TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

AUSTRIAN INSTITUTE OF TECHNOLOGY

M A S T E R T H E S I S

A Bioinformatic Approach for Detecting Genetic Mutations leading to Multidrug Resistance in Fasciola Hepatica

submitted to the Institute of Chemical Engineering and Analysis at the
Vienna University of Technology in cooperation with the
Austrian Institute of Technology

under supervision of

Ao.Univ.Prof. Mag.rer.nat. Dr.rer.nat Johann LOHNINGER

by

Nina HOFER
Liechtensteinstraße 11
2193 Wilfersdorf, Austria

Vienna, 14^{th} November 2015

.....
Nina Hofer

Kurzfassung

Die Bioinformatik ist eine Wissenschaft, die Modelle, Techniken und Methoden der Informatik in spezifischen Fachgebieten der Biologie, wie Genetik und Molekularbiologie anwendet. Die enormen Fortschritte auf dem Gebiet der Gensequenzierung führen zum stetigen Anwachsen biologischer Datenbanken. Der Einsatz der Bioinformatik ist somit unerlässlich um Sequenzen zu vergleichen und neue Informationen aus den Datenmengen zu gewinnen. In der Regel werden neue Sequenzen mit Datenbanken verglichen, die aus Sequenzen bekannter Struktur und Funktionen bestehen. Bei relativer Ähnlichkeit der Proteine oder Nukleotide zueinander können Funktionsinformationen transferiert werden. Der Sequenzvergleich findet auf der Ebene einzelner Basen oder Aminosäuren statt und ermöglicht so die Identifizierung genetischer Veränderungen. [90]
Heutzutage werden zur Datenbanksuche vor allem heuristische Suchalgorithmen verwendet, da diese weit weniger rechenintensiv sind. BLAST ist der am weitesten verbreitete Algorithmus dieser Kategorie und stellt eine Annäherung an die genaue Berechnung von Sequenzalignments mit dem Smith-Watermann und Needleman-Wunsch Algorithmus dar. Zunächst werden in einer schnellen Indexsuche Abschnitte in der Sequenz bestimmt, die Ähnlichkeiten aufweisen. Diese Bereiche werden dann mit Hilfe einer Substitutionsmatrix sensitiv untersucht und die lokalen Alignments berechnet. Ein Alignment ist eine Zuordnung von zwei oder mehreren Sequenzen, die es ermöglicht identische und ähnliche Positionen zu identifizieren. [96]
Die einfachste Möglichkeit diese Ähnlichkeiten grafisch darzustellen bietet bei einem paarweisen Alignment der Dotplot. Hierbei werden die Sequenzen auf jeweils einer Achse aufgetragen und identische Positionen mit Punkten markiert. Bei absoluter Identität der Sequenzen führt dies zum Ausbilden einer Diagonalen. Die Dotplot-Methode ist jedoch in vielerlei Hinsicht limitiert. Sie stellt kein Alignment dar und beschränkt sich auf den Vergleich zweier Sequenzen. Weiters sind die Plots oft stark verrauscht und erst die Anwendung verschiedener Filter und Schwellenwerte ermöglicht die Extraktion der gewünschten Informationen.
Werden mehr als 2 Sequenzen verglichen, spricht man von einem multiplen Alignment, welches ein zentrales Hilfsmittel z.B. zur Analyse von Verwandtschaftsverhältnissen, zur Konstruktion von Stammbäumen, aber auch zur Analyse von 3D-Strukturen und zur Genomanalyse, darstellt. [77] Im Laufe der Zeit haben sich viele Multiple Alignment Hilfsprogramme entwickelt, die sich vor allem bezüglich ihrer Genauigkeit und Laufzeit unterscheiden. [90]

Im Rahmen dieser Diplomarbeit wurden cDNA contigs von *Fasciola hepatica* mithilfe der Alignment-Programme BLAST und Geneious untersucht. Zur Verfügung stehen drei Datensätze, welche genetisches Material von Medikamenten-sensitiven Leberegeln enthalten und 1 Datensatz bestehend aus Medikamenten-resistentem Material.
Bereits durchgeführte Studien legen nahe, dass Membranproteine (ABC Transporter) als Effluxpumpen agieren und zu einer verminderten Aufnahme des Medikaments führen. Beim Menschen konnte nachgewiesen werden, dass 14 der 48 ABC Transporter an Erbkrankheiten beteiligt sind. Es wird vermutet, dass Mutationen die Funktionsweise der Pumpen verändern und die Entwicklung des Organismus zur Medikamentenresistenz unterstützen. [23]
Zu Beginn wurden die Leberegel Datensätze mit einer Datenbank, bestehend aus ABC Transporter Sequenzen von *Homo sapiens*, *Drosophila melanogaster* und *Caenorhabditis elegans* mithilfe von BLAST, verglichen. Dies ermöglichte die Identifikation der Anzahl von ABC Transportern, welche in *Fasciola hepatica* verfügbar sind. Weiters konnte auch die Zuteilung zu den einzelnen Unterkategorien der Transporter vorgenommen werden, mit dem Resultat, dass alle 3 bekannten Multi Drug Resistance (MDR) ABC Transporter im Leberegel vertreten sind. Zur Visualisierung und zum Vergleich der sensitiven mit den resistenten ABC Transporter-Sequenzen des Leberegels wurden diese in Geneious importiert und multiple Alignments durchgeführt. Dadurch konnten Punktmutationen identifiziert werden, welche möglicherweise wertvolle Informationen über die Adaptionsmechanismen bei der Entwicklung von MDR liefern. Zusätzlich wurden einige funktionell wichtige Proteinmotive untersucht und teilweise stark degeneriert vorgefunden.

Abstract

This master thesis deals with a bioinformatic approach for detecting genetic mutations leading to multidrug resistance in the trematode *Fasciola hepatica*. Enhanced anthelmintic drug efflux by ABC transporters has been assumed to be involved in developing resistance [109]. Therefore, cDNA contigs of ABC transporters of drug sensitive versus resistant individuals were analyzed to identify functional significant mutations. The Basic Local Alignment Tool (BLAST) was used to compare the *Fasciola hepatica* contigs with a database consisting of genetic material of *Homo sapiens*, *Drosophila melanogaster* and *Caenorhabditis elegans*. Therefore, the liver fluke contigs could be characterized and uniquely classified to the single ABC transporter subfamilies. In addition, sequence alignments were conducted by the use of Geneious to organize, visualize and compare the classified contigs. Several point mutations could be detected probably leading to functional changes of the transporters. Furthermore, biological motifs were identified and found to be highly degenerate. These results may provide new insights into the adaption mechanisms the parasites have developed to survive drug exposure.

Acknowledgements

This thesis owes its existence to the help, support and inspiration of several people.

First of all, I would like to express my sincere appreciation and gratitude to my supervisor Prof. Lohninger for his support during my research. Thanks for all the helpful discussions, advices and knowledge he passed on to me and also for the correction of my thesis.
My sincere thanks goes to Silvia Fluch, who provided me the opportunity to be part of this project and to gain new experiences and knowledge.
I'm also thankful to Elisabeth Witschnitzky for her guidance and close collaboration.

Finally, my deepest gratitude goes to my family for their understanding and unconditional support throughout my studies and my whole life. Their encouragement and dedication enabled me to achieve my aims and provide the foundation for this work.

Contents

1. Introduction

Bioinformatics has gradually become an important area of science using computational approaches to answer biological questions. With the enormous increase of biological data and structural information available, the highly multidisciplinary subject recruits not only biologists but also mathematicians, physicists and computer scientists.
The ultimate goal is to predict structure and further function from gene sequences usually by applying comparative analyses. New sequences are typically compared with libraries or databases of sequences with known functional properties to annotate the unknown genes. [13] The similarities and differences are analyzed at the level of individual bases or amino acids to detect genetic differences (chapter 2) [81].

For database searching two main heuristic approaches to identify homologs have been deployed. The FASTA suite of programs and the BLAST programs, which are described in chapter 3 [13]. Basically, an input sequence is continuously aligned to each subject sequence in the database to detect significant hits followed by individual sequence alignments [81].
The dot plot, a two dimensional similarity matrix can be used as simple graphical representation of two sequences (section 4.1). The matrix cells are shaded black, if residues are identical resulting in diagonal lines. Insertions and deletions disrupt the diagonal and regions of local similarity or repetitive sequences give rise to additional diagonal matches. [13] To reduce the chance of random hits a single pathway with most biological significance is searched through the dot plot. Therefore, most likely equivalent residues are determined and scored concerning their similarity.
This lead to the introduction of scoring matrices and to the more sophisticated residue substitution matrices PAM and BLOSUM for proteins, explained in section 4.3.
For nucleotides also different constraints on substitution rates can be defined leading to various nucleotide substitution models specified in section 4.2. [13] [81]

Generally, alignment methods attempt to determine the optimal alignment between sequences by modeling the mutational process that has given rise [11]. The algorithm tries to match the maximum number of identical or similar residue pairs, whereas it tolerates the minimum number of insertions or deletions in the sequences [13]. Needleman-Wunsch and Smith-Waterman originally developed sequence alignment algorithms based on dynamic programming, which result in long calculation periods but yield the most reliable alignments of protein sequences (section 4.4). Nevertheless, they become replaced by heuristic algorithms due to reduced execution time and acceptable alignment results. [87]

Concerning the number of aligned sequences it can be distinguished between pairwise and multiple sequence alignment. An alignment of more than two sequences constitutes a multiple alignment widely used in many different areas in bioinformatics including homology searches, genomic annotation, structure prediction and functional genomics [87]. In addition, it enables to search for patterns of highly conserved residues and to detect functional important motifs and their variations (section 5.4.1) [13].
Dynamic programming cannot easily be extended for applying it to multiple sequence alignments and therefore progressive alignments have to be performed [11]. Diverse approaches use slightly different methods and vary in accuracy and speed, resulting in several multiple sequence alignment tools available on the market [13]. A short overview of their algorithms is given in section 4.5.

This master thesis has been carried out at the Austrian Institute of Technology within the framework of the *Fasciola hepatica* project.
The helminth parasite (chapter 5) affects livestock and humans leading to the disease fascioliasis [9]. According to WHO, at least 2.4 million people are infected in more than 70 countries worldwide. Fascioliasis is currently the most widespread disease known in terms of latitude, longitude and altitude [72]. The economic losses due to *Fasciola hepatica* including veterinary costs and production losses exceed 3 billion $ globally [12] [83]. In consideration that more than 55% of all farm animal diseases are caused by parasitic helminths, it is estimated that every year 400 million € are spend on drugs [76]. In the absence of effective vaccines against fascioliasis, therapeutic drugs are the mainstays of prevention and control [52]. The benzimidazole triclabendazole (TCBZ) is the only effective known pharmacological treatment and highly effective against the immature and mature trematode [41]. The complete reliance on TCBZ and the continuous use result in adaption of the parasites developing drug resistance (section 5.2) [22]. The first case was reported in Australia in 1983 followed by incidences worldwide [43].
Previous studies suggest that genomic mutations in drug efflux pumps may lead to a decreased uptake, increased efflux or metabolic change [109]. In humans, it could be proven that 14 out of 48 of these pumps, encoding ABC transporter are involved in hereditary diseases [23].

With this biological motivation, cDNA contigs of putative ABC transporters of drug sensitive and resistant parasitic helminths were analyzed.
BLAST was used to compare the genetic information of the liver flukes to ABC transporters of *Homo sapiens*, *Drosophila melanogaster* and *Caenorhabditis elegans* to detect similarities. Therefore, the contigs could be characterized and uniquely classified to the single ABC transporter subfamilies and their subgroups. From this, a first estimation of the amount of ABC transporters available in the trematode could be received. Furthermore, all ABC transporter efflux pumps, known to be involved in multidrug resistance could be detected in *Fasciola hepatica*.

To organize and compare the identified ABC transporter contigs of the drug senstive to the resistant individuals they get arranged with the multiple sequence alignment program Geneious. Single nucleotide polymorphisms, which distinguish the resistant cDNA dataset from the sensitive ones were searched. This may offer valuable information about the adaption changes involved in developing drug resistance.
Moreover, important sequence motifs were investigated in Geneious to detect functional significant variations.

2. Gene Mutation

Gene mutations are defined as alteration in the nucleotide sequence in DNA. The permanent change can affect a single nucleotide pair or larger gene segments of a chromosome. Mutations cause alterations in the genetic code, which lead to genetic variation and the potential to develop diseases.
The consequence of a mutation may be a protein change, which further can influence the function of the protein. Further, they may cause variations of the gene, called *allele*, which basically carry out the same function in the cell.
Humans and nearly all mammals have two copies of each gene and therefore two alleles of each chromosome, which is called a diploid organism.
Alleles can be classified in heterozygous, homozygous or rather codominance. In the first case the dominant alleles show their effect even if the individual only has one copy of the allele. One dominant "green eye" allele and a single copy can determine the green eye color of a human.
In the second case, recessive alleles only show their effect if the individual has two copies of the allele. If both alleles are dominant it is called codominance, there both alleles are expressed equally. For example, the blood group AB results of codominance of the A and B dominant alleles.
Allele is a short form of allelomorph ("other form"), which was used in early days of genetics to describe variant forms of a gene detected as different phenotype. Today alleles are understood to be alternative DNA sequences at the same physical gene locus, which may or may not result in different phenotypic traits. In any particular diploid organism the genotype for each gene comprises the pair of alleles present at that locus, which are the same in homozygotes and different in heterozygotes. A population or species of organisms typically includes multiple alleles at each locus among various individuals. Allelic variation at a locus is measurable as the number of alleles (polymorphism) present or the proportion of heterozygotes in the population. [108]

The most common type of genetic variation is the single nucleotide polymorphisms (SNPs) occurring when a single nucleotide in the genome differs between members of a species or between paired chromosomes of an individual. SNPs are point mutations including nucleotide substitutions, insertions or deletions occurring in DNA with a 0.1% frequency. The mutational process thought to be governing by the evolution of SNPs is that of nucleotide substitution. [74] 99,9% of the DNA sequences of all human are identical. 80% of the remaining 0,1% correspond to SNPs [2]. This type of mutation may alter the reading frame of the gene and can have a profound impact on the individuals in which they are present. The sickle cell disease is caused by a point mutation of an

adenine A for a thymine T, that changes a hydrophilic amino acid glutamic acid to a hydrophobic amino acid valine. The crohn's disease is the consequence of an insertion of cytosine C causing frameshift.
The nonsense mutation alters the nucleotide sequence in such a way that a stop codon (UAA UAG UGA) is coded in place of an amino acid causing an incomplete protein. For example, cystic fibrosis is caused by a non-synonymous nonsense mutation. The change of guanine G for thymine T (GGA $\rightarrow$ TGA) changes the protein glycine (Gly) to a STOP codon. [5] [104]
The 20 amino acids found in proteins are built up by three nucleotide sets called codons and because most amino acids have multiple codons, a number of possible DNA sequences might represent the same protein sequence. Although a change in the DNA sequence occurs, silent point mutations do not change the protein [16].

2.1. The Cause of Mutations

Mutations arise spontaneously and often occur naturally due to errors during the DNA replication. Replication errors can result from failure of three separate processes, namely base selection, proof-reading and DNA mismatch repair (MMR), which act sequentially to ensure the fidelity of replication. The first two processes allow DNA replication to proceed with a fidelity of 10^{-7} per bp replicated. The final step, MMR, recognizes DNA base mispairs and initiates a DNA repair cascade, contributing to genomic fidelity and yielding a final error rate of 10^{-10} per bp.
Natural exposure of an organism to certain environmental factors, such as ultraviolet light, chemical carcinogens or ionizing radiation can also cause mutations. [111]
On the one hand, mutations can be beneficial and lead to an evolutionary advantage of a certain genotype. Otherwise, mutations may lead to changes in the structure of an encoded protein or to a decrease or complete loss in its expression. [66]

2.1.1. Multidrug Resistance

The emergence of mutations in nucleic acids is one of the major factors underlying evolution, providing the working material for natural selection. Also pathogens tend to adopt through mutations various mechanisms to survive unfavorable conditions and enable these organisms to develop drug resistance.
Resistance to drugs is a heritable increase in the frequency of individuals in a population able to tolerate doses of a compound following exposure to the drug. [21] Microorganisms have evolved a multitude of mechanisms to overcome the effectiveness of drugs, thereby surviving exposure to the drug. The resistance among various microbial species to for example different antimicrobial drugs has emerged as a cause of public health threat all over the world at a terrifying rate.
Antimicrobial resistance associated with high mortality rates and high medical costs has

a significant impact on the effectiveness of antimicrobial agents. Expansion of global trade and tourism lead to increased potential of multidrug resistance to spread all over the world. [100]
In former studies different mechanism (Figure 2.1) could be found the parasites have developed to protect themselves against drugs. The authors focused on identification of single nucleotide polymorphisms in genes, which were predicted either to be drug targets or to be involved in the uptake, metabolism or efflux of drugs. [26] [100]
Antimicrobials bind to the cell wall inhibiting its synthesis and blocking the cell growth and division. Thus, alteration in the cell membrane composition may lead to a decreased permeability and uptake of the drug into the cell. In addition, these changes result in a lack of active target sites for the drugs to bind. Another multidrug resistance (MDR) mechanism constitutes an overexpression of drug target enzymes leading to target bypass and the production of alternate target molecules affecting the access of drugs to the target sites. Also enzymatic degradation or inactivation of antimicrobials may play a role in MDR. However, MDR mediated by drug efflux pumps remains the predominant mechanism and represents the focus in our study. The overexpression of ATP - binding cassette (ABC) transporter membrane proteins (section 5.2.1), known as multidrug efflux pumps, generates MDR and continues cellular functions without any interference. MDR proteins affect the fluidity and permeability leading to an ATP dependent efflux of the drug and therefore to a decrease of the intracellular concentration.
This mechanism is also known from cancer cells, which limits the long-term use of chemotherapy. [100]

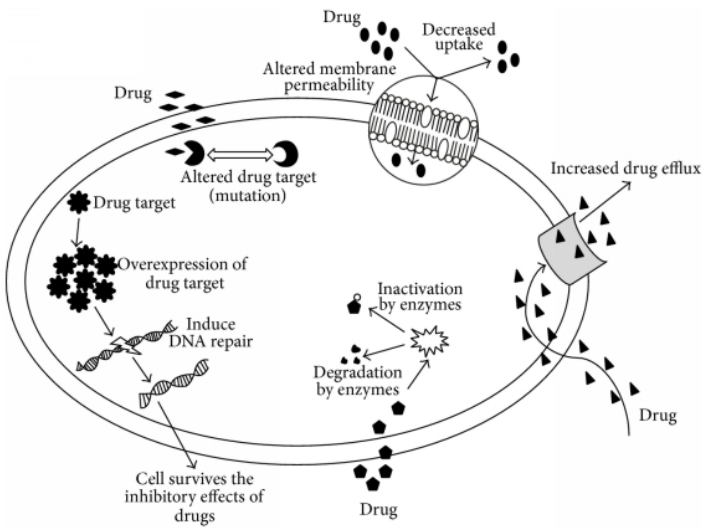


Figure 2.1.: MDR mechanisms: an alteration in the cell membrane composition leads to a decreased drug uptake or to drug target changes. Enzymes and efflux pumps cause a decreased drug concentration within the cell. Mutations in target genes may retain cellular function by reducing susceptibility to inhibition. [100]

3. Basic Local Alignment Search Tool

The Basic Local Alignment Search Tool[1] (BLAST) was developed by Altschul et al. in 1990 and has become one of the most popular local alignment tools (section 4.4) in biology. It is mainly applied for database searches to compare sequences and to identify homology. [81] BLAST is a heuristic method, which rapidly aligns a query DNA or nucleotide sequence to a library or database of sequences.
The word based algorithm identifies short segments of high similarity. This simplified approach is 10 to 50 times faster than applying the standard dynamic programming algorithm of Smith-Waterman (section 4.4). [11]
BLAST finds statistically significant similarities between sequences by evaluating alignments [14]. This is accomplished by integrating gap costs (section 3.2.2) and scoring matrices (section 4.2 and 4.3). [13]

3.1. BLAST Algorithm

The BLAST algorithm tries to find a short fragment of the query sequence (input sequence) that aligns perfectly with a fragment of the subject sequence found in the database. Then the alignment is extended in both directions until the score is at least equal to the cutoff score threshold S (section 3.2). Therefore the algorithm attempts to find short lengths of exact matches (Figure 3.1). [81]

Low complexity regions might produce high scores and confuse the program. They are filtered out and marked with an X for protein sequences or N for nucleic acid sequences. Primarily, the BLAST algorithm divide the query sequence into words of length w. For proteins w = 3 and w = 11 for DNA. [46]

[1] http://blast.ncbi.nlm.nih.gov/Blast.cgi

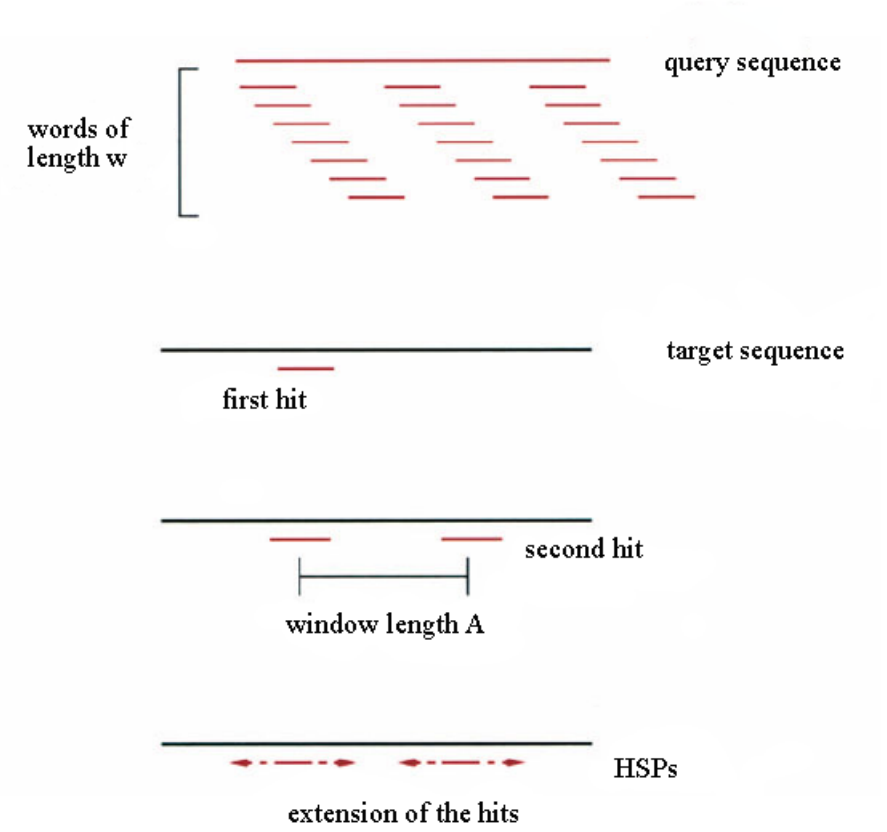


Figure 3.1.: BLAST search algorithm: two hit method. (cp. [46])
1) Listing of all words of length w in the query sequence.
2) Scanning the target sequence for matches (hits).
3) Detection of a second hit within a threshold distance A.
4) Bidirectional extension.

