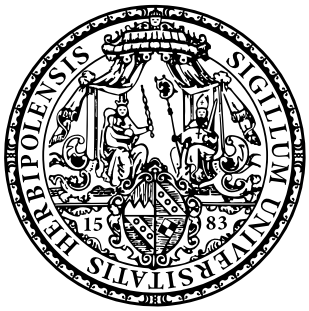
Diploma Thesis

The ITS2 Database - Application And Extension



Christian Selig

submitted on February 22, 2007

Department of Bioinformatics, Faculty of Biology
Julius-Maximilians-University Würzburg

Supervisors
Dr. Matthias Wolf
Prof. Dr. Jörg Schultz

Primary Referee
Prof. Dr. Jörg Schultz

Secondary Referee
Prof. Dr. Thomas Dandekar

Contents

1 Introduction

1.1 Biological Data and Databases

The unpredecented growth of biological data from gene and genome sequencing (Benson et al., 2007, GenBank has doubled in size about every 18 months) is a foundation for a more systematic research on gene and protein sequence and structure, regulatory networks and cascades of gene expression, functional roles, protein interactions and, of course, a more detailed insight into the molecular fundament of evolution.

Even though the exponential and unstoppable growth of primary databases such as NCBI (Benson et al., 2000), KEGG (Kanehisa, 2002) and EMBL (Kulikova et al., 2007) on the one hand challenges scientists to

- perform more careful selection and evaluation of data,
- to integrate knowledge from a large number of databases (Baxevanis, 2003),
- to utilize advanced mathematical methods and models (Cohen, 2004),
- to publish precise algorithms and programs and
- to choose valid methods of statistical analysis

more than ever, on the other hand, it also offers vast opportunities for analyses of varying scale and scope that hopefully contribute to broader, more systematic and quantitative models for life sciences.

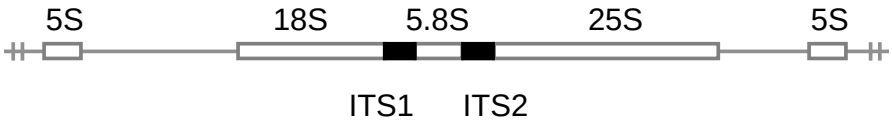


Figure 1.1: rDNA repeat from yeast (adapted from Venema and Tollervey, 1999)

Contributing to the latter aim while taking careful measures to adhere to the former prerequisites, the internal transcribed spacer 2 database (Wolf et al., 2005a; Schultz et al., 2006) aims to contribute to these goals.

1.2 Ribosomal DNA and ITS2

DNA sequences coding for the ribosomal RNA (rDNA repeat, see figure 1.1 for the yeast rDNA) make valuable and reliable datasets for phylogenetic analysis (Olsen and Woese, 1993). Countless analyses have been applied to the sequences that make up the structural part of the ribosome, usually the 18S region (which has been critized, see Petrov and Aleshin, 2002). In the realm of structural biology, ribosomal secondary structures of species from a diverse set of phylogenetic groups is continously revealed and understood more closely (such as the ribosome of the honeybee, *Apis mellifera*, cf. Gillespie et al., 2006).

The internal transcribed spacer 2 region (ITS2) is located between the 5.8S and 28S rDNA region (see figure 1.1) and is removed from the primary transcript. It appeared as a potential and potent marker for phylogeny (Coleman and Mai, 1997b; Coleman and Vacquier, 2002; Coleman, 2003). What makes ITS2 especially interesting and distinct from i.e. ITS1 is the observation that its structure (cf. figure 1.2 on the following page) is conserved in green algae and flowering plants (Coleman and Mai, 1997a) as well as other eukaryotes, a finding that is supported by large scale analysis from Schultz et al. (2005). Comparing two ore more ITS2 secondary structures, another interesting feature becomes accessible: from the viewpoint of the biospecies

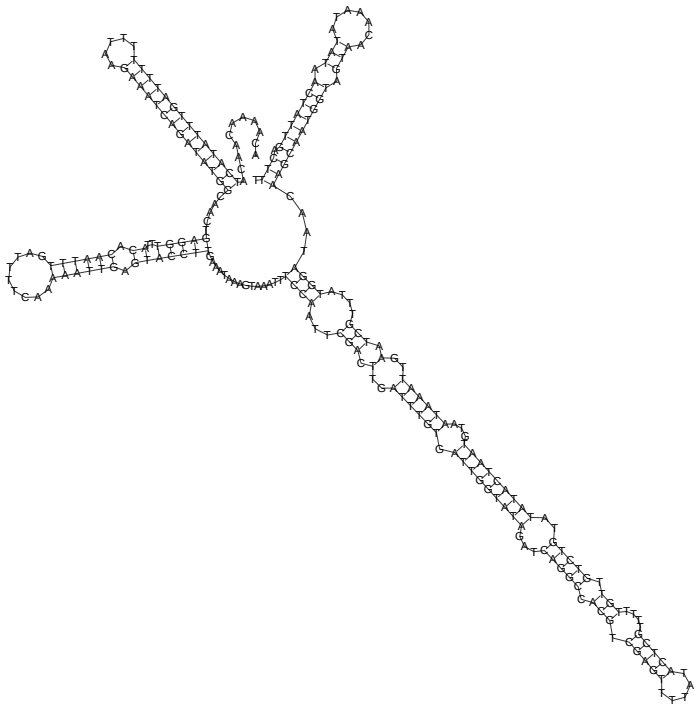


Figure 1.2: Typical ITS2 secondary structure (sequence from Genbank Accession AY652557, *Trichoplax adhaerens*): four helices, the third being the longest

concept (Mayr, 1942, 1967, 1996) whose definition, or indicator hypothesis, of a species includes the production of fertile offspring, it seems that at least one CBC — compensatory base change[1] — between two species' ITS2 secondary structures is an indicator of the two belonging to distinct species (Coleman, 2000; Coleman and Vacquier, 2002; Coleman, 2003), given the large evolutionary distance needed for a CBC to occur. A large-scale study on this correlation is on the way (Müller et al., 2007), with tools for CBC analysis available (Wolf et al., 2005b; Seibel et al., 2006).

It can be well reasoned that the structural conservation of ITS2, while remaining variable on the sequence level, has implications on functional roles in the splicing process of precursor ribosomal RNA (for a host of reviews on

[1] A CBC is an evolutionary event in RNA secondary structures where the basepairing itself is conserved, but both bases were exchanged in the course of evolution.

cleavage and recognition sites in ITS2, see references in Coté and Peculis, 2001).

The usability of ITS2 sequences and their structures as a phylogenetic marker raised the question of the possibility of deriving them automatically in an efficient manner, which over a short period of time led to the implementation of a web-accessible internal transcribed spacer 2 database with fully computer-predicted structural information.

1.3 Previous work on the ITS2 database

The initial work on ITS2 started out by using minimum free energy (MFE) folding of ITS2 sequences downloaded from NCBI, revealing 5092 structures (Schultz et al., 2005) with a computationally checked-for common structure as described by Coleman and Mai (1997a).

From this original dataset, secondary structures were homology modelled (Achtziger, 2005). Simplified, this process goes through 3 steps (first two steps shown in figure 1.3 on the next page):

1. Compute best alignment of sequence without known structure against sequences with known structure.
2. Transfer conserved base pairs.
3. Do quality checking and postfolding[2].

This process revealed more than 20,000 new valid internal transcribed spacer 2 secondary structures that were organized and stored in a relational database management system and made accessible through a web interface (Wolf et al., 2005a) as well as a web service (Schultz et al., 2006). The database allows downloading of sequences for later use, i.e. alignment with MARNA (Siebert and Backofen, 2005), RNAforester (Höchsmann et al., 2003, 2004) or 4SALE (Seibel et al., 2006).

[2] adapted Nussinov folding algorithm (Nussinov and Jacobson, 1980)

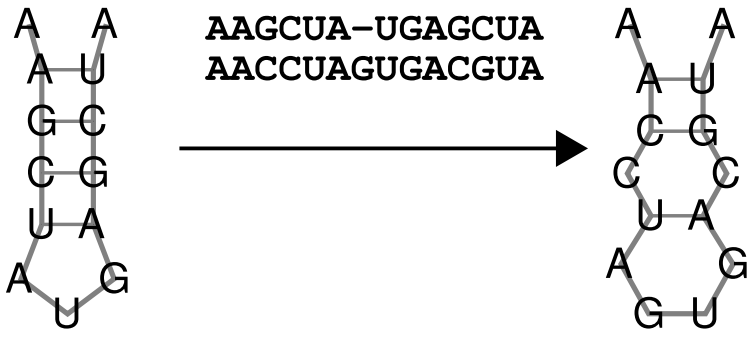


Figure 1.3: Basic principle of homology modelling: Transfer of conserved base pairs from a sequence alignment

1.4 Problem Scope

The internal transcribed spacer 2 database is a valuable resource for phylogenetic analysis. Though not apparent to the end user, it has some drawbacks which stem from the course of its development. Storage of results was taken care of after the actual computations, which despite the clear design of the relational model led to semantic inconsistencies. The database also included intermediate results that were useful when the project started; as the homology modelling method and associated procedures matured, this data remained and added unnecessary complexity to the database, rendering the extraction and display of extended information on stored structures increasingly difficult. Furthermore, some details could not have been and subsequently were not planned for. There was no shared code base for building and updating the database; the result was a duplication of common code.

It became clear that a rewrite, emphasizing on the data structures instead of the used algorithms (Raymond, 2001; Brooks, 1995, Chapter 9), tangibly the relational database, would ensure semantic clarity and extensibility of both the database *and* the code — back-end scripts as well as front-end applications — around it.

1.5 HM alternatives

Minimum free energy folding and homology modelling are very straightforward mechanisms of RNA structure discovery. Yet MFE folding often delivers an RNA conformation that is apparent to a human observer as biologically valid, but cannot readily be classified as a valid structure by a computer. The option here would be searching through the space of energetically suboptimal structures, a computationally highly expensive procedure.

Homology modelling bypasses this problem by pulling an RNA secondary structure from one sequence to a similar sequence, thereby ignoring the folding space problem.

Yet there are other ways of overcoming the folding space problem, two of which will be discussed briefly: RNAshapes (Giegerich et al., 2004; Reeder and Giegerich, 2005; Steffen et al., 2006) and mFold (Zuker, 1989; Zuker et al., 1999; Zuker, 2003).

The question is whether the ITS2 database could have been generated from these algorithms instead of minimum free energy folding and/or homology modelling and whether they would deliver quantitatively and qualitatively better structures.

2 Rebuilding the ITS2 Database

2.1 Materials and Methods

For both backend and frontend, mostly the same programs were used:

Software	Backend	Frontend
Operating System: Generic GNU/Linux (SuSE 9.3 and 10.1)	yes	yes
Database: PostgreSQL 8.1.4 (PostgreSQL Global Development Group, 2006)	yes	yes
Remote Access: OpenSSH 4.2p1	yes	no
Remote Access: GNU wget 1.10.2	yes	no
Scripting Language: Perl 5.8.8	yes	yes
Database Access: Perl DBI 1.50 / Perl DBD 1.43	yes	yes
Biological Data Access: BioPerl 1.4.0 and 1.5.2 (Stajich et al., 2002)	yes	yes
Compiled Language: GNU Compiler Collection 4.1.0 (Stallman, 2003)	yes	no
Compiled Language Library: GNU C Library 2.4	yes	no
RNA Structure Software: ViennaRNA 1.6.1 (Hofacker et al., 1994)	yes	yes

Fast Sequence Search: WU BLAST 2.0[1] (Gish, 2004)	yes	yes
Sequence Search: Paracel BioView Toolkit 5.2.3	yes	no
EMBOSS 8.0[2] (Rice et al., 2000)	no	yes

In addition, mFold Zuker (1989) has been used for a comparative analysis.

2.2 Results

2.2.1 Overall design

The software and process design has been kept as simple as possible (see figure 2.1 on the following page). A set of backend scripts described below fetch data from NCBI, organize them into the database and perform calculations. The frontend accesses the database independently and presents data to the user in a query-driven manner. Code duplication between backend and frontend was not entirely avoided for practical reasons, only code that has general usefulness was swapped to shared libraries.

The backend scripts themselves are functionally separate entities with specific input requirements and clear output definitions. Instead of taking one GenBank entry and going through all steps (writing to database, MFE folding, homology modelling etc.), the scripts can take everything (or a subset thereof) that has been „left over" by the previous step.

[1] chosen over NCBI BLAST (Altschul et al., 1997) for reasons of speed, substitution matrix exchangeability and proficiency (Cha and Rouchka, 2005)

[2] only the program „needle" (statically linked) for pairwise global alignment was used

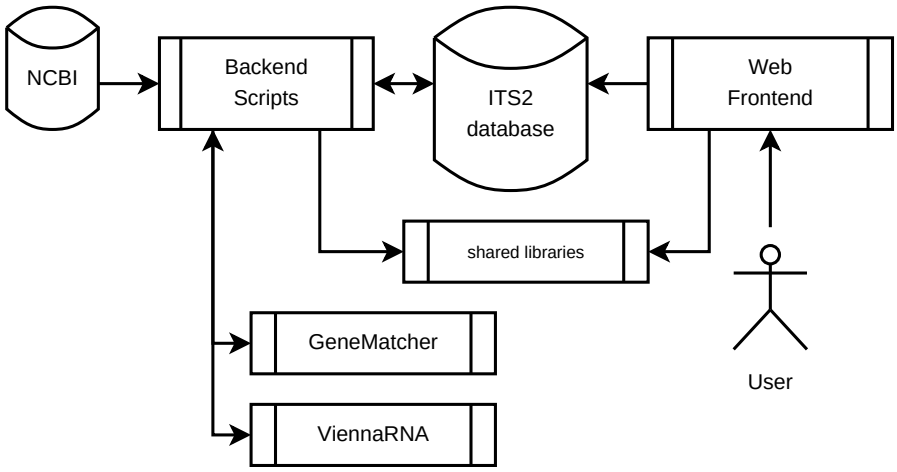


Figure 2.1: A simplied view on the ITS2 database and applications

2.2.2 Custom Back-End Software and Data Structures

2.2.2.1 ITS2 validity checker and feature extractor

A checker of formal ITS2 validity (as described by Coleman and Mai, 1997a) has been implemented in plain ANSI C (Institute, 1999). Input is enhanced FASTA with one ore more sequences where each is followed by one or more structures in dot-bracket notation. The program then checks for the number of helices, the longest helix, the UGGU motif and the UU mismatch motif (cf. appendix D.1 on page 59).

