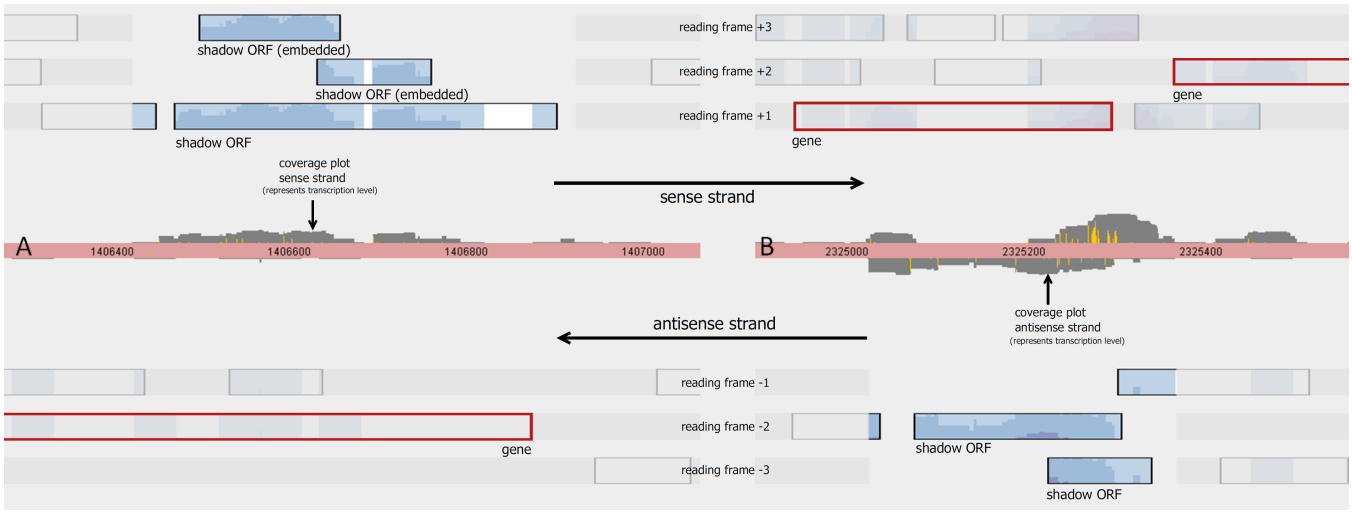


Figure 7: **A** On the sense strand two ORFs are shown which may be good overlapping gene candidates for future wet lab experiments (in reading frames $+1$ and $+3$). The longer ORF in reading frame $+1$ might be the better candidate, but the sloping coverage and the gap at the end of the ORF is indicative for the smaller ORF in reading frame $+3$. **B** On the antisense strand in reading frame -2 an ORF is shown which is completely covered with reads. Since this ORF is located opposite to a gene (in reading frame $+1$), this is also a good candidate for future examination by wet lab experiments.

able to display additional meta data of different kinds, e. g. NGS-reads. However, only few tools are able to deal with large size NGS-data, to show six frames separately, or to distinguish between reads in sense and antisense (compare to section 2). Besides, there are tools whose graphical representations aim at supporting special tasks. LookSeq [9] for instance uses a stack view which enables the identification of deletions and insertions by using paired-end reads. The Integrative Genomics Viewer (IGV) [12] integrates different data types and supports among others sequence alignments, microarrays, and genomic annotations.

For sophisticated analyses, development of advanced NGS viewers is indispensable, especially to solve the problem of having too much data to evaluate them manually. Our new tool with enhanced functionality in terms of automatic algorithms visually supports the expert analyst in getting insight into the data, thereby following the central paradigm of visual analytics.

7 Conclusions

In this paper, we introduced a system for analyzing next generation sequencing data from transcriptome sequencing (RNA-seq) with visual analysis techniques. Specific emphasis was put on an expressive visualization of the sequence coverage with a low mental load to determine ORF-fits. Furthermore, with the help of an adjustable interestingness function, the search of the user is guided. An overview representation guides the experimenters in the analysis and gives insight with respect to the values of the interestingness function for the different regions. A case study shows how the tool can be used to detect overlapping ORFs that potentially encode previously unknown genes. In the future, we will continue to improve the tool by integrating meta data as outlined above.