The sequences of the database are searched with the index of the length w. The words whose scores are greater than a certain threshold will remain in the possible matching words list, while those with lower scores will be discarded. Therefore not only identical but also similar positions are recognized by BLAST depending on the substitution matrix used. Only if a second hit can be found on the same diagonal in the dot plot (section 4.1), both hits are taken into account in the further procedure of the search. The maximum distance between the two hits are limited by the window length A.
Both hits are extended bidirectionally until the score stops to increase. Hits with a score above the cutoff score threshold are called *High Scoring Pairs* (HSPs). In the old version of BLAST (Atschul et al., 1990) no gaps were allowed at this step. The newer version (Atschul et al., 1997) allows gaps and thus linking the HSPs. [46]

3.2. Significance of BLAST Hits

3.2.1. Scores

The score of the HSPs consists of the sum of the individual scores. It depends on the substitution matrix and the amount of the penalty points for the formation and the extension of the gaps (section 3.2.2). To normalize the scores they are converted into the bit score

$$S' = \frac{\lambda S - lnK}{ln2} \tag{3.1}$$

where S is the score and S' the bit score. The bit score enables to compare the alignments of different calculations, because it is independent of the substitution matrix. The better the quality of the alignments the higher is the bit score. [14]
However, the magnitude of the bit score says less about the statistical significance of the alignment. It could also be just a random hit. Therefore, the expectation value E is calculated with the sequence length of the search sequence m and the sum of the sequence length of all comparative sequences n.

$$E = Kmn \cdot e^{-\lambda S} \sim mn \cdot 2^{-S'} \tag{3.2}$$

K is the parameter adjusting for the search space size, and λ is the scaling parameter for the scoring system. The expectation value E represents the number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The smaller the E value, the more significant the score.
Doubling the length of either sequence should double the number of HSPs attaining a given score. Therefore, the expectation value E is a function of the score (respectively of the bit score) and the database size.
As rule of thumb, E values below 10^{-6} are most probably statistically significant. Values between 10^{-6} and 10^{-2} absolutely deserve a second look and values between 10^{-2} and 1 do not indicate a good homology.
Always consider that BLAST is e heuristic method for local alignment and the significance of a hit is highly dependent on the size of the alignments and the size of the sequence database. [14]

The chance of finding zero HSPs with score >= S is e^{-E}, hence the probability of finding at least one such HSP is $1 - e^{-E}$, which is given by the p value. p values of $p < 0.05$ are usually considered statistically significant.
This parameter is not listed in the BLAST output and therefore the user has to consider the E value. However, if $E < 0,01$, p and E values are quite similar. [80]

3.2.2. Gap Penalties

Gap penalties including penalizing for insertions and deletions indicate another important part of the scoring process. Two completely different sequences can always be aligned without any mismatches by insertions of gaps. Using gaps in such an unconstrained manner leads to excessive gaps of little biological meaning. [11] Insertions and deletions should be assumed as rare events and thus the use of gaps is penalized in alignments. Various mechanisms have been developed for introducing gap penalties. Some algorithms use a length-independent penalty, whereas others define a fixed penalty for opening a gap, which increases according to the length of the insertion. [81]
Increasing the gap penalty will improve the statistical significance of shorter, or closely related alignments. Excessive penalties will enforce a gapless alignment, which would be biologically inaccurate.
Although the recommended combinations of scoring matrices (section 4.3) and gap penalties have been described in the literature, there is no formal theory available how gap penalties should be chosen. Therefore, they have to be set empirically. [87]

3.3. BLAST Output

BLAST uses statistical methods to compare a nucleotide or protein query sequence to a database of sequences. The algorithm calculates similarity scores for local alignments between the query sequence and the subject sequences using specific scoring matrices and returns a table of the best matches (hits) from the database. The sequences found are sorted by statistical significance given as E - value (Figure 3.3).
An accepted input format of BLAST is e.g. FASTA, which is characterized by a definition line, beginning with a ”>” symbol usually containing identifiers and descriptive information.

BLAST takes a large number of parameters that influence the way BLAST performs its search and formats its output. The most efficient way to conduct a BLAST search with your own database is to download the software and conduct it via the command line. BLAST is available on the NCBI homepage[2].

First of all, depending on the database the right BLAST subprogram has to be chosen. The *tblastx* program, which was used in Figure 3.3 and 3.2, compares a DNA translated (six frame translation) into protein with a DNA database translated into protein. If it is not desired to translate the sequences into proteins *blastn* can be used. To find similarities between a nucleotide query sequence and a protein database *blastx* has to be chosen. *blastp* performs a protein against protein search and *tblastn* compares a protein query against a nucleotide sequence database dynamically translated in all six reading frames.

[2]http://www.ncbi.nlm.nih.gov/

All subprograms offered by BLAST are summed up and listed in Table 3.1. [34]

Program	**Query Seq. Type**	**Database Seq. Type**	**Alignment Level**
blastn	nucleotide	nucleotide	nucleotide
blastp	protein	protein	protein
blastx	nucleotide	protein	protein
tblastn	protein	nucleotide	protein
tblastx	nucleotide	nucleotide	protein

Table 3.1.: BLAST subprograms [80]

It is important to realize that there is no set cutoff, which determines whether a match is considered significant or "similar enough" - this has to be set by the user. To find extremely similar sequences in closely related species the cut off value has to be relatively small, e.g. $1e^{-50}$ or even $1e^{-100}$.
In the example of use (chapter 5) the ABC transporter of *Fasciola hepatica* were compared to the ABC transporter of *Homo sapiens*, *Drosophila melanogaster* and *Caenorhabditis elegans*. [39] Especially the human but also the fruit fly and the nematode are evolutionarily very distant from *Fasciola hepatica*. Therefore relatively high cutoff values of about $1e^{-20}$ were used to detect similarities between the sequences.

After specifying the BLAST subprogram, database, query sequence and the cut off value, the alignment view options can be chosen from eleven possibilities.
The tabular alignment view represents a clear listing with additional parameters such as sequence identity [%], alignment length [*bp*], mismatches and gap openings.
The first column gives the identifiers and the description for the query sequences, which produced a significant hit with the subject sequences found in column two. In Figure 3.2 a *tblastx* search was performed comparing the *Fasciola hepatica* contigs of dataset 19929 (subject sequence) to the ABC transporter sequences of *Homo sapiens* (query sequence). The alignment view (Figure 3.3) shows the complete alignment for each hit. The alignment line between the query and the subject sequence indicates identities by inserting the identical amino acid or nucleotide, mismatches by a gap and similarities are marked with a plus (see Figure 3.3).

```
# TBLASTX 2.2.22 [Sep-27-2009]
# Query: gi|4128032|emb|AJ012376.1| Homo sapiens mRNA for ATP-binding cassette transporter-1 (ABC-1)
# Database: 19929_paired_assembly.Trinity.fasta.min5reads
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
gi|4128032|emb|AJ012376.1|      comp24099_c1_seq1       50.45   222     110     0       5713    6378    791     1456    3e-78    273
gi|4128032|emb|AJ012376.1|      comp24099_c1_seq1       51.52   33      16      0       6505    6603    1715    1813    3e-78   84.5
gi|4128032|emb|AJ012376.1|      comp24099_c1_seq1       36.19   105     67      0       2671    2985    788     1102    3e-31   84.5
gi|4128032|emb|AJ012376.1|      comp24099_c1_seq1       42.31   52      30      0       3175    3330    1292    1447    3e-31   83.1
gi|4128032|emb|AJ012376.1|      comp24099_c1_seq1       32.61   46      31      0       3059    3196    1173    1310    3e-31   83.1
gi|4128032|emb|AJ012376.1|      comp24099_c1_seq1       45.35   86      47      0       6353    6096    1431    1174    2e-30   83.1
gi|4128032|emb|AJ012376.1|      comp24099_c1_seq1       47.25   91      48      0       5984    5712    1062    790     2e-30   71.2
gi|4128032|emb|AJ012376.1|      comp30310_c0_seq7       44.64   224     124     0       5728    6399    23      694     3e-61    236
gi|4128032|emb|AJ012376.1|      comp30403_c3_seq1       57.81   128     54      0       5977    6360    698     315     2e-55    194
gi|4128032|emb|AJ012376.1|      comp30403_c3_seq1       60.61   33      13      0       6502    6600    176     78      2e-55   48.7
gi|4128032|emb|AJ012376.1|      comp30493_c0_seq2       37.14   105     66      0       5902    6216    551     237     2e-29   98.7
gi|4128032|emb|AJ012376.1|      comp30493_c0_seq2       32.69   52      35      0       6223    6378    227     72      2e-29   46.0
gi|4128032|emb|AJ012376.1|      comp30493_c0_seq2       33.33   48      32      0       2677    2820    749     606     2e-29   35.0
gi|4128032|emb|AJ012376.1|      comp12072_c0_seq1       49.48   97      49      0       5665    5955    27      317     2e-28    127
# TBLASTX 2.2.22 [Sep-27-2009]
# Query: gi|609355|gb|U18235.1|HSU18235 Human ATP-binding cassette protein (ABC2) mRNA HFBCD04 clone, partial cds
# Database: 19929_paired_assembly.Trinity.fasta.min5reads
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
# TBLASTX 2.2.22 [Sep-27-2009]
# Query: gi|1699037|gb|U78735.1|HSU78735 Human ABC3 mRNA, complete cds
# Database: 19929_paired_assembly.Trinity.fasta.min5reads
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
gi|1699037|gb|U78735.1|HSU78735 comp24099_c1_seq1       51.36   220     107     0       4751    5410    800     1459    2e-90    281
gi|1699037|gb|U78735.1|HSU78735 comp24099_c1_seq1       48.78   41      21      0       5525    5647    1700    1822    2e-90    144
gi|1699037|gb|U78735.1|HSU78735 comp24099_c1_seq1       29.09   55      39      0       4121    4285    50      214     2e-90    144
gi|1699037|gb|U78735.1|HSU78735 comp24099_c1_seq1       37.50   32      20      0       5423    5518    1589    1684    2e-90    144
gi|1699037|gb|U78735.1|HSU78735 comp24099_c1_seq1       31.03   29      20      0       4295    4381    227     313     2e-90    144
gi|1699037|gb|U78735.1|HSU78735 comp24099_c1_seq1       39.77   171     103     0       2201    2713    800     1312    3e-35    144
gi|1699037|gb|U78735.1|HSU78735 comp24099_c1_seq1       68.75   16      5       0       5274    5321    1323    1370    3e-35   85.4
gi|1699037|gb|U78735.1|HSU78735 comp24099_c1_seq1       51.14   88      43      0       5013    4750    1062    799     1e-31   85.4
gi|1699037|gb|U78735.1|HSU78735 comp24099_c1_seq1       45.00   80      44      0       5358    5119    1407    1168    1e-31   72.5
gi|1699037|gb|U78735.1|HSU78735 comp30310_c0_seq7       44.32   264     147     0       4736    5527    2       793     3e-70    266
gi|1699037|gb|U78735.1|HSU78735 comp30310_c0_seq7       40.13   157     94      0       2210    2680    26      496     6e-43    145
```

Figure 3.2.: *Tblastx* search output with a cutoff value of $1e^{-20}$ and a tabular alignment view. In the first column the query sequences (ABC transporter of *Homo sapiens)* are listed to the subject sequences (*Fasciola hepatica* sequences of dataset 19929), which produced significant hits. In the same line further parameters are listed, giving information about the quality of the alignment and the significance of the hits.

```
                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value  N

comp24099_c1_seq1 len=1944 path=[2469:0-1565 4035:1566-1669 ...   273   3e-78  3
comp30310_c0_seq7 len=1176 path=[1:0-594 596:595-598 600:599...   236   3e-61  1
comp30403_c3_seq1 len=714 path=[1620:0-152 6556:153-157 1778...   194   2e-55  2
comp30493_c0_seq2 len=1020 path=[358:0-54 413:55-484 843:485...    99   2e-29  3
comp12072_c0_seq1 len=331 path=[357:0-330]                        127   2e-28  1

>comp24099_c1_seq1 len=1944 path=[2469:0-1565 4035:1566-1669
            5481:1670-1768 4238:1769-1943]
          Length = 1944

 Score =  273 bits (590), Expect(2) = 3e-78
 Identities = 112/222 (50%), Positives = 156/222 (70%)
 Frame = +1 / +2


Query: 5713 RKPAVDRICVGIPPGECFGLLGVNGAGKSSTFKMLTGDTTVTRGDAFLNRNSILSNIHEV 5892
          R PAVDRI + + PGECFGLLGVNGAGK++TF+M+TGD   + G    N   +     +
Sbjct: 791  RPPAVDRITMAVLPGECFGLLGVNGAGKTTTFRMITGDLDPSDGLILTNGYDMNLEWRQA 970
            1       2  3

Query: 5893 HQNMGYCPQFDAITELLTGREHVEFFALLRGVPEKEVGKVGEWAIRKLGLVKYGEKYAGN 6072
           Q++GYCPQFDA+   LTGRE +EF+  LRG  +  +    E  + +L L  + +
Sbjct: 971  QQSIGYCPQFDALLTYLTGRETLEFYGRLRGQHDGLLRVEVERLLEELHLTHHADVAVKY 1150


Query: 6073 YSGGNKRKLSTAMALIGGPPVVFLDEPTTGMDPKARRFLWNCALSVVKEGRSVVLTSHSM 6252
          YSGG +RKLS A+AL+G  P++ LDEPT G+DP +RR +WN  +   + GR+V+L+SHSM
Sbjct: 1151 YSGGKRRKLSVAVALLGDSPLLCLDEPTAGVDPISRRRVWNAIIRHNQRGRTVLLSSHSM 1330


Query: 6253 EECEALCTRMAIMVNGRFRCLGSVQHLKNRFGDGYTIVVRIA 6378
          EECE LC+R+AIMVNGRF+CLG+ QHLK+RFG GY++ ++++
Sbjct: 1331 EECEVLCSRVAIMVNGRFKCLGTCQHLKDRFGRGYSLAIQVS 1456
```

Figure 3.3.: BLAST output with a cutoff value of $1e^{-20}$ and a pairwise alignment view between query and subject sequence of a certain hit. At the top, subject sequences are listed, which produce a significant alignment with the query sequences. To the right the parameters expectation value E and the normalized bit score S' are shown to get information about the significance and the quality of the hit. The hit table shows further information including the number of identities and the number of positives (fractions of residues that are either identical or similar). The alignment is mapped below, whereas an identity is marked with 1, a mismatch with 2, and a similarity with 3. Here, the ABC transporter sequences of *Fasciola hepatica* (dataset 19929) were used as subject sequence and the ABC transporter sequences of *Homo sapiens* as query database.

4. Sequence Alignment

The sequence alignment procedure forms the backbone of comparative and evolutionary genomics. It enables to identify regions of high similarity or functionally important sequence motifs.
Furthermore, the structure and the function of DNA are closely related and by comparing new sequences to those with known function conclusions can be drawn on the function of the unknown organism. [88]

An alignment program tries to find the best alignment between the sequences or in other words, it attempts to detect a path through the dot plot diagram including all (or the most visible) diagonals [59].
There are many possible ways to align two sequences, and in order to select the best one, means are needed to quantify their relative quality. The idea is to assign a score to each alignment, and then choose the one with the optimal score. The scoring schemes used for alignments typically include a substitution matrix and a gap penalty function. The substitution matrix is used to score matches and mismatches and the gap penalty function scores insertions and deletion events. [11]
Dot plots can also be used to assess repetitiveness in a single sequence. A sequence can be plotted against itself and regions that share significant similarities will appear as lines of the main diagonal.

4.1. Dot Plot

One way to visualize the similarity between sequences is to use a similarity matrix, known as dot plot. If nothing is known about the evolutionary relationship between the sequences, a dot plot provides a graphical illustration of the level of similarity, and the location of conserved elements in the sequences.
Each axis of the plot represents one of the two sequences to be compared. A dot is placed at each position where two residues match. In the resulting plot, regions of similarity will appear as diagonal stretch of dots. If two sequences are identical, the diagonals show a line. Insertions and deletions between the sequences will appear as lateral displacements of the diagonals and duplications appear as parallel diagonal lines in the plot. [81]

The dot plot is a visual aid, which helps to rapidly identify similar regions in sequences but it does not provide an alignment. In terms of revealing the best alignment in some sense, the dot plot method is limited. For this, a scoring scheme is needed, which is able to quantify the different possible alignments.

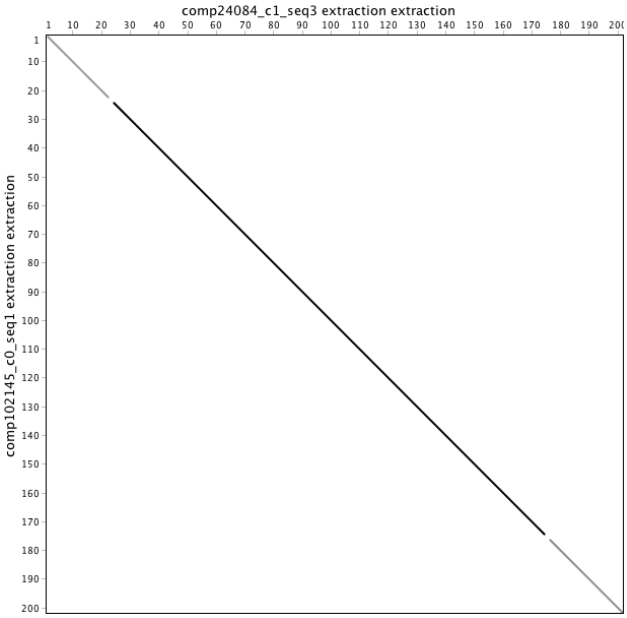


Figure 4.1.: The dot plot, generated by Geneious, shows a comparison of two C transporter sequences with a word size of 8. On the x - coordinate the sequence of the dataset 19931 is represented and on the y - coordinate the sequence of 19932 is shown. The sequences are identical, expect for the two substitutions cytosine to thymine and guanine to adenine. The identical parts of the sequences are represented by a diagonal line, which is twice interrupted at the positions of the substitutions.

```
>Nucleotide alignment 2 Alignment of 2 sequences: comp24084_c1_seq3 extraction
extraction, comp102145_c0_seq1 extraction extraction

Score = 987.0, Identities = 199/201 (99%),
Positives = 199/201 (99%), Gaps = 0/201 (0%)

comp24084_c1_seq3 extraction extraction    336 GCTCATCCAATCGGTCATTCTACTTGACCGTGTACGGAAGCATCGTGGGTGTTAATGTTC 395
                                               GCTCATCCAATCGGTCATTCTA TTGACCGTGTACGGAAGCATCGTGGGTGTTAATGTTC
comp102145_c0_seq1 extraction extraction   184 GCTCATCCAATCGGTCATTCTATTTGACCGTGTACGGAAGCATCGTGGGTGTTAATGTTC 243

comp24084_c1_seq3 extraction extraction    396 TGGCCACCACTTTCCGTGCAGTTCTGTTCGCTTTTGGTGGTTTAGCTGCCGCAGCCATCA 455
                                               TGGCCACCACTTTCCGTGCAGTTCTGTTCGCTTTTGGTGGTTTAGCTGCCGCAGCCATCA
comp102145_c0_seq1 extraction extraction   244 TGGCCACCACTTTCCGTGCAGTTCTGTTCGCTTTTGGTGGTTTAGCTGCCGCAGCCATCA 303

comp24084_c1_seq3 extraction extraction    456 TTCACGGGAATGCCTTGGACACAGTCCTCCATGCACGTGTCTCATTCTTCGATAGGACTC 515
                                               TTCACGGGAATGCCTTGGACACAGTCCTCCATGCACGTGTCTCATTCTTCGATA GACTC
comp102145_c0_seq1 extraction extraction   304 TTCACGGGAATGCCTTGGACACAGTCCTCCATGCACGTGTCTCATTCTTCGATAAGACTC 363

comp24084_c1_seq3 extraction extraction    516 CACAGGGTCGCATTCTGAATC 536
                                               CACAGGGTCGCATTCTGAATC
comp102145_c0_seq1 extraction extraction   364 CACAGGGTCGCATTCTGAATC 384
```

Figure 4.2.: The sequence alignment in text view of the sequences of Figure 4.1 allows a more detailed view of the substitutions. The contig *comp24084_ c1_ seq3* is part of the dataset 19931 and the contig *comp102145_ c0_ seq1* of 19932. The 2 differences (C $\rightarrow$ T, G $\rightarrow$ A) found between the resistant *Fasciola hepatica* 19931 and the sensitive *Fasciola hepatica*19932 are marked.

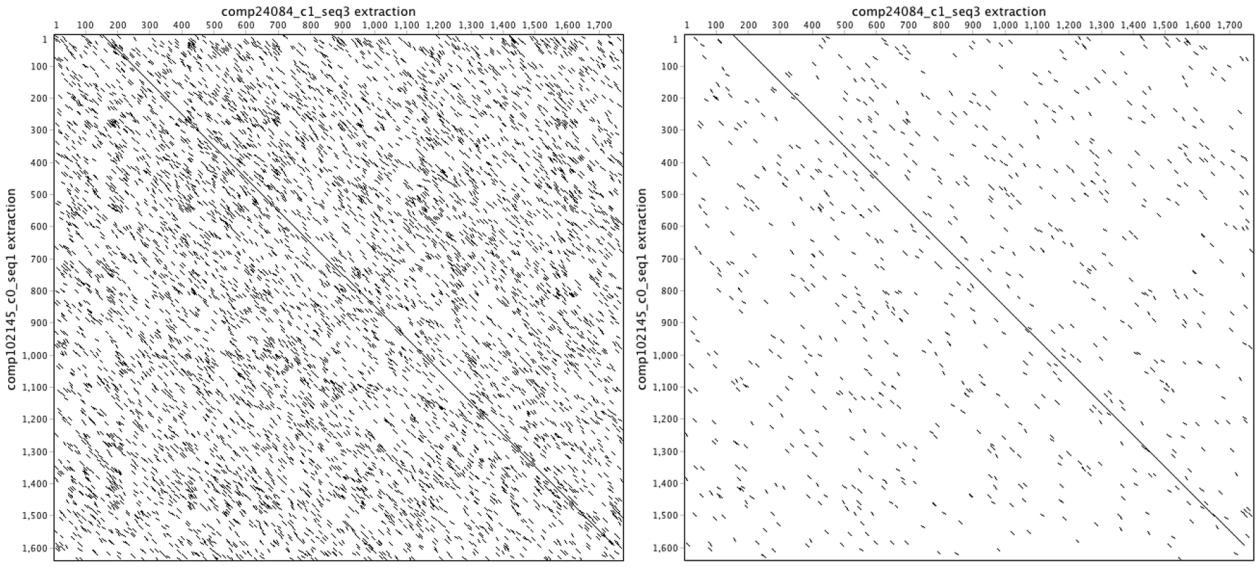


Figure 4.3.: Dot plot with high amount of background noise. By adapting the window size and the thresholds the noise can be reduced. The optimal result is shown in Figure 4.1 whereas not only the diagonal becomes visible but also the two nucleotide substitutions.

Moreover, the amount of background noise on a typical dot plot is huge and impedes the extraction of information (Figure 4.3). To distinguish dot patterns arising from background noise from significant dot patterns it is necessary to apply a filter. [11] Maizel and Lenk developed a filter method using overlapping fixed length windows and required that the comparison achieved some minimum identity score summed over that window before being considered. Therefore, only diagonals of a certain length will survive the filter. Most dot plot software provides a default window length and this is sufficient for an initial analysis. When searching for internal repeats, the length of the repeat can be used to cut out background noise, or the window length can be set, for example, to the length of an exon when comparing coding sequences. [70] [97]
However, an additional issue when comparing protein sequences is that we might not only want to highlight exact matches, but also take into account chemical and structural similarities between amino acids.

