We are IntechOpen,
the world's leading publisher of
Open Access books
Built by scientists, for scientists

6,900
Open access books available

186,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Predicting Tandemly Arrayed Gene Duplicates with WebScipio

Klas Hatje and Martin Kollmar
Abteilung NMR basierte Strukturbiologie, Max-Planck-Institut für Biophysikalische Chemie, Am Fassberg 11, Göttingen
Germany

1. Introduction

Since the first high-quality eukaryotic genome assemblies became available the large scale analysis of the origin of new genes came into the focus of many studies (Shoja & Zhang, 2006; Zhou et al., 2008). New genes can originate through multiple mechanisms including gene duplication, gene fusion/fission, exon shuffling, retroposition, horizontal gene transfer, and de novo from noncoding sequences (Long et al., 2003). Although initial models proposed that new copies of genes soon become nonfuntional (Nei & Roychoudhury, 1973; Ohno, 1970) it has since been shown for numerous genes that they retain function through creating redundancy, subfunctionalization, and neofunctionalization (Hahn, 2009; Li et al., 2005; Massingham et al., 2001). While de novo origination from noncoding sequence has been shown to play an unexpectedly important role (Zhou et al., 2008) most of the new genes are derived through duplications. Gene duplicates are normally classified into dispersed and tandem duplicates. Tandem duplications of clusters of genes, single genes, groups of exons, or single exons are thought to be formed by unequal crossing-over events, or misaligned homologous recombinational repair (Babushok et al., 2007; Zhang, 2003). A comparative analysis of the human, mouse, and rat genome has shown that about 15 % of all genes represent tandemly arrayed genes (Shoja & Zhang, 2006). A similar number of about 20 % has been found for the fruit fly *Drosophila melanogaster* (Quijano et al., 2008). All these analyses rely on the particular dataset of annotated genes used and the specific methods for defining genes as tandem genes. However, first annotations of genomes are in most cases done by automatic gene prediction programs, nowadays often supported by incorporating additional EST data, and therefore miss many genes, include artificially fused neighbouring genes, and contain mis-predicted exons and introns. Although these errors seem small, in the case of distinguishing tandem gene duplicates from genomic region duplication and *trans*-spliced genes they are essential. In addition, defining tandem genes by a certain number of nucleotides appearing in-between cannot separate tandem gene duplicates from duplications of small genomic regions. Tandemly arrayed gene duplicates are often conserved between species. Examples are the olfactory receptor genes that constitute a very large gene family of several hundred genes per species in vertebrates (Aloni et al., 2006) and the HOX genes (Garcia-Fernandez, 2005; Zhang & Nei, 1996). While algorithms have been developed to reconstruct the history and evolution of tandemly arrayed genes (Bertrand et al., 2008; Elemento et al., 2002) specific programs are not available for the prediction and local reconstruction of these gene arrays.

WebScipio is a web application to reconstruct genes based on a given protein query sequence and a genomic DNA target sequence (Odronitz et al., 2008). The reconstruction is done with Scipio (Keller et al., 2008), a post-processing script for the output of a BLAT run (Kent, 2002). BLAT is a very fast tool for the alignment of protein or DNA sequences if these sequences are almost identical. However, BLAT is not able to reconstruct intron and exon borders, it does not identify very short exons and very divergent exons, and it is not able to reconstruct genes spread on several pieces of contiguous DNA (contigs), which is very common in low-coverage genome assemblies. Furthermore, BLAT is not able to identify sequencing and assembly errors like additional or missing bases in exon regions or base substitutions leading to in-frame stop codons. Scipio is able to correct all these errors and extend the BLAT output for the missing sequences of short or divergent exons and of exon borders. In addition, Scipio assembles genes spread on several contigs. WebScipio has been developed as a web interface to Scipio so that the user does not have to install scripts and libraries. Moreover, WebScipio offers access to about 2300 genome assembly files of more than 650 sequenced eukaryotes (July 2011), and provides graphical and human-readable analyses of the results.

Here, we present an extension to the WebScipio web application to search for and predict tandemly arrayed gene duplicates for a given query sequence. This extension is not available via the Scipio command-line script. The user can search for gene duplicates in hundreds of species for which reliable annotations are not available yet, because WebScipio provides access to thousands of genome files.

2. Implementation

The new algorithm to predict tandemly arrayed gene duplicates is fully integrated into the web application WebScipio to make it usable for the inexperienced user and to visualize the results for immediate analysis. It was implemented in the Ruby programming language (Ruby Programming Language, 2011) using the BioRuby library (Goto et al., 2010) to handle sequences. WebScipio is based on the web framework Ruby on Rails (Ruby on Rails, 2011), which includes the Javascript libraries Prototype (Prototype JavaScript framework: Easy Ajax and DOM manipulation for dynamic web applications, 2011) and Scriptaculous (script.aculo.us - web 2.0 javascript, 2011). To keep the web application responsive, the search algorithm runs in the background with the help of the Ruby on Rails plug-ins Workling (purzelrakete's workling at master - GitHub, 2011) and Spawn (tra's spawn at master - GitHub, 2011). To store the user session data, the database backend Tokyo Tyrant is used in combination with Tokyo Cabinet (Tokyo Cabinet: a modern implementation of DBM, 2011). The results of the search are presented as SVG pictures (W3C SVG Working Group, 2011) and several human-readable representations, most notably a detailed alignment of protein query, target DNA sequence, and target translation. The raw results can be downloaded as General Feature Format (GFF) files or as YAML files (The Official YAML Web Site, 2011) for future upload and analysis. Specific results are available in various formats for further inspection, like the human-readable log-files, or publication quality figures, like the SVGs.

