The ARB Project

An integrated, non-commercial software solution for Phylogenetic Treeing, Sequence Data Analysis and Molecular Probe Design

Presentation by Yadhu Kumar, ARB Group

Lehrstuhl für Mikrobiologie
Lehrstuhl für Rechnertechnik & Rechnerorganisation
Technische Universität München, München, Germany
Ideas

◆ Central Database to maintain a structured integrative secondary data in combination with processed primary structures (aligned sequences) and any additional data assigned to the individual sequences.
Ideas

◆ Central Database to maintain a structured integrative secondary data in combination with processed primary structures (aligned sequences) and any additional data assigned to the individual sequences.

◆ Comprehensive selection of software tools directly interacting with one another and as well as with the central database facilitating in depth analysis of molecular data.
Central Database to maintain a structured integrative secondary data in combination with processed primary structures (aligned sequences) and any additional data assigned to the individual sequences.

Comprehensive selection of software tools directly interacting with one another and as well as with the central database facilitating in depth analysis of molecular data.

Common Graphical User Interface
ARB DATABASE
Aligned Sequences, Descriptive Data, User Data, Profiles, Filters, Trees
ARB Main Window & Import Window
ARB Main Window & Import Window

Import Various Databases

Enter file name of foreign database (may contain * or ? wildcards)

```
/nFshome/yadhu/GeneBankSequenceData.seq
```

Directories (D) and Files (F)

- CONTENTS OF `/nFshome/yadhu`
- `D 'PARENT DIR' (..)`
- `D '$PT_SERVER_HOME' (/nFshome/yadhu/ARB/lib/pts)`
- `D .AbiSuite`
- `D .SuSENautilusPage`
- `D .ansn`

- Import genon data in GENDANK format
- Import genon data in EMBL format
- Import Foreign data format (or press AUTO DETECT)

Enter alignment name + type (e.g. "16s" / ma for 16S rRNA)

<table>
<thead>
<tr>
<th>name</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>16s RNA</td>
<td>rna</td>
</tr>
</tbody>
</table>

AUTO DETECT

```
F gsc-rsf.ift         1k Feb 06 1
F gsc_seq_only.ift    Ok Nov 23 1
F gde.ift             Ok Nov 23 1
F gde_Flat.ift        Ok Feb 06 1
F genbank.ift         Ok Feb 06 1
F genprot.ift         Ok Nov 23 1
```
ARB DATABASE
Aligned Sequences, Descriptive Data, User Data, Profiles, Filters, Trees
Search, Query & Modify ARB Database

Search fields:
- full_name
- nuc_term
- ebi_check

Search strings:
- full_name: *Salmonella*
- nuc_term: >1500
- ebi_check: 

Hits: 18

- SalTyp44: Salmonella typhi
- SalPara2: Salmonella paratyphi A
- SalEnt22: Salmonella paratyphi B
- SalEnt23: Salmonella paratyphi C
- SalEnt25: Salmonella blockley
- SalEnt30: Salmonella muenchen
- SalEnt31: Salmonella weltevreden

[MARK LISTED]
[UNMARK REST]
[UNMARK LISTED]
[MARK REST]
[DELETE LISTED]
[WRITE TO FIELDS OF LISTED]

[REFRESH]
Search, Query & Modify ARB Database

This module modifies the contents of fields. You can:
- substitute substrings
- copy one field to another
- extract and calculate sequence information

DATABASE FIELDS

- full_name
- strain
- next_rel
- lb_name
- acc
- db_acc
- version
- aligned
- author
- title
- journal
- seqlist
- seqlist_count
- seqcheck
- seqcheck_count
- nuc
- nuc_term
- id
- tmp
- name

DATABASE FIELDS

Or select a predefined program:
- remove all " ***
- copy (full_name) to this field
- append (full_name) to this field
- calculate sequence checksum
- count ambiguities (nry, ...) (filter: ECOLI)
- truncate seq. ranges outside ecoli range
- increase G+C content in helical regions by 5%
- SIMPLPRO Simplified AminoAcid
- Count nucleotides
- Count nucleotides (incl. IUPAC)
ARB DATABASE
Aligned Sequences, Descriptive Data, User Data, Profiles, Filters, Trees