4.2. Nucleotide Substitution Models

An important problem in biological sequence analysis is to determine the evolutionary distance between sequences. The distance between two sequences is defined as the expected number of nucleotide substitutions per site. To estimate the number of nucleotide substitution between the common ancestor and the current sequences, models have to be developed. The easiest way would be the direct measure of the proportional number of mismatches in ungapped alignment of the two sequences. To get the evolutionary distance, the number of differences simply has to be divided by the sequence length. However, this description is only sufficient for very closely related sequences, because the probability of having a second substitution in an already changed position increases with time [11]. To estimate the number of substitutions, a probabilistic model is needed to describe changes between nucleotides. Commonly, continuous-time Markov chains are used for this purpose. [114] Often substitution rates between nucleotides are further constrained, leading to different models of nucleotide substitution described in chapter 4.2.1 and 4.3. [94]

4.2.1. Markov Models

A common model of substitution is to use a homogeneous, continuous-time, time reversible, stationary Markov chain. They form the basis of the likelihood and Bayesian analysis of multiple sequences on a phylogeny, when used in distance calculations. The basic assumption of a Markov chain is that the probability with which the chain jumps into other nucleotide states depends on the current sate, but not on how the current state is reached. Therefore, it depends only on the current but not on the past state. [94]
Each state represent a single nucleotide and the transition probability P_{jk}, gives the probability of replacing nucleotide j with nucleotide k.

$$P = \begin{array}{c} \\ A \\ T \\ G \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & T & G & C \end{array} \\ \begin{pmatrix} P(A|A) & P(A|T) & P(A|G) & P(A|C) \\ P(T|A) & P(T|T) & P(T|G) & P(T|C) \\ P(G|A) & P(G|T) & P(G|G) & P(G|C) \\ P(C|A) & P(C|T) & P(C|G) & P(C|C) \end{pmatrix} \end{array}$$

$P(T|A)$ is for example the probability that the next character will be T while the current character is A. [47] These probabilities can be weighted differently resulting in diverse models. The most important DNA models are described below starting with the simple Jukes Cantor model. More general models can be obtained by allowing unequal substitution rates and base frequencies. [63]

The Jukes - Cantor Model

The simplest Markov model of substitution in a single time step is the JC69 (Jukes and Cantor 1969) model [31]. All substitutions are independent and the rates are set to be equal and therefore substitutions among the four types of nucleotides occur randomly. The consequence of this assumption is that the overall rate of substitution is $\lambda = 3\alpha$.

$$Q = q_{ij} = \begin{array}{c} \\ A \\ T \\ G \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & T & G & C \end{array} \\ \begin{pmatrix} - & \alpha & \alpha & \alpha \\ \alpha & - & \alpha & \alpha \\ \alpha & \alpha & - & \alpha \\ \alpha & \alpha & \alpha & - \end{pmatrix} \end{array}$$

The rows sum to zero and consequently the diagonal elements are given by taking the negative sum of the other, namely -3α. Using the rates in matrix Q, the probability of each nucleotide substitution occurring when $t > 0$ can be determined, creating another matrix called the transition probability matrix $P(t) = p_{ij}(t) = e^{Qt}$. [25]

The stationary distribution[1] of this Markov chain $\varphi = (0.25, 0.25, 0.25, 0.25)$
Using the Markov chain, expressions can be derived describing how changes accumulate at site i over a period of time Δt. After one time step the probability of observing e.g. an A at site i is given by

$$\varphi_{A,t+1} = (1 - 3\alpha)\varphi_{A,t} + \alpha\varphi_{C,t} + \alpha\varphi_{G,t} + \alpha\varphi_{T,t} \tag{4.1}$$

where $\varphi_j^{(t)}$ is the probability of being in state E_j at time t. Equation (4.1) can be reduced to

$$\varphi_{A,t+1} = (1 - 3\alpha)\varphi_{A,t} + \alpha\left[1 - \varphi_{A,t}\right] \tag{4.2}$$

The first term represents the probability of observing A at time $t + 1$ if the residue at site i at time t was an A. The second term is the probability of observing A if the residue at time t was not an A. [31]Equation (4.2) can be applied equally to the other 3 nucleotides, since the model is symmetric.
After some algebraic manipulation an expression for the evolutionary distance between two sequences can be defined. The Jukes - Cantor distance is given by

$$d_{JC}(a, b) = -\frac{3}{4} log\left(1 - \frac{4}{3}D\right) \tag{4.3}$$

where D represents the number of observed nucleotide differences divided by the total number of nucleotides of the two sequences a and b. [101]

[1] The stationary distribution of a Markov Chain with transition matrix P is some vector ψ, such that $\psi P = \psi$

The K80 Model

The 2 parameter Kimura model (1980) does not consider the probability of mutations of all 4 nucleotides as random and distinguishes between transitions and transversions. Substitutions between the two pyrimidines (C $\leftrightarrow$ T) or the two purines (A $\leftrightarrow$ G) are called transitions whereas transversions include substitution of a pyrimidine by a purine or reverse (C,T $\leftrightarrow$ G,A). Transversions, representing substitutions across types of nucleotides occur less frequent than substitutions between the same type. [98] Therefore, the transition matrix of the two-parameter model using uniform base frequencies but different rates for transitions and transversions adapts to

$$\begin{array}{c} \\ A \\ T \\ G \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & T & G & C \end{array} \\ \begin{pmatrix} - & \beta & \alpha & \beta \\ \beta & - & \beta & \alpha \\ \alpha & \beta & - & \beta \\ \beta & \alpha & \beta & - \end{pmatrix} \end{array}$$

where α is the rate of transitions and β the rate of transversions. [31] If R is the expected ratio of transition changes to transversions, the following relations can be expressed:

$$\alpha = \frac{R}{R+1} \tag{4.4}$$

$$\beta = \left(\frac{1}{2}\right) \frac{1}{R+1} \tag{4.5}$$

Therefore, the probability of transitions is

$$P_{Transition} = \frac{1}{4} - \frac{1}{2}e^{\left(-\frac{2R+1}{R+1}t\right)} + \frac{1}{4}e^{\left(-\frac{2}{R+1}t\right)} \tag{4.6}$$

and the transversion probability is

$$P_{Transversion} = \frac{1}{2} - \frac{1}{2}e^{\left(-\frac{2}{R+1}t\right)} \tag{4.7}$$

The K80- 2 parameter model results in a corrected distance

$$D_{K2p}(a,b) = -\frac{1}{2}log(1-2p-q) - \frac{1}{4}log(1-2q) \tag{4.8}$$

where p and q represent the number of differences of the sequences concerning transitions and transversions divided by the total number of nucleotides of the two sequences a and b. [4] [79]

HKY85

Hasegawa, Kishino and Yano developed a model in 1985 known as the HKY85 [25]. It is a combination of the Kimura80 and Felsenstein81 models[2] and should describe the nucleotide sequence behavior more realistic with the aid of additional parameters [11].
As the K80 model (section 4.2.1) the HKY85 model distinguishes between the rate of transitions and transversions. Additionally, the base frequencies $\pi = (\pi_A, \pi_T, \pi_G, \pi_C)$ are considered as unequal. [79]
The rate matrix converts to

$$\begin{array}{c} \\ A \\ T \\ G \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & T & G & C \end{array} \\ \begin{pmatrix} - & \beta\pi_T & \alpha\pi_G & \beta\pi_C \\ \beta\pi_A & - & \beta\pi_G & \alpha\pi_C \\ \alpha\pi_A & \beta\pi_T & - & \beta\pi_C \\ \beta\pi_A & \alpha\pi_T & \beta\pi_G & - \end{pmatrix} \end{array}$$

GTR Model

The General time-reversible model is the most general model which was first proposed by Tavaé. It further extended the models to allow all six pairs of substitutions to have differing rates. The substitution rates differ between each pair of nucleotides and are consistent within a pair indicating time reversibility. [11]
The substitution rate matrix adapts to

$$\begin{array}{c} \\ A \\ T \\ G \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & T & G & C \end{array} \\ \begin{pmatrix} - & \alpha\pi_T & \beta\pi_G & \gamma\pi_C \\ \alpha\pi_A & - & \sigma\pi_G & \rho\pi_C \\ \beta\pi_A & \sigma\pi_T & - & \epsilon\pi_C \\ \gamma\pi_A & \rho\pi_T & \epsilon\pi_G & - \end{pmatrix} \end{array}$$

There exist further models, but these are the most commonly used. They allow for unequal expected frequencies of bases and for inequalities of transitions and transversions.

[2] Felsenstein suggested that the substitution rate only depends on its base frequency

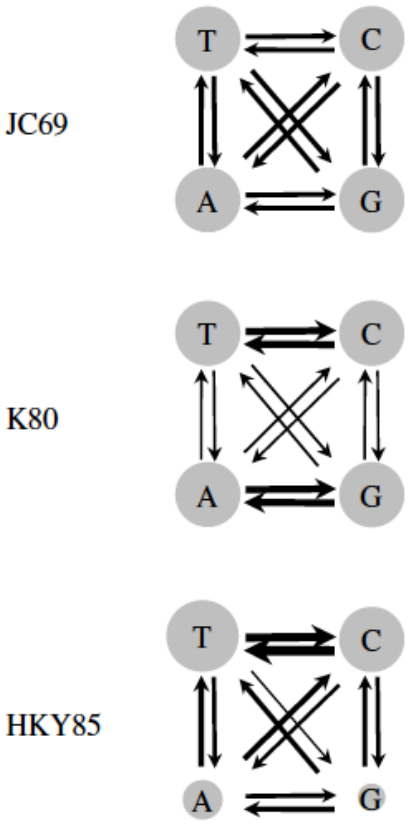


Figure 4.4.: Relative substitution rates between nucleotides of 3 different Markov chain models: Jukes and Cantor (JC69), Kimura (K80) and Hasegawa et al. (HKY85). The thickness of the lines represents the substitution rates while the sizes of the circles indicate the base frequencies. [114]

4.3. Amino Acid Substitution Models

Not all amino acids have the same probability to mutate. Some are more easily replaced or preserved than others. Amino acids with similar codons or properties and structures have a higher probability of exchange.
To evaluate individual point mutations, substitution matrices can be used. [14]
In 1978 the first substitution matrix, called PAM matrix (section 4.3.1) was introduced followed by many similar approaches to score point mutations. The PAM substitution matrices are based on an underlying dataset consisting of closely related aligned protein sequences. [27] PAM1 assumes the sequences to be aligned are 99% identical, hence the accepted point mutation rate is 1% [87]. In order to deal with more divergent sequences the evolutionary information based on the PAM1 matrix is extrapolated to higher matrix levels (e.g. PAM250) by multiplying the PAM1 by itself [19].
Other popular substitution matrices are for example the Henikoff matrix and the BLOSUM matrix (section 4.3.2). Later on the GONNET matrix, which is based on the dataset of the entire Swiss - Prot database and also an update of the PAM matrix, the JTT matrix, become popular. [11]

The information content in scoring matrices is given in terms of *relative entropy*. The higher the entropy is, the higher is the evolutionary distance between sequences. PAM250 has a relative entropy of about 0.36 bits per aligned residue, whereas that of PAM120 is 0.98. [85] Therefore, the relative entropy can be used to compare different scoring matrices and is given by

$$H = \sum_{i=1}^{20} \sum_{j=1}^{20} q_{ij} s_{ij} \tag{4.9}$$

where s_{ij} represents the scores scaled in bit units and q_{ij} the target frequencies from the aligned amino acid pairs.
Equation (4.9) calculates the average mutual information per amino acid pair. [85]
Gap penalties are neither considered in BLOSUM nor in PAM matrices [10].

PAM100 $\sim$ BLOSUM90
PAM120 $\sim$ BLOSUM80
PAM160 $\sim$ BLOSUM60
PAM200 $\sim$ BLOSUM52
PAM250 $\sim$ BLOSUM45

Table 4.1.: Entropy of scoring matrices PAM and BLOSUM. To compare close related sequences PAM100 or BLOSUM90 are appropriate. For more distant sequences PAM250 or BLOSUM45 are available. [27]

4.3.1. Point Accepted Mutation Scoring Matrix

The point (or percent) accepted mutation scoring matrix, short PAM, was the first scoring matrix, developed by Margaret Dayhoff et al. in 1978 and was used to evaluate individual point mutations based on experimental data [19]. The dataset contains 71 gapless alignments of sequences having at least 85% similarity [27].
Dayhoff built a phylogenetic tree of the evolutionary closely related proteins to find out accepted mutations and stored the data in a matrix. The matrix entries represent the likelihood of replacing an amino acid i by an amino acid j in an given evolutionary time. This matrix is called the PAM matrix and has twenty rows and columns representing the amino acids translated by the genetic code. [114]

1 PAM unit of time is the amount of time one amino acid takes in every hundred to undergo an accepted mutation.
Thus, the PAM1 scoring matrix can only be used in comparing relatively evolutionary close sequences. The diagonal in a PAM1 matrix represents the probability to still observe the same residue after 1 PAM. It doesn't not mean that there was no mutation but maybe a succession of two or more mutations ending at the initial residue.
As the divergence of the sequences increases, PAM50, PAM100 or PAM250 are possibilities to further score the substitutions. [27] With the assumption that repeating mutations follow the same pattern as those in the PAM1, these matrices are all based on the PAM1 [103]. The PAM250 for example is produced by multiplying the PAM1 by itself 250 times, representing 250 accepted point mutations [85].

PAM matrices are used as a scoring matrix when comparing DNA sequences or protein sequences to judge the quality of the alignment. This form of scoring system is utilized by a wide range of alignment software.

So far, just a mutational probability matrix M was defined, given the probability of amino acid i being replaced by the amino acid j over a given evolutionary time. The entries of the non-diagonal elements of the unsymmetric probability matrix can be calculated with

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_k A_{kj}} \tag{4.10}$$

and the diagonal elements with

$$M_{ii} = 1 - \lambda m_i \tag{4.11}$$

in these equations m is the relative mutability[3], A the matrix of accepted point mutations and λ represents constant of proportionality. To obtain a scoring matrix a measure has to be introduced, which reflects the significance of an alignment with respect to what could happen randomly. [27]

[3] relative mutability $m_j = \frac{\text{number of changes of j}}{\text{number of occurrences of j}}$

Thus, the score involves the ratio between the probability derived from non random to random models and can be described by

$$P_{ji,n} = f_i M^n_{ji} \tag{4.12}$$

$$r_n(i,j) = \frac{M^n_{ji}}{f_j} = \frac{P_{ji,n}}{f_i f_j} \tag{4.13}$$

M^n_{ji} represents the mutational probability matrix at PAM level n and gives the probability of point accepted mutations replacing the jth amino acid with the ith amino acid. The denominator in equation (4.13) expresses the probability of these amino acids being aligned by chance. f represents the effective frequency. [27]

Essentially, all substitution matrices are log-odds matrices. A log-odds score is the logarithm of the likelihood ratio of two models. [11] Thus, with the aid of log-odd scores the PAM matrix can be defined by

$$s_n(i,j) = log\frac{M^n_{ji}}{f_j} = log\frac{P_{ji,n}}{f_i f_j} \tag{4.14}$$

A positive value ($s_n > 0$) as result of equation (4.14) indicates a positive score and characterizes the accepted mutations. $s_n < 0$ indicates a negative score and therefore an unfavorable mutation.

The PAM Matrices work well with similar sequences but for evolutionarily distant sequences the results become less realistic. Furthermore, the assumption of constant substitution rates throughout the sequences is definitively not true in reality. A new approach was the block substitution matrix which should deliver more realistic scores, especially for distant related sequences.

4.3.2. Blocks Substitution Matrix

The Blocks Substitution Matrix, short BLOSUM, is also designed for scoring protein alignments. [87] In contrast to the PAM matrix, the BLOSUM matrix is constructed empirically from multiple alignments of evolutionarily more distant, but homologous protein sequences. [11] BLOSUM matrices use a larger amount of sequence data than PAM matrices and consider local alignments blocks or highly conserved regions rather than independent residue alignments. [87]

First of all, the protein blocks have to be retrieved from the BLOCKS database, which consists of ungapped multiple alignments of highly conserved protein regions. The next step is to cluster the sequences according to the given matrix level (BLOSUM90, BLOSUM80, etc.) in each block. Afterwards, sequence weights have to be set in a way that the contribution of each cluster in a block is equal to one. For example, for the BLOSUM60 matrix, sequences with at least 60% identity are clustered.
Next, the number of matches and mismatches are determined in each block column,

resulting in a table of frequencies f_{ij} of observed pairs of amino acids i and j. Thus, the probability occurrence for each amino acid pair (q_{ij}) can the estimated and additionally calculated for a random model (e_{ij}) [11]

$$q_{ij} = \frac{f_{ij}}{\sum_{i=1}^{20} \sum_{j=1}^{i} f_{ij}} \tag{4.15}$$

$$e_{ij} = \begin{cases} p_i p_j & \text{if } i = j \\ 2 p_i p_j & \text{if } i \neq j \end{cases} \tag{4.16}$$

where p_i is the probability of amino acid i occurring in a pair

$$p_i = q_{ij} + \sum_{i \neq j} \frac{q_{ij}}{2} \tag{4.17}$$

The log-odd ratio can be calculated by

$$S_{ij} = 2 log_2 \frac{q_{ij}}{e_{ij}} \tag{4.18}$$

where q_{ij} represents the observed frequencies of occurrence of a pair i, j and e_{ij} gives the expected frequencies of i, j. [27]

4.4. Pairwise and Multiple Sequence Alignment

To compare sequences and to identify significant mutations sequence alignments are conducted. In 1970, Needleman and Wunsch published a dynamic programming method to produce global pairwise alignments. There, optimal alignments of partial sequences are searched and stepwise composed to an optimum global alignment resulting in one-to-one comparison of two sequences. Global alignments are useful when comparing sequences that have not diverged substantially, or when the sequences constitute a single element, such as a protein domain. [32]
If the sequences are highly diverged or have become rearranged during evolution, a local alignment might be more suitable. [11] Smith and Waterman extended the ideas of Needleman and Wunsch to develop a local alignment algorithm called the Smith-Waterman algorithm, which searches partial paths in dot plots (section 4.1), which are no longer extended once the marginal sequences of the partial path don't match well and the similarity score goes below the threshold. [29] Both algorithms belong to a class of algorithms called dynamic programming algorithms. They allow to find optimal solutions to problems divided into subproblems but can take a long time to run. [32]
In sequence alignments it can be distinguished between pairwise alignments, in which sequences, even if they are part of a larger set, are aligned only in pairs, and multiple alignments, in which more than two sequences are aligned simultaneously. [59]

Till now, the Needleman-Wunsch and the Smith-Waterman algorithms are the most accurate pairwise alignment algorithms in existence [32].

To compare the sequences and to quantify similarity, the sequences are arranged in rows on top of each other such that matched residues are arranged in successive columns. With the aid of insertions the number of matching residues can be optimized and the resulting alignment is an assembly of matches, mismatches, insertions and deletions.
As there are always many different possibilities to align sequences and to select the best one, similarity scores have to be introduced to quantify their relative quality. [11] The scoring schemes normally include substitution matrices to score matches and mismatches and gap penalty functions to judge insertions and deletions. By summing up the single scores, a complete alignment score is received and gives a measure of the quality of the current alignment.

Multiple sequence alignment is an extension of pairwise alignment to incorporate more than two sequences. In contrast to the pairwise algorithm, most of the multiple alignment algorithms are heuristic rather than exact solutions and based on progressive alignment (section 4.4.1), such as CLUSTALW. [86] Otherwise, the execution time and the memory requirements would be excessive. Multiple sequence alignments have many uses in molecular biology including e.g. genome sequence annotation, gene prediction, phylogeny reconstruction, RNA and protein structure analysis and functional classification of proteins. [24]

During the last years, lots of different multiple alignment programs appeared on the market because of the continuously increasing amount of sequences in public databases like NCBI.
In section 4.5 some of these popular tools for multiple alignment are described briefly. Furthermore, multiple alignments may be hard to interpret in the text view without additional software tools. Thus, a number of visualization tools have been developed providing graphical representation of the alignments. [24] For example the Geneious[4] tool (section 4.5) which was used to visualize the alignments of the *Fasciola hepatica* dataset.

4.4.1. Progressive Alignment Construction

The most commonly used approach to multiple sequence alignment is progressive alignment. Initially, all possible pairwise alignments are constructed and used to estimate a phylogenetic tree, often called the guide tree, by using a distance-based algorithm. Using the tree, the most similar sequences are aligned to each other using a pairwise algorithm. Then further sequences are added based on the structure of the phylogenetic tree. The most widely used progressive alignment algorithm is currently *CLUSTALW*.

[4] http://www.geneious.com/

Progressive alignment works well especially for close related sequences, but it suffers also from many drawbacks.
The resulting progressive alignment is depended on the guide tree, which is just a rough approximation to the true phylogenetic tree, as it is based on pairwise sequence distances only. In addition, any mistakes made in early steps can not be corrected by later steps.
Other approaches to avoid this problems are for example the iterative methods used by MultAlin (Corpet 1988) and DIALIGN (Morgenstern 1999; Morgenstern, Frech, et al. 1998). Here, the alignment generated from one pass of an algorithm is used to construct a new guide tree, which can then be used for a new alignment. [86]

4.5. Sequence Alignment Tools

In this section the most popular multiple sequence alignment tools on the market are described briefly. Geneious is explored a little bit more in detail because it is used for all visual analysis for the example of use (chapter 5).

All of these tools for multiple alignment, explored below, are available online or for download from their websites.

Geneious Alignment

Geneious is a feature rich software for visualization and analysis of protein or nucleotide sequences [63]. The interface is very user friendly and with the aid of the sources panel and the document table a breakdown of the sequences is possible [67]. It organizes and stores data, provides graphical outputs, visualization of 3D structures and much more.
Geneious also offers the possibility to get a first impression of the similarity of two sequences by using the dot plot option and to determine whether the similarity is global or local.
In addition, several alignment tools are incorporated in Geneious to enable an easy access to different analysis approaches of the same sequences without data shifting. [69] The Geneious multiple alignment algorithm uses progressive pairwise alignment. To develop the guide tree the neighbor-joining method is used. [67]
It provides the alignment software CLUSTALW (section 4.5), MUSCLE (section 4.5), MAFFT, Mauve and LASTZ.
To run a multiple alignment in Geneious, all the sequences have to be chosen and the favored algorithm has to be selected. The execution time is quite low and after finishing the Geneious alignment software offers the option of refining the multiple sequence alignment. Therefore, sequences can be removed from the alignment and realigned very simply.
Concerning the substitution matrices used in Geneious alignment, protein sequences offer the choice of either PAM or BLOSUM matrices. For nucleotide sequences match and mismatch costs can be determined, which allow to set different scores for transition and

transversion resulting in various substitution models. Geneious is also able to indicate the amount of similarity of the sequences and to calculate the optimal score by itself. For both protein and nucleotide alignments gap penalties can be set by the user. [67]

To build a tree, protein or nucleotide sequences can be chosen and different options are available. One possibility is the genetic distance model, which allows the user to determine the substitution model used to estimate the branch lengths. If DNA sequences are used, the Jukes Cantor, HKY (section 4.2.1) and Tamura Nei models are choices and if amino acids sequences are considered only the Jukes Cantor distance correction can be used. The tree building method in Geneious is the neighbor-joining method or the UPGMA method. [67]

CLUSTALW Alignment

CLUSTALW is an improvement of the original CLUSTAL program introduced in 1990 by Higgins and Sharp and is the most commonly used multiple sequence alignment tool. The 'W' in CLUSTALW stands for 'weights' because CLUSTALW uses a sophisticated scheme in order to prevent very similar sequences from dominating the multiple sequence alignment. [20] It uses substitution matrices and gap penalties depending on several parameters such as local sequence similarity and amino acid composition of protein sequences.If for example the PAM or BLOSUM substitution matrix is chosen by the user, CLUSTALW automatically chooses the most adapted matrix level [20].
The algorithm uses the progressive method, explained in section 4.4.1, to build the alignments. Therefore, the software clusters the sequences by similarity to produce a phylogenetic tree. Following the dendrogram topology and starting to align the single branches new alignments are received. These new alignments are treated by CLUSTALW like single sequences and are aligned to each other. [20]
One drawback of CLUSTALW is that it is a strictly global alignment program and always aligns a sequences from the beginning to the end. TCoffe is an alternative to CLUSTALW, also using progressive alignment, providing more accurate alignments but to a higher execution time [20].