2.1 Search algorithm

The overall workflow of the search algorithm is shown in Fig. 1. The search for tandem gene duplications is based on the exon-intron structure of a gene generated by Scipio. Thus the first step of the algorithm includes a WebScipio run generating a new gene structure or the upload of an existing Scipio result.
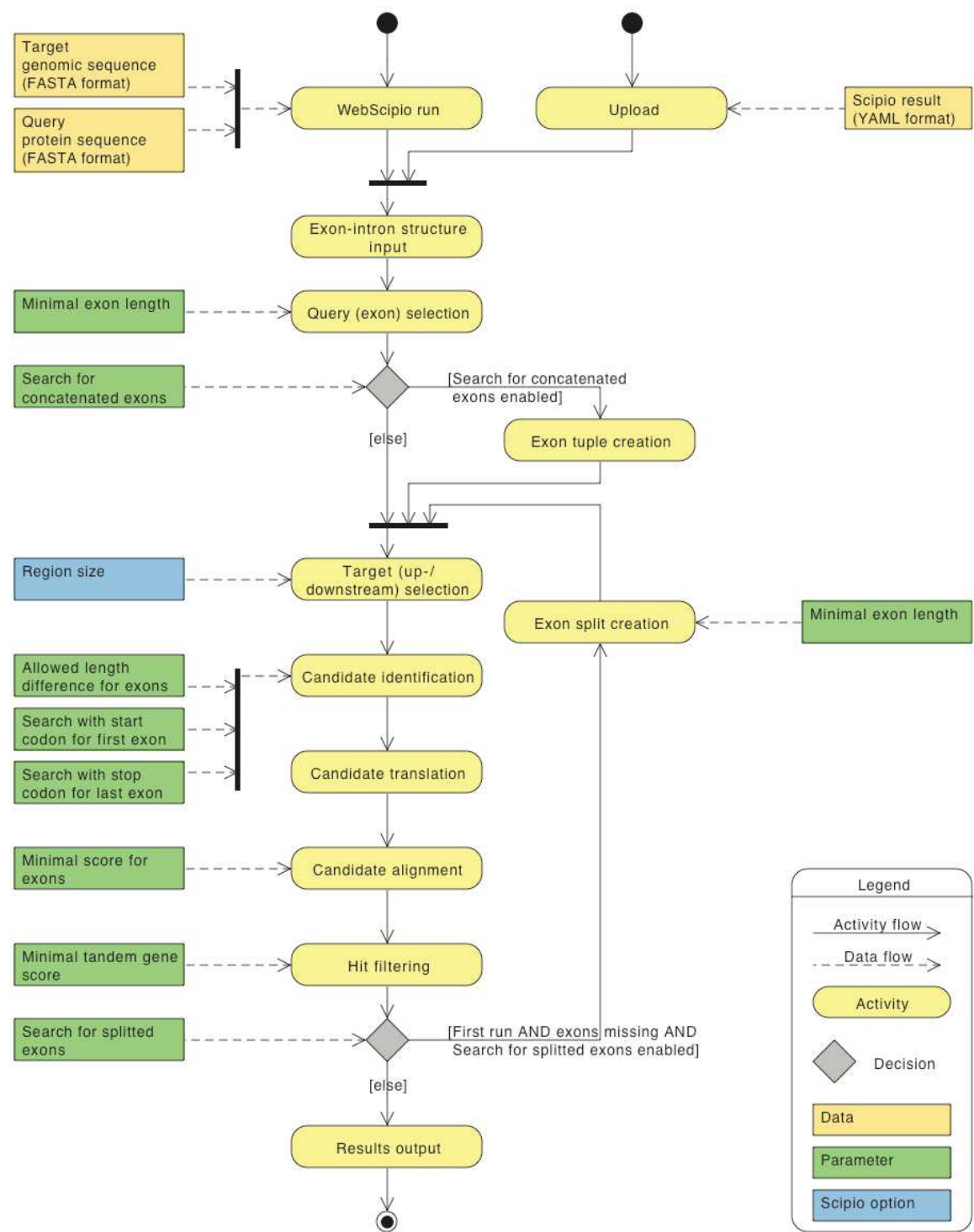
Fig. 1. Activity flow diagram of the search for tandem gene duplications: The activity diagram shows the processing steps of the search algorithm and the influence of the parameters on each step. The run starts with an exon-intron gene structure determined by Scipio. Based on the chosen parameters the exons and up- and downstream regions are selected and searched for candidate exons of gene duplicates. The candidates are processed and filtered. These steps are repeated for exons that have not been found. Those exons are splitted and the search is repeated with fragments. In the end, the algorithm outputs the exon-intron structure of the original gene and all gene duplicates.

2.1.1 Query and target selection

The next steps are the selection of the query and the target for the search. All exons, which are longer than a minimal length, are selected as query. The minimal length can be adjusted by the *minimal exon length* parameter, which is given in number of amino acids coded by the exon. In addition, the algorithm is able to generate exon tuples by the fusion of neighbouring exons to one exon. This means that all pairs (2-tuples) of consecutive exons, triplets (3-tuples), 4-tuples, 5-tuples, up to all exons are concatenated and used as query exons. This option can be enabled by the *search for concatenated exons* parameter. The nucleotide sequences of the up- and downstream regions of the gene are used as target sequences. The lengths of these sequences are determined by the Scipio parameter *region size* in number of nucleotides. The up- and downstream sequences are scanned in forward and reverse direction. For the reverse strand the reverse complements of the given target sequences are created.

2.1.2 Candidate identification

The query and target selection steps are followed by the search for exon candidates in the target sequences. The search algorithm assumes that exons of gene duplications have a similar length, share sequence similarity, are translated in the same reading frame and have conserved splice sites. Candidate exons are determined in the target sequences for each exon of the original gene and each exon tuple. The target nucleotide sequences are scanned for sequence sections, which do not differ more than a maximal number of nucleotides from the original exon length. This maximal difference is given by the *allowed length difference for exons* parameter in number of amino acids. In addition, the sequence section, which determines an exon candidate, must be flanked by a splice site pattern that corresponds to the introns surrounding the original exon or exon tuple. Allowed splice site patterns for the first two and last two nucleotides of these introns are GT---AG, GC---AG, GG---AG, and AT---AC. The first exon of a gene must start with the start codon ATG and the last exon must be followed by one of the stop codons TAG, TAA, or TGA. To allow searches for partial genes, the algorithm is able to find candidates corresponding to the first and last exon of the gene fragment that share splice site patterns instead of having a start codon or stop codon. This behaviour can be adjusted by the *search with start codon for first exon* and *search with stop codon for last exon* parameters.

2.1.3 Candidate translation and alignment

Candidate sequences are translated to amino acids in the same reading frame as the original exon. If a candidate sequence includes a stop codon, the candidate is rejected immediately. The translations of the candidate exons are aligned to the original exon translations by a global alignment algorithm. The pair_align tool of the SeqAn package (Doring et al., 2008) is used for this task. The resulting alignment score is divided by the score resulting from the alignment of the original exon translation to itself. This normalised score makes exons of different lengths and amino acid compositions comparable. Finally, exon candidates having a score lower than the score given by the *minimal score for exons* parameter are rejected.

2.1.4 Hit filtering

The resulting candidate hits are filtered. If candidate sequences are overlapping, the lower scoring candidates are rejected. Neighbouring candidate exons are combined to genes if they

are in the same order as the original exons. For each identified tandem gene a score is calculated that reveals how many residues of the original gene were found in the tandem gene duplication. The score is calculated as the number of residues of the original gene that are aligned to residues in the tandem gene duplicate (and not to gaps) divided by the number of all residues of the original gene. The tandem gene duplications that have a low score are rejected. This behaviour can be adjusted by the *minimal tandem gene score* parameter.

2.1.5 Exon split run

If exons of a duplicated gene are missing, either in between two neighbouring exons, at the start of the gene or at the end, the search is repeated for these exons by splitting the missing original exons into pieces. The original exon sequences are split in two parts at each nucleotide as long as the smaller part is longer than the minimum exon length. The algorithm scans the intron regions of the duplicated genes that miss exons for candidates corresponding to these exon splits, each composed of two parts. Thus, exons, which are split by an intron in the duplicated gene, are found too. This option can be enabled by the *search for splitted exons* parameter.

2.1.6 Results output

The output of the search algorithm is the exon-intron structure of all identified tandem gene duplications combined in one result, and the exon-intron structure of each duplicated gene alone. For every result a gene structure drawing is shown, as well as several options to further examine gene details like the alignment of the query sequence to the translation of the hit and the hit itself (Fig. 2).

2.2 WebScipio integration

The search algorithm is fully integrated into the web interface of WebScipio. The search for tandem gene duplications can be enabled in the Advanced Options section. WebScipio provides an interface to easily set the parameters, suggests default parameters, which will be suitable for most cases, and offers documentation at several help pages and examples. The raw results for the gene cluster can be downloaded all together in one YAML file or the result for each gene of the cluster in a separate file. In addition to the raw data, the SVG figures of the gene structures and FASTA files of the sequences (cDNA, genomic DNA, exons, introns, target translation) are available for download. WebScipio provides an upload option for downloaded YAML files to let the user analyse his results at a later date.