2.2.2.2 Database structure

The relational database model was modified compared to the original one from Achtziger (2005). Taxon count and CBC tables were removed, alignment information now includes the structural alignment. A „run" table contains information on build and update runs. It enables comprehensibility of the timeline of sequence and structure insertion into the database.
The database consists of mostly independent modules[3] (see figure 2.2).

[3]Technically, the database structure can be at least partially used for storage of other RNA secondary structures and for protein secondary structures, mostly within the constraints

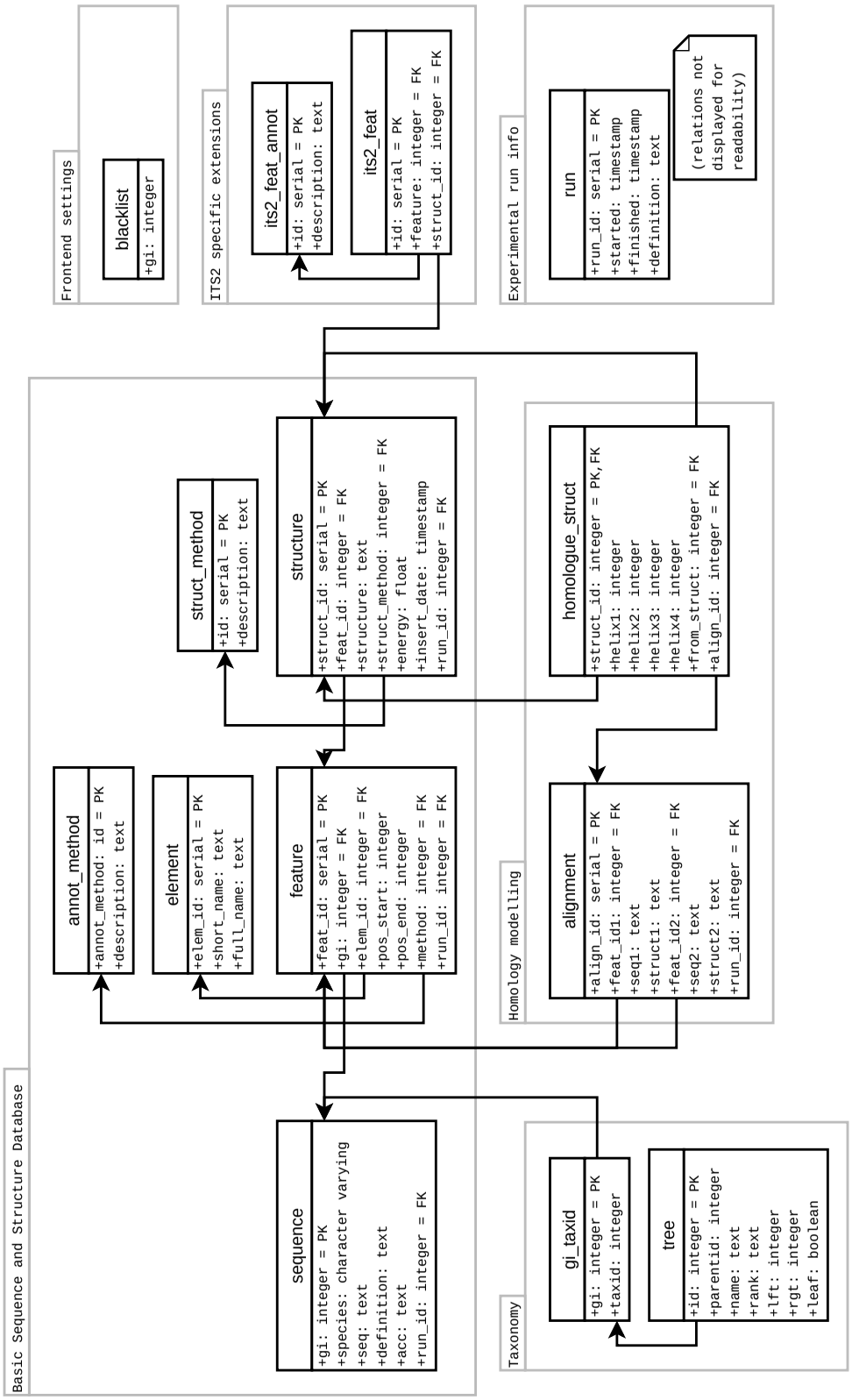


Figure 2.2: Database schema

ID	Structures found . . .
1	. . . by RNAfold of GenBank-annotated features
2	. . . in first homology run
3	. . . in second homology run
4	. . . by RNAfold of BLAST-reannotated features
5	. . . in first homology run of BLAST-reannotated features
6	. . . in second homology run of BLAST-reannotated features
7	. . . as partial structures in GenBank-annotated features

Table 2.2: Structure discovery methods

2.2.2.3 Database rebuild and update scripts

For precise descriptions of the scripts, their accepted input formats, theory of operation, output on files/database, standard command line output, error messages and parameters see the appendix on page 59 et seq.

Beforehand, two sets are defined: Structure discovery methods (see table 2.2) and annotation methods (see table 2.3 on page 18).

NCBI data retrieval The script retrieves data from the NCBI Nucleotide database. The query is *„internal transcribed spacer 2" OR ITS2*. Sequences are stored in a large file and later split up into separate files, one per GenBank identifier.

Database transfer of NCBI sequences Sequences from GenBank files are extracted and written either with full feature annotation or without annotation to the ITS2 database, whereas only ribosomal features are accepted and all others, such as primer binding sites, are ignored. Sequences whose features are annotated as reverse complement („opposite strand") are reversed together with their features. Sequences with features that have different directions are rejected. Features that lead to BioPerl parser errors or that have missing taxon information are also rejected.

of having a database based on NCBI Nucleotide. Taxonomy is optional, depending on the frontend.

Taxonomy update The latest NCBI taxonomy is downloaded from its official FTP site and written to temporary tables in the database. A recursive algorithm adds nested set information. The regenerated tree is transferred to the static tables.

Direct Vienna RNA fold ITS2 features are folded by RNAfold from the Vienna RNA package, evaluated for their correctness by the checker and in case of correctness written to the ITS2 database. The GenBank identifiers of putative ITS2 sequences can either be all selected from the entire database (in the event of a rebuild) or from a file that contains raw GenBank identifiers without or with putative differing annotation.

Folds without reannotation are assigned a „1" as structure discovery method whereas folds from a reannotated source are assigned a „4" and a new feature is inserted into the feature table.

Modelling database writer (accessory script) This script fetches all ITS2 features from the database that for which either the initial MFE folding or the first homology modelling run found a valid structure. It writes them to a standard FASTA file for later use by the GeneMatcher and BLAST loader scripts.

GeneMatcher loader (accessory script) This script uploads the file from the modelling database writer to the GeneMatcher.

Homology run preparation ITS2 features that do not yet have a valid structure in the database can be all selected from the database or supplied by a file with aforementioned format, written to a query file and submitted to the Paracel GeneMatcher through an SSH call to the GeneMatcher primary gateway machine (Parameters: global alignment, gap open penalty -25, gap extension penalty -6, ITS2PAM50 matrix as published by Wolf et al., 2005a). Results in BLAST output format are then split up into separate files.

annotation method	structure method constraints
1: from NCBI	only 1, 2, 3 and 7
2: cutting after homology modelling	only 2 and 3
3: BLAST reannotated positions	only 4, 5 and 6

Table 2.3: Annotation methods

source	new
1	2,5
2	3,6
1,2	7

Table 2.4: Assigned new structure discovery method depending on source structure

Homology modelling From one or more alignment files, the homology modelling process is applied to each alignment. In case of passing the quality parameters ($E \leq 10^{-16}$, all four helices $\geq 75\%$ transferrable, total feature length ≥ 130 basepairs), resulting structures are postfolded and exceeding ends are cutted. Sequence and structure alignment, the new structure, its features, its source structure and if necessary, new feature positions are written to database (see table 2.3), assigning a structure discovery method to the new structure (see table 2.4).
Partial structures that do not fulfill the four-helix and length criteria can optionally be added to the database whenever at least two consecutive helices can be transferred ($\geq 75\%$ transferrability). This depends on setting the correct parameter(s).

BLAST database loader (accessory script) This script is basically the same as the GeneMatcher loader and uses the identical input file for setting up a WU BLAST database.

BLAST run Assembles sequences without valid structures either from a list of GenBank identifiers or the whole database and runs them against BLAST

(Parameters: gapped BLAST, E value threshold 10^{-16}, gap open costs -10, gap extension costs -2, ITS2PAM50 matrix as published by Wolf et al., 2005a). Differing or new ITS2 feature positions in sequences are written to a reannotation file.

2.2.2.4 Rebuild and Update protocols

The scripts explained above can be combined to either rebuild or to update the ITS2 database. The protocol — the detailed way and order of calling the scripts for the two purposes — is summed up in a flow chart (see figure 2.3 on the next page). It is to be noted that both the rebuild process and update processes use exactly the same process, only with differing input and parameter settings.

2.2.3 Custom Front-End Software

The ITS2 database front-end, the „web interface", is a set of Perl scripts and static HTML files. Though based on the original web interface, most of the code underwent a rewrite.

The overall design premise was easy accessibility of functions, implemented through intuitive tabs on top of the page with lesser-used functions on the left side bar. A high emphasis was put on consistent look and feel: Links within dynamic pages are always textual links and cross-references within the database, wheras outlinks (to NCBI Nucleotide and NCBI Taxonomy) are always small icons.

The color scheme of the old ITS2 logo was employed as a Corporate Design throughout the frontend.

A screenshot of the new welcome page is presented in figure 2.4 on page 22.

Significant differences to the old web interface are:

- details for each entry with information on

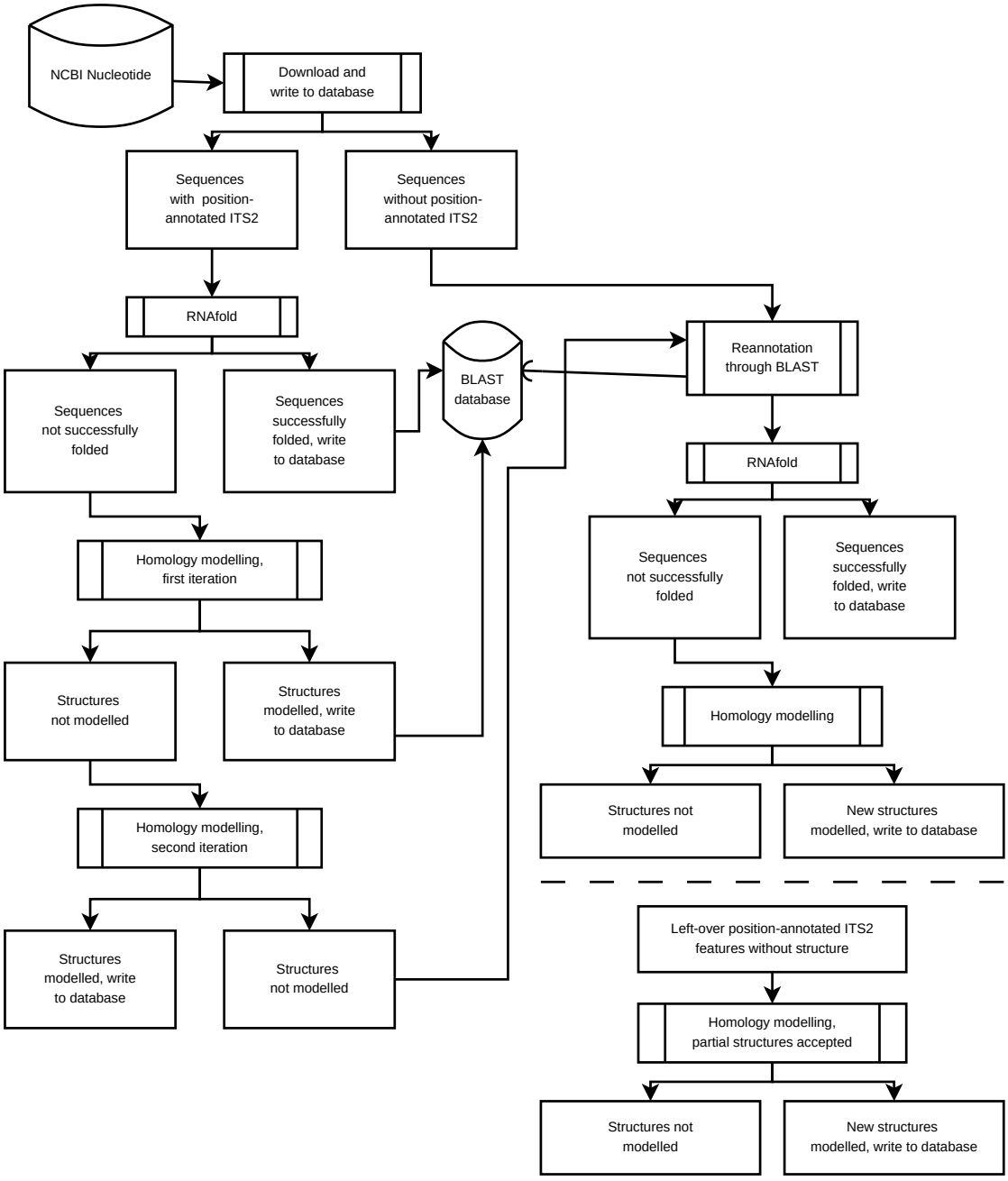


Figure 2.3: Database Rebuild and Update Process

    - GenBank identifier, NCBI accession number, insertion date, taxonomy lineage, sequence, structure, figure of secondary structure, free energy, structure discovery method, annotation method
    - if the structure is a homology modelled structure: source sequence, full alignment including structure, CBC and half-CBC information, alignment E value, helix transfer percentages
    - if the structure is a model for other structures: a tree of structures modelled from currently viewed structure
    - if the structure is a partial structure: figure of template with homologous positions marked red
- a custom modelling service where template structures can be specified by the user or retrieved from the database
- database statistics and latest updates information
- new browse and search mode
- ambiguity checking for the taxonomy[4]
- export to RNAStructML

[4]Taxons were handled as if they had unique names, now their database identifier is used.