DATABASE Querying

IMPORT Data

EBI / GENBANK / RDP / ANTWERPEN
LABORATORY
OTHER
ARB DATABASE
Aligned Sequences, Descriptive Data, User Data, Profiles, Filters, Trees

DATABASE Querying

IMPORT Data

LABORATORY

OTHER

EBI / GENBANK / RDP / ANTWERPEN

Automated Sequence Aligner
Primary Structure Editor
Secondary Structure Editor

ClustalW
Fast Aligner
Primary & Secondary Structure Editors
Primary & Secondary Structure Editors
Primary & Secondary Structure Editors
ARB DATABASE
Aligned Sequences, Descriptive Data, User Data, Profiles, Filters, Trees

Automated Sequence Aligner
Primary Structure Editor
Secondary Structure Editor

DATABASE Querying

IMPORT Data

LABORATORY
OTHER

EBI / GENBANK / RDP / ANTWERPEN
ARDB DATABASE
Aligned Sequences, Descriptive Data, User Data, Profiles, Filters, Trees

DATABASE Querying

IMPORT Data

LABORATORY

OTHER

Automated Sequence Aligner
Primary Structure Editor
Secondary Structure Editor

Phylogenetic Treeing

EBI / GENBANK / RDP / ANTWERPEN

OTHER

LABORATORY
Phylogenetic Tree Building Methods

Phenetic Methods
- Distance based
  - Neighbor Joining Method
  - Minimum Evolution Method

Cladistic Methods
- Character Based
  - Maximum Likelihood Method
  - Maximum Parsimony Method
Phylogenetic Tree Building Methods

Phenetic Methods
- Distance based

Neighborhood Joining Method
- Clustering Algorithm
  Builds a small tree and keeps adding the sequences to arrive at a full desired tree

Minimum Evolution Method
- Optimality Criterion
  Selects the tree whose sum of branch lengths is the minimum

Cladistic Methods
- Character Based

Maximum Likelihood Method
- Optimality Criterion
  Selects the tree that is most likely to have produced the observed data

Minimum Evolution Method
- Optimality Criterion
  Selects the tree whose sum of branch lengths is the minimum

Maximum Parsimony Method
- Optimality Criterion
  Selects the tree that require fewer evolutionary changes
Treeing Methods in ARB

- Neighbor Joining Method
  - Distance Based
  - Phylip Distance Matrix Method
- FastDNAml Method
  - Character Based
- Parsimony Method

Character Data:

```
1 2 3 4
T T A T T A A
A A T T T A A
A A A A A T A
A A A A A A T
```

Distance Matrix:

```
1  0  3  5  5
2  0  4  4  0
3  0  2  0  0
4  0  0  0  0
```
The ARB Parsimony Tool

- Able to handle big trees (e.g. >30,000 16S/18S rRNA sequences)
- Allows optimization of trees and sub-trees with different parameters.
- Adding sequences is possible without changing initial topology.
Phylogenetic Treeing using ARB Software

- Import Sequences
- Alignment of Sequences with Automated Aligner
- Manual Correction of Aligned Sequences
- Primary Structure Editor
- Secondary Structure Editor
- Phylogenetic Treeing
- Visualization and Inference of Trees
EBI / GENBANK / RDP / ANTWERPEN

LABORATORY

OTHER

IMPORT Data

DATABASE
Querying

ARB DATABASE
Aligned Sequences, Descriptive Data, User Data, Profiles, Filters, Trees

Automated Sequence Aligner

Primary Structure Editor

Secondary Structure Editor

Phylogenetic Treeing
Visualization of Sequence Data
Visualization of Sequence Data
Visualization of Sequence Data
Phylogenetic Treeing

Automated Sequence Aligner
Primary Structure Editor
Secondary Structure Editor

LABORATORY

IMPORT Data

ARB DATABASE
Aligned Sequences, Descriptive Data, User Data, Profiles, Filters, Trees

