

# Indonesia Manual

Indonesia is an integrated program package for biological sequence analysis. The program is written in Java2 and a graphical user interface is an integrated part of the program. Instead of reading the manual online, you may [download it in pdf](#)

## Supported computer platforms

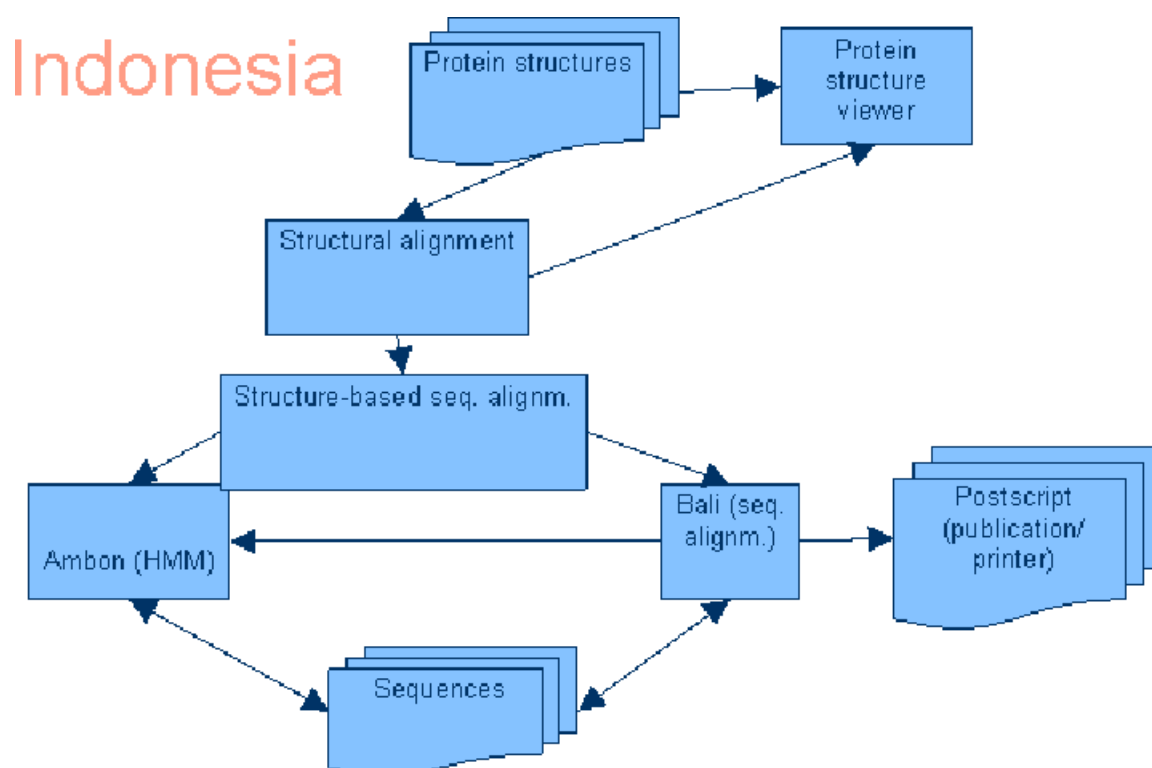
The program runs on any platform supporting Java2, e.g. Windows (XP,2000,98,95), Linux, Irix 6.3, Tru64 Unix, Solaris, Mac (OS X), etc.

## Overview

The program package consists of three parts:

- Bali: Multiple sequence alignment. Either from scratch (like CLUSTALW) or starting from a set of pre-aligned sequences.
- Structure-based multiple sequence alignment. Based on the 3D coordinates from X-ray or NMR studies, the structures are aligned. The structural alignment forms the basis for the subsequent sequence alignment.
- Ambon: Hidden Markov Model (HMM) building. Based on a set of pre-aligned sequences, a HMM is built. The model may be used to test whether or not other sequences are likely to belong to the HMM.

An overview of the flow of information is shown in the Figure below.



## Installation

### Step 1:

First you need to install the Java Runtime Environment (JRE) on your computer – it is available free of charge for all the platforms I know of. If only the Software Development Kit (SDK) is available – no problem – the JRE is included in the SDK.