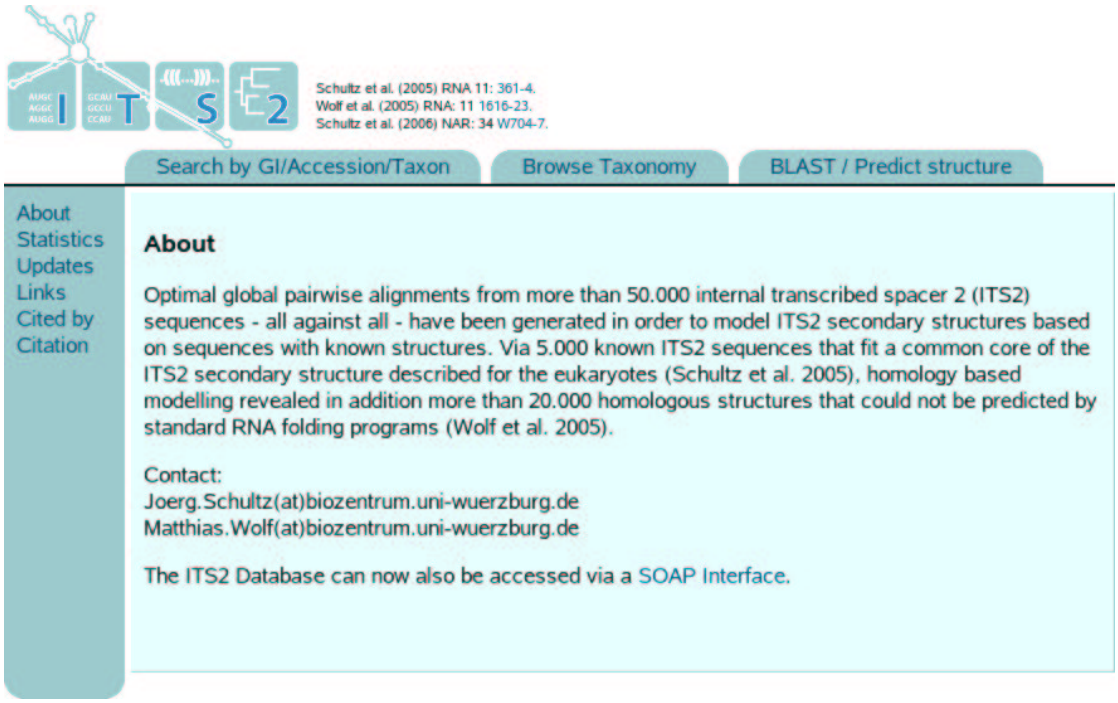


Figure 2.4: The new web interface

2.2.4 Data from database rebuild run

138,753 sequences were downloaded from NCBI on January 22, 20:33 local time. Exactly this number was processed by the download script.

For the database insert, 138,731 sequences were added and 22 were rejected (see table 2.5 on the next page). The distribution of features shows that ITS2 outnumbers all others (see table 2.6 on the following page). There were also numerous sequences that do not have precise feature start and end annotation (see table 2.7 on page 25).

The initial MFE fold revealed 9,883 structures (see table 2.8 on page 25). The left-over sequences were homology modelled, yielding 25,900 new structures. From these, a second HM run yielded 8,978 structures (see tables 2.9 on page 25, 2.10 on page 25, 2.11 on page 26). The quality of homology modelling was measured by plotting the energy of each sequence's MFE fold to

the one given by RNAeval (see figure 2.5 on the following page).

The BLAST run applied on all sequences for which no ITS2 structure could be determined (no matter whether annotated or not) delivered 49,023 putative (re-)annotated ITS2 regions. These newly-selected regions were homology modelled and were (re-)annotated in the database when a full structure could be found. 20,565 new structures were found, 6,355 thereof by cutting, enlarging or sliding an existing ITS2 feature. 14,210 were found in sequences that had no previous annotation.

This left 48,181 annotated ITS2 features for which no structure could be determined. A homology modelling process that accepted partial structures (at least two consecutive helices) revealed additional 11,395 incomplete structures.

The database now contains 76,721 ITS2 sequences and structures, thereof 65,326 complete and 11,395 partial.

Sequences downloaded from NCBI	131,753
rejected features: mixed strand directions in annotation	16
rejected features: error reported by BioPerl	6
Sequences transferred to database	131,731

Table 2.5: Results from database insert on sequences

18S rRNA	37,601
ITS 1	82,282
5.8S rRNA	90,408
ITS 2	99,297
28S rRNA	29,053
5S rRNA	226
Total	338,867

Table 2.6: Results from database insert on features

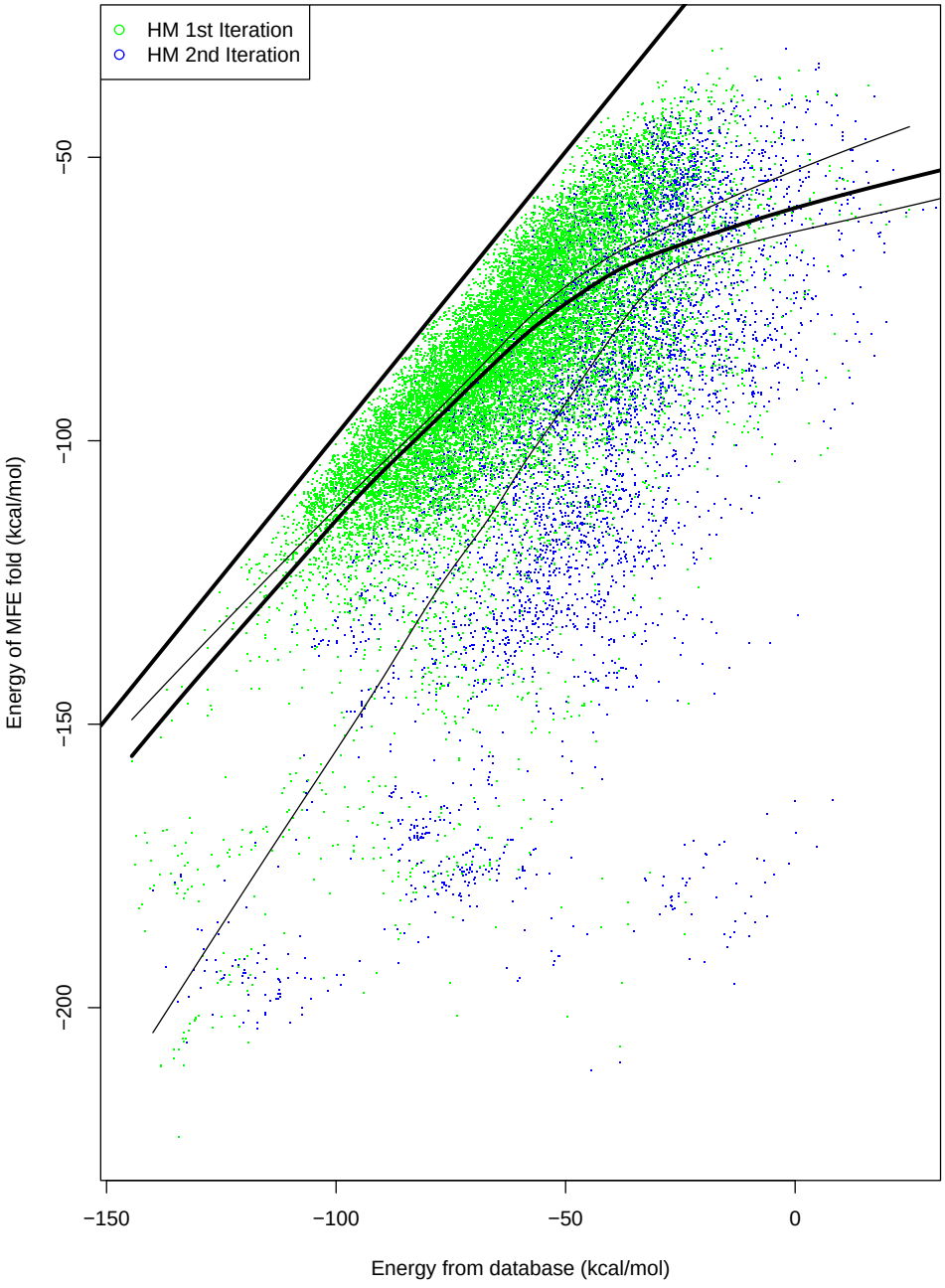


Figure 2.5: Comparison of the energies of homology modelled structures versus energies of MFE fold on their sequences (local non-linear regressions shown)

Sequences that do have at least one annotated feature	102,513
Sequences that do not have annotated features	36,218
Total sequences	138,731

Table 2.7: Results from database insert on provided feature annotation

Sequences with annotated ITS2 feature	99,297
Formally valid MFE fold	9,883
Sequences left	89,414

Table 2.8: Results from MFE fold

Sequences not validly MFE folded	89,414
Successful HM 1st iteration	25,900
Successful HM 2nd iteration	8,978
Sequences left	54,536

Table 2.9: Results from homology modelling runs

Reason	Count
No result / hit / HSP found	30,853
Wrong strand in alignment	64
Missing source structure	0
Structure for GI exists	0
Prediction crosses kingdom	19
Insignificant alignment	9,316
Resulting structure too short	3
Not a valid ITS2 structure	218
Less than 4 helices transferred	23,041
Total	63,514

Table 2.10: Results from first homology modelling run, rejected sequences or structures

Reason	Count
No result / hit / HSP found	29,017
Wrong strand in alignment	68
Missing source structure	0
Structure for GI exists	0
Prediction crosses kingdom	21
Insignificant alignment	6,910
Resulting structure too short	3
Not a valid ITS2 structure	229
Less than 4 helices transferred	18,288
Total	54,536

Table 2.11: Results from second homology modelling run, rejected sequences or structures

2.2.5 Comparative Analysis

To evaluate differences between the old database and the new database, counts for valid structures in both databases were retrieved and compared (total numbers and itemized by structure methods, see tables 2.12 on the next page and 2.13 on the following page).

A comparative analysis of the entire database rebuild was conducted with two experimental settings: First, a use of GeneMatcher and BLAST identity matrices instead of ITS2PAM50 (settings for GeneMatcher: gap open -15, gap extension -2; settings for BLAST: gap open -10, gap extension -2). Second, usage of mFold (Zuker, 1989) instead of Vienna's RNAfold (Hofacker et al., 1994). Data shown in tables 2.14 on the next page and 2.15 on page 28 were taken from the front-end statistics page. An apportioned view can be found in tables 2.16 on page 28 and 2.17 on page 28.

Structures in old database	27,410
Structures also in new database	24,962
not in new database	2,448

Structures in new database	65,326
Structures also in old database	24,962
not in old database	40,364

Table 2.12: Total number of structures in old and new database, not including partial structures

	1	2	3	4	5	6	7
1	4,976	0	0	0	0	0	0
4	1,179	13,027	862	26	16	90	305
7	203	1,445	1,610	22	3	138	700
9	2	63	36	93	109	392	34
11	3	62	39	90	46	430	68

Table 2.13: Number of structures per method in the old database (structure methods on the left) and the new database (structure methods on top)

Method	orig	ident	mFold
1: MFE fold	9,883	9,883	34,287
2: HM first iteration	25,900	20,984	25,197
3: HM second iteration	8,978	4,863	5,230
4: MFE fold after BLAST	4,333	4,967	12,930
5: HM first iteration after BLAST	1,477	1,548	1,660
6: HM second iteration after BLAST	14,755	10,065	10,082
7: Partial structures	11,395	5,370	7,557
Total	76,721	57,680	96,943

Table 2.14: Structures found in comparative analysis, grouped by structure discovery method

Method	orig	ident	mFold
by Genbank	52,913	39,269	68,767
large end cutting	3,243	1,831	3,504
BLAST: cutting/enlarging/sliding	6,355	5,640	5,903
BLAST: without previous annotation	14,210	10,940	24,672
Total	76,721	57,680	96,943

Table 2.15: Structures found in comparative analysis, grouped by feature discovery method

Structures in mFold-derived database	96,943
Structures also in default database	74,978
not in default database	21,965

Structures in default database	76,721
Structures also in mFold-derived database	74,978
not in mFold database	1,743

Table 2.16: Total number of structures in mFold-derived and default (RNAfold-derived) database, partial structures included

	1	2	3	4	5	6	7
1	9,804	12,786	2,289	464	154	1,361	3,819
2	66	12,052	4,148	94	42	1,047	2,758
3	1	587	1,802	11	5	309	565
4	0	30	82	3,460	885	3,873	344
5	0	14	22	24	292	652	28
6	0	104	130	102	55	6,108	266
7	12	278	416	26	15	165	3,431

Table 2.17: Number of structures per method in the mFold-derived database (structure methods on the left) and the RNAfold-derived database (structure methods on top)

3 Alternative Methods of ITS2 Secondary Structure Discovery

3.1 Materials and Methods

Perl with BioPerl and DBI (cf. section 2.1 on page 12 et seq.) was used as the scripting language for the evaluation scripts. RNAeval from the Vienna RNA package, version 1.6.1 (Hofacker, 2003) was used for energy-based quality comparison. GNU R (R Development Core Team, 2006) was used for data visualization.

3.1.1 RNAshapes

RNAshapes (Steffen et al., 2006) is an RNA structure discovery software which is based on structural models, called shapes — abstract representatives of structures (Giegerich et al., 2004; Reeder and Giegerich, 2005). According to the authors, this approach integrates well with dynamic programming and offers computation with a reasonable speed as opposed to the original Sankoff algorithm (Sankoff, 1985).

