The Eukaryotic ITS2 Database – A workbench for modelling RNA sequence-structure evolution

* * *

Die Eukaryotische ITS2 Datenbank - Eine Plattform zur Modellierung von RNA Sequenzstruktur Evolution

Doctoral thesis for a doctoral degree at the
Graduate School of Life Sciences,
Julius-Maximilians-Universität Würzburg

PhD thesis submitted by:	*Christian Koetschan*
Place of birth:	*Würzburg*

March 15, 2012

Submitted on: . . . .

Office stamp

Members of the Promotionskomitee:

Chairperson:	Prof. Dr. Matthias Frosch
Primary Supervisor:	Prof. Dr. Jörg Schultz
Supervisor (Second):	Prof. Dr. Thomas Dandekar
Supervisor (Third):	Prof. Dr. Dietmar Seipel

Additional supervisors:

Group Leader:	Dr. Tobias Müller
Group Leader:	Dr. Matthias Wolf

Date of Public Defence: . . . .

Date of receipt of Certificates: . . . .

Affidavit / Eidesstattliche Erklärung

English:

I hereby confirm that my thesis entitled ”The Eukaryotic ITS2 Database – A workbench for modelling RNA sequence-structure evolution” is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Signature: ______________________

(Christian Koetschan)

____

Place, Date

Deutsch:

Hiermit erkläre ich an Eides statt, die Dissertation „Die Eukaryotische ITS2 Datenbank - Eine Plattform zur Modellierung von RNA Sequenzstruktur Evolution“ eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Unterschrift: ______________________

(Christian Koetschan)

____

Ort, Datum

Acknowledgements

Many people have supported me during the last years at the Faculty of Bioinformatics, and made work feel like pleasure. First of all, I want to thank all members of the Department for a great and enjoyable time during my studies and the time of my PhD.

Further, I want to thank my mentor Prof. Dr. Jörg Schultz, for the guidance of my PhD thesis, inspiring discussions in the group meetings, and for maintaining a comfortable working environment. Also I thank Prof. Dr. Thomas Dandekar and Prof Dr. Dietmar Seipel for supporting me as members of the Graduate School supervisor committee during the last years.

I owe a debt of gratitude to Dr. Tobias Müller for spending so much time and contributing with his profound knowledge and advice. Thank you for the time-consuming discussions – even exceeding the working hours, suggestions and mathematical assistance in many cases! I really appreciate all this effort, which improved my thesis "significantly" – *** without any doubts or further needed statistical tests!

I would also like to thank Dr. Matthias Wolf, Dr. Torben Friedrich, Dr. Frank Förster and Thomas Hackl for their collaborative teamwork and assistance, answers and explanations to biological questions and discussions – even until 5 am in the morning.

Special Thanks go to Daniela Beißer, Gaby Wangorsch, Santosh Nilla and Astrid Fieselmann. Without extensive mountain-biking tours, badminton, speedminton, slack-lining and regular walks to the chocolate vending machine, time would have passed only half as fast and I would have missed some of the most pleasant moments in my life. Thank you for the great time, I will definitely miss it!

Finally I want to thank Desislava Boyanova and my family who supported me especially through the time of writing and during my studies, for which I owe them the biggest "Thank you"!

Table of Contents

1 Summary / Zusammenfassung

1.1 English:

During the past years, the internal transcribed spacer 2 (ITS2) was established as a commonly used molecular phylogenetic marker for the eukaryotes. Its fast evolving sequence is predestinated for the use in low-level phylogenetics. However, the ITS2 also consists of a very conserved secondary structure. This enables the discrimination between more distantly related species. The combination of both in a sequence-structure based analysis increases the resolution of the marker and enables even more robust tree reconstructions on a broader taxonomic range.

But, performing such an analysis required the application of different programs and databases making the use of the ITS2 non trivial for the typical biologist. To overcome this hindrance, I have developed the ITS2 Workbench, a completely web-based tool for automated phylogenetic sequence-structure analyses using the ITS2 (http://its2.bioapps.biozentrum.uni-wuerzburg.de). The development started with an optimization of length modelling topologies for Hidden Markov Models (HMMs), which were successfully applied on a secondary structure prediction model of the ITS2 marker. Here, structure is predicted by considering the sequences' composition in combination with the length distribution of different helical regions. Next, I integrated HMMs into the sequence-structure generation process for the delineation of the ITS2 within a given sequence. This re-implemented pipeline could more than double the number of structure predictions and reduce the runtime to a few days. Together with further optimizations of the homology modelling process I can now exhaustively predict secondary structures in several iterations. These modifications currently provide 380,000 annotated sequences including 288,000 structure predictions. To include these structures in the calculation of alignments and phylogenetic trees, I developed the R-package "treeforge". It generates sequence-structure alignments on up to four different coding alphabets. For the first time also structural bonds were considered in alignments, which required

the estimation of new scoring matrices. Now, the reconstruction of Maximum Parsimony, Maximum Likelihood as well as Neighbour Joining trees on all four alphabets requires just a few lines of code. The package was used to resolve the controversial chlorophyceaen dataset and could be integrated into future versions of the ITS2 workbench. The platform is based on a modern, feature-rich Web 2.0 user interface equipped with the latest AJAX and Web-service technologies. It performs HMM-based sequence annotation, structure prediction by energy minimization or homology modelling, alignment calculation and tree reconstruction on a flexible data pool that repeats calculations according to data changes. Further, it provides sequence motif detection to control annotation and structure prediction and a sequence-structure based BLAST search, which facilitates the taxon sampling process. All features and the usage of the ITS2 workbench are explained in a video tutorial. However, the workbench bears some limitations regarding the size of datasets. This is caused mainly due to the immense computational power needed for such extensive calculations. To demonstrate the validity of the approach also for large-scale analyses, a fully automated reconstruction of the Chlorophyta (Green Algal) Tree of Life was performed. The successful application of the marker even on large datasets underlines the capabilities of ITS2 sequence-structure analysis and suggests its utilization on further datasets. The ITS2 workbench provides an excellent starting point for such endeavours.

1.2 Deutsch:

In den vergangenen Jahren etablierte sich der Marker „internal transcribed spacer 2”(ITS2) zu einem häufig genutzten Werkzeug in der molekularen Phylogenetik der Eukaryoten. Seine schnell evolvierende Sequenz eignet sich bestens für den Einsatz in niedrigeren phylogenetischen Ebenen. Die ITS2 faltet jedoch auch in eine sehr konservierte Sekundärstruktur. Diese ermöglicht die Unterscheidung weit entfernter Arten. Eine Kombination aus beiden in einer Sequenzstrukturanalyse verbessert die Auflösung des Markers und ermöglicht die Rekonstruktion von robusteren Bäumen auf höherer taxonomischer Breite. Jedoch war die Durchführung solch einer Analyse, die die Nutzung unterschiedlichster Programme und Datenbanken vorraussetzte, für den klassischen Biologen nicht einfach durchführbar. Um diese Hürde zu umgehen, habe ich den „ITS2 Workbench“ entwickelt, eine im Internet nutzbare Arbeitsplattform zur automatisierten sequenzstrukturbasierten phylogenetischen Analyse basierend auf der ITS2 (http://its2.bioapps.biozentrum.uni-wuerzburg.de). Die Entwicklung begann mit der Längenoptimierung unterschiedlicher „Hidden Markov Model“ (HMM)-Topologien, die erfolgreich auf ein Modell zur Sequenzstrukturvorhersage der ITS2 angewandt wurden. Hierbei wird durch die Analyse von Sequenzbestandteilen in Kombination mit der Längenverteilung verschiedener Helixregionen die Struktur vorhergesagt. Anschließend konnte ich HMMs auch bei der Sequenzstrukturgenerierung einsetzen um die ITS2 innerhalb einer gegebenen Sequenz zu lokalisieren. Dieses neu implementierte Verfahren verdoppelte die Anzahl vorhergesagter Strukturen und verkürzte die Laufzeit auf wenige Tage. Zusammen mit weiteren Optimierungen des Homologiemodellierungsprozesses kann ich nun erschöpfend Sekundärstrukturen in mehreren Interationen vorhersagen. Diese Optimierungen liefern derzeit 380.000 annotierte Sequenzen einschließlich 288.000 Strukturvorhersagen. Um diese Strukturen für die Berechnung von Alignments und phylogenetischen Bäumen zu verwenden hab ich das R-Paket „treeforge“ entwickelt. Es ermöglicht die Generierung von Sequenzstrukturalignments auf bis zu vier unterschiedlich kodierten Alphabeten. Damit können erstmals auch strukturelle Basenpaarungen in die Alignmentberechnung mit einbezogen werden, die eine Schätzung neuer Scorematrizen vorraussetzten. Das R-Paket ermöglicht zusätzlich die Rekonstruktion von „Maximum Parsimony“, „Maximum Likelihood“ und „Neighbour Joining“ Bäumen auf allen vier Alphabeten mittels weniger Zeilen Programmcode. Das Paket wurde eingesetzt, um die noch umstrittene Phylogenie der „chlorophyceae“ zu rekonstruieren und könnte in zukünftigen Versionen des ITS2 workbench verwendet werden. Die ITS2 Plattform basiert auf einer modernen und sehr umfangreichen Web 2.0 Oberfläche und beinhaltet neuste AJAX und Web-Service Technologien. Sie umfasst die HMM basierte Sequenzannotation, Strukturvorhersage durch Energieminimierung bzw. Homologiemodellierung, Alignmentberechnung und Baumrekonstruktion basierend auf einem flexiblen Datenpool, der Änderungen am Datensatz automatisch aktualisiert. Zusätzlich wird eine Detektion von Sequenzmotiven ermöglicht, die zur Kontrolle von Annotation und Strukturvorhersage dienen kann. Eine BLAST basierte Suche auf Sequenz- und Strukturebene bietet zusätzlich eine Vereinfachung des Taxonsamplings. Alle Funktionen sowie die Nutzung der ITS2 Webseite sind in einer kurzen Videoanleitung dargestellt. Die Plattform lässt jedoch nur eine bestimmte Größe von Datensätzen zu. Dies liegt vor allem an der erheblichen Rechenleistung, die bei diesen Berechnungen benötigt wird. Um die Funktion dieses Verfahrens auch auf großen Datenmengen zu demonstrieren, wurde eine voll automatisierte Rekonstruktion des Grünalgenbaumes (Chlorophyta) durchgeführt. Diese erfolgreiche, auf dem ITS2 Marker basierende Studie spricht für die Sequenz-Strukturanalyse auf weiteren Daten in der Phylogenetik. Hier bietet der ITS2 Workbench den idealen Ausgangspunkt.

2 Introduction

From "On the origin of species" to molecular biology

What is a species? Are same looking individuals also belonging to the same species? Which one evolved first? These questions have given food for thought to biologist for several centuries and different theses were raised. But even today – more than 150 years after Charles Darwins (* 12.2.1809; † 19.4.1882) "On the Origin of Species" (Darwin, 1859), the answers are not always trivial to give. Whereas at first, species were distinguished by their morphological features, one of the most recognized statements on the species concept was introduced by Ernst Mayr (* 5.7.1904; † 3.2.2005) in 1942 in his book Mayr (1942): "Species are groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups." This statement "gained almost universal acceptance because it explained concisely the role of the species in biology" (Mayr, 1942). Nevertheless, a discrimination is not very easy, especially in difficult cultivatable or solitary habitats. Furthermore it leaves the question unanswered regarding asexual individuals. But even so, large progress has been made after Darwin's sketch of the probably first phylogenetic tree. With the emergence of computational biology, large scale sequence analysis, exponentially growing databases and high-trough-put sequencing, today tons of data became available in a relatively short time. Since the studies of Watson and Crick on the structure of DNA: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material" (Watson and Crick, 1953), it is clear where genetic information is coded.

The molecular phylogenetic pipeline for automated tree reconstruction

However, the question arose on how and where to find comparable regions inside these differing genomes. This led to the assignment of phylogenetic markers. A phylogenetic marker is a sequence fragment of same origin, available in a large taxonomic unit with differing nucleic manifestations for each species. Based on these fragments, an alignment – containing columns with equal and distinct positions can be created. First, these were constructed by hard and exhausting manual work, nowadays tough, the alignment algorithms of MUSCLE (Edgar, 2004a,b) and ClustalW2 (Thompson et al., 1994; Larkin et al., 2007) simplify this task. From the alignment, it is only one more step to a molecular phylogeny. Saitou and Nei (1987); Gascuel (1997) proposed a method called Neighbour Joining (NJ) to calculate a tree very quickly. His method counts the number of differences in the alignment-columns for each sequence and computes a distance based on the results. Competing treeing methods like maximum parsimony (Camin and Sokal, 1965) and maximum likelihood (Felsenstein, 1981, 1985, 2004) are not less popular and widely applied in this field.

The ITS2 – a predestinated marker for tree inference

However, a phylogenetic tree reconstruction strongly relies on the quality of the underlying marker. This is firstly, its exact annotation, and secondly, the range of its resolution and sequence variability. To the latter, a plethora of debates had been started which resulted in the use of chloroplast specific (e.g. matK, rbcL, rpoC1, psbA-trnH), mitochondrial (e.g. CO1) or ribosomal (e.g. ITS1/2, 5.8S) markers (Moniz and Kaczmarska, 2009, 2010; Chen et al., 2010; Yao et al., 2010). Where chloroplast markers perform reasonably well on plants, they can't be applied on animal species. Here, CO1 provides good results instead (Astrin et al., 2006; Kitahara et al., 2010; Smith et al., 2008). This discrepancy can be explained by the fact that fast evolving markers suit best for low-level phylogenies with a high sequence variability needed to distinguish between two closely related species. In the case of high-level phylogenies instead, a lower variability is preferred to distinguish between far related species. Here, fast sequence variations would cause a high level of noise.

Our work focuses on the ribosomal RNA cistron (Figure 1), especially the Internal Transcribed Spacer 2 (ITS2). The rRNA cistron contains genes forming the smaller (18S) and larger (28S, 5.8S and 5S) subunit of the ribosome in eukaryotes. Located in between are the genetic spacers ITS1 and ITS2. Ribosomes occur in prokaryotes as well as eukaryotes for the synthesis of proteins during translation, however the ITS2, which exact function is not unravelled yet is only available in the latter.

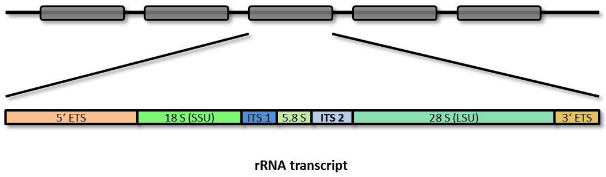


Figure 1: A graphical view of the ribosomal RNA cistron. It consists of the 18S small subunit (SSU), ITS1, 5.8S, ITS2, 28S large subuni(LSU).

At first, only sequence data was accessible to distinguish between different species (Baldwin et al., 1995; Powers et al., 1997; Suh et al., 1993) and it turned out that the range of this marker is often too weak to cover also the discrimination in higher – order or family levels (Coleman, 2003). Interestingly, the ITS2 folds into a secondary structure which revealed a common, very conserved core throughout all eukaryotes (Coleman, 2003; Schultz et al., 2005; Mai and Coleman, 1997). With the conserved structure enabling discrimination at higher ranks, also the accuracy and robustness of trees increased (Keller et al., 2010; Telford et al., 2005). This finding quickly led to the integration of secondary structure into phylogenetic databases like the ITS2-DB (Schultz et al., 2006; Selig et al., 2008; Koetschan et al., 2010). With an exact Hidden Markov Model annotation of the conserved flanking regions 5.8S and 28S of the ITS2

(Keller et al., 2009), a huge amount of highly reliable individual secondary structures became available. Due to the advent of sequence-structure analysis software, the calculation of alignments (Seibel et al., 2006, 2008; Bauer et al., 2007; Siebert and Backofen, 2005) and reconstruction of phylogenetic trees (Wolf et al., 2008) could benefit from both integrated features.

An integrated platform for ITS2-based phylogenies

The presence of such a large set of phylogenetic tools requires a broad knowledge for their meaningful application and can often be time-consuming. Especially after two decades of controversy in taxon sampling (Nabhan and Sarkar, 2011), one has to be prepared that calculations might result in repetitions. Further, erroneous sequences like false classifications often become visible just after the final tree has been produced. To overcome this time-consuming process, the ITS2 workbench has been developed. This working suite unifies all formerly required stand-alone tools for a sequence-structure based analysis online and follows the work flow proposed by Schultz and Wolf (2009). It requires no additional installations and integrates an updated version of the ITS2 database with nearly 300,000 sequences and structures. Beside an interactive way of adding or removing sequence-structures at different stages, quality control mechanisms have been implemented allowing an automated repetition of previous calculations, as soon as changes to the dataset have been made.

In a nutshell, the workbench remains compatible to the former software, however additionally providing very fast insights into a phylogeny by a complete automation of a predefined pipeline and producing reliable results within just a few mouse clicks.

3 Optimization of length modelling topologies for HMMs

3.1 Introduction

To achieve high quality predictions, the first part – the annotation of the marker, is already a crucial step. Here, the ITS2 marker benefits from its very conserved neighbouring genes. These can be targeted by a simple Hidden Markov Model approach. HMMs – first emerged in the IT and mainly used for speech recognition (Rabiner, 1989; Juang and Rabiner, 1991), nowadays play an important role in software for spam deobfuscation (Lee and Ng, 2005), image processing (Willsky, 2002; Rossi et al., 2010) or in bioinformatics applications. To the latter, the work of Jean Eddy and the development of the HMMER suite (Eddy, 1998, 2008) had a big influence, which simplified the integration and distribution of HMMs in a large variety of programs. Especially when it comes to the automated detection of specific regions or motifs inside a strand of typically Nucleotides, Aminoacids or Proteins, HMMs benefit from their statistical capabilities. Being successfully applied in Genescan (Lukashin and Borodovsky, 1998) for the detection of genes, the HMMER webserver for interactive sequence similarity search (Finn et al., 2011), the prediction of interaction sites by Friedrich et al. (2006) or signal peptide (Juncker et al., 2003; Schneider and Fechner, 2004; Zhang and Wood, 2003; Käll et al., 2004) and transmembrane predictions (Sonnhammer et al., 1998; Tusnády and Simon, 1998; Krogh et al., 2001; Kahsay et al., 2005; Martelli et al., 2002; Liu et al., 2003; Bagos et al., 2004) are just some of the highlights in this field. Regarding the ITS2, the exact annotation of its sequence is of major importance for an accurate structure prediction (Keller et al., 2009). Interestingly, the ITS2 also consists of different motifs (Koetschan et al., 2010) and structural regions, such as stems, loops or inter-helical areas. This formed the idea whether – based on the composition of the ITS2 sequence, a full structure prediction can be achieved. In the following study, HMMs were applied to identify start and

end regions of the ITS2, as well as to model a complete secondary structure. However, with different structural regions, a large variety of the underlying source data is coherent, especially when regarding the length distribution of sequences. This makes it difficult to elicit optimal performance using state of the art software like the HMMER suite. Typically, an HMM consists of several states, each connected by transitions. The retention time within one state containing a self-transition then usually follows a geometric law (Durbin et al., 1998). But this must not automatically reflect the length distribution of source data, e.g. loop or stems inside the ITS2. A bell-shaped length distribution e.g. cannot be modelled by only one self-transitional state, but by a sequential replication of those (Durbin et al., 1998), which would promise a better HMM-based prediction. In the following, HMM-based optimization methods are described by adapting the number of states with its probabilities of an HMM to the length distribution of the underlying data according to the publications of Koetschan (2008); Friedrich et al. (2010); Friedrich (2009).
In comparison to the diploma thesis of Koetschan (2008), the publication (Friedrich et al., 2010) contains a full validation of optimisation methods on artificial test scenarios, confirmed by statistical tests, cross validation and receiver operating characteristic (ROC) curves. Further a performance estimation of ML (Maximum Likelihood) and MM (Method of Moments) was newly included. Beside a novel calculation method for ITS2 secondary structure error predictions, the HMM optimization was successfully applied on various interesting biological sequences and motifs in a quick screening.

3.2 Methods

3.2.1 Hidden Markov Models

In a few words, an HMM can be explained as a network of nodes or states $\mathcal{Q} = \{q_1, \ldots, q_m\}$, described in more detail by Durbin et al. (1998). These States q_i, q_j are typically connected to each other by a transition probability τ_{ij}. Thus, all outgoing transitions of a state sum to one. States may emit an alphabet of symbols $\mathcal{O} = \{\omega_1, \ldots, \omega_n\}$, or can even be silent. To add emission and transition probabilities to an HMM, different learning algorithms are known. Supervised learning, e.g. relies on state path and emission values to calculate the according probabilities. When states paths are unknown, the Baum-Welch (Baum et al., 1970; Welch, 2003) training estimates this information by a maximum likelihood approach. Further important algorithms like Viterbi (Viterbi, 1967; Forney, 1973) and posterior decoding (Durbin et al., 1998) are able to provide at least one best state path through the model.

3.2.2 Maximum likelihood

The likelihood L to the density function $p(X|\Theta)$ with the random variable X and the unknown parameters Θ of a given a data set of size N is described by

$$p(X|\Theta) = \prod_{i=1}^{N} p(x_i|\Theta) = L(\Theta|x_1, \ldots, x_N). \tag{3.1}$$

To estimate the unknown parameters Θ, the set of parameters $\hat{\Theta}$ which maximizes the overall likelihood is calculated

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\ L(\Theta|x_1, \ldots, x_N). \tag{3.2}$$

The parameters maximizing the likelihood equal the parameters maximizing the log-likelihood function. By transferring calculations to the log-space, the product of 3.1 is replaced by a sum, which avoids calculations with very high numbers.

$$\mathcal{L}(\Theta|x_1, \ldots, x_N) = log(L(\Theta|x_1, \ldots, x_N)). \tag{3.3}$$

Here, $log(L(\Theta|x_1, \ldots, x_N))$ is described as the log-likelihood of a model with parameters Θ. When maximizing the log-likelihood, its estimates Θ_i ($\Theta = \Theta_1, \ldots, \Theta_n$) are derived by locating the zero point of partial derivatives of this function (Johnson et al., 2006)

$$\frac{\delta}{\delta\,\Theta_i}\mathcal{L}(\Theta_1, \ldots, \Theta_m|x_1, \ldots, x_N) = 0, \qquad i = 1, \ldots, m. \tag{3.4}$$

3.2.3 Method of moments

With this method, the moments of a function are used as parameters to describe a distribution. This idea was first published by Pearson (1902) and gives an analytical approach for parameter estimation.

For a random variable X the moments of a function can be calculated by

$$E\left[X^k\right] = \begin{cases} \sum\limits_x x^k p(x) & \text{if X is discrete,} \\ \int\limits_{-\infty}^{+\infty} x^k f(x)\, dx & \text{if X is continuous.} \end{cases} \tag{3.5}$$

Here, $p(x)$ is referred to as the probability mass function, $f(x)$ as the probability density function (PDF) of x. Then,

$$M(t) = E(e^{tX}) \tag{3.6}$$

is the according moment generating function of X.

When differentiating $M(t)$ at $t = 0$, the k^{th} differentiation corresponds to the k^{th} moment

$$\frac{d^k}{dt} M(t)|_{t=0} = E[X^k e^{tX}] \mid_{t=0} = E(X^k). \tag{3.7}$$

Now, parameters can be estimated by replacing moments by sample moments and resolving the equations

$$\begin{aligned} E[X^k] &= \frac{1}{n}\sum_{i=1}^{n} x_i^k = \overline{x^k}, \\ \overline{x^k} &= \bar{x}, \overline{x^2}, \ldots, \overline{x^k} := \text{sample moment k.} \end{aligned} \tag{3.8}$$

3.2.4 Goodness of fit and model choice

The likelihood of a model can be regarded under the influence of model parameters and its complexity. A measure for model choice which punishes model complexity by the number of parameters on a logarithmic scale is the BIC criteria by Schwarz (1978). It is calculated according to

$$BIC = -2\mathcal{L}(\theta, x) + k\, ln(n). \tag{3.9}$$

$\mathcal{L}(\theta, x)$ here denotes the likelihood of the model where k equals the number of estimated parameters and n is the sample size. Estimators which don't compute a likelihood can be directly compared by determining the L_1 distance of two discrete distributions x and y

$$d_1(x,y) = ||x-y||_1 := \sum_{i=1}^{n} |x_i - y_i|. \tag{3.10}$$

3.3 Results

3.3.1 Length modelling – topologies

Each self-transitional state of an HMM is coupled to a specific holding time while resting in the self-transition. Thus, at each (re)-visit, the length of the state path is incremented by one symbol resulting from this specific state. Sequences, emitted by this state then follow a geometric length distribution (Durbin et al., 1998). To model even bell-shaped source data (Figure 3), in the following, three length modelling topologies are described in a nested complexity. A state, which becomes substituted by several chained-linked states in the next complexity level is further called a macro state (MS).

The p macro state

emits geometrically length distributed sequences ($geo(p)$). The likelihood function for these sequence lengths $x_1, \dots, x_n$ is given by

$$L(p|x_1,\dots,x_n) \quad = \quad \prod_{i=1}^{n} p(1-p)^{x_i-1}. \tag{3.11}$$

To estimate p, the log-likelihood of the equation above is then maximized

$$\begin{aligned} \hat{p} &= \operatorname*{argmax}_p \mathcal{L}(p|x_1,\dots,x_n) && (3.12)\\ &= \operatorname*{argmax}_p \log(\prod_{i=1}^{n} p(1-p)^{x_i-1}) \\ &= \operatorname*{argmax}_p n\log p + \sum_{i=1}^{n}(x_i-1)\log(1-p). \end{aligned}$$

By solving the equation

$$\frac{\delta\mathcal{L}(p|x_1,\dots,x_n)}{dp} \quad = \tfrac{n}{p} - \textstyle\sum_{i=1}^{n}(x_i-1)\tfrac{1}{1-p} \quad = 0 \tag{3.13}$$

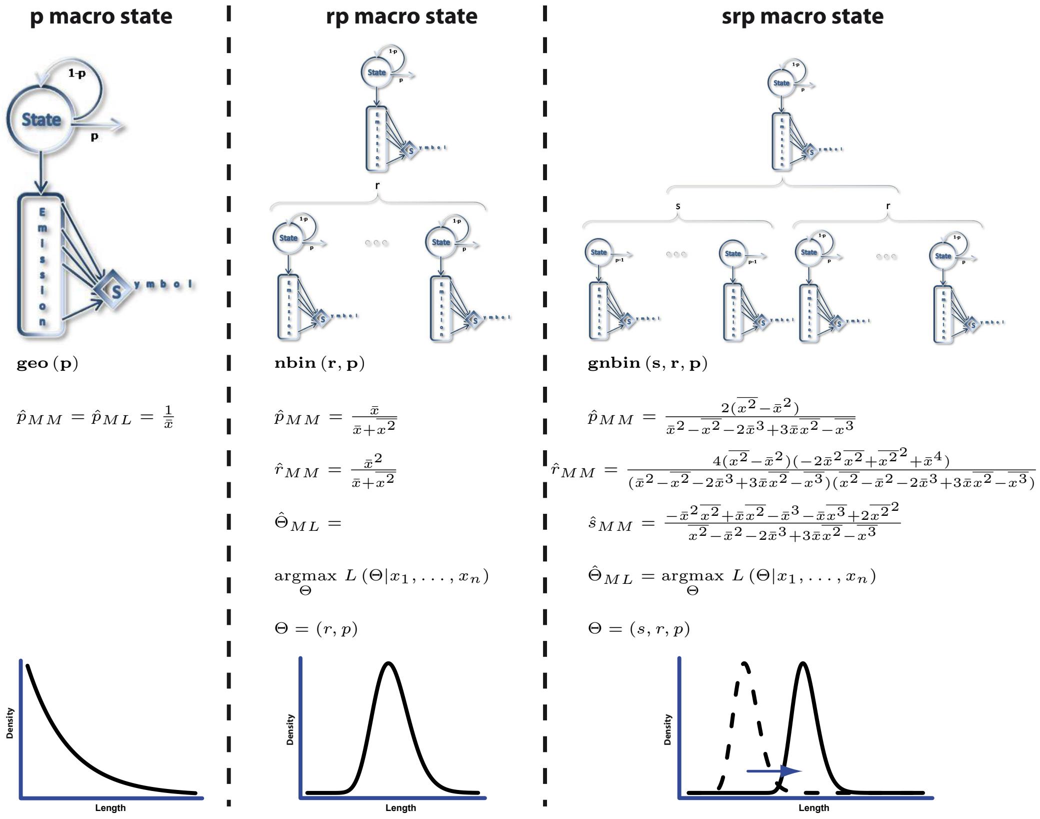

p macro state
1-p
State
p
Emission
Symbol
geo (p)
$\hat{p}_{MM} = \hat{p}_{ML} = \frac{1}{\bar{x}}$
Density
Length
rp macro state
r
State
1-p
p
Emission
Symbol
nbin (r, p)
$\hat{p}_{MM} = \frac{\bar{x}}{\bar{x}+\overline{x^2}}$
$\hat{r}_{MM} = \frac{\bar{x}^2}{\bar{x}+\overline{x^2}}$
$\hat{\Theta}_{ML} =$
$\underset{\Theta}{\text{argmax}}\ L\,(\Theta|x_1, \ldots, x_n)$
$\Theta = (r, p)$
Density
Length
srp macro state
s
r
State
1-p
p
p=1
Emission
Symbol
gnbin (s, r, p)
$\hat{p}_{MM} = \frac{2(\overline{x^2}-\bar{x}^2)}{\bar{x}^2-\overline{x^2}-2\bar{x}^3+3\bar{x}\overline{x^2}-\overline{x^3}}$
$\hat{r}_{MM} = \frac{4(\overline{x^2}-\bar{x}^2)(-2\bar{x}^2\overline{x^2}+\overline{x^2}^2+\bar{x}^4)}{(\bar{x}^2-\overline{x^2}-2\bar{x}^3+3\bar{x}\overline{x^2}-\overline{x^3})(\overline{x^2}-\bar{x}^2-2\bar{x}^3+3\bar{x}\overline{x^2}-\overline{x^3})}$
$\hat{s}_{MM} = \frac{-\bar{x}^2\overline{x^2}+\bar{x}\overline{x^2}-\bar{x}^3-\bar{x}\overline{x^3}+2\overline{x^2}^2}{\overline{x^2}-\bar{x}^2-2\bar{x}^3+3\bar{x}\overline{x^2}-\overline{x^3}}$
$\hat{\Theta}_{ML} = \underset{\Theta}{\text{argmax}}\ L\,(\Theta|x_1, \ldots, x_n)$
$\Theta = (s, r, p)$
Density
Length

Figure 2: Regarding the graphical representation of p, rp and srp macro states, the nested topology becomes visible easily. For rp, states are chained in a series of r repetitions. For the shifted srp model, s fixed copies lacking self transitions are prepended to the rp macro state. For each topology, estimators of distribution parameters s,r and p are given by maximum likelihood and the method of moments. Finally, an exemplary length distribution is depicted below each macro state.

towards p, the ML estimator $\hat{p} = \frac{1}{\bar{x}}$ is received. This p macro state represents the less complex form of length modelling and is concordant to a single self-transitive state. In Figure 2, the p macro state (left) as well as its ML and MM estimators are depicted together with a rough shape of the geometric distribution.

The rp macro state

emits negative binomial length distributed sequences ($nbin(r, p)$). As one might be aware of, the geometric distribution is contained in the negative binomial distribution when setting the parameter r to one. Thus, the rp macro state allows to model a geometric length distribution or the bell shape typically for the negative binomial distribution, regulated by the parameter r. Transferred to HMMs, r specifies the number of repetitive, chain-linked copies (see the middle part of Figure 2) of a self-transitional state with a holding time influenced by p. As the explicit mathematical calculation of an ML estimator for this case is very demanding, here the MM estimator is explained instead. The method of moments – as the name implies makes use of central and empirical moments to form an unbiased estimator. However for ML, still the maximal likelihood can be calculated by varying distribution parameters in a discrete space. The first and second moment, expectation (μ^1_{nbin}) and variance (μ^2_{nbin}) can be described by the distributions parameters r and p as

$$\mu^1_{nbin} \;=\; \frac{r}{p}, \qquad \mu^2_{nbin} = r\frac{1-p}{p^2}. \tag{3.14}$$

By exchanging theoretical with empirical moments,

$$\bar{x} \;=\; \frac{r}{p}, \qquad \overline{x^2} = r\frac{1-p}{p^2} \tag{3.15}$$

and resolving both equations to r and p

$$\hat{p} \;=\; \frac{\bar{x}}{\overline{x^2}+\bar{x}}, \qquad \hat{r} = \frac{\bar{x}^2}{\overline{x^2}+\bar{x}}. \tag{3.16}$$

both parameter estimators are obtained. Floating point estimations of r were rounded to the next integer value for a smooth integration into HMMs.

The srp macro state

emits shifted negative binomial length distributed sequences ($gnbin(s, r, p)$). This distribution is similar to the negative binomial distribution, except that an extension of exactly s characters or states increases the length of a sequence and so, shifts the distribution on the x-axis. Here, the srp macro state is equal to the rp macro state when setting the shift – s to zero. The moment generating function

$$M_{nbin}(t) \quad = \quad e^{t(r+s)}p^r(1 - e^t + e^t p)^{-r}. \tag{3.17}$$

was used to generate MM estimators for the parameters s, r and p similar to the rp macro state.

In the following, a short table summarizes macro state specific information on moments and PDFs for all three model variants.

	geometric	negative binomial	generalized negative binomial
macro state	p	rp	srp
PDF	$p(1-p)^{n-1}$	$\binom{n-1}{r-1}p^r(1-p)^{n-r}$	$\binom{n-s-1}{r-1}p^r(1-p)^{n-s-r}$
$E[X]$	$\frac{1}{p}$	$\frac{r}{p}$	$\frac{r}{p}+s$
$Var[X]$	$\frac{1-p}{p^2}$	$r\frac{1-p}{p^2}$	$r\frac{1-p}{p^2}$

Table 1: This table summarizes properties of different modelling topologies by providing information about the modelled length distribution, PDF, variance and expectation value.

3.3.2 Application on biological examples

Screening biological sequences in terms of length distributions revealed a variety of data sets following a thoroughly negative binomial law. Here, exemplary the length distribution of 232 β-sheet core regions from transmembrane proteins received from the TMPDB database (Figure 3 A) (Ikeda et al., 2003), signal peptides (Figure 3 B), 3'-UTR sequences of *C. elegans* from the UTRome database (Figure 3 C) (Mangone et al., 2008) and opening stem regions of the ITS2 from helix 3 obtained from *Asteraceae* and the ITS2 database (Figure 3 D) (Schultz et al., 2006; Selig et al., 2008; Koetschan et al., 2010) are illustrated. Each histogram represents the length distribution of data (black line), the estimated parameters including the corresponding distribution modelled by each macro state (dotted lines) as well as BIC and L_1 distance values. In addition, the optimal model choice is marked by a star.

Figure 3 A shows that rp and srp macro states are capable of modelling the transmembrane β sheets length distribution far better than the p macro state. Here, ML and MM estimators propose quite similar values.

In Figure 3 B, ML and MM estimators similarly propose 8 to 11 copies for the rp macro state to model the bell shape of signal peptides. Additionally, one can see the mismatch to length distributed data modelled by the p macro state. Käll et al. (2004) also modelled signal peptides by breaking the topology into several chained structural regions.

Due to the increasing number of sequenced genomes, the accurate identification of prominent parts like introns/exons, intergenic regions, splice sites or untranslated flanking regions gets further attention. Interestingly, the next example of 3'-UTR sequences (Figure 3 C) does not show a bell-shaped curve. Here, clearly a geometric distributed topology was preferred by all estimators. Thus, these data could be modelled best by a simple p macro state.

Figure 3 D shows slight differences between ML and MM estimates for the srp macro state on ITS2 secondary structure data. However, one can see that a bell shape is favoured from a geometric distribution model. This is also underlined when regarding BIC and L1 distance.

ITS2 secondary structure prediction

Focusing on this data, a more complex HMM was developed modelling not only helical parts but the whole structural conformation of the ITS2. Therefore, different structural parts with varying base compositions were inspected on their length distribution and parameters were estimated according to the previously introduced length modelling topologies (Chapter 3.3.1). According to the conserved core structure of the ITS2 (Coleman, 2003; Schultz et al., 2005), these parts were mainly categorized by opening/closing stem regions, loops or unpaired helix connecting regions for all four helices in the structure (Figure 4

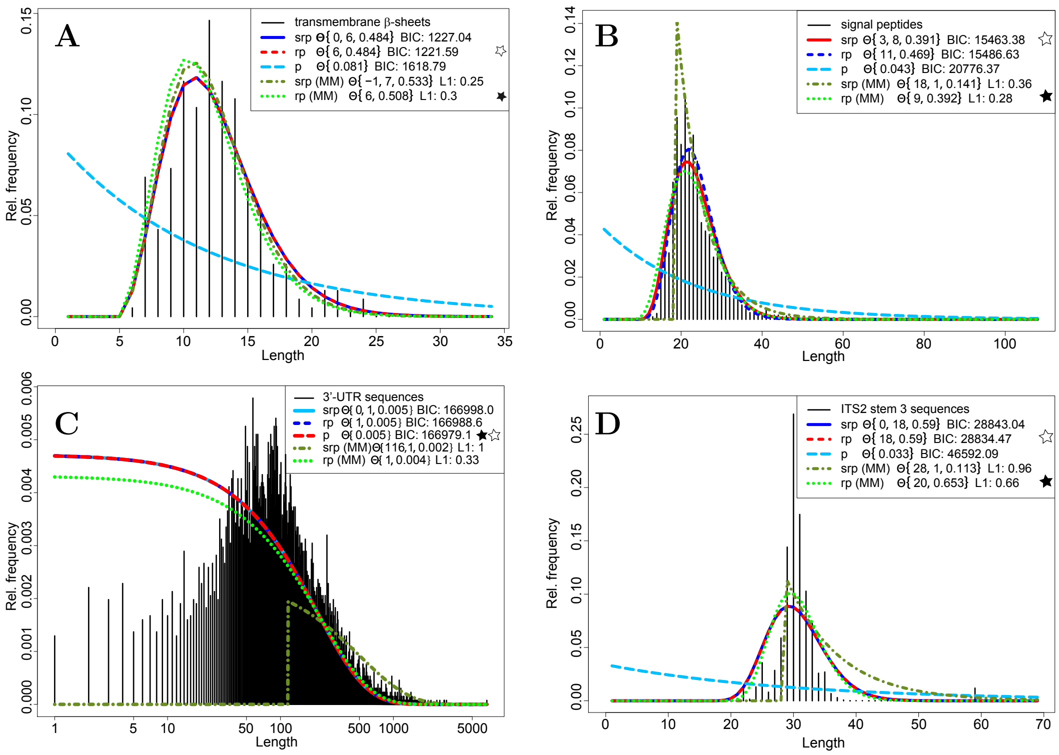


Figure 3: This Figure exhibits four different histograms, presenting the length distribution of 232 β-sheet core regions (A), signal peptides (B), 3'-UTR sequences of *C. elegans* (C) and opening stem regions from helix 3 of ITS2 sequences of family *Asteraceae* (D). Each graph shows estimated parameters by ML and MM for the macro states p, rp and srp. Further, optimal BIC and L1 distance values are highlighted by a star. When regarding the bell shapes of Figures A,B and D and comparing them to all three model topologies, one can see clearly a preference for the rp macro state. Where data is less bell shape distributed (C), the conventional HMM topology seems to perform best in this case.

illustrates the according macro states). Results of ML and MM estimations are highlighted in Table 2 for p - and rp macro states. Here, the preferred model topology is marked bold. As one can see, in most cases the optimized rp macro state is favoured by a better BIC (ML) or L_1 distance (MM). In the following, a ten-fold 90/10 cross validation on structure predictions of the p and rp macro state was performed on an asteraceaen ITS2 data set. The log ratio between errors made by p and rp macro states in comparison to reference structures of

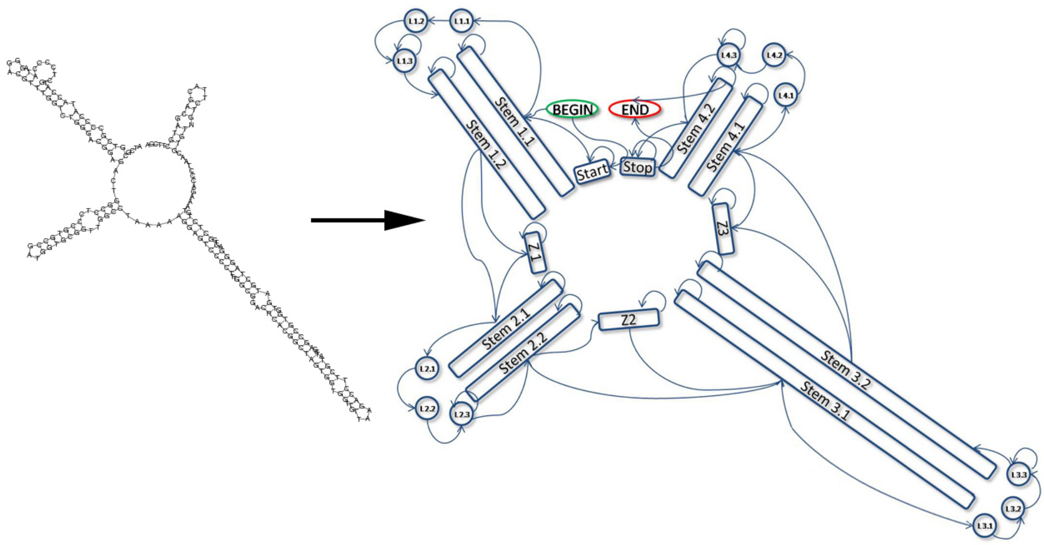


Figure 4: This HMM (right) reflects a mapping of structural regions from *Centaurea alba* ITS2 secondary structure (left), here synonymously representing the conserved core structure of this spacer. Opening and closing stem regions ("Stem X.1"/"Stem X.2"), unpaired loops and inter helix regions are modelled by different macro states. These sometimes may be skipped and were not always present in the training data set of *Asteraceae*. All states except first and second loop states are self-transitive. This guarantees loops with a length of at least three nucleotides.

the ITS2 database are visualized in Figure 5. Here, one can clearly see the improved accuracy of the rp macro state resulting from the HMM optimization. However, also some (blue) regions exist where the estimator favoured a model which turned out to be rather suboptimal. The significance of this topology optimization was further documented by a Wilcoxon rank sum test - proving a significant error reduction averaged over all cross fold estimations of 34,25% for the rp model. Here, all ten p-values were very significant and smaller than 2.26×10^{-4}.