MUSCLE Alignment

MUSCLE is one of the fastest multiple alignments methods. It uses progressive alignment and is quite accurate. It is even more accurate than TCoffe and faster than CLUSTALW. First of all, MUSCLE generates a rough draft of the alignment, using a very simple guide tree. In the next step a more accurate guide tree is developed, based on the initial alignment and a second progressive alignment is generated. [115]

5. Fasciola Hepatica - Example of Use

5.1. Introduction

Fasciola hepatica is a brown flatworm helminth and belongs together with the *Fasciola gigantica* to the liver flukes [53]. The parasitic trematode may infect livestock and humans leading to the disease fascioliasis, affecting the biliary canals and gallbladder, causing enormous economic losses and high medical costs [9].
The WHO estimates that at least 2.4 million people are infected in more than 70 countries worldwide.
Additionally, the estimated number of unreported cases is much higher. It is appraised that more than 180 million people are at risk to get infected. [112]
Countries with rather damp climate and mild temperatures, like Ireland, present suitable conditions for the liver fluke's intermediate host, the freshwater snail [33].
Nevertheless, *Fasciola hepatica* can be found on all continents, except the Antarctic, which indicates its high adaptability to external conditions [1].

Humans get rather accidentally hosts of *Fasciola hepatica* by transmission from animal to humans (zoonosis) [75].
In regions where infected animals are in contact with vegetation consumed by humans, transmission can happen easily. Nearly all known cases result from watercress consumption or through the ingestion of water lettuce or alfalfa [95]. The primary hosts of *Fasciola hepatica* are wild or domesticated mammals, especially cattle and sheep, but also goats, horses, buffaloes, alpacas, camels, deer and rabbits may be affected [107] [9].

In 1379, the first references of infections with *Fasciola hepatica* are made by Jehan de Brie [95]. France, was the first country in which a modern epidemic of human fascioliasis occurred in 1956 [102].
At the beginning, fascioliasis was completely underestimated, but with the enormous increase in the number of infections of animals and humans and the distribution of the helminth all over the world, effective strategies were researched to regain control of the disease [72] [71].
Fascioliasis became not only an economic problem, due to decreased meat and milk production, decreased female fertility of the ruminants and increased veterinary costs, but also a worldwide human health problem [18].

Since 1983 an anthelmintic drug, triclabendazole (TCBZ) exists, which helps to prevent the spread of the disease [43]. Till now it is still the only medication recommended by the WHO and additionally, the only effective known pharmacological treatment against fascioliasis [112] [1].
Triclabendazole is administered once orally or twice in more severe cases and shows high efficacy against the immature and the adult liver flukes [60].

However, shortly after the introduction of TCBZ as treatment for infected animals with fascioliasis, resistance to TCBZ has been reported. The first case of resistance was documented in Australia in 1995, followed by reports in South America and Europe. [43]
The mechanism underlying TCBZ resistance probably results from active efflux or reduced uptake of the drug [17]. Furthermore, possible changes in the target molecule or drug modification may also play a role in TCBZ resistance [51].

Lots of researches started to find alternatives to TCBZ. Hernandez et al. attempt to develop vaccines to treat fascioliasis, but the lack of knowledge of the immunological processes, e.g. how and when the parasite initiates control of the host immune response, limits the success [75].
Availability of new anthelmintic drugs has become a pressing need. To discover and develop alternatives to TCBZ, the mechanism of drug resistance and the genetic adaption of *Fasciola hepatica* have to be understood first [99].

5.1.1. Epidemiology

Fascioliasis is currently the most widespread disease known in terms of latitude, longitude and altitude [72].
In Australia, the prevalence of fascioliasis in different regions is reported from 41% in Victoria, 31% to 55% in Gippsland and 0.4% to 50% in Queensland. In northern Victoria researches in sheep reported a fluke intensity up to 72 flukes per animal. [33]
Studies in the UK show an enormous increase of the prevalence of fascioliasis from 48% to 72% in just 3 years from 2003 to 2006 [75]. In Belgium the prevalence is estimated to be 37%, in Germany 50% and in Spain even 61% [73] [15] [58].
Also serological surveys in Austria were conducted investigating bulk milk samples of domestic dairy farms. Of the total round 5.000 tested samples, 15.5% respond clearly positive and 29.5% weakly positive. Concerning the frequency of positive serological reactions significant regional differences in Austria could be found. In the pasture areas of the Upper Styria, the West and East Styria the infection frequency is expectedly higher than in the South plains and hill land. [40]
A study conducted in Tyrol revealed, that in 73% of the tested cows antibodies against the liver fluke *Fasciola hepatica* were found [1].

In Europe, more than 55% of all farm animal diseases are caused by parasitic helminths, resulting in production losses and high veterinary costs. Estimations claim, that every year 400 million € are spent on anthelmintic drugs. [76]

The economic losses concerning the parasite *Fasciola hepatica* exceed 3 billion $ globally [12] [83]. In Australia, it is estimated that 40 million sheep and 6 million cattle are infected with the liver fluke, leading in economic losses of about 50 to 80 million $ per year due to production losses and further 10 million $ per year due to treatment of the infected animals.

Fascioliasis diminishes the proceeds concerning the milk and meat production. The reduction in milk yield in dairy cattle depends rather on the intensity of the infection and further on the level of animal nutrition. [33]
Basically, the reduction concerning the milk output of infected animals has been estimated from 3.8% to 15.2% [56].
Furthermore, the contamination of the liver and the decrease of wool output contribute to the high losses due to fascioliasis.
However, far more damaging effects are the decreased fertility in dairy cattle, lower calf birth weight and the reduced growth of infected animals [55].
Taken as a whole, the production losses are estimated to be around 17% per animal concerning an infection of livestocks with helminth parasites [35].

5.1.2. Morphology and Life Cycle

The *Fasciola hepatica* is a parasitic trematode, which affects mammals and leads to the disease fascioliasis [112].
An adult flatworm (Figure 5.1) has a bay leaf form of about 3 cm length and 1.5 cm width. They appear pink - grayish to dark red and the body surface is covered by various spines. Each *Fasciola hepatica* possesses ovaries and testes and allow individual flukes to reproduce independently. [95]

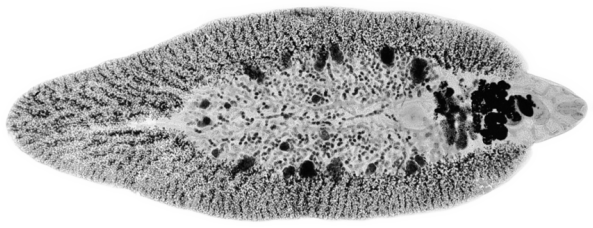


Figure 5.1.: Adult stage of *Fasciola hepatica*. [61]

The life cycle of the liver fluke (Figure 5.3) represents a complex interplay between the parasite and the hosts and takes around 14 to 23 weeks [75].
The resulting prepatent period[1] is dependent on the temperature. Higher temperatures (25 degrees) can reduce it to 38 days and lower temperatures (15 degrees) extend it.
The adult fluke parasitizes the larger biliary passages and the gallbladder of its hosts. Especially cattle and sheep serve as primary hosts for the large liver fluke, but also other herbivorous domestic and wild animals get infected. Cases of fascioliasis in horses, alpacas, donkeys, mules, buffalos, deer, wild boars, rabbits and further grazing animals have been reported. [102]

The adult *Fasciola hepatica* stays in the bile ducts and produces 20 000 to 24 000 eggs per fluke per day, which are released by way of bile and intestine with the faeces [75]. The eggs can only hatch and develop if they reach water of appropriate quality and physiochemical characteristics. Temperatures of about 15 to 25 degrees enables the development of the miracidia[2] and after 9 to 21 days it may hatch. At unsuitable climatic conditions or absence of water the parasite is able to stay viable for several months. [102]
However, the motile miracidia swims rapidly until discovering a mud snail (aquatic or amphibious), its second host and invades via chemotaxis [1].
Inside the intermediate snail host the development of the next three typical lifecycle stages of a trematode take place, including the sporocyst, redial generations and the production of cercariae.
Afterwards, the cercariae is released by the snail, shedded into the water and after a short swim period (usually below an hour), it attaches to water plants above or below the water line. [102]
Under laboratory conditions, a single snail is able to shed around 2000 cercariae before it dies [28]. The cercariae encyst as metacercariae (Figure 5.2), loses its tail and become infective within 24 hours and is able to survive for up to a year [8]. Following ingestion by the definitive host, the metacercariae excyst within an hour and emerge from the cysts in the intestine. The excysted trematodes are now called juvenile flukes and by penetrating through the intestinal wall they arrive at the abdominal cavity by about 2 hours after ingestion. After additional six days the trematode reaches the liver via the abdominal cavity, where they stay for about 5 to 6 weeks.
In this acute phase, the affected patients complain of severe abdominal pain, nausea, skin rashes, fever and respiratory disturbances [112]. The destruction of the liver cells results in internal bleeding and a swollen liver causing symptoms related to haemorrhage and inflammation [1].

Finally, the parasite migrates into the bile duct, where it becomes sexually mature, produces eggs and initates another cycle of infection.
This chronic phase is characterized by jaundice, anaemia and an intermittent pain. [102]
In addition, the patients often suffer from pancreatitis, gallstones and bacterial superinfections and as a result of the long term inflammation also liver fibrosis occurs [112].

[1] Period between infection of the host and the first appearance of eggs in the faeces.
[2] The first stage larva of a trematode.

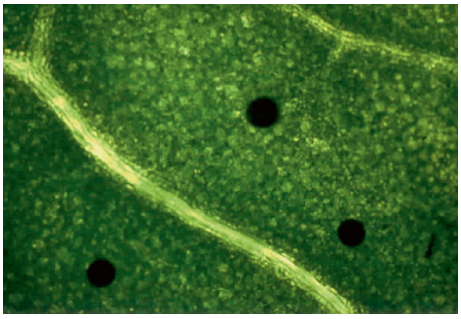


Figure 5.2.: Encysted metacercariae attached to a leaf of a freshwater plant. The definitive host gets infected by ingestion. (Orig. S. Mas-Coma)

The final prepatent period differs depending on the type of host and also on the severity of the parasite infestation in the liver. In sheep and cattle it is about 6 to 13 weeks, whereas in humans rather 3 to 4 months.
Furthermore, the trematode is able to survive for long periods, including anatomical and probably immunological long term damages. The longest recorded lifespan of the liver fluke was 11 years in a sheep and 9 to 13 years in humans [75]. In cattle surviving periods of 9 to 12 months are documented. [102]

5.2. Triclabendazole Resistance

The benzimidazole triclabendazole presents the drug of choice since 20 years against the chronic and acute fascioliasis in ruminant livestock throughout the world [45]. Its widespread and extensive use arises from the superior efficacy against both the immature and mature parasites [22]. The adult liver fluke *Fasciola hepatica* as well as all its immature stages down to 2 days post-infection in the definitive host can be killed by the treatment with triclabendazole [45].
Other standard anthelmintic drugs like praziquantel, used e.g. for the treatment of *Schistosoma mansoni*, are for unknown reasons not effective against fascioliasis [41] [1].

The mechanism of action regarding triclabendazole is not completely understood yet. Based on the known effects of other benzimidazole drugs, it is believed that TCBZ might bind to the β tubulin molecule and therefore prevent the formation of microtubules. [17] It could be shown, that the entry of the drug into the fluke is based mainly on diffusion across the tegumental syncytium rather than by oral ingestion [78].

However, only 12 years after TCBZ was introduced as veterinary drug in 1983 the first case of drug resistance was reported in Australia. Since then further resistance cases in liver fluke populations were documented worldwide. The continuous, inadequate use and

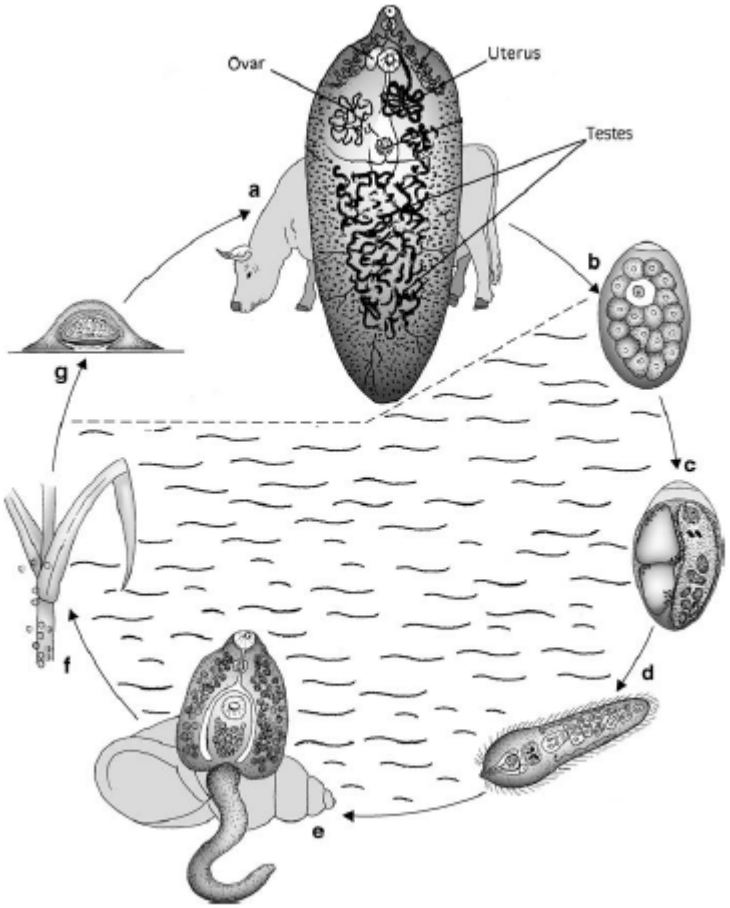


Figure 5.3.: The life cycle of *Fasciola hepatica* [68].
a) the adult flatworm affects the liver and gallbladder of ruminants.
b) Egg without an embryo.
c) Embryo develops within the egg.
d) Miracidium, the first stage larva of a trematode, searching an intermediate host.
e) Cercaria developing within the redia of the intermediate host Galba truncatula.
f) Metacercariae attaches to water plant leafs. The final host gets infected by ingestion.
g) cross section of the metacercariae.

the complete reliance on TCBZ led to the development of resistance of the parasites. [44] Till now, the mechanism underlying TCBZ resistance have not been thoroughly elucidated. Research shows, that it is probably a multifactorial process including changes in drug uptake, drug efflux and metabolism. [109]
In nematodes, benzimidazole resistance is linked to changes in the target molecule, which could not be confirmed for TCBZ resistant liver flukes [17]. In heterologous expression systems, recent studies show an interaction of anthelmintic drugs with homologues of permeability glycoprotein (P-gp) and an increase of intracellular accumulation of fluorescent ABCB1 substrate rhodamine 123 caused by TCBZ [54] [30].
It is known that *Fasciola hepatica* expresses a P-gp like ABC transporter and further it could be proven that their resistant phenotype can be reversed, from resistant to susceptible by applying P-gp inhibitors [109] [17].
The function of ABC transporter (section 5.2.1) is to import or export a wide spectrum of different substrates and the conducted studies suggest that these transporters may be responsible for an active efflux or an reduced uptake of the drug [60].
In addition, a recent survey of Wilkinson et al. examined single nucleotide polymorphisms (SNPs) of a potentially ABC transporter of *Fasciola hepatica*. They discovered a more frequently incidence of the allele specifying codon S1144R in TCBZ resistant isolates [109].

Taken together all this awarenesses, P-gp like ABC transporter might play a significant role in mechanism of drug resistance as well as in detoxification processes in helminths to survive inside their hosts [60].
Thus, polymorphisms in ABC transporters have been increasingly studied over the last few years to detect SNPs leading to transporter dysfunction, to diseases, or resistances to drugs.

5.2.1. ATP - Binding Casette Transporter

ATP Binding Cassette (ABC) transporters are membrane proteins that either import or export various substrates across the cellular membrane [60]. They represent the largest transmembrane protein family and are found in all organisms [84]. Hydrolysis of ATP to ADP provides the energy to drive the active transport of a variety of substrates, including ions, sugars, amino acids, polypeptides, toxic metabolites, xenobiotics and drugs [113].
Based on phylogenetic analysis, the ABC transporters are classified into eight subfamilies (ABCA to ABCH) concerning their sequence and organization of their ATP binding domain.
Within one subfamily, the transporters are further subdivided into subgroups concerning their similarity, additionally marked with numbers, e.g. ABCA1. [23]
Full and functional transporters typically consist of two hydrophilic ATP binding domains (Nucleotide Binding Domains, NBDs) in combination with two hydrophobic

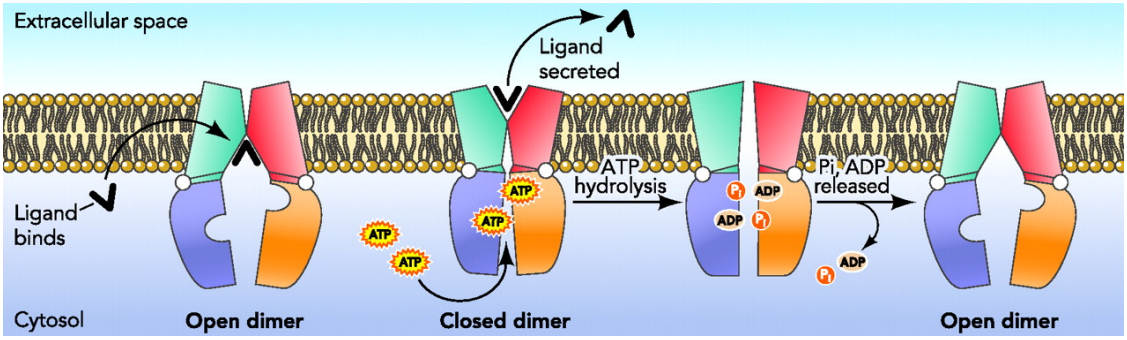


Figure 5.4.: A simple ATP-switch mechanism powers ABC transporters. [64]

transmembrane domains (TMDs) [92]. The ABCE and ABCF subfamilies contain 2 NBDs, but no TMDs [41].
For the transport cycle of ABC transporter (Figure 5.4) and the exact process of hydrolysis several hypotheses are available. Through structural and biochemical analyses it turned out that the ATP binding and hydrolysis is coupled to conformational changes in the transporter based on high and low affinity states for ligands on different sides of the membrane. [64]
In the resting state, the NBDs form an open dimer with low affinity for ATP. By binding of a ligand to the high affinity site on the TMDs, conformational changes are induced in the NBD, transduced via the intracellular loops of the TMDs interdigitating with the NBDs. Thus, the binding probably enhances the affinity to ATP of the NBDs. Previous structural studies prove that unusual conformations of the Walker A motif may preclude binding of ATP.
In the next step, 2 ATP molecules bind and form the closed dimer configuration, which generates a significant amount of free energy, possibly used to induce conformational changes in the TMDs. This change may include e.g. breaking interactions between TM α helices and may alter the position and affinity of the ligand binding site. Consequently the TMDs open, the substrate is released and hydrolysis of ATP follows. In P-gps the hydrolysis occurs nonsimultaneously and necessarily of both ATPs to complete a transport cycle. Through the release of phosphate (P_i) and ADP the transporter snaps back to its initial configuration.
This ATP switch model consists of 4 steps, including 4 conformational changes. First associated with the binding of ligand, then ATP binding, ATP hydrolysis and ADP and P_i release. However, the transport mechanism is rather complex and many details remain to be elucidated. [64]

The highly conserved cytosolic and L-shaped nucleotide binding domains contain characteristic motifs e.g. the Walker A and Walker B. These 2 motifs, found in all ATP binding proteins, are separated by the Q loop and the Walker C motif (the signature), which makes up approximately 90 to 120 amino acids [106].

In contrast to the TMDs, the NBDs are homologous throughout the family including the named motifs and loops (D, H and Q loop), which represent a unique characteristic for investigating ABC transporters. Whereas the NBDs are located in the cytoplasm and are responsible for the energy transfer, the TMDs form the ligand binding sites and provide substrate specificity. [64]

The domains in eukaryotic ABC transporters are organized as either full transporters combining all four required domains (2 TMDs and 2 NBDs) in one polypeptide, or half transporters consisting only of 1 TMD and 1 NBD, requiring homo- or heterodimerization for full functionality [64] [109] [65]. In eukaryotes, substrates are moved from the cytoplasm to the outside of the cell or into an intracellular compartment like the endoplasmatic reticulum (ER), mitochondria or peroxisome. The ABC pumps in bacteria work predominantly as import pumps of essential substrates that cannot be obtained by diffusion into the cell. [23]

The human genome encodes 48 ABC transporters, representing seven of the eight families (A-G) (Table A.1). Up to date 14 of these transporters are proven to be involved in human hereditary diseases, including cystic fibrosis (CF), adrenoleukodystrophy (ALD) and cholesterol metabolism disorders (Table A.2). [23] In addition, several members of the ABC transporters have been associated with drug resistance in parasites. [65]

Some of the transporters seem to be specific for their endogenous substrates, others function as multidrug efflux pumps and are able to transport a variety of different drugs and substrates [23]. The latter are responsible for the transport of cytotoxic compounds out of the cell and play an important role in the uptake and distribution of drugs. [64]
The three most important transporters known to be involved in multidrug resistance are the P - glycoprotein (P-gp/ABCB1), the Breast Cancer Resistance Protein (BCRP/ABCG2) and the Multidrug Resistance Associated Protein 2 (MRP2/ABCC2). These transporters are highly expressed in the human gut, limiting the intestinal absorption of foreign substances. Furthermore, they are localized in the canalicular membrane of the liver and kidney regulating the secretion of drugs and metabolites to bile. [60] The P-gp transporter was the first characterized eukaryotic ABC transporter discovered in 2006, based on the prevention of effective cytotoxic chemotherapy [64]. Other members of the ABCC subfamily and possibly ABCA2 are also associated to multidrug resistance [23].
However, parasites and nematodes seem to possess a high amount of ABC transporters and some of those appear very similar to P-gps. *Caenorhabditis elegans* genomes code for approximately 60 transporters, 13 P-gp homologues of subfamily B and multidrug resistance proteins (MRP) of subfamily C. [92]
In *Haemonchs conchrotus*, the red stomach worm 9 P-gp homologues could be identified and the nematode *Schistosoma mansoni* code for approximately 20 transporters including 7 of class ABCB [110]. Possibly the parasites have developed to express many ABC transporters to improve their adaptability [1].

5.3. Material and Methods

5.3.1. Fluke Isolates

Fluke isolates of known drug sensitivity and treatment history were studied originating from cattle or sheep of various european countries. In total four datasets (19929, 19930, 19931 and 19932) were analyzed, each including the cDNA contigs of two flukes.
The data of 19931 indicate genetic material of the Dutch isolate collected on a farm in North Holland [36]. Tests confirmed the presence of resistance to the anthelmintic drug TCBZ [37]. The others, 19929, 19930 and 19932 are susceptible to TCBZ and are obtained from Down in Northern Ireland, from Cullompton in South West England and from Austria.
The Cullompton isolate was received in 1998 from sheep and is known to be sensitive to several fasciolicides [36].
The Austrian adult liver flukes derive from cattle, fresh slaughtered in Traisen, Lower Austria. In cooperation with the Medical University of Vienna, RNA was isolated to generate full length cDNA.
Basically, the Dutch TCBZ resistant isolate was used for comparisons with the TCBZ sensitive ones to detect functionally significant mutations.

In Table 5.1 information concerning the dataset name, reads, contigs, origin and phenotype of the material is summed up and listed.