DATABASE Querying

EBI / GENBANK / RDP / ANTWERPEN

OTHER

IMPORT Data

Visualization of Sequence Data
ARB DATABASE
Aligned Sequences, Descriptive Data, User Data, Profiles, Filters, Trees

DATABASE Querying

Automated Sequence Aligner
Primary Structure Editor
Secondary Structure Editor

Phylogenetic Treeing

Positional Tree SERVER
Probe Match
Probe Design

Visualization of Sequence Data

IMPORT Data

LABORATORY
OTHER

EBI / GENBANK / RDP / ANTWERPEN

OTHER

LABORATORY

IMPORT Data
Probe Design and Probe Match
Probe Design and Probe Match

This module searches for specific oligonucleotides in the database.

The PT_SERVER's (not the current) database is used searching probe targets.

Enter some parameters (press help to get more information):

<table>
<thead>
<tr>
<th>Parameter</th>
<th>MIN</th>
<th>MAX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of output</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Max. non group hits</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Max. hairpin bonds</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Min group hits (%)</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Length of probe</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>Temperature</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>G+C-content</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>ECOLI-position</td>
<td>0</td>
<td>10000</td>
</tr>
</tbody>
</table>
### Probe Design and Probe Match

#### Probe Design Parameters:
- **Length of probe**: 18
- **Temperature** [50.0 - 100.0]
- **GC-Content** [50.0 - 100.0]
- **E.Coli Position** [0 - 100000]

#### Max Non Group Hits:
- 0

#### Min Group Hits:
- 50%

#### Target:
<table>
<thead>
<tr>
<th>Target</th>
<th>le apos</th>
<th>ecol grps</th>
<th>C+C4GC+2AT</th>
<th>Decrease T by n*3C</th>
<th>probe matches n non group species</th>
<th>Probe sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>CACCAACACACGACAC</td>
<td>18 A-10376</td>
<td>63</td>
<td>55.6 56.0</td>
<td>0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 1; 1; 3; 5; 7; 9;</td>
<td>CACCUCCCUCCCUCCU</td>
<td></td>
</tr>
<tr>
<td>ACGGACACACGACUCC</td>
<td>18 A* 8</td>
<td>68</td>
<td>55.6 56.0</td>
<td>0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 3; 1; 3; 5; 6; 8; 9;</td>
<td>GACACCUUCUCCCUCL</td>
<td></td>
</tr>
<tr>
<td>CACCAACACACGACAC</td>
<td>18 A* 6</td>
<td>67</td>
<td>55.6 56.0</td>
<td>0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 1; 1; 1; 4; 5; 5; 5;</td>
<td>GACACCUUCUCCUCCU</td>
<td></td>
</tr>
<tr>
<td>UCCCAACACACGACAC</td>
<td>18 A- 3</td>
<td>61</td>
<td>55.6 56.0</td>
<td>0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 2; 2; 2;</td>
<td>CUCUCUCUCCUCUCCU</td>
<td></td>
</tr>
<tr>
<td>ACGGACACACGACUCC</td>
<td>18 B-10394</td>
<td>68</td>
<td>55.6 56.0</td>
<td>0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 1; 2; 2; 2; 4;</td>
<td>GACACCUUCUCCCUCCU</td>
<td></td>
</tr>
<tr>
<td>CCGGACACACGACUCC</td>
<td>18 B* 2</td>
<td>69</td>
<td>61.1 58.0</td>
<td>0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 2; 2; 4; 7;15;16;17;17;17; GACACCUUCUCCUCCU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGUCGACACACGACGA</td>
<td>18 A* 6</td>
<td>59</td>
<td>55.6 56.0</td>
<td>0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 1; 1; 1; 1; 2; 4;29;30;36;62;</td>
<td>UCCUCUCUCCUCUCCU</td>
<td></td>
</tr>
<tr>
<td>CUGACACACACGACGA</td>
<td>18 A- 4</td>
<td>60</td>
<td>55.6 56.0</td>
<td>0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 1; 1; 1; 2; 4; 5; 8; 9;</td>
<td>UCCUCUCUCCUCUCCU</td>
<td></td>
</tr>
<tr>
<td>ACGGACACACGACAC</td>
<td>18 A* 1</td>
<td>64</td>
<td>55.6 56.0</td>
<td>0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 1; 1; 2; 2; 2; 2; 4; 4; 5; 5;10;</td>
<td>GACACCUUCUCCUCCU</td>
<td></td>
</tr>
</tbody>
</table>