State	r_{MM}, p_{MM}	$L_1^{MM}(p)$	$L_1^{MM}(rp)$	r_{ML}, p_{ML}	BIC_p^{ML}	BIC_{rp}^{ML}
Start	(2, 0.62)	1.37	**1.20**	(1, 0.24)	**14866.57**	14875.33
Stem 1.1	(19, 0.89)	2.00	**0.80**	(19, 0.88)	26262.90	**11962.26**
Loop 1.3	(2, 0.87)	**1.14**	1.14	(1, 0.46)	**9694.14**	9702.91
Stem 1.2	(22, 0.93)	2.00	**1.13**	(21, 0.90)	26598.26	**11159.77**
Z1	(2, 0.87)	**1.25**	1.29	(1, 0.38)	**11304.23**	11313.00
Stem 2.1	(13, 0.98)	2.00	**0.78**	(13, 0.98)	22871.83	**4286.04**
Loop 2.3	(1, 0.75)	**0.49**	0.49	(1, 0.68)	**5893.03**	5901.80
Stem 2.2	(13, 0.98)	2.00	**0.78**	(13, 0.98)	22872.92	**4303.24**
Z2	(4, 0.88)	1.81	**0.96**	(3, 0.69)	15138.06	**9136.57**
Stem 3.1	(38, 0.99)	2.00	**0.95**	(37, 0.97)	29828.17	**7933.07**
Loop 3.3	(3, 1.00)	1.59	**0.59**	(2, 0.66)	12310.04	**7853.56**
Stem 3.2	(43, 0.96)	2.00	**0.77**	(43, 0.95)	30942.04	**11300.02**
Z3	(11, 0.91)	1.93	**1.14**	(9, 0.77)	22013.25	**11110.74**
Stem 4.1	(8, 0.92)	2.00	**1.03**	(8, 0.95)	19777.04	**5853.66**
Loop 4.3	(2, 1.00)	1.79	**0.36**	(1, 0.50)	**8939.28**	8948.05
Stem 4.2	(8, 0.93)	2.00	**1.11**	(8, 0.96)	19726.27	**5433.01**
Stop	(4, 0.79)	1.50	**1.07**	(3, 0.59)	16245.67	**11259.91**

Table 2: This Table illustrates different macro states of an HMM modelling ITS2 secondary structure together with corresponding ML and MM parameter estimates. In addition, BIC and L1 distance represent optimal model choice and best values are marked bold. One can clearly see a favour for the rp macro state topology which models, inter alia, bell-shaped length distributions.

3.4 Discussion

In this section, three HMM topologies were introduced modelling different length distributions. The first method, the p macro state equals to a conventional self transitional HMM state emitting geometrically distributed sequence lengths. The final optimized srp macro state is not only capable of modelling geometrically distributed data but also models bell-shaped negative binomial length distributions even in a shifted form. This allows to describe and model data more accurately in terms of HMMs. Further, no changes in decoding

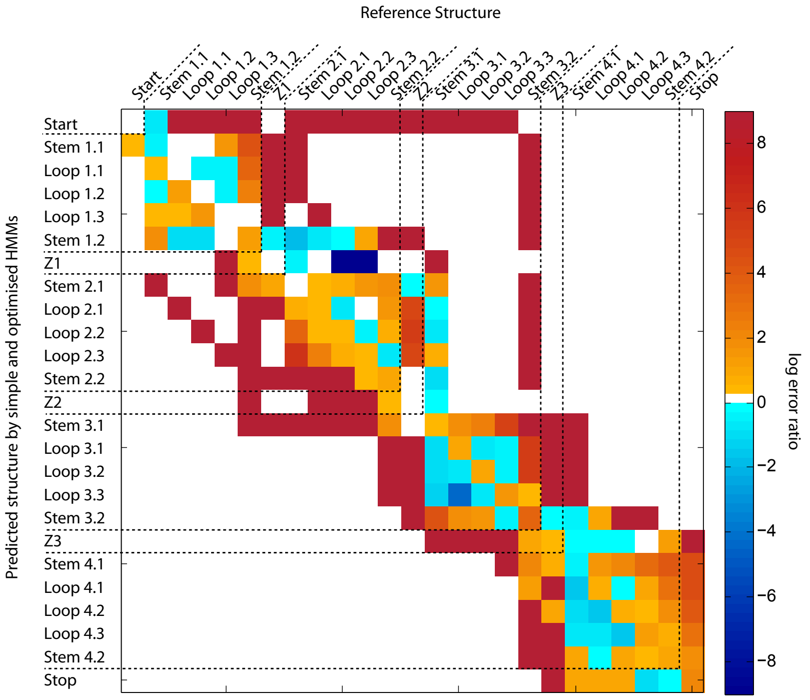


Figure 5: This heat map illustrates structural regions resulting from a ten fold cross validation test between p and rp macro states, modelling ITS2 secondary structure from *Asteraceae*. Here, the log error ratio of p and rp macro states in comparison to a high quality reference structure from the ITS2 database is highlighted in different colors. Reddish marked fields indicate a better, less error-prone modelling of the rp macro state, while blueish fields mark structural positions where a simple p macro state topology returned more accurate predictions.

or training algorithms have to be made, as this optimization changes just the HMMs topology - thus can be integrated into present applications with a minimum effort. To receive optimal model parameters, two estimators, the method of moments and the maximum likelihood method were derived. The latter is known to be asymptotically efficient, however sometimes is still very time consuming. In contrast to ML, the method of moments is able to provide all estimates in a relatively quick time.

To underline the performance of topology optimizations, applications on artificial data (not shown) as well as on biological examples were evaluated. Figure 3 points out the relevance to actual data by providing several biological examples implicating a bell-shaped length distribution from current databases. The further reviewed/examined novel method for secondary structure prediction from ITS2 sequences revealed a statistically very significant performance gain when modelling structure by an optimized HMM topology. Although this method does not provide valid base pairings at the moment, it could be further augmented to a level of support vector machines for stochastic context-free grammars which might be capable of retaining base pairings by emitting both binding nucleotides of a helix within one atomic step.

In conclusion, this method allows an easy design of new, as well as adaptation of current established HMM topologies by providing two estimators which fit the HMM topology to the underlying length distribution of data. Bilmes (2004) even describes bimodal HMM topologies which leaves room for more accurate predictions and complex length distributions.

4 Sequence-structure phylogeny

4.1 Introduction

Having the necessary requirements to annotate and predict a structure, the next steps in a sequence-structure based phylogeny are the alignment and the tree reconstruction. A set of tools was developed that deals already with sequence-structure based alignments (Seibel et al., 2006, 2008; Bauer et al., 2007; Siebert and Backofen, 2005), however different alphabets are imaginable for this task. For the reconstruction of phylogenetic trees instead, only one distance based tree inference method (ProfDistS) (Wolf et al., 2008) is currently able to handle sequence-structure data. In the following, an R-package (treeforge) was created which provides Neighbour Joining (Saitou and Nei, 1987; Gascuel, 1997), Maximum Parsimony (Camin and Sokal, 1965) and Maximum Likelihood (Felsenstein, 1981, 1985, 2004) tree reconstruction on four different sequence-structure alphabets. Further, different substitution models like jukes cantor (Jukes and Cantor, 1969), GTR (Tavaré, 1986) - as well as a marker specific ITS2 model (Müller et al., 2002; Wolf et al., 2005a) are available. To cope with the newly developed sequence-structure alphabets, new scoring matrices were evaluated and integrated into treeforge, resulting in a rapid prototyping R-package for phylogenetic analysis.

4.2 Tree inference with the R-package treeforge

The treeforge R package enables a user to align sequence and sequence-structure data in FASTA (Ncbi, 2007) or xFASTA (xFASTA enhances FASTA by an additional line containing structure information in bracket dot bracket notation below a sequence) format either with the alignment algorithm of ClustalW2 (Thompson et al., 1994; Larkin et al., 2007) or MUSCLE (Edgar, 2004a,b). Additionally, a various number of different sequence-structure coding alphabets (Chapter 4.2.1) are available. For phylogenetic tree reconstruction, the following three methods are implemented and available for all types of alphabets: Maximum Parsimony, Neighbour-Joining using the BIONJ algorithm (Gascuel, 1997) and Maximum Likelihood. For these methods, different substitution models are available. Finally, trees can easily be visualized or exported to standard Newick format (Felsenstein et al., 1986). The R-package, including a vignette explaining the basic usage of the package is available on CD with this dissertation.

4.2.1 Sequence-structure coding alphabets

Four different alphabets to model sequence or sequence-structure data were defined. The alphabet 'DNA' merely codes the four base nucleotides A,C,G,T. Any other IUPAC characters are translated into the ambiguity character N, except U - which also codes for T. DNA is the only alphabet that doesn't contain any structure information.

'RNA10' is the most sparse alphabet containing sequence and structure information together. It codes the four base nucleotides A,C,G,T in combination with an unpaired structure, plus the structural bonds between A ↔ T T ↔ A, C ↔ G G ↔ C and G ↔ T T ↔ G. Additionally, Uracil and other IUPAC characters are handled equivalently to the DNA alphabet.

An alphabet that doesn't consider horizontal dependencies between structural bonds is 'RNA12'. Here, each of the four base nucleotides A,C,G,T is coupled with one of the three structural conformations '.','(' and ')'. This results in 12 base substitution plus some additional ones for Uracil and the other less specific IUPAC characters. This alphabet is best evaluated and implemented in the standalone software 4SALE (Seibel et al., 2006, 2008), ProfDistS (Wolf et al., 2008) and the ITS2 workbench.

'RNA16' is the alphabet containing most, even though redundant information and was first described by Smith et al. (2004). Similar to 'RNA10', horizontal dependencies between structural bonds are encoded. In addition to 'RNA10', here each side of a bond is coded by a different character. This results in the four unpaired structural base substitutions A,C,G,T plus the six ones from 'RNA10' - A ↔ T T ↔ A, C ↔ G G ↔ C and G ↔ T T ↔ G - each one coded twice for the 3' and the 5' side of a secondary structure. Additionally, further IUPAC characters as well as Uracil are treated equivalently to the other alphabets.

4.2.2 Substitution models for phylogenetic reconstructions

For each phylogenetic tree reconstruction method, different evolutionary models are available. These models specify different substitution rates for alphabet characters and are represented by a substitution matrix Q which fulfils the following requirements:

$$Q_{ii} = -\sum_{i \neq j} Q_{ij} \tag{4.1}$$

$$\pi Q = 0 \tag{4.2}$$

$$\pi_i q_{ij} = \pi_j q_{ji} \tag{4.3}$$

with $\pi = \{\pi_1 \dots \pi_n\}$ representing the stationary distribution of the alphabet.

The Jukes Cantor (JC) rate matrix is one of the simplest substitution models and assumes an equal appearance of nucleotides as well as an equal distribution of nucleotide mutations (Jukes and Cantor, 1969). Thus, a rate matrix describing the rate q_{ij} which is a point mutation from character i into character j can be easily described by:

$$Q = \begin{pmatrix} -(n-1)\lambda & \dots & \lambda \\ \vdots & \ddots & \vdots \\ \lambda & \dots & -(n-1)\lambda \end{pmatrix} \tag{4.4}$$

Only one free parameter λ controls the overall substitution rate. When calibrated to 1 PEM (Percent of Expected Mutations),

$$-0.01 = \sum_{i=1}^{n} \pi_i Q_{ii} = -(n-1)\lambda \tag{4.5}$$

$$\Rightarrow \lambda = \frac{1}{(n-1)100}. \tag{4.6}$$

JC relies on a symmetric substitution matrix and thus is time reversible. In more detail, a mutation from A → T occurs with the same rate as a mutation from T → A.

A distance based on this model is described by

$$d_{jc} = -\frac{n-1}{n} \log_e(1 - \frac{n-1}{n} p) \tag{4.7}$$

where p is the proportion of sites with different characters. In addition to an uncorrected hamming distance matrix, Juces Cantor also regards silent (synonymous A → T → A) mutations in his correction model.

ITS2 is a substitution model which is explicitly based on the according marker (Müller et al., 2002; Wolf et al., 2005a). Thus, distributions for base pair mutations and nucleotide occurrences were pre-calculated on a set of reliable ITS2 sequences.

GTR, the General Time Reversible model similar to JC assumes a symmetric substitution matrix and thus is called time reversible (Tavaré, 1986). It is the most general model possible and in contrast to JC, here, the rate between different substitution pairs, as well as the nucleotide frequencies can vary. Exemplary, Q is shown for a four character alphabet

$$Q = \begin{pmatrix} -(x_1 + x_2 + x_3) & x_1 & x_2 & x_3 \\ \frac{\pi_1 x_1}{\pi_2} & -(\frac{\pi_1 x_1}{\pi_2} + x_4 + x_5) & x_4 & x_5 \\ \frac{\pi_1 x_2}{\pi_3} & \frac{\pi_2 x_4}{\pi_3} & -(\frac{\pi_1 x_2}{\pi_3} + \frac{\pi_2 x_4}{\pi_3} + x_6) & x_6 \\ \frac{\pi_1 x_3}{\pi_4} & \frac{\pi_2 x_5}{\pi_4} & \frac{\pi_3 x_6}{\pi_4} & -(\frac{\pi_1 x_3}{\pi_4} + \frac{\pi_2 x_5}{\pi_4} + \frac{\pi_3 x_6}{\pi_4}) \end{pmatrix} \tag{4.8}$$

Due to the large number of free parameters (in this case four base frequencies and six substitution rates except one substitution which serves as normalization constant for the others and is equal to one), a PAM (Percent of Accepted Mutations) distance and parameters of the GTR model are typically calculated by a maximum likelihood approach.

In addition to JC, ITS2 and GTR, for the maximum likelihood tree calculation, a more advanced JC_IG, ITS2_IG and GTR_IG substitution model also allows to model invariable sites (I) and rate heterogeneity of the substitution process (G) by allowing up to four different speeds of mutation rates.

Finally, bootstrap support is available for all of the methods and models.

4.3 Calculating evolutionary rates and scoring sequences

A score matrix is used to score the similarity of two nucleotides, amino acids, proteins or encoded sequence-structure tuples and is applied in programs for sequence alignment or BLAST (Altschul et al., 1997, 1990), etc.. It goes back to the calculations by Dayhoff et al. (1978), which first defined a scoring matrix for amino acid similarities and the work of Henikoff and Henikoff (1992) and the BLOSUM62 matrix. Typically, the main diagonal - scoring identical characters gets a positive, other diagonals a more negative score (exemplary, the tuple 'A.' ↔ 'A(' should be scored higher than 'A.' ↔ 'T('). A very simple score matrix is the identity matrix, scoring equal positions with 1, others with 0.

But as nucleotides and evolutionary rates are not equally distributed in nature, Müller et al. (2002); Wolf et al. (2005a) have developed a specific score matrix specifically for the marker ITS2. In this study, the method was further applied on the newly developed alphabets 'RNA10' and 'RNA16', and reimplemented for the old alphabets 'RNA12' and 'DNA'.

The whole process was repeated iteratively, meaning a first calculated score-matrix served as input for new alignments in the next iteration until

$$|S_{ij}^{old} - S_{ij}^{new}| < \epsilon. \tag{4.9}$$

Gap-costs were set to (0,0) according to 4SALE. The dataset was based on 400/1200 manually inspected sequences and structures of the same genus/species taken from the CBC analysis of Müller et al. (2007).

Detailed estimation process:

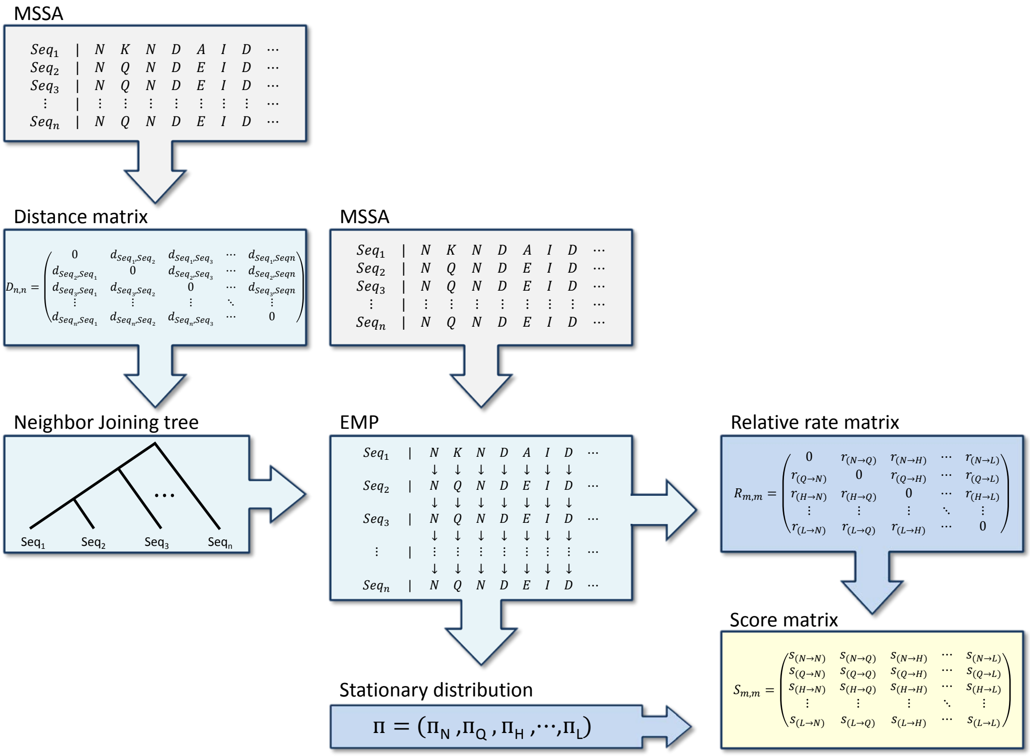


Figure 6: Estimation of a score matrix based on an evolutionary markov process. From a multiple sequence-structure alignment, a distance matrix and neighbour joining tree are calculated. This tree together with the alignment serves as input for calculating an EMP. The EMP models point mutations in a column of the alignment and uses the tree distances for time calibration. From this EMP, a stationary distribution of the alphabet, as well as the relative rate matrix are derived. They are prerequisites for the score matrix calculation.

- The estimation starts with the generation of Multiple Sequence-Structure Algnments (MSSAs) for the species and genus datasets with a specific alphabet. For these alignments, the identity matrix or a known matrix from literature is suitable.

- After the alignment, a distance matrix for each sequence tupel is calculated by counting and normalizing the differences in both sequences.

From this matrix, a Neighbour Joining tree is calculated easily.

- This tree, together with the MSSA is needed for modelling an Evolutionary Markov Process (EMP) (Dayhoff et al., 1978; Müller, 2001; Müller and Vingron, 2000; Müller et al., 2002; Schliep, 2011) along the columns of the alignment. The EMP allows to describe the development at one fixed position of a sequence, influenced by point mutations under evolutionary circumstances. The distance of two sequences in the tree serves as parameter for describing the speed of one sequence mutating into another. This further allows to calculate a rate for each possible substitution by calibrating the time to one mutation at every 100 positions in a sequence (1 PEM).

First, a transition-matrix $P(t)$ describes the time discrete markov chain with transitions of character i mutating into character j. Further, the condition of the process is assumed to be ergodic, which means that

$$\pi = \pi P(t). \tag{4.10}$$

All outgoing transition probabilities must be positive and sum to one

$$\begin{aligned} \sum_j p_{i,j} &= 1 && (4.11) \\ p_{ij} &> 0. && (4.12) \end{aligned}$$

The rate matrix Q is then defined as the change of P in an infinitely small time period

$$\begin{aligned} \lim_{t \to \infty} \frac{P(t) - I}{t} &= Q && (4.13) \\ \Rightarrow P(t) &= e^{tQ}. && (4.14) \end{aligned}$$

The estimation of the EMP is implemented in the R-package 'phangorn' by Schliep (2011).

- With the EMP, a relative rate matrix R, which is basically similar to Q but excludes base frequencies, and the stationary distribution finally are the last requirements for calculating the score matrix

$$\begin{aligned} q_{ij} &= \pi_i r_{ij} & (4.15)\\ S(t) &= \log(\frac{M(t)}{M(\infty)}) & (4.16)\\ \text{with } M(t) &= \text{diag}(\pi)e^{tQ} & (4.17) \end{aligned}$$

  S(t) is derived according to the log-odds formula by Dayhoff et al. (1978). $M(T)$ is the matrix corresponding to the distribution of character substitutions with $t = 50$ PAM. From both, genus and species datasets, only the medians were used to build one overall genus/species relative rate matrix and one overall genus/species base frequency distribution which served as input for the score matrix calculation.

Finally, the score matrix is visualized in the bubble plots of Figures 40 and 41. Here, red dots indicate a positive, green dots a negative score. The size of each bubble corresponds to the value of its absolute number. Thus, the main diagonal axis typically scores equal positions with a positive score. Other diagonals are scored more negatively, depending on the number of exchanges which occurred for the specific tuples. When comparing species and genus matrices, one might easily recognize a difference in the G-T rates for example. This might give motivation to further analyse the differences between species and genus specific evolutionary rates.

4.4 Reconstruction of the chlorophyceaen tree with different coding alphabets

Based on the newly estimated score matrices for each alphabet, a data set of Chlorophyta, equal to Buchheim et al. (2011b) was used for phylogenetic tree reconstruction of the chlorophyceaen class of green algae with treeforge. This phylogeny is known to be not easy reconstructible due to several jumping clades (Keller et al., 2008). In total, twelve trees with 100 bootstrap replicas were reconstructed resulting from the combination of the three methods mp_phylo, ml_phylo and dist_phylo and the four alphabets 'DNA', 'RNA10', 'RNA12' and 'RNA16'. Gap-costs were set to (0,0) while using the model GTR_IG for the ML reconstruction, the model GTR for the BIONJ tree. The GTR_IG model resulted as the optimal choice from a model-test (AIC) between all implemented substitution models (data not shown). The reconstruction of a tree with treeforge just requires a few lines of code:

```
alignment <- file2phy(file="Chlorophyta.xfasta",type='xfasta',
                      alphabet='RNA10',go=0,ge=0)
fitobj <- ml_phylo(phydat=alignment,nbs=100,model="GTR_IG")
ml_tree_rna10 <- fitobj@tree
```

To compare the final twelve trees they were visualized by a multidimensional scaling (MDS) plot (Figure 7). Here, the topological differences between multiple trees become visible on the two main axes x and y by calculating an all-against-all distance matrix and transferring this information onto two axes. A clustering on tree reconstruction methods can be clearly seen on the y-axis. However, this method visualizes the distances between branches weighted equally which makes it difficult to compare the main topological differences on higher ranks. Therefore, the trees were sorted by their alliances of the main clades which are the OCC group consisting of the orders Oedogoniales, Chaetopeltidales and Chaetophorales, the CW group of Chlamydomonadales and the DO group of Sphaeropleales. Additionally, two Trebouxiophyceaen sequences (*Parachlorella beijerinckii*, *Chlorella sorokiniana*) and the outgroup of Ulvophyceae (*Acrochaete sp. 6 BER 2007*, *Ulva laetevirens*) where included in the dataset. In the following, the alphabet 'RNA16' with methods ML and BIONJ and the alphabet 'RNA10' with method MP reconstructed a tree with a similar topology to Buchheim et al. (2011b); Brouard et al. (2010); Turmel et al. (2008). For example the MP tree of the alphabet 'RNA10' is visualized below (Figure 8).

It consists of the DO group (Sphaeropleales) and the CW group (Chlamydomonadales) as a monophyletic sister group with a high bootstrap support of

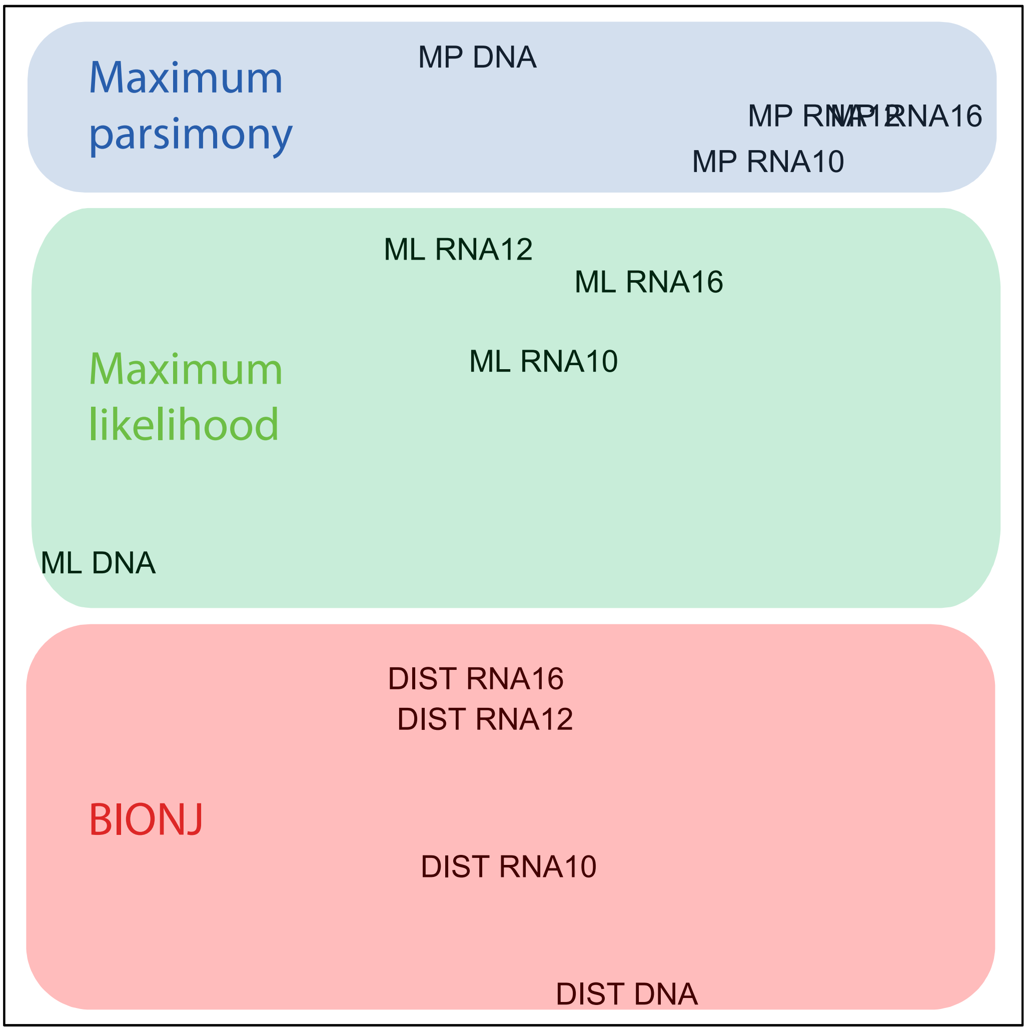


Figure 7: A multidimensional scaling of the alphabets 'DNA', 'RNA10', 'RNA12' and 'RNA16' in combination with the tree reconstruction methods MP, ML and BIONJ is visualized above. Here, similar methods cluster on the y-axis.

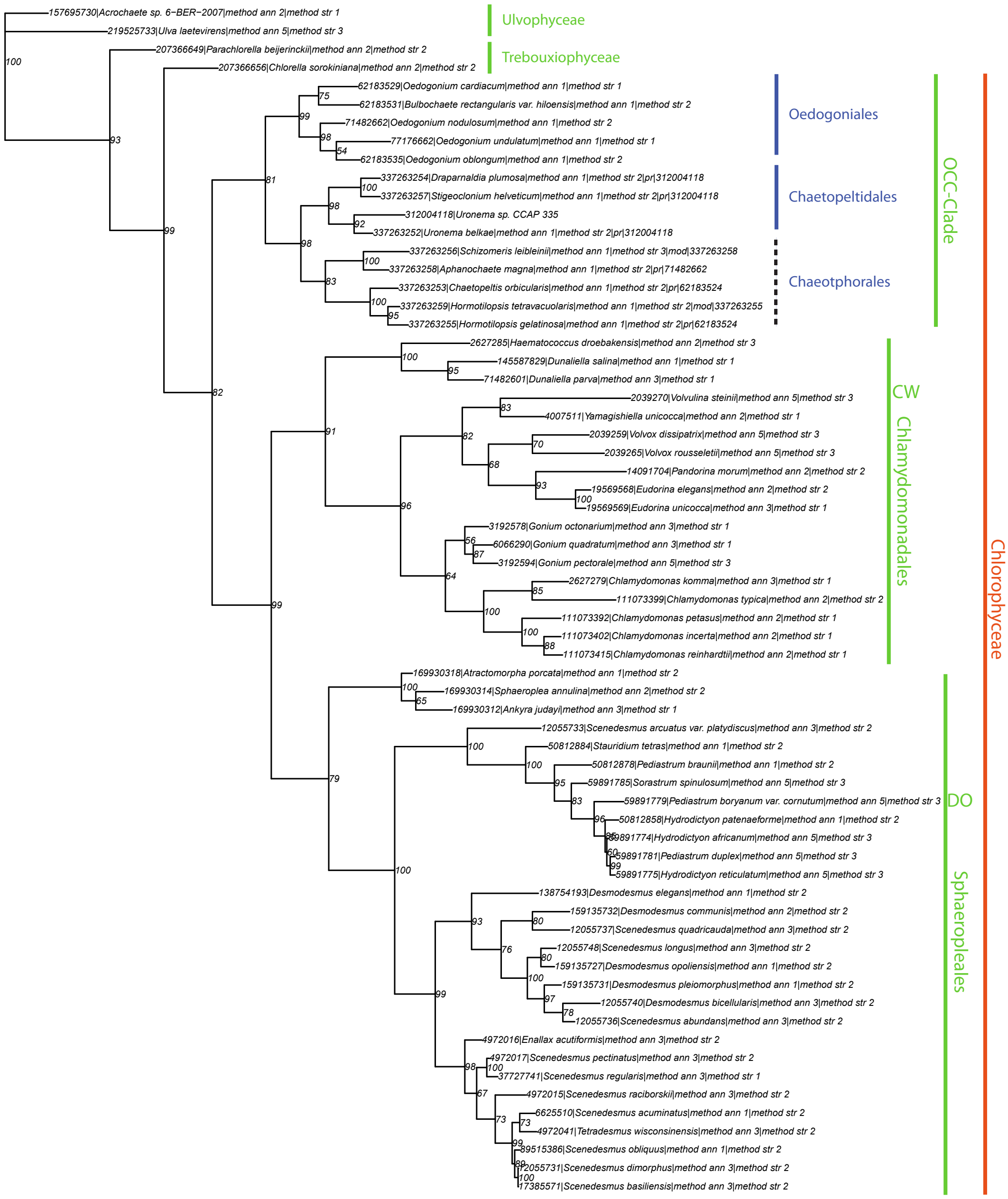

157695730|Acrochaete sp. 6−BER−2007|method ann 2|method str 1
219525733|Ulva laetevirens|method ann 5|method str 3
207366649|Parachlorella beijerinckii|method ann 2|method str 2
207366656|Chlorella sorokiniana|method ann 2|method str 2
62183529|Oedogonium cardiacum|method ann 1|method str 1
62183531|Bulbochaete rectangularis var. hiloensis|method ann 1|method str 2
71482662|Oedogonium nodulosum|method ann 1|method str 2
77176662|Oedogonium undulatum|method ann 1|method str 1
62183535|Oedogonium oblongum|method ann 1|method str 2
337263254|Draparnaldia plumosa|method ann 1|method str 2|pr|312004118
337263257|Stigeoclonium helveticum|method ann 1|method str 2|pr|312004118
312004118|Uronema sp. CCAP 335
337263252|Uronema belkae|method ann 1|method str 2|pr|312004118
337263256|Schizomeris leibleinii|method ann 1|method str 3|mod|337263258
337263258|Aphanochaete magna|method ann 1|method str 2|pr|71482662
337263253|Chaetopeltis orbicularis|method ann 1|method str 2|pr|62183524
337263259|Hormotilopsis tetravacuolaris|method ann 1|method str 2|mod|337263255
337263255|Hormotilopsis gelatinosa|method ann 1|method str 2|pr|62183524
2627285|Haematococcus droebakensis|method ann 2|method str 3
145587829|Dunaliella salina|method ann 1|method str 1
71482601|Dunaliella parva|method ann 3|method str 1
2039270|Volvulina steinii|method ann 5|method str 3
4007511|Yamagishiella unicocca|method ann 2|method str 1
2039259|Volvox dissipatrix|method ann 5|method str 3
2039265|Volvox rousseletii|method ann 5|method str 3
14091704|Pandorina morum|method ann 2|method str 2
19569568|Eudorina elegans|method ann 2|method str 2
19569569|Eudorina unicocca|method ann 3|method str 1
3192578|Gonium octonarium|method ann 3|method str 1
6066290|Gonium quadratum|method ann 3|method str 1
3192594|Gonium pectorale|method ann 5|method str 3
2627279|Chlamydomonas komma|method ann 3|method str 1
111073399|Chlamydomonas typica|method ann 2|method str 2
111073392|Chlamydomonas petasus|method ann 2|method str 1
111073402|Chlamydomonas incerta|method ann 2|method str 1
111073415|Chlamydomonas reinhardtii|method ann 2|method str 1
169930318|Atractomorpha porcata|method ann 1|method str 2
169930314|Sphaeroplea annulina|method ann 2|method str 2
169930312|Ankyra judayi|method ann 3|method str 1
12055733|Scenedesmus arcuatus var. platydiscus|method ann 3|method str 2
50812884|Stauridium tetras|method ann 1|method str 2
50812878|Pediastrum braunii|method ann 1|method str 2
59891785|Sorastrum spinulosum|method ann 5|method str 3
59891779|Pediastrum boryanum var. cornutum|method ann 5|method str 3
50812858|Hydrodictyon patenaeforme|method ann 1|method str 2
59891774|Hydrodictyon africanum|method ann 5|method str 3
59891781|Pediastrum duplex|method ann 5|method str 3
59891775|Hydrodictyon reticulatum|method ann 5|method str 3
138754193|Desmodesmus elegans|method ann 1|method str 2
159135732|Desmodesmus communis|method ann 2|method str 2
12055737|Scenedesmus quadricauda|method ann 3|method str 2
12055748|Scenedesmus longus|method ann 3|method str 2
159135727|Desmodesmus opoliensis|method ann 1|method str 2
159135731|Desmodesmus pleiomorphus|method ann 1|method str 2
12055740|Desmodesmus bicellularis|method ann 3|method str 2
12055736|Scenedesmus abundans|method ann 3|method str 2
4972016|Enallax acutiformis|method ann 3|method str 2
4972017|Scenedesmus pectinatus|method ann 3|method str 2
37727741|Scenedesmus regularis|method ann 3|method str 1
4972015|Scenedesmus raciborskii|method ann 3|method str 2
6625510|Scenedesmus acuminatus|method ann 1|method str 2
4972041|Tetradesmus wisconsinensis|method ann 3|method str 2
89515386|Scenedesmus obliquus|method ann 1|method str 2
12055731|Scenedesmus dimorphus|method ann 3|method str 2
17385571|Scenedesmus basiliensis|method ann 3|method str 2
Ulvophyceae
Trebouxiophyceae
Oedogoniales
Chaetopeltidales
Chaeotphorales
OCC-Clade
CW
Chlamydomonadales
DO
Sphaeropleales
Chlorophyceae

Figure 8: This tree of Chlorophyceae was calculated by maximum parsimony on the alphabet 'RNA10'. It resolves the DO and CW group as monophyletic sister groups with a high bootstrap support of 99. Both are further in alliance with the OCC clade and a bootstrap support of 82. This topology has a similar conformation to the published phylogeny of Buchheim et al. (2011b); Brouard et al. (2010); Turmel et al. (2008).

99. Both are in alliance with the OCC clade with a lower bootstrap support of 82. In this clade, Chaetopeltidales and Chaeotphorales are resolved as sister group with a high bootstrap support of 98.

Similarly, the alphabet 'RNA16' with method MP, the alphabet 'RNA10' with method ML, and the 'DNA' alphabet with BIONJ algorithm reconstructed trees resolving a slight difference to the previous reconstructed tree when regarding the upper ranks. Here, as example the MP tree of the alphabet 'RNA16' is represented for this group in Figure 9.
In this case, the OCC group forms a sister clade with the CW group of Chlamydomonadales, albeit the bootstrap support is not very robust for all three trees at this position (79 for MP with 'RNA16', 53 for ML with 'RNA10' and 28 for the 'DNA' BIONJ tree). For the method ML with alphabet 'RNA10', the OCC clade was resolved in congruence to the first group of trees, classifying the order of Oedogoniales as sister group to the orders of Chaetopeltidales and Chaetophorales, however with a low bootstrap support of 64. For the other two methods – MP with alphabet 'RNA16' and BIONJ with alphabet 'DNA' these clades were resolved differently.