Sample ID	**19929**	**19930**	**19931**	**19932**
Origin	Down	Cullompton	Dutch	Austrian
Phenotype	Sensitive	Sensitive	Resistant	Sensitive
Raw Reads	45.925.010	49.833.866	36.095.878	42.419.544
HQ Contigs	97.766	157.300	93.375	256.577
Contig Nucleotides	50.476.973	78.064.169	105.570.474	52.785.237
Included Reads	16.226.440	19.685.638	17.737.410	20.175.394
Reads in HQ contigs	48%	52%	62%	57%
Average Contig length	516	496	1131	206

Table 5.1.: Data concerning the origin, phenotype and the information content of the cDNA of the samples. (cp.[3])

5.3.2. Data Preparation

RNA was prepared and isolated from the Austrian adult liver flukes. It was then converted into the far more stable complementary DNA (cDNA) to enable further analyses. The cDNA sequencing was outsourced and conducted by the Next Generation Sequencing Core Facility, CSF, Vienna, Austria.
The underlying idea of sequencing experiments is to add a nucleotide to an extending primer strand such that the base that is added to the end of the strand is complementary to the corresponding base on the template.
One of the most popular next generation sequencing systems are Illumina, SOLiD, 454 (Roche), Heliscope and SMRT. The quality of the sequencing is mainly dependent on the read length and the sequencing depth. Illumina, SOLiD and Heliscope produce a high amount of short sequence reads of an average length of about $\sim$ 100-mer. The 454 pyrosequencing technology and the SMRT sequencing from Pacific Biosciences establish smaller numbers of longer reads ($>$ 400-mer for 454 and 1.000 - 3.000-mer for SMRT).
On the one hand, for generating a deNovo assembly longer reads are easier to handle concerning sequence assembly. On the other hand Illumina and also SOLiD work with a high sequencing depth and therefore are ideal for semi quantitative experiments such as gene expression.
However, the cost effectiveness of shorter read technologies and the availability of sophisticated genome assembly algorithms result in the standard use in case of deNovo assembly and reference assisted genome characterisation. [91]

The sequence technology provides the raw reads, which were further used to produce an assembly of the genome. In this study, 36.000.000 raw reads for dataset 19931 to 45.000.000 for dataset 19929 were available. Moreover, read pairs were formed, which halves the amount of raw reads and to increase the quality, erroneous and very short reads were excluded. The threshold of eliminating short length reads was set to 50 bp.
In addition, the base calling accuracy, which indicates the probability that a given base is called incorrectly by the sequencer was determined to produce High Quality (HQ) reads. A Phred Quality Score of 30 to a base is equivalent to a base call accuracy of about 99.9%, meaning that every 1000 bp sequencing read will likely contain an error. [50]
In the next step, all pairs of reads are compared to identify overlaps and are assembled as HQ contigs by the use of Trinity. The arising challenge as mentioned before is the shortness of the sequence reads generally only around 300bp long.
An experimental strategy, the paired-end sequencing, facilitates to obtain contiguous assemblies. The DNA fragments are sequenced from both ends, whereas the distance between each paired read is known. Therefore alignment algorithms can use this information to map the reads over repetitive regions more precisely. [91] [50]

Dataset 19932 contains the highest amount of HQ contigs ($\sim$ 257.000) of lowest average length, including round 21.000.000 reads and 53.000.000 nucleotides.
The dataset of the putative resistant liver fluke possesses round 93.000 HQ contigs of 106.000.000 nucleotides leading to the largest average length of 1.130 bp.

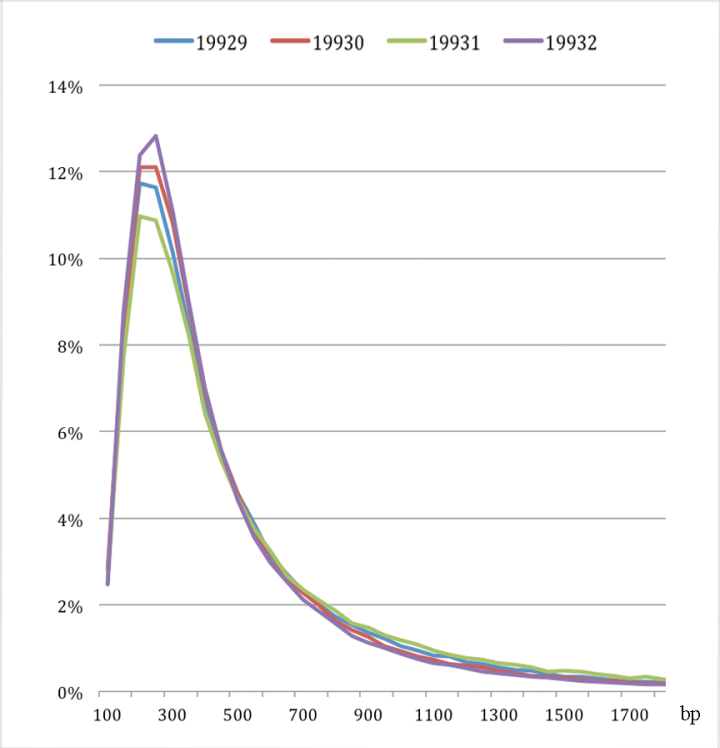


Figure 5.5.: The contig length [*bp*] plotted against the normalized amount of contigs [%] of the *Fasciola hepatica* datasets. [3]

The two other sensitive datasets 19929 and 19930 consist of 98.000 and 158.000 HQ reads of an average length of round 500 bp. The average amount of reads, which are included in the HQ contigs makes up 55% of the raw reads. [3] The exact values concerning the information content of the datasets can be found in Table 5.1. All data preparation, except the sequencing process, was conducted at the department Health and Environment, Bioresources[3] of the Austrian Institute of Technology in Tulln [3].

In this thesis, phylogenetic analyses with the received cDNA ABC transporter datasets are described to further discover single nucleotide polymorphisms and therefore detect ABC transporter dependent resistance.
In addition, comparative analyses of differently expressed genes in resistant versus sensitive specimen are conducted to get insights of the adaption changes developing drug resistance in parasites.

[3]DNA Bank and Genotyping Services

5.3.3. Identification of ABC Transporters

ABC transporters typically encode structural proteins and single SNPs may result in severe function changes or even lead to a complete function loss. Several transporters are associated with genetic disorders and with multidrug resistance in chemotherapy by pumping out anticancer drugs as well as in pathogenic microorganisms [23] [48] [60]. Within parasites the underlying mechanism contributing to the resistance against anthelmintic drugs like triclabendazole seems to be increased drug efflux mediated by ATP binding cassettes transporters. By blocking these transporters it should be possible to restore susceptibility to anthelmintic drugs.

ABC transporters may support the survival of the liver fluke in the hostile environment of the bile duct even though many of the bile salts are conjugated. These free bile salts reach concentrations which are still high enough to be toxic to most cells.
Possibly, the liver fluke gets protected by its outer tegumental layers, but it must further possesses a detoxification mechanism for its gut. The export of bile salts is known to be mediated by the Bile Salt Export Pump (BSEP/ABCB11) in mammals. Therefore, *Fasciola hepatica* expresses maybe a similar protein which provides a biochemical barrier against bile acids.
Underlying this hypotheses, Kumkate et al. discovered in 2002 a protein in *Fasciola gigantica* recognized by a ABCB1 antibody lining the digestive tract. [1]

Furthermore, a recent study of Wilkinson et al. detected 3 SNPs in an putative ABC transporter of *Fasciola hepatica* resulting in an amino acid change. One of these SNPs could be seen more frequently in the TCBZ resistant isolates. [109]

The resistance to anthelmintic drugs may develop due to a mutation in a transcription factor resulting in overexpression of an otherwise normal ABC transporter.
Based on this, we tried to identify ABC transporters of *Fasciola hepatica* and further detect SNP candidate regions involved in drug resistance. The severity of a SNP in a transporter depends on the occurring location leading to a protein change which may affect the function of the transporter [6].

Phylogenetic analyses were conducted for each ABC transporter subfamily in comparison to those of *Homo sapiens*, the fruit fly *Drosophila melanogaster* and the nematode *Caenorhabditis elegans*. Furthermore, sequence alignments were conducted with Geneious to detect SNPs.

5.4. Detection of Significant Hits

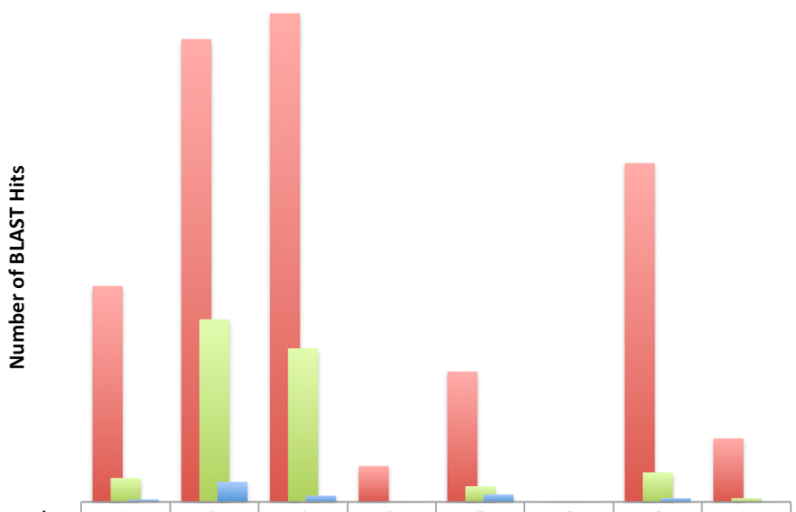
To detect regions of similarity the *Fasciola hepatica* sequences are examined by BLAST search (chapter 3) to a database consisting of ABC transporters of different organisms. The database GenBank was used to retrieve 49 sequences of nucleotides of *Homo sapiens*, 55 protein sequences of *Drosophila melanogaster* ABC tranpsporter and 70 sequences of proteins of *Caenorhabditis elegans*. The ABC transporter sequences of these three organisms are run against the datasets of *Fasciola hepatica* consisting of nucleotide sequences of putative ABC transporters.
Based on the use of nucleotides or protein datasets, the right BLAST program has to be chosen. The programs available for BLAST are listed in chapter 3, Table 3.1.
The *Homo sapiens* dataset consists of nucleotides whereas the *Drosophila melanogaster* and the *Caenorhabditis elegans* are protein databases. Therefore, for the comparison of *Homo sapiens* and the liver flukes, a *tblastx* search was perfomed, searching translated nucleotides databases using translated nucleotide queries. For the other two organisms the BLAST program *tblastn* was conducted, searching translated nucleotide databases using protein queries.

As amino acid substitution matrix the BLOSUM62 matrix (section 4.3.2) is used by default to align the sequences with no more than 62% similarity and to count the relative frequencies of amino acids and their substitution probabilities. In terms of nucleotides, each identical match is scored in the same way whereas mismatches are penalized with negative scores. [105]

Furthermore, a statistical significance threshold for reporting matches against the database sequences was set with an E value of e^{-20}. Matches with ascribed statistical significance greater than the expectation threshold won't be reported. Therefore, the lower the E value, the more stringent matches are filtered out with little similarity [80]. The organisms compared in our study are evolutionary not closely related. Thus, the E value should not be set too low to still receive matches to detect homology.

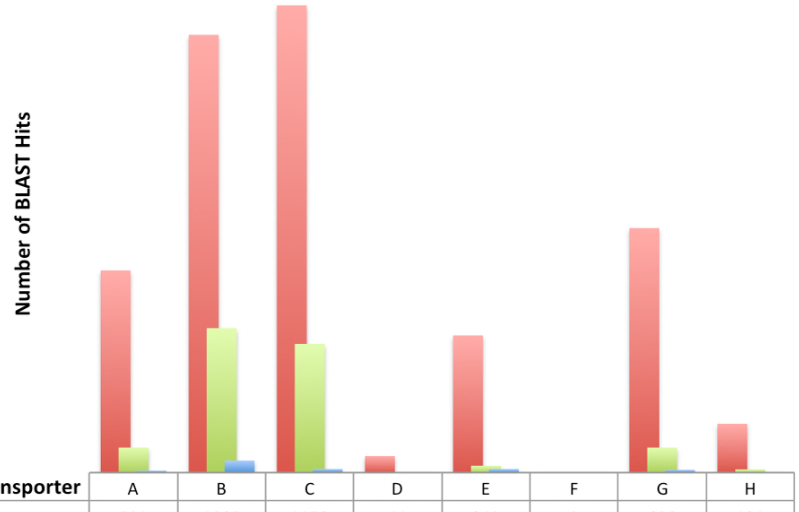
To demonstrate the impact of the E value the comparison of the ABC transporter dataset of *Drosophila melanogaster* was conducted with no threshold and with one of e^{-20} (Figure 5.6). With no statistical significance threshold BLAST detects many hits with low similarities. Specifying the BLAST search by introducing an cutoff E value, homologies between the sequences may be detected. Lots of ABC transporters of *Fasciola hepatica* show similarities to several ABC transporter subfamilies. The unique classification of the contigs to one transporter sharing the most identity provides a first estimation of the amount of ABC transporter existing in the liver flukes (Table 6.1).
The assignment of the single contigs of *Fasciola hepatica* to the subfamilies of the transporters is listed in the Appendix (A.1).

1)

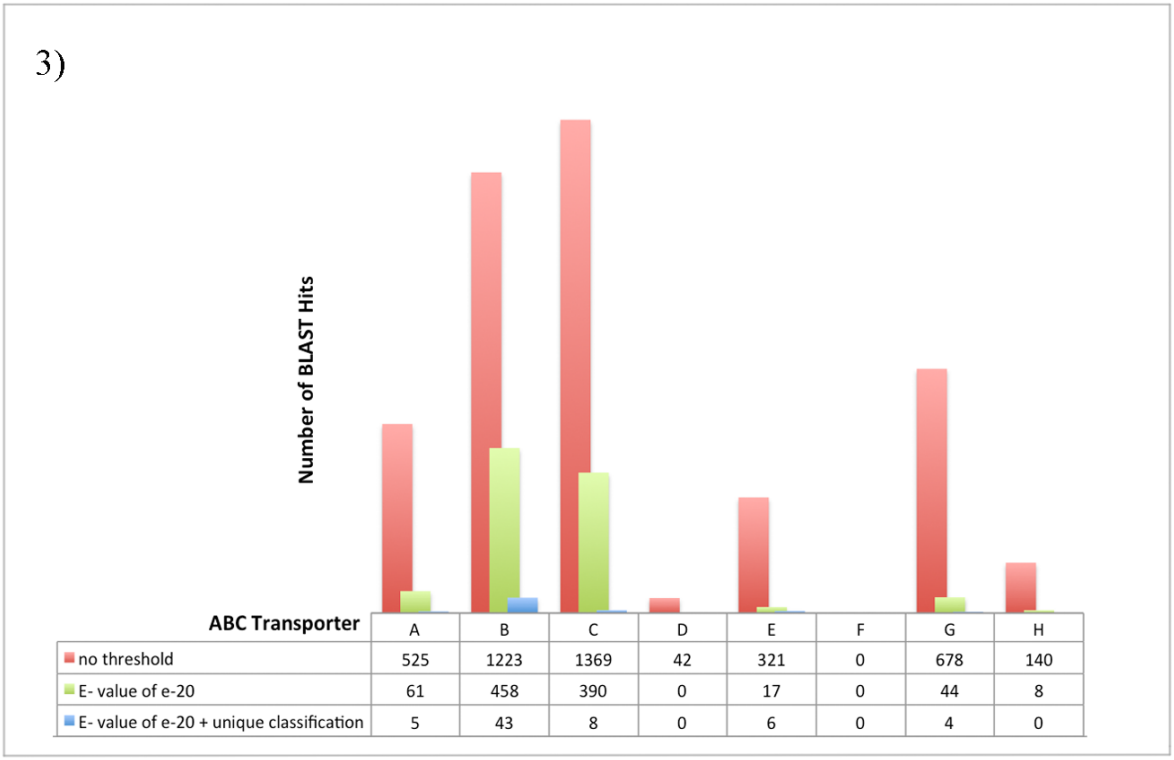


ABC Transporter	A	B	C	D	E	F	G	H
no threshold	343	735	776	57	207	0	538	101
E- value of e-20	38	290	244	0	25	0	47	6
E- value of e-20 + unique classification	4	32	10	0	12	0	6	0

2)



ABC Transporter	A	B	C	D	E	F	G	H
no threshold	501	1085	1158	41	340	0	606	121
E- value of e-20	62	358	319	0	17	0	62	8
E- value of e-20 + unique classification	5	30	9	0	9	0	7	0



ABC Transporter	A	B	C	D	E	F	G	H
no threshold	525	1223	1369	42	321	0	678	140
E- value of e-20	61	458	390	0	17	0	44	8
E- value of e-20 + unique classification	5	43	8	0	6	0	4	0

Figure 5.6.: Conducted BLAST searches (*tblastn*) against the ABC transporter sequences of *Drosophila melanogaster* using 1) the 19929, 2) 19931 and 3) the 19932 *Fasciola hepatica* dataset as query. The red bars indicate a BLAST analysis with no statistical significance threshold resulting in larger numbers of hits. By introducing a threshold of e^{-20} the hits reduces enormously (green), displaying only sequences with high similarity. ABC transporters of *Fasciola hepatica* which have commonalities with e.g. A and B transporter are listed in both. The data of the unique classification where each liver fluke transporter is assigned to only one ABC transporter (A-H) is displayed by the blue bars. The dataset 19930 was excluded from analysis because of the suspicion of being a triploid form.

Organism	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Caenorhabditis elegans	7	24	9	5	1	3	9	2
Drosophila melanogaster	10	10	12	2	1	3	15	3
Homo sapiens	12	11	12	4	1	3	5	0

Table 5.2.: The number of ABC transporters of the organism the liver fluke was compared with and the composition in each ABC subfamily. The evolutionary most closely related nematode possesses its majority of transporter in the ABCB subfamily, consistent with the putative amount of ABC transporter in the liver fluke.

The vast majority of ABC transporters identified in *Fasciola hepatica* are in the ABCB subfamily ($\sim 50\%$), followed by the ABCC subfamily ($\sim 14\%$). Furthermore, no transporters of the subfamily H could be found.
After visualization and analysis of the parasite sequences in Geneious, the dataset turned out to be fragmented. Therefore, the numbers of ABC transporters of the liver flukes are probably a bit lower in reality because of several transporter fragments, which are constituents of the same transporter are counted separately.
The high amount of hits received for the subfamily D is due to the detection of many splice variants of the contigs *comp28795_c0* and *comp28985_c1* (Table A.5 and A.3). The splicing process removes introns from the mRNA and joins the exons to create mature mRNA. It may also lead to several mature mRNAs from one mRNA, resulting in several proteins from a single gene termed "alternative splicing variants". [82]

The dataset 19930 delivers results of expression and polygenetic analyses different to the other 3 *Fasciola hepatica* datasets. Probably the fluke is triploid[4] and to avoid errors, the dataset 19930 was completely excluded from further analyses.

In three cases the contigs showed a high similarity to two different subfamilies (ABCB and ABCC). Contig *comp106670_c1_seq16* is classified as a transporter of subfamily ABCB (E value of $4,00e^{-27}$) but delivers also for subfamily C an E value of $4,00e^{-26}$. Additionally, contig *comp106670_c1_seq17* and *comp27410_c0_seq1* are assigned to subfamily C and show high similarity to the members of subfamily B. To classify these exceptional cases correctly additional parameters such as the bit score are considered. However, all the other contigs could be assigned clearly to the single subfamilies by the E value.
The ABC subfamilies are further subdivided into subgroups concerning their expression and function. This classification based on the comparison of the fully described NCBI sequences of the ABC transporter of *Homo sapiens* with the dataset. The results of the identified ABC transporter subgroups in the liver flukes can be found in Table 6.4.

[4] 3 sets of chromosomes

5.4.1. Characteristic Nucleotide Binding Domain Motifs

Motifs are short, conserved regions of peptide or nucleic acid sequences, which possess specific functions. By producing a multiple alignment of distant related sequences, gaps are necessary to arrange the alignment in a correct way. Accordingly, 'islands' of conservation tend to appear surrounded by mutational change. These conserved regions (motifs) are helpful in comparing widely different genomes and to conclude on their common functions or appearance. [42]

A full ABC transporter typically shows a length of 200 to 220 amino acids, including 2 NBDs and 2 TMDs (Figure 5.7) [7]. In contrast to the TMDs, the NBDs are highly conserved and composed of 3 (Walker A, Walker B and the signature) characteristic protein sequence motifs that are involved in binding and hydrolyzing ATP. Previous reports indicated that intact ATP is preferentially bound at NBD1, whereas trapping of the ATP hydrolysis product, ADP, occurs predominantly at NBD2 [38]. Therefore, ATP interaction with NBD1 increases ATP or ADP binding at NBD2 [49].

The Walker A motif, also known as P loop (phosphate binding loop) consists of 8 amino acids of following pattern GXXGXGKS/T[5] [62]. It is considered to be a fundamental and ancient functional motif in biological systems [57]. Downstream the ABC transporter signature[6] motif LSGGQ is located followed by the Walker B motif.
Surveys of studying mutations of the signature motif in bacterial ABC proteins suggest, that it is involved in ATP hydrolysis and not in ATP binding leading to a change in ATPase activity [93] [89]. In addition, its consensus sequence is highly conserved and only few variations (e.g. Q to E in BtuD) could be detected. The signature motif is a unique sequence characterizing all ABC transporters and distinguishes them from other proteins containing the NBD. [84]
The Walker B motif is composed of 7 amino acids resembling $\phi\phi\phi$DEXX where ϕ represent any hydrophobic residue. The amino acids XX are not well conserved but often AT forming the "DEAT box" [62].

In the liver fluke contigs of subfamily ABCB and ABCC transporters the motifs Walker A, Walker B and the signature were identified to detect mutations.
Basically, only one transporter consisting of 2 NBDs could be found. Otherwise the contigs contain only 1 NBD resembling a half transporter or representing only a fragment of the transporter sequence. It may indicate partially sequenced cDNA or maybe the NBDs are encoded on different polypeptides.
For the ABCC subfamily 5 variations of the Walker A and Walker B and 3 variations of the signature could be detected. The liver fluke contigs of ABCB contain 7 variations of Walker A, 5 of Walker B and 6 variations of the signature. The signature mutation LSGGE, known from BtuD, appeared frequently.

[5] G, K, T, S and X denote glycine, lysine, threonine, serine and an arbitrary amino acid.
[6] The ABC transporter signature is also called the C motif.