3.1.2 mFold

MFold by Zuker (1989) is, just as RNAfold (Hofacker et al., 1994) an MFE-based method of RNA secondary structure prediction. Suboptimal folds are computed by a sampling method:

1. Each possible base pairing is held constant.
2. For each, the MFE structure is calculated.
3. Structures that are too similar are purged a posteriori by a distance criterion.

3.2 Results

3.2.1 Evaluation protocols

3.2.1.1 Energy-based evaluation

The method employed for comparing discovered structures for mFold and RNAshapes is the same: MFE and HM structures from the database that possess the original GenBank annotation constitute the reference set, structure data calculated from each other method constitutes the experimental set; reference and experiment structures are compared by their free energies. The theory of operation of the evaluation protocol consists of four steps:

1. Retrieve all sequences with their valid ITS2 structures and energies from the ITS2 database as a reference dataset.
2. Retrieve all sequences with a valid ITS2 structure from the ITS2 database as experimental dataset.
3. Run RNAshapes[1] respectively mFold[2] on the experimental dataset.
4. Run the evaluation script. Its output is a table consisting of

[1]Parameters: `-a -t 5 -m [][][][] -u -c 10.0 -M 30 -# 100` (shape folding; most abstract = highest shape specifity; match four-helix shape pattern; ignore unstable structures; set energy range to 10%; set maximum loop length to 30; print only the first 100 structures)
[2]Parameters: `P=30` (suboptimal search space 30% below MFE structure)

a) (in case of RNAshapes) GI (integer), energy from database (float), structure discovery method from database (integer), energy from best shape (float).

b) (in case of mFold) GI (integer), structure discovery method from database (integer), is MFE structure and valid (boolean), is first valid suboptimal structure (boolean), is non-first valid suboptimal structure (boolean), energy from RNAeval call (float), energy from ITS2 database (float).

3.2.1.2 Further evaluation of suboptimal folds

MFold returns multiple structures per sequence. Evaluation protocols as two Perl scripts were implemented: (I) The number of valid ITS2 secondary structures per fold (II) The overall difference of the resulting structures measured by a simple distance criterion between the thermodynamically best fold to all other valid structures. The distance measure was 2 for mismatches, defined as bracket open in the best structure versus bracket close in each other structure et vice versa and 1 for putative mismatches, defined as bracket open or close in the best structure versus an unpaired base.

3.2.2 RNAshapes

Correctly folded ITS2	40,531
Incorrect or non-computable structures	12,379
Total	52,910

Results from mFold were compared on the basis of free energy. Nonlinear regression were calculated.

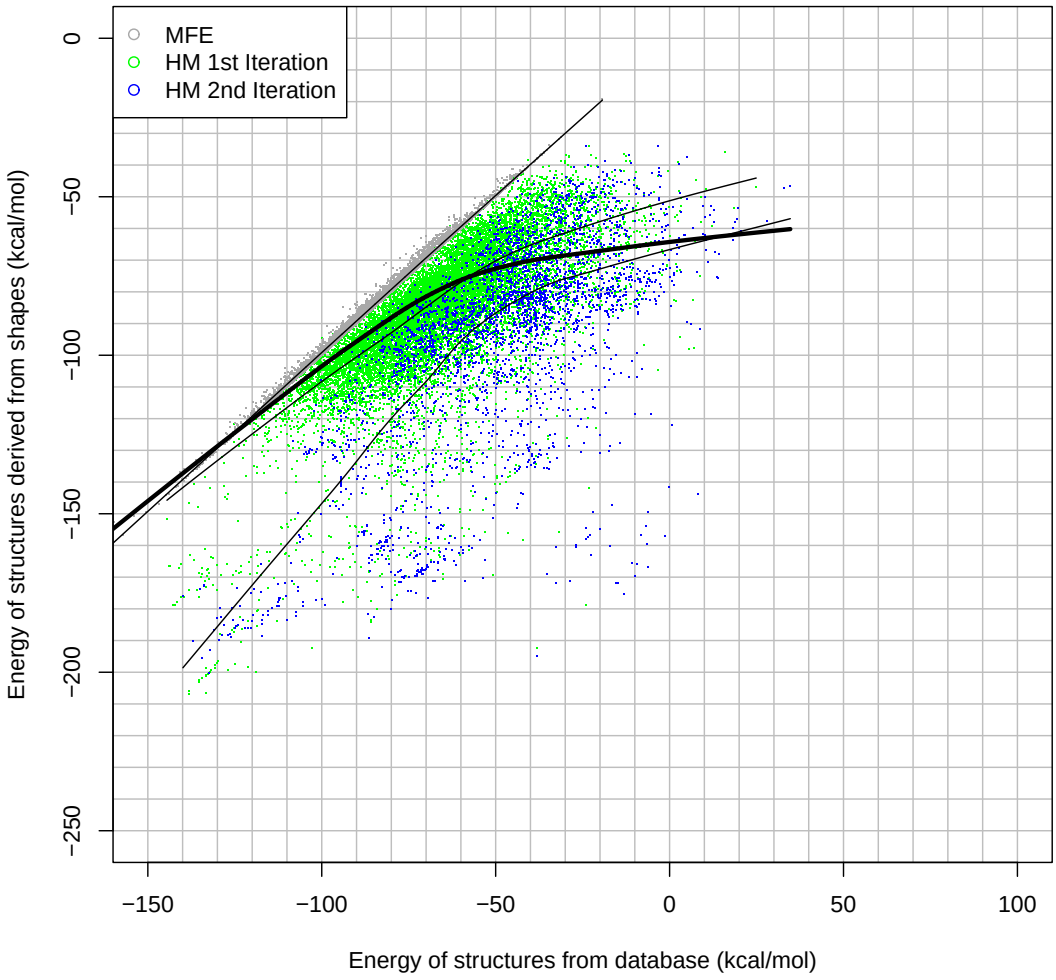


Figure 3.1: Scatterplot of energies from RNAshapes-derived structures versus energies of the structure for the same sequence in the ITS2 database

3.2.3 mFold

Overall and apportioned statistics for all folds on 52,910 sequences (table 3.1 on the following page) and valid folds (tables 3.2 on the next page and 3.3 on the following page) were compiled.

Valid folds	276,824
Invalid folds	2,055,168
$\sum$	2,331,992

Table 3.1: Overall statistics from mFold evaluation

Database method	Valid structures	unique sequences
1	205,439	9,801
2	55,911	12,091
3	5,460	1,876
7	10,014	3,244
$\sum$	276,824	26,992

Table 3.2: Overall statistics from mFold evaluation

mFold position	database method				
	1	2	3	7	$\sum$
MFE	8,246	384	18	52	8,700
first suboptimal	9,798	12,091	1,876	3,224	26,989
other suboptimals	187,395	43,436	3,566	6,738	241,135
$\sum$	205,439	55,911	5,460	10,014	276,824

Table 3.3: Counts from mFold evaluation, only valid structures included

Results from mFold were compared on the basis of free energy. Nonlinear regressions were calculated (see figure 3.2 on the next page). The number of valid ITS2 structures were plotted as a histogram (see figure 3.3 on page 36; data: 26,992 secondary structures, containing at least one correct ITS2, with a total of 276,824 valid ITS2 structures, mean 10.26, median 5, maximum 48). The distances applied to sequences with at least two correct ITS2 were plotted as a box-whisker plot (see figure 3.4 on page 36; data: 23,894 secondary structures, minimum distance: 0, mean: 2.9, median: 2, maximum: 101).

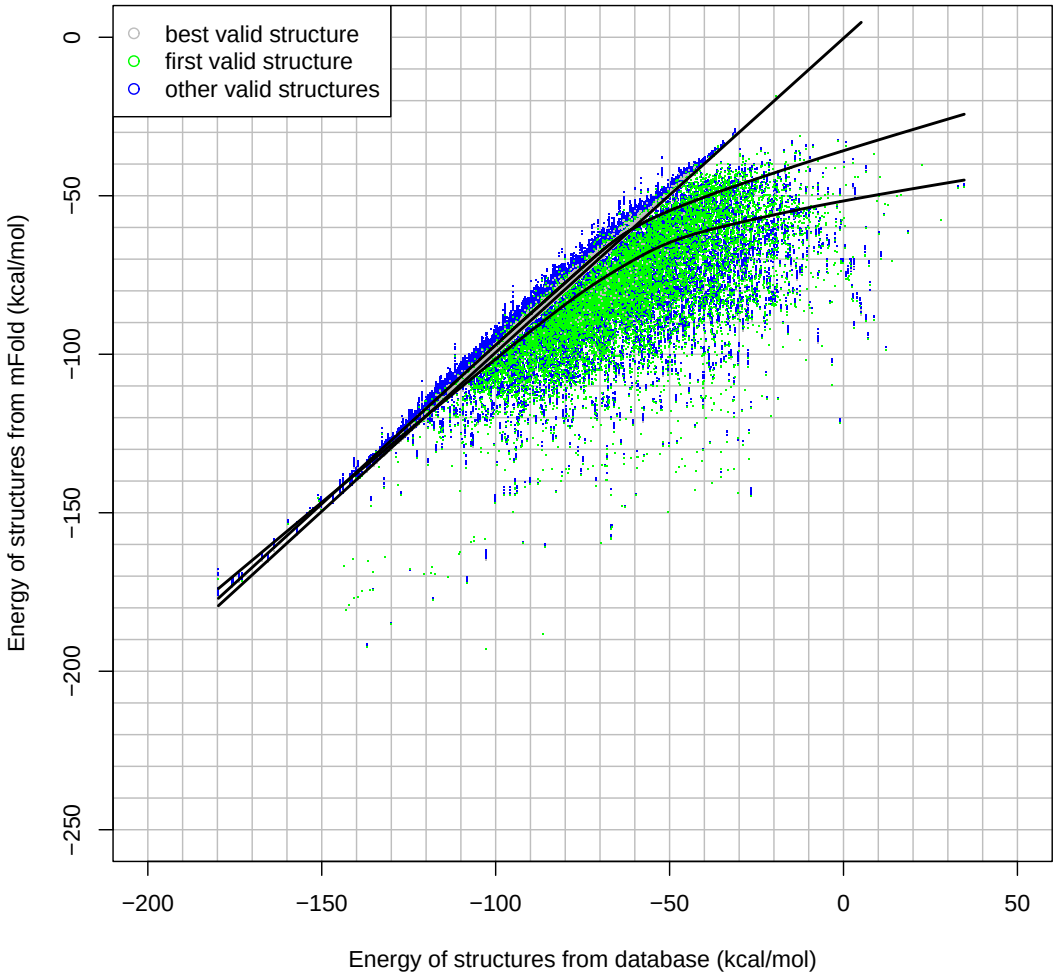


Figure 3.2: Scatterplot of energies from mFold-derived structures versus those ITS2 database versus energies of the structure for the same sequence in the ITS2 database

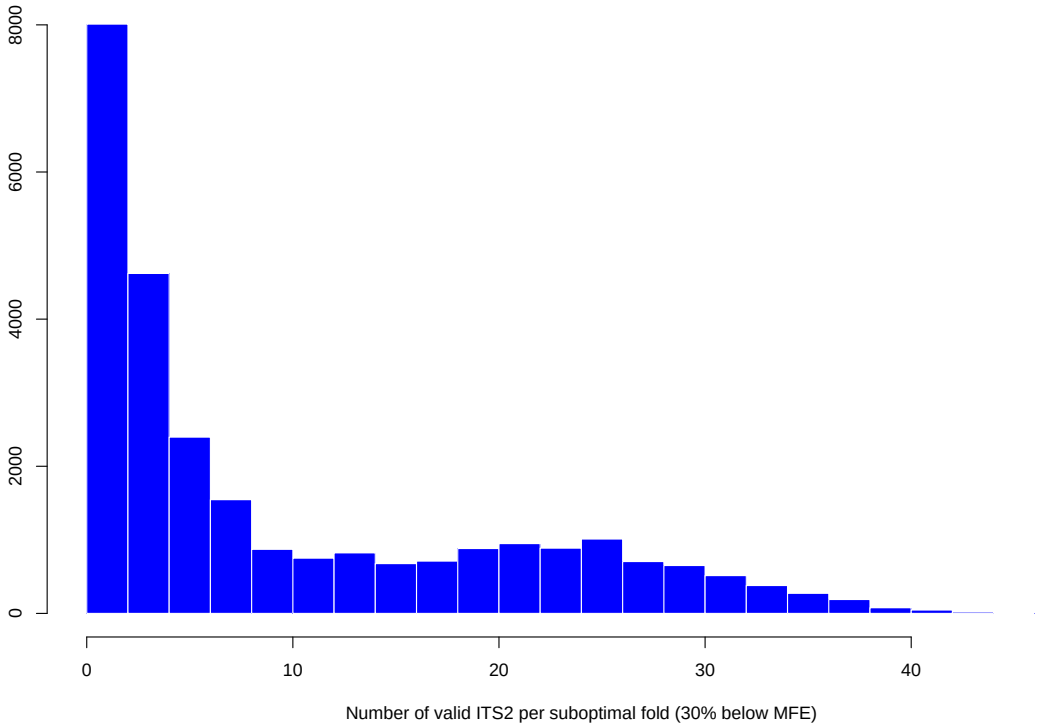


Figure 3.3: Number of valid mFold-derived ITS2 structures per folded sequence

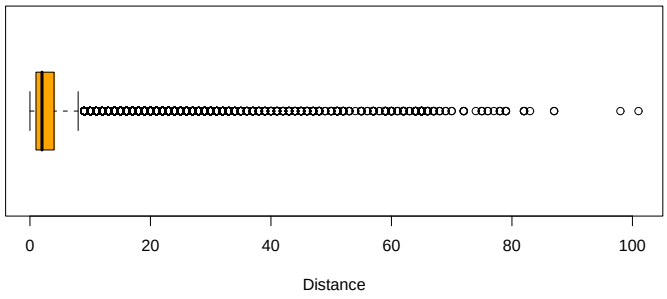


Figure 3.4: Distances of valid best free energy structure from each fold versus other valid structures in the same fold

4 A Case Study: Phylogeny of Placozoans

In nature's infinite book of secrecy A litte I can read.
– William Shakespeare, „Antony and Cleopatra"

4.1 Introduction

Trichoplax adhaerens SCHULZE 1883 is classified in its own phylum, Placozoa (Grell, 1971). Often described as the simplest known animal (cf. Miller and Ball, 2005), *Trichoplax* is a marine amoeba-like organism with a size of up to three millimeters. Its autapomorphic characters comprise (a) of contractile fiber cells in the body's middle layer, (b) cilia of ventral cylinder cells with two horizontal ciliar roots in addition to the vertically upright main root and (c) a lack of extracellular matrix and collagens (Ax, 1995). Despite the fact that *Trichoplax* has no recognizable body axes, it has two different epithelia one of which points to the ground and is seen as a precursor of body asymmetry. Genes associated with this basic bauplan have homologs in *Trichoplax*, such as Brachyury and Tbx2/3 (Martinelli and Spring, 2003), Not (Martinelli and Spring, 2004), Trox-2 (Jakob et al., 2004) and Pax-B (Hadrys et al., 2005).