3. Results and discussion

WebScipio uses the command-line tool Scipio to reconstruct the gene structures of given protein sequences based on the available eukaryotic genome assemblies. Scipio has been developed for the case that protein sequences and target genome sequence are from the same organism. Nevertheless, Scipio allows several mismatches that might result from sequencing and assembly errors like missing or additional bases, which lead to frame-shifts, or in-frame stop codons that would lead to premature gene stops. Mismatches might also be the result from differences in the source of the protein sequence, which might have been obtained from cDNA libraries of a certain strain, and the specific sequenced strain of the species. To accomplish this task, Scipio relies on BLAT, which is one of the fastest tools available for
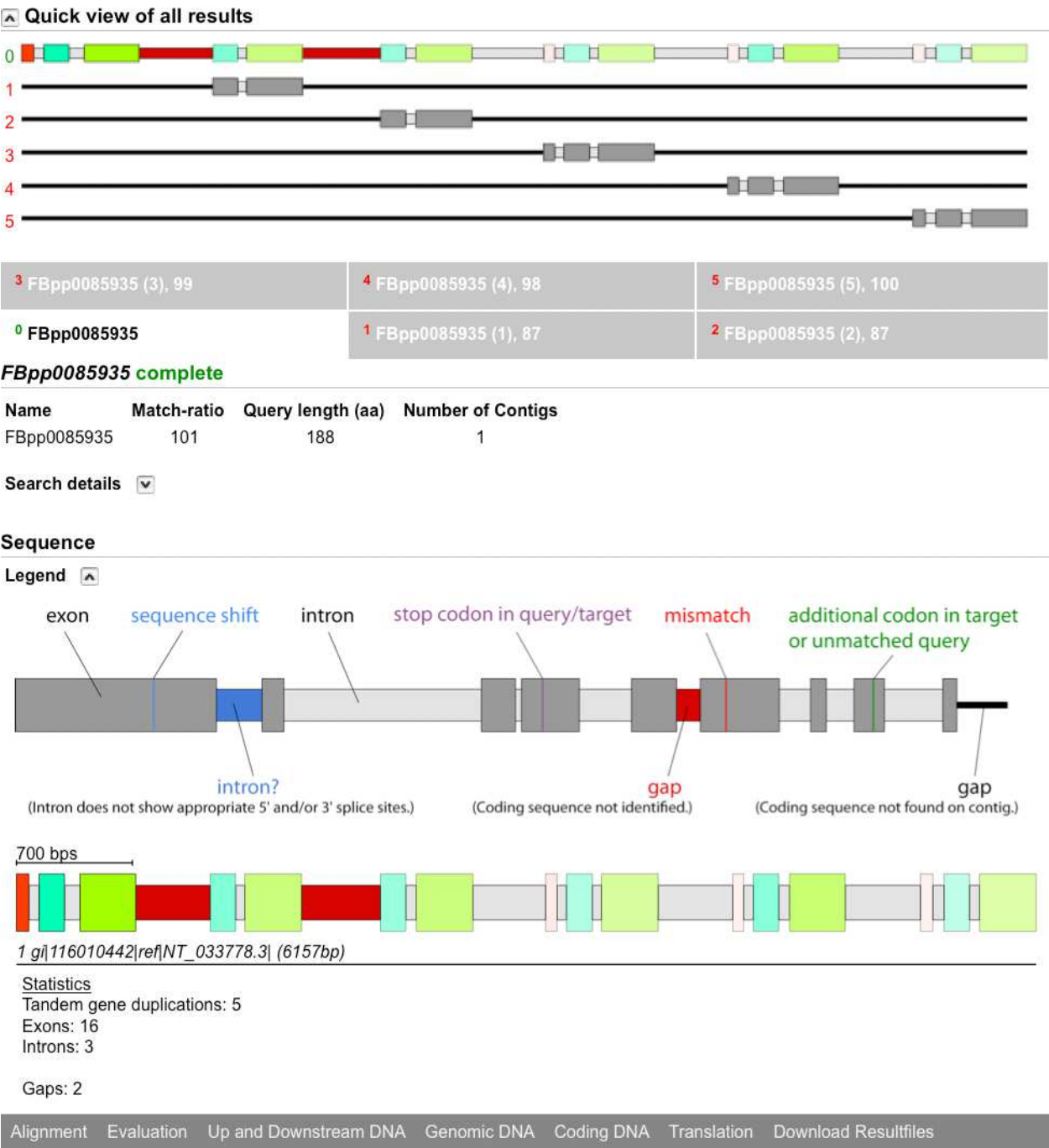
Fig. 2. WebScipio result view of the search for tandem gene duplications of the *Drosophila melanogaster* CG14502 gene (Flybase sequence accession FBpp0085935): Exons are illustrated as coloured rectangles, introns as grey narrow rectangles, and gaps as red narrow rectangles. Gaps indicate missing exons of the tandem gene duplicates. For the search the default parameters were used except for the *minimal score for exons* parameter that was set to 5 % to find some exon duplicates of the first exon.

the alignment of almost similar protein or DNA sequences. As Scipio tolerates a certain amount of mismatches between query and target sequence it can also successfully be used for cross-species gene reconstructions and predictions (Odronitz et al., 2008). Because Scipio relies on BLAT the success of the cross-species search depends on the difference between the

query and target gene. If genes are highly conserved in evolution Scipio is able to correctly reconstruct genes in species that diverged hundreds of million years ago. If genes evolve fast Scipio can predict genes only in very related organisms. This behaviour can also be used to predict gene duplicates in the same organism, and is implemented as *multiple results* parameter in the Scipio options. Again, because Scipio relies on BLAT, only those duplicates will be identified that are very similar. An advantage of this option is that Scipio is able to find dispersed as well as tandem duplicates.

In an analysis of the origin of new genes in the *Drosophila* species complex (Zhou et al., 2008) it has been shown that the majority of the constrained functional new genes are dispersed duplicates. In contrast, tandem duplications were found to be young events and to lead to lower survival rates. Thus, tandem duplicates are often pseudogenes most probably because the introduction of frame shifts and in-frame stop codons does not demand too many mutations to destroy the transcription and expression of the new gene. If duplicates are kept in the genome they acquire new functions through neofunctionalization and subfunctionalization by accumulation of many substitutions (Ohno, 1970). Those genes are too divergent to be identified by the *multiple results* option of Scipio. However, although accumulating many substitutions tandem duplicates very often retain the gene structure of the original gene including intron splice sites and reading frames of exons. Occasionally further introns might be introduced or prior existing introns lost because these changes would not destroy transcription and translation. To use this knowledge in tandem gene duplicate identification we developed an algorithm that searches for duplicates of a query sequence based on the restrictions imposed by its gene structure. Every piece of DNA in the up- and downstream region of the original exon that has the same splice sites and shares sequence homology to the original exon, when translated in the same reading frame, is thought to be a candidate for an exon of a duplicated gene. In the case that introns have been lost or gained in the duplicated genes the splice site restrictions apply to the outer borders of the fused or split exons. WebScipio is able to correctly reconstruct the gene structure for a given protein sequence and is thus very suited as starting point for searches for candidate exons of duplicated genes.

To search for tandem gene duplicates an extension to WebScipio was implemented providing several parameters to adjust the search according to users or genome-specific needs. In most cases, however, the standard parameters will provide reasonable and interpretable results. As soon as the search is done, WebScipio shows an overview of the results as small gene structure pictures (Quick View), which reveal the exon regions of the found tandem genes (Fig. 2). For convenient analysis the genomic region comprising the gene structure of the query sequence and the exons of the predicted tandem genes is shown in a combined graph and provided as one YAML file. The exons of the original gene are dark coloured and the corresponding predicted exons have the same but lighter colour. The darkness of the colour relates to the similarity of the predicted exon to the original one. The same colour scheme is used to highlight the various exons in the Alignment view of the genomic regions (Fig. 3). The Alignment view shows the nucleotide sequence of the gene ordered in exons and introns. For every exon the genomic DNA and the corresponding translation are shown, as well as the alignment of the query sequence to the translation.