However, the C motif is heavily degenerated not only for the resistant liver fluke dataset of 19931 but also for 19929 and 19932.
The consensus sequence of Walker A was highly conserved, only threonine often mutates to serine or to alanine.
For Walker B either the "DEAT box" could be identified or the amino acid sequence DDPL[7]. The D loop located right after the Walker B showed 3 variations, LD, VD and DP. All variations of the motifs found in the *Fasciola hepatica* datasets in ABCB and ABCC are listed in Table 6.2 and Table 6.3

[7]DDP, represents aspartic acid D and proline P

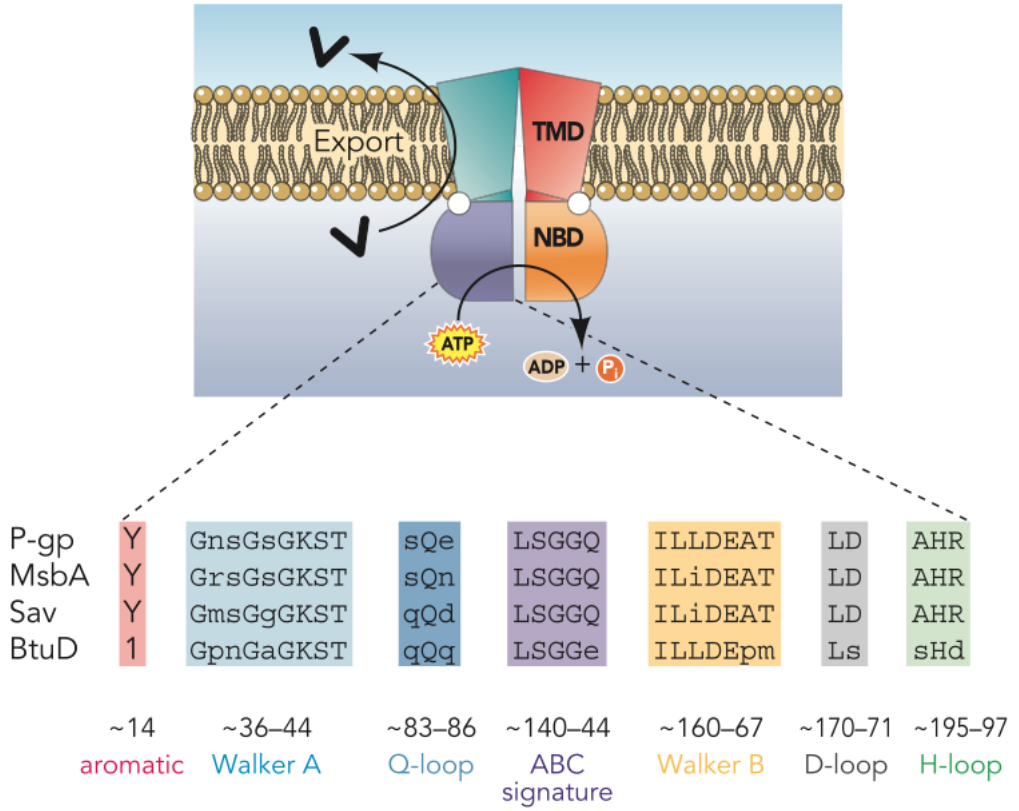


Figure 5.7.: General protein structure of a typical full ABC transporter consisting of 2 transmembrane domains (TMDs) and 2 nucleotide binding domains (NBDs). The conserved NBDs consist of different characteristic motifs represented in the Figure indicating their common protein sequences listed for different ABC multidrug transporter. The P - glycoprotein is associated with multidrug resistance and belongs to the human ABCB subfamily. MsbA is a MDR efflux ABC transporter representing an ATPase found in bacteria. Sav was detected in *Staphylococcus aureus* and is a homolog of MDR ABC transporter like BtuD, the vitamin B12 importer from *Escherichia coli*. [64]

5.5. Multiple Alignment Geneious

Contigs of the *Fasciola hepatica* datasets could be detected with the aid of BLAST, which show high similarity to the subfamilies of the ABC transporter of *Homo sapiens*, the fruit fly and the nematode. Furthermore, they get classified into subfamilies A to G and imported in Geneious for further analyses. Geneious allows to group the single contigs into lists or to keep the sequences separately. In the Source Panel, folders named from A to H were produced representing the individual subfamilies. These folders include the contigs, which produce significant hits grouped by the organism they were compared with. The sequences were imported as sequence lists named with the letter of the subfamily, followed by the *Fasciola hepatica* dataset name and the organism the comparison was conducted. For example, A_19931_HomoSapiens include all contigs of subfamily A of the dataset 19931, which produced hits by comparing it with the ABC transporters of *Homo sapiens*.

For each ABC subfamily and liver fluke dataset a multiple sequence alignment was performed using the incorporated Geneious logarithm. Either the BLOSUM matrices were used as cost matrix to align protein sequences or match/mismatch costs for nucleotide sequences. Geneious indicates the target sequence similarity for the alignment scores. Thus, it determines the amount of similarity between the sequences for which those scores are optimal. Furthermore, both protein and nucleotide alignments have the opportunity to introduce gap open/gap extension penalties/costs. [67]
Basically, the gap open penalty was set to twelve and the extension penalty to three.
In addition, the alignment with free end gaps was chosen to avoid that gaps at either end of the alignment are penalized.
After visualization of the sequences it could be seen that some of the transporter are fragmented. In Figure 5.8 an alignment of the ABCC subfamily contigs of the dataset 19932 is displayed, pointing out that the last 4 contigs from bp 3.800 to bp 9.275 are probably constituent of one transporter.

The contigs of one subfamily of all *Fasciola hepatica* datasets are then compared to identify SNP regions and mutations. For ABCB around 100 sequences get aligned simultaneously, which makes it really difficult to detect dissimilarities. Therefore, the contigs are further grouped visually in subfamilies concerning their similarity and are aligned again. Also genetic variations of transporters within a subfamily could be discovered.
Figure 5.9 pictures all contigs of the ABCG subfamily of all liver fluke datasets. The differences of the contigs are visible as black bars in the grey sequence representation. By zooming into the sequences (Figure 5.10) two transporter variations become visible. We tried to reduce the amount of sequences to a minimum by keeping only one sequence for each transporter variation. The sequences within one transporter variation may not be of the same length and to get a - possibly full length sequence - consensus sequences are generated. With the reduced amount of sequences the contigs of one subfamily become comparable. This enables the detection of differentially expressed genes or SNPs leading to functional changes between the triclabendazole resitant versus sensitive individuals.

Several SNPs could be detected in the ABC transporter Subfamilies B,C, E and G. The putative B7 transporter shows 12 amino acid differences between the sensitive (19932) and the resistant (19931) *Fasciola hepatica* contigs, including 5 transversions and 7 transitions. The transversion cytosine to adenine downstream of Walker B definitely leads to a protein change. Furthermore, in B10 one transition could be identified located between Walker A and the ABC signature, suggesting that it may have functional significance. Both contigs of the sensitive liver flukes possess a thymine in contrast to the resistant one, which expresses a cytosine.
For comparison with the ABC transporter of *Drosophila melanogaster* and *Caenorhabditis elegans*, 3 contigs[8] showing many expression differences between all liver fluke datasets were found.
In subfamily C a putative C2 transporter was identified indicating 4 transitions. These SNPs are found not only by the comparison of ABC transporter of *Homo sapiens* but also the other 2 organism.
In addition, a single SNP occurred in ABCE, not known to be involved in multidrug resistance.
For ABCG, 2 subfamilies G1 and G2 are found in the liver fluke contigs showing SNPs. G1 contains 1 transition from thymine to cytosine. In G2 two SNPs right next to each other are identified not found in any other subfamily. The contig of 19931 possesses an adenine and a thymine, whereas the contigs of 19929 and 19932 contain a thymine and a cytosine. The same pattern is shown for the comparison with the nematode sequences, but in different liver fluke contigs with a different nucleotide exchange. Again the contig of the resistant dataset differs in two subsequent nucleotides from the contigs of the sensitive individuals (GG changes to TC).
All identified nucleotide differences between the resistant individual and the sensitive flukes are listed in Table 6.5. They may lead to a protein change and further to a functional alteration of the transporter. The exact location and their significance concerning TCBZ resistance still has to be investigated.

[8] *comp30171_c0_seq5, comp25903_c0_seq3, comp104119_c1_seq2*

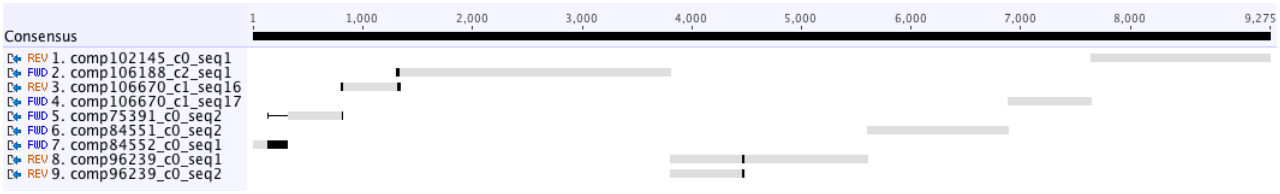


Figure 5.8.: Multiple alignment of the identified ABCC contigs of *Fasciola hepatica* dataset 19932 established with Geneious. The grey bars represent the sequences whereas differences are marked black. The 4 contigs from bp 3800 to bp 9275 seem to be the constituents of one transporter implicating a fragmented dataset.

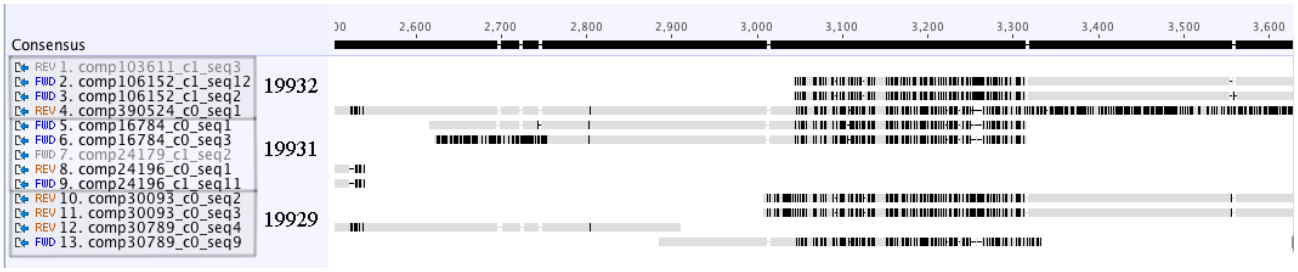


Figure 5.9.: Multiple alignment of the ABCG Transporter of all *Fasciola hepatica* datasets (19929, 19931 and 19932) conducted with Geneious. The differences between the contigs are visible as black bars in the grey sequence representation. On closer inspection (Figure 5.10) 2 transporter variations could be identified.

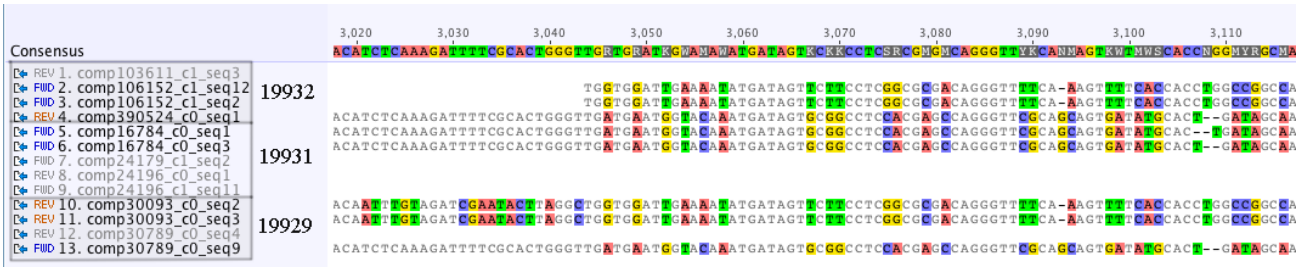


Figure 5.10.: Multiple alignment of the ABCG Transporter of all *Fasciola hepatica* datasets (19929, 19931 and 19932) conducted with Geneious. The detailed view enables the identification of 2 different variations of the transporter within the subfamily G. The number of contigs for each transporter was reduced to the minimum on variations.

6. Results

Comparative analyses have been conducted with the Basic Local Alignment Search tool BLAST (chapter 3) and the sequence alignment tool Geneious (section 4.5). The received results suggest how much and which ABC transporter are available in the liver fluke (Table 6.1 and 6.4). The assignment of the single contigs of *Fasciola hepatica* to the subfamilies of the ABC transporter can be found in the Appendix (A.1). In addition, the characteristic motifs of the transporter including their variations are identified and listed (Table 6.2 and 6.3). The detected SNPs (Table 6.5) may affect the function of the ABC transporter leading to drug resistance. Their exact location in the protein and functional significance still has to be investigated.

Organism	**Sample ID**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Caenorhabditis elegans	19929	4	31	13	21	3	7	4	0
	19931	5	30	10	3	5	6	5	0
	19932	6	41	5	0	2	5	4	0
Drosophila melanogaster	19929	4	32	10	0	12	0	6	0
	19931	5	30	9	0	9	0	7	0
	19932	5	43	8	0	6	0	4	0
Homo sapiens	19929	7	35	16	18	1	6	4	0
	19931	5	33	12	2	2	4	16	0
	19932	8	46	6	3	2	5	15	0

Table 6.1.: The liver fluke datasets were compared with ABC transporter sequences of the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster* and the *Homo sapiens*. The number of contigs detected by the use of the local alignment search tool BLAST for each subfamily of ABC transporter of *Fasciola hepatica* are listed above. All contigs could be clearly assigned to one subfamily. By comparing these results to Table 5.2, it has to be considered that each *Fasciola hepatica* dataset contains the cDNA of two flukes. Moreover, the amount of detected contigs are probably a little bit higher than in reality, because the dataset is fragmented and constituents of one transporter are probably counted separately.

Dataset	Walker A	Signature	Walker B	D loop
19929	GQSGCGKST	LSGGQ	LLLDEAT	LD
	GPSGCKST	LSGGE	FLLDEAT	
		LSAGQ	LIYDEAT	
19931	GQSGCGKST	LSSGQ	LLLDEAT	LD
	GGSGAGKST	LSGGE	FLLDEAT	
	GHSGCGKTS	YGGGQ	LIYDEAT	
	GPSGSGKST		ILLDEAT	
	GQSGAGKST			
19932	GQSGCGKST	LSGGQ	LLLDEAT	LD
	GGSGAGKST	LSGGE	ILLDEAT	
	GPNGSGKST	LSSGQ	LIYDEAT	
	GPSGSGKST	LSVGQ	LILDEAT	

Table 6.2.: Identified motifs detected in the contigs of *Fasciola hepatica* subfamily ABCB. The findings refer to the contigs received by the comparison with the nematode *Caenorhabditis elegans*. The Walker A normally resembling GXXGXGKS/T, whereas X denotes an arbitrary amino acid, is highly conserved. Only the known mutation of threonine to serine could be found frequently. In the putative resistant liver fluke (19931) a permutation of the last two amino acids threonine and serine could be observed. The signature motif LSGGQ often appears as LSGGE and is partially highly degenerated. For Walker B, 16 amino acids downstream of the signature, the characteristic DEAT motif could be identified followed by the conserved D loop LD.

Dataset	Walker A	Signature	Walker B	D loop
19929	GQSGAGKST	LSGGE	LVLDEAT	LD
	GGSGAGKST		ILLDEAT	
	GRTGSGKSS			
19931	GTVGSGKSS	LSGGQ	LVLDEAT	VD
	GPVGSGKSA	FSTGQ	YLLDDPL	
	GRTGSGKSS			
19932	GTVGSGKSS	LSGGQ	LVLDEAT	LD
	GPVGSGKSA	FSTGQ	YLLDDPL	VD
	GRTGSGKSS		LVVDEAT	DP

Table 6.3.: Identified motifs of the liver fluke subfamily ABCC detected in the contigs found by comparison with the ABC transporters of *Caenorhabditis elegans*. The last amino acid of Walker A deviants at times from threonine or serine to alanine. About 99 amino acids downstream 2 mutations of the characteristic ABC signature motif could be identified. Like in the subfamily ABCB the LSGGE mutation occur frequently in contrast to the FSTGQ mutation only found in ABCC. The Walker B motif shows either the common "DEAT box" or the mutation DDPL. Also variations of the D loop could be identified.

ABC Subfamily	19929	19931	19932
A	A1	A1	A1
	A3	A3	A3
	A4	A4	A4
	A8	A8	A8
	A13		A13
B	B1	B1	B1
	B4		B4
	B6	B6	B6
	B7	B7	B7
		B8	B8
	B10	B10	B10
C	C1	C1	
	C2	C2	C2
	C3	C3	C3
	C6		
D	D4	D4	D4
E	E1	E1	E1
F	F1	F1	F1
	F3		F3
G	G1	G1	G1
	G2	G2	G2

Table 6.4.: Identified subfamilies of *Fasciola hepatica* based on the comparison with the ABC transporter of *Homo sapiens*. All three transporters (ABCB1, ABCC2 and ABCG2) known to be involved in drug resistance seem to be present in the liver fluke.

ID	ABC	Contig	SNP													
19931	B7	comp25017_c0_seq1	A	G	C	C	C	G	T	A	G	G	C	T		
19932		comp102748_c0_seq2	G	A	T	G	G	C	C	C	C	A	T	C		
19929	B10	comp25157_c0_seq1	T													
19931		comp25345_c3_seq5	C													
19932		comp105929_c0_seq3	T													
19929	B	comp29030_c0_seq7	C													
19931		comp23802_c0_seq1	T													
19932		comp101188_c0_seq1	C													
19929	B	comp30171_c0_seq5	C	T	T	A	C	G	G	C	G	T	A	C	C	G	C
19931		comp25903_c0_seq3	G	T	C	G	T	A	T	T	A	C	G	C	T	G	T
19932		comp104119_c1_seq2	G	C	T	A	T	A	T	C	G	C	A	T	C	A	T
19929	B	comp29483_c2_seq2	A													
19931		comp24532_c0_seq1	C													
19932		comp99094_c1_seq1	A													
19929	C2	comp27910_c0_seq2	T	A	A	C										
19931		comp24084_c1_seq1	C	G	G	T										
19932		comp102145_co_seq1	T	A	A	C										
19929	C	comp27910_c0_seq2	T	G												
19931		comp24084_c1_seq3	C	A												
19932		comp102145_c0_seq1	T	A												
19929	E	comp30906_c0_seq5	G													
19931		comp26245_c5_seq4	A													
19932		comp103624_c1_seq1	G													
19929	G1	comp30789_c0_seq4	T													
19931		comp24196_c0_seq1	C													
19932		comp103611_c1_seq3	T													
19929	G2	comp29155_c0_seq1	T	C												
19931		comp24811_c1_seq12	A	T												
19932		comp103372_c0_seq1	T	C												
19929	G	comp30789_c0_seq4	A													
19931		comp24196_c1_seq11	G													
19932		comp103611_c1_seq3	A													
19929	G	comp30789_c0_seq4	G	G	T											
19931		comp16784_c0_seq1	T	C	C											
19932		comp390524_c0_seq1	G	G	T											

Table 6.5.: Nucleotide differences between the putative resistant *Fasciola hepatica* individual (19931) and the sensitive liver flukes (19929 and 19932). The contigs are aligned in Geneious and all detected SNPs may lead to a protein change with functional significance.

7. Discussion

In order to compare drug sensitive flukes (dataset 19929, 19930, 19932) to the putative resistant fluke (dataset 19931) genomic DNA was prepared and sequenced by next generation sequencing methods to obtain transcriptomes reflecting ABC transporter of *Fasciola hepatica* adults.
The liver fluke database was searched using protein sequences corresponding to 48 ABC transporters of *Homo sapiens* as well as 56 and 60 nucleotide sequences of the fruit fly *Drosophila melanogaster* and the nematode *Caenorhabditis elegans*.
The Basic Local Alignment Search Tool, BLAST, provides a suitable tool for the identification of specific gene sequences as long as they don't get too short. The heuristic search algorithm enables to handle high amount of data in acceptable calculation time.
The liver fluke sequences are compared to already characterized ones to detect sequence similarity and to infer both the structure and the function of the genes. [11]
By setting an appropriate cutoff threshold concerning the evolutionary distance of the sequences significant hits may be found. Thus, the identification of ABC transporters and further the classification of the contigs to the corresponding subfamily of the transporters is possible with BLAST.

In the case of the *Fasciola hepatica* database, the ABCB subfamily clearly contain the most ABC transporters followed by the subfamily ABCC. Consistent to this, in *Caenorhabditis elegans*, the evolutionary closest organism the liver fluke was compared with, also subfamily B clearly presents the largest subfamily.
Tblastx and *tblastn* BLAST analyses revealed, that the liver fluke transporters represent seven subfamilies from ABCA to ABCG.
For subfamily H, present in the nematode and *Drosophila melanogaster*, no hits could be found with BLAST by setting an E value of e^{-20}.

However, the number of identified putative ABC transporters in Table 6.1 may be over-represented, because the database was shown to be fragmented and the partial transcripts are probably counted separately. Moreover, each *Fasciola hepatica* dataset contains the cDNA of two flukes. These two facts make it rather difficult to compare the obtained numbers of putative ABC transporters to the amount of transporter present in *Homo sapiens*, the fruit fly and the nematode (Table 5.2).
Nevertheless, a first estimation of the number of ABC transporters available in *Fasciola hepatica* could be achieved by the use of BLAST.

Based on the comparison of the fully described ABC transporter sequences of *Homo sapiens* the subdivision of the single ABC subfamilies concerning their expression and function was conducted. All known multidrug efflux pumps involved in protecting tissue from toxic xenobiotics and endogenous metabolites (ABCB1, ABCC2 and ABCG2) seem to be present in the liver fluke.

The first time identification of the biologically and functionally important motifs, which uniquely characterizes ABC transporter is not recommendable to conduct with BLAST. The motifs normally consist of only 5 to 10 amino acids and are partially degenerated. To find unknown mutations of the motifs, it is useful to completely visualize the sequences with an sequence alignment program. Geneious not only provides a user-friendly interface to organize and visualize the sequences but also different incorporated sequence alignment algorithm enabling the validation of the alignment without data shifting. The option to produce dot plots gives a first impression of the sequence similarity and with the text and alignment view further analyses and the detection of SNPs can be conducted. The Geneious algorithms are also available in the test version whereas for the use of MUSCLE, CLUSTALW or Realign algorithm a license is needed. It further offers a translational align algorithm to test if a detected amino acid mutation leads to a protein change and enables to determine the significance of the SNP right away.
The investigation of the motifs delivers essential knowledge about the amount and structure of the ABC transporter domains. SNPs in these regions may lead to drastic changes of the transporter function and may play a role in developing drug resistance.

Comparative analysis of the contigs in triclabendazole resistant versus sensitive individuals were conducted with Geneious to get insight into the nature of the adaptive changes causing drug resistance.
Because further studies prove that ABCB and ABCC transporters are mainly involved in drug resistance we concentrated especially on these two subfamilies to survey the motifs and identify single nucleotide polymorphisms.

Basically, only one transporter could be found, which consists of 2 complete NBDs including the Walker A, followed by the ABC signature and the Walker B. Otherwise the contigs contain one NBD, parts of it or even no NBD, arguing again for fragmented parts of transporters or maybe also for partially sequenced cDNA.
Each Walker A motif showed the universally conserved lysine and glycine. Only the known variation of threonine to serine occur frequently and in some cases the threonine/serine changes to an alanine.
About 99 amino acids downstream of the Walker A, the signature motif is located, showing partially nonconservative substitutions. The LSGGE variation, known from the MDR transporter BtuD, the vitamine B12 importer of *Escherichia coli*, appeared several times in the liver fluke datasets.
Not rarely 2 of 5 amino acids in the normally highly conserved signature motif are degenerate.

The Walker B is separated by 16 amino acids from the ABC signature and often shows the typical DEAT motif. Only in subfamily C variations from DEAT to DDPL[1] could be detected. The motifs are involved in binding and hydrolyzing ATP and therefore are very important for the function of the ABC transporter. Although no significant differences between the motifs of the resistant and the sensitive *Fasciola hepatica* datasets could be found, mutations in these regions may have functional significance and should be further investigated.

However, several SNPs could be identified irrespectively of the walker motifs, in the ABC transporter subfamilies B,C, E and G.
In ABCB7, a difference between the putative resistant liver fluke dataset 19931 and the sensitive one 19932 definitely leading to a protein change could be detected. All other discovered SNPs may lead to a protein change and thus to a functional alteration of the transporter. The right translation frame and the exact positions of the mutations still have to be identified to evaluate the significance of the SNPs.
Transporter G2 showed twice, two subsequent nucleotide changes not seen in any other subfamily. Recent studies suggest that the up regulated expression of ABCG2 in Oct - 3/4 expressing cells result in higher resistance to chemotherapeutic drugs. [48]
Therefore, this transporter may play an important role beside transporter B and C in developing drug resistance.

[1]Glutamic acid, alanine and threonine changes to aspartic acid proline and leucine.

8. Conclusion

Comparative analyses of drug sensitive versus resistant ABC transporter sequences of *Fasciola hepatica* were conducted to identify single nucleotide polymorphisms. The local alignment tool BLAST was used to compare cDNA contigs of the trematode to a database consisting of ABC transporter of 3 characterized organisms. This resulted in new information about the ABC transporters available in the liver fluke. It could be shown, that all known multidrug efflux pumps (ABCB1, ABCC2 and ABCG2) are expressed in *Fasciola hepatica*. Additionally, single nucleotide polymorphisms were detected in ABCC2 and ABCG2 probably leading to functional changes of the transporters. The ABCG2 transporter showed two subsequent nucleotide changes not discovered in any other transporter. Irrespectively of the known MDR efflux pumps, several SNPs could be found in the ABC transporter subfamilies B, C, E and G by the use of the multiple alignment program Geneious. Furthermore, investigations of the functionally important sequence motifs discovered highly degenerated motifs most likely leading to functional changes.
The exact positions of the SNPs still have to be identified and functional studies will elucidate the significance of the mutations. All these results provide new insights into the diversity of the entire ABC subfamilies of the liver fluke and may deliver knew knowledge about the adaption mechanisms the parasites have developed to survive drug exposure.