Acknowledgements

This work has been partly funded by the German Research Society (DFG) under the grant SPP 1395 (Informations- und Kommunikationstheorie in der Molekularbiologie, InKoMBio), project 'Finding new overlapping genes and their theory (FOG-Theory)'.

References

[1] Artemis: Genome Browser and Annotation Tool. `http://www.sanger.ac.uk/resources/software/artemis/`.

[2] Genome Studio. `http://www.illumina.com/software/genomestudio_software.ilmn`.

[3] UCSC Browser. `http://genome.ucsc.edu/`.

[4] SOLiD™ System High-throughput Analysis of Differential Gene Expression, 2008. Applied Biosystems, Application Note.

[5] J. Heer, N. Kong, and M. Agrawala. Sizing the Horizon: The Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations. In *ACM Human Factors in Computing Systems (CHI)*, 2009.

[6] M. R. Hemm, B. J. Paul, J. Miranda-Rios, A. Zhang, N. Soltanzad, and G. Storz. Small stress response proteins in *Escherichia coli*: Proteins missed by classical proteomic studies. *J. Bacteriol.*, 192(1):46–58, 2010.

[7] W. Huang and G. Marth. Eagleview: A genome assembly viewer for next-generation sequencing technologies. *Genome Research*, 18(9):1538–1543, 2008.

[8] W. Kim, M. W. Silby, S. O. Purvine, J. S. Nicoll, K. K. Hixson, M. Monroe, C. D. Nicora, M. S. Lipton, and S. B. Levy. Proteomic Detection of Non-Annotated Protein-Coding Genes in *Pseudomonas fluorescens* Pf0-1. *PLoS ONE*, 4(12):e8455, 12 2009.

[9] H. M. Manske and D. P. Kwiatkowski. Lookseq: A browser-based viewer for deep sequencing data. *Genome Research*, 19(11):2125–2132, 2009.

[10] C. B. Nielsen, M. Cantor, I. Dubchak, D. Gordon, and T. Wang. Visualizing genomes: techniques and challenges. *Nature methods*, 7(3 Suppl), March 2010.

[11] C. B. Nielsen, S. D. Jackman, I. Birol, and S. J. M. Jones. Abyss-explorer: Visualizing genome sequence assemblies. *IEEE Transactions on Visualization and Computer Graphics*, 15:881–888, 2009.

[12] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, Jan. 2011.

[13] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M.-A. Rajandream, and B. Barrell. Artemis: sequence visualization and annotation. *Bioinformatics*, 16(10):944–945, 2000.

[14] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya, and T. Kaseda. Two-Tone Pseudo Coloring: Compact Visualization for One-Dimensional Data. In *Proceedings of the 2005 IEEE Symposium*

on Information Visualization, pages 23–30. IEEE Computer Society, 2005.

[15] M. Schatz, A. Phillippy, B. Shneiderman, and S. Salzberg. Hawkeye: an interactive visual analytics tool for genome assemblies. *Genome Biology*, 8:1–12, 2007.

[16] G. O. S. Thomassen, R. Weel-Sneve, A. D. Rowe, J. A. Booth, J. M. Lindvall, K. Lagesen, K. I. Kristiansen, M. Bjrs, and T. Rognes. Tiling Array Analysis of UV Treated *Escherichia coli* Predicts Novel Differentially Expressed Small Peptides. *PLoS ONE*, 5(12):e15356, 12 2010.

[17] S. Tunca, C. Barreiro, J.-J. R. Coque, and J. F. Martn. Two overlapping antiparallel genes encoding the iron regulator DmdR1 and the Adm proteins control sidephore and antibiotic biosynthesis in *Streptomycescoelicolor* A3(2). *FEBS Journal*, 276(17):4814–4827, 2009.