To demonstrate the application, quality, and limitations of the new algorithm we provide some example searches in the following sections. Tandemly arrayed gene duplicates have several characteristics that need to be considered. Gene duplications can be found on both the forward and the reverse strand. The duplicated genes might contain fused exons or

Intron 7 ⊻

Duplicated Exon 3.1 ⌃   Tandem Gene Score: 99 %, Exon Score: 11.2 %

```
     ATGTCTGTCCGGTGTTTTATTGCTGAACTT------TGTATGGCACTCCTGCTCGTAGGTGAATGCCAAACG    14184640
      M   S   V   R   C   F   I   A   E   L   █  █   C   M   A   L   L   L   V   G   E   C   Q   T        |
      |   X   X   |   X   X   X   X   X   X   █  █   X   X   X   |   |   X   |   X   X   |   X   X        |
      M   L   P   R   L   L   L   P   I   I   G   G   V   L   I   L   L   I   V   D   Q   C   N   G       24

     GCG                                                                                 14184643
      A                                                                                    |
      X                                                                                    |
      K                                                                                   25
```

Intron 8 ⊻

Duplicated Exon 3.2 ⌃   Tandem Gene Score: 99 %, Exon Score: 36.43 %

```
     CGGAAATGGGACTACGAACCCATATCCATCGAAACCCACTCATCCGATGAGAGTTTAATAAAATTAGATATG    14184772
      R   K   W   D   Y   E   P   I   S   I   E   T   H   S   S   D   E   S   L   I   K   L   D   M        |
      |   |   |   |   |   |   |   X   |   |   |   X   X   X   |   |   X   X   X   |   |   X   X   X        |
      R   K   W   D   Y   E   P   L   S   I   E   T   Y   T   T   D   E   T   K   L   K   I   A   A       49

     AAAATTGAGCGAATCAATCGCGGAGTGTTTGGATTGACGCCAAGGTTGAGTGGAACTACGATACGACTGATG    14184844
      K   I   E   R   I   N   R   G   V   F   G   L   T   P   R   L   S   G   T   T   I   R   L   M        |
      |   X   |   |   X   |   |   X   X   |   X   X   X   X   X   X   X   X   X   X   X   X   X   X        |
      K   V   E   R   V   G   R   G   E   Y   A   L   S   G   N   I   E   █  F   K   Y   T   P   E       72

     AAACCATGGTAC                                                                        14184856
      K   P   W   Y                                                                        |
      X   X   X   X                                                                        |
      D   N   T   M                                                                       76
```

Intron 9 ⊻

Duplicated Exon 3.3 ⊻   Tandem Gene Score: 99 %, Exon Score: 44.21 %

Intron 10 ⊻

Fig. 3. Alignment view of the first two exons of the third gene duplicate of the
Drosophila melanogaster CG14502 gene (see also Fig. 2): Each exon is named by the
tandem gene number and the exon number. In addition, the tandem gene score and the
exon score is given for each exon in percentage. The alignment indicates the positions of
the sequences in the genome and the protein. The first line of the alignment represents
the nucleotide sequence of the gene and the second line the translation of this sequence.
The third line shows how the amino acids of this translation match the amino acids
coded by the original exon, which are shown in the forth line. Mismatches are
represented by an X in dark red or, if amino acids are chemically similar, in light red.
Gaps in the alignment are shown as hyphens in green. The Duplicated Exon 3.3
alignment has been closed for representation purposes.

might contain additional introns. In the case of retroposed genes, which are derived from
the reverse transcription and insertion of processed genes, gene duplications do not have

any introns. Although gene duplications are more often found for small genes consisting of one or only a few exons, gene duplicates can also be identified for genes consisting of dozens of exons spanning large genomic regions. Because tandem gene duplicates are defined by being located next to each other in the genome, intergenic regions are expected to be short. This is also the reason why the parameter for bordering the search in up- and downstream regions of the original gene limits this region to 300,000 nucleotides. However, WebScipio cannot exclude that there may be additional genes in-between gene duplicates. An example for such a scenario would be the duplication or multiple duplications of small genomic regions that encode several genes. In most cases we considered examples from the fruit fly *Drosophila melanogaster* and sequences from Flybase (Tweedie et al., 2009), because the corresponding genome is of high quality and the annotation of the genome is already at a very advanced stage. Fragmented genomes, like draft genomes for which only short contigs are available, or chromosome assemblies containing many gaps, are useful to screen for interesting candidates but do not provide the reliability needed for tests of the algorithms quality and limitations. An advanced annotation provides the advantage that genomic locations of most genes have already been identified. Thus the gene order is already established although there might still be errors in the annotation of single exons.

3.1 Examples of tandemly arrayed gene duplicates
3.1.1 Gene duplicates on both the forward and the reverse strand
The WebScipio tandem gene duplication extension has been developed to find tandem gene duplications on the forward as well as on the reverse strand in relation to the query gene. The example in Fig. 4 shows five gene duplicates of the *Drosophila melanogaster* heat shock protein 23 gene (Hsp23), which consists of one exon. The first duplicate (Hsp67Bc) and the forth duplicate (Hsp26) in the genomic region are on the reverse strand, the other duplications Hsp22, CG4461, and Hsp27 are in the same reading direction as Hsp23. This search was performed with default parameters except increasing the *allowed length difference for exons* parameter to 30 amino acids. The most divergent gene duplication Hsp67Ba (Table 1), which is encoded in the genomic region between Hsp26 and Hsp23, was not found. This example shows that although the sequence identity is very low between the duplicates and the Hsp23 search sequence (Table 1), five duplicates could be identified. The length difference between Hsp23 and Hsp67Ba was too large so that candidates of the length of Hsp67Ba were not included in the search with the given search parameters.

	Hsp67Bc	Hsp22	CG4461	Hsp26	Hsp67Ba	Hsp23	Hsp27
Length [aa]	199	174	200	208	445	186	213
Identity to Hsp23	0.29	0.31	0.26	0.49	0.15	1.00	0.41
Strand	rev	for	for	rev	rev	for	for

Table 1. Comparison of the length, similarity, and reading direction of the genes of the *Drosophila melanogaster* heat shock protein cluster.

Drosophila melanogaster heat shock protein 23 gene and duplicated genes on both strand

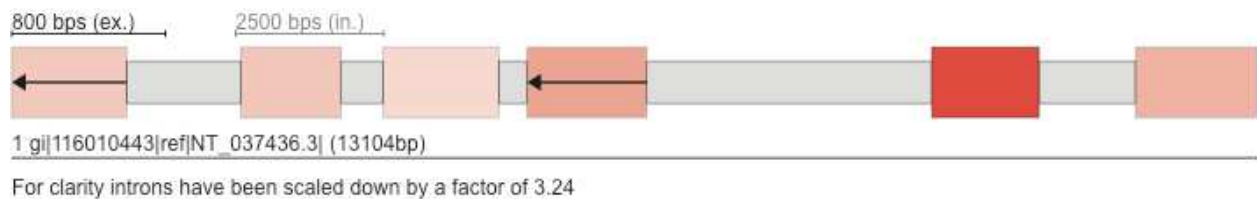For clarity introns have been scaled down by a factor of 3.24

Fig. 4. _Drosophila melanogaster_ heat shock protein gene duplicates: The figure shows the duplications found by the algorithm with Hsp23 as query. The genomic region contains, from the left to the right side in the drawing, the identified genes Hsp67Bc, Hsp22, CG4461, Hsp26, the query gene Hsp23, and another gene duplicate Hsp27. Gene duplications on the reverse strand are marked by an arrow in reverse direction.

3.1.2 Duplicated exons in six tandemly arrayed genes including a lost intron and a pseudogene

The new algorithm is able to reconstruct tandemly arrayed gene duplications containing many exons and gene duplicates. The _Drosophila melanogaster_ CG30047 gene includes 12 exons. Five duplicates of this gene could be identified with the algorithm (Fig. 5, top). In the second duplicated gene an intron loss could be identified. The exons 11 and 12 of CG30047 are translated as one exon in this duplicate (Fig. 5, bottom). To find such lost introns the option to _search for concatenated exons_ has been enabled. The third duplicate most probably represents a pseudogene, because exon 11 contains a frame shift and could thus not be found. Other reasons for the frame shift could be sequencing and assembly errors. However, the _Drosophila melanogaster_ genome (Adams et al., 2000) is one of the best available and a lot of effort has been spent in the finishing process. Thus, it is more probable that the third duplicate is a pseudogene. Exon 1, which codes for seven amino acids, has low complexity and could therefore only be identified in the second gene duplication by setting the _minimal exon length_ parameter to 7 aa.