On the various Unix platforms that might require a super user password. Ask your local system administrator for help.

On Windows, you can install the JRE without administrator rights.

For Windows, Linux and Solaris, the JRE is available at: <http://java.sun.com/j2se/downloads.html> (click on the J2SE 1.4.0 or later link).

For Tru64 Unix (Compaq), the JRE is available at: <http://www.compaq.com/java/download/index.html>

For Irix (Silicon Graphics Unix), the JRE is available at: <http://www.sgi.com/developers/devtools/languages/java.html>

### Step 2:

Download the indonesia.jar file [here](#) (download in Netscape – click with the right mouse button and choose Save link as – download in Internet Explorer – right click on the link and choose Save target as). NOTE! If you are using MAC OS 10.2 or an older version of the Java Runtime, you may want to install this version of the [indonesia.jar](#) file instead. Put the file in a folder called something like: c:\indonesia on Windows or /usr/local/indonesia on Unix.

Create a small script (batch file) that will launch the program.

On Windows call the file indonesia.bat:

```
C:\path_to_java\java -jar c:\indonesia\indonesia.jar
```

On Unix call the file indonesia.sh:

```
#!/bin/tcsh
```

```
/path/to/java/java -jar /usr/local/Indonesia/indonesia.jar
```

On Unix you should furthermore make the file executable with: `chmod a+x indonesia.sh`.

## Bali features

Bali is a multiple sequence alignment program for protein sequences with the possibility to use *a priori* information. *A priori* information might be information stemming from biochemical experiments or 3D structural information. There is furthermore a built-in multiple sequence alignment editor. The features include:

- Reads sequences from any number of files in any of the common sequence formats (even extracts sequence from PDB files).
- Add any number of sequences to the existing alignment from any number of files in any of the common sequence formats.
- Choose between many of the common amino acid substitution matrices like blosum, gonnet, and pam.
- Save the alignment in several common sequence alignment formats or in PostScript format for publication.
- Drag parts of the sequences using the mouse to correct for misaligned parts.
- Often, there are a few so-called ‘fingerprint’ residues in the sequences. These marker residues must be aligned because they are important functionally and/or structurally. Clicking the right mouse button in a column locks that column. This is symbolized with a small lock under the column. This allows the use of prior information as constraints. The lock is effective with respect to both dragging the sequences and re-iterating the alignment.
- There is a free-edit mode in which the sequences may be edited (delete/insert amino acid residues) without any restrictions. This is useful if there are extra residues in some of the sequences that are not interesting (e.g. a signal peptide or a domain that is not shared by any of the other sequences).
- The order of the sequences may be changed and any sequence may be deleted.
- Choose between several coloring schemes or define a custom coloring scheme.
- Calculate consensus sequence with your own defined fraction as cutoff to define ‘consensus’ in a given column.
- Calculate sequence identity and similarity between all pairs of sequences. Two amino acid residues are considered similar when their score in the substitution matrix is greater than zero. The result is presented in a table with the identities in the top-left corner, similarities in the lower-right corner and the number of amino acid residues in the sequence in the diagonal.
- Calculate the entropy in any of the columns in the alignment. Low entropy means little variation of amino acid residue types in that particular column. The result is presented as a graph under the alignment.
- Annotate the alignment with secondary structure elements (alpha-helices or beta-strands).
- Annotate the alignment with a comment line e.g. with \*, \$, letters, number etc. to accentuate features in the alignment.
- Re-iterate the alignment.

## Indonesia screen

When you start Indonesia, the first screen you see is the Bali program.



By clicking on “Open a sequence file” a window pops up and you can choose one or more (ctrl+left mouse click selects the next file etc.).



When you have selected the file(s) (in all formats except PDB files that will be treated separately) the sequences are read and the individual sequences are shown in a window.



You may select and deselect the sequences that you actually want to align. You may also repeat the procedure and open more sequence files to add sequences to your list. Or you can start all over by clicking ‘delete sequence list’.

You can also add new sequences in the text window: Delete the text “Replace this text with your sequences” and write your sequences in the window (mouse right click gives you access to paste the contents of the clipboard from e.g. Word into the text window). Separate sequences with an asterix, \*. Click “add the sequence(s) in the textbox to the list”.