Though apparently *Trichoplax* can reproduce sexually, successful mating has never been achieved under laboratory conditions (Miller and Ball, 2005).

The phylogenetic placement is under debate (see Ender and Schierwater, 2003, for a set of references therein). The most recent publication places

them as the most basal lower metazoan phylum (Dellaporta et al., 2006). Despite its unclear overall phylogenetic position, new sequencing efforts — including the internal transcribed spacer 2 (ITS2) region — of placozoans from different global locations suggest that this phylum consists of more than one species (Voigt et al., 2004).

Coleman and Vacquier (2002) found an interesting correlation: „When sufficient evolutionary distance has accumulated to produce even one CBC in the relatively conserved pairing positions of the ITS2 transcript secondary structure, taxa differing by the CBC are observed experimentally to be totally incapable of intercrossing" (see also Coleman, 2000, 2003). A large scale study analysing this correlation is in the works (Müller et al., 2007); first results indicate that a lack of CBCs in ITS2 secondary structures is not an indictator of two organisms belonging to the same species. However, at least one CBC is a statistically significant indicator that predicts two organisms belonging to distinct species. This classifier works highly accurate in 91% in all cases.

Mayr (1942) put forward a famous definition – or better, an indicator hypothesis – that species are „groups of interbreeding natural populations that are reproductively isolated from other such groups" (Mayr, 1996). Given that CBCs in ITS2 secondary structures are found to correlate strongly with distinct species, one can use this molecular indicator for giving the minimal number of species from a set of secondary structures. This new criterium should be independent of reproduction.

The question in regard to placozoans is whether a CBC analysis can divide the available ITS2 secondary structures into two ore more significantly distinct groups.

4.2 Materials and Methods

In GenBank (Benson et al., 2007), 48 entries for the query *Trichoplax AND „internal transcribed spacer 2"* were retrieved, including one entry whose

ITS2 feature was defined in a non-standard manner. Of these, 34 *Trichoplax adhaerens* full secondary structures have been found in and downloaded from the ITS2 database (Wolf et al., 2005a). 12 partial secondary structures and the non-standard defined sequence had – as evidenced by a CLUSTAL W alignment – a high similarity to the original GenBank entry (Accession U65478) so that a model of this entry was created by the ITS2 database prediction facility from the best hit. The model's third and fourth helix were corrected with the structural alignment and editing tool 4SALE (Seibel et al., 2006) and clues from partial MFE folding with RNAfold (Hofacker et al., 1994; Hofacker, 2003). All 13 sequences could be modelled in high quality with the ITS2 database custom modelling service (parameters: ITS2PAM50 matrix as published in Wolf et al., 2005a, 75% transfer percentage).
Through these steps, high quality structural models for all 48 sequences were determined (see table 4.1 on the following page for numbers, GenBank accession numbers and structure source).

Number	GenBank Accession Numbers	Secondary Structure Source
8	AY652557, AY652558, AY652559, AY652560, AY652573, AY652574, AY652575, AY652576	ITS2 Database, MFE structure
26	AY652543, AY652544, AY652545, AY652546, AY652547, AY652548, AY652549, AY652550, AY652551, AY652552, AY652553, AY652554, AY652555, AY652556, AY652561, AY652562, AY652563, AY652564, AY652565, AY652566, AY652567, AY652568, AY652569, AY652570, AY652571, AY652572	ITS2 Database, Homology Modelling
1	U65478	Predict from AY652563, 4SALE, RNAfold
13	AY652530, AY652531, AY652532, AY652533, AY652534, AY652535, AY652536, AY652537, AY652538, AY652539, AY652540, AY652541, AY652542	ITS2 Database, Custom Modelling

Table 4.1: Sequences and structures for *Trichoplax* from the ITS2 database

The resulting 48 secondary structures were aligned with 4SALE, minor visible errors in secondary structures corrected. The 4SALE alignment uses CLUSTAL W with pseudo-protein coding that combines sequence and structure and a specific scoring matrix from the general time-reversible model as described by Müller and Vingron (2000) and Müller et al. (2002). The CBC matrix of the alignment was exported and used as the dataset for further computation (see figure 4.1 on the next page). GNU R (R Development Core Team, 2006) was used for matrix visualization. For the sake of further discussion, a „pure group" is defined as a group of secondary structures who do not have CBCs to each other but to all others not in that group.

SplitsTree (Huson, 1998) was used for plotting neighbor net. TreeIllustrator (Trooskens et al., 2005) was used for drawing a CBC tree.

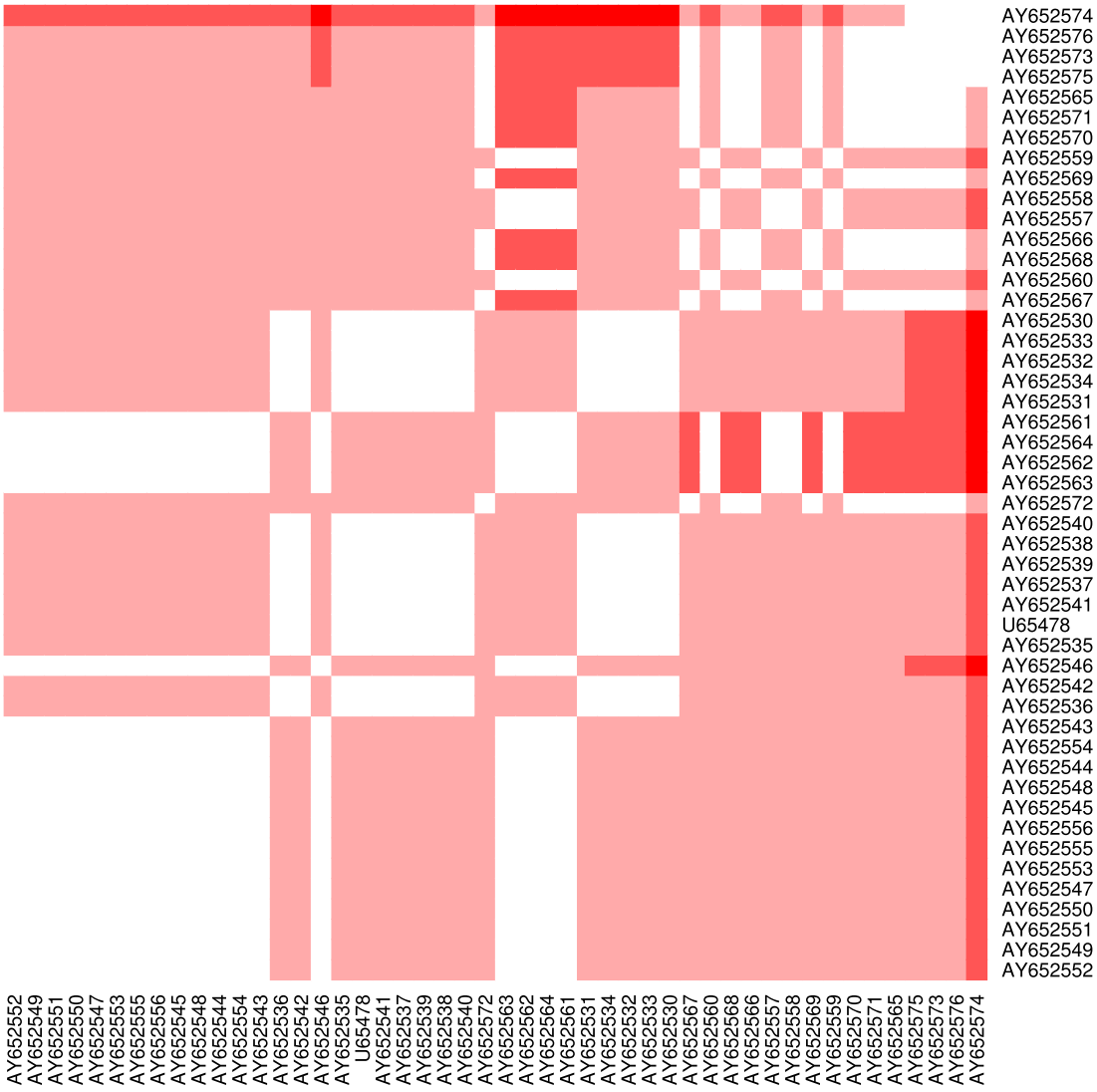


Figure 4.1: Symmetrical CBC matrix for all placozoans hierarchically clustered (white: no CBCs, dark red: 3 CBCs)

4.3 Results

The cluster analysis on the adjacency matrix revealed that there are three pure groups:

- Group I: AY652543, AY652544, AY652545, AY652546, AY652547, AY652548, AY652549, AY652550, AY652551, AY652552, AY652553, AY652554, AY652555, AY652556
- Group II: U65478, AY652530, AY652531, AY652532, AY652533, AY652534, AY652535, AY652536, AY652537, AY652538, AY652539, AY652540, AY652541, AY652542
- Group III: AY652565, AY652566, AY652567, AY652568, AY652569, AY652570, AY652571, AY652572

While group II is definitely pure, group I may either contain four additional sequences (Ia: AY652561, AY652562, AY652563, AY652564) or these belong to a separate group together with four others (Ib: AY652557, AY652558, AY652559, AY652560). Group III has an intersection of three sequences (IIIa: AY652573, AY652575, AY652576) with sequence AY652574 (IIIb).

The eight H groups as defined by Voigt et al. (2004) mostly fit the groups found in figure 4.4 on page 46. The biogeography of the probes does not strictly follow the phylogenetic groups, supporting the notion that placozoans are cosmopolites. This may be due to their size (Finlay, 2002).

4.4 Conclusions

From the cluster analysis (see figure 4.2 on the next page), it is revealed that group II is pure. Thus it can be safely said that the placozoans consist of at least two species. Groups I and III are also pure but have close neighbors that may belong to other groupings. We can conclude that Group Ia either belongs to Group I or Group Ib; that Group IIIa belongs to Group III or

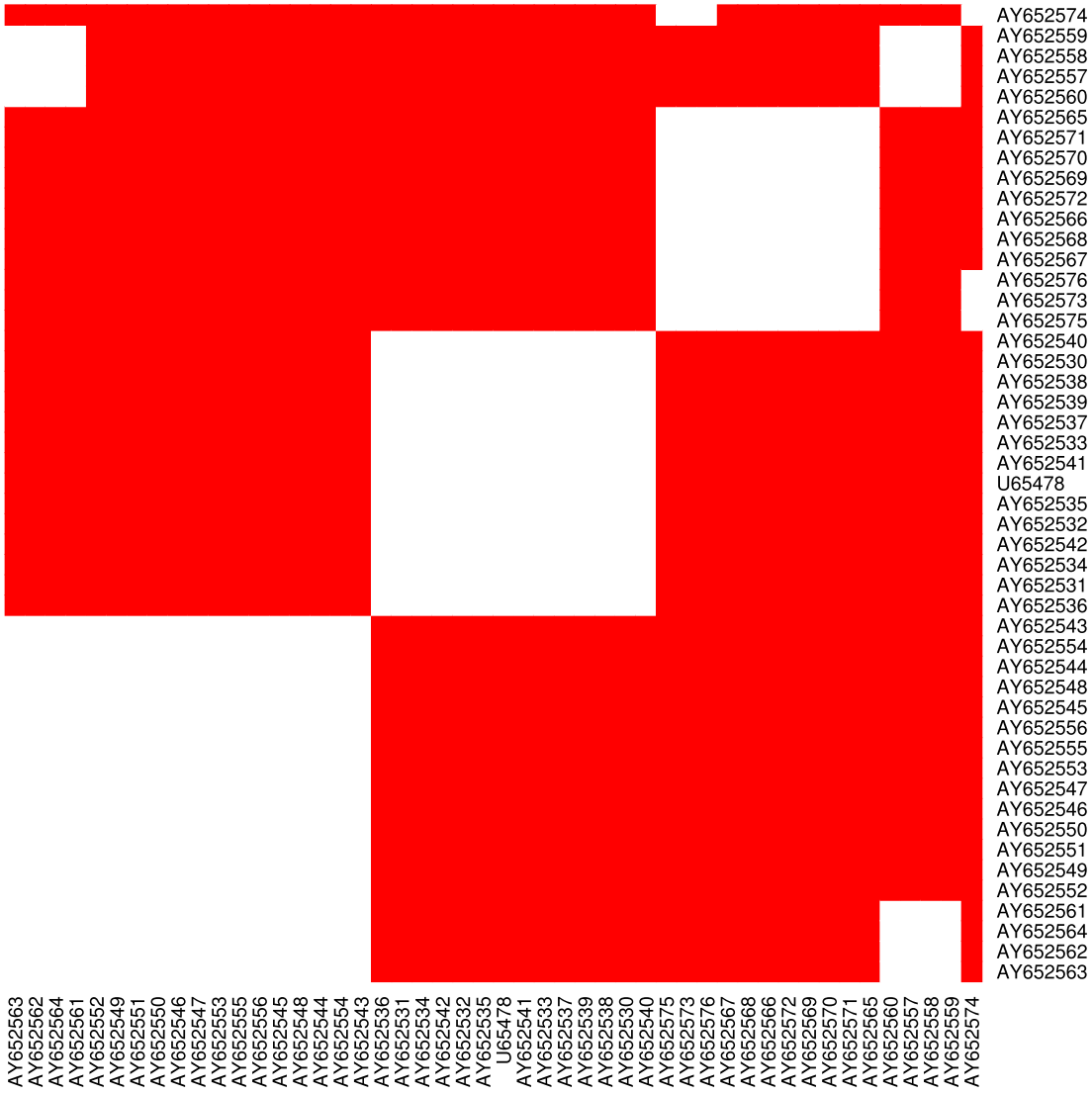


Figure 4.2: Symmetrical CBC adjacency matrix all placozoans as a hierarchical cluster (white: no CBCs, dark red: at least one CBC)

to Group IIIb; or that Groups Ia and Ib as well as IIIa and IIIb constitute distinct species. Given that the three clusters are sufficiently distinct, it can be well claimed that at least three species are a likely scenario, which is also supported by the SplitsTree (see figure 4.3 on the following page) and a distance tree (see figure 4.4 on page 46). Both trees show a tendency towards three main branches, with the SplitsTree opening up less likely yet not rejectable alternatives on a number of sequences.