3.1.3 Myosin heavy chain gene duplicates

Mammals encode two clusters of muscle myosin heavy chain genes, one cluster containing the α- and β-cardiac muscle myosin heavy chain genes (Saez et al., 1987; Weydert et al., 1985), and one cluster containing six skeletal muscle myosin heavy chain genes in the order embryonic, 2a, 2x, 2b, perinatal, and extraocular (Sun et al., 2003; Weydert et al., 1985). These myosin genes consist of 38 exons each. Based on their gene size and number of exons the genes of the muscle myosin gene cluster should be on the upper limit of the complexity of a search for tandem gene duplicates. With the new WebScipio extension all genes of the muscle myosin cluster in _Homo sapiens_ could be identified (Fig. 6). For the search the _region size_ parameter was set to 300,000 nucleotides and the _minimal score for exons_ to 50 %. This example also shows the advantage of the new WebScipio extension compared to the _multiple results_ option in Scipio. When searching with the _multiple results_ option of Scipio and the 2a gene as starting sequence, mixed genes are found for every additional gene candidate (Fig. 6). Scipio does not know about gene borders and analyses all BLAT hits according to their score. Therefore, Scipio combines the highest scoring hits to gene candidate one (2a), the next highest scoring hits to gene candidate two (2x), and so on. The third gene
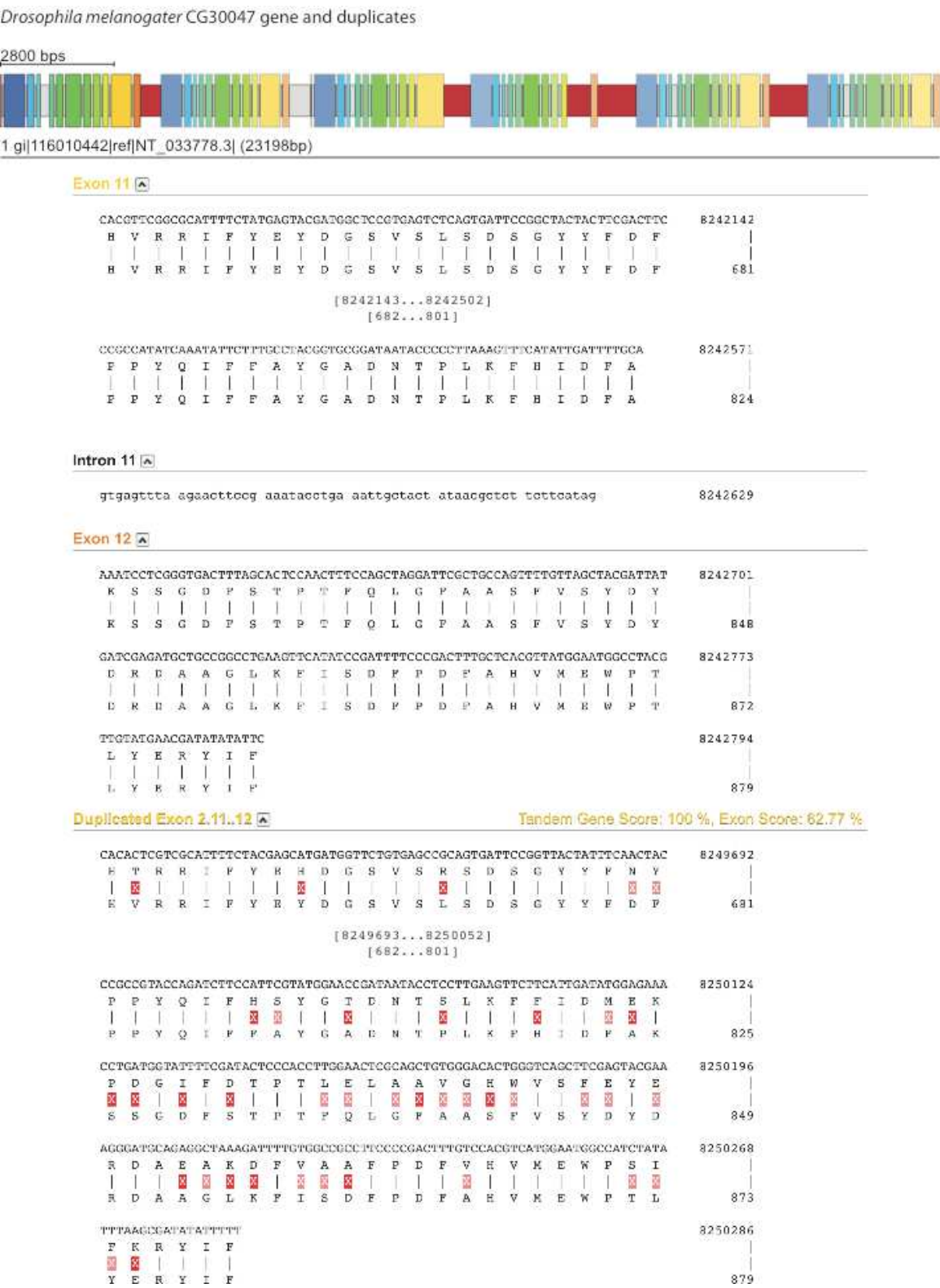
Fig. 5. *Drosophila melanogaster* CG30047: Five gene duplications were found for the CG30047 gene. In the second duplication the intron between exon 11 and 12 was lost as shown in the alignment. The alignment of exon 11 (CG30047) and the alignment of the corresponding region in the duplicated gene were shortend by amino acids 682 to 801 for representation purposes.

Homo sapiens muscle myosin heavy chain gene cluster

Genes found with new algorithm

Genes found with Scipio (*multiple results* parameter enabled)

Genes found with new algorithm (scaled)

For clarity introns have been scaled down by a factor of 9.17
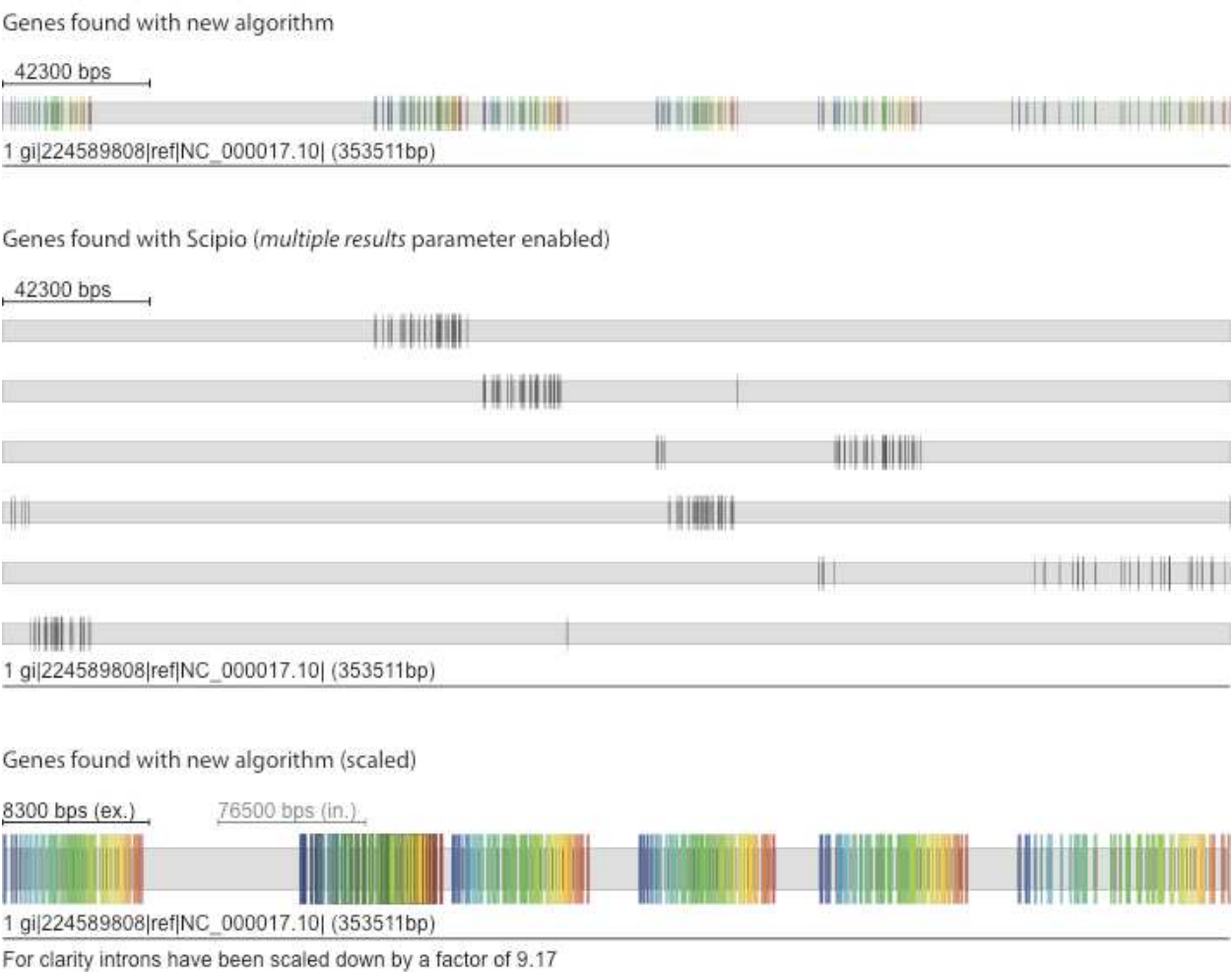
Fig. 6. *Homo sapiens* muscle myosin heavy chain gene cluster: The skeletal muscle myosin heavy chain cluster consists of the genes embryonic, 2a, 2x, 2b, perinatal, and extraocular, from left (5' end) to the right (3' end). The WebScipio search for tandem gene duplicates based on the 2a gene identifies all other genes of the cluster. The Scipio search with the parameter *multiple results* also identifies six gene candidates but only the search sequence (the 2a gene) is found correctly while the other gene candidates consist of fusions of different parts of the other muscle myosin heavy chain genes.