Bibliography

[1] Austrian Institute of Technology, department Health and Environment, Bioresources.

[2] R. Mach, Technical University of Vienna.

[3] Witschnitzki E., Austrian Institute of Technology, department Health and Environment, Bioresources.

[4] Distance Methods, DNA and Protein Models. Department of Genome Sciences, University of Washington. pages 1–34, 2012.

[5] About Education, Gene Mutation. http://biology.about.com/od/basicgenetics/ss/genemutation.htm.

[6] Ahcene Boumendjel, Jean Boutonnat, Jacques Robert. *ABC Transporters and Multidrug Resistance*. 2009.

[7] Suresh V. Ambudkar, In-Wha Kim, Di Xia, and Zuben E. Sauna. The A-loop, a novel conserved aromatic acid subdomain upstream of the Walker A motif in ABC transporters, is critical for ATP binding. *FEBS Lett.*, 580(4):1049–55, 2006.

[8] S. Anderson, P. S. Coulson, S. Ljubojevic, A. P. Mountford, and R. A. Wilson. The radiation-attenuated schistosome vaccine induces high levels of protective immunity in the absence of B cells. *Immunology*, 96(1):22–28, 1999.

[9] Keyhan Ashrafi, M. Adela Valero, Raquel V. Peixoto, Patricio Artigas, Miroslava Panova, and Santiago Mas-Coma. Distribution of Fasciola hepatica and F. gigantica in the endemic area of Guilan, Iran: Relationships between zonal overlap and phenotypic traits. *Infect. Genet. Evol.*, 31(0):95–109, 2015.

[10] Information Resources Management Association. *Bioinformatics: Concepts, Methodologies, Tools, and Applications*.

[11] Marina Axelson-Fisk. *Comparative Gene Finding Models, Algorithms and Implementation*. 2015.

[12] M. D. Bargues, M. A. Valero, and S. Mas-Coma. Fascioliasis and other plant-borne trematode zoonoses. 35:1255–1278, 2005.

[13] Andreas Baxevanis and B. F. Francis Ouellette. *Bioinformatics - A Practical Guide to the Analysis of Genes and Proteins*, volume XXXIII. 2001.

[14] Joseph Bedel, Ian Kor, and Mark Yandel. *BLAST*, volume 53. 2003.

[15] S. Bennema, J. Vercruysse, E. Claerebout, T. Schnieder, C. Strube, E. Ducheyne, G. Hendrickx, and J. Charlier. The use of bulk-tank milk ELISAs to assess the spatial distribution of Fasciola hepatica, Ostertagia ostertagi and Dictyocaulus viviparus in dairy cattle in Flanders (Belgium). *Vet. Parasitol.*, 165(1-2):51–57, 2009.

[16] Berg J.M., Tymoczko J.L., Stryer L. *Biochemistry, Amino Acids Are Encoded by Groups of Three Bases Starting from a Fixed Point, http://www.ncbi.nlm.nih.gov/books/NBK22358/.*

[17] G. P. Brennan, I. Fairweather, A. Trudgett, E. Hoey, McCoy, M. McConville, M. Meaney, M. Robinson, N. McFerran, L. Ryan, C. Lanusse, L. Mottier, L. Alvarez, H. Solana, G. Virkel, and P. M. Brophy. Understanding triclabendazole resistance. *Exp. Mol. Pathol.*, 82(2):104–109, 2007.

[18] C. Calléja, K. Bigot, C. Eeckhoutte, P. Sibille, C. Boulard, and P. Galtier. Comparison of hepatic and renal drug-metabolising enzyme activities in sheep given single or two-fold challenge infections with Fasciola hepatica. *Int. J. Parasitol.*, 30(8):953–8, 2000.

[19] Supratim Choudhuri. *Bioinformatics for Beginners Genes, Genomes, Molecular Evolution, Databases and Analytical Tools.* 2014.

[20] Jean-Michel Claverie and Cedric Notredame. *Bioinformatics for Dummies*, volume 1. 2007.

[21] Gerald C Coles and George K Kinoti. Defining resistance in Schistosoma. *Parasitol. Today*, 13(4):157–158, 1997.

[22] Krystyna Cwiklinski, Katherine Allen, James LaCourse, Diana J. Williams, Steve Paterson, and Jane E. Hodgkinson. Characterisation of a novel panel of polymorphic microsatellite loci for the liver fluke, Fasciola hepatica, using a next generation sequencing approach. *Infect. Genet. Evol.*, 32:298–304, 2015.

[23] Michael Dean. The human ATP-binding cassette (ABC) transporter superfamily. *J. Lipid Res.*, 42(7):1007–17, 2001.

[24] Paul H. Dear. Bioinformatics. 2007.

[25] Department of Statistics University of Oxford. Basic Models of Nucleotide Evolution. Technical report, 1993.

[26] Eileen Devaney. Genetic and genomic approaches to understanding drug resistance in parasites. *Parasitology*, 140(12):1451–4, 2013.

[27] Gonze Didier. PAM and BLOSUM substitution matrices Computing Centre Universiteit Brussel. pages 1–9.

[28] Dreyfuss G., Rondelaud D. A study of the shedding of cercariae from Lymnaea truneatula raised under constant conditions of temperature and photoperiod. 1994.

[29] Martin Dugas and Karin Schmidt. *Medizinische Informatik und Bioinformatik*. 2003.

[30] Jacques Dupuy, Michel Alvinerie, Cecile Ménez, and Anne Lespine. Interaction of anthelmintic drugs with P-glycoprotein in recombinant LLC-PK1-mdr1a cells. *Chem. Biol. Interact.*, 186(3):280–286, 2010.

[31] Dannie Durand. Markov Models of Sequence Carnegie Mellon University School of Computer Science. pages 1–7, 2014.

[32] David Edwards, Jason Stajich, and David Hansen. *Bioinformatics, Tools and Applications*, volume XXXIII. 2009.

[33] T. P. Elliott, J. M. Kelley, G. Rawlin, and T. W. Spithill. High prevalence of fascioliosis and evaluation of drug efficacy against Fasciola hepatica in dairy cattle in the Maffra and Bairnsdale districts of Gippsland, Victoria, Australia. *Vet. Parasitol.*, 209(1-2):117–24, 2015.

[34] Entelechon. http://www.entelechon.com/2008/08/blast-parameters.

[35] FABRE Technology Platform Working Group. Sustainable Farm Animal Breeding and Reproduction. A Vision for 2025. www.fabretp.org.

[36] I. Fairweather. Liver fluke isolates: a question of provenance. *Vet. Parasitol.*, 176(1):1–8, 2011.

[37] C. P. H. Gaasenbeek, L. Moll, J. B. W. J. Cornelissen, P. Vellema, and F. H. M. Borgsteede. An experimental study on triclabendazole resistance of Fasciola hepatica in sheep. *Vet. Parasitol.*, 95:37–43, 2001.

[38] M. Gao, H. R. Cui, Douglas W. Loe, Caroline E. Grant, Kurt C. Almquist, Susan P. Cole, and Roger G. Deeley. Comparison of the Functional Characteristics of the Nucleotide Binding Domains of Multidrug Resistance Protein 1. *J Biol Chem*, 275(17):13098–13108, 2000.

[39] Generation Challenge Programme - Partnerships in modern crop breeding for food security. http://www.generationcp.org/.

[40] Landesregierung Fachabteilung Gesundheit and Pflegemanagement Veterin. Veterinärbericht. 2014.

[41] Robert M. Greenberg. Schistosome ABC multidrug transporters: From pharmacology to physiology. *Int. J. Parasitol. Drugs drug Resist.*, 4(3):301–9, 2014.

[42] John M. Hancock and Marketa J. Zvelebil. *Concise Encyclopaedia of Bioinformatics and Computational Biology*, volume 53. 2014.

[43] R. E. B. Hanna, C. McMahon, S. Ellison, H. W. Edgar, P.-E. Kajugu, A. Gordon, D. Irwin, J. P. Barley, F. E. Malone, G. P. Brennan, and I. Fairweather. Fasciola hepatica: a comparative survey of adult fluke resistance to triclabendazole, nitroxynil and closantel on selected upland and lowland sheep farms in Northern Ireland using faecal egg counting, coproantigen ELISA testing and fluke histology. *Vet. Parasitol.*, 207(1-2):34–43, 2015.

[44] R.E.B. Hanna, H.W.J. Edgar, S. McConnell, E. Toner, M. McConville, G.P. Brennan, C. Devine, A. Flanagan, L. Halferty, M. Meaney, L. Shaw, D. Moffett, M. McCoy, and I. Fairweather. Fasciola hepatica: Histological changes in the reproductive structures of triclabendazole (TCBZ)-sensitive and TCBZ-resistant flukes after treatment in vivo with TCBZ and the related benzimidazole derivative, Compound Alpha. *Vet. Parasitol.*, 168(3-4):240–254, 2010.

[45] R.E.B. Hanna, F.I. Forster, G.P. Brennan, and I. Fairweather. Fasciola hepatica: Histological demonstration of apoptosis in the reproductive organs of flukes of triclabendazole-sensitive and triclabendazole-resistant isolates, and in field-derived flukes from triclabendazole-treated hosts, using in situ hybridisatio. *Vet. Parasitol.*, 191(3-4):240–251, 2013.

[46] Andrea Hansen. *Bioinformatik Ein Leitfaden für NaturwissenschaftIer*. 2001.

[47] Charlie Hodgman, Andrew French, and David Westhead. *Bioinformatics*, volume XXXIII. 2010.

[48] Y. Hosokawa, H. Takahashi, A. Inoue, Y. Kawabe, Y. Funahashi, K. Kameda, K. Sugimoto, H. Yano, H. Harada, S. Kohno, S. Ohue, T. Ohnishi, and J. Tanaka. Oct-3/4 modulates the drug-resistant phenotype of glioblastoma cells through expression of ATP binding cassette transporter G2. *Biochim Biophys Acta*, 1850(6):1197–1205, 2015.

[49] Yue Xian Hou, Liying Cui, John R. Riordan, and Xiu Bao Chang. ATP binding to the first nucleotide-binding domain of multidrug resistance protein MRP1 increases binding and hydrolysis of ATP and trapping of ADP at the second domain. *J. Biol. Chem.*, 277(7):5110–5119, 2002.

[50] Illumina. Quality Scores for Next-Generation Sequencing. pages 1–2, 2011.

[51] P.M. Jones and A.M. George. Multidrug resistance in parasites: ABC transporters, P-glycoproteins and molecular modelling. *Int. J. Parasitol.*, 35(5):555–566, 2005.

[52] George A.M. Jones P.M. Multidrug resistance in parasites: ABC transporters, P-glycoproteins and molecular modelling. 2005.

[53] Mustafa Kasim Karahocagil, Hayrettin Akdeniz, Mahmut Sunnetcioglu, Muttalip Cicek, Rafet Mete, Nevzat Akman, Ebubekir Ceylan, Hasan Karsen, and Kubilay Yapici. A familial outbreak of fascioliasis in Eastern Anatolia: A report with review of literature. *Acta Trop.*, 118(3):177–183, 2011.

[54] Ravi S. Kasinathan, Tinopiwa Goronga, Shanta M. Messerli, Thomas R. Webb, and Robert M. Greenberg. Modulation of a Schistosoma mansoni multidrug transporter by the antischistosomal drug praziquantel.

[55] Hassan Khoramian, Mohsen Arbabi, Mahmood Mahami Osqoi, Mahdi Delavari, Hossein Hooshyar, and Mohammarreza Asgari. Prevalence of ruminants fascioliasis and their economic effects in Kashan, center of Iran. *Asian Pac. J. Trop. Biomed.*, 4(11):918–922, 2014.

[56] G. Knubben-schweizer, S. Rüegg, P. R. Torgerson, C. Rapsch, F. Grimm, M. Häss-sig, P. Deplazes, and U. Braun. Control of bovine fasciolosis in dairy cattle in Switzerland with emphasis on pasture management. *Vet. J.*, 186(2):188–191, 2010.

[57] Eugene V Koonin, Roman L Tatusov, and Kenneth E Rudd. Sequence similarity analysis of Escherichia coli proteins : Functional and evolutionary implications. *Proc. Natl. Acad. Sci. USA*, 92(December):11921–11925, 1995.

[58] Birte Kuerpick, Thomas Schnieder, and Christina Strube. Seasonal pattern of fasciola hepatica antibodies in dairy herds in Northern Germany. *Parasitol. Res.*, 111(3):1085–1092, 2012.

[59] Sudhir Kumar and Alan Filipski. Multiple sequence alignment : In pursuit of homologous DNA positions. pages 127–135, 2007.

[60] S. Kumkate, S. Chunchob, and T. Janvilisri. Expression of ATP-binding cassette multidrug transporters in the giant liver fluke Fasciola gigantica and their possible involvement in the transport of bile salts and anthelmintics. *Mol. Cell. Biochem.*, 317(1-2):77–84, 2008.

[61] Lehrmittel online. http://lehrmittelonline.de/.

[62] Detlef D. Leipe, Yuri I. Wolf, Eugene V. Koonin, and L. Aravind. Classification and evolution of P-loop GTPases and related ATPases. *J. Mol. Biol.*, 317(1):41–72, 2002.

[63] Phillippe Lemey, Marco Salemi, and Anne Mieke Vandamme. *The Phylogenetic Handbook*, volume XXXIII. 2009.

[64] Kenneth J. Linton. Structure and function of ABC transporters. *Physiology (Bethesda).*, 22(27):122–30, 2007.

[65] Shikai Liu, Qi Li, and Zhanjiang Liu. Genome-Wide Identification, Characterization and Phylogenetic Analysis of 50 Catfish ATP-Binding Cassette (ABC) Transporter Genes. *PLoS One*, 8(5):1–17, 2013.

[66] Lodish H, Berk A, Zipursky SL, et al., Molecular Cell Biology. http://www.ncbi.nlm.nih.gov/books/NBK21578/.

[67] Biomatters Ltd. Geneious 8.1. pages 1–228, 2015.

[68] Richard Lucius and Brigitte Loos-Frank. *Biologie von Parasiten*, volume 1. 2008.

[69] Marjorie A. Hoy. Insect Molecular Genetics: An Introduction to Principles and Applications. 2013.

[70] Marketa J. Zvelebil, Jeremy O. Baum. *Understanding Bioinformatics*. 2008.

[71] Mas-Coma S., Bargues M.D., Valero M.A. Diagnosis of human fascioliasis by stool and blood techniques: Update for the present global scenario. Parasitology. 2014.

[72] Mas-Coma S., Valero M.A., Bargues M.D. Fasciola, lymnaeids and human fascioliasis, with a global overview on disease transmission, epidemiology, evolutionary genetics, molecular epidemiology and control. Adv. Parasitol. 69, http://www.who.int/foodborne_trematode_infections/fascioliasis/en/. 2009.

[73] Mercedes Mezo, Marta González-Warleta, José Antonio Castro-Hermida, and Florencio M. Ubeira. Evaluation of the flukicide treatment policy for dairy cattle in Galicia (NW Spain). *Vet. Parasitol.*, 157(3-4):235–243, 2008.

[74] Charles J. Mode and Candace K. Sleeman. *Stochastic Processes in Genetics and Evolution*. 2012.

[75] Verónica Molina-Hernández, Grace Mulcahy, Jose Pérez, Álvaro Martínez-Moreno, Sheila Donnelly, Sandra M O'Neill, John P Dalton, and Krystyna Cwiklinski. Fasciola hepatica vaccine: we may not be there yet but we're on the right road. *Vet. Parasitol.*, 208(1-2):101–11, 2015.

[76] Eric Morgan, Johannes Charlier, Guy Hendrickx, Annibale Biggeri, Dolores Catalan, Georg von Samson-Himmelstjerna, Janina Demeler, Elizabeth Müller, Jan van Dijk, Fiona Kenyon, Philip Skuce, Johan Höglund, Padraig O'Kiely, Bonny van Ranst, Theo de Waal, Laura Rinaldi, Giuseppe Cringoli, Hubertus Hertzberg, Paul Torgerson, Adrian Wolstenholme, and Jozef Vercruysse. Global Change and

Helminth Infections in Grazing Ruminants in Europe: Impacts, Trends and Sustainable Solutions. *Agriculture*, 3(3):484–502, 2013.

[77] Burkhard Morgenstern. Einführung in die Bioinformatik, Algorithmen zur Sequenzanalyse. 2005.

[78] L. Mottier, L. Alvarez, L. Ceballos, and C. Lanusse. Drug transport mechanisms in helminth parasites : Passive di V usion of benzimidazole anthelmintics. 113:49–57, 2006.

[79] Luay K. Nakhleh. Phylogenetics: Distance Methods University Rice. 2015.

[80] NCBI BLAST help. http://www.ncbi.nlm.nih.gov/blast.

[81] C.A. Orengo, D.T. Jones, and J.M. Thornton. *Bioinformatics Genes, Proteins and Computers*. 2003.

[82] Osama M Ouda, Sara El-Metwally, and Mohamed Helmy. *Next Generation Sequencing Technologies and Challenges in Sequence Assembly*. 2014.

[83] D. Piedrafita, T. W. Spithill, R. E. Smith, and H. W. Raadsma. Improving animal and human health through understanding liver fluke immunology. (March):572–581, 2010.

[84] Alicia Ponte-Sucre, Emilia Diaz, and Maritza Padron-Nieves. *Drug Resistance in Leishmania Parasites Consequences, Molecular Mechanisms and Possible Treatments*. 2013.

[85] Frederic M. Richards, David S. Eisenberg, Peter S. Kim, and Peer Bork. *Analysis of Amino Acid Sequences*, volume 33. 2000.

[86] Michael Rosenberg. *Sequence Alignment Methods, Models, Concepts and Strategies*, volume 53. 2009.

[87] David J Russell. *Multiple Sequence Alignment Methods*. 2014.

[88] Salam Al Karadaghi. Introduction to protein structure and structural bioinformatics, http://www.proteinstructures.com/index.html.

[89] Günter Schmees, Anke Stein, Sabine Hunke, Heidi Landmesser, and Erwin Schneider. Functional consequences of mutations in the conserved 'signature sequence' of the ATP-binding-cassette protein MalK. *Eur. J. Biochem.*, 266(2):420–430, 1999.

[90] P.M Selzer, R.J. Marhöfer, and A. Rohwer. *Angewandte Bioinformatik*. 2004.

[91] Aswin Sai Narain Seshasayee. *Bacterial Genomics: Genome Organization and Gene Expression Tools*, volume 16. 2015.

[92] Jonathan A. Sheps, Steven Ralph, Zhongying Zhao, David L. Baillie, and Victor Ling. The ABC transporter gene family of Caenorhabditis elegans has implications for the evolutionary dynamics of multidrug resistance in eukaryotes. *Genome Biol.*, 5(3):R15, 2004.

[93] V. Shyamala, V. Baichwal, E. Beall, and G. F. Ames. Structure-function analysis of the histidine permease and comparison with cystic fibrosis mutations. *J. Biol. Chem.*, 266(28):18714–9, 1991.

[94] Adam Siepel and David Haussler. Phylogenetic Hidden Markov Models. *Engineering*, (12):325–351, 2005.

[95] Stanford University. http://web.stanford.edu. 2001.

[96] Ryan J. Sullivan. *BRAF Targets in Melanoma Biologcal Mechanisms, Resistance, and Drug Discovery.* 2015.

[97] SIB Swiss Institute of Bioinformatics/EMBnet Node Switzerland, Volker Flegel, and Vassilios Ioannidis. Pairwise Sequence Alignments.

[98] T. Strachan, A. Read. *Human Molecular Genetics.* 2004.

[99] Amira Taman and Manar Azab. Present-day anthelmintics and perspectives on future new targets. *Parasitol. Res.*, 113(7):2425–2433, 2014.

[100] Jyoti Tanwar, Shrayanee Das, Zeeshan Fatima, and Saif Hameed. Multidrug resistance: an emerging crisis. *Interdiscip. Perspect. Infect. Dis.*, 2014:541340, 2014.

[101] Alan R. Templeton. *Population Genetics and Microevolitionary Theory*, volume XXXIII. 2006.

[102] Rafael Toledo and Bernard Fried. *Digenetic Trematodes*, volume 7. 2014.

[103] Joo Chuan Tong and Shoba Ranganathan. Computer-Aided Vaccine Design. *Woodhead Publ. Ser. Biomed.*, 2013.

[104] Understanding Evolution, Berkeley. http://evolution.berkeley.edu/.

[105] Vrije Universiteit. D Ecentralization and L Ocal a Utonomy :. *Public Adm.*, (September):538–540, 2004.

[106] C. van der Does, Chiara Presenti, Katrin Schulze, Stephanie Dinkelaker, and Robert Tampe. Kinetics of the ATP Hydrolysis Cycle of the Nucleotide-binding Domain of Mdl1 Studied by a Novel Site-specific Labeling Technique. *J. Biol. Chem.*, 281(9):5694–5701, 2005.

[107] Severo Vázquez-prieto, Román Vilas, Florencio M. Ubeira, and Esperanza Paniagua. Veterinary Parasitology Temporal genetic variation of Fasciola hepatica from sheep in. *Vet. Parasitol.*, 209(3-4):268–272, 2015.

[108] Welcome Genome Campus. http://www.yourgenome.org/facts.

[109] Richard Wilkinson, Christopher J. Law, Elizabeth M. Hoey, Ian Fairweather, Gerard P. Brennan, and Alan Trudgett. An amino acid substitution in Fasciola hepatica P-glycoprotein from triclabendazole-resistant and triclabendazole-susceptible populations. *Mol. Biochem. Parasitol.*, 186(1):69–72, 2012.

[110] Williamson, S.M. and A.J. Wolstenholme. P-glycoprotein of Haemonchus contortus: development of real time PCR assays for gene expression studies. 2012.

[111] N. Woodford and M. J. Ellington. The emergence of antibiotic resistance by mutation. *Clin. Microbiol. Infect.*, 13(1):5–18, 2007.

[112] World Health Organization. http://www.who.int. 2015.

[113] Jie Xiong, Lifang Feng, Dongxia Yuan, Chengjie Fu, and Wei Miao. Genome-wide identification and evolution of ATP-binding cassette transporters in the ciliate Tetrahymena thermophila: A case of functional divergence in a multigene family. *BMC Evol. Biol.*, 10(1):330, 2010.

[114] Ziheng Yang. *Computational Molecular Evolution.* 2006.

[115] Shui Quing (Ed.) Ye. Bioinformatics - A Pratical Approach. page 646, 2008.