In the Bali window you may also choose the amino acid substitution matrix and the gap penalty parameters. The gap penalty parameters have been chosen depending on the chosen substitution matrix. You may set your own values by deselecting the checkbox “default gap parameters” and adjusting the slider or write the value directly in the text. The other checkbox “positive elements in matrix only” is a shift of the values in the substitution matrix that have been shown to improve the alignment results (reference here).

The last two checkboxes in the Bali window is whether or not to order the sequences so that the most similar sequences will be aligned first and whether or not to weight the sequences according to similarity to the other sequences. The first feature should always be on except if you have pre-ordered the sequences. The second option downweights two sequences if they are very similar so that a particular sequence pattern is not given a too prominent weight – during the alignment the other sequences without that particular pattern might be “forced” into that pattern giving a sub-optimal result.

Finally you may click the “align sequences” to align the sequences from scratch. If you have loaded a file with pre-aligned sequences and you want to use Indonesia to annotate and/or edit the alignment just press “display sequences”.

### The alignment editor window (AEW)

The mouse actions in the AEW are summarized in the table below. On a two button mouse, the middle mouse button is usually emulated by pressing both left and right mouse button. One button mouse – hmm?

Left-click in alignment	Get information on a residue.
Right-click in alignment	Lock column.
Middle-click in alignment	Create all-gap column.
Left-click on sequence name.	Select sequence name (for moving the delete key) – several sequence
Right-click on sequence name.	Change sequence name dialog pops
Drag in alignment (any mouse button pressed)	Drags sequence (any sequences sel
Drag in the area above the alignment (left button)	Set helix/strand annotation to alpha
Drag in the area above the alignment (right button)	Set helix/strand annotation to beta-
Drag in the area above the alignment (middle button)	Delete annotation.

When the sequences have been aligned (or just a display of pre-aligned sequences), the alignment editor window appears.



In the AEW, you can save the alignment in sequence formats FASTA, PIR and CLUSTAL or in PostScript format (menu File->save or save as).

By clicking with the left mouse button on any residue in the alignment, you will get the sequence it belongs to and the position in the sequence.

You can copy the alignment to the clipboard in various forms (full alignment, just sequences etc.) and paste it into e.g. a Word document or any other program (menu Edit).



You can enter the free edit mode (menu Edit->Edit Alignment) – and get out of the free edit mode by selecting the menu item again.

You can change the order of the sequences or delete one or more sequences. By clicking with the left mouse button on the sequence name, the sequence is selected (this can be repeated). If you press the up or down arrow keys, the sequence(s) will move up or down. If you press the delete key, the sequence(s) will be deleted. Be careful – there is NO undo!



An important feature is the ability to drag the sequences with the left mouse button in the AEW. Two different drag/push modes exist: one where the amino acid residues behave like pearls on a string and one where the sequence behaves like a rigid rod so that inserting a gap will move the whole

sequence. You can toggle between the two drag/push modes in the menu item (Edit->Sequences move like pearls on a string).

You can lock any of the columns in the alignment by clicking with the right mouse button in the column. A small lock appears under the column.



When you have edited the alignment to correct for errors and maybe locked the columns you **know** should be aligned (like a protein family fingerprint), you can re-iterate the alignment in the menu (Alignment->Iterate!). After a while, the alignment appears in a new AEW (there is a number on each AEW so that you can see which is which).



In the AEW, you can LOAD new sequences to the existing alignment (File->Load sequences). After the sequences have been loaded a new AEW appears. The new sequences are **not** aligned – so press (Alignment->Iterate!) to align the new sequences to the existing alignment.

In the alignment menu item, you can also calculate the identity and similarity matrix (see description above) (menu Alignment->Calc. Seq. similarity).



If you have edited the alignment some columns with just gaps might have been introduced, they can be removed (menu Alignment->Remove gap columns).

## Annotation

For almost all of the annotation, you can choose to show it when you save the alignment in PostScript format - except the entropy display in each column.

In the AEW, you have many different options to annotate your alignment. To calculate the consensus sequence, you input the threshold for a column to be in consensus (between 0.51 and 1.0 – i.e. 51% and 100%). In the menu select (Annotation->Calculate consensus). The consensus sequence is shown below the alignment.