#### Min group hits (%): 50

#### ECOLI-position: D 10000

---

**Note:** The provided screenshot is a visual representation of the probe design and match interface, showing various parameters and results related to probe design and selection.
 Probe Design and Probe Match

**Probe Design Parameters:**
- Length of probe: 18

**Probe Match**
- Target String: `CAACCAACCCACCAACAC`
- PT_SERVER: `localhost: LSU_rRNA.arb`
- Search depth: `SEARCH UP TO 1 MISMATCHES`
- Use weighted mismatches: ☐
- Check complement too: ☑
- Mark in database: ☑
- Write Result to field `tmp`: ☐

**MATCH**

<table>
<thead>
<tr>
<th>name</th>
<th>full name</th>
<th>mis</th>
<th>N_mis</th>
<th>umis</th>
<th>pos</th>
<th>ecoli</th>
<th>rev</th>
<th><code>CAACCAACCCACCAACAC</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>* LstMon24</td>
<td>Listeria monocytogenes</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>10376</td>
<td>63</td>
<td>0</td>
<td>AUCCAGUC--------------------CUUCCUCLUU</td>
</tr>
<tr>
<td>* LstMon19</td>
<td>Listeria monocytogenes</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>10376</td>
<td>63</td>
<td>0</td>
<td>AUCCAGUC--------------------CUUCCUCLUU</td>
</tr>
<tr>
<td>* LstInn15</td>
<td>Listeria innocua</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>10376</td>
<td>63</td>
<td>0</td>
<td>AUCCAGUC--------------------CUUCCUCLUU</td>
</tr>
<tr>
<td>* LstMon23</td>
<td>Listeria monocytogenes 6</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>10376</td>
<td>63</td>
<td>0</td>
<td>AUCCAGUC--------------------CUUCCUCLUU</td>
</tr>
<tr>
<td>* LstSee13</td>
<td>Listeria seeligeri</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>10376</td>
<td>63</td>
<td>0</td>
<td>AUCCAGUC--------------------CUUCCUCLUU</td>
</tr>
<tr>
<td>* LstWel2</td>
<td>Listeria welshimeri</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>10376</td>
<td>63</td>
<td>0</td>
<td>AUCCAGUC--------------------CUUCCUCLUU</td>
</tr>
<tr>
<td>* LstIvan2</td>
<td>Listeria ivanovii</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>10376</td>
<td>63</td>
<td>0</td>
<td>AUCCAGUC--------------------CUUCCUCLUU</td>
</tr>
<tr>
<td>* LstInn16</td>
<td>Listeria innocua</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>10376</td>
<td>63</td>
<td>0</td>
<td>AUCCAGUC--------------------CUUCCUCLUU</td>
</tr>
<tr>
<td>* LstMon21</td>
<td>Listeria monocytogenes</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>10376</td>
<td>63</td>
<td>0</td>
<td>AUCCAGUC--------------------CUUCCUCLUU</td>
</tr>
<tr>
<td>* LstInn17</td>
<td>Listeria innocua</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>10376</td>
<td>63</td>
<td>0</td>
<td>AUCCAGUC--------------------CUUCCUCLUU</td>
</tr>
<tr>
<td>* LstIvan1</td>
<td>Listeria ivanovii</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>10376</td>
<td>63</td>
<td>0</td>
<td>AUCCAGUC--------------------CUUCCUCLUU</td>
</tr>
<tr>
<td>* LstIvan3</td>
<td>Listeria ivanovii</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>10376</td>
<td>63</td>
<td>0</td>
<td>AUCCAGUC--------------------CUUCCUCLUU</td>
</tr>
</tbody>
</table>
Combination of the Protargol Method according to Foissner with FISH on *Epistylis* sp
Hybridised DNA Chips