A. Appendix

Gene	Alias	Location	Function
ABCA1	ABC1	9q31.1	Cholesterol efflux onto HDL
ABCA2	ABC2	9q34.3	Drug resistance
ABCA3	ABC3	16p13.3	Surfactant secretion?
ABCA4	ABCR	1p21.3	N-Retinylidiene-PE efflux
ABCA5		17q24.3	
ABCA6		17q24.3	
ABCA7		19p13.3	
ABCA8		17q24.3	
ABCA9		17q24.3	
ABCA10		17q24.3	
ABCA12		2q34	
ABCA13		7p12.3	
ABCB1	PGY1,MDR	7q21.12	Multidrug resistance
ABCB2	TAP1	6p21.3	Peptide transport
ABCB3	TAP2	6p21.3	Peptide transport
ABCB4	PGY3	7q21.12	PC transport
ABCB5		7p21.1	
ABCB6	MTABC3	2q35	Iron transport
ABCB7	ABC7	Xq21-q22	Fe/S cluster transport
ABCB8	MABC1	7q36.1	
ABCB9		12q24.31	
ABCB10	MTABC2	1q42.13	
ABCB11	SPGP	2q24.3	Bile salt transport

Gene	Alias	Location	Function
ABCC1	MRP1	16p13.12	Drug resistance
ABCC2	MRP2	10q24.2	Organic anion efflux
ABCC3	MRP3	17q21.33	Drug resistance
ABCC4	MRP4	13q32.1	Nucleoside transport
ABCC5	MRP5	3q27.1	Nucleoside transport
ABCC6	MRP6	16p13.12	
ABCC7	CFTR	7q31.31	Chloride ion channel
ABCC8	SUR	11p15.1	Sulfonylurea receptor
ABCC9	SUR2	12p12.1	K(ATP) channel regulation
ABCC10	MRP7	6p21.1	
ABCC11		16q12.1	
ABCC12		16q12.1	
ABCD1	ALD	Xq28	VLCFA transport regulation
ABCD2	ALDL1, ALDR	12q11	
ABCD3	PXMP1,PMP70	1p22.1	
ABCD4	PMP69, P70R	14q24.3	
ABCE1	OABP, RNS4I	4q31.31	Oligoadenylate binding protein
ABCF1	ABC50	6p21.1	
ABCF2		7q36.1	
ABCF3		3q27.1	
ABCG1	ABC8, White	21q22.3	Cholesterol transport?
ABCG2	ABCP, MXR, BCRP	4q22	Toxin efflux, drug resistance
ABCG4	White2	11q23	
ABCG5	White3	2p21	Sterol transport
ABCG8		2p21	Sterol transport

Table A.1.: Human ABC transporters [23]

Gene	Mendelian disorder	Complex disease
ABCA1	Tangier disease, FHDLD	
ABCA4	Stargardt/FFM, RP, CRD, CD	AMD
ABCB1	Ivermectin susceptibility	Digoxin uptake
ABCB2	Immune deficiency	
ABCB3	Immune deficiency	
ABCB4	PFIC3	ICP
ABCB7	XLSA/A	
ABCB11	PFIC2	
ABCC2	Dubin-Johnson Syndrome	
ABCC6	Pseudoxanthoma elasticum	
ABCC7	Cystic Fibrosis, CBAVD	Pancreatitis, bronchiectasis
ABCC8	FPHHI 600509	
ABCD1	ALD 300100	
ABCG5	Sitosterolemia 605459	
ABCG8	Sitosterolemia 605460	

Table A.2.: Diseases and phenotypes caused by ABC genes. [23]

A.1. Classification of *Fasciola hepatica* contigs to ABC transporters

A	B	C	D	E	F	G
comp12072_c0_seq1	comp15585_c0_seq1	comp15489_c0_seq1	comp28795_c0_seq1	comp22706_c0_seq2	comp28071_c0_seq1	comp29155_c0_seq1
comp17178_c0_seq1	comp17705_c0_seq1	comp2681_c0_seq1	comp28795_c0_seq10		comp28071_c1_seq3	comp30093_c0_seq2
comp189385_c0_seq1	comp17730_c0_seq1	comp27410_c0_seq1	comp28795_c0_seq12		comp28071_c1_seq4	comp30093_c0_seq3
comp24099_c1_seq1	comp18746_c0_seq1	comp27410_c0_seq3	comp28795_c0_seq13		comp30906_c0_seq1	comp30789_c0_seq4
comp30310_c0_seq7	comp24009_c0_seq2	comp27410_c0_seq7	comp28795_c0_seq14		comp30906_c0_seq5	
comp30403_c3_seq1	comp24009_c0_seq3	comp27910_c0_seq2	comp28795_c0_seq15		comp30906_c2_seq1	
comp30493_c0_seq2	comp25157_c0_seq1	comp27910_c1_seq1	comp28795_c0_seq17			
	comp25157_c0_seq3	comp30390_c0_seq1	comp28795_c0_seq19			
	comp27826_c0_seq10	comp30390_c0_seq2	comp28795_c0_seq2			
	comp27826_c0_seq2	comp30390_c0_seq3	comp28795_c0_seq3			
	comp27826_c0_seq5	comp30390_c0_seq6	comp28795_c0_seq5			
	comp27826_c0_seq7	comp30390_c0_seq8	comp28795_c0_seq7			
	comp27951_c0_seq8	comp30390_c2_seq1	comp28795_c0_seq8			
	comp28343_c0_seq2	comp30390_c2_seq3	comp28795_c0_seq9			
	comp28343_c1_seq1	comp30390_c3_seq1	comp28795_c0_seq20			
	comp28343_c1_seq2	comp508273_c0_seq1	comp28985_c1_seq10			
	comp29030_c0_seq1		comp28985_c1_seq15			
	comp29030_c0_seq2		comp28985_c1_seq16			
	comp29030_c0_seq4					
	comp29030_c0_seq7					
	comp29483_c0_seq1					
	comp29483_c0_seq2					
	comp29483_c0_seq3					
	comp29483_c2_seq1					
	comp29483_c2_seq2					
	comp29651_c0_seq1					
	comp30171_c0_seq5					
	comp30879_c2_seq10					
	comp30879_c2_seq11					
	comp30879_c2_seq12					
	comp30879_c2_seq2					
	comp30879_c2_seq6					
	comp30879_c2_seq7					
	comp30879_c2_seq8					
	comp656746_c0_seq1					

Table A.3.: Assignment of the 19929 *Fasciola hepatica* contigs to the ABC transporter based on the comparison to the *Homo sapiens* dataset.

A	B	C	E	G
comp24099_c1_seq1	comp15585_c0_seq1	comp15489_c0_seq1	comp22706_c0_seq2	comp30093_c0_seq2
comp30310_c0_seq7	comp17705_c0_seq1	comp27410_c0_seq1	comp22706_c0_seq3	comp30093_c0_seq3
comp30403_c3_seq1	comp17730_c0_seq1	comp27410_c0_seq3	comp22706_c0_seq4	comp30789_c0_seq2
comp30493_c0_seq2	comp18746_c0_seq1	comp27410_c0_seq7	comp28071_c0_seq1	comp30789_c0_seq4
	comp24009_c0_seq2	comp27910_c0_seq2	comp28071_c1_seq2	comp30789_c0_seq8
	comp24009_c0_seq3	comp27910_c1_seq1	comp28071_c1_seq3	comp30789_c0_seq9
	comp25157_c0_seq1	comp30390_c0_seq3	comp28071_c1_seq4	
	comp25157_c0_seq3	comp30390_c0_seq6	comp30906_c0_seq1	
	comp25157_c0_seq4	comp30390_c2_seq1	comp30906_c0_seq2	
	comp27410_c0_seq1	comp30390_c2_seq3	comp30906_c0_seq4	
	comp27826_c0_seq10		comp30906_c0_seq5	
	comp27826_c0_seq2		comp30906_c2_seq1	
	comp27826_c0_seq5			
	comp27826_c0_seq7			
	comp28343_c0_seq2			
	comp28343_c1_seq1			
	comp28343_c1_seq2			
	comp29030_c0_seq1			
	comp29030_c0_seq2			
	comp29030_c0_seq4			
	comp29030_c0_seq7			
	comp29483_c0_seq2			
	comp29483_c2_seq1			
	comp29483_c2_seq2			
	comp29651_c0_seq1			
	comp30171_c0_seq5			
	comp30879_c2_seq10			
	comp30879_c2_seq12			
	comp30879_c2_seq2			
	comp30879_c2_seq7			
	comp30879_c2_seq8			
	comp656746_c0_seq1			

Table A.4.: Assignment of the 19929 *Fasciola hepatica* contigs to the ABC transporter based on the comparison to the *Drosophila melanogaster* dataset.

A	B	C	D	E	F	G
comp24099_c1_seq1	comp15585_c0_seq1	comp27410_c0_seq1	comp28795_c0_seq1	comp22706_c0_seq2	comp28071_c0_seq1	comp30093_c0_seq2
comp30310_c0_seq7	comp17705_c0_seq1	comp27410_c0_seq3	comp28795_c0_seq10	comp22706_c0_seq3	comp28071_c1_seq2	comp30093_c0_seq3
comp30403_c3_seq1	comp17730_c0_seq1	comp27410_c0_seq7	comp28795_c0_seq12	comp22706_c0_seq4	comp28071_c1_seq3	comp30789_c0_seq4
comp30493_c0_seq2	comp18746_c0_seq1	comp27910_c0_seq2	comp28795_c0_seq13		comp28071_c1_seq4	comp30789_c0_seq9
	comp24009_c0_seq2	comp27910_c1_seq1	comp28795_c0_seq14		comp30906_c0_seq1	
	comp24009_c0_seq3	comp30390_c0_seq1	comp28795_c0_seq15		comp30906_c0_seq5	
	comp24009_c0_seq4	comp30390_c0_seq2	comp28795_c0_seq17		comp30906_c2_seq1	
	comp25157_c0_seq1	comp30390_c0_seq3	comp28795_c0_seq2			
	comp25157_c0_seq3	comp30390_c0_seq6	comp28795_c0_seq3			
	comp27826_c0_seq10	comp30390_c2_seq1	comp28795_c0_seq5			
	comp27826_c0_seq2	comp30390_c2_seq3	comp28795_c0_seq8			
	comp27826_c0_seq5	comp30390_c3_seq1	comp28795_c0_seq9			
	comp27826_c0_seq7	comp508273_c0_seq1	comp28985_c1_seq10			
	comp28343_c0_seq2		comp28985_c1_seq15			
	comp28343_c1_seq1		comp28985_c1_seq16			
	comp28343_c1_seq2		comp28985_c1_seq2			
	comp29030_c0_seq1		comp28985_c1_seq3			
	comp29030_c0_seq2		comp28985_c1_seq4			
	comp29030_c0_seq4		comp28985_c1_seq5			
	comp29030_c0_seq7		comp28985_c1_seq7			
	comp29483_c0_seq2		comp28985_c1_seq8			
	comp29483_c2_seq1					
	comp29483_c2_seq2					
	comp29651_c0_seq1					
	comp30171_c0_seq5					
	comp30879_c2_seq10					
	comp30879_c2_seq12					
	comp30879_c2_seq2					
	comp30879_c2_seq7					
	comp30879_c2_seq8					
	comp656746_c0_seq1					

Table A.5.: Assignment of the 19929 *Fasciola hepatica* contigs to the ABC transporter based on the comparison to the *Caenorhabditis elegans* dataset.

A	B	C	D	E	F	G
comp12908_c0_seq1	comp1056102_c0_seq1	comp16301_c0_seq3	comp22914_c0_seq1	comp19148_c0_seq1	comp22457_c1_seq1	comp1572814_c0_seq1
comp16085_c0_seq3	comp18261_c1_seq1	comp24084_c1_seq1	comp22914_c0_seq4	comp19148_c1_seq2	comp26245_c2_seq1	comp16784_c0_seq1
comp22725_c2_seq1	comp18261_c1_seq2	comp24084_c1_seq2			comp26245_c5_seq2	comp16784_c0_seq3
comp23081_c1_seq1	comp19729_c0_seq2	comp24084_c1_seq3			comp26245_c5_seq4	comp24179_c1_seq2
comp26521_c0_seq2	comp19729_c0_seq3	comp24084_c1_seq4				comp24179_c2_seq1
	comp213476_c0_seq1	comp24667_c0_seq1				comp24179_c2_seq2
	comp223058_c0_seq1	comp24667_c0_seq2				comp24196_c0_seq1
	comp223058_c0_seq2	comp26258_c1_seq1				comp24196_c1_seq11
	comp23802_c0_seq1	comp26258_c1_seq1				comp24811_c1_seq10
	comp23802_c0_seq3	comp26258_c2_seq2				comp24811_c1_seq11
	comp24270_c0_seq1	comp460578_c0_seq1				comp24811_c1_seq12
	comp24532_c0_seq1	comp7199_c0_seq1				comp24811_c1_seq4
	comp24532_c0_seq2					comp24811_c1_seq6
	comp24532_c0_seq3					comp24811_c1_seq7
	comp24532_c1_seq2					comp24811_c1_seq8
	comp25017_c0_seq1					comp24811_c1_seq9
	comp25017_c0_seq3					
	comp25017_c0_seq4					
	comp25345_c2_seq1					
	comp25345_c2_seq3					
	comp25345_c3_seq1					
	comp25345_c3_seq2					
	comp25345_c3_seq3					
	comp25345_c3_seq4					
	comp25345_c3_seq5					
	comp25903_c0_seq3					
	comp26692_c0_seq1					
	comp26692_c0_seq3					
	comp26692_c2_seq1					
	comp417307_c0_seq1					
	comp659818_c0_seq1					
	comp772312_c0_seq1					
	comp807734_c0_seq1					

Table A.6.: Assignment of the 19931 *Fasciola hepatica* contigs to the ABC transporter based on the comparison to the *Homo sapiens* dataset.

A	B	C	E	G
comp12908_c0_seq1	comp150338_c0_seq1	comp24084_c1_seq1	comp19148_c0_seq1	comp16784_c0_seq1
comp16085_c0_seq3	comp18261_c1_seq1	comp24084_c1_seq3	comp19148_c0_seq2	comp16784_c0_seq3
comp22725_c2_seq1	comp18261_c1_seq2	comp24084_c1_seq4	comp19148_c1_seq1	comp16784_c1_seq1
comp23081_c1_seq1	comp19729_c0_seq2	comp24667_c0_seq1	comp19148_c1_seq2	comp24179_c2_seq1
comp26521_c0_seq2	comp19729_c0_seq3	comp24667_c0_seq2	comp19148_c1_seq3	comp24179_c2_seq2
	comp213476_c0_seq1	comp26258_c1_seq1	comp22457_c1_seq1	comp24196_c0_seq1
	comp223058_c0_seq1	comp26258_c2_seq2	comp26245_c2_seq1	comp24196_c1_seq11
	comp223058_c0_seq2	comp460578_c0_seq1	comp26245_c5_seq2	
	comp23802_c0_seq1		comp26245_c5_seq4	
	comp23802_c0_seq3			
	comp24270_c0_seq1			
	comp24532_c0_seq1			
	comp24532_c0_seq2			
	comp24532_c0_seq3			
	comp25017_c0_seq1			
	comp25017_c0_seq3			
	comp25017_c0_seq4			
	comp25345_c2_seq1			
	comp25345_c2_seq3			
	comp25345_c3_seq1			
	comp25345_c3_seq2			
	comp25345_c3_seq3			
	comp25345_c3_seq4			
	comp25345_c3_seq5			
	comp25903_c0_seq3			
	comp26692_c0_seq1			
	comp26692_c0_seq3			
	comp26692_c2_seq1			
	comp659818_c0_seq1			
	comp772312_c0_seq1			

Table A.7.: Assignment of the 19931 *Fasciola hepatica* contigs to the ABC transporter based on the comparison to the *Drosophila melanogaster* dataset.

A	B	C	D	E	F	G
comp12908_c0_seq1	comp150338_c0_seq1	comp24084_c1_seq1	comp22914_c0_seq1	comp19148_c0_seq1	comp22457_c1_seq1	comp16784_c0_seq1
comp16085_c0_seq3	comp18261_c1_seq1	comp24084_c1_seq3	comp22914_c0_seq4	comp19148_c0_seq2	comp26245_c2_seq1	comp16784_c0_seq3
comp22725_c2_seq1	comp18261_c1_seq2	comp24084_c1_seq4	comp25572_c0_seq4	comp19148_c1_seq1	comp26245_c5_seq2	comp24179_c1_seq2
comp23081_c1_seq1	comp19729_c0_seq2	comp24667_c0_seq1		comp19148_c1_seq2	comp24179_c2_seq1	comp24196_c0_seq1
comp26521_c0_seq2	comp19729_c0_seq3	comp24667_c0_seq2		comp19148_c1_seq3	comp24179_c2_seq2	comp24196_c1_seq11
	comp213476_c0_seq1	comp26258_c1_seq1			comp26245_c5_seq4	
	comp223058_c0_seq1	comp26258_c2_seq1				
	comp223058_c0_seq2	comp26258_c2_seq2				
	comp23802_c0_seq1	comp460578_c0_seq1				
	comp25017_c0_seq4	comp7199_c0_seq1				
	comp23802_c0_seq3					
	comp24270_c0_seq1					
	comp24532_c0_seq1					
	comp24532_c0_seq2					
	comp24532_c0_seq3					
	comp24532_c1_seq2					
	comp25017_c0_seq1					
	comp25017_c0_seq3					
	comp25017_c0_seq4					
	comp25345_c2_seq1					
	comp25345_c2_seq3					
	comp25345_c3_seq1					
	comp25345_c3_seq3					
	comp25345_c3_seq2					
	comp25345_c3_seq4					
	comp25345_c3_seq5					
	comp25903_c0_seq3					
	comp26692_c0_seq1					
	comp26692_c0_seq3					
	comp26692_c2_seq1					
	comp772312_c0_seq1					

Table A.8.: Assignment of the 19931 *Fasciola hepatica* contigs to the ABC transporter based on the comparison to the *Caenorhabditis elegans* dataset.

A	B	C	D	E	F	G
comp104111_c1_seq2	comp101188_c0_seq2	comp102145_c0_seq1	comp104242_c0_seq1	comp104419_c0_seq1	comp101486_c0_seq1	comp103372_c0_seq1
comp105393_c2_seq1	comp101188_c0_seq1	comp106188_c2_seq1	comp105531_c0_seq1	comp104419_c0_seq2	comp101486_c1_seq2	comp103372_c0_seq5
comp451556_c0_seq1	comp102748_c0_seq2	comp106670_c1_seq17	comp105531_c0_seq3		comp103624_c0_seq2	comp103372_c0_seq6
comp72662_c1_seq1	comp102748_c0_seq5	comp84551_c0_seq2			comp103624_c0_seq3	comp103611_c1_seq3
comp902966_c0_seq1	comp103196_c1_seq2	comp96239_c0_seq1			comp103624_c1_seq1	comp106152_c1_seq10
comp99674_c0_seq1	comp104119_c1_seq1	comp96239_c0_seq2				comp106152_c1_seq12
comp99674_c0_seq2	comp104119_c1_seq2					comp106152_c1_seq13
comp99674_c0_seq3	comp105592_c2_seq1					comp106152_c1_seq14
	comp105592_c2_seq2					comp106152_c1_seq2
	comp105592_c2_seq3					comp106152_c1_seq22
	comp105592_c2_seq4					comp106152_c1_seq6
	comp105592_c2_seq6					comp106152_c1_seq7
	comp105692_c0_seq2					comp106152_c1_seq9
	comp105692_c0_seq4					comp390524_c0_seq1
	comp105692_c0_seq5					comp658573_c0_seq1
	comp105692_c0_seq6					
	comp105692_c0_seq7					
	comp105804_c0_seq1					
	comp105804_c0_seq2					
	comp105804_c0_seq3					
	comp105804_c0_seq4					
	comp105804_c1_seq1					
	comp105804_c2_seq1					
	comp105929_c0_seq1					
	comp105929_c0_seq12					
	comp105929_c0_seq13					
	comp105929_c0_seq15					
	comp105929_c0_seq17					
	comp105929_c0_seq19					
	comp105929_c0_seq2					
	comp105929_c0_seq3					
	comp105929_c0_seq4					
	comp105929_c0_seq5					
	comp105929_c0_seq6					
	comp105929_c0_seq7					
	comp105929_c0_seq8					
	comp105929_c1_seq1					
	comp1125873_c0_seq1					
	comp331540_c0_seq1					
	comp659574_c0_seq1					
	comp89166_c0_seq2					
	comp89166_c0_seq5					
	comp99094_c0_seq1					
	comp99094_c0_seq2					
	comp99094_c0_seq3					
	comp99094_c1_seq1					

Table A.9.: Assignment of the 19932 *Fasciola hepatica* contigs to the ABC transporter based on the comparison to the *Homo sapiens* dataset.

A	B	C	E	G
comp104111_c1_seq2	comp101188_c0_seq1	comp102145_c0_seq1	comp101486_c0_seq1	comp103611_c1_seq3
comp105393_c2_seq1	comp101188_c0_seq2	comp106188_c2_seq1	comp101486_c1_seq2	comp106152_c1_seq12
comp451556_c0_seq1	comp102748_c0_seq2	comp106670_c1_seq17	comp103624_c1_seq1	comp106152_c1_seq2
comp99674_c0_seq1	comp102748_c0_seq5	comp75391_c0_seq2	comp103624_c1_seq2	comp390524_c0_seq1
comp99674_c0_seq2	comp103196_c1_seq2	comp84551_c0_seq2	comp103624_c1_seq3	
	comp104119_c1_seq2	comp96239_c0_seq1	comp104419_c0_seq1	
	comp105592_c2_seq1	comp96239_c0_seq2		
	comp105592_c2_seq2	comp96239_c0_seq3		
	comp105592_c2_seq3			
	comp105592_c2_seq4			
	comp105592_c2_seq6			
	comp105692_c0_seq2			
	comp105692_c0_seq4			
	comp105692_c0_seq5			
	comp105692_c0_seq6			
	comp105692_c0_seq7			
	comp105804_c0_seq1			
	comp105804_c0_seq2			
	comp105804_c0_seq3			
	comp105804_c0_seq4			
	comp105804_c1_seq1			
	comp105804_c2_seq1			
	comp105929_c0_seq1			
	comp105929_c0_seq12			
	comp105929_c0_seq13			
	comp105929_c0_seq15			
	comp105929_c0_seq17			
	comp105929_c0_seq19			
	comp105929_c0_seq2			
	comp105929_c0_seq4			
	comp105929_c0_seq7			
	comp105929_c0_seq8			
	comp105929_c1_seq1			
	comp106670_c1_seq16			
	comp106670_c1_seq2			
	comp106670_c1_seq7			
	comp659574_c0_seq1			
	comp89166_c0_seq2			
	comp99094_c0_seq1			
	comp99094_c0_seq2			
	comp99094_c0_seq3			
	comp99094_c1_seq1			

Table A.10.: Assignment of the 19932 *Fasciola hepatica* contigs to the ABC transporter based on the comparison to the *Drosophila melanogaster* dataset.

A	B	C	E	F	G
comp104111_c1_seq2	comp101188_c0_seq1	comp102145_c0_seq1	comp104419_c0_seq1	comp101486_c0_seq1	comp103611_c1_seq3
comp105393_c2_seq1	comp101188_c0_seq2	comp106188_c2_seq1	comp104419_c0_seq2	comp101486_c1_seq2	comp106152_c1_seq12
comp451556_c0_seq1	comp102748_c0_seq2	comp106670_c1_seq17		comp103624_c0_seq2	comp106152_c1_seq2
comp99674_c0_seq1	comp102748_c0_seq5	comp84551_c0_seq2		comp103624_c0_seq3	comp390524_c0_seq1
comp99674_c0_seq2	comp103196_c1_seq2	comp96239_c0_seq1		comp103624_c1_seq1	
comp99674_c0_seq3	comp104119_c1_seq2				
	comp105592_c2_seq1				
	comp105592_c2_seq2				
	comp105592_c2_seq3				
	comp105592_c2_seq4				
	comp105592_c2_seq6				
	comp105692_c0_seq2				
	comp105692_c0_seq4				
	comp105692_c0_seq5				
	comp105692_c0_seq6				
	comp105692_c0_seq7				
	comp105804_c0_seq1				
	comp105804_c0_seq2				
	comp105804_c0_seq3				
	comp105804_c0_seq4				
	comp105804_c1_seq1				
	comp105929_c0_seq1				
	comp105929_c0_seq12				
	comp105929_c0_seq13				
	comp105929_c0_seq15				
	comp105929_c0_seq17				
	comp105929_c0_seq19				
	comp105929_c0_seq2				
	comp105929_c0_seq4				
	comp105929_c0_seq5				
	comp105929_c0_seq6				
	comp105929_c0_seq7				
	comp105929_c0_seq8				
	comp105929_c1_seq1				
	comp331540_c0_seq1				
	comp659574_c0_seq1				
	comp89166_c0_seq2				
	comp99094_c0_seq1				
	comp99094_c0_seq2				
	comp99094_c0_seq3				
	comp99094_c1_seq1				

Table A.11.: Assignment of the 19932 *Fasciola hepatica* contigs to the ABC transporter based on the comparison to the *Caenorhabditis elegans* dataset.