Similar to the topology above, the alphabet 'RNA12' for the methods ML, MP and BIONJ resolved trees with the OCC clade being a sister group to the CW group of Chlamydomonadales, and both being in alliance with the DO group of Sphaeropleales, however the group of *Ankyra judayi*, *Sphaeroplea annulina* and *Atractomorpha porcata* jumps between the clades of OCC (MP, BIONJ) and CW (ML).

The remaining methods, MP and ML with alphabet 'DNA', as well as BIONJ with alphabet 'RNA10' mostly did not resolve the OCC clade as monophyletic group. Further, all trees are available on CD with this dissertation.

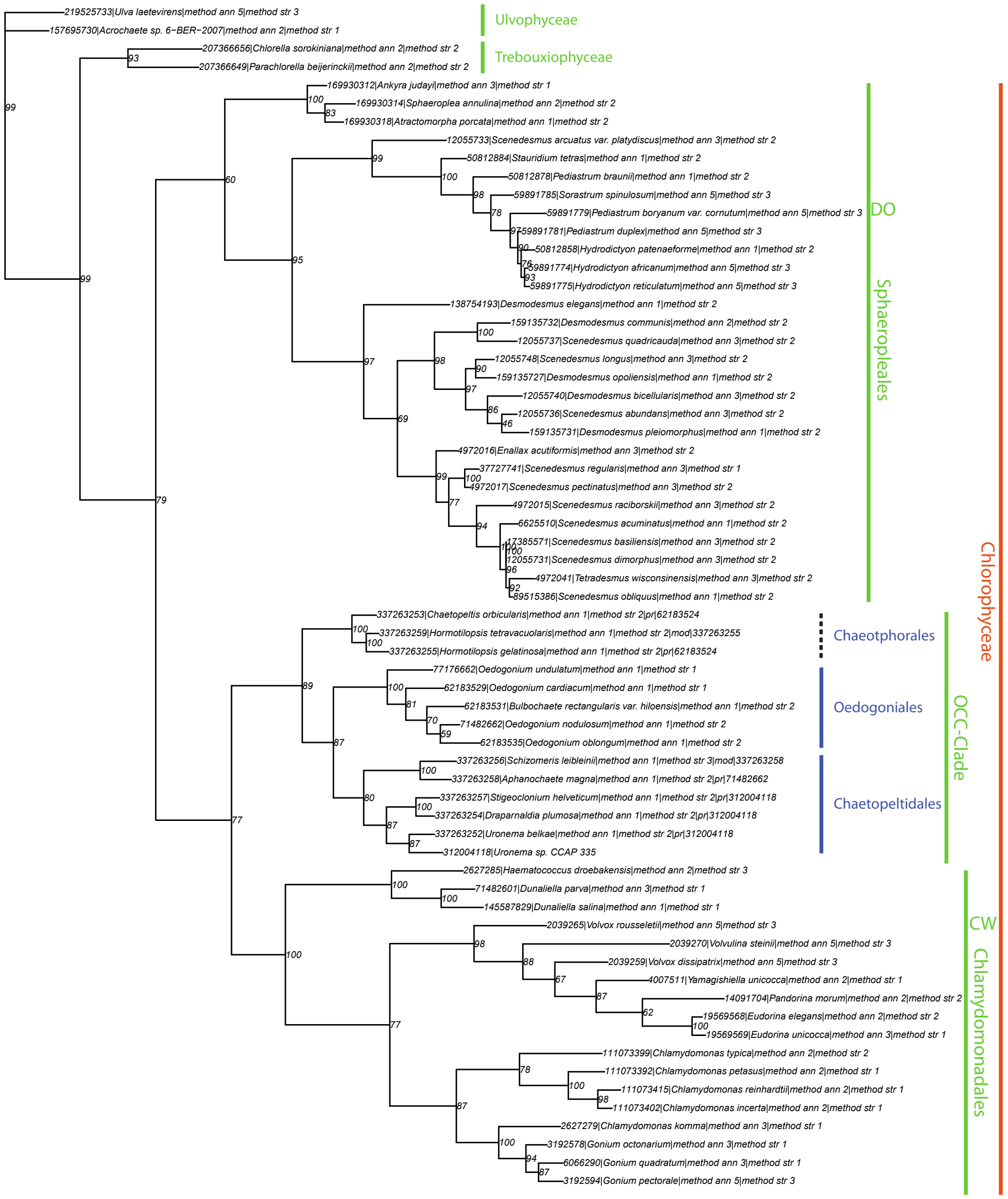

219525733|Ulva laetevirens|method ann 5|method str 3
157695730|Acrochaete sp. 6−BER−2007|method ann 2|method str 1
207366656|Chlorella sorokiniana|method ann 2|method str 2
93
207366649|Parachlorella beijerinckii|method ann 2|method str 2
169930312|Ankyra judayi|method ann 3|method str 1
100
169930314|Sphaeroplea annulina|method ann 2|method str 2
99
83
169930318|Atractomorpha porcata|method ann 1|method str 2
12055733|Scenedesmus arcuatus var. platydiscus|method ann 3|method str 2
99
50812884|Stauridium tetras|method ann 1|method str 2
100
50812878|Pediastrum braunii|method ann 1|method str 2
60
98
59891785|Sorastrum spinulosum|method ann 5|method str 3
78
59891779|Pediastrum boryanum var. cornutum|method ann 5|method str 3
97
59891781|Pediastrum duplex|method ann 5|method str 3
90
50812858|Hydrodictyon patenaeforme|method ann 1|method str 2
95
76
59891774|Hydrodictyon africanum|method ann 5|method str 3
93
99
59891775|Hydrodictyon reticulatum|method ann 5|method str 3
138754193|Desmodesmus elegans|method ann 1|method str 2
159135732|Desmodesmus communis|method ann 2|method str 2
100
12055737|Scenedesmus quadricauda|method ann 3|method str 2
12055748|Scenedesmus longus|method ann 3|method str 2
98
97
90
159135727|Desmodesmus opoliensis|method ann 1|method str 2
97
12055740|Desmodesmus bicellularis|method ann 3|method str 2
86
12055736|Scenedesmus abundans|method ann 3|method str 2
69
46
159135731|Desmodesmus pleiomorphus|method ann 1|method str 2
4972016|Enallax acutiformis|method ann 3|method str 2
37727741|Scenedesmus regularis|method ann 3|method str 1
99
100
4972017|Scenedesmus pectinatus|method ann 3|method str 2
79
77
4972015|Scenedesmus raciborskii|method ann 3|method str 2
6625510|Scenedesmus acuminatus|method ann 1|method str 2
94
17385571|Scenedesmus basiliensis|method ann 3|method str 2
100
100
12055731|Scenedesmus dimorphus|method ann 3|method str 2
96
4972041|Tetradesmus wisconsinensis|method ann 3|method str 2
92
89515386|Scenedesmus obliquus|method ann 1|method str 2
337263253|Chaetopeltis orbicularis|method ann 1|method str 2|pr|62183524
100
337263259|Hormotilopsis tetravacuolaris|method ann 1|method str 2|mod|337263255
100
337263255|Hormotilopsis gelatinosa|method ann 1|method str 2|pr|62183524
77176662|Oedogonium undulatum|method ann 1|method str 1
89
100
62183529|Oedogonium cardiacum|method ann 1|method str 1
81
62183531|Bulbochaete rectangularis var. hiloensis|method ann 1|method str 2
70
71482662|Oedogonium nodulosum|method ann 1|method str 2
59
87
62183535|Oedogonium oblongum|method ann 1|method str 2
337263256|Schizomeris leibleinii|method ann 1|method str 3|mod|337263258
100
337263258|Aphanochaete magna|method ann 1|method str 2|pr|71482662
80
337263257|Stigeoclonium helveticum|method ann 1|method str 2|pr|312004118
100
337263254|Draparnaldia plumosa|method ann 1|method str 2|pr|312004118
77
87
337263252|Uronema belkae|method ann 1|method str 2|pr|312004118
87
312004118|Uronema sp. CCAP 335
2627285|Haematococcus droebakensis|method ann 2|method str 3
100
71482601|Dunaliella parva|method ann 3|method str 1
100
145587829|Dunaliella salina|method ann 1|method str 1
2039265|Volvox rousseletii|method ann 5|method str 3
98
2039270|Volvulina steinii|method ann 5|method str 3
100
88
2039259|Volvox dissipatrix|method ann 5|method str 3
67
4007511|Yamagishiella unicocca|method ann 2|method str 1
87
14091704|Pandorina morum|method ann 2|method str 2
62
19569568|Eudorina elegans|method ann 2|method str 2
100
77
19569569|Eudorina unicocca|method ann 3|method str 1
111073399|Chlamydomonas typica|method ann 2|method str 2
78
111073392|Chlamydomonas petasus|method ann 2|method str 1
100
111073415|Chlamydomonas reinhardtii|method ann 2|method str 1
98
87
111073402|Chlamydomonas incerta|method ann 2|method str 1
2627279|Chlamydomonas komma|method ann 3|method str 1
100
3192578|Gonium octonarium|method ann 3|method str 1
94
6066290|Gonium quadratum|method ann 3|method str 1
87
3192594|Gonium pectorale|method ann 5|method str 3
Ulvophyceae
Trebouxiophyceae
DO
Sphaeropleales
Chlorophyceae
Chaeotphorales
Oedogoniales
Chaetopeltidales
OCC-Clade
CW
Chlamydomonadales

Figure 9: This tree of chlorophyceae was calculated by maximum parsimony on the alphabet 'RNA16'. In contrast to Figure 8, here the CW clade forms a sister group to the OCC clade, albeit with a lower bootstrap support of 77. Both are then allied to the DO group with a bootstrap support of 79.

4.5 Discussion

Compared to Buchheim et al. (2011b); Brouard et al. (2010); Turmel et al. (2008), the alphabets 'RNA10' and 'RNA16' in most cases reconstructed plausible trees, partially with only small differences and sometimes even better bootstrap support. However, Buchheim et al. (2011b) reconstructed his trees' topology based on the alphabet 'RNA12', which provided a few more topological differences here. This might be due to the absence of hand-crafted optimizations in the alignment, or the mixture of directly folded and homology modelled structures, which was needed for a full automation during this study. Interestingly, 'RNA10' and 'RNA16' resolved this tree topology without manual artefact corrections. This suggests a smaller sensitivity to artefacts when using these alphabets, which mainly differ from 'RNA12' by the additional coding of the horizontal dependencies between structural bonds. But, as this is the first evaluation of both alphabets and even performed on a complicated dataset, one should further evaluate these alphabets on a more certain phylogeny or simulated data to confirm their performance in comparison to 'RNA12' or 'DNA'. A differentiation based on treeing methods, which was first proposed in Figure 7, doesn't seem to influence the higher ranks. These are still uncertain in literature and hard to resolve in a robust way. However, a more detailed analysis of the phylum Chlorophyta is given in Chapter 8. The new scoring and rate matrices were explicitly calculated separately for species and genus based alignments. Although just the latter ones were included into treeforge, one might use the matrices of both types in future analysis to detect discrepancies between the evolutionary rates in species and genus. Finally, this handy package covers all important treeing methods including a large variety of models, together with the major aligning algorithms. It comes with four different sequence-structure alphabets, leaving room for a large variety of further analysis on sequence and secondary structure.

5 Generating ITS2 sequence-structure data

5.1 Introduction

With sequence annotation, secondary structure prediction, alignment and tree reconstruction methods, now the algorithmic framework is complete to run a full phylogenetic sequence-structure analysis. Nevertheless, nucleotide sequences are stored in large sequence databases like GenBank (Benson et al., 2011). GenBank updates and shares sequence data with other international collaborating databases like EMBL (Kulikova et al., 2007) or DDBJ (Tateno et al., 2002). Although ITS2 sequences are partially annotated here, this information is not as accurate as needed for a precise secondary structure prediction. Some annotations include imprecise location descriptors, like "in between", "before" or "uncertain" in GenBank data files. These were ignored by previous database versions (Schultz et al., 2006; Selig et al., 2008), which ran a less accurate BLAST based annotation approach. Therefore, a new pipeline was developed generating a larger quantity of reliable sequence-structure data (Koetschan et al., 2010). This is based on ideas of the previous versions of the ITS2 database with about 110,000 structures. However, the new workflow incorporates HMM-based sequence annotation, includes more databases, performs an iterative homology modelling process and makes use of several smaller additions and speed improvements which result currently in about 380,000 sequences in total including 288,000 structure predictions.

5.2 Database creation and taxonomic tree storage (A)

The basic workflow of the pipeline is visualized in Figure 10. It starts with the creation of an emtpy ITS2 database according to the database schema in Figure 16. For storing taxonomic information, e.g. a representation of the taxonomic tree of life, the database contains the table 'taxons'. This is filled by obtaining data from the NCBI taxonomy database (Federhen, 2011). The taxonomy database is a curated set of names and classifications which are available per FTP download and updated every two hours. To store the tree, a data structure in form of a nested set is used.

5.3 Sequence download and indexing (B)

The acquisition of sequence data is visualized in part B of Figure 10. Genbank holds a set of sub-databases which are available per FTP download. Searching for ITS2 sequences, we process the following ones listed in Table 3.

gbenv	Environmental samples
gbinv	Invertebrates
gbmam	Other mammals
gbpln	Plants
gbpri	Primates
gbrod	Rodents
gbsts	Sequence tagged sites
gbuna	Unannotated
gbvrt	Other vertebrates
nc	Complete genomic molecules including genomes, chromosomes, organelles, plasmids
daily	Daily updates, unsorted

Table 3: This partial overview of GenBank sub-databases includes all that are used in Figure 10 B for sequence download. Additionally, daily sequences are added to this data set as well.

In a next step these sequences in form of GenBank flat files are indexed. This is necessary as the download contains several 100,000 zipped GenBank files which are piped into one big datafile containing several hundred GB. While downloading this data, its length, GI and offset are simultaneously stored in the database (table ’gb_indexes’). By additionally indexing the ’gi’ column of the PostgreSQL table, this method allows to find and read a complete GenBank entry from such a large data file just within milliseconds.
Subsequently, files are prepared for HMM annotation. By making use of the index, GenBank flat files are read and further parsed. For this purpose, a new GenBank parser was developed. Compared to the BioPerl toolkit (Stajich et al., 2002), it allows the extraction of relevant information in about 1/3rd of the time. Finally, sequences are stored in split FASTA formatted files in preparation of the HMM annotation. For a Hidden Markov Model based de-

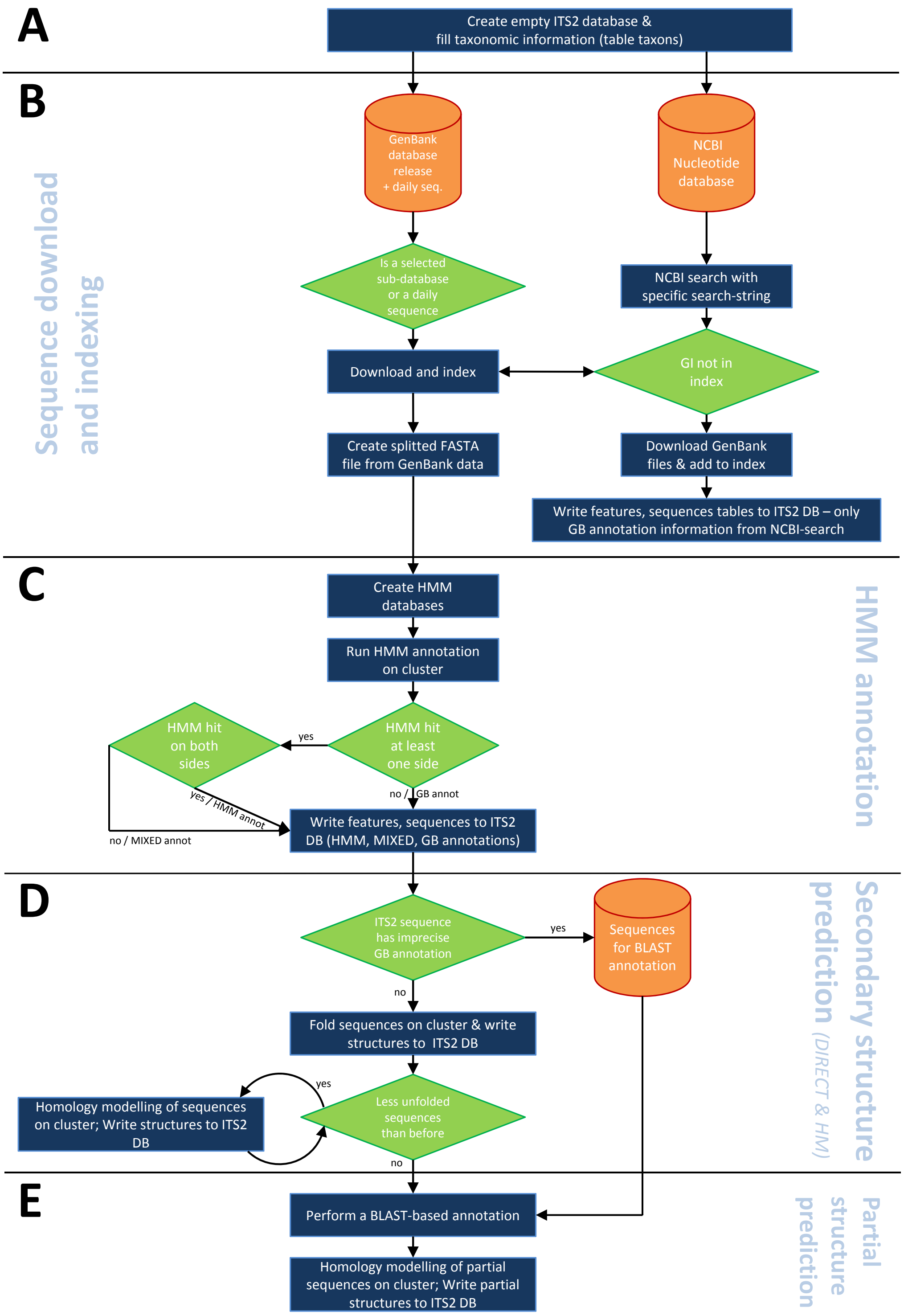

A
Create empty ITS2 database & fill taxonomic information (table taxons)
B
Sequence download and indexing
GenBank database release + daily seq.
NCBI Nucleotide database
Is a selected sub-database or a daily sequence
NCBI search with specific search-string
Download and index
GI not in index
Create splitted FASTA file from GenBank data
Download GenBank files & add to index
Write features, sequences tables to ITS2 DB – only GB annotation information from NCBI-search
C
HMM annotation
Create HMM databases
Run HMM annotation on cluster
HMM hit on both sides
yes
HMM hit at least one side
yes / HMM annot
no / GB annot
no / MIXED annot
Write features, sequences to ITS2 DB (HMM, MIXED, GB annotations)
D
Secondary structure prediction (DIRECT & HM)
ITS2 sequence has imprecise GB annotation
yes
Sequences for BLAST annotation
no
Fold sequences on cluster & write structures to ITS2 DB
yes
Homology modelling of sequences on cluster; Write structures to ITS2 DB
Less unfolded sequences than before
no
E
Partial structure prediction
Perform a BLAST-based annotation
Homology modelling of partial sequences on cluster; Write partial structures to ITS2 DB

Figure 10: Flowchart explaining data generation of ITS2 sequences and secondary structures. First, the database is filled with a taxonomic tree (A). In the next step, sequences are retrieved from Genbank and indexed for further processing (B). Afterwards, HMM annotation and secondary structure prediction are the crucial steps in the workflow (C,D). Finally, partial structures are predicted (E).

tection, the flanking regions of the ITS2 are indispensable. Hence, available 5.8S and 28S strands are added to the ITS2, if its annotation is imprecise (a list of possible values and examples is given in Table 4).

EXACT	(5..100)
BEFORE	(<5..100)
AFTER	(>5..100)
WITHIN	((5.10)..100)
BETWEEN	(99^100)
UNCERTAIN	(99.?100)

Table 4: This Table lists all possible values that an annotated sequence might contain in GenBank data files. Additionally, a short example is given on the right.

Beside the download of GenBank databases, a file download based on a marker specific search terms is performed additionally, targeting the complete nucleotide database of NCBI. In the following Table 5 a list of specific search terms is given. For future database versions, which might not only incorporate the marker ITS2, the search is performed for all subunits of the rRNA cistron.

internal+transcribed+spacer
ITS1 ITS+1 internal+transcribed+spacer+1 ITS-1 ITS_1
ITS2 ITS+2 internal+transcribed+spacer+2 ITS-2 ITS_2
5.8S 5.8+S 5.8-S 5.8_S
28S 28+S 28-S 28_S

Table 5: Marker specific search terms used for NCBI query against the nucleotide database.

Once a list containing GIs is retrieved from the NCBI search, a match to the index is performed. Only files which are not indexed yet are fully downloaded afterwards and written to the ITS2 database. Here, it should be mentioned that it may be quite possible that same features with same GI but different annotation become added to the database during the whole precess.

5.4 HMM annotation (C)

Having downloaded all necessary sequence information, now profile HMM databases need to be generated. These rely on manually curated sets of 5.8S and 28S bordering regions of the ITS2 with a length of 25 nucleotides from Keller et al. (2009). Both form a very conserved hybridizing stem with typically two free nucleotides. Sequence sets are available for 'Eukaryota', 'Diptera', 'Viridiplantae', 'Fungi' and 'Metazoa'. Based on those and their reverse complements, in total ten HMMs are created for usage in HMMER2 (Eddy, 1998).

Applied on the split FASTA sequences, all hits are evaluated and sequences are sorted into three bins of 'GenBank' only, 'Mixed' or fully 'HMM' annotated. Here we trust in particular HMM annotated hits and prefer them before a mixed or GenBank annotation. During evaluation, the ITS2 from all scoring combinations is extracted and folded. If the resulting structure proves to be a valid ITS2, it is tagged. From all tagged foldings, the HMM annotated ones with lowest energy are again preferred before mixed or GenBank annotations. This seemingly complicated process ensures that a maximum of fold features is retrieved in the next step.

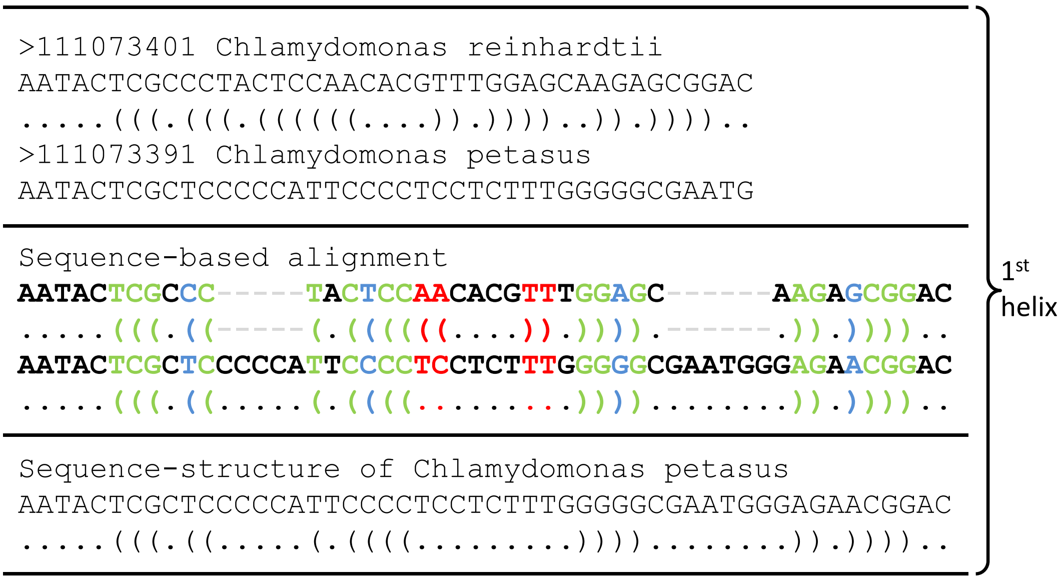
5.5 Secondary structure prediction (D)

After annotation, all sequences with preference for HMM, then mixed and finally Genbank annotation are loaded from the database and folded via Unafold (Markham and Zuker, 2008) by energy minimization. This resulted in about 100,000 directly folded structures during the last database updates. Based on the folded sequences, structures are iteratively homology modelled (Wolf et al., 2005a) until the number of unfolded sequences remains constant. Here, some constraints have to be set to retain a certain level of quality:

- All four helices must be modelled with at least 75% of transfer for each helix.

- The overall length must not be shorter than in the 0.1 % quantile of directly folded sequences.
- The length of unpaired sequence ends must not be shorter than in the 99.0 % quantile.
- The number of N-characters contained in a sequence must not be larger than in the 99.9 % quantile.

The process of homology modelling (Wolf et al., 2005a; Selig et al., 2008) is explained by a small example in Figure 11. It starts with two sequences where only one secondary structure is known. As the term homology suggests, the sequences are aligned first, to visualize homologue positions. These are marked in light green (Figure 11). Based on those conforming columns, the structure is transferred from the template. However, there might appear the case that structural bonds ('(' or ')') occur at positions where template and target nucleotides differ (Figure 11 blue or red marked positions). In this circumstance, structure is transferred only if the corresponding target nucleotides can form a valid base-pairing (blueish marked positions). All other/uncertain positions become unpaired ('.'). This results in a transfer of $\frac{\text{20 brackets target}}{\text{24 brackets template}}100 = 83,33$ % for the first helix. Finally, a post-folding process closes remaining bulges – resulting from the gaps, as example (not illustrated in the Figure).



```
>111073401 Chlamydomonas reinhardtii
AATACTCGCCCTACTCCAACACGTTTGGAGCAAGAGCGGAC
.....(((.(((.((((((....)).))))..)).))))..
>111073391 Chlamydomonas petasus
AATACTCGCTCCCCCATTCCCCTCCTCTTTGGGGGCGAATG

Sequence-based alignment
AATACTCGCCC-----TACTCCAACACGTTTGGAGC------AAGAGCGGAC
.....(((.((-----(.((((((....)).)))).------.)).))))..
AATACTCGCTCCCCCATTCCCCTCCTCTTTGGGGGCGAATGGGAGAACGGAC
.....(((.((.....(.((((.........))))........)).))))..

Sequence-structure of Chlamydomonas petasus
AATACTCGCTCCCCCATTCCCCTCCTCTTTGGGGGCGAATGGGAGAACGGAC
.....(((.((.....(.((((.........))))........)).))))..
```

Figure 11: This Figure illustrates the homology modelling (Wolf et al., 2005a) process of the first helix between the template *Chlamydomonas reinhardtii* and the target sequence *Chlamydomonas petasus*. Green columns indicate homologue positions where structure information is transferred directly. Blueish columns allow a structure transfer, however nucleotides between template and target sequence are not equal here. Red and black positions mark columns where no structural transfer is possible.

Notice: Target sequences which obtained a homology modelled structure may become re-annotated by the templates if their annotation method is HMM on that side, and the target sequences' method is not.

5.6 Partial structure prediction (E)

Partial structures are mainly classified in types. The first type is yielded from a BLAST-based annotation. This annotation method performs a target sequence-based BLAST search against all fold structures already in the database. The best hit then serves as template for homology modelling. As BLAST is only a "local alignment" search tool, this annotation method is not very reliable. Hence, this type of structures is classified as partial structure, although it might contain four helical regions. The second type are structures remaining from yet unfolded sequences where only two concatenated helices could be transferred with at least 75 % per helix. These are typically missing larger fragments of the ITS2 and should not be included into standard phylogenetic analyses. They might be useful for studies on specific helices only, however should be handled with special care.

Having completed this task, the database is nearly ready for use. By running the internal PostgreSQL function 'makeproductionready()' (Table 7) some final clean-ups are performed, e.g. the deletion of additional markers which were already integrated for future database versions, the pre-calculation of structure counts for each taxonomic rank or the deletion of redundantly annotated features using the previously explained priority rules.

5.7 General overview of files and folders

The database generation follows a set of Perl and bash scripts which are located under 'rRNA_DB/db/trunk' and follow the sketch visualized in Figure 10. The 'generate.pl' – directly located in this directory basically calls numbered scripts placed in the 'scripts' directory. A more detailed overview of folders and scripts is given in the appendix (Chapter 15.L, 15.M).

5.8 Discussion

In this study, two main objectives were faced: (i) a larger number of secondary structures, and (ii) a better quality of annotations leading to more accurate structure predictions. As sequence databases grow daily, there is a natural increase within each update. However, with the ability to annotate sequences by ourselves, we do not have to rely just on search terms and fault-prone sequence tags any more. This achievement enabled the inclusion of a large set of relevant sequence databases which can be scanned using HMMs. Of course, HMMs also do not cover a 100 % rate of detections, thus the NCBI search is still present in the pipeline.

Next, the topic of secondary structure prediction was addressed. We want to rely on state-of-the-art folding algorithms like energy minimization, which is close to the ITS2's chemical folding process. Therefore, the folding algorithm itself was left unchanged. However, one could further enhance the homology modelling step. Having predicted new secondary structures in one iteration, these could serve as templates for further iterations as well. It turned out that this procedure must be controlled very strictly to avoid reduction of sequence lengths with increasing number of iterations. Now, the majority of structures are modelled in the first 3 rounds.

The improved prediction quality of structures mainly originates from the accuracy of the new HMM-based annotation approach. By allowing also mixed annotations for sequences where a flanking region is cut or unpredictable by the HMM, we gain the maximum performance out of this method and could significantly increase the number of direct folds. However, one can argue whether direct folds lead to better phylogenetic predictions compared to homology modelled structures. It seems that the answer to this question is data-dependent and must be addressed in each case individually, even tough "homology modelling tends to dampen the influence of artefacts (...) and thus yields more robust tree topologies" (Markert et al., 2012). Finally, our database is a central place to go when working on ITS2-based phylogenetics, and so far the only database that incorporates secondary structure predictions for this marker.

6 Technical implementation of the ITS2 workbench and the ITS2 database

6.1 Introduction

Having now tools available for ITS2 annotation, a large database of already predicted ITS2 sequences with their secondary structures as well as the algorithms to calculate a sequence-structure based phylogeny, the last missing part is an environment for making this pipeline available to the public in a simple and intuitive way. Where first only stand-alone software was able to deal with secondary structure, and a user had to manually download data from the ITS2 database, import and export it into and from several tools, today the ITS2 workbench provides a self-contained working suite, online and easy to use. To implement such high usability, a complete redesign of the outmoded CGI scripts had to take place. The ITS2 workbench is now equipped with the Perl based web development framework Catalyst which runs on a stand-alone Apache web server in combination with a PostgreSQL database server and handles requests from a modern JavaScript frontend.

6.2 The frontend of the ITS2 workbench

The frontend itself was developed using the JavaScript Framework ExtJS in combination with HTML and CSS. JavaScript is a scripting language which is applied on the client side, mainly implemented in web browsers and was used in the beginning for updating DOM structures inside an HTML document or for showing simple user notifications. It allows an object orientated, imperative as well as functional programming style. During the last 6 years, JavaScript has developed from its shadowy existence into a prominent programming language for interactive website design and stands side by side with the modern terms of Web 2.0 and Asynchronous JavaScript and XML (AJAX).

6.2.1 AJAX and XHR with Web 2.0

The basic concept behind AJAX is an additional AJAX-engine (Figure 12 bottom) between client and server. This engine is typically based on JavaScript and becomes initialized when the pages loads for the first time. The engine itself cares about page rendering as well as interactive communication with a server. Compared to the classical web application model (Figure 12 top), this has a major advantage:

A user initiating an event by performing a mouse-click for example, does not generate an HTTP-request that loads a complete website again (which would require the user to just wait), but instead triggers the AJAX-engine to load only necessary parts from the server. While data transmission takes place, the AJAX-engine is fully operational and could render further information on the website, start another data transmission or response to a different user interaction.

Further positive side-effects are a reduced traffic overhead, thus a faster response from the server and a more elegant programming style which brings the website one step closer to a desktop-like environment but on the Web. This development initiated a race in speed optimization (e.g. DOM inserts) of JavaScript engines between the major browser manufactures, and further led to a battle in the development of JavaScript frameworks. These orientated on desktop GUIs, which resulted in the implementation of typical design styles like layout-containers, panels, event handlers or even drawing-suites.

6.2.2 The ExtJS JavaScript framework

In the following, a short description of open source JS frameworks, available in 2009 is given (Table 6), where finally ExtJS was chosen for implementing the ITS2 workbench.

ExtJS provides the functionality to extend from objects, which are typically based on the class 'Component'. This class supports automated creation, rendering and destruction of objects. Using the 'ComponentManager', one can easily access a component by its 'id'-property. A 'Container' class directly

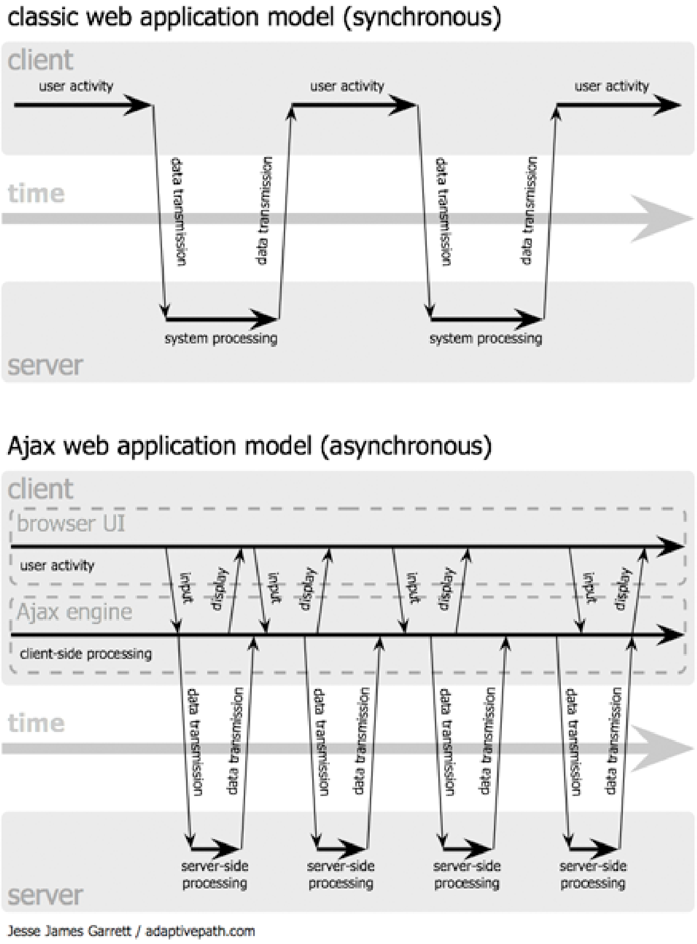


Figure 12: The classical web application model on the top clearly visualizes the time passing until a server responds with data transmission after a user interaction (typically a page click) took place. When regarding the AJAX model on the bottom, the time of a server response can be reduced by eliminating redundant traffic. Further, the rendering of some components can already take place before a data transfer from the server is completed.

Dojo toolkit	several widgets; less organized, documented, structured
jQuery	very good concept; focussed on events and animation
Prototype & Scriptaculous	focus on animations; no GUI options
Yahoo / YUI	GUI elements and animation; weak documentation
ExtJS	Very rich GUI; animations and clear API documentation

Table 6: Short overview of major public licensed JavaScript frameworks available in 2009.

extended from 'Component' supports the possibility to add items using the 'items'-property. By applying this concept to different layouts, components can be nested easily.

An alternative way of placing a component however, is to render it directly into a DOM element. When specifying the property 'renderTo' this task is executed automatically by ExtJS on component initialization. A simple example shows this concept for the "Cited by" page of the workbench by using a panel in combination with an XMLHttpRequest.

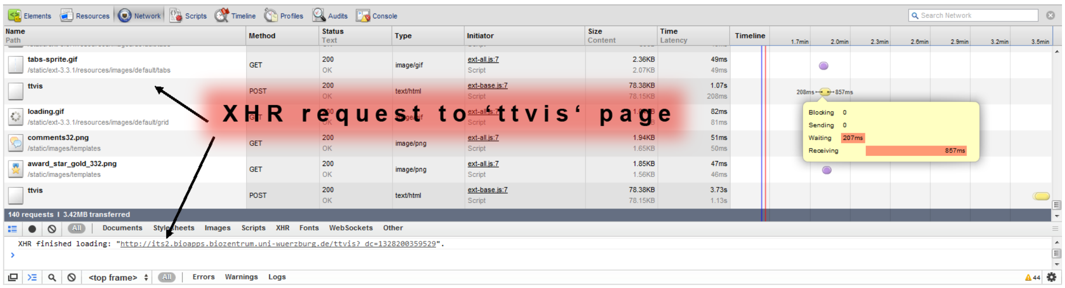
HTML:

```
<html>
    <head>Example Citations</head>
    <body>
        <div id="myCitations">
    </body>
</html>
```

JavaScript:

```
var citationspanel = new Ext.Panel({
    id:'citedby',
    bodyStyle: 'padding:15px',
    title:'Cited by',
    iconCls:'icon-citedby',
    autoLoad:{url:'ttvis',method'POST',params:{page:'citedby'}},
    renderTo:'myCitations'
});
```

These lines create a panel, render it to the div-element with the id 'myCitations' and automatically invoke an XHR request to the 'ttvis' controller (Figure 13). Data content (the citations) is finally delivered from the server after a few milliseconds. Typically, POST requests are created that enable the transfer of variables as message content. The controller 'ttvis' is explained later in the appendix, Table 17.



Name Path	Method	Status Text	Type	Initiator	Size Content	Time Latency
tabs-sprite.gif /static/ext-3.3.1/resources/images/default/tabs	GET	200 OK	image/gif	ext-all.js:7 Script	2.36KB 2.07KB	49ms 49ms
ttvis	POST	200 OK	text/html	ext-base.js:7 Script	78.38KB 78.15KB	1.07s 208ms
loading.gif /static/ext-3.3.1/resources/images/default/grid	GET	200 OK	image/gif	ext-all.js:7 Script	[illegible]	82ms 81ms
comments32.png /static/images/templates	GET	200 OK	image/png	ext-all.js:7 Script	1.94KB 1.65KB	51ms 50ms
award_star_gold_332.png /static/images/templates	GET	200 OK	image/png	ext-all.js:7 Script	1.85KB 1.56KB	47ms 46ms
ttvis	POST	200 OK	text/html	ext-base.js:7 Script	78.38KB 78.15KB	3.73s 1.13s

Blocking 0
Sending 0
Waiting 207ms
Receiving 857ms

140 requests I 3.42MB transferred

```
XHR finished loading: "http://its2.bioapps.biozentrum.uni-wuerzburg.de/ttvis?_dc=1328200359529".
>
```

Figure 13: Visualisation of a data request (XHR) in Google Chrome, initiated by ExtJS. The server's response is received after a few milliseconds.

However, data received from XHR requests must not always be rendered on a website immediately. Therefore, ExtJS provides a substantial element inside its framework: The 'Store' class which is integrated into several components. A 'store' allows to collect data in records, to filter and to modify them. Typically, components e.g. grids load data into a store and display them in specific columns or formats. To exchange data with the 'store', a specific encoding must be applied. This is implemented by different data readers/writers that handle, for example, XML or JSON (JavaScript Object Notation) encoded data content.