candidate, for example, mainly consists of the exons of the perinatal muscle myosin heavy chain gene, but the N-terminus of the 2b gene has a higher homology to the 2a gene than the N-terminus of the perinatal gene and therefore the 2b N-terminus is combined with the C-terminus of perinatal.

The nile tilapia *Oreochromis niloticus* contains another type of a muscle myosin heavy chain gene cluster (Fig. 7). Here, two genes (Mhc6 and Mhc13) are encoded on the forward strand, and Mhc7 is encoded on the reverse strand. Nevertheless, WebScipio correctly reconstructed the complete cluster when searching with the Mhc13 gene. When searching with Mhc6 or Mhc7, the small C-terminal exons of the respective other genes could not be identified. These examples demonstrate that WebScipio with the new extension is able to correctly identify arrays of very large and complex genes. For this search the minimal score for exons parameter was set to 30 % and the region size parameter to 50,000 nucleotides.

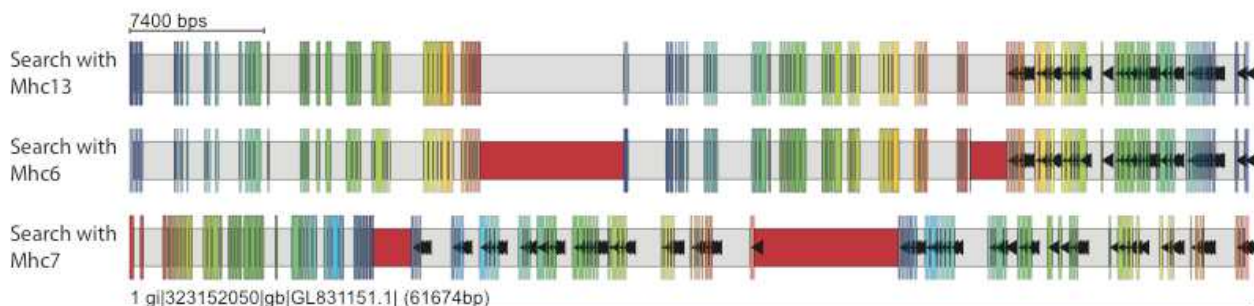*Oreochromis niloticus* myosin heavy chains 6, 7 and 13

Fig. 7. *Oreochromis niloticus* muscle myosin heavy chain gene cluster: The nile tilapia contains a cluster of three muscle myosin heavy chain genes (Mhc13, Mhc6, and Mhc7) of which Mhc7 is encoded in the opposite direction. The last exon is too divergent to be identified in most cases. Only when searching with the Mhc13 gene, the tandem genes Mhc6 and Mhc7 are reconstructed completely.

3.1.4 Revealing a pseudogene

For the *Drosophila melanogaster* CG3397 gene the first exon is splitted into two exons in the prediction of the gene duplication. For this search the default parameters were used and the option to *search for splitted exons* was enabled. The predicted gene is most probably a pseudogene, because either the predicted intron between the two splitted exons is too short to be spliced, or the exon translation results in a frame shift if both parts are considered as one potential exon. The details are shown in the alignment (Fig. 8).

3.2 Examples of non-tandemly arrayed gene duplicates
3.2.1 Duplicated gene regions

Tandemly arrayed genes evolve through unequal recombination. In this process not only single genes might be duplicated but small genomic regions containing several genes. The result would be a tandemly arrayed group of genes. Because WebScipio is searching for each gene separately it cannot separate a group of duplicated genes from a tandem array of single genes. An example for duplicated genomic regions is the region in *Drosophila melanogaster* containing genes coding for histones (Fig. 9). The new algorithm identified many duplicates for each of the His1, His2A, His2B, His3, and His4 genes in the *Drosophila* genome. As query the genes CG33825 (His1), CG33826 (His2A), CG33894 (His2B), CG33827 (His3), and CG33893 (His4) were used. The His2B and His4 genes are on the reverse strand in comparison to the other genes. The genes are very similar (some code for the same protein sequence) resulting in alignment scores between 99 % and 100 %. Only two more divergent gene duplicates were found for the His2A gene. The first two gene duplicates of His2A have alignment scores of 79 %.

3.2.2 Trans-spliced genes

Tandem gene duplicates and *trans*-spliced genes could evolve through the same gene duplication process during evolution, except that only part of the gene is duplicated instead of the complete gene. The exon-intron structure of tandem gene duplicates and *trans*-spliced genes look very similar, which complicates their differentiation during the process of gene identification. If, for example, the constitutive part of the *trans*-spliced gene consists of only

Drosophila melanogaster gene CG3397

400 bps

1 gi|56411841|ref|NT_033777.2| (3712bp)