From this initial analysis it is evident that *Placozoa* is indeed „no longer a phylum of one" (Voigt et al., 2004). It is a phylum of at least two species, quite probably three or even more. As has been noted before, while sexual reproduction has not been achieved in the laboratory, molecular phylogenetic analysis made this discovery possible. A morphological approach would take much longer and would be more laboursome.
The inferences from secondary structure data are not a unidirectional procedure: the hypothesis put forward can now be tested for on a morphological level.

The ITS2 database and 4SALE in conjunction with standard phylogenetic utilities are viable tools for phylogenetic analysis. Manual intervention (besides editing the alignment) was only needed once to handcraft a structural template in the step of modelling the remaining sequences.

Of course, this analysis is only a preliminary result. Further research in vivo and in silico will have to fortify the findings.

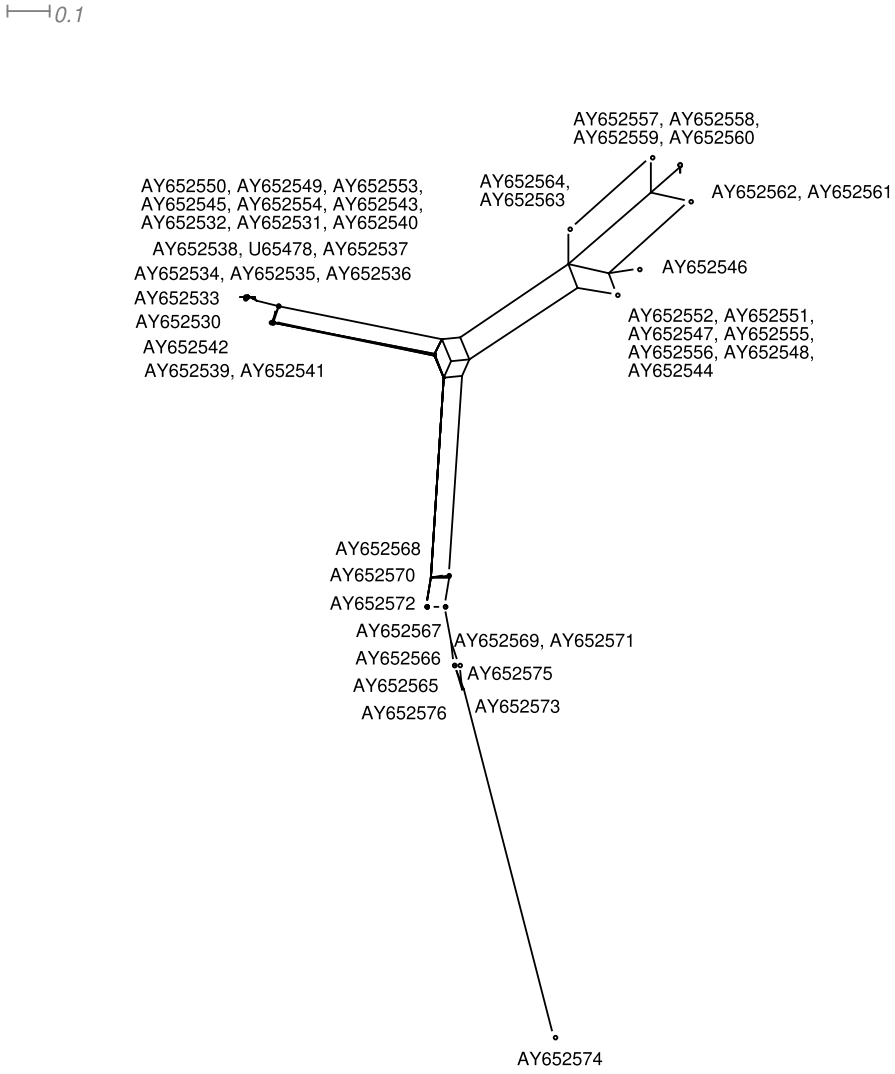


Figure 4.3: SplitsTree on the matrix from figure 4.1 on page 41

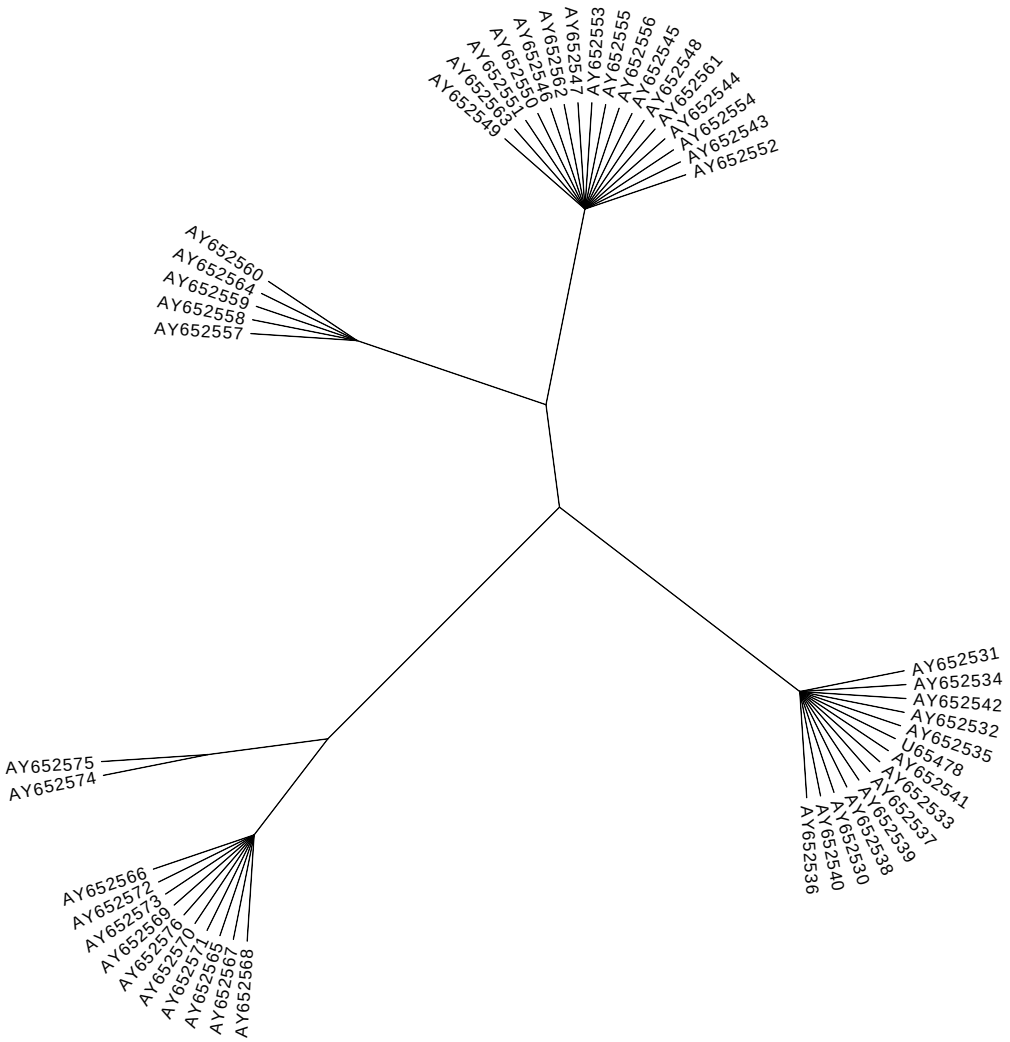


Figure 4.4: Distance tree on the matrix from figure 4.1 on page 41

5 Discussion

5.1 Database rebuild

5.1.1 Technical considerations

The new code shows certain features that do have direct influence on the practicality of the ITS2 database.

The ITS2 validity checker *its2check* is a simple but powerful program. It is well fit for large-scale analysis from high-output discovery algorithms. It is also used by J. Jablczynski for evaluating an evolutionary model of ITS2 (personal communication).
The objective was not to just replace the existing code used by Schultz et al. (2005) but to enhance speed for later use in the evaluation of mFold suboptimal structures (see section 3.1.2 on page 29).

The rebuild code for the database itself is stable, robust and understandable. It rejects wrong parameters and checks for a number of common problems that appear when building upon sequences from primary databases. These problems include checking on definitely erroneous or missing taxonomy, formatting errors of GenBank entries or annotations not unambigously understandable (i.e. mixed strands).

Error messages were designed to be helpful in detecting the source of unwanted behavior. Each tool has a short help that can be accessed from the command line. Log output is printed by every single tool, so a GenBank entry can be traced on its way through the various processing steps.

Another valuable feature of the rebuild code is its reusability. It can be used for rebuilding and updating the database by using different parameters. Thus no code duplication for these seemingly distinct processes has taken place, except for the two separate protocol scripts that call the tools with different parameters.
Given its simple way of initiation, comparative experiments can be conducted on a large scale, which might include checking the quality gain of new matrices and other alignment parameters or the use of other algorithms. The use of mFold instead of RNAfold for ab initio structure discovery (discussed later) was easily implemented with only minor modifications to the rebuild system.

The front-end of the database has also advanced. Modifications to its look-and-feel are a twist towards intuitive interfaces, as inspired by databases such as SMART (Letunic et al., 2006) and PFAM (Finn et al., 2006). Scientific information on the database and its build process was included as static information. The interface now makes data visible to the user that were indiscernible before. Two advantages arise when breaking down this wall of opacity: the user can grasp the structure of and the process behind the database, while the developer has an implicit double-check code that requires all data in the database to be consistent so that no display errors appear (at least none that stem from wrong data).
Furthermore, the custom modelling facility is the first step towards a completely user-driven homology modelling process. Though the homology modelling method has so far only been published and validated for ITS2, there is no compelling reason why the process shouldn't be usable for other RNA structures.

5.1.2 Biological Aspects

When changing the model for ITS2, even when it is only done on the implementation level, logic requires that all entries found in the old database should be found in the new database. For the 5,092 Vienna folded structures

as published by Schultz et al. (2005), this was achieved with 5,065 of them found in the new database (26 were replaced by NCBI with updated entries, 1 was retracted by the author). The new database was able to cover the old one very good (see tables 2.12 on page 27 and 2.13 on page 27).

The large numbers in the Results section compared to the original work by Achtziger (2005) may irritate at first: Why were so many structures not found before? Some reasons can be considered more or less obvious:

- The ITS2 specific nucleotide scoring matrix which was published by Wolf et al. (2005a) was computed from ITS2 alignments, but not used for them, as this methodical iteration was not possible at that time. A comparison of the database built with and without the ITS2PAM50 maxtrix show a structure discovery gain of $> 20\%$ (see tables 2.14 on page 27 and 2.15 on page 28).

- The use of BLAST (Altschul et al., 1990, 1997), more precisely WU-BLAST (Gish, 2004) instead of CLUSTAL W (Thompson et al., 1994) provides a simple and well-tested fast method for finding regions of significant similarity. This process implies that only sequences which are so similar to existing ones are used as new input sequences for folding and homology modelling. Despite this high quality requirement, for 36% of sequences that had no previous annotation an ITS2 including structure could be found, compared to 45% of annotated features. Another argument for its quality is that for both annotated and BLAST-discovered ITS2 regions, the percentage of MFE structures amounts to about 10.

- The inclusion of partial structures also contributed to database size — which should not be taken as a comparison against the old database, as it did not include such structures.
  ITS2 being not (yet) the most important phylogenetic marker, it is often nothing more than an annotated sequencing remnant. Another problem is that of helix quality, where helix IV (or sometimes also helices I or II) consists of such a low number of base pairs that even one or two

base changes or alignment circumstances can plunge the transfer percentage below 75%. While this is unfortunate for users of ITS2, there is no reason why those parts of the sequence for which a homology modelling can be calculated shouldn't be included in the database.

Despite these improvements, there is one significant underlying issue that will be observed every time the database is rebuilt in a way that require multi-step processing such as homology modelling. While the first build of the database is always consistent, updates contribute new MFE structures that, in hindsight, could have been templates for other sequences that already exist in the database. In theory, this could only be overcome by rebuilding the database for every new sequence available on GenBank. This is not possible for two reasons:

- The build process is computationally expensive. Computing all structures currently takes more than 2 days, notwithstanding database preparation, sequence retrieval, analysis and taxonomy transfer. Alignments can only be done fast because of the Paracel GeneMatcher supercomputer. If the latter was not available, rebuilding the database would only be a very rare luxury. Even if doing such a recomputation regularly, the machine would be blocked for other scientific uses.
- Scientists relying on the database work on a specific dataset. If this dataset were to change regularly with identifiers staying the same while changing the associated structures (or even sequences), results could not be replicated independently or the database would require a complex versioning system.

Furthermore, always including all sequences that do not have a structure from the database in the update runs is computationally expensive, also feeding potentially wrong or very short sequences to the alignment process over and over again.

With the update process, it is indeed possible that to a low degree, biological inconsistencies may be introduced. Let's presume A is an MFE structure,

B is homology modelled from A and C is homology modelled from B. Inconsistency is created whenever a new structure D is found in the update process that predicts the structure of C better than B (or D predicts predicts B better than A). While this may not be pure from a logical perspective, it is acceptable due to the high quality standards that have to be met by new structures.