6.2.3 Web 2.0 data formats

To transfer data from the server to the client, former applications were based on the SOAP networking protocol. This allows to deliver XML in a SOAP envelope using XHRs which influenced the definition of the term AJAX. The main benefits of this protocol are a well-defined standard, several available parsers and writers in different programming languages as well as an easy validation using a DOM or XML Schema Definitions (XSD). Some of the major drawbacks are a large traffic overhead and a longer development cycle for setting up definition and validation files. Here, the JSON format benefits from a very sparse and easy declaration, however lacks any automated validation and specifications before and after transferring data.

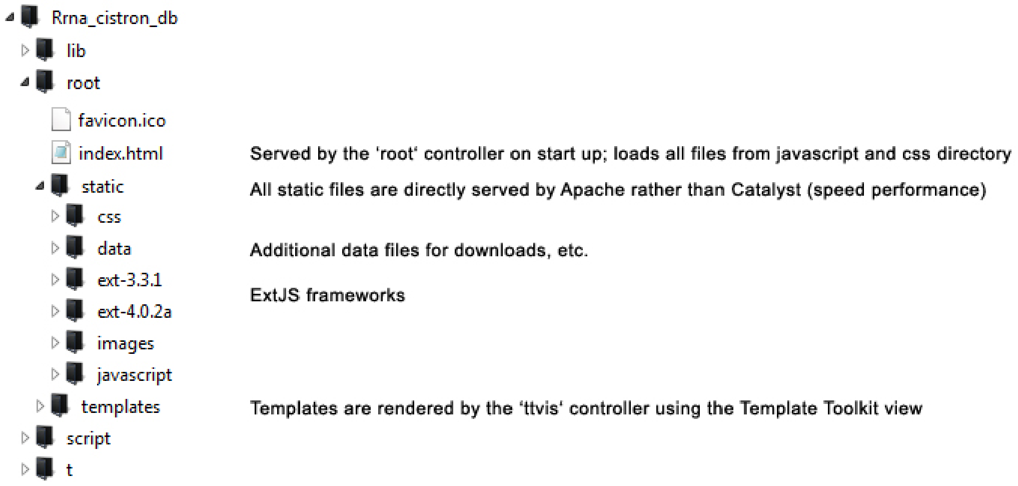
In the following, an extract from a JSON transfer loading sequence information into a data-grid is given (sequences and structures are shortened):

```
{"totalCount":3,"jsonresponse":[

{"energy":"-31.6","sequence":"TCTGC","method_str":"2","acc":"AB190265",
"group":"Alveolata","gi":"52546182","structure":".....",
"method_ann":"3","specname":"Symbiodinium sp. Kokubu1a"},

{"energy":"-31.6","sequence":"TCTGC","method_str":"2","acc":"AB190276",
"group":"Alveolata","gi":"52546193","structure":".....",
"method_ann":"3","specname":"Symbiodinium sp. Sakurajima2c"},

{"energy":"-35.1","sequence":"TTTCA","method_str":"2","acc":"AB190280",
"group":"Alveolata","gi":"52546197","structure":".....",
"method_ann":"2","specname":"Symbiodinium sp. Amami1a"}

]}
```

The JSON notation which is the JavaScript object notation allows to obtain a JavaScript object by just applying the 'eval()' function on the JSON string. JSON data is usually transferred by XMLHttpRequests. To not be misleading, XHR was formerly developed for Microsofts Internet Explorer and enables data transfer for all kinds of HTTP requests in combination with JavaScript. Also plain text or HTML code can be delivered.

6.2.4 Frontend files and locations

The workbench frontend is organized under 'Rrna_cistron_db/root', see Figure 14. Here, the 'index.html' is the direct entry point for the ITS2 workbench and is rendered by the 'root' controller. All necessary JavaScript and CSS files are directly loaded from the 'index.html'.



```
Rrna_cistron_db
  lib
  root
    favicon.ico
    index.html        Served by the 'root' controller on start up; loads all files from javascript and css directory
    static            All static files are directly served by Apache rather than Catalyst (speed performance)
      css
      data            Additional data files for downloads, etc.
      ext-3.3.1       ExtJS frameworks
      ext-4.0.2a
      images
      javascript
    templates         Templates are rendered by the 'ttvis' controller using the Template Toolkit view
  script
  t
```

Figure 14: Important files and folders of the frontend of the ITS2 workbench.

As soon as JavaScript and CSS files are transferred, the file 'applayout.js' provides the basic border layout on startup. This integrates a large set of additional functions and objects. In the appendix, a short overview of those files and their functionalities is given (Table 15):

6.3 The backend of the ITS2 workbench

6.3.1 Catalyst: A Perl-based Web development framework

Catalyst is a MVC (Model View Controller) Web development framework based on the programming language Perl. Compared to the old ITS2 Web interface, this framework enables a structured organization of files and a separation of programming, visualization and database logic. Using this standard, it follows the principles of Don't Repeat Yourself (DRY) and is fully object orientated. It further comes with a set of CPAN hosted plugins and runs on all major Web servers.

The MVC architecture of Catalyst allows a flexible program design including reusable components, an organized structure and programming logic. The 'Model' represents a full database scheme by encapsulating each table into an own Perl object. When fetching data from the database, the corresponding Perl objects can be returned, providing easy access to the equally named columns. The ITS2 workbench provides two database models. One represents the sequence-structure database (see also Figure 16), a second one represents the administration database depicted in Figure 17. A 'View' instead is responsible for the data encoding when sending data back from the server to the client. This format is typically one of the previously described Web 2.0 data formats in chapter 6.2.3. The ITS2 workbench currently provides a JSON, a file download and a Template Toolkit (TT) view. Finally, a 'Controller' connects model and view and further delegates all necessary actions. By calling a website in a specific way:

http://name.tld/controller/function

a controller, and even function calls provided by the controller are accessible for the client (see also the chapter about the RESTful programming paradigm 6.3.1). This enables a very flexible way of programming and eases the development of reusable code components.

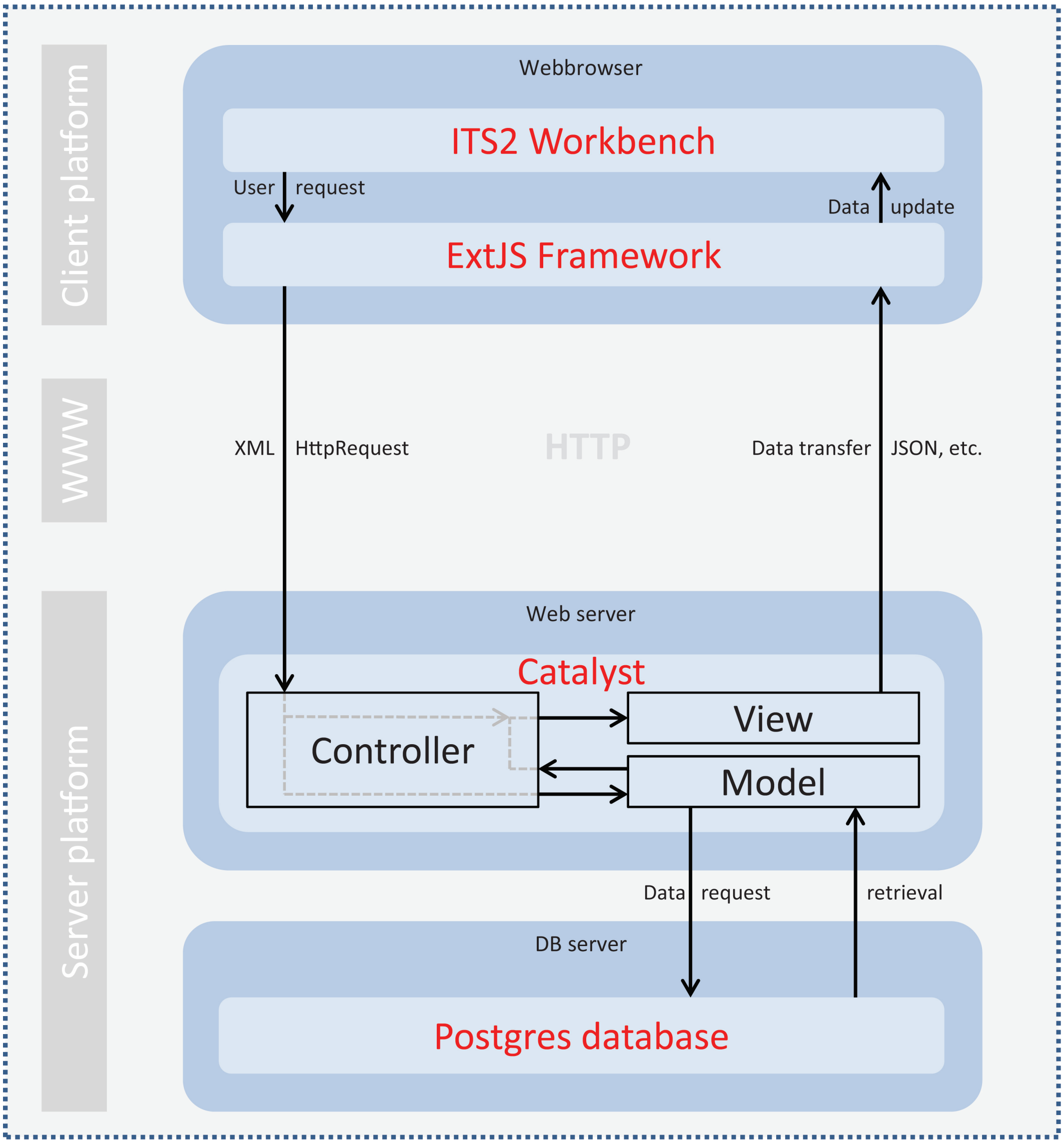


Figure 15: The Figure illustrates a typical data request in the ITS2 workbench, established to the MVC Web development framework Catalyst. The XHR request from the client side causes a controller to contact the PostgreSQL database if necessary. Finally, information is rendered by a specific view, and transferred back to the client in a view-dependent format. The client then further processes its received data.

Figure 15 visualizes the basic concept behind the MVC architecture. A client request is first accepted by the controller. It decides whether a database connection is required, and in such case uses the database model to fetch relevant entries. The view finally regulates how or in which format data has to be encoded before it is sent back to the client, where it can then be further processed. With the XHR request of course, beside the URL for the controller, also further information in form of variables can be delivered by GET or POST.

The RESTful programming paradigm and CRUD: Regarding the controller in more detail, a common implementation is the Restful design. The term representational state transfer (REST) was first defined by Fielding (2000) in his dissertation in 2000. Although it is not standardised, it became integrated into most today's web development frameworks in different ways. The basic principle behind this concept is to access or modify the representation of a specific resource, referenced by an URI. To interact with this resource, only two things have to be known. This is first its identifier, typically in a form of:

http://example.com/resources or
http://example.com/resources/item

and secondly the action to be performed. These actions, often go in coherence with basic database operations. These can be summarized under the term CRUD, standing for:

```
Create (INSERT DB statement, HTTP transfer by POST request)
Read (SELECT DB statement, HTTP transfer by GET request)
Update (UPDATE DB statement, HTTP transfer by PUT request)
Delete (DELETE DB statement, HTTP transfer by DELETE request)
```

The database actions are triggered by corresponding HTTP POST, GET, PUT and DELETE requests. Especially the ITS2 administration interface is a typical implementation of RESTful CRUD web services. It allows to create, show, modify, or delete information on the admin interface for managing citations, news, staff-entries and more.

Template Toolkit: Having finished all calculations by the controller, different views are available in Catalyst that send information back to the client. Beside the common formats XML and JSON (see Chapter 6.2.3), Catalyst provides a special view for transferring parts of a website in form of templates, the Template Toolkit view. TT is a very frequently used module in the Perl community. It provides a way to separate website design from programming logic. A template is a sketch of code, usually HTML which provides the complete design and layout of parts of a website. Beside a replacement of variables with corresponding values, TT can also provide basic programming logic like loops, conditions and further simple constructs that ease the design process. In the following a short extract of the 'Cited by' template is given as example. It produces an unordered list containing information about citations like authors, title, journal etc.:

```
<span class="main_header">Cited by</span><br><br>
[% IF ttcitations %]
    <ul class="striped-ul" style="list-style-type:none;">
        [% FOREACH citation IN ttcitations %]
            <li class="shadow">
                [% citation.authors %] ([% citation.pubyear %]):<br>
                [% citation.title %]<br><i>[% citation.journal %]</i>
            </li>
        [% END %]
</ul>
[% ELSE %]
        <p>No citations found!</p>
[% END %]
```

All templates are located in the templates folder, see Figure 14. Further, a short description of each template is given in the appendix (Table 16).

6.3.2 The PostgreSQL database including PL/pgSQL

PostgreSQL is an object-relational database providing typical relational database mechanisms in combination with object orientated extends like the inheritance of tables. With its Procedural Language/PostgreSQL Structured Query Language (PL/pgSQL), it allows to declare stored procedures, using variables, functions, conditions and loops. PL/pgSQL code can be called directly from SQL statements as well as from triggers. The ITS2 database implements many reasonable PL/pgSQL functions, views and triggers which are listed in details in the appendix (see Tables 7, 8 and 11).
Regarding the databases itself, the workbench connects to two separate ones. The first one is filled during the update procedure in section 5 and provides all information related to ITS2 sequences, structures, taxonomy information etc.. The second database is filled using the ITS2 admin interface and provides basic information about the website like staff, citations or RSS feeds.

Database Schema of the ITS2 database To be conform with the Catalyst naming-conventions, all table names end with a plural 's'. Indexed columns are marked with a white symbol in Figure 16. Indices on primary keys are auto generated by PostgreSQL.

alignments Stores the 'neelde'-alignment (EMBOSS package) of homology modelled sequence-structures. Includes model and template (transferred) information as well as alignment statistics.

annotation_methods Holds the annotation methods 'HMM' (both sides HMM hit), 'MIXED' (only one side HMM hit, other side GB entry), 'GB' (both sides GB entry), 'HM' (re-annotation by homology modelling) and 'BLAST' (BLAST search based annotation).

elements Refers to the basic elements of the rRNA cistron which are: 18S, ITS1, 5.8S, ITS2 and 28S.

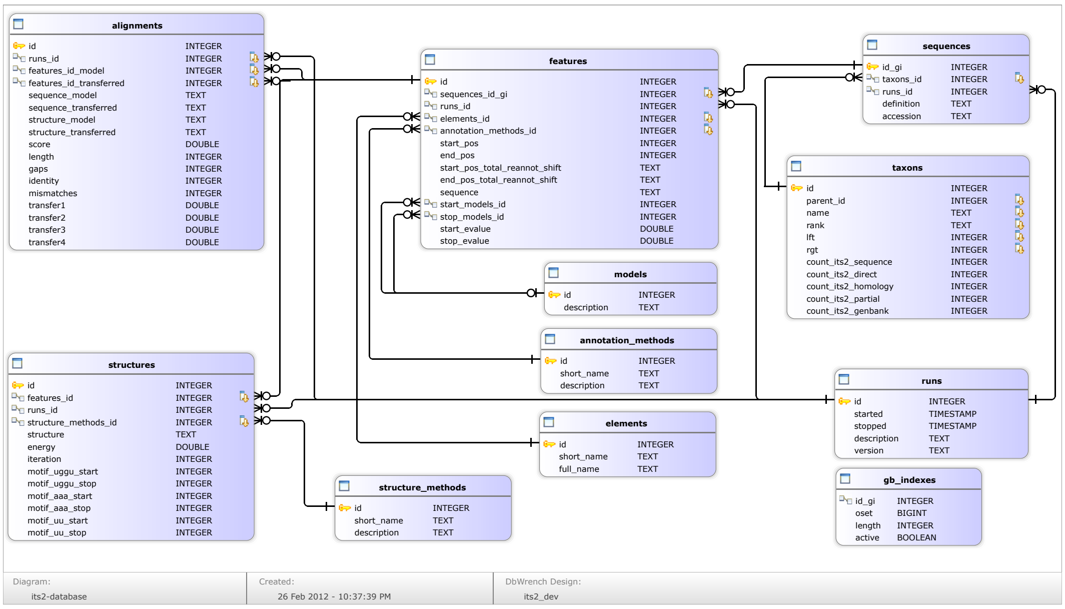


Figure 16: Database Schema of the ITS2 database.

features Stores sequence information, element type, and annotation specific information.

gb_indexes This table only exists while the database is being filled (see Chapter 5). It holds the GIs index position and length for a fast extraction of GenBank flat files from the index data file. Although GIs are unique, they might not be unique in this table due to the inclusion of daily GenBank updates. However 'old' GI references are flagged as inactive.

models A representation of models that were used for ITS2 annotation. These can be for GB/MIXED annotation: 'Genbank' or 'Genbank imprecise'; for MIXED/HMM annotation: 'Diptera', 'Eukaryota', 'Fungi', 'Metazoa' or 'Viridiplantae'.

runs This table contains information about the database version, a short description as well as the duration of the update process.

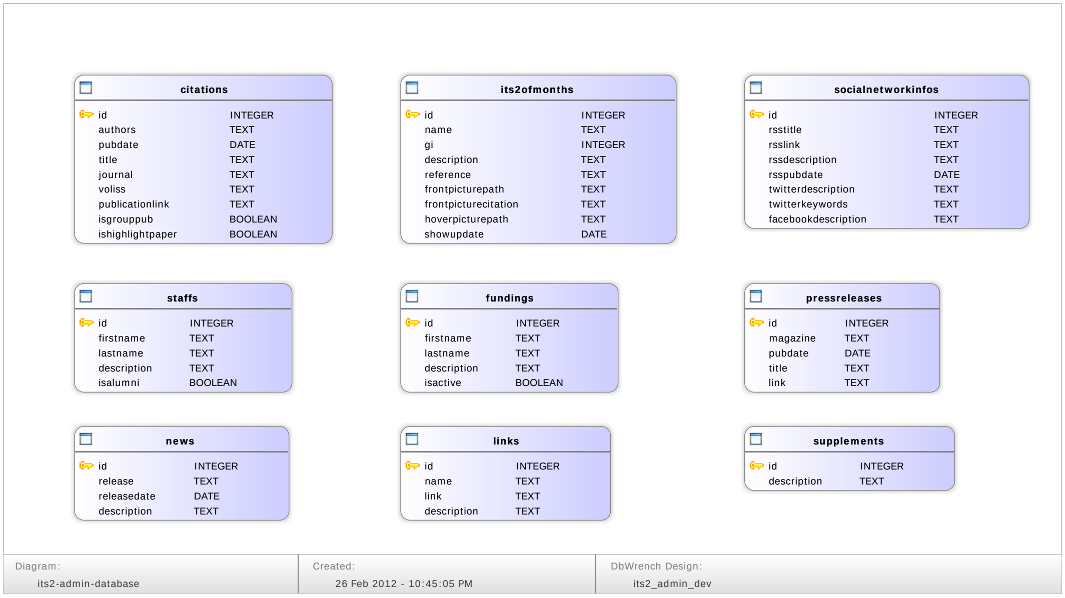
sequences Holds information about a sequence, though not a specific feature or a sequence data itself! It stores mainly its GI, Accession and definition.

structure_methods The structure methods of the ITS2 can be 'DIRECT' for direct folds, 'HM' for homology modelled or 'PARTIAL' for incomplete structures.

structures Saves the structure of a sequence itself in bracket-dot-bracket notation. Further energy and HM iteration is stored if available. This table also contains fields for sequence or sequence/structure motifs. However the detection of motifs is not automated yet.

taxons This table stores the taxonomic tree in form of a nested set. Further, for each taxonomy, the number of sequence counts depending on the structure method are available. These were added for performance speed-ups only.

Database Schema of the ITS2 workbench – admin interface The schema of the ITS2 admin database (Figure 17) mainly contains static not connected tables. This is information about citations, RSS feeds or staff as example.



citations	
id	INTEGER
authors	TEXT
pubdate	DATE
title	TEXT
journal	TEXT
voliss	TEXT
publicationlink	TEXT
isgrouppub	BOOLEAN
ishighlightpaper	BOOLEAN

its2ofmonths	
id	INTEGER
name	TEXT
gi	INTEGER
description	TEXT
reference	TEXT
frontpicturepath	TEXT
frontpicturecitation	TEXT
hoverpicturepath	TEXT
showupdate	DATE

socialnetworkinfos	
id	INTEGER
rsstitle	TEXT
rsslink	TEXT
rssdescription	TEXT
rsspubdate	DATE
twitterdescription	TEXT
twitterkeywords	TEXT
facebookdescription	TEXT

staffs	
id	INTEGER
firstname	TEXT
lastname	TEXT
description	TEXT
isalumni	BOOLEAN

fundings	
id	INTEGER
firstname	TEXT
lastname	TEXT
description	TEXT
isactive	BOOLEAN

pressreleases	
id	INTEGER
magazine	TEXT
pubdate	DATE
title	TEXT
link	TEXT

news	
id	INTEGER
release	TEXT
releasedate	DATE
description	TEXT

links	
id	INTEGER
name	TEXT
link	TEXT
description	TEXT

supplements	
id	INTEGER
description	TEXT

Diagram:	Created:	DbWrench Design:
its2-admin-database	26 Feb 2012 - 10:45:05 PM	its2_admin_dev

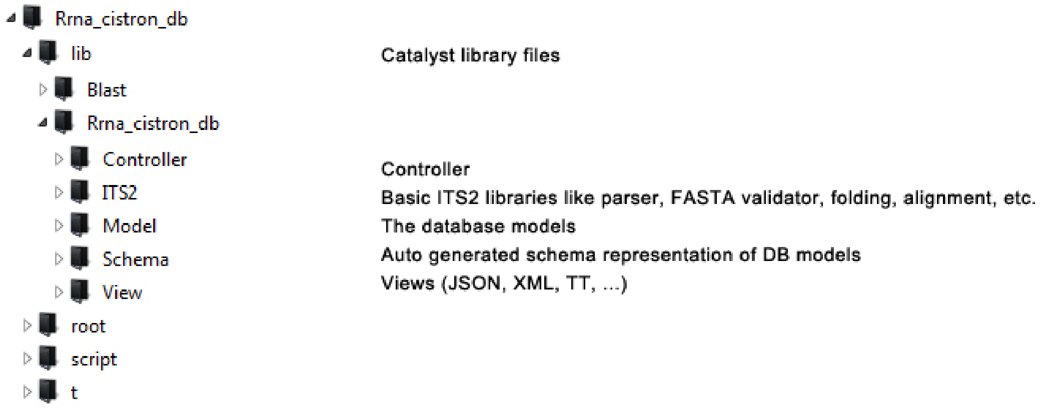
Figure 17: Database Schema of the ITS2 admin database.

6.3.3 Apache and Fcgid communication

To run the ITS2 workbench on an Apache web server, Catalyst provides two methods located in the
'Rrna_cistron_db/script' directory: CGI and FastCGI. Beside both options, Catalyst also provides its own web and debugging server, which is not recommended in production environments. FastCGI – the performance optimized version of CGI – provides an intelligent caching mechanism for CGI programs and allows Catalyst to handle even parallel requests in a quick way and without reloading Perl interpreters, etc. on each new request. The workbench runs a maximum of 10 parallel instances and allows an execution time of 1000 seconds per process.

6.3.4 Backend files and locations

The workbench backend is organized under 'Rrna_cistron_db/lib/Rrna_cistron_db', see Figure 18. Here, controller, models and views are located in their corresponding folders. In addition, the folder ITS2 contains specific Perl modules for parsing and validating data, or for small encapsulated algorithms.



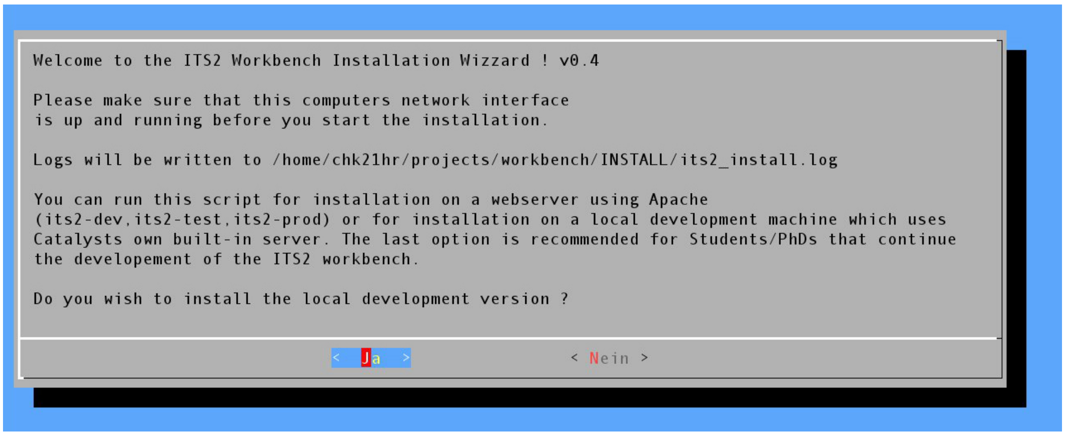
```
◢ Rrna_cistron_db
  ◢ lib                    Catalyst library files
    ▷ Blast
    ◢ Rrna_cistron_db
      ▷ Controller         Controller
      ▷ ITS2               Basic ITS2 libraries like parser, FASTA validator, folding, alignment, etc.
      ▷ Model              The database models
      ▷ Schema             Auto generated schema representation of DB models
      ▷ View               Views (JSON, XML, TT, ...)
  ▷ root
  ▷ script
  ▷ t
```

Figure 18: Important files and folders of the backend of the ITS2 workbench.

A detailed overview of all controller and specific modules is given in the appendix (see Table 17, 18).

6.4 ITS2 workbench installation script

To install the ITS2 database with all relevant dependencies and Perl modules, an ITS2 installer (Figure 19) was developed. It distinguishes mainly between two installation types. One for a production environment on a webserver, and a second one for developer use. All output messages are logged to the file 'its2_install.log' during installation. For specific information about the installation process, please follow the guided steps provided by the installer itself, and read the tech-documents in the SVN 'INSTALL' folder.



```
Welcome to the ITS2 Workbench Installation Wizzard ! v0.4

Please make sure that this computers network interface
is up and running before you start the installation.

Logs will be written to /home/chk21hr/projects/workbench/INSTALL/its2_install.log

You can run this script for installation on a webserver using Apache
(its2-dev,its2-test,its2-prod) or for installation on a local development machine which uses
Catalysts own built-in server. The last option is recommended for Students/PhDs that continue
the developement of the ITS2 workbench.

Do you wish to install the local development version ?
```

Figure 19: The ITS2 workbench installer in this Figure, provides an installation for production environments, as well as for developers.

6.5 Discussion

The ITS2 workbench, technically speaking, is equipped with the latest web technologies. This is especially its feature-rich JavaScript user interface and the web services provided by the Catalyst web development framework. However, when focussing on the backend, Catalyst clearly plays a role as an outsider in the community. It lacks the extensive support that professional platforms like Ruby on Rails or Oracles Java EE in combination with Hibernate and Spring can provide. Nevertheless, Catalyst is so far the most advanced web development framework for Perl, which has been established as a commonly used programming language in the field of Bioinformatics. With a large variety of plugins, it provides all basic features needed for typical web services.

7 The ITS2 workbench

7.1 Introduction

Having solved all technical aspects, the ITS2 workbench (Koetschan et al., 2012) is finally available at http://its2.bioapps.biozentrum.uni-wuerzburg.de. The website unifies all necessary tools that are required for a first ITS2 based phylogenetic analysis on sequence and structure. This ideally starts with the creation of a dataset – the taxon sampling, and ends with the calculation of a phylogenetic tree. However, this process is not always as straightforward as one would assume. During the analysis of data, especially when regarding the alignment, or even worse – after having calculated a ”final” tree, one might realize that something went wrong. This might be a surprising classification of a sequence inside an inappropriate looking clade, long branches between closely related species or a sequence that doesn't fit into the alignment at all. When narrowing down these issues, it often turns out that a false taxonomic classification, wrong annotation or sequencing error has caused the artefact. Nevertheless, the whole time-consuming analysis has to be repeated. The ITS2 workbench assists here with a finesse. All sequence-structures that form the taxon sampling are accessible from a pool at any time. As soon as one detects an error in the alignment or tree, the erroneous sequence-structure can be deleted from the pool, and previously performed calculations are automatically repeated (see Figure 20). This also affects later added sequence-structures. Together with additional tools that complete the pipeline of Schultz and Wolf (2009), the workbench offers a quick glance into an ITS2 based sequence-structure phylogeny and hereby simplifies the time-consuming tasks of taxon sampling.

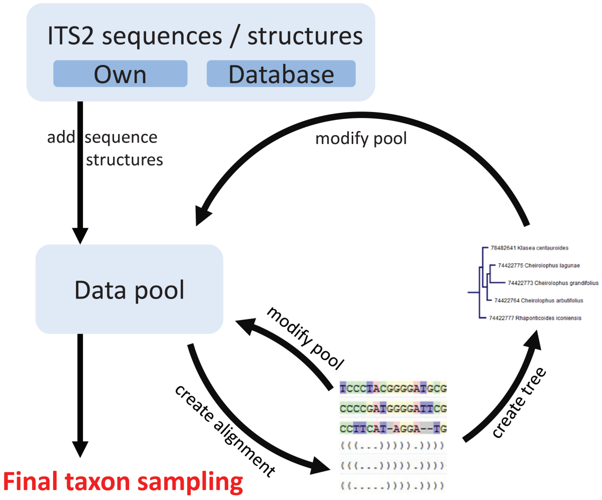


Figure 20: This Figure illustrates the process of taxon sampling using the ITS2 workbench. It starts with own sequence-structures or ones retrieved from the ITS2 database. These are added to a data pool which is the starting point for an alignment. By deleting sequences from the alignment, they are deleted from the pool as well, and a new alignment is recalculated automatically. The same procedure is applied when further adding sequences to the pool, or deleting them from the tree. Having repeated these steps until no further error-prone sequence-structures are visible, the taxon sampling is finished.

7.2 Overview

When regarding the website (Figure 21), the workbench is organized in a border-layout style and divided into header, west and center panel. The header enables access to the ITS2 database during all times using the live search function and provides links to basic information around the website. The west panel instead contains typical tools around the marker, like the HMM annotation, its secondary structure prediction or sequence motifs visualization. Further, it shows an accordion menu, guiding through the basic workflow of creating, managing and analysing a dataset. On the bottom left, the pool receives sequence and structure information from different locations of the website per drag & drop. Finally, a large tab panel in the center of the website shows all relevant information and allows to close or switch between relevant tabs during the whole work process.

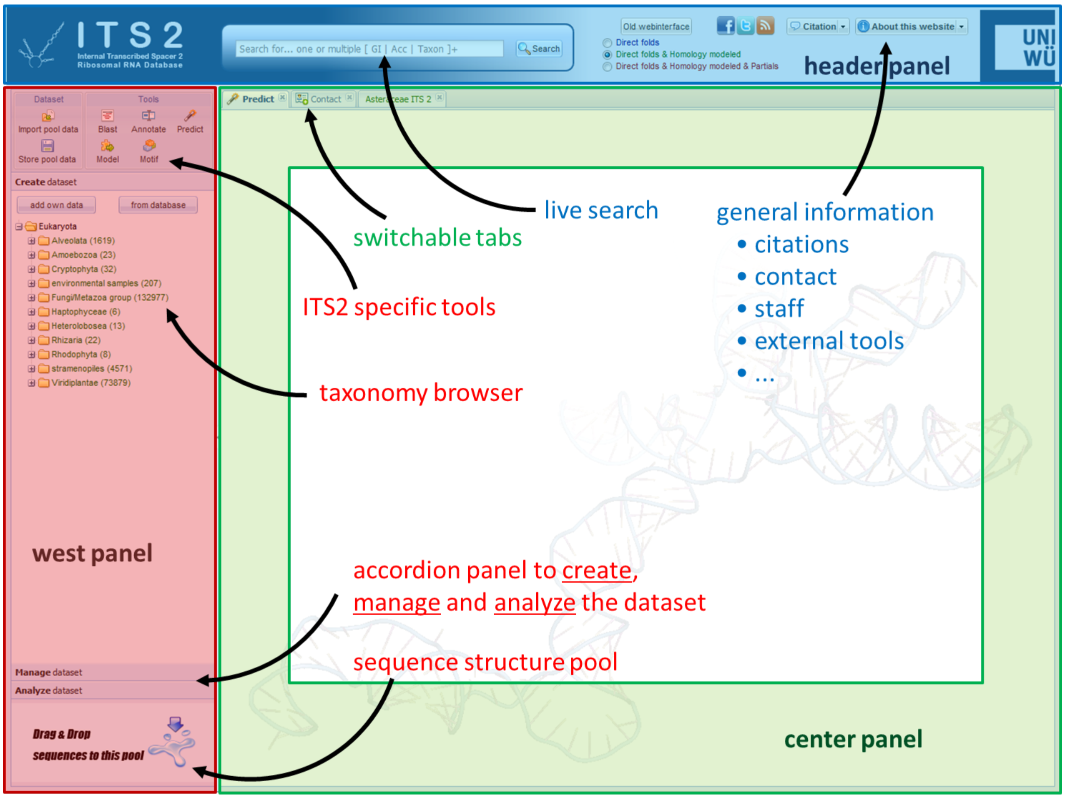


Figure 21: The ITS2 workbench is divided into header, west and center panel. The header contains a live search and custom information about the website as well as citations. The west panel gives access to typical marker-specific tools and shows the basic workflow for sequence-structure analysis with the data pool on the bottom left. The center panel allows several tabs to be opened simultaneously while working with the website.

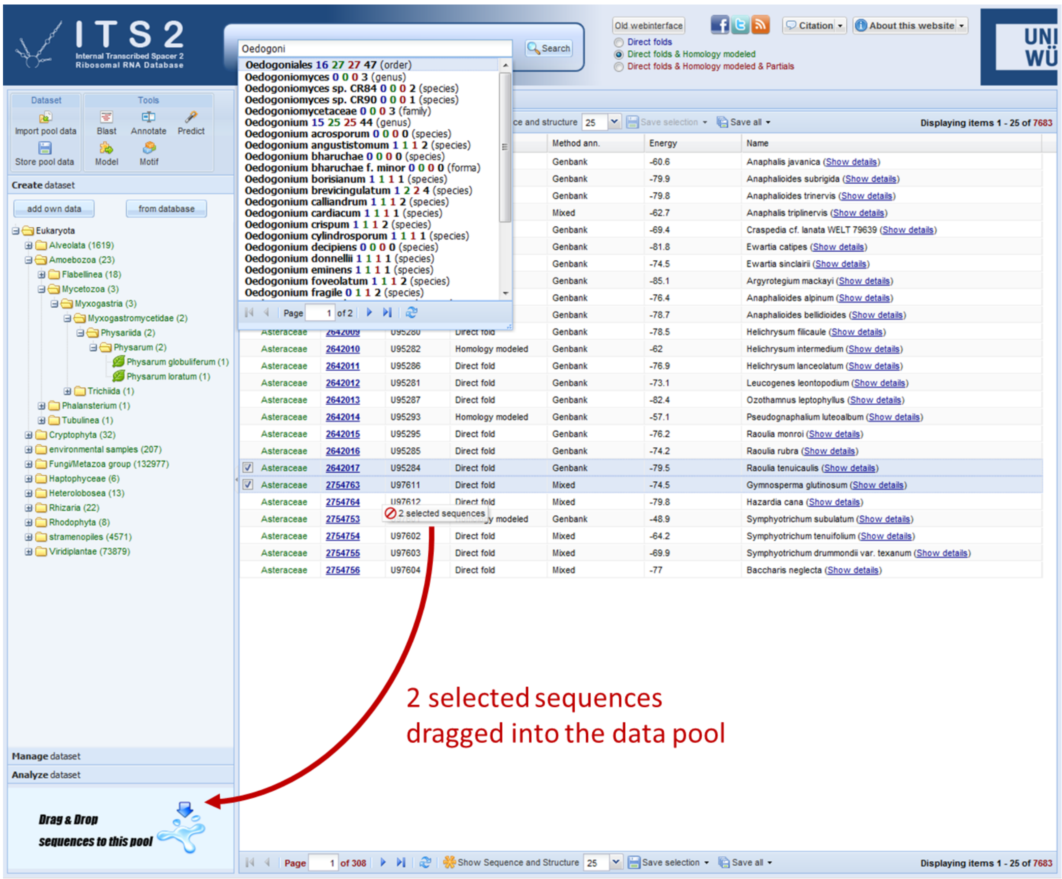
7.3 Creating a sequence-structure dataset

7.3.1 Datasets based on the ITS2 database

Typically, biologists aiming for a phylogenetic analysis, mainly a phylogenetic tree, start with the creation of a dataset. This means, they are interested in a taxonomic group and want to search for sequences, fitting into the taxon sampling. As prominently visible on the top, this can easily be accessed by the search function. The search accepts one search term, or a comma separated list. This list may contain GeneInfo Identifiers, Accession numbers or even several taxonomic names. To facilitate the input of taxonomic names, a live search is performed, completing the name based on a prefix matching of the first entered letters. Once the selection is finished and the search button is pressed, a large set of information is fetched from the database and becomes visible in the center panel. This is by default the searched group, the GeneInfo Identifier, the taxonomic name, the structure prediction method and of course sequence and structure itself. Depending on the radio buttons beside the search field, different categories are available that return sequences only, directly folded and homology modelled structures or all folded ones including partials. Furthermore, less frequently needed information such as the annotation method or sequence motifs can be displayed at any time by expanding the columns of the grid view. But performing a search based on pre-known search terms is only one of the ways to begin the creation of a dataset.

Beside the specific search function, a user might also want to browse through the tree of life on the left. This will give an overview of different groups available in the ITS2 database, while showing their number of sequences/structures in brackets next to the taxonomic name. By switching between the previously mentioned radio buttons for different structural types in the header, the tree itself changes its color and structure according to the available sequences. This allows to browse through sequences only, or to include partial structures in the tree.

Finally, by clicking on a taxonomic name, a new tab (like after the direct search) is opened in the center panel. Now, the user is able to view sequences and structures. The next step is adding them to a dataset. Therefore, a data pool is included at the website on the bottom left. The data pool is able to store sequences, structures as well as alignments or trees. To add sequence-structures, a user just needs to drag and drop columns from a grid onto the pool (see Figure 22).