Duplicated Exon 1.1[0...89] ⌃   Tandem Gene Score: 98 %, Exon Score: 72.16 %

```
ATGTCGGGATCACTGCCAGCTACTTTCGTCAAGGGTTTCCACGATGAGGAGAAGGTGCGTCGTATGGAGTAC   7807671
M   S   G   S   L   P   A   T   F   V   K   G   F   H   D   E   E   K   V   R   R   M   E   Y            |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |            |
M   S   G   S   L   P   A   T   F   V   K   G   F   H   D   E   E   K   V   R   R   M   E   Y            24
```

```
CGCCAACTCGGATCCACGGGTCTGCACGTGTCTAAATTAGCTATAGGTGGCTCCCCCTTTGCAATCTCTTC     7807599
R   Q   L   G   S   T   G   L   H   V   S   K   L   A   I   G   G   S   P   L   C   N   L   F            |
|   |   |   |   |   |   |   X   |   |   |   X   |   X   |   |   X   X   |   X   X   |   |                |
R   Q   L   G   S   T   G   L   R   V   S   K   I   A   L   G   G   A   T   L   S   K   L   F            48
```

```
TTTGATGACTATGATCGCGAGGAGGGCATCCTTATGGTGCAGGAGGCCATTAGATCTGGCATTAACTATATA   7807527
F   D   D   Y   D   R   E   E   G   I   L   M   V   Q   E   A   I   R   S   G   I   N   Y   I            |
X   |   |   X   |   |   |   |   |   |   |   X   |   |   |   |   |   |   |   |   |   |   |   |            |
S   D   D   F   D   R   E   E   G   I   L   T   V   Q   E   A   I   R   S   G   I   N   Y   I            72
```

```
GACACAGCTCCATTCTATCGGAGG------------------TGCTGCTTG                           7807494
D   T   A   P   F   Y   R   R   ▪ ▪ ▪ ▪ ▪ ▪ ▪   C   C   L                                           |
|   |   |   |   |   |   X   X   ▪ ▪ ▪ ▪ ▪ ▪ ▪   X   X   X                                           |
D   T   A   P   F   Y   G   Q   G   K   S   E   E   L   L   G   Q                                   89
```

Intron 4 ⌃

gcca g   7807489

Duplicated Exon 1.1[89...143] ⌃   Tandem Gene Score: 98 %, Exon Score: 88.81 %

```
GCCCTAAAGGATGTACCCCGGGAGGCCTATTATATTGCCACTAAAGTGGCGCGCTACGGGCTGGATCCGAAG   7807417
A   L   K   D   V   P   R   E   A   Y   Y   I   A   T   K   V   A   R   Y   G   L   D   P   K            |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   X   |   |   |   X                    |
A   L   K   D   V   P   R   E   A   Y   Y   I   A   T   K   V   A   R   Y   E   L   D   P   N            113
```

```
AATATGTTTGACTATTCGGCTGACAAAGCTCGGGAGAGTGTGAAGCGGAGTCTGGAGCGGCTCCAGTTGGAC   7807345
N   M   F   D   Y   S   A   D   K   A   R   E   S   V   K   R   S   L   E   R   L   Q   L   D            |
|   |   |   |   |   X   |   X   |   |   |   |   |   |   |   |   |   |   X   |   |   |   |                |
N   M   F   D   Y   T   A   A   K   A   R   E   S   V   K   R   S   L   E   L   L   Q   L   D            137
```

```
AGGGTGGACATACTACAG                                                                      7807327
R   V   D   I   L   Q                                                                               |
|   |   |   X   |   |                                                                               |
R   V   D   V   L   Q                                                                               143
```

Fig. 8. *Drosophila melanogaster* CG3397 gene: A gene duplication could be identified downstream of the CG3397 gene, which, however, most probably is a pseudogene.

one exon while the *trans*-spliced part consists of groups of similar alternative exons the correct reconstruction of the *trans*-spliced gene would not look different compared to a partial reconstruction of a cluster of duplicated genes for which the first (or last) exons were not found because of low similarity. The gene CG1637 of Drosphila is a *trans*-spliced gene (McManus et al., 2010). The WebScipio algorithm predicts tandemly arrayed genes for isoform A and B of CG1637, although the first exons of the potential tandem gene candidates were not found (Fig. 10). The close inspection of the three isoforms shows that the predicted exons do not belong to duplicated genes, but to *trans*-spliced variants of the same gene. Another type of problem is demonstrated by the dynein intermediate chain gene of *Drosophila melanogaster*. Here, the dynein intermediate chain gene is annotated as four separate genes (Sdic1, Sdic2, Sdic3 and Sdic4) in Flybase (version of June 24th, 2011). The problem is, however, that the real first two exons of the gene are not annotated in Flybase.
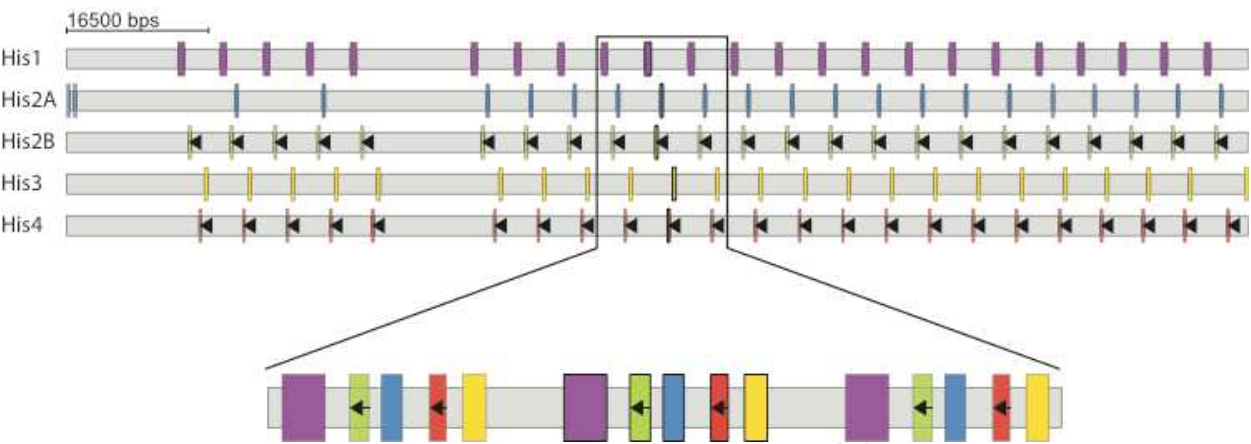
Fig. 9. *Drosophila melanogaster* histones: The results for the separate searches for gene duplicates of the histones His1, His2A, His2B, His3, and His4 are shown. Based on the results of the search for each single gene it is not possible to distinguish between a gene and a genomic region duplication. The results of all searches at the same scale shows that not single genes but a genomic region containing all five histone genes has been duplicated several times.

Fig. 10. *Drosophila melanogaster* CG1737 gene and *Drosophila melanogaster* dynein intermediate chain: The algorithm identified duplicated exons in the *trans*-spliced CG1737 and dynein intermediate chain genes. The search was done with default parameters and the *search for concatenated exons* and *search for splitted exons* options were enabled. To reveal the last and most divergent exon the *region size* parameter was set to 35,000 nucleotides and the *allowed length difference* parameter to 30 amino acids for the dynein intermediate chain gene.

The sequence encoded by the true first exons is conserved throughout all major branches of the eukaryotic tree of life that express a cytoplasmic dynein, in chromalveolates, Excavata, and Opisthokonta. In addition, this N-terminal part of the dynein intermediate chain is of high functional importance because it connects dynein to dynactin by interacting with the

dynactin p150 gene. Based on these facts and the found exon order of the genomic region, we expect the gene to be *trans*-spliced (Fig. 10, bottom).

4. Conclusion

Our algorithm provides a method to consistently predict and reconstruct tandemly arrayed gene duplicates. It has been integrated into the web interface of WebScipio allowing the search for gene duplicates of a given query protein sequence in the respective genome assemblies. WebScipio provides access to more than 2300 genome assembly files from more than 650 eukaryotes (July 2011) and is updated as soon as further genome assemblies become available whether from newer versions of already sequenced species or from newly sequenced genomes. The search results are presented in drawings coloured according to the sequence similarity of the gene duplicate to the search sequence, and in several human-readable formats like detailed alignments of the found exons to the genomic DNA. Sequences and figures can be downloaded, as well as the complete raw data for later upload or further computational analysis. The new algorithm is based on the precondition that gene duplicates rather retain the gene structure of the original gene than the sequence. We could show that the new extension to WebScipio is able to correctly predict and reconstruct gene duplicates on both the forward and the reverse strand. Also, the new algorithm is able to correctly reconstruct complicated gene structures spread over hundreds of thousands of nucleotides like the skeletal muscle myosin heavy chain gene cluster in mammals. Gene duplications often accumulate gene function destroying mutations that lead to frame shifts and in-frame stop codons. Those potential pseudogenes are identified by WebScipio but the user has to carefully inspect the results to distinguish between sequencing errors and real pseudogenes. WebScipio cannot distinguish between gene duplicates and duplications of small genomic regions that might encode several genes. Here, WebScipio can identify and reconstruct the duplicates of one gene but does not provide any hints about other genes in the intergenic regions. *Trans*-spliced genes often contain clusters of alternative exons. Those clusters will be identified by WebScipio, but again the user needs to evaluate the results to distinguish between cases of *trans*-spliced genes, where the constitutive part is encoded by just a few exons, or real gene duplications, for which some terminal exons could not be identified because of very low sequence similarity or even assembly gaps. Altogether, WebScipio provides an easy to use way to analyse the genomic region of every gene of interest for the very common event of tandem gene duplication.