The only practical measure to overcome (or more precisely, circumvent) these practical problems is doing a full recomputation of database on a regular basis while keeping the old datasets available.

5.2 Alternative Methods

RNAshapes and mFold delivered a sufficent number of correct ITS2 folds that can be used as a comparison to the currently used methods.

RNAshapes first seemed an interesting contender for fast and accurate structure discovery. Given its novel way of non-thermodynamical folding, it circumvents the folding-space problem. Yet, having neither a root in physics nor in evolution, RNAshapes computes structures that may be thermodynamically better than homology modelled ones (see figure 3.1 on page 32), but are not biologically viable. This includes introduction of large unpaired regions between or inside of helices and other effects as long as the formal criteria are satisfied.
With regard to the case study, according to Coleman (2003) CBC analysis can only take place from conserved positions in the structure, which makes the homology modelling process a more natural data source than shape folding.
Yet RNAshapes cannot be fully dismissed. Further comparative analysis with tools such as RNAdistance (from the Vienna Package, see Hofacker et al., 1994; Hofacker, 2003) on coarse grained weighted structures (Shapiro, 1988; Margalit et al., 1989) and basepair-identity statistics might assist in this venture.

MFold from Zuker (1989) principally could have been used for the database and would have revealed a larger number of structures than RNAfold (see tables 2.14 on page 27 and 2.15 on page 28). Yet questions remained: First, is mFold better than RNAfold because one or a handful possible ITS2 structures are not the MFE fold and instead are hidden in the suboptimal space; or are there a host of valid ITS2 structures for each fold. A histogram of valid ITS2 structures per fold (see figure 3.3 on page 36) revealed that there are many valid ITS2 per fold. Second, if there are many valid ITS2, do they have significantly different conformations that could introduce errors in sequence/structure alignment. By using a simple distance criterion (see figure 3.4 on page 36), it was shown that most ITS2 found in suboptimal folds are largely of the same structure.

Still there are technical aspects and drawbacks that make mFold an unsatisifying option:

- It is slow. Rebuilding the database with mFold instead of RNAfold is four times slower. Of course, Zuker's implementation cannot be blamed for this, it is a direct result of the algorithm.
- It does not integrate well in automated processes, as it does not conform to the underlying operating system philosophy (no Unix pipe support, wrong output devices, ...).
- The Vienna bracket-dot-bracket format is not yet supported and no converter is available.

6 Conclusions

RNA homology modelling proves to be a viable method of structure discovery. In the past it was shown that thermodynamics is not always the best means (Doshi et al., 2004). For practical use, own secondary structure models should be designed from the data made available through the ITS2 database, but not blindly accepting its contents. Homology modelling quality is influenced by experience, tertiary structure knowledge (if available) and manual alignment control (Gutell et al., 1992). As CBCs occur mostly in conserved secondary structure positions (Coleman, 2003), they may be a guide and help in setting modelling constraints.
Data from the comparative rebuild run and alternative methods evaluation make it recommendable to use mFold as soon as aforementioned mFold-specific problems are solved. Also, the new dataset should be checked on whether results from mFold need different quality filters. Speed could be improved by using a computing cluster. Integration can achieved with wrapper scripts and converters. As soon as this is done, mFold should be used for the next version of the ITS2 database.

The case study showed that the ITS2 database is a good starting point for phylogenetic analysis. 4SALE (Seibel et al., 2006) proved useful for aligning and editing secondary structures. CBC analysis offers a number of vistas for phylogeneticists.

The ITS2 database is a foundation for more work in the field of computional RNA biology. It integrates thermodynamics, evolution and structural biology. The reliable dataset it provides is a starting point and a reference set for further research.

7 Perspectives

Though the ITS2 database is now vastly improved and may be said to be close to maximum discovery rate, there are always points that can be worked on on a varying timescale:

- Müller (personal communication) created matrices that weigh sequence and structure against sequence. This is closer to the envisioned process than the currently used matrix. Furthermore, Karlin-Altschul parameters for BLAST have not yet been determined.
- The ITS2 database was not heavily optimized. The main loss of speed on the web interface side mostly can be traced back to the nested set method used for the taxonomy. Though faster than implementing recursive queries on the software side, there are probably better options.
- In the future, two projects may be desirable:
  - A generalized RNA homology modelling package that can handle more complex RNA secondary and supersecondary structure features. A good software package might be of great help in this usually laboursome process and allow for phylogeny on secondary structure of more complex RNA molecules. This is quite natural as homology modelling has been used in protein world for a long time.
  - The most interesting piece of software, replacing the former, would be one that can take a given secondary structure and try to fit a sequence „as good as possible" into it. What „good" in this context means has not yet been determined.

8 Acknowledgements

This work would not have been possible without a number of people that I would like to mention personally. My supervisors have been guides throughout the whole work on this thesis. Prof. Jörg Schultz' clear vision on large scale analysis, data retrieval and organization and algorithms empowered me to work fearlessly with giant datasets. Dr. Matthias Wolf was the biological conscience of the venture, giving meaning to the eight characters[1] the entire database revolves around. Dr. Tobias Müller was always helpful when it came to questions of statistical evaluation and data visualization. Prof. Thomas Dandekar managed to set up a comfortable work environment at the department. My friendly colleagues never minded complicated questions as well as spontaneous outbreaks of happiness and desperation on my side.

I'd like to thank all of them very much.

Last but not least I'd like to thank the anonymous proofreaders and anyone who didn't keep me from working on this thesis in a more than necessary manner.

[1] A U C G ( ) . -

A Summary

The internal transcribed spacer 2 (ITS2) of the ribosomal gene repeat is an increasingly important phylogenetic marker whose RNA secondary structure is widely conserved across eukaryotic organisms. The ITS2 database aims to be a comprehensive resource on ITS2 sequence and secondary structure, based on direct thermodynamic as well as homology modelled RNA folds.

Results: (a) A rebuild of the original ITS2 database generation scripts applied to a current NCBI dataset reveal more than 60,000 ITS2 structures. This more than doubles the contents of the original database and triples it when including partial structures. (b) The end-user interface was rewritten, extended and now features user-defined homology modelling. (c) Other possible RNA structure discovery methods (namely suboptimal and shape folding) prove helpful but are not able to replace homology modelling. (d) A use case of the ITS2 database in conjunction with other tools developed at the department gave insight into molecular phylogenetic analysis with ITS2.

B Zusammenfassung

Der internal transcribed spacer 2 (ITS2) des ribosomalen Genrepeats ist ein zunehmend wichtiger phylogenetischer Marker, dessen RNA-Sekundärstruktur innerhalb vieler eukaryontischer Organismen konserviert ist. Die ITS2-Datenbank hat zum Ziel, eine umfangreiche Ressource für ITS2-Sequenzen und -Sekundärstrukturen auf Basis direkter thermodynamischer als auch homologiemodellierter RNA-Faltung zu sein.

Ergebnisse: (a) Eine komplette Neufassung der ursprünglichen die ITS2-Datenbank generierenden Skripte, angewandt auf einen aktuellen NCBI-Datensatz, deckte mehr als 65.000 ITS2-Strukturen auf. Dies verdoppelt den Inhalt der ursprünglichen Datenbank und verdreifacht ihn, wenn partielle Strukturen mit einbezogen werden. (b) Die Endbenutzer-Schnittstelle wurde neu geschrieben, erweitert und ist jetzt in der Lage, benutzerdefinierte Homologiemodellierungen durchzuführen. (c) Andere möglichen RNA-Strukturaufklärungsmethoden (subop-timales und formenbasiertes Falten) sind hilfreich, können aber Homologiemodellierung nicht ersetzen. (d) Ein Anwendungsfall der ITS2-Datenbank in Zusammenhang mit anderen am Lehrstuhl entwickelten Werkzeugen gab Einblick in die Verwendung von ITS2 für molekulare Phylogenie.

C List Of Abbreviations

CBC compensatory base change

MFE minimum free energy

HM homology modelling

ITS2 internal transcribed spacer 2

D Input, Parameters, Output and Data Formats of Custom Programs

D.1 ITS2 Validity Checker

Call its2check

Theory of Operation Takes extended FASTA input with multiple structures per sequence allowed and checks each sequence-structure pair for four helices, the third being the longest and the UU mismatch and UGGU motifs.

External Calls None.

Input Extended FASTA through Unix standard input.

Parameters

0: print original FASTA, but only those structures which are valid ITS2.

1: print only statistics for all structures in the supplied data.

2: print original FASTA with all structures and space-delimited validity information.

3: Works like setting 2, but with extended statistics on helices.

'none': Without any parameter called, its2check prints a help message.

Output

standard output The output is as specified by the parameter. For the validity information, the space-separated fields are:

1-6: number of helices (integer), longest helix (integer), UGGU motif (boolean), UU motif (boolean), number of outer helices removed for analysis (integer), valid ITS2 (boolean; is 1 if number of helices is 4, longest helix is 3 and number of outer helices removed is 0).

7-10: (only applicable for parameter 3) prints a count for helices 1-4 irregarding of the number of actual helices as given in the first field.

D.2 NCBI Download

Call 01_ncbi_download_new.pl

Theory of Operation

1. Download ITS2 GenBank identifiers from NCBI
2. Download one ITS2 GenBank file from NCBI by using downloaded identifiers in step 1
3. Split up the downloaded GenBank file into multiple files, one per GenBank identifier
4. Delete temporary files

External Calls GNU wget.

Input None.

Parameters

–nodelete Do not delete temporary files (its2-ids.xml, its2-ncbi.dat) created by the script.

–nodownload Do not download data from NCBI, instead use the temporary files; works only if they are not deleted.

–nosplit do not split up GenBank file retrieved from NCBI into separate, GI-wise files.

–days x Download data only from the last x days.

–help Show help message.

Output

standard output (depending on parameters) Messages from GNU wget; one message for each separate file.

D.3 Database transfer of NCBI sequences

Call 02_gb_to_db.pl

Theory of operation

1. Assemble list of GenBank identifiers
2. Open file(s) for reading
3. Select and store relevant features
4. Check for consistency of features
5. Write to database

External calls None.

Input Either one GI corresponding to one GenBank file in the GenBank directory or a directory with zero or more GenBank files.

Parameters

–fromdir dir Read all GenBank files from directory *dir*.

–nocheckexist Do not check for existence of entry in ITS2 database upon writing.

–nowritedb Do process the data, but do not actually write to database.

–verbose Set verbose mode (only necessary for debugging information).

–help Show help message.

Output

standard output Error messages upon failure; one-line descriptions for each GenBank file inserted, with reasons for rejecting GenBank entries described.

database Additions to the sequence and feature tables.s

D.4 Taxonomy update

Call 03_update_taxonomy.pl

Theory of Operation

1. Download full taxonomy from NCBI FTP
2. Transfer taxonomy information to database
3. Select necessary information (ID, parent ID, name, rank) into table
4. Calculate nested set values
5. Cleanup (delete temporary tables and downloaded files)

External Calls GNU wget, GNU tar, GNU gzip, psql.

Input None.

Parameters

–nodownload Do not fetch taxonomy from NCBI.

–nosetuptables Do not set up temporary taxonomy tables.

–nowritetree Do not write nested set information for the temporary tree table.

–notransfer Do not transfer temporary tree to live database tree.

–nodeletetemp Do not delete temporary files and tables.

–help Show help message.

Output

database tree table.

standard output Messages from utilities.

D.5 Direct Vienna RNA fold

Call 04_vienna_fold.pl

Theory of operation

1. Assemble (list of) GenBank identifiers with or without reannotation
2. Retrieve their sequences exclusively from the database
3. Fold sequence
4. Check for formal validity
5. Write to ITS2 database
6. Delete temporary files

External calls RNAfold.

Input format GenBank identifier, GenBank identifier with differing feature start and end information (=reannotation)

Input Either one or more GenBank identifiers according to the input format as parameters or all sequences without structures from database or GenBank identifiers according to input format from a file (one per line)

Parameters

–fromdb Retrieve sequences to be folded from ITS2 database.

–annotation x (multiple) (only applicable if –fromdb used) Retrieve only sequences with given annotation method *x* from ITS2 database.

–fromfile file Retrieve sequences to be folded from *file*, one per line, according to the input format.

–nodelete Do not delete temporary files.

–nocheckexist Do not check for existence of already valid structure in the database upon writing.

–nowritedb Process all data, but do not write to database.

–verbose Set verbose mode (only necessary for debugging information).

–help Show help message.

Output

standard output Error messages upon failure; one-line descriptions for each structure inserted, with reasons for rejecting structures.

database structure and its2_feat tables; feature table if reannotation took place.

D.6 Modelling database writer

Call create_modelling_db.pl

Theory of operation Select all sequences from database who has a Vienna folded or first run homology modelled structure; write all these sequences to one FASTA formatted file.

External calls None.

Input None.

Parameters None.

Output

files its2.fasta, containing sequences as described above.

standard output one line message of sequences written to file.

D.7 GeneMatcher loader

Call load_genematcher.sh

Theory of operation Upload its2.fasta to the GeneMatcher database.

External calls btk command line utilities.

Input its2.fasta

Parameters None.

Output

standard output messages from the btk utilities.

D.8 Homology run preparation

Call 05_homologyrun.pl

Theory of operation

1. Read list of GenBank identifiers, either all without structure in database or from file to gmrun.fsa.
2. Fetch their sequences.
3. Run the GeneMatcher with the generated query, global alignment mode.
4. Split big BLAST-formatted result file into one file per alignment.

External calls Secure Shell, BioView ToolKit btk.

Input format GenBank identifier, GenBank identifier with differing feature start and end information (=reannotation)

Input Either one or more GenBank identifiers according to the input format from a file or all sequences without structures from database.

Parameters

–fromfile file Read the list of GenBank identifiers from *file*.

–annotation x (multiple argument) (only applicable if –fromfile not used) only select ITS2 sequences with feature annotation of type(s) *x* (defaults to 1 and 2)

–nocheckexist Do not exclude sequences that already have valid structure in the database.

–nowritequery Do not write the query file gmrun.fsa.

–norun Do not run GeneMatcher.

–nosplitresult Do not split up the result file gmrun.fsa from GeneMatcher.

–verbose Set verbose mode (only necessary for debugging information).