○ Direct folds
● Direct folds & Homology modeled
○ Direct folds & Homology modeled & Partials

Oedogoni

Oedogoniales 16 27 27 47 (order)
Oedogoniomyces 0 0 0 3 (genus)
Oedogoniomyces sp. CR84 0 0 0 2 (species)
Oedogoniomyces sp. CR90 0 0 0 1 (species)
Oedogoniomycetaceae 0 0 0 3 (family)
Oedogonium 15 25 25 44 (genus)
Oedogonium acrosporum 0 0 0 0 (species)
Oedogonium angustistomum 1 1 1 2 (species)
Oedogonium bharuchae 0 0 0 0 (species)
Oedogonium bharuchae f. minor 0 0 0 0 (forma)
Oedogonium borisianum 1 1 1 1 (species)
Oedogonium brevicingulatum 1 2 2 4 (species)
Oedogonium calliandrum 1 1 1 2 (species)
Oedogonium cardiacum 1 1 1 1 (species)
Oedogonium crispum 1 1 1 2 (species)
Oedogonium cylindrosporum 1 1 1 1 (species)
Oedogonium decipiens 0 0 0 0 (species)
Oedogonium donnellii 1 1 1 1 (species)
Oedogonium eminens 1 1 1 1 (species)
Oedogonium foveolatum 1 1 1 2 (species)
Oedogonium fragile 0 1 1 2 (species)

Page 1 of 2

- Eukaryota
  - Alveolata (1619)
  - Amoebozoa (23)
    - Flabellinea (18)
    - Mycetozoa (3)
      - Myxogastria (3)
        - Myxogastromycetidae (2)
          - Physariida (2)
            - Physarum (2)
              - Physarum globuliferum (1)
              - Physarum loratum (1)
        - Trichiida (1)
    - Phalansterium (1)
    - Tubulinea (1)
  - Cryptophyta (32)
  - environmental samples (207)
  - Fungi/Metazoa group (132977)
  - Haptophyceae (6)
  - Heterolobosea (13)
  - Rhizaria (22)
  - Rhodophyta (8)
  - stramenopiles (4571)
  - Viridiplantae (73879)

ce and structure 25 Save selection Save all Displaying items 1 - 25 of 7683

					Method ann.	Energy	Name
					Genbank	-60.6	Anaphalis javanica (Show details)
					Genbank	-79.9	Anaphalioides subrigida (Show details)
					Genbank	-79.8	Anaphalioides trinervis (Show details)
					Mixed	-62.7	Anaphalis triplinervis (Show details)
					Genbank	-69.4	Craspedia cf. lanata WELT 79639 (Show details)
					Genbank	-81.8	Ewartia catipes (Show details)
					Genbank	-74.5	Ewartia sinclairii (Show details)
					Genbank	-85.1	Argyrotegium mackayi (Show details)
					Genbank	-76.4	Anaphalioides alpinum (Show details)
					Genbank	-78.7	Anaphalioides bellidioides (Show details)
	Asteraceae	2642009	U95280	Direct fold	Genbank	-78.5	Helichrysum filicaule (Show details)
	Asteraceae	2642010	U95282	Homology modeled	Genbank	-62	Helichrysum intermedium (Show details)
	Asteraceae	2642011	U95286	Direct fold	Genbank	-76.9	Helichrysum lanceolatum (Show details)
	Asteraceae	2642012	U95281	Direct fold	Genbank	-73.1	Leucogenes leontopodium (Show details)
	Asteraceae	2642013	U95287	Direct fold	Genbank	-82.4	Ozothamnus leptophyllus (Show details)
	Asteraceae	2642014	U95293	Homology modeled	Genbank	-57.1	Pseudognaphalium luteoalbum (Show details)
	Asteraceae	2642015	U95295	Direct fold	Genbank	-76.2	Raoulia monroi (Show details)
	Asteraceae	2642016	U95285	Direct fold	Genbank	-74.2	Raoulia rubra (Show details)
☑	Asteraceae	2642017	U95284	Direct fold	Genbank	-79.5	Raoulia tenuicaulis (Show details)
☑	Asteraceae	2754763	U97611	Direct fold	Mixed	-74.5	Gymnosperma glutinosum (Show details)
	Asteraceae	2754764	U97612	Direct fold	Mixed	-79.8	Hazardia cana (Show details)
	Asteraceae	2754753	U97601	Homology modeled	Genbank	-48.9	Symphyotrichum subulatum (Show details)
	Asteraceae	2754754	U97602	Direct fold	Mixed	-64.2	Symphyotrichum tenuifolium (Show details)
	Asteraceae	2754755	U97603	Direct fold	Mixed	-69.9	Symphyotrichum drummondii var. texanum (Show details)
	Asteraceae	2754756	U97604	Direct fold	Mixed	-77	Baccharis neglecta (Show details)

2 selected sequences dragged into the data pool

Page 1 of 308 Show Sequence and Structure 25 Save selection Save all Displaying items 1 - 25 of 7683

Figure 22: Using the live search function on the top, a user can easily scan for organisms. In addition, a taxonomic tree on the left enables the user to browse through all eukaryotes contained in the database. Clicked nodes inside the tree or searched organisms are presented in a new tab in the center panel. Here, specific information for each sequence is visible, such as its annotation methods, folding energy, Accession number, etc.. In order to create a dataset, these sequences can be added to the data pool by a simple selection in the grid, and a drag & drop onto the pool on the bottom left.

Per default, sequences are added including their secondary structures. If these are not available, just an unpaired structure is added to avoid conflicts when creating the alignment. Of course, sequence-structures can also be included by using the context menu which appears by a right click on a grid column.

7.3.2 Datasets from own sequences

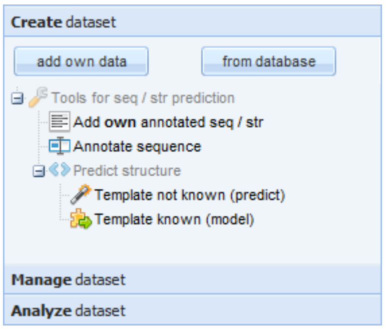
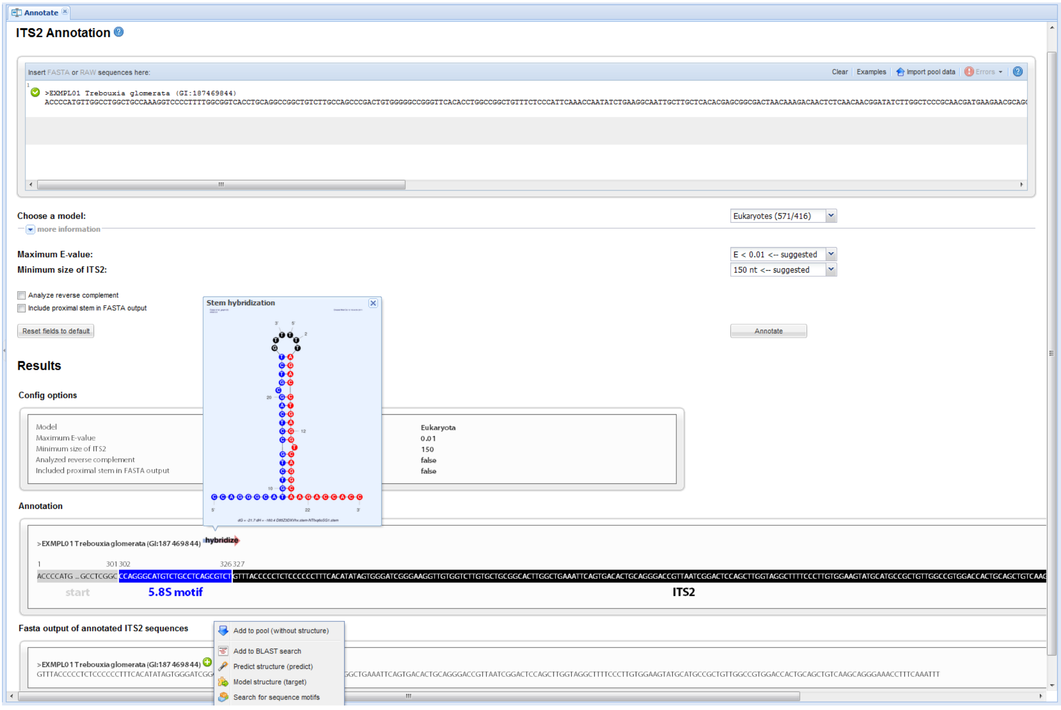


Figure 23: The accordion panel on the left side of the window allows to create, manage and analyse a dataset. To create a dataset, own sequences or ones from the database may be used.

Beside including sequences from the database directly, the "Create dataset" frame also offers to add own data. However, when importing user data, this might not be annotated yet or may lack a secondary structure. Therefore, a user possessing both, an annotated sequence and a secondary structure can use an upload form to directly add data to the pool, which is checked for consistency, before it is added. But in case that the structure is missing, or sequences are not yet annotated, a few more steps are required.

The ITS2 "Annotation" tool: Having sequenced own data, or obtained a complete rRNA cistron, the HMM annotation tool assists the user to locate the exact borders of the ITS2.



ITS2 Annotation

Insert FASTA or RAW sequences here:

```
>EXMPL01 Trebouxia glomerata (GI:187469844)
ACCCCATGTTGGCCTGGCTGCCAAAGGTCCCCTTTTGGCGGTCACCTGCAGGCCGGCTGTCTTGCCAGCCCGACTGTGGGGGCCGGGTTCACACCTGGCCGGCTGTTTCTCCCATTCAAACCAATATCTGAAGGCAATTGCTTGCTCACACGAGCGGCGACTAACAAAGACAACTCTCAACAACGGATATCTTGGCTCCCGCAACGATGAAGAACGCAG
```

Choose a model: Eukaryotes (571/416)

Maximum E-value: E < 0.01 <-- suggested

Minimum size of ITS2: 150 nt <-- suggested

☐ Analyze reverse complement

☐ Include proximal stem in FASTA output

Results

Config options

Model	Eukaryota
Maximum E-value	0.01
Minimum size of ITS2	150
Analyzed reverse complement	false
Included proximal stem in FASTA output	false

Annotation

>EXMPL01 Trebouxia glomerata (GI:187469844)

```
1         301 302                  326 327
ACCCCATG ...GCCTCGGC CCAGGGCATGTCTGCCTCAGCGTCT GTTTACCCCCTCTCCCCCCTTTCACATATAGTGGGATCGGGAAGGTTGTGGTCTTGTGCTGCGGCACTTGGCTGAAATTCAGTGACACTGCAGGGACCGTTAATCGGACTCCAGCTTGGTAGGCTTTTCCCTTGTGGAAGTATGCATGCCGCTGTTGGCCGTGGACCACTGCAGCTGTCAAG
```

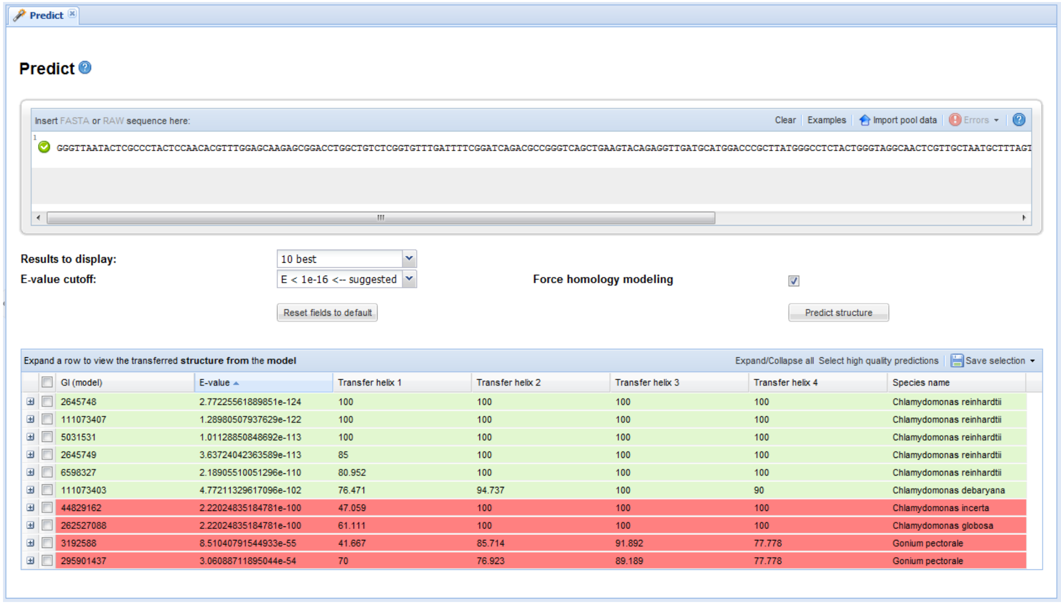
Fasta output of annotated ITS2 sequences

```
>EXMPL01 Trebouxia glomerata (GI:187469844)
GTTTACCCCCTCTCCCCCCTTTCACATATAGTGGGATCGG GGCTGAAATTCAGTGACACTGCAGGGACCGTTAATCGGACTCCAGCTTGGTAGGCTTTTCCCTTGTGGAAGTATGCATGCCGCTGTTGGCCGTGGACCACTGCAGCTGTCAAGCAGGGAAACCTTTCAAATTT
```

Figure 24: Results from an HMM based ITS2 sequence annotation are illustrated here. One can easily recognize the 5.8S motif identified on the left side of the ITS2 sequence. Further, the hybridizing proximal stems of both HMM motifs are visualized as control.

Therefore, a sequence is pasted into the sequence editor. After an automated validation of the sequence content and format, the green check mark appears on the left. Now the users may choose a model based on the sequence origin. Five different models are available for 'Eukaryota', 'Fungi', 'Diptera', 'Viridiplantae' or 'Metazoa' data. Additionally, choice boxes for e-value cut-off and minimum ITS2 length allow to reduce the probability of false hits during annotation. After pressing the "Annotate" button, the results are displayed below (see Figure 24). The flanking borders of the ITS2, the 5.8S and 28S motif, are depicted in blue and red. However, the annotation without these motifs (each containing 25 nucleotides) is not possible. To control the process, both hybridizing stems are displayed which should contain the typical free nucleotides on both stem sides. In case of inverse strand directions or no results, a user has also the possibility to analyse the reverse complement of the sequence by enabling the appropriate check box. With the HMM annotated ITS2 sequence, the next step towards a sequence-structure based phylogeny is the secondary structure prediction. The user clicks with the right button on the "add" - symbol, and transfers the sequence to the structure prediction tool.

Easy prediction of secondary structure – The "Predict" tool: There are two different types of structure prediction. To facilitate the decision for the user, again, the annotated sequence just needs to be copied into the sequence editor. Then, the best method for structure prediction is chosen automatically. First, a prediction based on energy minimization is performed. If this is unsuccessful, an automated sequence-based BLAST search starts against the whole ITS2 database with the predefined e-value cut-off listed below the input field. According to the given number of best hits, results from the homology modelling step are displayed in a grid (Figure 26).



```
1
GGGTTAATACTCGCCCTACTCCAACACGTTTGGAGCAAGAGCGGACCTGGCTGTCTCGGTGTTTGATTTTCGGATCAGACGCCGGGTCAGCTGAAGTACAGAGGTTGATGCATGGACCCGCTTATGGGCCTCTACTGGGTAGGCAACTCGTTGCTAATGCTTTAGT
```

Results to display: 10 best

E-value cutoff: E < 1e-16 <-- suggested

Force homology modeling ☑

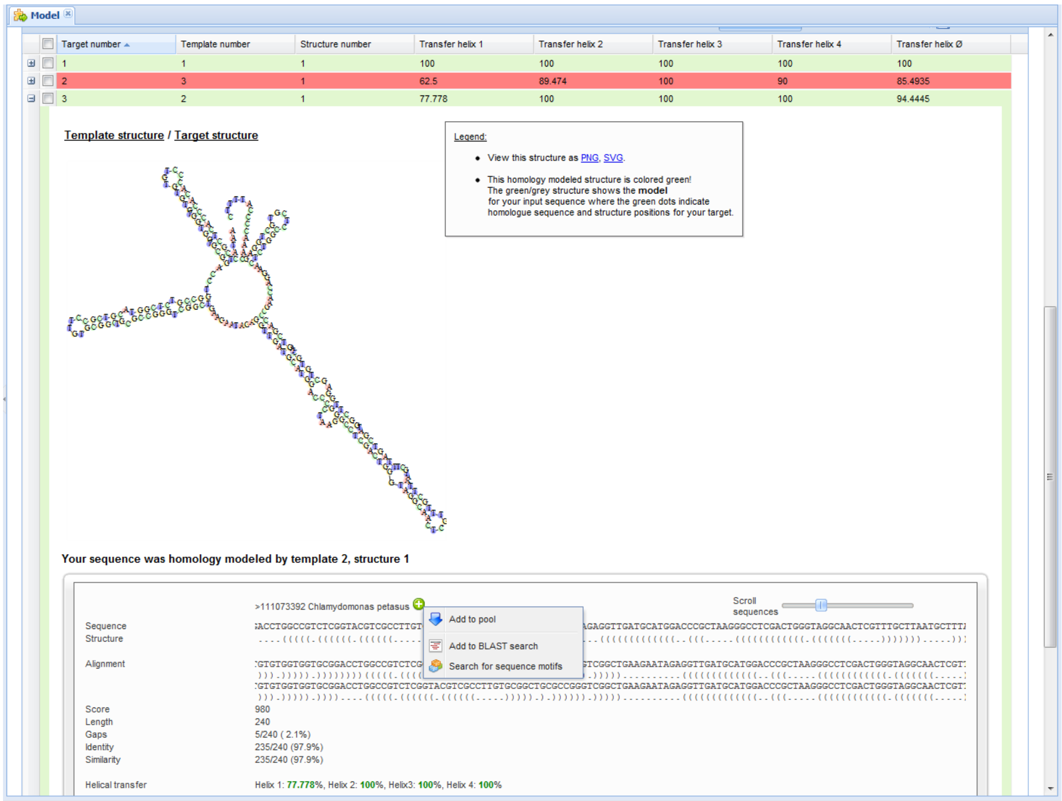
Expand a row to view the transferred **structure from** the **model**

	GI (model)	E-value	Transfer helix 1	Transfer helix 2	Transfer helix 3	Transfer helix 4	Species name
□	2645748	2.77225561889851e-124	100	100	100	100	Chlamydomonas reinhardtii
□	111073407	1.28980507937629e-122	100	100	100	100	Chlamydomonas reinhardtii
□	5031531	1.01128850848692e-113	100	100	100	100	Chlamydomonas reinhardtii
□	2645749	3.63724042363589e-113	85	100	100	100	Chlamydomonas reinhardtii
□	6598327	2.18905510051296e-110	80.952	100	100	100	Chlamydomonas reinhardtii
□	111073403	4.77211329617096e-102	76.471	94.737	100	90	Chlamydomonas debaryana
□	44829162	2.22024835184781e-100	47.059	100	100	100	Chlamydomonas incerta
□	262527088	2.22024835184781e-100	61.111	100	100	100	Chlamydomonas globosa
□	3192588	8.51040791544933e-55	41.667	85.714	91.892	77.778	Gonium pectorale
□	295901437	3.06088711895044e-54	70	76.923	89.189	77.778	Gonium pectorale

Figure 25: This page shows the secondary structure prediction on the ITS2 workbench. Having pasted a sequence into the editor, it might either be folded directly or become homology modelled. Here, the results of homology modelling are visible inside a grid. Green rows indicate a high quality prediction, red ones a low quality prediction with at least one helix below 75 %.

This grid is sorted by e-value and lists information about helix transfer and the taxonomic name of the template in green or red. Here, green rows indicate a high quality prediction, red rows indicate a prediction with at least one helix below 75 %. For more details, each row can be expanded, and shows the secondary structure, alignment statistics and helical transfer. For low quality models, the BLAST identified template structure is visible and coloured with green and gray. The green dots indicate equal sequence and structure positions between template and target. Having identified an appropriate template, the modelled sequence-structure can now be added to the data pool by a simple drag & drop or via the context menu.

Homology Modelling of structure – the "Model" tool: Homology modelling of a structure using the "Model" tool is quite similar to the previously explained prediction tool, however it has a major difference. Whereas in the previous section a template was unknown and identified by a BLAST search, here the user has the possibility to specify an own template. Further, not only one, but even several templates in combination with multiple secondary structures are imaginable as model input. The algorithm then automatically chooses the best out of all possible combinations. If wished, also the best consensus template can be calculated.



Target number	Template number	Structure number	Transfer helix 1	Transfer helix 2	Transfer helix 3	Transfer helix 4	Transfer helix Ø
1	1	1	100	100	100	100	100
2	3	1	62.5	89.474	100	90	85.4935
3	2	1	77.778	100	100	100	94.4445

Template structure / Target structure

Legend:

- View this structure as PNG, SVG.
- This homology modeled structure is colored green! The green/grey structure shows the **model** for your input sequence where the green dots indicate homologue sequence and structure positions for your target.

Your sequence was homology modeled by template 2, structure 1

>111073392 Chlamydomonas petasus

```
Sequence      :ACCTGGCCGTCTCGGTACGTCGCCTTGT            AGAGGTTGATGCATGGACCCGCTAAGGGCCTCGACTGGGTAGGCAACTCGTTTGCTTAATGCTTTA
Structure     ....(((((.((((((.((((((.....            ..((((((((((((..(((.....(((((((((((.(((((((.....)))))))).....))

Alignment     :GTGTGGTGGTGCGGACCTGGCCGTCTCG            TCGGCTGAAGAATAGAGGTTGATGCATGGACCCGCTAAGGGCCTCGACTGGGTAGGCAACTCGT
              )))).)))))).))))))))(((((.((((            .)))))..........((((((((((((..(((.....(((((((((((.(((((((.....
              :GTGTGGTGGTGCGGACCTGGCCGTCTCGGTACGTCGCCTTGTGCGGCTGCGCCGGGTCGGCTGAAGAATAGAGGTTGATGCATGGACCCGCTAAGGGCCTCGACTGGGTAGGCAACTCGT
              )))).)))))).))))....(((((.((((((.((((((.....)))))).).)))))).)))))..........((((((((((((..(((.....(((((((((((.(((((((.....

Score         980
Length        240
Gaps          5/240 ( 2.1%)
Identity      235/240 (97.9%)
Similarity    235/240 (97.9%)

Helical transfer   Helix 1: 77.778%, Helix 2: 100%, Helix3: 100%, Helix 4: 100%
```

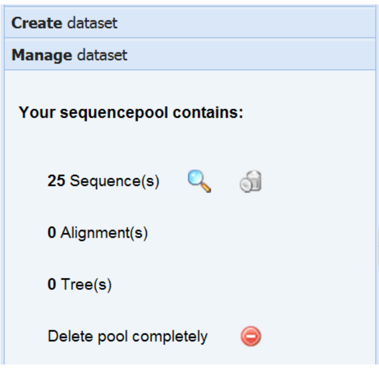
Figure 26: The Figure shows a homology modelled structure on the ITS2 workbench. High quality predictions are marked green, low quality predictions are coloured red. By expanding a row, the secondary structure of the target sequence becomes visible. Additionally, further information such as the alignment is available.

Beside these extensions, further advanced options are available for the more experienced user. These are alignment specific gap costs, different score matrices as well as adjustable thresholds for high quality models.

The visualization of results is similar to the "Predict" tool. A grid separates high quality (green) from low quality (red) predictions. Each grid row can be expanded to view the modelling details (Figure 26). High quality predictions can further be selected automatically for download, or for the transfer into the pool.

7.4 Managing the dataset

Having created a dataset, the next step allows to manage it. This means visual inspection, correction or deletion of sequence-structures, alignment or tree (Figure 27).



Create dataset

Manage dataset

Your sequencepool contains:

25 Sequence(s)

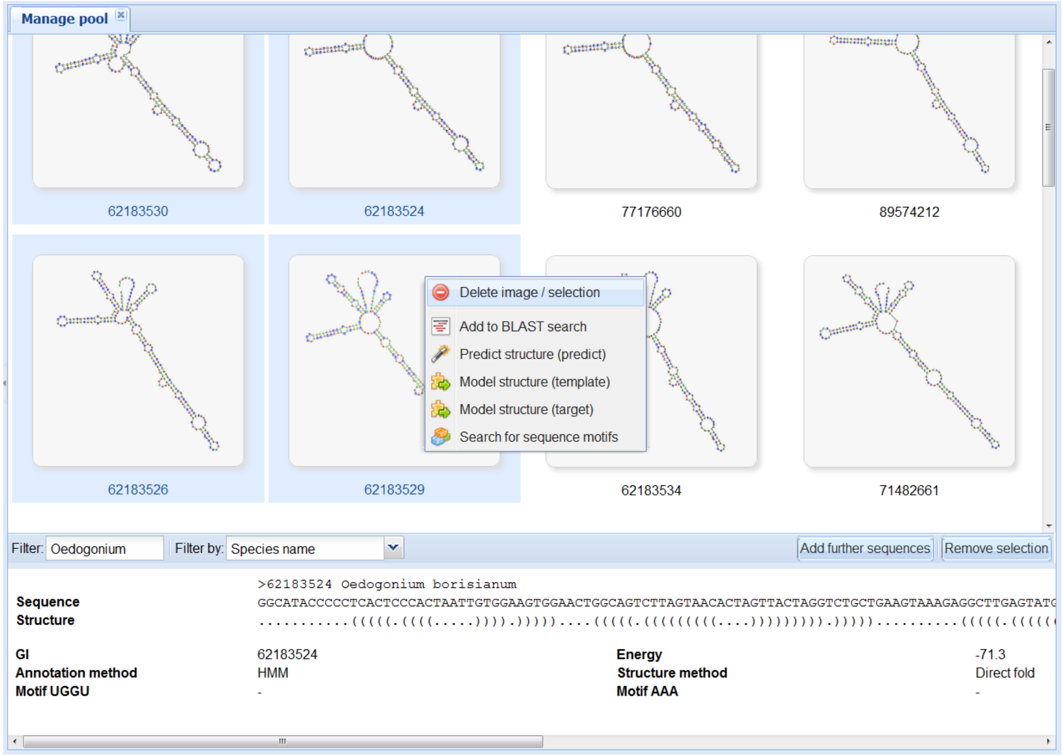
0 Alignment(s)

0 Tree(s)

Delete pool completely

Figure 27: Managing a dataset in the ITS2 workbench means to view structures in the pool, see an alignment or to delete a tree for example. However, not all possibilities are available at any time.

By now, only sequence-structures have been added to the pool. These can be displayed by a click on the loupe, or the pool itself. When a structure is available, the folded conformation of the ITS2 is illustrated in small pictograms (Figure 28). Otherwise, icons for sequence only or partial structures are available. With a left click on an image, it becomes highlighted, and information about sequence motifs, energy, etc. appears at the bottom. To be able to deal with big datasets, a filter option allows to keep track of data. It permits to focus on sequences matching specific search-criteria like species name, GeneInfo Identifier, energy, structure or annotation method, respectively.



Filter: Oedogonium Filter by: Species name

```
                      >62183524 Oedogonium borisianum
Sequence              GGCATACCCCCTCACTCCCACTAATTGTGGAAGTGGAACTGGCAGTCTTAGTAACACTAGTTACTAGGTCTGCTGAAGTAAAGAGGCTTGAGTATG
Structure             ...........(((((.((((.....)))).)))))....(((((.(((((((((....))))))))).)))))..........(((((.((((

GI                    62183524                        Energy              -71.3
Annotation method     HMM                             Structure method    Direct fold
Motif UGGU            -                               Motif AAA           -
```

Figure 28: The ITS2 workbench allows to manage sequence-structures, and illustrates the structural conformation in small overview pictures. By applying a filter, specific structures can be focussed. When marking a structure, information about sequence motifs or the annotation method becomes visible in a small panel below. Further, structures can be deleted or transferred to different tools in other tabs.

Finally, sequences can be marked by a mouse click in combination with the control key. A right click on a marked sequence then shows a context menu enabling the deletion of structures from the pool, or their transfer to other tools in different tabs.

7.5 Analysing the dataset

Having managed the dataset and maybe already deleted some structures by manual inspection in view of the taxon sampling, the next step in a phylogenetic analysis is the creation of an alignment, and further the calculation of a phylogenetic tree (see Figure 29).

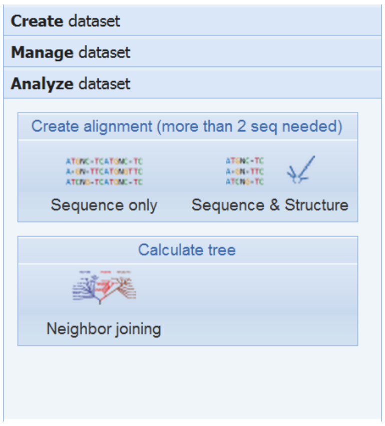
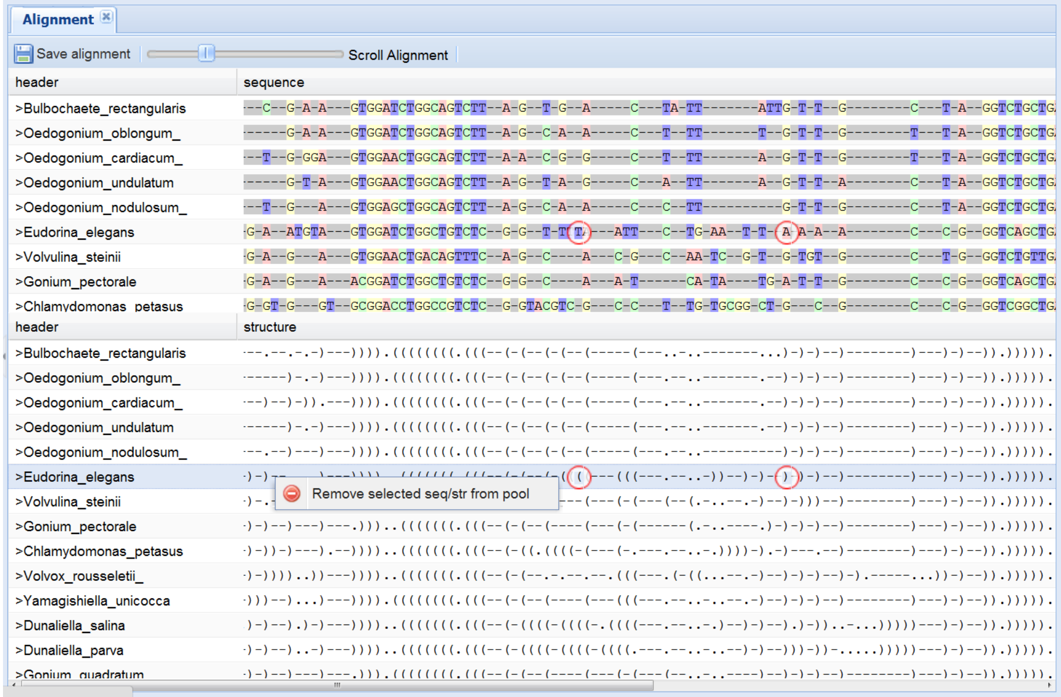


Figure 29: When analysing a dataset on the ITS2 workbench, different options are available. If the pool contains secondary structure information, a sequence or sequence-structure based alignment can be calculated. Afterwards, the option for calculating a Neighbour Joining tree becomes highlighted.

7.5.1 Alignment on sequence or sequence and structure

The alignment can be performed on sequence only, or, if secondary structure is included in the dataset, on both characteristics of the marker. Not visible to the user, ITS2 specific gap costs and score matrices (see Chapter 4.3) are integrated during the alignment process. Having finished calculating, the alignment opens in a new tab (Figure 30). This shows the sequence headers on the left, the sequence alignment itself on the top, and if available the secondary structure alignment on the bottom. By clicking on a nucleotide or structural bond, the corresponding ones become highlighted with a red circle. Further, columns can be rearranged by drag & drop to compare sequences or structures directly. Having identified an erroneous sequence or structure in the alignment, this can be deleted directly with a right click in the context menu. Naturally, the user is asked whether he wants the alignment to be recalculated afterwards. Finally, the alignment can be stored for the use in external programs, or the study can be continued with the calculation of a tree.



header	sequence
>Bulbochaete_rectangularis	---C--G-A-A---GTGGATCTGGCAGTCTT--A-G--T-G--A-----C---TA-TT-------ATTG-T-T--G--------C---T-A--GGTCTGCTG
>Oedogonium_oblongum_	-----G-A-A---GTGGATCTGGCAGTCTT--A-G--C-A--A-----C---T--TT-------T--G-T-T--G--------T---T-A--GGTCTGCTG
>Oedogonium_cardiacum_	---T--G-GGA---GTGGAACTGGCAGTCTT--A-A--C-G--G-----C---T--TT-------A--G-T-T--G--------T---T-A--GGTCTGCTG
>Oedogonium_undulatum	-----G-T-A---GTGGAACTGGCAGTCTT--A-G--T-A--G-----C---A--TT-------A--G-T-T--A--------C---T-A--GGTCTGCTG
>Oedogonium_nodulosum_	---T--G---A---GTGGAGCTGGCAGTCTT--A-G--C-A--A-----C---C--TT----------G-T-T--G--------C---T-A--GGTCTGCTG
>Eudorina_elegans	-G-A--ATGTA---GTGGATCTGGCTGTCTC--G-G--T-TTTA---ATT---C--TG-AA--T-T--A-A-A--A--------C---C-G--GGTCAGCTG
>Volvulina_steinii	-G-A--G---A---GTGGAACTGACAGTTTC--A-G--C----A---C-G---C--AA-TC--G-T--G-TGT--G--------C---T-G--GGTCTGTTG
>Gonium_pectorale	-G-A--G---A---ACGGATCTGGCTGTCTC--G-G--C----A---A-T------CA-TA----TG-A-T-T--G--------C---C-G--GGTCAGCTG
>Chlamydomonas_petasus	-G-GT-G---GT--GCGGACCTGGCCGTCTC--G-GTACGTC-G---C-C---T--TG-TGCGG-CT-G---C--G--------C---C-G--GGTCGGCTG

header	structure
>Bulbochaete_rectangularis	---.--.-.-)---)))).((((((((.(((--(-(--(-(--(-----(---..-..-------...)-)-)--)--------)---)-)--)).))))).
>Oedogonium_oblongum_	------)-.-)---)))).((((((((.(((--(-(--(-(--(-----(---.--..-------.--)-)-)--)--------)---)-)--)).))))).
>Oedogonium_cardiacum_	---)--)-)).---)))).((((((((.(((--(-(--(-(--(-----(---.--..-------.--)-)-)--)--------)---)-)--)).))))).
>Oedogonium_undulatum	------)-.-)---)))).((((((((.(((--(-(--(-(--(-----(---.--..-------.--)-)-)--)--------)---)-)--)).))))).
>Oedogonium_nodulosum_	---.--)---)---)))).((((((((.(((--(-(--(-(--(-----(---.--..----------)-)-)--)--------)---)-)--)).))))).
>Eudorina_elegans	-)-)--)---)))) ((((((((((--(-(--(-(((---(((---.--..-))--)-)-))-)--)--------)---)-)--)).))))).
>Volvulina_steinii	-)-. ---(---(-(---(--(.-..--.-)--)-)))--)--------)---)-)--)).))))).
>Gonium_pectorale	-)-)--)---)---.)))..(((((((.(((--(-(--(----(---(-(------(.-..----.)-)-)-)--)--------)---)-)--)).))))).
>Chlamydomonas_petasus	-)-))-)---).--))))..(((((((.(((--(-((.((((-(---(-.---.--..-.))))-).-)---.--)--------)---)-)--)).))))).
>Volvox_rousseletii_	-)-))))..))---))))..(((((((.(((--(-(--.-.--.--.(((---.(-((...--.-)--)-)-)--)-).-----...))-)--)).))))).
>Yamagishiella_unicocca	-)))--)...)---)))).((((((((.(((--(-(--(----(---(((---.--..-..--.-)--)-)-)--)--------)---)-)--)).))))).
>Dunaliella_salina	)-)--).)-)---))))..(((((((.(((--(-((((-((((-.((((---.--..-.)--)-)--).)-))..-...)))))---)-)--)).))))).
>Dunaliella_parva	-)-)--)..-)---))))..(((((((.(((--(-((((-((((-((((.---.--..-..--)-)--)))-))-.....)))))---)-)--)).))))).
>Gonium_quadratum	-)-)--)---)---.))).((((((((.(((--(-(--(----(---(-(---(--..-..----)--)-)-)--)--------)---)-)--)).))))).

Figure 30: In this Figure, a sequence-structure alignment of the ITS2 workbench is presented. The alignment of sequence information is visible on the top, whereas the structure alignment can be found at the bottom. Corresponding bonds can be highlighted with a simple click and are marked with red circles. To remove a sequence from the alignment and thereby also from the pool, a context menu is available.

7.5.2 Tree calculation with the Neighbour Joining algorithm

The tree icon is visible on the ”Analyze” tab, as soon as the alignment is created. Currently, the user can only calculate a tree based on the fast Neighbour Joining algorithm using ProfDistS (Wolf et al., 2008). The tree opens in a new tab (Figure 31), and allows zooming and scrolling to different nodes. Here, a user aiming for a good taxon sampling can easily inspect the tree, and delete a node just by opening the context menu. Further, a re-rooting of the tree at each node is possible. As soon as the dataset has changed, calculations are repeated automatically when switching between tabs. Although the calculation of the tree is primarily offered as first impression on the dataset, a download in Newick format is available. It must be stated however, that this tree is by no means a publication ready tree. Therefore, further treeing methods should be taken into consideration, and at least a bootstrap analysis would be required.

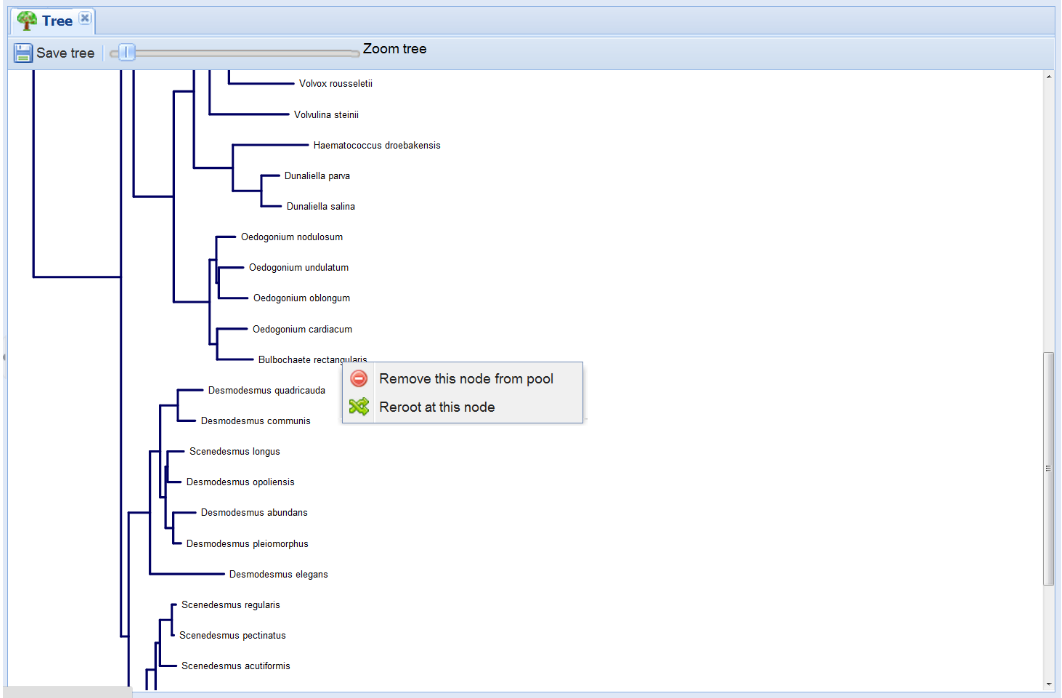


Figure 31: This graphic shows a tree calculated on the ITS2 workbench. With a click on a node, the context menu allows to delete or re-root this position. The tree can be scaled and moved inside the panel, and is available for download in Newick format.