5. Acknowledgments

6. References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F. et al. (2000). The genome sequence of Drosophila melanogaster, *Science*, Vol.287, No.5461, pp. 2185-2195

Aloni, R., Olender, T. & Lancet, D. (2006). Ancient genomic architecture for mammalian olfactory receptor clusters, *Genome Biol*, Vol.7, No.10, pp. R88

Babushok, D. V., Ostertag, E. M. & Kazazian, H. H., Jr. (2007). Current topics in genome evolution: molecular mechanisms of new gene formation, *Cell Mol Life Sci*, Vol.64, No.5, pp. 542-554

Bertrand, D., Lajoie, M. & El-Mabrouk, N. (2008). Inferring ancestral gene orders for a family of tandemly arrayed genes, *J Comput Biol*, Vol.15, No.8, pp. 1063-1077

Doring, A., Weese, D., Rausch, T. & Reinert, K. (2008). SeqAn an efficient, generic C++ library for sequence analysis, *BMC Bioinformatics*, Vol.9, pp. 11

Elemento, O., Gascuel, O. & Lefranc, M. P. (2002). Reconstructing the duplication history of tandemly repeated genes, *Mol Biol Evol*, Vol.19, No.3, pp. 278-288

Garcia-Fernandez, J. (2005). The genesis and evolution of homeobox gene clusters, *Nat Rev Genet*, Vol.6, No.12, pp. 881-892

Goto, N., Prins, P., Nakao, M., Bonnal, R., Aerts, J. & Katayama, T. (2010). BioRuby: Bioinformatics software for the Ruby programming language, *Bioinformatics*

Hahn, M. W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates, *J Hered*, Vol.100, No.5, pp. 605-617

Keller, O., Odronitz, F., Stanke, M., Kollmar, M. & Waack, S. (2008). Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species, *BMC Bioinformatics*, Vol.9, pp. 278

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool, *Genome Res*, Vol.12, No.4, pp. 656-664

Li, W. H., Yang, J. & Gu, X. (2005). Expression divergence between duplicate genes, *Trends Genet*, Vol.21, No.11, pp. 602-607

Long, M., Betran, E., Thornton, K. & Wang, W. (2003). The origin of new genes: glimpses from the young and old, *Nat Rev Genet*, Vol.4, No.11, pp. 865-875

Massingham, T., Davies, L. J. & Lio, P. (2001). Analysing gene function after duplication, *Bioessays*, Vol.23, No.10, pp. 873-876

McManus, C. J., Duff, M. O., Eipper-Mains, J. & Graveley, B. R. (2010). Global analysis of trans-splicing in Drosophila, *Proc Natl Acad Sci U S A*, Vol.107, No.29, pp. 12975-12979

Nei, M. & Roychoudhury, A. K. (1973). Probability of fixation and mean fixation time of an overdominant mutation, *Genetics*, Vol.74, No.2, pp. 371-380

Odronitz, F., Pillmann, H., Keller, O., Waack, S. & Kollmar, M. (2008). WebScipio: an online tool for the determination of gene structures using protein sequences, *BMC Genomics*, Vol.9, pp. 422

The Official YAML Web Site, (2011). Available from http://www.yaml.org/

Ohno, S. (1970). Evolution by Gene Duplication, Berlin, *Springer*

Prototype JavaScript framework: Easy Ajax and DOM manipulation for dynamic web applications, (2011). Available from http://www.prototypejs.org

purzelrakete's workling at master - GitHub, (2011). Available from http://github.com/purzelrakete/workling

Quijano, C., Tomancak, P., Lopez-Marti, J., Suyama, M., Bork, P., Milan, M., Torrents, D. & Manzanares, M. (2008). Selective maintenance of Drosophila tandemly arrayed duplicated genes during evolution, *Genome Biol*, Vol.9, No.12, pp. R176

Ruby on Rails, (2011). Available from http://rubyonrails.org

Ruby Programming Language, (2011). Available from http://www.ruby-lang.org/

Saez, L. J., Gianola, K. M., McNally, E. M., Feghali, R., Eddy, R., Shows, T. B. & Leinwand, L. A. (1987). Human cardiac myosin heavy chain genes and their linkage in the genome, *Nucleic Acids Res*, Vol.15, No.13, pp. 5443-5459

script.aculo.us - web 2.0 javascript, (2011). Available from http://script.aculo.us

Shoja, V. & Zhang, L. (2006). A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat, *Mol Biol Evol*, Vol.23, No.11, pp. 2134-2141

Sun, Y. M., Da Costa, N. & Chang, K. C. (2003). Cluster characterisation and temporal expression of porcine sarcomeric myosin heavy chain genes, *J Muscle Res Cell Motil*, Vol.24, No.8, pp. 561-570

Tokyo Cabinet: a modern implementation of DBM, (2011). Available from http://fallabs.com/tokyocabinet/

tra's spawn at master - GitHub, (2011). Available from http://github.com/tra/spawn

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R. et al. (2009). FlyBase: enhancing Drosophila Gene Ontology annotations, *Nucleic Acids Res*, Vol.37, Database issue, pp. D555-559

W3C SVG Working Group, (2011). Available from http://www.w3.org/Graphics/SVG/

Weydert, A., Daubas, P., Lazaridis, I., Barton, P., Garner, I., Leader, D. P., Bonhomme, F., Catalan, J., Simon, D., Guenet, J. L. et al. (1985). Genes for skeletal muscle myosin heavy chains are clustered and are not located on the same mouse chromosome as a cardiac myosin heavy chain gene, *Proc Natl Acad Sci U S A*, Vol.82, No.21, pp. 7183-7187

Zhang, J. (2003). Evolution by gene duplication: an update, *Trends Ecol Evol*, Vol.18, pp. 292-298

Zhang, J. & Nei, M. (1996). Evolution of Antennapedia-class homeobox genes, *Genetics*, Vol.142, No.1, pp. 295-303

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S. & Wang, W. (2008). On the origin of new genes in Drosophila, *Genome Res*, Vol.18, No.9, pp. 1446-1455
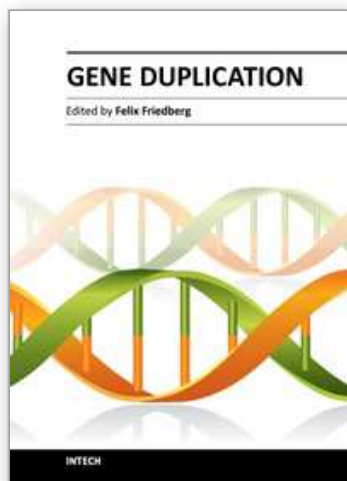
Gene Duplication

Edited by Prof. Felix Friedberg

The book Gene Duplication consists of 21 chapters divided in 3 parts: General Aspects, A Look at Some Gene Families and Examining Bundles of Genes. The importance of the study of Gene Duplication stems from the realization that the dynamic process of duplication is the "sine qua non" underlying the evolution of all living matter. Genes may be altered before or after the duplication process thereby undergoing neofunctionalization, thus creating in time new organisms which populate the Earth.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

INTECH
open science | open minds