E. sulfureus

T. halophilus

E. asini

DNA chip images from Dr. Thomas Behr
Design of multiple probes for *Glaucoma scintillans*
In situ hybridization of *Glaucoma scintillans* with multiple probes

A: Species Specific Probe

B: Genus Specific Probe 1

C: Genus Specific Probe 2

D: *G. scintillans* under light microscope

Pictures from Dr. Johannes Fried
EBI / GENBANK / RDP / ANTWERPEN

LABORATORY

IMPORT Data

OTHER

DATABASE Querying

ARB DATABASE
Aligned Sequences, Descriptive Data, User Data, Profiles, Filters, Trees

Visualization of Sequence Data

Automated Sequence Aligner

Primary Structure Editor

Secondary Structure Editor

Phylogenetic Treeing

Positional Tree SERVER

Probe Match

Probe Design
The ARB Genome Project

- Descriptive Data
- Database Query
- Import Filters
- Maps
- Sequences

ORGANISM

EXPERIMENT

GENES
ARB Genome Window: Displaying List of Organisms and associated information.
Genome map of *Listeria monocytogenes* displaying rRNA Operons
Genome map in block view and Individual Gene Information display
The ARB Genome Project

EXPERIMENT

GENES
- Descriptive Data
- Database Query
- Primary Structure Editor
- Automated Aligner
- Probe & Primer Design

ORGANISM
- Descriptive Data
- Database Query
- Import Filters
- Maps
- Sequences
Searching for Genes in the Genome of Organisms and displaying respective Gene Information
The ARB Genome Project

Organism Descriptive Data

Experiment

Genes Descriptive Data

Primary Structure Editor

Database Query

Probe & Primer Design

Automated Aligner

Import Filters

Maps

Sequences

Database Query

Descriptive Data

ORGANISM

EXPERIMENT
The ARB Genome Project

**ORGANISM**
- Descriptive Data
- Database Query
- Import Filters
- Maps
- Sequences

**EXPERIMENT**
- Analysed Data
- Results
- Protocols

**GENES**
- Descriptive Data
- Database Query
- Primary Structure Editor
- Probe & Primer Design
- Automated Aligner

**PHYSIOLOGICAL PATHWAYS**
Currently Maintained ARB Databases

(Eucarya, Archaea, Bacteria)

- Small subunit rRNA  – 16S, 18S rRNA (41,737)
- Large subunit rRNA  – 23S, 28S rRNA (7,312)
- Elongation – initiation factors
- Proton translocation ATPase subunits
- Heat shock proteins
- recA
- RNA polymerases
- DNA gyrase
- Cytochrome oxidase
Operating Systems

- LINUX / Unix Operating System
- Mac OS
- Windows

Programming Languages

- C, C++, Perl and other scripting languages
- GUI is based on X Windows & Open Motif Library
ARB running on Vmware, a Linux emulation software under Windows
Operating Systems

- LINUX / Unix Operating System
- Mac OS

Programming Languages

- C, C++, Perl and other scripting languages
- GUI is based on X Windows & Open Motif Library
"As we enjoy great Advantages from the Inventions of others, we should be glad of an Opportunity to serve others by any Invention of ours; and this we should do freely and generously."

- Benjamin Franklin
People Behind The ARB Project

Group Leader

Dr. Wolfgang Ludwig
Lehrstuhl fuer Mikrobiologie
Technische Universitaet Muenchen
ludwig@mikro.biologie.tu-muenchen.de

Programmers and Curators

Future Goals

- Online Probe Design using ARB Positional Tree server
- Multiple probe sets for selected phylogenetic groups (chip design)
- Chip data analysis and evaluation tool
- Further Development of ARB Genome Analysis Software