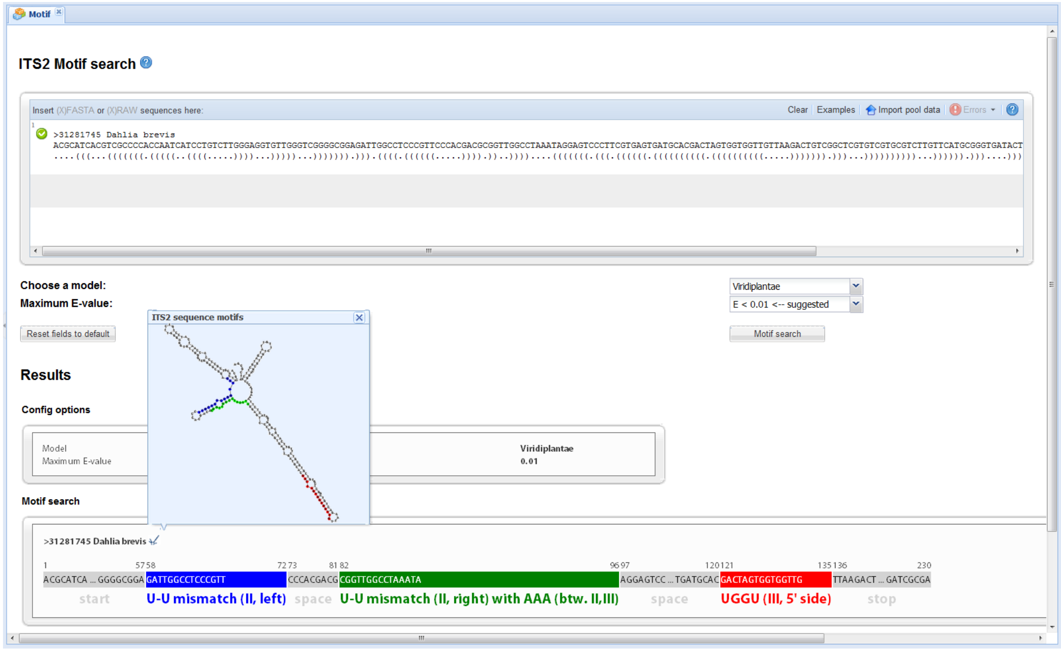
7.6 Additional tools

7.6.1 Motif detection in ITS2 sequences

Beside the typical pipeline for the creation of a taxon sampling, the ITS2 workbench offers some additional tools around the marker. One is the ITS2 sequence-motif detection. Sequence motifs are conserved parts or regions which occur in a large proportion of a dataset, here, inside the ITS2. Although these are not of major importance for phylogenetic analyses, they can assist in the review of sequence annotation or secondary structure prediction. Three basic sequence motifs were identified with the motif search tool MEME (Bailey et al., 2009), based on the dataset of Keller et al. (2009):

- U-U mismatch (helix II, left),
- U-U mismatch (helix II, right) with AAA (btw. helix II and III),
- UGGU (helix III, 5' side)

Together with the two free nucleotides of the hybridizing stem regions 5.8S and 28S, five characteristics are available that ensure the correctness of a sequence annotation. The motif prediction on the workbench follows the typical procedure. A sequence is pasted into the sequence editor, or transferred to the "Motif" tool from a different tab. When providing additional structure information, the motifs become highlighted in their folded conformation later.



ITS2 Motif search

Insert (X)FASTA or (X)RAW sequences here:

```
>31281745 Dahlia brevis
ACGCATCACGTCGCCCCACCAATCATCCTGTCTTGGGAGGTGTTGGGTCGGGGCGGAGATTGGCCTCCCGTTCCCACGACGCGGTTGGCCTAAATAGGAGTCCCTTCGTGAGTGATGCACGACTAGTGGTGGTTGTTAAGACTGTCGGCTCGTGTCGTGCGTCTTGTTCATGCGGGTGATACT
....(((...((((((.(((((..((((.....))))...)))))...))))))).))).((((.((((((.....)))).))..))))....(((((((.(((.((((((.((((((((((.((((((((((.....)))))).)))...))))))))))...))))))).))).....)))
```

Choose a model: Viridiplantae

Maximum E-value: E < 0.01 <-- suggested

Results

Config options

Model Viridiplantae

Maximum E-value 0.01

Motif search

>31281745 Dahlia brevis

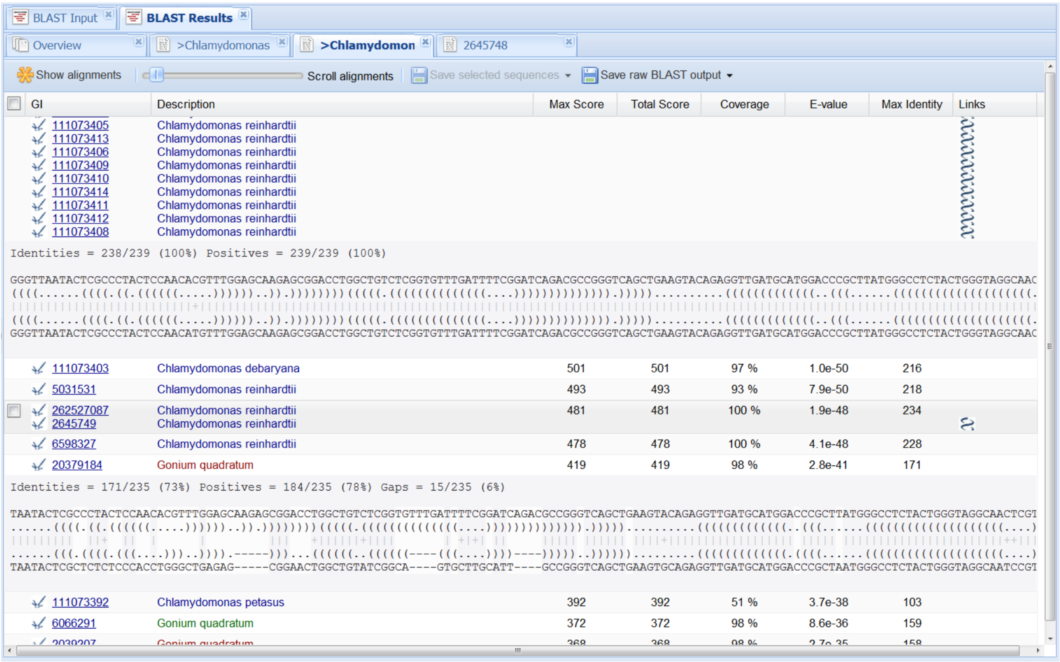
1 57	58 72	73 81	82 96	97 120	121 135	136 230
ACGCATCA ... GGGGCGGA	GATTGGCCTCCCGTT	CCCACGACG	CGGTTGGCCTAAATA	AGGAGTCC ... TGATGCAC	GACTAGTGGTGGTTG	TTAAGACT ... GATCGCGA
start	U-U mismatch (II, left)	space	U-U mismatch (II, right) with AAA (btw. II,III)	space	UGGU (III, 5' side)	stop

Figure 32: The motif search tool on the ITS2 workbench detects three sequence motifs. When entering additional structure information, the motifs are highlighted in a small coloured pictogram. This is ideal for reviewing the correctness of a folded structure.

However, the presence of a secondary structure is no prerequisite for motif detection. Two choice boxes for e-value threshold and model (possible models are 'Fungi' and 'Viridiplantae') are available to configure the HMM based motif search. Having clicked on the search button, the sequence is scanned for all three motifs, and results are displayed in a detailed table below (see Figure 32). If, in addition to sequence, also structure was added as input, the according sequence motifs are now coloured accordingly when opening the structure view.

7.6.2 BLAST on sequence and structure

A tool that completes the taxon sampling process is the newly developed sequence-structure BLAST. The ITS2 sequence is a fast evolving marker, and thus provides a high resolution on lower taxonomic groups. Thus, a BLAST search on sequence only, will mainly detect closely related species. For a well balanced taxon sampling instead, also more distantly related species are needed. The sequence-structure BLAST search fulfils this requirement by combining sequence and secondary structure into its search. Furthermore, this enables the identification of distantly related species inside the ITS2 workbench. When opening the according tab, one can choose between sequence only or sequence-structure BLAST and paste the sequence-structures into the editor. After the search has finished, results are available in new tabs, one for each sequence. Here, all relevant information about score, e-value or coverage are visualized inside a grid. With a double click, a row expands and the alignment becomes visible (Figure 33).



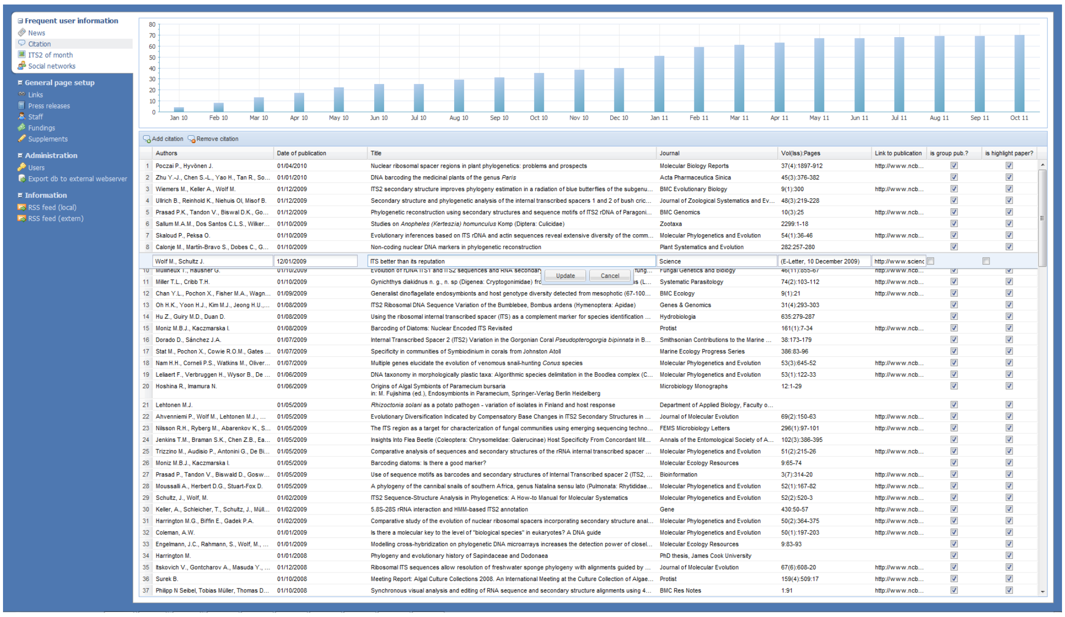
```
GI          Description                  Max Score  Total Score  Coverage  E-value  Max Identity  Links
111073405   Chlamydomonas reinhardtii
111073413   Chlamydomonas reinhardtii
111073406   Chlamydomonas reinhardtii
111073409   Chlamydomonas reinhardtii
111073410   Chlamydomonas reinhardtii
111073414   Chlamydomonas reinhardtii
111073411   Chlamydomonas reinhardtii
111073412   Chlamydomonas reinhardtii
111073408   Chlamydomonas reinhardtii
Identities = 238/239 (100%) Positives = 239/239 (100%)

GGGTTAATACTCGCCCTACTCCAACACGTTTGGAGCAAGAGCGGACCTGGCTGTCTCGGTGTTTGATTTTCGGATCAGACGCCGGGTCAGCTGAAGTACAGAGGTTGATGCATGGACCCGCTTATGGGCCTCTACTGGGTAGGCAAC
((((......((((.((.((((((.....))))))..)).))))))))(((((.((((((((((((((....)))))))))))))).))))).........((((((((((((..(((......(((((((((((((((((((((.
||||||||||||||||||||||||||+||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
((((......((((.((.((((((.....))))))..)).))))))))(((((.((((((((((((((....)))))))))))))).))))).........((((((((((((..(((......(((((((((((((((((((((.
GGGTTAATACTCGCCCTACTCCAACATGTTTGGAGCAAGAGCGGACCTGGCTGTCTCGGTGTTTGATTTTCGGATCAGACGCCGGGTCAGCTGAAGTACAGAGGTTGATGCATGGACCCGCTTATGGGCCTCTACTGGGTAGGCAAC

111073403   Chlamydomonas debaryana      501        501          97 %      1.0e-50  216
5031531     Chlamydomonas reinhardtii    493        493          93 %      7.9e-50  218
262527087   Chlamydomonas reinhardtii    481        481          100 %     1.9e-48  234
2645749     Chlamydomonas reinhardtii
6598327     Chlamydomonas reinhardtii    478        478          100 %     4.1e-48  228
20379184    Gonium quadratum             419        419          98 %      2.8e-41  171
Identities = 171/235 (73%) Positives = 184/235 (78%) Gaps = 15/235 (6%)

TAATACTCGCCCTACTCCAACACGTTTGGAGCAAGAGCGGACCTGGCTGTCTCGGTGTTTGATTTTCGGATCAGACGCCGGGTCAGCTGAAGTACAGAGGTTGATGCATGGACCCGCTTATGGGCCTCTACTGGGTAGGCAACTCGT
......((((.((.((((((.....))))))..)).))))))))(((((.((((((((((((((....)))))))))))))).))))).........((((((((((((..(((......(((((((((((((((((((((....)
||||||||   ||+  ||  |      |        |||   +||||||+||||        |  +|+|  ||     |||     |||||  ||||||  ||||+||||||||||||||||||  |||||  |||||||||||||||||||++|||
......(((.((((.(((....)))..)))).-----)))...((((((..((((((----(((....))))----)))))..))))))..........(((((((((((((.((((......(((((((((((((((((((((....)
TAATACTCGCTCTCTCCCACCTGGGCTGAGAG-----CGGAACTGGCTGTATCGGCA----GTGCTTGCATT----GCCGGGTCAGCTGAAGTGCAGAGGTTGATGCATGGACCCGCTAATGGGCCTCTACTGGGTAGGCAATCCGT

111073392   Chlamydomonas petasus        392        392          51 %      3.7e-38  103
6066291     Gonium quadratum             372        372          98 %      8.6e-36  159
2039207     Gonium quadratum             368        368          98 %      2.7e-35  158
```

Figure 33: This Figure shows the results of a sequence-structure BLAST search on the ITS2 workbench. This BLAST type allows to detect also more distantly related species due to its integration of secondary structure into the search. With a double click on a row, the full alignment appears on the screen.

Finally, the alignment rows can also be transferred into the pool per drag & drop, whereas buttons on the top allow to store selected sequences or even the complete BLAST output in different formats.

7.7 The ITS2 admin interface

The last tool completing the ITS2 workbench is its administration interface. This website allows to manage most dynamic contents like citations (Figure 34) in simple, editable grids.



	Authors	Date of publication	Title	Journal	Vol(Iss):Pages	Link to publication	is group pub.?	is highlight paper?
1	Poczai P., Hyvönen J.	01/04/2010	Nuclear ribosomal spacer regions in plant phylogenetics: problems and prospects	Molecular Biology Reports	37(4):1897-912	http://www.ncb...	☑	☑
2	Zhu Y.-J., Chen S.-L., Yao H., Tan R., So...	01/01/2010	DNA barcoding the medicinal plants of the genus *Paris*	Acta Pharmaceutica Sinica	45(3):376-382		☑	☑
3	Wiemers M., Keller A., Wolf M.	01/12/2009	ITS2 secondary structure improves phylogeny estimation in a radiation of blue butterflies of the subgenu...	BMC Evolutionary Biology	9(1):300	http://www.ncb...	☑	☑
4	Ullrich B., Reinhold K., Niehuis Ol, Misof B.	01/12/2009	Secondary structure and phylogenetic analysis of the internal transcribed spacers 1 and 2 of bush cric...	Journal of Zoological Systematics and Ev...	48(3):219-228		☑	☑
5	Prasad P.K., Tandon V., Biswal D.K., Go...	01/12/2009	Phylogenetic reconstruction using secondary structures and sequence motifs of ITS2 rDNA of Paragoni...	BMC Genomics	10(3):25	http://www.ncb...	☑	☑
6	Sallum M.A.M., Dos Santos C.L.S., Wilker...	01/10/2009	Studies on *Anopheles (Kerteszia) homunculus* Komp (Diptera: Culicidae)	Zootaxa	2299:1-18		☑	☑
7	Skaloud P., Peksa O.	01/10/2009	Evolutionary inferences based on ITS rDNA and actin sequences reveal extensive diversity of the comm...	Molecular Phylogenetics and Evolution	54(1):36-46	http://www.ncb...	☑	☑
8	Calonje M., Martín-Bravo S., Dobes C., G...	01/10/2009	Non-coding nuclear DNA markers in phylogenetic reconstruction	Plant Systematics and Evolution	282:257-280		☑	☑
	Wolf M., Schultz J.	12/01/2009	ITS better than its reputation	Science	(E-Letter, 10 December 2009)	http://www.scienc	☐	☐
10	Mullineux T., Hausner G.	01/10/2009	Evolution of rDNA ITS1 and ITS2 sequences and RNA secondary ... fung...	Fungal Genetics and Biology	46(11):855-67	http://www.ncb...	☑	☑
11	Miller T.L., Cribb T.H.	01/10/2009	Gynichthys diakidnus n. g., n. sp (Digenea: Cryptogonimidae) fro... s (L...	Systematic Parasitology	74(2):103-112	http://www.ncb...	☑	☑
12	Chan Y.L., Pochon X., Fisher M.A., Wagn...	01/09/2009	Generalist dinoflagellate endosymbionts and host genotype diversity detected from mesophotic (67-100...	BMC Ecology	9(1):21	http://www.ncb...	☑	☑
13	Oh H.K., Yoon H.J., Kim M.J., Jeong H.U.,...	01/08/2009	ITS2 Ribosomal DNA Sequence Variation of the Bumblebee, Bombus ardens (Hymenoptera: Apidae)	Genes & Genomics	31(4):293-303		☑	☑
14	Hu Z., Guiry M.D., Duan D.	01/08/2009	Using the ribosomal internal transcribed spacer (ITS) as a complement marker for species identification ...	Hydrobiologia	635:279-287		☑	☑
15	Moniz M.B.J., Kaczmarska I.	01/08/2009	Barcoding of Diatoms: Nuclear Encoded ITS Revisited	Protist	161(1):7-34	http://www.ncb...	☑	☑
16	Dorado D., Sánchez J.A.	01/07/2009	Internal Transcribed Spacer 2 (ITS2) Variation in the Gorgonian Coral *Pseudopterogorgia bipinnata* in B...	Smithsonian Contributions to the Marine ...	38:173-179		☑	☑
17	Stat M., Pochon X., Cowie R.O.M., Gates ...	01/07/2009	Specificity in communities of Symbiodinium in corals from Johnston Atoll	Marine Ecology Progress Series	386:83-96		☑	☑
18	Nam H.H., Corneli P.S., Watkins M., Oliver...	01/07/2009	Multiple genes elucidate the evolution of venomous snail-hunting *Conus* species	Molecular Phylogenetics and Evolution	53(3):645-52	http://www.ncb...	☑	☑
19	Leliaert F., Verbruggen H., Wysor B., De ...	01/06/2009	DNA taxonomy in morphologically plastic taxa: Algorithmic species delimitation in the Boodlea complex (C...	Molecular Phylogenetics and Evolution	53(1):122-33	http://www.ncb...	☑	☑
20	Hoshina R., Imamura N.	01/06/2009	Origins of Algal Symbionts of Paramecium bursaria in: M. Fujishima (ed.), Endosymbionts in Paramecium, Springer-Verlag Berlin Heidelberg	Microbiology Monographs	12:1-29		☑	☑
21	Lehtonen M.J.	01/05/2009	*Rhizoctonia solani* as a potato pathogen - variation of isolates in Finland and host response	Department of Applied Biology, Faculty o...			☑	☑
22	Ahvenniemi P., Wolf M., Lehtonen M.J., ...	01/05/2009	Evolutionary Diversification Indicated by Compensatory Base Changes in ITS2 Secondary Structures in ...	Journal of Molecular Evolution	69(2):150-63	http://www.ncb...	☑	☑
23	Nilsson R.H., Ryberg M., Abarenkov K., S...	01/05/2009	The ITS region as a target for characterization of fungal communities using emerging sequencing techno...	FEMS Microbiology Letters	296(1):97-101	http://www.ncb...	☑	☑
24	Jenkins T.M., Braman S.K., Chen Z.B., Ea...	01/05/2009	Insights Into Flea Beetle (Coleoptera: Chrysomelidae: Galerucinae) Host Specificity From Concordant Mit...	Annals of the Entomological Society of A...	102(3):386-395		☑	☑
25	Trizzino M., Audisio P., Antonini G., De Bi...	01/05/2009	Comparative analysis of sequences and secondary structures of the rRNA internal transcribed spacer ...	Molecular Phylogenetics and Evolution	51(2):215-26	http://www.ncb...	☑	☑
26	Moniz M.B.J., Kaczmarska I.	01/05/2009	Barcoding diatoms: Is there a good marker?	Molecular Ecology Resources	9:65-74		☑	☑
27	Prasad P., Tandon V., Biswald D., Gosw...	01/05/2009	Use of sequence motifs as barcodes and secondary structures of Internal Transcribed spacer 2 (ITS2, ...	Bioinformation	3(7):314-20	http://www.ncb...	☑	☑
28	Moussalli A., Herbert D.G., Stuart-Fox D.	01/05/2009	A phylogeny of the cannibal snails of southern Africa, genus Natalina sensu lato (Pulmonata: Rhytididae...	Molecular Phylogenetics and Evolution	52(1):167-82	http://www.ncb...	☑	☑
29	Schultz, J., Wolf, M.	01/02/2009	ITS2 Sequence-Structure Analysis in Phylogenetics: A How-to Manual for Molecular Systematics	Molecular Phylogenetics and Evolution	52(2):520-3	http://www.ncb...	☑	☑
30	Keller, A., Schleicher, T., Schultz, J., Müll...	01/02/2009	5.8S-28S rRNA interaction and HMM-based ITS2 annotation	Gene	430:50-57	http://www.ncb...	☑	☑
31	Harrington M.G., Biffin E., Gadek P.A.	01/02/2009	Comparative study of the evolution of nuclear ribosomal spacers incorporating secondary structure anal...	Molecular Phylogenetics and Evolution	50(2):364-375	http://www.ncb...	☑	☑
32	Coleman, A.W.	01/01/2009	Is there a molecular key to the level of "biological species" in eukaryotes? A DNA guide	Molecular Phylogenetics and Evolution	50(1):197-203	http://www.ncb...	☑	☑
33	Engelmann, J.C., Rahmann, S., Wolf, M., ...	01/01/2009	Modelling cross-hybridization on phylogenetic DNA microarrays increases the detection power of closel...	Molecular Ecology Resources	9:83-93		☑	☑
34	Harrington M.	01/01/2008	Phylogeny and evolutionary history of Sapindaceae and Dodonaea	PhD thesis, James Cook University			☑	☑
35	Itskovich V., Gontcharov A., Masuda Y., ...	01/12/2008	Ribosomal ITS sequences allow resolution of freshwater sponge phylogeny with alignments guided by ...	Journal of Molecular Evolution	67(6):608-20	http://www.ncb...	☑	☑
36	Surek B.	01/10/2008	Meeting Report: Algal Culture Collections 2008. An International Meeting at the Culture Collection of Algae...	Protist	159(4):509:17	http://www.ncb...	☑	☑
37	Philipp N Seibel, Tobias Müller, Thomas D...	01/10/2008	Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4...	BMC Res Notes	1:91	http://www.ncb...	☑	☑

Figure 34: This graphic shows the ITS2 workbench administration website that allows to manage dynamic contents of the ITS2 workbench. On the left, a list of topics is available, which can be edited directly inside the grid rows in the center of the website. As example, the citations page is depicted.

On the left side, a list of manageable topics is available. A click on those, opens the corresponding table in the center. Here, rows can be added, deleted or updated easily using the grid-editor. Changed values are quickly synchronized with the database and become visible on the ITS2 workbench immediately.

7.8 Discussion

During the past ten years, ITS2 sequence-structure phylogeny has been discovered as an interesting concept that improves phylogenetic analyses (Coleman, 2003; Schultz et al., 2005; Telford et al., 2005; Müller et al., 2007; Keller et al., 2010), and was applied in several studies (more than 150 ITS2 database citations). Following the proposed workflow of Schultz and Wolf (2009), however, required the use of a large variety of different programs (Seibel et al., 2006; Wolf et al., 2008; Larkin et al., 2007; Page, 1996; Wolf et al., 2005b) in combination with the ITS2 database. Further, these tools needed to be maintained for a variety of different operating systems. The ITS2 workbench, is so far the first website, that embeds the complete phylogenetic pipeline for a sequence-structure based analysis. However, it also bears some limitations. Having created an alignment though, one might find out that some sequence parts or structural bonds need to be corrected. Such editing is currently impossible in the workbench. Here, it was decided in favour for a full automation and reproducibility. Nevertheless, one could think about implementing a few semi-automated features like the cropping of borders for frayed alignments. Another aspect is the calculation of the tree: currently, only the Neighbour Joining method is provided, which runs in a relatively short time. But even here, the number of taxa had to be limited and no bootstrap support is given. This can be disregarded when concentrating on a taxon sampling, but for more advanced tree calculation, one has to download the alignment, and run standalone treeing software like ProfDistS. However, there exists no publication about an MP or ML based sequence-structure treeing program yet. A solution to this problem could be the integration of the newly developed R-package described in Chapter 4.2. Although it is not evaluated on large scale-data yet, it provided promising results on the chlorophyceaen dataset (Chapter 4.4). By running this package with parallel bootstraps on a cluster, it might be able to provide trees for MP, ML and BIONJ even in a reasonably short time.

8 ITS2 application on large scale-data - automated reconstruction of the Green Algal Tree of Life

8.1 Indroduction

Continuing the discussion towards a large-scale analysis, one might question beside the technical hindrances, the potential of the ITS2 itself (Alvarez and Wendel, 2003; Sang, 2002) for a successful application on larger taxonomic units. This might be due to the known genetic heterogeneity of the ITS2 which resulted in a large discussion over several years (Alvarez and Wendel, 2003; Sang, 2002; Bezzhonova and Goryacheva, 2008; Wang and Yao, 2005; Nickrent et al., 1994; Powers et al., 1997; Feliner and Rossello, 2007; Wolf and Schultz, 2009) or the questioned range of this marker when applied on such big datasets. Nevertheless, the increasing number of published manuscripts using ITS/ITS2 as phylogenetic marker (Feliner and Rossello, 2007) and the proposal of ITS/ITS2 as a DNA barcode in several recent studies (Li et al., 2011; Chen et al., 2010; Yao et al., 2010; Gao et al., 2010; Sass et al., 2007) underlines its potential, especially when incorporating both of its features into one's analysis.

To further demonstrate the automated workflow of (Schultz and Wolf, 2009) on large datasets, the phylogenetic tree of Chlorophyta (green algea) with about 2270 taxa was reconstructed automatically, using techniques, equivalently implemented in the ITS2 workbench. This study, published by Buchheim et al. (2011a) included the creation of datasets, alignments, and a short technical chapter which were part of this thesis.

8.2 Materials and Methods

The analysis was based on ITS2 sequences and structures obtained from the ITS2 database v3.0 (2009/09/29). For all available Chlorophyceaen (591), Trebouxiophyceae (741) and Ulvophyceae (938) samples, a sequence structure based global multiple alignment was generated using 4SALE v1.5 and ClustalW2 with an ITS2 sequence-structure specific scoring matrix. Out of each alignment, a class-specific tree was calculated by Profile Neighbor Joining (PNJ) and ProfDistS v0.9.8 using a General Time Reversible (GTR) substitution model. Here, in each case *Micromonas* (Prasinophyceae) was added as outgroup to the data set before. Additionally, a global Chlorophyta tree containing all (2270) taxa from the class specific trees was calculated accordingly. All trees were finally rooted and visualized by FigTree v1.2.3.

For bootstrap support, manual profiles were set in ProfDistS with the aid of Cartoon2Profile (http://profdist.bioapps.biozentrum.uni-wuerzburg.de/cgi-bin/index.php?section=cart2prof) for the most important clades inside the trees. Cartoon2Profile simplifies profile definition by exporting cartoons from FigTree into a profiles definition file, compatible to ProfDistS. For online visualization, all three trees of each class were concatenated and visualized by HyperGeny, a hyperbolic tree browser and are available at:

http://hypertree.bioapps.biozentrum.uni-wuerzburg.de.

Using conventional hardware, a 2GHz computer took less than one hour for calculating each class-specific alignment and about 3.5h for the calculation of the whole Chlorophyta alignment. Determining the bootstrap support took approximately 10 minutes for each tree.

8.3 Results

The class of Chlorophyceae (Figure 37) shows the three orders of Oedogoniales, Sphaeropleales and Chlamydomonadales/Volvocales. Oedogoniales are categorized by the genera of *Oedogonium*, *Bulbochaete* and *Oedocladium*. Sphaerophleales, grouped into *Desmodesmus*, *Scenedesmus*, *Atractomorpha* and *Sphaeroplea* have a good bootstrap support of (94%) and were placed as a monophyletic unit. Chlamydomonadales/Volvocales consist of *Chlamydomonas*, *Yamagishiella*, *Pandorina*, *Eudorina*, *Astrephomene*, *Gonium*, *Phacotus* and *Dunaliella*. In this case, Chlamydomonadales were separaded into two distinct groups (Chlamydomonadales I and Chlamydomonadales II), albeit with low bootstrap support.

The class of Trebouxiophyceae (Figure 38) contains the three clades, Microthamniales I (*Trebouxia* alliance), Microthamniales II (*Asterochloris* alliance) and the Chlorellales with the genera of *Chlorella*, *Parachlorella*, *Coccomyxa*, *Micractinium* and *Didymogenes.* Each clade has a high bootstrap support of 99% (Microthamniales I), 94% (Microthamniales II) and 96% (Chlorellales).

The class of Ulvophyceae (Figure 39) is resolved with the four clades of Bryopsidales, *Urospora/Acrosiphonia*, Ulvales I and Ulvales II. Bryopsidales consist of the orders Halimeda and Caulerpa and show a high bootstrap support of 92%. The *Urospora/Acrosiphonia* clade is supported by 79%. Ulvales I consist of the taxa *Bolbocoelon*, *Blidingia*, *Monostroma*, *Umbraulva*, *Acrochaete* and *Ulva I*, whereas Ulvales II reveal a second ulvean group – Ulva II. Ulvales I and Ulvales II have low bootstrap support with the latter forming a sister group to Urospora/Acrosiphonia.

The phylum Chlorophyta (Figure 35) reveals the three classes of green algae described above, however in some aspects there are differences compared to the class-based analysis. Per default, the class-based analysis treats each class as monophyletic, the phylum-level analysis though questions this assumption slightly. Oedogoniales are grouped together with Chlorellales III (*Coccomyxa*) as sister to Ulvales I (*Urospora/Acrosiphonia*). Further, Sphaeropleales II (Sphaeropleaceae) are in alliance with Chlorellales I (*Chlorella*, *Parachlorella*, *Micractinium*, *Didymogenes*, *Diacanthos*, *Closteriopsis*, *Actinastrum*, *Dictyosphaerium*, *Auxenochlorella*, *Lobosphaeropsis*), Chlorellales II (*Pseudochlorella*, *Koliella*) and Microthamniales II. Additionally, Sphaeropleales I (*Desmodesmus*, *Scenedesmus*) are resolved as sister group to Ulvales I. Regarding all these taxa, the Chlamydomonadales is classified as a monophyletic sister group.

The Trebouxiophyceae as well as the Ulvophyceae form four non-monophyletic clades. For the Trebouxiophyceae, these are Microthammniales I, Microthamniales II, Chlorellales III and the group of Microthamniales II, Chlorellales I and Chlorellales II. For the Ulvophyceae, these are the Bryopsidales II (*Caulerpa*), the Ulvales I and the group of Ulvales, *Urospora/Acrosiphonia* and Bryopsidales I (*Halimeda*).

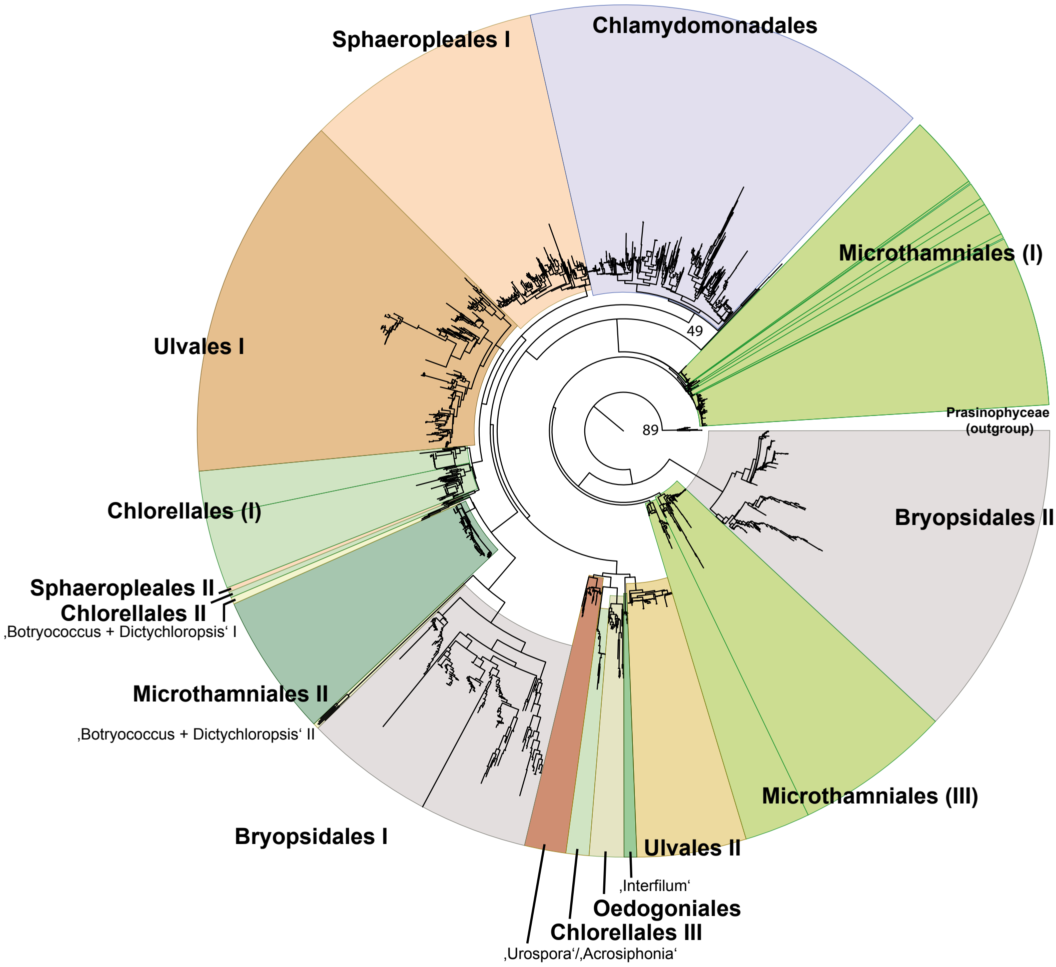


Figure 35: Profile Neighbour joining tree calculated on the phylum Chlorophyta (with 100 bootstrap replicas) using all ITS2 sequences and structures available from the ITS2 database version 3.0 (2009/09/29). The image is taken from Buchheim et al. (2011a).

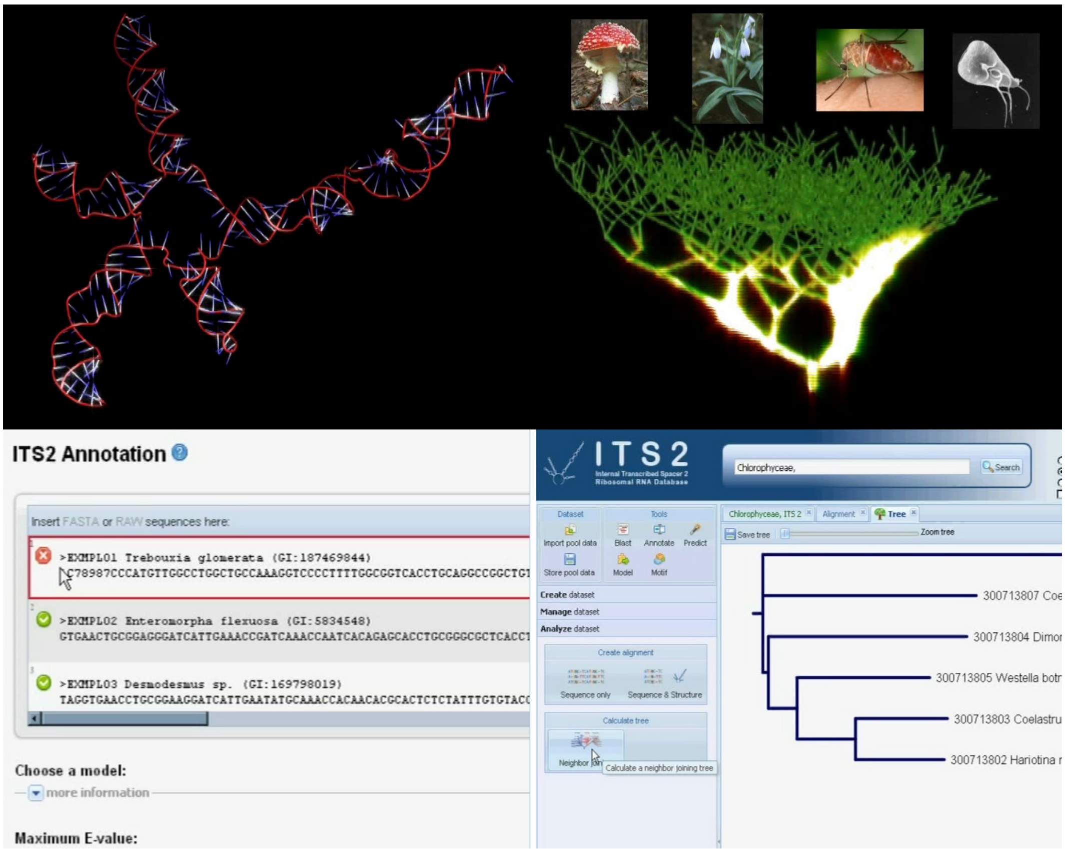
8.4 Discussion

Regarding the reconstructed trees of the classes Chlorophyceae, Trebouxiophyceae and Ulvophyceae, they are basically in accordance with several published manuscripts using a variety of different markers, for example 18S rRNA (Wolf et al., 2002; Hepperle et al., 2000; Lewis and Flechtner, 2004; Nakada et al., 2008), rbcL (Nozaki et al., 2000; Loughnane et al., 2008; Zechman, 2003; Nozaki et al., 2003) or atpB (Buchheim et al., 2010; Nozaki et al., 2000, 2003). However, some differences exist to other studies, for example when regarding the Chlorophyceae (Buchheim et al., 2001, 2002). Figure 37 places Chlamydomonadales as a basal, paraphyletic unit, which differs from both manuscripts, where Oedogoniales, Chaetophorales and / or Chaetopeltidales were reconstructed at this position. The differences might be due to a weak support in those datasets, variations in the taxon sampling – e.g. during time of analysis there existed no ITS2 data for Chaetopeltidales / Chaetophorales, or differences in rooting the outgroup. Further, discrepancies become visible when comparing the phylum-level analysis to reconstructions on class-level. Here, for example, the group of Chlamydomonadales is resolved as monophyletic in the phylum-level tree (Figure 35), but compared to the class-level analysis (Figure 37), Chlamydomonadales II forms a sister group to the groups of Chlamydomonadales I, Oedogoniales and Sphaeropleales, albeit with a low bootstrap support of 47 and 57, respectively. Beside these small irregularities, the ITS2 marker unravelled a large phylogeny from species to phylum-level in a full automated pipeline and without manual intervention. Taking into account also the short calculation period, this suggests an application of ITS2 sequence-structure analysis on similarly large datasets.