–help Show help message.

Output

files gmrun.bl which is split up into GI-wise BLAST-formatted alignment files. Query structures may contain reannotation format as described previously, for later use by the homology modelling script.

D.9 Homology modelling

Call 06_homology_modelling.pl

Theory of operation

1. Read list of alignment files either as a parameter or from a directory.
2. Extract alignment information, including identifier-encoded reannotation.
3. Retrieve source structure from database.
4. Check for E-Value and non-existence of target structure in database.
5. Transfer structure.
6. Analyze for helix transfer percentage.
7. Postfold transferred structure.
8. Given four helices, check for ITS2 validity, reannotate either from previous BLAST result or cut ends, determine structure method type, determine energy and write to database.

External calls its2check, RNAeval.

Input format BLAST standard alignment.

Input BLAST-formatted global alignment files.

Parameters

–fromdir dir Reads alignment files with file suffix .bl from directory *dir*.

–nocheckexist Do not check for existence of already valid structure in the database upon writing.

–nowritedb Process all data, but do not write to database.

–verbose Set verbose mode (only necessary for debugging information).

–help Show help message.

Output

standard output One line per processed alignment, giving either an error or a conformation of database modification. The latter always includes full information on number of helices transferred, transfer percentages, annotation type (GenBank, GenBank cutted and BLAST) and structure method type; anything else can be considered an error message.

database structure, its2_feat, homologue_struct and alignment tables; feature table if the source feature of the newly derived structure has been reannotated.

D.10 BLAST database loader

Call load_blastdb.sh

Theory of operation Build BLAST database from its2.fasta.

External calls xdformat from WU BLAST.

Input its2.fasta.

Parameters None.

Output

standard output Messages from xdformat.

BLAST database Index files in the BLAST database directory.

D.11 BLAST run

Call 07_blast.pl

Theory of Operation

1. Retrieve a number of GIs.
2. For each GI, fetch their full sequence.
3. Assemble a FASTA file.
4. Run BLAST with that FASTA file.
5. For each hit, take the one the highest p-value and write it to a reannotation file.

External Calls WU-BLAST (blastn executable).

Input None.

Parameters List of GIs.

Output

standard output Messages from BLAST and status information.

files blastrun.fsa is the generated FASTA file for the run. blastresult.txt is the tab-formatted result from BLAST. From this file, reannotation.txt is computed and in the reannotation format described above.

Bibliography

M. Achtziger. Master's thesis, Julius-Maximilians-Universität, Würzburg, Germany, 2005.

S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

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 Res*, 25:3389–3402, 1997.

P. Ax. *Das System der Metazoa: ein Lehrbuch der phylogenetischen Systematik*. G. Fischer Verlag, Stuttgart; Jena; New York, 1995. ISBN 3-437-30803-3.

A.D. Baxevanis. The Molecular Biology Database Collection: 2003 update. *Nucleic Acids Res*, 31(1):1–12, 2003.

D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, B.A. Rapp, and D.L. Wheeler. GenBank. *Nucleic Acids Res*, 28(1):15–18, 2000.

D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler. GenBank. *Nucleic Acids Res*, 35:21–25, 2007.

F.P. Brooks. *The Mythical Man-Month: Essays on Software Engineering*. Addison-Wesley Professional, Boston, USA, 20th anniversary edition edition, August 1995. ISBN 0-2018-3595-9.

I.E. Cha and E.C. Rouchka. Comparison of Current BLAST Software on Nucleotide Sequences. In *IPDPS '05: Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05) - Workshop 7*, page 197.1, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7695-2312-9.

J.E. Cohen. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biol*, 2(12):e439, 2004.

A.W. Coleman. The significance of a coincidence between evolutionary landmarks found in mating affinity and a DNA sequence. *Protist*, 151(1):1–9, 2000.

A.W. Coleman. ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends Genet*, 19(7):370–375, 2003.

A.W. Coleman and J.C. Mai. The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants. *J Mol Evol*, 44(3):258–271, 1997a.

A.W. Coleman and J.C. Mai. Ribosomal DNA ITS-1 and ITS-2 sequence comparisons as a tool for predicting genetic relatedness. *J Mol Evol*, 45 (2):168–177, 1997b.

A.W. Coleman and V.D. Vacquier. Exploring the Phylogenetic Utility of ITS Sequences for Animals: A Test Case for Abalone (*Haliotis*). *J Mol Evol*, 54 (2):246–257, 2002.

C.A. Coté and B.A. Peculis. Role of the ITS2-proximal stem and evidence for indirect recognition of processing sites in pre-rRNA processing in yeast. *Nucleic Acids Res*, 29(10):2106–2116, 2001.

S.L. Dellaporta, A. Xu, S. Sagasser, W. Jakob, M.A. Moreno, L.W. Buss, and B. Schierwater. Mitochondrial genome of *Trichoplax adhaerens* supports Placozoa as the basal lower metazoan phylum. *Proc Natl Acad Sci*, 103 (23):8751–8756, 2006.

K.J. Doshi, J.J. Cannone, C.W. Cobaugh, and R.R. Gutell. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, 5:105, 2004.

A. Ender and B. Schierwater. Placozoa are not derived cnidarians: evidence from molecular morphology. *Mol Biol Evol*, 20(1):130–134, 2003.

B.J. Finlay. Global dispersal of free-living microbial eukaryote species. *Science*, 296(5570):1061–1063, 2002.

R.D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L.L. Sonnhammer, and A. Bateman. Pfam: clans, web tools and services. *Nucleic Acids Res*, 34:247–251, 2006.

R. Giegerich, B. Voß, and M. Rehmsmeier. Abstract shapes of rna. *Nucleic Acids Res*, 32(16), 2004.

J.J. Gillespie, J.S. Johnston, J.J. Cannone, and R.R. Gutell. Characteristics of the nuclear (18S, 5.8S, 28S and 5S) and mitochondrial (12S and 16S) rRNA genes of *Apis mellifera* (*Insecta*: *Hymenoptera*): structure, organization, and retrotransposable elements. *Insect Mol Biol*, 15(5):657–686, 2006.

W. Gish. WU BLAST, 2004. URL `http://http://blast.wustl.edu`.

K.G. Grell. *Trichoplax adhaerens*: F.E. Schulze und die Entstehung der Metazoen. *Naturwiss Rundschau*, 24:160–161, 1971.

R.R. Gutell, A. Power, G.Z. Hertz, E.J. Putz, and G.D. Stormo. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res*, 20(21):5785–5795, 1992.

T. Hadrys, R. DeSalle, S. Sagasser, N. Fischer, and B. Schierwater. The *Trichoplax* PaxB gene: a putative Proto-PaxA/B/C gene predating the origin of nerve and sensory cells. *Mol Biol Evol*, 22(7):1569–1578, 2005.

M. Höchsmann, T. Töller, R. Giegerich, and S. Kurtz. Local Similarity in RNA Secondary Structures. In *Proceedings of the IEEE Bioinformatics Conference (CSB 2003)*, pages 159–168, 2003.

M. Höchsmann, B. Voss, and R. Giegerich. Pure Multiple RNA Secondary Structure Alignments: A Progressive Profile Approach. In *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 53–62, 2004.

I.L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res*, 31(13):3429–3431, 2003.

I.L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie*, 125(2):167–188, 1994.

D.H. Huson. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68–73, 1998.

American National Standards Institute. *ANSI/ISO/IEC 9899-1999: Programming Languages — C*. American National Standards Institute, 1430 Broadway, New York, NY 10018, USA, 1999.

W. Jakob, S. Sagasser, S. Dellaporta, P. Holland, K. Kuhn, and B. Schierwater. The Trox-2 Hox/ParaHox gene of *Trichoplax* (Placozoa) marks an epithelial boundary. *Dev Genes Evol*, 214(4):170–175, 2004.

M. Kanehisa. The KEGG database. *Novartis Found Symp*, 247:91–101, 2002.

T. Kulikova, R. Akhtar, P. Aldebert, N. Althorpe, M. Andersson, A. Baldwin, K. Bates, S. Bhattacharyya, L. Bower, P. Browne, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, G. Hoad, C. Kanz, C. Lee, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, D. Lorenc, H. McWilliam, G. Mukherjee, F. Nardone, M.P.G. Pastor, S. Plaister, S. Sobhany, P. Stoehr, R. Vaughan, D. Wu, W. Zhu, and R. Apweiler. EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res*, 35:D16–20, 2007.

I. Letunic, R.R. Copley, B. Pils, S. Pinkert, J. Schultz, and P. Bork. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res*, 34:D257–60, 2006.

H. Margalit, B.A. Shapiro, A.B. Oppenheim, and J.V. Maizel. Detection of common motifs in RNA secondary structures. *Nucleic Acids Res*, 17:4829–4845, 1989.

C. Martinelli and J. Spring. Distinct expression patterns of the two T-box homologues Brachyury and Tbx2/3 in the placozoan *Trichoplax adhaerens*. *Dev Genes Evol*, 213(10):492–499, 2003.

C. Martinelli and J. Spring. Expression pattern of the homeobox gene Not in the basal metazoan *Trichoplax adhaerens*. *Gene Expr Patterns*, 4(4): 443–447, 2004.

E. Mayr. *Systematics and the origin of species*. Columbia University Press, New York, 1942.

E. Mayr. *Artbegriff und Evolution*. Parey, Hamburg / Berlin, 1967.

E. Mayr. What is a species and what is not? *Philosophy of Science*, 63: 262–277, 1996.

D.J. Miller and E.E. Ball. Animal Evolution: The Enigmatic Phylum Placozoa Revisited. *Curr Biol*, 15(1):R26–R28, 2005.

T. Müller, N. Philippi, J. Schultz, T. Dandekar, and M. Wolf. personal communication, 2007.

T. Müller, R. Spang, and M. Vingron. Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol Biol Evol*, 19(1):8–13, 2002.

T. Müller and M. Vingron. Modeling amino acid replacement. *J Comput Biol*, 7(6):761–776, 2000.

R. Nussinov and A.B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci*, 77(11):6309–6313, 1980.

G.J. Olsen and C.R. Woese. Ribosomal RNA: a key to phylogeny. *FASEB J*, 7(1):113–123, 1993.

N. B. Petrov and V.V. Aleshin. [Conditionally neutral phylogenetic markers of major taxa: a new aspect of the evolution of macromolecules]. *Genetika*, 38(8):1043–1062, Aug 2002. English Abstract.

The PostgreSQL Global Development Group. PostgreSQL 8.2.1 Documentation, 2006. URL `http://www.postgresql.org/docs/8.2/static/`.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. URL `http://www.R-project.org`.

E.S. Raymond. *Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. O'Reilly & Associates, Inc., Sebastopol, CA, USA, 2001. ISBN 0596001312.

J. Reeder and R. Giegerich. Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics*, 21(17), 2005.

P. Rice, I. Longden, and A. Bleasby. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet*, 16(6):276–277, 2000.

D. Sankoff. Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. *SIAM J Appl Math*, 45(5):810–825, 1985.

J. Schultz, S. Maisel, D. Gerlach, T. Müller, and M. Wolf. A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. *RNA*, 11:361–364, 2005.

J. Schultz, T. Müller, M. Achtziger, P.N. Seibel, T. Dandekar, and M. Wolf. The internal transcribed spacer 2 database – a web server for (not only) low level phylogenetic analyses. *Nucleic Acids Res*, 34:W704–W707, 2006.

F.E. Schulze. *Trichoplax adhaerens*. *Zoologischer Anzeiger*, 6:92, 1883.

P.N. Seibel, T. Müller, T. Dandekar, J. Schultz, and M. Wolf. 4SALE – a tool for synchronous RNA sequence and secondary structure alignment and editing. *BMC Bioinformatics*, 7:498, 2006.

B.A. Shapiro. An algorithm for comparing multiple RNA secondary structures. *Comput Appl Biosci*, 4(3):387–393, 1988.

S. Siebert and R. Backofen. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, 21(16):3352–3359, 2005.

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. Lehväslaiho, 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 Res*, 12(10):1611–1618, Oct 2002.

R. Stallman. *Using the GNU Compiler Collection*. Free Software Foundation, Inc., Cambridge, Massachusetts, 2003.

P. Steffen, B. Voss, M. Rehmsmeier, J. Reeder, and R. Giegerich. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503, 2006.

J.D. Thompson, D.G. Higgins, and T.J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680, Nov 1994. Comparative Study.

G. Trooskens, D. De Beule, F. Decouttere, and W. Van Criekinge. TreeIllustrator, 2005. URL `http://nexus.ugent.be/geert/index.php`.

J. Venema and D. Tollervey. Ribosome synthesis in *Saccharomyces cerevisiae*. *Annu Rev Genet*, 33(1):261–311, 1999.

O. Voigt, A.G. Collins, V.B. Pearse, J.S. Pearse, A. Ender, H. Hadrys, and B. Schierwater. Placozoa – no longer a phylum of one. *Curr Biol*, 14(22): R944–R945, 2004.

M. Wolf, M. Achtziger, J. Schultz, T. Dandekar, and T. Müller. Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures. *RNA*, 11(11):1616–1623, 2005a.

M. Wolf, J. Friedrich, T. Dandekar, and T. Müller. CBCAnalyzer: inferring phylogenies based on compensatory base changes in RNA secondary structures. *In Silico Biology*, 5:0027, 2005b.

M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244(4900):48–52, 1989.

M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31(13):3406–3415, 2003.

M. Zuker, D.H. Mathews, and D.H. Turner. *Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide in RNA Biochemistry and Biotechnology*. NATO ASI Series. Kluwer Academic Publishers, 1999.

List of Tables

List of Figures

Hiermit erkläre ich, Christian Selig, geboren am 9.2.1983 in Bamberg, eidesstattlich, dass ich die vorliegende Diplomarbeit selbständig und unter ausschließlicher Verwendung der angegebenen Literatur und Hilfsmittel erstellt habe.

Diese Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt.

Hereby I, Christian Selig, born on February 2nd, 1983 in Bamberg/Germany, declare on oath that this diploma thesis was written autonomously, exclusively using cited literature and resources.

This work has not yet been submitted in this or similar form to another examination board.

Christian Selig, Würzburg, 22.02.2007