You also have the option to enter a line with your own comments – like signs, letters or numbers that you might want to refer to in the text in a paper. In the menu select (Annotation->Edit own comment) to edit the line and select it again to leave it as it is. The delete and backspace keys on the keyboard works.



You can have up to two different numbering schemes for the alignment. And each of them can be numbered just following the columns in the alignment or follow the amino acid residue numbering in any of the sequences. The two numbering schemes can be toggled on/off and will be displayed above and below the alignment, respectively.



To emphasize regions in the alignment, you can add annotation boxes, i.e. rectangular boxes in a

given color and with a given line style. Select Annotation->Add annotation box to add a new annotation box, the numbers are the numbers of the sequences in the alignment and the positions are the positions in the alignment. To edit the annotation boxes for any property (position, color or line style) select Annotation->edit annotation boxes in the menu. Here you can also choose to delete any of the annotation boxes.



The consensus sequence is not the only way to get an overview of the regions in the alignment with high conservation. The entropy graph, showing the entropy in each column in the alignment, is also a useful tool. You can toggle in it on/off in the menu item Annotation->Entropy graph. The total entropy for the alignment is a sum over the entropy in each column – Annotation->Calc total entropy.



Another means of annotation is to vary the color scheme. In the menu item color scheme, you will find many different pre-defined color schemes and with the possibility to re-define any of them to enhance the features you want. You can also choose between coloring boxes with the amino acid letters in black or color the letters or toggle coloring off.



You can select a different font size for the alignment - Menu Font size. And toggle between a normal font and bold.



A last useful annotation option is to assign secondary structure elements to the alignment. You can do it manually (later I will show ways to get it done automatically if one or more protein structures are available). Just go to the area above the alignment and numbering screen and press the left mouse button (alpha helix) or right mouse button (beta strand) and drag the mouse button. You can delete the secondary structure annotation at a position by clicking the middle mouse button. The minimum length for an element is set to 3 (alpha helices) and 2 (beta strand).



## Saving to PostScript

Many printers have direct capability to print PostScript files. If your printer does not support PostScript, you can e.g. use GhostView to print PostScript through various drivers. GhostView is also a very nice PostScript viewer. GhostView is also available for Windows not just Unix, e.g. at <http://www.cs.wisc.edu/~ghost/gsview/index.htm> - you need to install GhostScript as well. You can also use GhostView to convert the PostScript file to e.g. jpg, tiff or png format to include the figure in e.g. MS Word.

When you have chosen a file name a dialog box appears. You can choose which of the annotations you want to display (only those that you have active will appear), and whether to place them above or below the alignment. The sequences and the sequence names are separated by a couple of characters. The spacing might be too large – so in the box separation between sequence names and sequences set a **negative** number as the separation.



## Saving a Sequence Logo

A sequence logo is a concentrated presentation of the information a multiple alignment (T.D. Schneider and R. M. Stephens (1990). "Sequence logos: A New Way to Display Consensus Sequences", Nucl. Ac. Res., 18: 6097-6100). To save a sequence logo, choose "Sequence Logo" as

the file type in the save file menu. When you have chosen a file name a dialog box appears. Here you can choose paper size, margins etc. The sequence logo is a small part of the alignment, usually 5-20 positions. You can choose the mid-point of the sequence logo in “zero-point” to an absolute position in the sequence alignment (e.g. 50) and choose the width from the zero point to the left and right, respectively. The sequence logo is saved in PostScript format so you can view it with e.g. GhostView (as mentioned in “Saving to PostScript”).

### **Saving the Entropy in ODAT format (and mapping it onto a 3D structure)**

The entropy in each column can be saved to a file by choosing the file format “ODAT”. If you want to map the entropy onto a 3D structure follow these steps:

- 1) Make sure the 3D structure you want to map the entropy onto is the first in the alignment.
- 2) Save the entropy to an ODAT file.
- 3) Download the perl-script [bfactor.pl](#)
- 4) Make the perl-script executable with "chmod a+x bfactor.pl"
5. Apply the perl-script on the 3D structure  
./bfactor.pl -pdb mypdbfile.pdb -odat myodatfile.odat > newpdb.pdb
6. The file newpdb.pdb will contain the entropy at the place of the b-factor.