9 A video tutorial for the ITS2 workbench

9.1 Short summary

Finally, the ITS2 workbench is in use and accessible for the public. However, the inexperienced user might need some guidance on the way to the phylogenetic tree. Therefore, a short movie about the ITS2 workbench was created (Merget et al., 2012). It starts with a general introduction about the ITS2 marker, followed by a detailed explanation of the implemented pipeline. Part of this thesis was the 3D animated creation of a growing phylogenetic tree (Figure 36 top right).



```
>EXMPL01 Trebouxia glomerata (GI:187469844)
C78987CCCATGTTGGCCTGGCTGCCAAAGGTCCCCTTTTGGCGGTCACCTGCAGGCCGGCTGT
>EXMPL02 Enteromorpha flexuosa (GI:5834548)
GTGAACTGCGGAGGGATCATTGAAACCGATCAAACCAATCACAGAGCACCTGCGGGCGCTCACCT
>EXMPL03 Desmodesmus sp. (GI:169798019)
TAGGTGAACCTGCGGAAGGATCATTGAATATGCAAACCACAACACGCACTCTCTATTTGTGTACC
```

Figure 36: This illustration shows four different captures from the ITS2 workbench movie. On the top left, a 3D reconstruction of the ITS2 is visible. On the top right, a growing phylogenetic tree is depicted. The illustrations on the bottom show the website of the workbench with the ITS2 annotation tool on the left, and a calculated tree on the right.

The small movie is available at:

http://www.jove.com/video/3806/the-its2-database.

10 Conclusion and outlook

During the last years, the development of sequencing technologies was pushed towards the 4th generation and high-throughput sequencing became available. This progress guarantees the exponential growth of databases like GenBank for the years to come. However, with the raising number of sequences, the need for exact annotation, sorting and management of data becomes fundamental. Large databases like GenBank, EMBL or DDBJ are not capable of providing this service on such immense datasets accurately. Thus, the development of topic-related sub-databases is obvious. When focussing on sequence phylogeny, different databases exist for a variety of markers (Pruesse et al., 2007; Cole et al., 2009; Wang et al., 2009; Koetschan et al., 2010). However, the ITS2 workbench is - to our knowledge - the only database for ITS2 sequence-structure phylogeny. Its data storage and accuracy was significantly improved during this study by the inclusion of HMM sequences annotation and an exhaustive secondary structure prediction pipeline which more than doubled the number of structure predictions. But this is not even half of what the workbench can provide. Whereas at the beginning of this work, a large variety of tools were needed for a full sequence-structure analysis, today all these methods are integrated into the ITS2 workbench. This has the major advantage that time-consuming installations, data import and export along with simple handling errors could be reduced to a clear pipeline easily followed just with a few mouse-clicks on the Web. Therefore, a large number of different Web-services had to be implemented providing data for the AJAX driven Web 2.0 interface. Although the workbench currently bears some limitations, such as its repertoire of treeing methods, solutions like the new treeforge R-package are nearly ready for use. Beside a transfer of the currently established sequence-structure phylogeny to methods like MP and ML, it consists of newly developed sequence-structure alphabets, that performed reasonably well during first evaluations. However, there still remains the challenge in computing large-scale phylogenetic trees on bigger datasets online. Recently developments towards cloud computing on large clusters and computing farms could provide solutions here. In addition, Web design is pushing towards the further development of JavaScript, like Google's DART, local file storage mechanisms like the W3C draft of the Localstorage API (Hickson, 2011), or new browser databases such as the W3C draft of IndexDB (Mehta et al., 2011). C/C++ to JavaScript cross-compilers like Emscripten even today bring C/C++ programs to the web, just by running JavaScript. This suggests that in future, also the C++ algorithms of ClustalW2 or treeing software could be compiled for running in the Web browser using local CPU and memory resources. With the capability of calculating larger datasets, and further optimizations of the phylogenetic pipeline, in future, on might obtain a broader coverage in phylogenetic analyses.

Reconsidering the initial questions "What is a species?", "Are same looking individuals also belonging to the same species?", "Which one evolved first?" we might ask whether these are answered now? One has to admit, not completely – or the critic would say, not really. But one thing has drastically changed, and this work has definitely pushed the level one step further into this direction. Whereas just a few years ago, researchers had to undertake tedious manual work, analyse species with their morphological features in thousands of hours, to just gain an idea on how few fellows are related, today, with the use of the ITS2 workbench, this cumbersome work has been reduced to a minimum of effort and just requires few mouse-clicks.

11 References

S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. ISSN 00222836. doi: 10.1006/jmbi.1990.9999. URL http://www.ncbi.nlm.nih.gov/pubmed/2231712.

S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146917&tool=pmcentrez&rendertype=abstract.

I Alvarez and J F Wendel. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution*, 29(3):417–434, 2003. URL http://linkinghub.elsevier.com/retrieve/pii/S1055790303002082.

Jonas J Astrin, Bernhard A Huber, Bernhard Misof, and Cornelya F C Klutsch. Molecular taxonomy in pholcid spiders (Pholcidae, Araneae): evaluation of species identification methods using CO1 and 16S rRNA. *Zoologica Scripta*, 35(5):441–457, 2006. ISSN 03003256. doi: 10.1111/j.1463-6409.2006.00239.x. URL http://www.blackwell-synergy.com/doi/abs/10.1111/j.1463-6409.2006.00239.x.

P G Bagos, T D Liakopoulos, I C Spyropoulos, and S J Hamodrakas. A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC Bioinformatics*, 5:29, 2004. ISSN 14712105. doi: 10.1186/1471-2105-5-291471-2105-5-29. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15070403.

Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Re-*

search, 37(Web Server issue):W202–W208, 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19458158.

B G Baldwin, M J Sanderson, J M Porter, M F Wojciechowski, C S Campbell, and M J Donoghue. The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden*, 82(2):247–277, 1995. ISSN 00266493. doi: 10.2307/2399880. URL http://www.jstor.org/stable/2399880?origin=crossref.

Markus Bauer, Gunnar W Klau, and Knut Reinert. Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinformatics*, 8(1471-2105 (Electronic) LA - ENG PT - JOURNAL ARTICLE):271, 2007. URL http://dx.doi.org/10.1186/1471-2105-8-271.

L E Baum, T Petrie, G Soules, and N Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970. ISSN 00034851. doi: 10.1214/aoms/1177697196. URL http://www.jstor.org/stable/2239727.

Dennis a Benson, Ilene Karsch-Mizrachi, Karen Clark, David J Lipman, James Ostell, and Eric W Sayers. GenBank. *Nucleic acids research*, 40(December 2011):48–53, December 2011. ISSN 1362-4962. doi: 10.1093/nar/gkr1202. URL http://www.ncbi.nlm.nih.gov/pubmed/22144687.

O V Bezzhonova and I I Goryacheva. Intragenomic heterogeneity of rDNA internal transcribed spacer 2 in Anopheles messeae (Diptera: Culicidae). *Journal of Medical Entomology*, 45(3):337–341, 2008. URL http://www.ingentaconnect.com/content/esa/jme/2008/00000045/00000003/art00002.

J A Bilmes. What HMMs can't do. In *Invited paper and lecture ATR Workshop*. Citeseer, 2004. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.66.6329&amp;rep=rep1&amp;type=pdf.

Jean-Simon Brouard, Christian Otis, Claude Lemieux, and Monique Turmel. The exceptionally large chloroplast genome of the green alga Floydiella terrestris illuminates the evolutionary history of the Chlorophyceae. *Genome biology and evolution*, 2:240–256, 2010. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2997540&tool=pmcentrez&rendertype=abstract.

Mark A Buchheim, Eugenia A Michalopulos, and Julie A Buchheim. PHYLOGENY OF THE CHLOROPHYCEAE WITH SPECIAL REFERENCE TO THE SPHAEROPLEALES: A STUDY OF 18S AND 26S rDNA DATA. *Journal of Phycology*, 37(5):819–835, 2001. ISSN 00223646. doi: 10.1046/j.1529-8817.2001.00162.x. URL http://doi.wiley.com/10.1046/j.1529-8817.2001.00162.x.

Mark A Buchheim, Julie A Buchheim, Tracy Carlson, and Paul Kugrens. Phylogeny of Lobocharacium (Chlorophyceae) and allies: A study of 18S and 26S rDNA data. *Journal of Phycology*, 38(2):376–383, 2002. ISSN 00223646. URL http://www.refdoc.fr/Detailnotice?idarticle=9623255.

Mark A Buchheim, Andrea E Kirkwood, Julie A Buchheim, Bindhu Verghese, and William J Henley. Hypersaline Soil Supports a Diverse Community of Dunaliella (Chlorophyceae)1. *Journal of Phycology*, 46(5):1038–1047, 2010. ISSN 00223646. doi: 10.1111/j.1529-8817.2010.00886.x. URL http://doi.wiley.com/10.1111/j.1529-8817.2010.00886.x.

Mark A Buchheim, Alexander Keller, Christian Koetschan, Frank Förster, Benjamin Merget, and Matthias Wolf. Internal Transcribed Spacer 2 (nu ITS2 rRNA) Sequence-Structure Phylogenetics: Towards an Automated Reconstruction of the Green Algal Tree of Life. *PLoS ONE*, 6(2):10, 2011a. URL http://dx.plos.org/10.1371/journal.pone.0016931.

Mark A Buchheim, Danica M Sutherland, Tina Schleicher, Frank Förster, and Matthias Wolf. Phylogeny of Oedogoniales, Chaetophorales and Chaetopeltidales (Chlorophyceae): inferences from sequence-structure analysis of ITS2. *Annals of Botany*, 2011b. ISSN 10958290. doi: 10.1093/aob/mcr275. URL http://www.ncbi.nlm.nih.gov/pubmed/22028463.

Joseph H Camin and Robert R Sokal. A Method for Deducing Branching Sequences in Phylogeny. *Evolution*, 19(3):311, 1965. ISSN 00143820. doi: 10.2307/2406441. URL http://www.jstor.org/stable/2406441?origin=crossref.

Shilin Chen, Hui Yao, Jianping Han, Chang Liu, Jingyuan Song, Linchun Shi, Yingjie Zhu, Xinye Ma, Ting Gao, Xiaohui Pang, Kun Luo, Ying Li, Xiwen Li, Xiaocheng Jia, Yulin Lin, and Christine Leon. Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species. *PLoS ONE*, 5(1):8, 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20062805.

J R Cole, Q Wang, E Cardenas, J Fish, B Chai, R J Farris, A S Kulam-Syed-Mohideen, D M McGarrell, T Marsh, G M Garrity, and J M Tiedje. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, 37(Database issue):D141–D145, 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19004872.

A Coleman. ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends in Genetics*, 19(7):370–375, 2003. ISSN 01689525. doi: 10.1016/S0168-9525(03)00118-5. URL http://linkinghub.elsevier.com/retrieve/pii/S0168952503001185.

Charles Darwin. *On the Origin of Species*, volume 5. John Murray, 1859. ISBN 1551113376. doi: 10.1038/005318a0. URL http://www.ias.ac.in/resonance/February2009/p204-208.pdf.

M O Dayhoff, R M Schwartz, and B C Orcutt. A model of evolutionary change in proteins. *Atlas of protein sequence and structure*, 5(Suppl 3):345–352, 1978. doi: 10.1.1.145.4315. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.145.4315.

R Durbin, S Eddy, A Krogh, and G Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998. ISBN 0521629713. URL http://eisc.univalle.edu.co/cursos/web/material/750068/1/6368030-Durbin-Et-Al-Biological-Sequence-Analysis-CUP-2002-No-OCR.pdf.

S R Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

Sean R Eddy. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput Biol*, 4(5):e1000069, 2008. doi: 10.1371/journal.pcbi.1000069. URL http://dx.doi.org/10.1371/journal.pcbi.1000069.

Robert C Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1):113, 2004a. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=517706&tool=pmcentrez&rendertype=abstract.

Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004b. URL http://www.ncbi.nlm.nih.gov/pubmed/15034147.

Scott Federhen. The NCBI Taxonomy database. *Nucleic Acids Research*, pages 1–8, 2011. ISSN 13624962. doi: 10.1093/nar/gkr1178. URL http://www.ncbi.nlm.nih.gov/pubmed/22139910.

Gonzalo Nieto Feliner and Josep A Rossello. Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. *Molecular Phylogenetics and Evolution*, 44(2): 911–919, 2007. URL http://www.sciencedirect.com/science/article/B6WNH-4N2D2TM-2/2/b61d728843cc71cc3916f51e9c7562df.

J Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981. URL http://www.ncbi.nlm.nih.gov/pubmed/7288891.

J Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4):783–791, 1985. ISSN 00143820. doi: 10.2307/2408678. URL http://www.jstor.org/stable/2408678.

J Felsenstein. *Inferring Phylogenies*, volume 266. Sinauer Associates, 2004. ISBN 0878931775. URL http://www.ncbi.nlm.nih.gov/pubmed/11675604.

J Felsenstein, J Archie, W H E Day, W Maddinson, C Meacham, F J Rohlf, and D Swofford. The Newick tree format, 1986. URL http://evolution.genetics.washington.edu/phylip/newicktree.html.

Roy Thomas Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000. URL http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm.

Robert D Finn, Jody Clements, and Sean R Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(Web Server issue):W29–W37, 2011. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125773&tool=pmcentrez&rendertype=abstract.

G D Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973. ISSN 00189219. doi: 10.1109/PROC.1973.9030. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1450960.

Torben Friedrich. *New statistical Methods of Genome-Scale Data Analysis in Life Science - Applications to enterobacterial Diagnostics, Meta-Analysis of Arabidopsis thaliana Gene Expression and functional Sequence Annotation*. PhD thesis, University of Würzburg, 2009.

Torben Friedrich, Birgit Pils, Thomas Dandekar, Jörg Schultz, and Tobias Müller. Modelling interaction sites in protein domains with interaction profile hidden Markov models. *Bioinformatics*, 22(23):2851–2857, 2006. URL http://www.ncbi.nlm.nih.gov/pubmed/17000753.

Torben Friedrich, Christian Koetschan, and Tobias Müller. Optimisation of HMM topologies enhances DNA and protein sequence modelling. *Stat Appl Genet Mol Biol*, 9(1):Article 6, 2010. doi: 10.2202/1544-6115.1480. URL http://dx.doi.org/10.2202/1544-6115.1480.

Ting Gao, Hui Yao, Jingyuan Song, Chang Liu, Yingjie Zhu, Xinye Ma, Xiaohui Pang, Hongxi Xu, and Shilin Chen. Identification of medicinal plants in the family Fabaceae using a potential DNA barcode ITS2. *Journal of Ethnopharmacology*, 130(1):116–121, 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20435122.

O Gascuel. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7):685–95, 1997. ISSN 07374038. URL http://www.ncbi.nlm.nih.gov/pubmed/9254330.

S Henikoff and J G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915–10919, 1992. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=50453&tool=pmcentrez&rendertype=abstract.

Dominik Hepperle, Eberhard Hegewald, and L Krienitz. PHYLOGENETIC POSITION OF THE OOCYSTACEAE (CHLOROPHYTA) 1 The complete 18S rRNA gene sequences of three Oocystis A . Braun species (Oocystaceae) and three other chlorococcal algae , Tetrachlorella alternans (G . M . Smith) Kor ˇ Okada (Scenedesmacea. *Direct*, 595(3):590–595, 2000. URL http://onlinelibrary.wiley.com/doi/10.1046/j.1529-8817.2000.99184.x/full.

Ian Hickson. Web Storage, 2011. URL http://dev.w3.org/html5/webstorage/.

Masami Ikeda, Masafumi Arai, Toshikatsu Okuno, and Toshio Shimizu. TM-PDB: a database of experimentally-characterized transmembrane topologies. *Nucleic Acids Research*, 31(1):406–409, 2003. URL http://www.nar.oupjournals.org/cgi/doi/10.1093/nar/gkg018.

Norman L Johnson, Chapel Hill, and North Carolina. Univariate Discrete Distributions Univariate Discrete Distributions (3rd ed.), by Norman L. Johnson , Adrienne W. Kemp , and Samuel Kotz , Hoboken, NJ : Wiley , 2005 , ISBN 0-471-27246-9 , xix + 646 pp., 125.00 . *Technometrics*, 48(3): 450–450, 2006. ISSN 00401706. doi: 10.1198/tech.2006.s421. URL http://pubs.amstat.org/doi/abs/10.1198/tech.2006.s421.

B H Juang and L R Rabiner. Hidden Markov Models for Speech Recognition. *Technometrics*, 33(3):251, 1991. ISSN 00401706. doi: 10.2307/1268779. URL http://www.jstor.org/stable/1268779?origin=crossref.

T H Jukes and C R Cantor. Evolution of protein molecules. In H N Munro, editor, *Mammalian Protein Metabolism*, volume 3 of *Mammalian protein metabolism*, chapter 24, pages 21–132. Academic Press, 1969. URL http://www.citeulike.org/group/1390/article/768582.

Agnieszka S Juncker, Hanni Willenbrock, Gunnar Von Heijne, Sø ren Brunak, Henrik Nielsen, and Anders Krogh. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Science*, 12(8):1652–1662, 2003. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2323952&tool=pmcentrez&rendertype=abstract.

Robel Y Kahsay, Guang Gao, and Li Liao. An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. *Bioinformatics*, 21(9):1853–1858, 2005. URL http://www.ncbi.nlm.nih.gov/pubmed/15691854.

Lukas Käll, Anders Krogh, and Erik L L Sonnhammer. A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, 338(5):1027–1036, 2004. URL http://www.ncbi.nlm.nih.gov/pubmed/15111065.

Alexander Keller, Tina Schleicher, Frank Förster, Benjamin Ruderisch, Thomas Dandekar, Tobias Müller, and Matthias Wolf. ITS2 data corroborate a monophyletic chlorophycean DO-group (Sphaeropleales). *BMC Evolutionary Biology*, 8:218, 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18655698.

Alexander Keller, Tina Schleicher, Jörg Schultz, Tobias Müller, Thomas Dandekar, and Matthias Wolf. 5.8S-28S rRNA interaction and HMM-based ITS2 annotation. *Gene*, 430(1-2):50–7, 2009. ISSN 18790038. doi: 10.1016/j.gene.2008.10.012. URL http://www.ncbi.nlm.nih.gov/pubmed/19026726.

Alexander Keller, Frank Förster, Tobias Müller, Thomas Dandekar, Jörg Schultz, and Matthias Wolf. Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees. *Biology Direct*, 5(1):4, 2010. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2821295&tool=pmcentrez&rendertype=abstract.

Marcelo V Kitahara, Stephen D Cairns, Jarosław Stolarski, David Blair, and David J Miller. A Comprehensive Phylogenetic Analysis of the Scleractinia (Cnidaria, Anthozoa) Based on Mitochondrial CO1 Sequence Data. *PLoS ONE*, 5(7):9, 2010. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2900217&tool=pmcentrez&rendertype=abstract.

Christian Koetschan. *Enhanced length modelling methods for Hidden Markov Models*. PhD thesis, University of Würzburg, 2008.

Christian Koetschan, Frank Förster, Alexander Keller, Tina Schleicher, Benjamin Ruderisch, Roland Schwarz, Tobias Müller, Matthias Wolf, and

Jörg Schultz. The ITS2 Database III—sequences and structures for phylogeny. *Nucleic Acids Research*, 38(Database issue):D275–D279, 2010. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2808966&tool=pmcentrez&rendertype=abstract.

Christian Koetschan, Thomas Hackl, Tobias Müller, Matthias Wolf, Frank Förster, and Jörg Schultz. ITS2 database IV: Interactive taxon sampling for internal transcribed spacer 2 based phylogenies. *Molecular phylogenetics and evolution*, February 2012. ISSN 1095-9513. doi: 10.1016/j.ympev.2012.01.026. URL http://www.ncbi.nlm.nih.gov/pubmed/22366368.

A Krogh, B Larsson, G Von Heijne, and E L Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, 305(3):567–580, 2001. URL http://www.ncbi.nlm.nih.gov/pubmed/11152613.

Tamara Kulikova, Ruth Akhtar, Philippe Aldebert, Nicola Althorpe, Mikael Andersson, Alastair Baldwin, Kirsty Bates, Sumit Bhattacharyya, Lawrence Bower, Paul Browne, Matias Castro, Guy Cochrane, Karyn Duggan, Ruth Eberhardt, Nadeem Faruque, Gemma Hoad, Carola Kanz, Charles Lee, Rasko Leinonen, Quan Lin, Vincent Lombard, Rodrigo Lopez, Dariusz Lorenc, Hamish McWilliam, Gaurab Mukherjee, Francesco Nardone, Maria Pilar Garcia Pastor, Sheila Plaister, Siamak Sobhany, Peter Stoehr, Robert Vaughan, Dan Wu, Weimin Zhu, and Rolf Apweiler. EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Research*, 35 (Database issue):D16–20, 2007. ISSN 13624962. doi: 10.1093/nar/gkl913. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1897316&tool=pmcentrez&rendertype=abstract.

M A Larkin, G Blackshields, N P Brown, R Chenna, P A McGettigan, H McWilliam, F Valentin, I M Wallace, A Wilm, R Lopez, J D Thompson, T J Gibson, and D G Higgins. Clustal W and Clustal X version 2.0.

Bioinformatics, 23(21):2947–2948, 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17846036.

Honglak Lee and Andrew Y Ng. Spam Deobfuscation using a Hidden Markov Model. *English*, 2005. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.79.5218&amp;rep=rep1&amp;type=pdf.

Louise A Lewis and Valerie R Flechtner. Cryptic Species of Scenedesmus (Chlorophyta) From Desert Soil Communities of Western North America. *Journal of Phycology*, 40(6):1127–1137, 2004. ISSN 00223646. doi: 10.1111/j.1529-8817.2004.03235.x. URL http://doi.wiley.com/10.1111/j.1529-8817.2004.03235.x.

De-Zhu Li, Lian-Ming Gao, Hong-Tao Li, Hong Wang, Xue-Jun Ge, Jian-Quan Liu, Zhi-Duan Chen, Shi-Liang Zhou, Shi-Lin Chen, Jun-Bo Yang, Cheng-Xin Fu, Chun-Xia Zeng, Hai-Fei Yan, Ying-Jie Zhu, Yong-Shuai Sun, Si-Yun Chen, Lei Zhao, Kun Wang, Tuo Yang, and Guang-Wen Duan. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences of the United States of America*, pages 1104551108–, 2011. ISSN 10916490. doi: 10.1073/pnas.1104551108. URL http://www.pnas.org/cgi/content/abstract/1104551108v1.

Qi Liu, Yi-Sheng Zhu, Bao-Hua Wang, and Yi-Xue Li. A HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins. *Computational Biology and Chemistry*, 27(1):69–76, 2003. URL http://www.ncbi.nlm.nih.gov/pubmed/12798041.

Ciar N J Loughnane, Lynne M McIvor, Fabio Rindi, Dagmar B Stengel, and Michael D Guiry. Morphology, rbcL phylogeny and distribution of distromatic Ulva (Ulvophyceae, Chlorophyta) in Ireland and southern Britain. *Phycologia*, 47(4):416–429, 2008. doi: 10.2216/07-61.1. URL http://www.phycologia.org/perlserv/?request=get-abstract&amp;doi=10.2216/PH07-61.1.

A V Lukashin and M Borodovsky. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*, 26(4):1107–1115, February 1998.

J C Mai and A W Coleman. The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants. *Journal of Molecular Evolution*, 44(3):258–271, 1997. URL http://www.ncbi.nlm.nih.gov/pubmed/9060392.

Marco Mangone, Philip MacMenamin, Charles Zegar, Fabio Piano, and Kristin C Gunsalus. UTRome.org: a platform for 3'UTR biology in C. elegans. *Nucleic Acids Research*, 36(Database issue):D57–D62, 2008. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2238901&tool=pmcentrez&rendertype=abstract.

S M Markert, T Müller, C Koetschan, T Friedl, and M Wolf. 'Y' Scenedesmus (Chlorophyta, Chlorophyceae): the internal transcribed spacer 2 rRNA secondary structure re-revisited. *Plant biology*, 2012. doi: 10.1111/j.1438-8677.2012.00576.x.

Nicholas R Markham and Michael Zuker. UNAFold: software for nucleic acid folding and hybridization. *Methods In Molecular Biology Clifton Nj*, 453(1):3–31, 2008. URL http://www.springerlink.com/index/10.1007/978-1-60327-429-6.

Pier Luigi Martelli, Piero Fariselli, Anders Krogh, and Rita Casadio. A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics*, 18 Suppl 1(NIL):S46–S53, 2002. URL http://www.ncbi.nlm.nih.gov/pubmed/12169530.

Ernst Mayr. *Systematics and the origin of species from the viewpoint of a zoologist*, volume 13 of *Columbia biological series ... No. XIII.* Columbia University Press, 1942. ISBN 0674862503. URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&amp;path=ASIN/0674862503.

Nikunj Mehta, Jonas Sicking, Eliot Graff, Andrei Popescu, and Orlow Jeremy. Indexed Database API, 2011. URL http://www.w3.org/TR/IndexedDB/.

Benjamin Merget, Christian Koetschan, Thomas Hackl, Frank Förster, Thomas Dandekar, Tobias Müller, Jörg Schultz, and Matthias Wolf. The ITS2 Database. *Journal of Visualized Experiments*, 61:e3806, 2012. doi: 10.3791/3806. URL http://www.jove.com/video/3806/the-its2-database.

Mónica B J Moniz and Irena Kaczmarska. Barcoding diatoms: Is there a good marker? *Molecular ecology resources*, 9 Suppl s1(s1):65–74, 2009. doi: 10.1111/j.1755-0998.2009.02633.x. URL http://www.ncbi.nlm.nih.gov/pubmed/21564966.

Mónica B J Moniz and Irena Kaczmarska. Barcoding of diatoms: nuclear encoded ITS revisited. *Protist*, 161(1):7–34, 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/19674931.

T Müller and M Vingron. Modeling amino acid replacement. *Journal of computational biology a journal of computational molecular cell biology*, 7(6): 761–776, 2000. URL http://www.ncbi.nlm.nih.gov/pubmed/11382360.

Tobias Müller. *Modellierung von Proteinevolution*. PhD thesis, 2001.

Tobias Müller, Rainer Spang, and Martin Vingron. Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Molecular Biology and Evolution*, 19(1):8–13, 2002. URL http://www.ncbi.nlm.nih.gov/pubmed/11752185.

Tobias Müller, Nicole Philippi, Thomas Dandekar, Jörg Schultz, and Matthias Wolf. Distinguishing species. *Rna New York Ny*, 13(9):1469–1472, 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17652131.

Ahmed Ragab Nabhan and Indra Neil Sarkar. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in Bioinformatics*, pages bbr014–, 2011. ISSN 14774054. doi:

10.1093/bib/bbr014. URL http://bib.oxfordjournals.org/content/early/2011/03/23/bib.bbr014.abstract.

Takashi Nakada, Kazuharu Misawa, and Hisayoshi Nozaki. Molecular systematics of Volvocales (Chlorophyceae, Chlorophyta) based on exhaustive 18S rRNA phylogenetic analyses. *Molecular Phylogenetics and Evolution*, 48 (1):281–291, 2008. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WNH-4S3G3X7-1&_user=10&_rdoc=1&_fmt=&_orig=search&_sort=d&_docanchor=&view=c&_searchStrId=1081008629&_rerunOrigin=scholar.google&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=e16db5ad61e3a7dfcb93791912e903fb.

Ncbi. FASTA format description, 2007. URL http://www.ncbi.nlm.nih.gov/blast/fasta.shtml.

Daniel L Nickrent, Kevin P Schuette, and Ellen M Starr. A molecular phylogeny of IArceuthobium (Viscaceae) based on nuclear ribosomal DNA internal transcribed spacer sequences. *American Journal of Botany*, 81 IS - (9):1149–1160, 1994. URL http://www.jstor.org/stable/2445477.

H Nozaki, K Misawa, T Kajita, M Kato, S Nohara, and M M Watanabe. Origin and evolution of the colonial volvocales (Chlorophyceae) as inferred from multiple, chloroplast gene sequences. *Molecular phylogenetics and evolution*, 17(2):256–68, November 2000. ISSN 1055-7903. doi: 10.1006/mpev.2000.0831. URL http://www.ncbi.nlm.nih.gov/pubmed/11083939.

Hisayoshi Nozaki, Osami Misumi, and Tsuneyoshi Kuroiwa. Phylogeny of the quadriflagellate Volvocales (Chlorophyceae) based on chloroplast multigene sequences. *Molecular Phylogenetics and Evolution*, 29(1):58–66, 2003. URL http://linkinghub.elsevier.com/retrieve/pii/S1055790303000897.

R D M Page. TreeView: An application to display phylogenetic trees on personal computers. *Computer applications in the biosciences CABIOS*, 12 (4):357–358, 1996. URL http://eprints.gla.ac.uk/394/.

Karl Pearson. on the Systematic Fitting of Curves To Observations and Measurements. *Biometrika*, 1(3):265–303, 1902. ISSN 00063444. doi: 10.2307/2331540. URL `http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/1.3.265`.

T O Powers, T C Todd, A M Burnell, P C B Murray, C C Fleming, A L Szalanski, B A Adams, and T S Harris. The rDNA Internal Transcribed Spacer Region as a Taxonomic Marker for Nematodes. *Journal of Nematology*, 29(4):441–450, 1997. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2619808&tool=pmcentrez&rendertype=abstract`.

Elmar Pruesse, Christian Quast, Katrin Knittel, Bernhard M Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank Oliver Glöckner. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35(21):7188–7196, 2007. URL `http://www.ncbi.nlm.nih.gov/pubmed/17947321`.

L R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. 77(2):257–286, 1989. doi: 10.1109/5.18626.

Lorenzo Rossi, Jacob Chakareski, Pascal Frossard, and Stefania Colonnese. Proceedings of 2010 IEEE 17th International Conference on Image Processing A NON-STATIONARY HIDDEN MARKOV MODEL OF MULTIVIEW VIDEO TRAFFIC. *Signal Processing*, pages 2921–2924, 2010.

N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987. URL `http://www.ncbi.nlm.nih.gov/pubmed/18343690`.

Tao Sang. Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical Reviews in Biochemistry and Molecular Biology*, 37(3):121–147, 2002. URL `http://www.ncbi.nlm.nih.gov/pubmed/12139440`.

Chodon Sass, Damon P Little, Dennis Wm Stevenson, and Chelsea D Specht. DNA Barcoding in the Cycadales: Testing the Potential of Proposed Barcoding Markers for Species Identification of Cycads. *PLoS ONE*, 2(11):9, 2007. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2063462&tool=pmcentrez&rendertype=abstract.

Klaus Peter Schliep. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21169378.

Gisbert Schneider and Uli Fechner. Advances in the prediction of protein targeting signals. *Proteomics*, 4(6):1571–1580, 2004. URL http://www.ncbi.nlm.nih.gov/pubmed/15174127.

Jörg Schultz and Matthias Wolf. ITS2 Sequence-Structure Analysis in Phylogenetics: A How-to Manual for Molecular Systematics. *Molecular Phylogenetics and Evolution*, 52(2):520–523, 2009. ISSN 10959513. doi: 10.1016/j.ympev.2009.01.008. URL http://www.ncbi.nlm.nih.gov/pubmed/19640446.

Jörg Schultz, Stefanie Maisel, Daniel Gerlach, Tobias Müller, and Matthias Wolf. A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. *Rna New York Ny*, 11(4):361–364, 2005. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1370725&tool=pmcentrez&rendertype=abstract.

Jörg Schultz, Tobias Müller, Marco Achtziger, Philipp N Seibel, Thomas Dandekar, and Matthias Wolf. The internal transcribed spacer 2 database—a web server for (not only) low level phylogenetic analyses. *Nucleic Acids Research*, 34(Web Server issue):W704–W707, 2006. URL http://dx.doi.org/10.1093/nar/gkl129.

G Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978. ISSN 00905364. doi: 10.2307/2958889. URL http://www.jstor.org/stable/2958889.

Philipp N Seibel, Tobias Müller, Thomas Dandekar, Jörg Schultz, and Matthias Wolf. 4SALE – A tool for synchronous RNA sequence and secondary structure alignment and editing. *BMC Bioinformatics*, 7(1):498, 2006. URL http://www.ncbi.nlm.nih.gov/pubmed/17101042.

Philipp N Seibel, Tobias Müller, Thomas Dandekar, and Matthias Wolf. Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4SALE. *BMC research notes*, 1:91, 2008. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2587473&tool=pmcentrez&rendertype=abstract.

Christian Selig, Matthias Wolf, Tobias Müller, Thomas Dandekar, and Jörg Schultz. The ITS2 Database II: homology modelling RNA structure for molecular systematics. *Nucleic Acids Research*, 36(Database issue):D377–D380, 2008. URL http://dx.doi.org/10.1093/nar/gkm827.

Sven Siebert and Rolf Backofen. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, 21(16):3352–3359, 2005. URL http://www.ncbi.nlm.nih.gov/pubmed/15972285.

Andrew D Smith, Thomas W H Lui, and Elisabeth R M Tillier. Empirical models for substitution in ribosomal RNA. *Molecular Biology and Evolution*, 21(3):419–427, 2004. URL http://dx.doi.org/10.1093/molbev/msh029.

M Alex Smith, Nikolai A Poyarkov, and Paul D N Hebert. DNA BARCODING: CO1 DNA barcoding amphibians: take the chance, meet the challenge. *Molecular ecology resources*, 8(2):235–246, 2008. doi: 10.1111/j.1471-8286.2007.01964.x. URL http://www.blackwell-synergy.com/doi/abs/10.1111/j.1471-8286.2007.01964.x.

E L Sonnhammer, G Von Heijne, and A Krogh. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings International Conference on Intelligent Systems for Molecular Biology ISMB Inter-*

national Conference on Intelligent Systems for Molecular Biology, 6(NIL): 175–182, 1998. URL http://www.ncbi.nlm.nih.gov/pubmed/9783223.

J E Stajich, D Block, K Boulez, S E Brenner, S A Chervitz, C Dagdigian, G Fuellen, J G Gilbert, I Korf, H Lapp, H Lehvaslaiho, C Matsalla, C J Mungall, B I Osborne, M R Pocock, P Schattner, M Senger, L D Stein, E Stupka, M D Wilkinson, and E Birney. The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, 12(10):1611–1618, 2002. ISSN 10889051. doi: 10.1101/gr.361602.1. URL http://dx.doi.org/10.1101/gr.361602.

Y B Suh, L B Thien, H E Reeve, and E A D A S E P Zimmer. MOLECULAR EVOLUTION AND PHYLOGENETIC IMPLICATIONS OF INTERNAL TRANSCRIBED SPACER SEQUENCES OF RIBOSOMAL DNA IN WINTERACEAE. *American Journal of Botany*, 80(9):1042–1055 ST – MOLECULAR EVOLUTION AND PHYLOGENET, 1993. ISSN 00029122. URL http://www.jstor.org/stable/2445752.

Y Tateno, T Imanishi, S Miyazaki, K Fukami-Kobayashi, N Saitou, H Sugawara, and T Gojobori. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Research*, 30(1):27–30, 2002. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=99140&tool=pmcentrez&rendertype=abstract.

S Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17(17):57–86, 1986. URL http://books.google.com/books?hl=en&amp;lr=&amp;id=8aI1phhOKhgC&amp;oi=fnd&amp;pg=PA57&amp;dq=Some+probabilistic+and+statistical+problems+in+the+analysis+of+DNA+sequences&amp;ots=rlLEaJIcPl&amp;sig=ScWJwRxj6YEd85wxsVoMSq7XNY8.

Maximilian J Telford, Michael J Wise, and Vivek Gowri-Shankar. Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: examples from the bilateria. *Molecular Biology* *and Evolution*, 22(4):1129–1136, 2005. URL http://discovery.ucl.ac.uk/10796/.

JD Thompson, DG Higgins, and TJ Gibson. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, posistion-specific gap penalites and weight matrix choice. *Nucleic Acids Research*, 22:4673*4680, 1994.

Monique Turmel, Jean-Simon Brouard, Cédric Gagnon, Christian Otis, and Claude Lemieux. Deep Division in the Chlorophyceae (Chlorophyta) Revealed By Chloroplast Phylogenomic Analyses. *Journal of Phycology*, 44(3):739–750, 2008. ISSN 00223646. doi: 10.1111/j.1529-8817.2008.00510.x. URL http://blackwell-synergy.com/doi/abs/10.1111/j.1529-8817.2008.00510.x.

G E Tusnády and I Simon. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *Journal of Molecular Biology*, 283(2):489–506, 1998. URL http://www.ncbi.nlm.nih.gov/pubmed/9769220.

A Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. 13(2):260–269, 1967. doi: 10.1109/TIT.1967.1054010.

D-M Wang and Y-J Yao. Intrastrain internal transcribed spacer heterogeneity in Ganoderma species. *Canadian Journal of Microbiology*, 51(2):113–121, 2005. URL http://www.ncbi.nlm.nih.gov/pubmed/16091769.

Norman Wang, Alison R Sherwood, Akira Kurihara, Kimberly Y Conklin, Thomas Sauvage, and Gernot G Presting. The Hawaiian Algal Database: a laboratory LIMS and online resource for biodiversity data. *BMC Plant Biology*, 9(1):117, 2009. URL http://www.biomedcentral.com/1471-2229/9/117.

J D Watson and F H Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953. URL http://www.ncbi.nlm.nih.gov/pubmed/17804965.

Lloyd R Welch. Hidden Markov Models and the Baum-Welch Algorithm. *IEEE Information Theory Society Newsletter*, 53(4):1,10–13, 2003. URL http://www.itsoc.org/publications/nltr/it_dec_03final.pdf.

Alan S Willsky. Multiresolution Markov models for signal and image processing. *Electrical Engineering*, 90(8):1396–1458, 2002. ISSN 00189219. doi: 10.1109/JPROC.2002.800717. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1037568.

M Wolf, M Buchheim, E Hegewald, L Krienitz, and D Hepperle. Phylogenetic position of the Sphaeropleaceae (Chlorophyta). *Plant Systematics and Evolution*, 230(3-4):161–171, 2002. ISSN 03782697. doi: 10.1007/s006060200002. URL http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s006060200002.

Matthias Wolf and Jorg Schultz. ITS Better Than Its Reputation. *Science E-Letter*, 2009. URL http://www.sciencemag.org/content/325/5941/682.full/reply#sci_el_12692.

Matthias Wolf, Marco Achtziger, Jörg Schultz, Thomas Dandekar, and Tobias Müller. Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures. *RNA*, 11(11):1616–1623, 2005a. ISSN 13558382. doi: 10.1261/rna.2144205. URL http://dx.doi.org/10.1261/rna.2144205.

Matthias Wolf, Joachim Friedrich, Thomas Dandekar, and Tobias Müller. CBCAnalyzer: inferring phylogenies based on compensatory base changes in RNA secondary structures. *In Silico Biology*, 5(3):291–294, 2005b. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15996120.

Matthias Wolf, Benjamin Ruderisch, Thomas Dandekar, Jörg Schultz, and Tobias Müller. ProfDistS: (profile-) distance based phylogeny on sequence–structure alignments. *Bioinformatics*, 24(20):2401–2402, 2008. ISSN 13674811. doi: 10.1093/bioinformatics/btn453. URL http://www.ncbi.nlm.nih.gov/pubmed/18723521.

Hui Yao, Jingyuan Song, Chang Liu, Kun Luo, Jianping Han, Ying Li, Xiaohui Pang, Hongxi Xu, Yingjie Zhu, Peigen Xiao, and Shilin Chen. Use of ITS2 Region as the Universal DNA Barcode for Plants and Animals. *PLoS ONE*, 5(10):9, 2010. URL http://dx.plos.org/10.1371/journal.pone.0013102.

Frederick W Zechman. PHYLOGENY OF THE DASYCLADALES (CHLOROPHYTA , ULVOPHYCEAE) BASED ON ANALYSES OF RUBISCO LARGE SUBUNIT (rbc L) GENE SEQUENCES. *Journal of Phycology*, 39(4):819–827, 2003. ISSN 00223646. doi: 10.1046/j.1529-8817.2003.02183.x. URL http://doi.wiley.com/10.1046/j.1529-8817.2003.02183.x.

Zemin Zhang and William I Wood. A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics*, 19(2):307–308, 2003. URL http://www.ncbi.nlm.nih.gov/pubmed/12538263.

12 List of Figures

13 List of Tables

14 List of Abbreviations

A	Adenine
AIC	Akaike's Information Criterion
AJAX	Asynchronous JavaScript and XML
BLAST	Basic Local Alignment Search Tool
C	Cytosine
CD	Compact Disc
CGI	Common Gateway Interface
CRUD	Create Read Update Delete
CSS	Cascading Style Sheets
DB	Database
DOM	Document Object Model
DRY	Don't Repeat Yourself
FCGI	Fast Common Gateway Interface
FTP	File Transfer Protocol
G	Guanine
GB	Gigabyte
GI	GenInfo Identifier
GTR	General Time Reversible
GUI	Graphical User Interface
HMM	Hidden Markov Model
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
ITS1	Internal Transcribed Spacer 1
ITS2	Internal Transcribed Spacer 2
IUPAC	International Union of Pure and Applied Chemistry
JC	Juces Cantor
JS	JavaScript
JSON	JavaScript Object Notation
LSU	Large Subunit

MDS	Multidimensional Scaling
ML	Maximum Likelihood
MS	Macro State
MSSA	Muliple Sequence Structure Alignment
MVC	Model View Controller
NCBI	National Center for Biotechnology Information
NJ	Neighbor Joining
PAM	Percent of Accepted Mutations
PDF	Probability Density Function
PEM	Percent of Expected Mutations
Perl	Practical Extraction and Report Language
PL/pgSQL	Procedural Language/PostgreSQL Structured Query Language
PNJ	Profile Neighbor Joining
REST	Representational State Transfer
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristics
rRNA	ribosomal Ribonucleic Acid
RSS	Really Simple Syndication
SOAP	originally: Simple Object Access Protocol
SQL	Structured Query Language
SSU	Small Subunit
SVN	Subversion
T	Thymine
TT	Template Toolkit
U	Uracil
XML	Extensible Markup Language
XSD	XML Schema Definition
YUI	Yahoo User Interface

15 Annex

15.A Phylogenetic tree of class Chlorophyceae

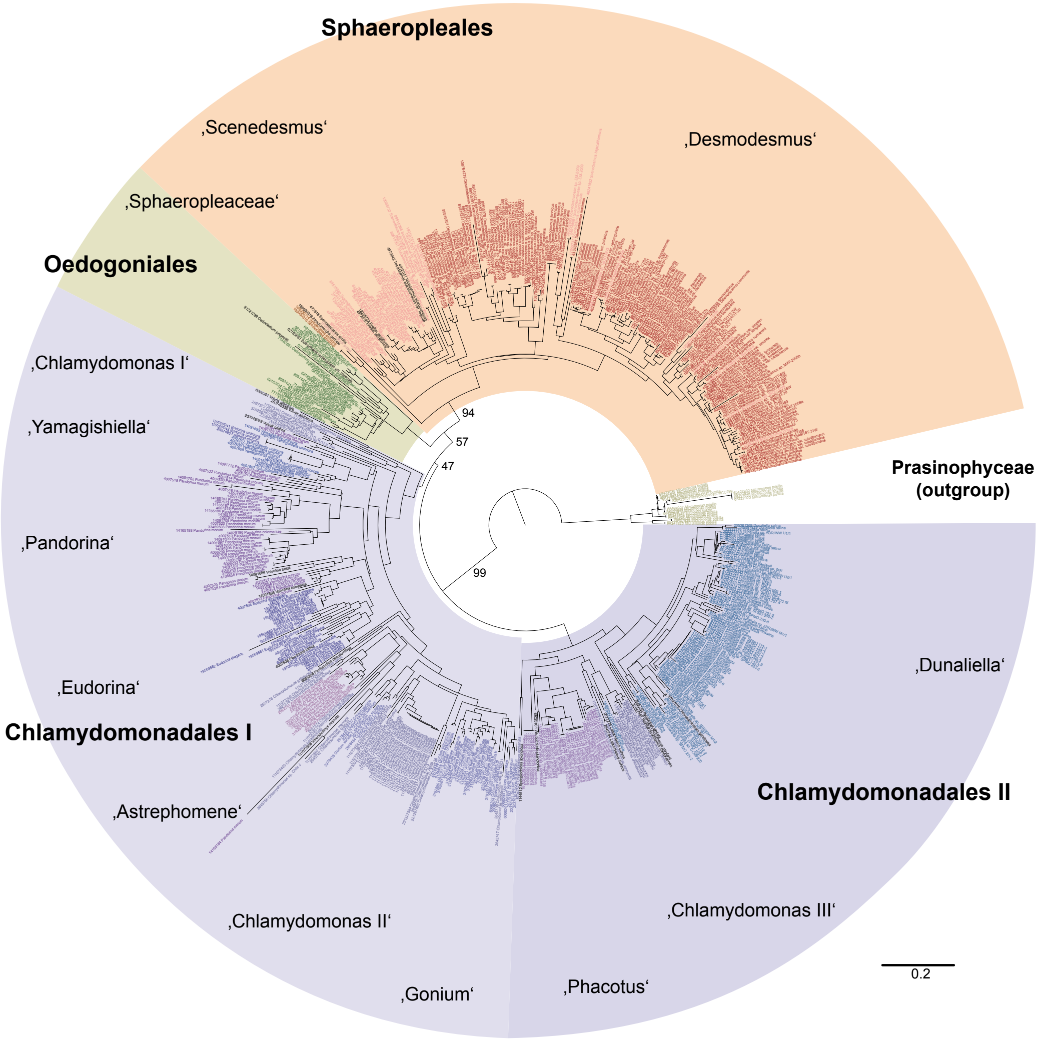


Figure 37: Profile Neighbour joining tree calculated on the class of Chlorophyceae (with 100 bootstrap replicas) using all ITS2 sequences and structures available from the ITS2 database version 3.0 (2009/09/29). The image was taken from Buchheim et al. (2011a).

15.B Phylogenetic tree of class Trebouxiophyceae

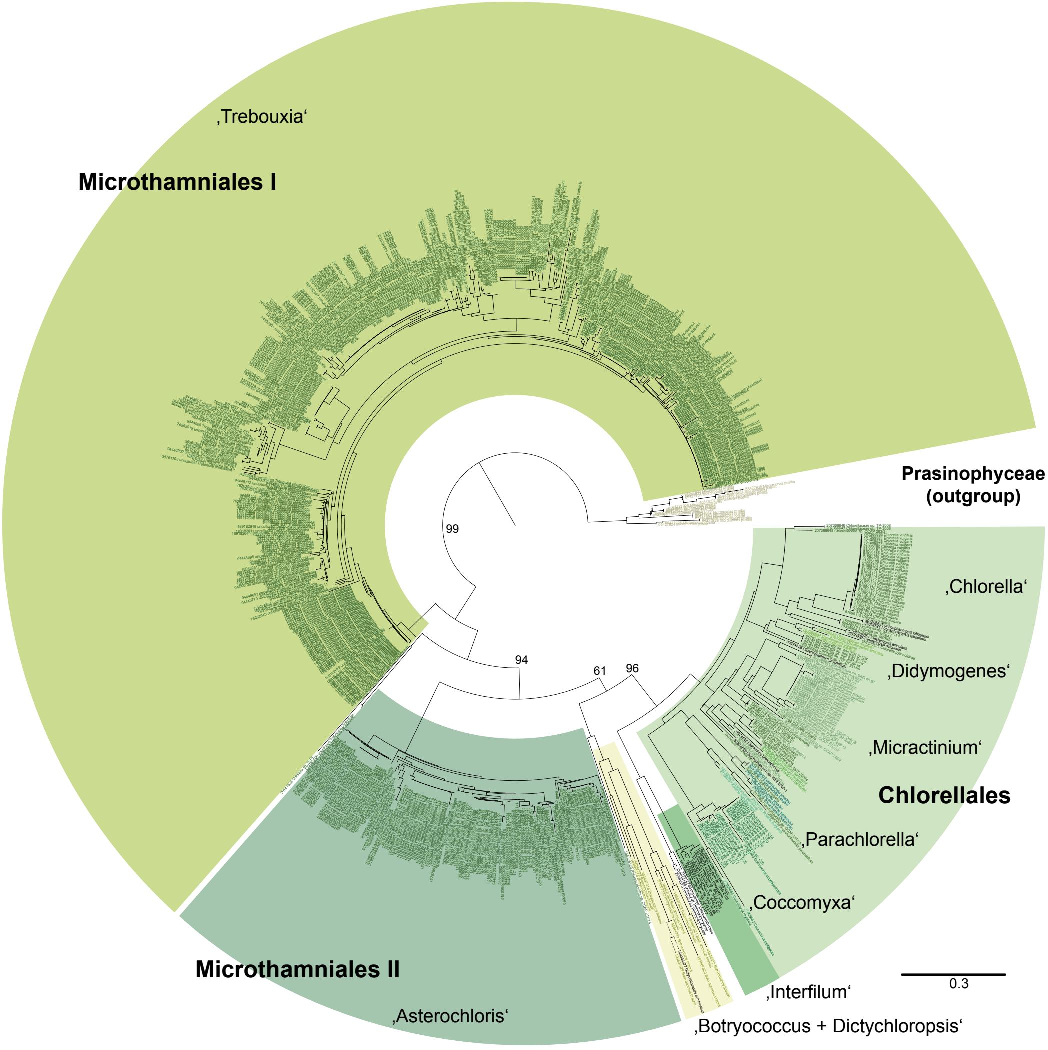


Figure 38: Profile Neighbour joining tree calculated on the class of Trebouxiophyceae (with 100 bootstrap replicas) using all ITS2 sequences and structures available from the ITS2 database version 3.0 (2009/09/29). The image was taken from Buchheim et al. (2011a).

15.C Phylogenetic tree of class Ulvophyceae

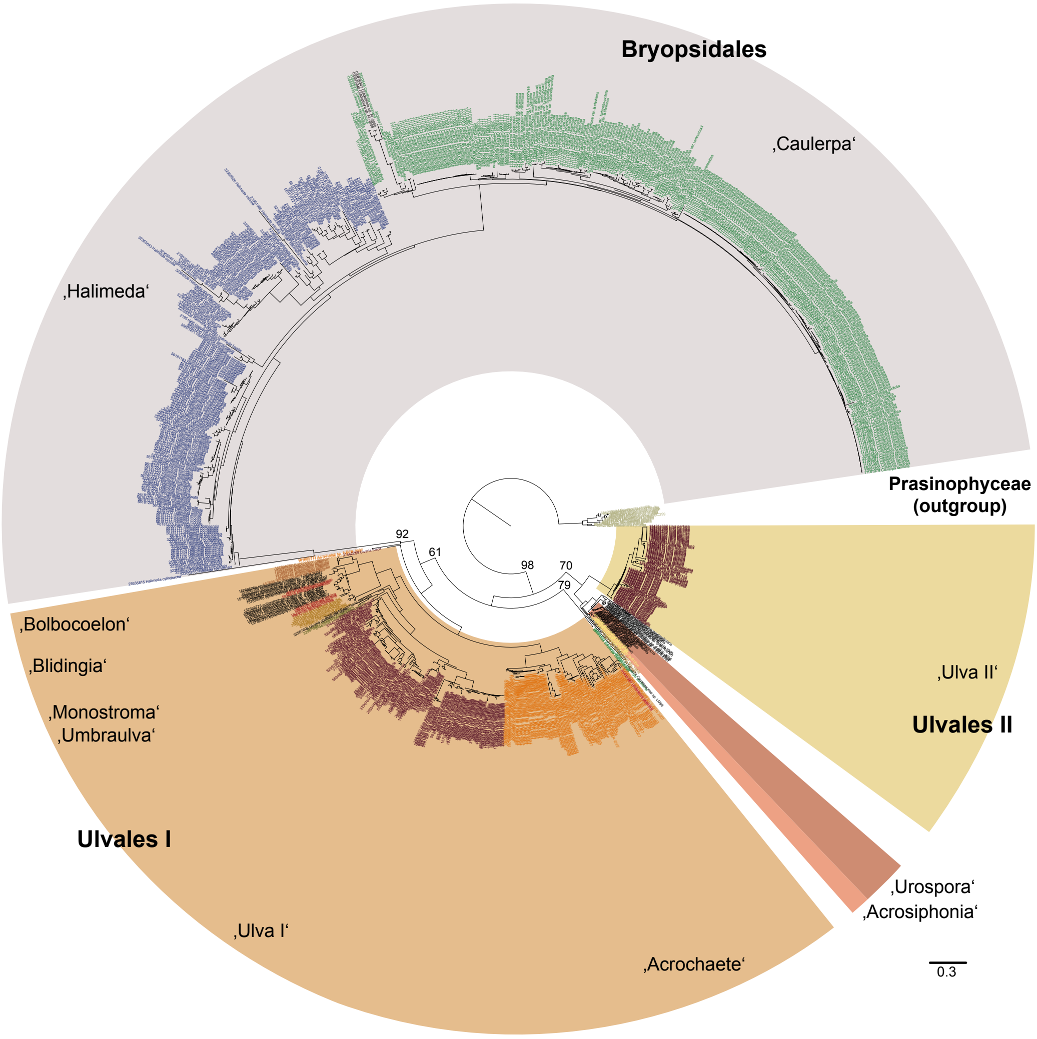


Figure 39: Profile Neighbour joining tree calculated on the class of Ulvophyceae (with 100 bootstrap replicas) using all ITS2 sequences and structures available from the ITS2 database version 3.0 (2009/09/29). The image was taken from Buchheim et al. (2011a).

15.D Sequence-structure alphabets

DNA coding sequence only with 4 letters

Compatible treeforge input format: fasta,xfasta

Sequence	A	C	G	T	U	else/N	-
Substitution	A	C	G	T	T	N	-

RNA10 coding sequence and structure with 10 letters

Compatible treeforge input format: xfasta

Sequence	A	C	G	T	U	N	A ↔ T	A ↔ U	T ↔ A	U ↔ A	C ↔ G	G ↔ C	G ↔ T	G ↔ U	T ↔ G	U ↔ G	N/else ↔ N/else	-
Structure	.	.	.	.	.	.	(↔)	(↔)	(↔)	(↔)	(↔)	(↔)	(↔)	(↔)	(↔)	(↔)	(↔)	-
Substitution	A	R	N	D	D	X	L	L	S	S	M	F	P	P	T	T	C	-

RNA12 coding sequence and structure with 12 letters

Compatible treeforge input format: xfasta

Sequence	A	C	G	T	U	else/N	A	C	G	T	U	else/N	A	C	G	T	U	else/N	-
Structure	.	.	.	.	.	.	(	(	(	(	(	(	)	)	)	)	)	)	-
Substitution	N	Q	H	K	K	X	A	D	E	I	I	Y	R	C	G	L	L	Z	-

RNA16 coding sequence and structure with 16 letters

Compatible treeforge input format: xfasta

Sequence	A	C	G	T	U	else/N
Structure	.	.	.	.	.	.
Substitution	A	R	N	D	D	X

Sequence	A ↔ T	A ↔ U	T ↔ A	U ↔ A	C ↔ G	G ↔ C	G ↔ T	G ↔ U	T ↔ G	U ↔ G	else/N ↔ else/N
Structure	(	(	(	(	(	(	(	(	(	(	(
Substitution	L	L	S	S	M	F	P	P	T	T	V

Sequence	A ↔ T	A ↔ U	T ↔ A	U ↔ A	C ↔ G	G ↔ C	G ↔ T	G ↔ U	T ↔ G	U ↔ G	else/N ↔ else/N	-
Structure	)	)	)	)	)	)	)	)	)	)	)	-
Substitution	C	C	H	H	Q	E	G	G	I	I	W	-

15.E Sequence and sequence-structure score matrices on species datasets

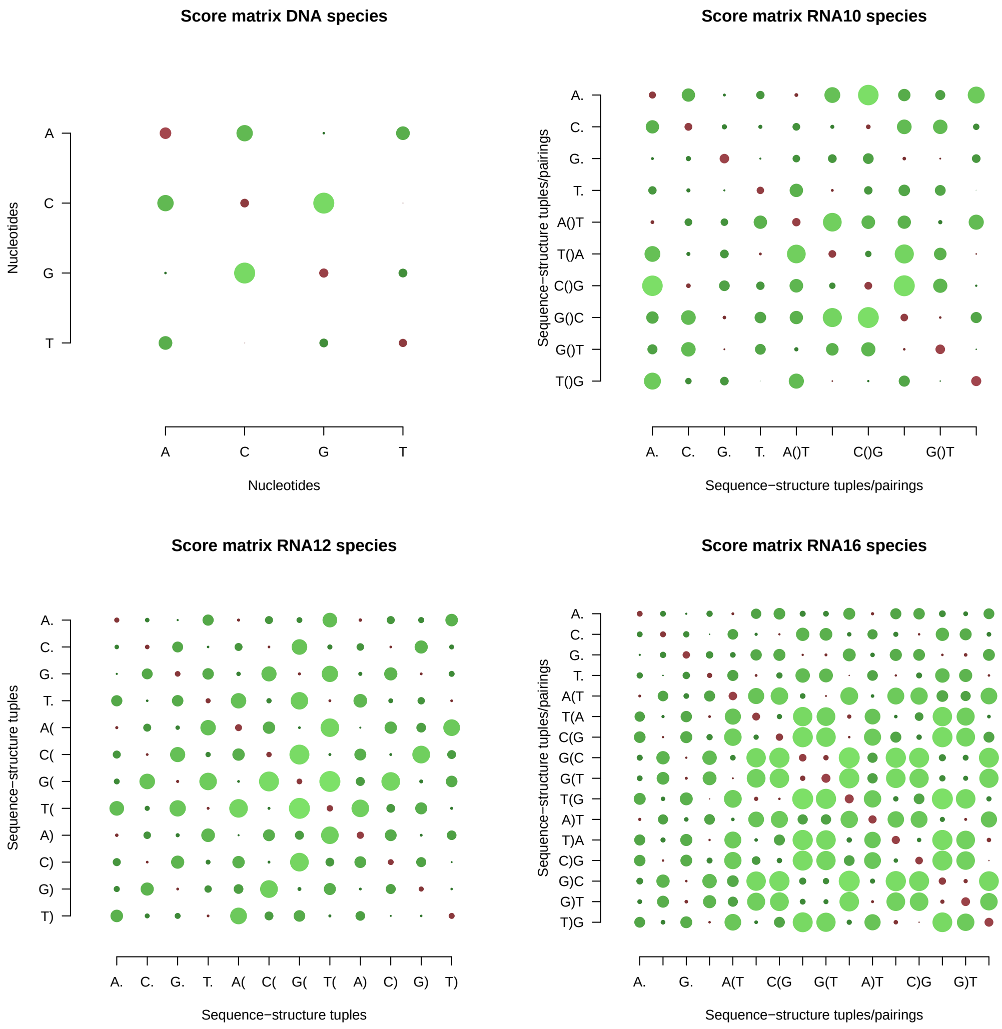


Figure 40: The following bubble plots visualize four different score matrices calculated on the alphabets ’DNA’, ’RNA10’, ’RNA12’ and ’RNA16’. The calculations were based on about 1200 manually inspected species alignments taken from Müller et al. (2007). Red dots indicate a positive, green dots a negative score. The size of each dot corresponds to the absolute size of its representing number. Thus, the main diagonal typically shows red, positive scoring bubbles for matching positions. Other diagonals show mostly a negative score, depending on the number of exchanges which occurred for the specific tuples.

15.F Sequence and sequence-structure score matrices on genus datasets

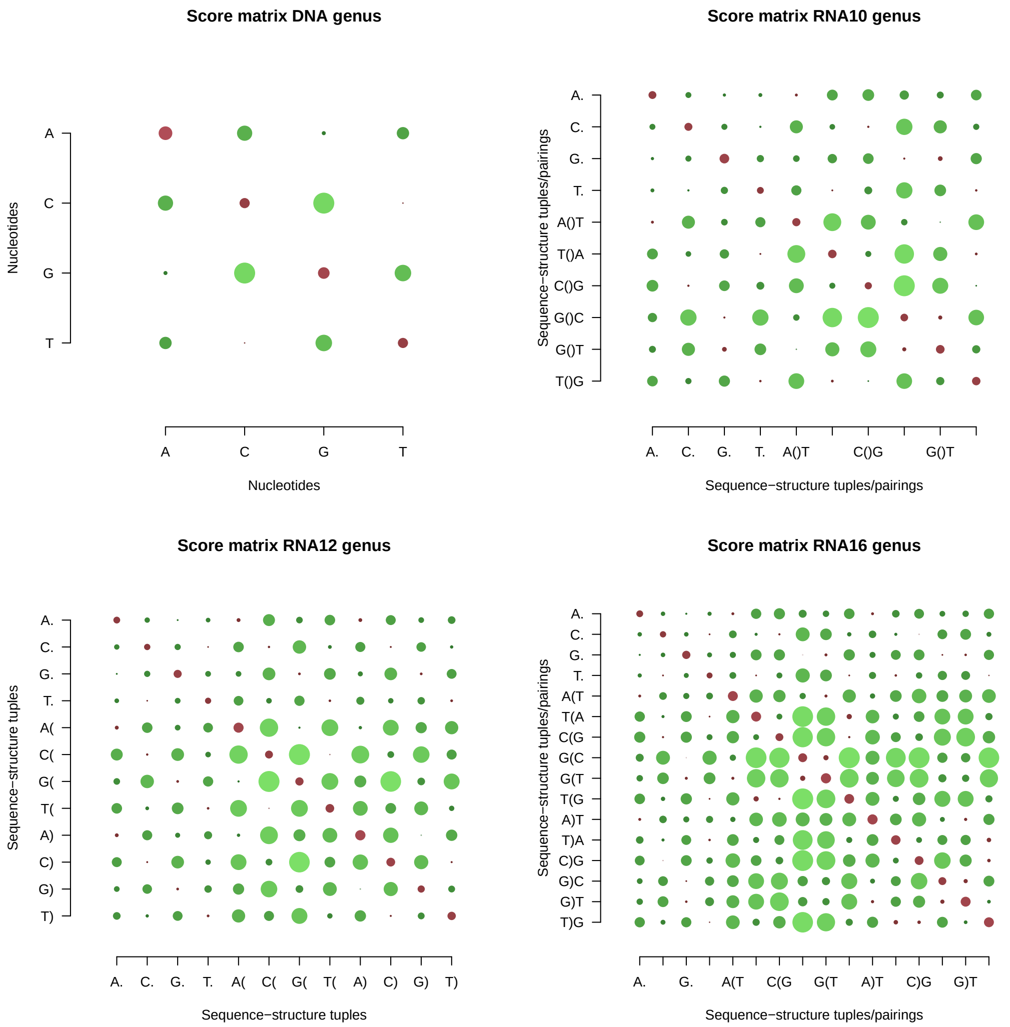


Figure 41: The following bubble plots visualize four different score matrices calculated on the alphabets 'DNA', 'RNA10', 'RNA12' and 'RNA16'. The calculations were based on about 400 manually inspected genus alignments taken from Müller et al. (2007). Red dots indicate a positive, green dots a negative score. The size of each dot corresponds to the absolute size of its representing number. Thus, the main diagonal typically shows red, positive scoring bubbles for matching positions. Other diagonals show mostly a negative score, depending on the number of exchanges which occurred for the specific tuples.

15.G PL/pgSQL functions of the ITS2 database

getGiForTaxon(text)	Returns all GIs for a case insensitive taxonomic name
getGiNameRankForTaxon(text)	Returns all GIs, names and ranks for a case insensitive taxonomic name
getLineage(int4)	Returns the taxonomic lineage to a given taxid
getLineageGI(int4)	Returns the taxonomic lineage to a given GI
createViews()	Creates all views specified in 15.H
getITS2OptimalFolds()	Returns fold ITS2 features in the priority order of HMM, MIXED, GB, HM, BLAST: ><gi> | <fid> |[<gi> | <fid> |] <sequence> <structure>
getLastITS2OptimalFolds()	Returns fold ITS2 features in the priority order of HMM, MIXED, GB, HM, BLAST from the last homology modelling iteration: ><gi> | <fid> |[<gi> | <fid> |] <sequence> <structure>
getITS2OptimalUnfolds()	Returns unfold ITS2 features in the priority order of HMM, MIXED, GB, HM, BLAST: ><gi> | <fid> | <start_pos lt end_pos> | <modstart> | <modstop> |[<gi> | <fid> | <start_pos lt end_pos> | <modstart> | <modstop> |] <sequence> <structure>
updateTaxonCountsRecursive()	This method updates all parent ids up to the root (not the root itself) with new taxon sums
makeProductionReady()	• Deletes all features which are not of element ITS2 • Deletes fold features which are redundant in the priority order of HMM, MIXED, GB, HM, BLAST • Deleting annotations which are in conflict with fold features in the priority order (for fold) of DIRECT, HM, PARTIAL • Deletes unfold features which are redundant in the priority order of HMM, MIXED, GB, HM, BLAST • Drops table gb_indexes; this table is not needed in production environment • Creates triggers: taxonCountsChanged, featureCountsChanged, structureCountsChanged • Updates taxonomy tree with structure counts
public.textcat_null(text, text)	Concatenates strings containing NULL

pg_grant(TEXT, TEXT, TEXT, TEXT)	Grants rights so user, e.g.
	select pg_grant('userreadonly ','select','%','public');
	select pg_grant('userall ','select,insert,update,delete','%','public');

pg_revoke(TEXT, TEXT, TEXT, TEXT)	Revokes rights from user
pg_drop()	Drops all tables

Table 7: Overview of PL/pgSQL functions available in the ITS2 database

15.H Views of the ITS2 database

ViewITS2Folds	Returns gi, fid, seq, str, ann, start_lt_stop, modstart, modstop, iter from all folded ITS2 features including duplicate annotations
ViewITS2Unfolds	Returns gi, fid, seq, str, ann, start_lt_stop, modstart, modstop from all unfolded ITS2 features including duplicate annotations
ViewITS2OptimalAnnotated	Returns gi, fid, seq, start_lt_stop, modstart, modstop from all ITS2 features in the priority order of HMM, MIXED, GB, HM, BLAST
ViewITS2OptimalFolds	Returns gi, fid, seq, str, ann, start_lt_stop, modstart, modstop, iter from all folded ITS2 features in the priority order of HMM, MIXED, GB, HM, BLAST
ViewITS2OptimalUnfolds	Returns gi, fid, seq, str, ann, start_lt_stop, modstart, modstop from all unfolded ITS2 features in the priority order of HMM, MIXED, GB, HM, BLAST

Table 8: Overview of views available in the ITS2 database

15.I Aggregate functions of the ITS2 database

textcat_all	Concatenates grouped strings containing NULL

Table 9: Overview of PL/pgSQL aggregate functions available in the ITS2 database

15.J Operators of the ITS2 database

<table>
<tr><td>||+</td><td>Concatenates VALUE ||+ NULL to VALUE</td></tr>
</table>

Table 10: Overview of PL/pgSQL operators in the ITS2 database

15.K Triggers of the ITS2 database

taxonCountsChanged	AFTER UPDATE ON taxons, updates taxon counts recursive
featureCountsChanged	AFTER INSERT OR DELETE ON features, updates taxon counts recursive
structureCountsChanged	AFTER INSERT OR DELETE ON structures, updates taxon counts recursive

Table 11: Overview of triggers, set in the ITS2 database

15.L ITS2 data generation – folders

sql	SQL-scripts creating and deleting the database with all functions, etc.
lib	Perl modules and different interfaces
scripts	All scripts executed for data generation, invoked by generate.pl
hmms	HMMER2 HMM database and source sequence files for HMM generation
log	Log-directory during the generation process
tmp	Temporary files
backup	PostgreSQL database backups from different stages
bin	Binary files like EMBOSS package, HMMER, Unafold, etc.

Table 12: Overview of folders for ITS2 database generation located under: ’rRNA_DB/db/trunk’

15.M ITS2 data generation – scripts and modules

generate.pl	Basic entry point for data generation
01-mk-pg-database.pl	An empty PostgreSQL database is created
02-fill-taxonomy-tree.pl	The taxonomy tree is downloaded and written to the database
03-download-genbank.pl	Download of important GenBank sub-databases
04-mk-genbank-index.pl	Index of GenBank files is created
05-parse-ncbi-search.pl	GenBank entries retrieved from NCBI-search are parsed
06-split-genbank.pl	GenBank databases are split for HMM annotation
07-write-ncbi-search.pl	Features and sequences from NCBI-search are written to the database
08 - 09	HMM annotation
10	Direct folding of sequences
11-*	Homology modelling
12-*	BLAST-based annotation
13-*	Partial structure prediction

Table 13: Overview of scripts for ITS2 database generation located under: ’rRNA_DB/db/trunk/scripts’

DbInterface.pm	Module for interfacing with a database, provides methods for e.g. writing features, sequences, receiving ITS2 optimal folds or unfolded features
GbInterface.pm	A comprehensive module for interfacing with GenBank data files; these can be read directly from the index file; contains accessors for obtaining sequence, lineage, gi, features and more
Index.pm	A module for creating, appending and reading from an index file; basically usable for any kind of files; independent from GenBank, just requires a specific file separator
FastaParser.pm	Robust parser for FASTA format, parses also 60 line-break FASTA
GbParser.pm	An extremely speed optimized parser for GenBank data files
NCBISearch.pm	Module for NCBI queries and downloads
Blast.pm	Module for BLAST search and parsing
Fold.pm	Module for folding sequences using UNAFold
Nussinov.pm	Module for post-folding homology modelled structures
HMM.pm	Parsing of various HMMER2 annotated features
ITS2check.pm	A module that tests whether the secondary structure of an ITS2 seems to be valid
globals.pm	This module contains all paths, constants and settings used during the update process

Table 14: Overview of Perl modules for ITS2 database generation located under: ’rRNA_DB/db/trunk/lib’

15.N JavaScript files of the ITS2 workbench frontend

accordion_analyse	Visualization of panel in the accordion under 'Analyse' tab
accordion_create	Visualization of panel in the accordion under 'Create' tab
accordion_manage	Visualization of panel in the accordion under 'Manage' tab
accordion_own_seq	Visualization of panel in the accordion under 'Create' tab under 'add own data' button
accordion_taxbrowser	Visualization of panel in the accordion under 'Create' tab under 'from database' button
analyse_alignmentvis2	Visualization of panel for showing a reduced and speed optimized version of the alignment
analyse_alignmentvis	Visualization of panel for showing the full version of alignment
analyse_treevis	Visualization of panel for showing trees
applayout	Direct JavaScript entry point, configures viewport with a border layout
blast	Visualization of panel when performing a BLAST search
COPYING	License information
create_annotate	Visualization of panel under 'Create' tab for own sequence annotation
create_known	Visualization of panel under 'Create' tab for own sequence upload
create_model	Visualization of panel under 'Create' tab for own sequence homology modelling
create_motif	Visualization of panel under 'Create' tab for own sequence motif search
create_predict	Visualization of panel under 'Create' tab for own sequence-structure prediction
create_seqvis	Visualization of panel under 'Create' tab and live-search for database seq-str search
devel	Visualization of panel for bug tracker - visible only on the development web server
its2_classes	Collection of general classes, specific to the ITS2, e.g. seq-str parser, etc.
its2_SequenceEditor	Class for ITS2 sequence-structure input
main_accordion	Visualization of accordion panel on the west side of the border layout
main_dragdrop	Drag & Drop management for the whole web interface
main_examples	ITS2 data examples for different components, mostly loaded into 'SequenceEditor'
main_functions	Basic unspecific functions used in the whole workbench e.g. error and alert messaging
main_handler	Basic event handlers of the workbench, specific event handling is included in each file
main_header	Visualization of the header menu in the north of the border layout
main_interactions	Functions for interaction of different components like sequence transfer between panels
main_overrides	Some basic overrides and extensions of the ExtJS framework
main_pool	Visualization of panel handling the pool and sequence-structure drag & drops
main_tools	Visualization of panel containing the tools section above the accordion on the west side
manage_poolvis	Visualization of panel for managing the data pool containing sequences and structures

Table 15: A short description of the major JavaScript files in the frontend of the ITS2 workbench.

15.O Templates of the Template Toolkit

about	Visualization of ’About’ page (currently hidden)
accordionAnalyse	Visualization of panel in the accordion under ’Analyse’ tab
accordionManage	Visualization of panel in the accordion under ’Manage’ tab
analysetreevis	Visualization of panel for phylogenetic tree
annotate	Visualization of annotation results from ’Annotate’ tab
citation	Visualization of ’Citations’ page (currently hidden)
citedby	Visualization of ’Cited by’ page
contact	Visualization of ’Contact’ page
createannotate	Visualization of panel in ’Annotate’ tab
createknown	Visualization of panel in ’Own sequences’ upload tab
createmodelHM	Visualization of homology modelling results from ’Model’ tab
createmodel	Visualization of panel in ’Model’ tab
createmotif	Visualization of panel in ’Motif’ tab
createpredictDF	Visualization of direct fold results from ’Predict’ tab
createpredictHM	Visualization of homology modelling results from ’Predict’ tab
createpredict	Visualization of panel in ’Predict’ tab
erroralert	Visualization of error messages
error	Visualization of error message when template is not found
flowchart	Visualization of ’Flow chart’ page
funding	Visualization of ’Funding’ page (currently hidden)
fun	Visualization of ’Fun’ page
grouppubbib	Visualization of ’Group Publications (BibTeX)’ page
grouppub	Visualization of ’Group Publications’ page
helpannotate	Visualization of help page showing usage information about ’Annotate’
helpblast	Visualization of help page showing usage information about ’Blast’
helpmodel	Visualization of model page showing usage information about ’Model’
helpmotif	Visualization of motif page showing usage information about ’Motif’
helppredict	Visualization of predict page showing usage information about ’Predict’
helpsequenceeditor	Visualization of sequence editor page showing usage information about sequence editor
highlight	Visualization of ’Highlight papers’ page
howtocite	Visualization of ’How to cite us’ page
links	Visualization of ’Links’ page (currently hidden)
motif	Visualization of motif search results from ’Motif’ tab
showdetails	Visualization of details page after sequence-structure search
staff	Visualization of ’Staff’ page
statistics	Visualization of ’Statistics’ page (currently hidden)
supplements	Visualization of ’Supplements’ page (currently hidden)
updates	Visualization of ’Updates’ page (currently hidden)
usagepolicy	Visualization of ’Usage policy’ page
whatsnew	Visualization of ’What’s new’ page

Table 16: A short description of the major template files of the ITS2 workbench.

15.P Controller of the ITS2 workbench backend

alignment	Creates a sequence or sequence-structure based alignment from all sequences in the pool; stores the alignment in the pool
alvis	Reads alignment from pool and prepares it for highlighting open/closing brackets of the secondary structure (this step is integrated in the backend for performance speed-ups)
annotate	Provides HMM sequence annotation; further calculates images of stem hybridization and returns newly annotated ITS2 positions
blast	Performs blast search on either sequence only, sequence-structure or sequence-structure database without partials; allows to cancel a Blast search or to download results
devel	Provides submission and retrieval of bugs to an XML file
livesearch	Provides a quick live search on taxonomic names and returns the number of available structures; also provides a method for fetching sequence and structure records for specific GIs
managedataset	Allows to manage the pool by removing all sequences, all alignments, all trees or even the whole pool
marker	Gives the short name of all markers stored in the ITS2 database
model	Runs a full homology modelling on template and target sequences; supports multiple templates, multiple structures and the option to find one best consensus template
motif	Performs motif search on sequences and returns start and stop position for each motif; if secondary structure is available, an SVG-file with coloured motifs is created
pool	Provides the possibility to add/remove sequence-structure entries or complete taxa to the pool; also returns information such as number of elements in the pool
poolvis	Returns all information available for sequences inside the pool; further calculates corresponding images of secondary structures; or exports the sequence part of the pool
predict	Folds a sequence, if folding fails in the web frontend, automatically a homology modelling step is performed by using the best hits of a prepended Blast search.
Root	The Root controller deletes expired sessions, creates a new session and redirects to the index.html file
seqdownload	A controller which handles the database download of selected/all sequences, alignments, trees when data is not fully available in the frontend yet
seqvis	Provides all details visible after a sequence search, either by the live-search option or from the taxonomic tree
session	Restores/saves a (previous) session by uploading/downloading a stored pool.xml file; also provides a method to add own sequences to a pool session-file
taxbrowser	Returns the taxonomic tree – visible on the left side of the workbench, including sequence/structure counts;
tree	Reads alignment, calculates tree, writes tree to pool
treevis	Visualization of a tree stored in the pool; uses Newick Utilities to create an SVG-tree which is then parsed by the ExtJS in the frontend. If specified, the tree is also (re)rooted here.
ttvis	Responsible for the visualization of **all** templates defined by the Template Toolkit

Table 17: A short description of controller integrated in the backend of the ITS2 workbench.

15.Q Modules of the ITS2 workbench backend

Alignment.pm	Creates MUSCLE or ClustalW2-based alignment
Base.pm	Contains basic paths and variables for the backend
Errors.pm	Provides text for error messages
Fold.pm	Folds and validates sequences/structures
Format.pm	Parses FASTA, XFASTA, XXFASTA, RAW, XRAW, XXRAW format
Hmm.pm	Parses HMMER2 output
HomologyModeling.pm	Provides the Homology Modelling
ITS2Check.pm	Checks correctness of folded ITS2 sequences
ITS2Motifs.pm	Annotates ITS2 motifs
Pool.pm	Provides methods to create a pool, add and remove sequences, alignments or trees
StructurePlot.pm	Provides colouring of different structures
Translateprot.pm	’RNA12’ encoding alphabet for sequence and structure
TransSeqStructPseudoProt.pm	’RNA12’ encoding alphabet for sequence and structure
Tree.pm	Module for tree calculation and re-rooting with Newick Utilities

Table 18: A short description of modules integrated in the backend of the ITS2 workbench.

16 Publications

Markert S. M., Müller T., **Koetschan C.**, Friedl T., and M. Wolf. "Y" Scenedesmus (Chlorophyta, Chlorophyceae): the internal transcribed spacer 2 (ITS2) rRNA secondary structure re-revisited. *Plant Biology*, (in press), 2012.

B. Merget, **Koetschan, C.**, T. Hackl, F. Förster, T. Dandekar, T. Müller, J. Schultz, and M. Wolf. The ITS2 Database. *Journal of Visualized Experiments*, 61:e3806, Mar 2012.

Koetschan, C., T. Hackl, Müller T., Wolf M., Förster F., and J. Schultz. ITS2 database IV: Interactive taxon sampling for internal transcribed spacer 2 based phylogenies. *Molecular Phylogenetics and Evolution*, Epub ahead of print, Feb 2012.

M. A. Buchheim, A. Keller, **Koetschan, C.**, F. Förster, B. Merget, and M. Wolf. Internal transcribed spacer 2 (nu ITS2 rRNA) sequence-structure phylogenetics: towards an automated reconstruction of the green algal tree of life. *PLoS ONE*, 6:e16931, Feb 2011.

Koetschan, C., F. Förster, A. Keller, T. Schleicher, B. Ruderisch, R. Schwarz, T. Müller, M. Wolf, and J. Schultz. The ITS2 Database III–sequences and structures for phylogeny. *Nucleic Acids Res.*, 38:D275–279, Jan 2010.

T. Friedrich, **Koetschan, C.**, and T. Müller. Optimisation of HMM topologies enhances DNA and protein sequence modelling. *Stat Appl Genet Mol Biol*, 9:6, Jan 2010.

17 Curriculum Vitae