## **Comparative Modeling**

If you have a protein structure with known 3D coordinates and would like a suggestion for a 3D model of a protein sequences that is homologous, you want to perform comparative modeling. Indonesia can help you in getting the best sequence alignment using the structural information in the known protein structure(s). Indonesia can read PDB files in various ways (cf. <http://www.rcsb.org/pdb/> for details on the PDB format and for download of PDB files).

### **One protein structure (PDB file)**

Start at the Bali main window and choose open file (I would recommend starting the program from scratch before you do this – to avoid any confusion). Locate the PDB file and open it. If there are any secondary structure records in the PDB file (which is the case for almost all PDB files obtained from RCSB) a prompt appears where you can choose whether you want to use the sequence information only or whether you would like to use the secondary structure information from one of the chains in the PDB file (each chain is treated as a separate sequence). If you select one of the chains an AEW will be opened immediately with the sequence and with secondary structure annotation AND with ‘=’ signs in the regions with secondary structure elements. The gap penalty has been increased in the regions with secondary structure elements. Now you can use the LOAD sequences to add the sequence(s) with unknown structure.



All the other features in the AEW described above are also available. Now you can save the alignment in e.g. FASTA format to use in MODELLER, or any other homology modeling program. This example was rather simple, but it works in more complicated cases too.

## More than one protein structure

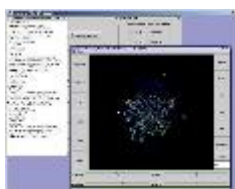
If you have more than one protein structure, it would be advantageous to align the 3D coordinates first and use this structural alignment to obtain a structure-based sequence alignment (SBSA). By pressing the other tab called structural alignment in the Bali window gives you this:



You can open any number of files in PDB format. Each chain in each structure is treated as an individual structure. Please note that the alignment is carried out on the C-alpha coordinates! A structure list appears similar to the sequence list for the sequence alignment part.

Three structural alignment methods are available, one by Levitt and Gerstein REFERENCE (fast), a brute force (slow), and the Gili method. The two first methods are based on pairwise superpositioning of structures where one structure (the first in the list) is chosen as the reference structure that the other structures are superpositioned onto. The third method, Gili, is based on local alignment of all structures. The Gili method is usually the most reliable, but it is recommended to try more methods.

The checkbox view aligned structures in Madura indicates whether you would like to see the result in the built-in protein structure viewer. Madura is actually also a separate application that has been coupled to all entries in PDB with published structure factors. More information at the Uppsala University Electron Density Service <http://portray.bmc.uu.se/eds>. You can choose between aligning the selected protein structure chains or just to display them (in Madura). If you press align structures, the following appears:



In Madura you can rotate the protein structure, zoom etc. In the window on the left, you get some alignment statistics. The  $P(z>Z)$  should be as close to zero as possible. If the value is 1.00, the structural alignment failed and you might want to play around with the various structural alignment methods and parameters.

The lower half of the Bali window now gives access to producing a SBSA. You can send the result directly to an AEW and/or to the Ambon (HMM builder) part of Indonesia. Click the button structure-based sequence alignment.



In the AEW, the alignment appears with secondary structure element annotation and '=' signs under the structurally aligned regions. The secondary structure annotation is only effective inside the aligned regions because it may vary from structure to structure outside the aligned regions. Now you may add the sequence(s) that you would like proposed 3D structures for (and press Annotation->Iterate!).



As for the single structure, you may save the result for use in MODELLER, or any other homology modeling program.

## Using Ambon (HMM builder)

Patrik.

## Comments etc.

Any comments, requests for features, bug reports etc. are more than welcome.

Dennis Madsen, Patrik Johansson, and Gerard J. Kleywegt.

Contact [dennis@xray.bmc.uu.se](mailto:dennis@xray.bmc.uu.se) [Patrik@xray.bmc.uu.se](mailto:Patrik@xray.bmc.uu.se) [gerard@xray.bmc.uu.se](mailto:gerard@xray.bmc.uu.se)

Citation:

D. Madsen, P. Johansson, and G. J. Kleywegt (2002) Indonesia: An integrated sequence analysis system. (In prep.)