Alfresco Manual, version 2.0

Introduction

Comparative sequence analysis is a powerful way of finding coding and regulatory regions in genomic DNA sequences. It does, however, require the use of several different types of analysis. The aim of Alfresco is to provide the user with a consistent interface to these analysis methods and to display the results in an intuitive way.

Please note that Alfresco is very much a program in development, which means that things might change quite rapidly so that what is described in this manual might not always be up to date. The contents of this manual reflects the functionality of ***Alfresco version 0.9****.*

Installation

For installation please consult the README file of the distribution.

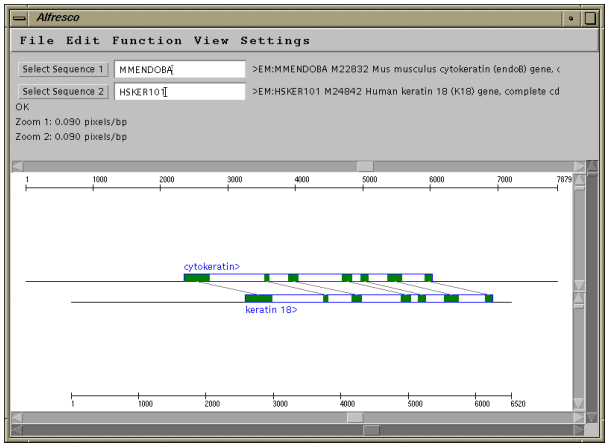
Quick Start

The distribution archive comes with some example files to get you going. These are sequence files in fasta format and cmp files, which are files that contains information about were exons are located and which exons in the two sequences that corresponds to each other.

1. Type

   ```
   alfresco
   ```

   to start the program.
2. Select "Read .cmp file" from the "File" menu.
3. Choose one of the .cmp files. This will read in the two sequence files and the exon information in the cmp file.
4. You should now see two lines with some green boxes drawn on them and lines drawn between the boxes on the two lines. Something like this (click on the image to enlarge it in a separate browser window):



5. There you are! Now you can poke around and try everything out by yourself, or go on reading the rest of this manual if you like.

Usage

```
alfresco [<pair file> | <cmp file> | options]

     options:
     -g <gff file> <seq file 1> <seqfile 2>   reads gff file and seq files
     -c <cgff file> <seq file 1> <seqfile 2>  reads cgff file and seq files
     -b <seq file 1> <seqfile 2>              batch mode
```

Terminology

Some terms that some people might find odd:

Entry
: A DNA sequence specified by a fasta file.

Entry pair
: A pair of entries. The 'working unit' of Alfresco.

Feature
: A feature of an entry, e.g. a coding exon, UTR, Repeat etc.

Reciprocal
: A feature indicating a homologous relationship between two features.

Tier
: The level above (or below) the line representing the entry on which a feature is drawn in the display area. The entry is drawn on tier 1 and the scale bar is drawn on tier 10.

Main Window

The upper part of the main window contains buttons and textfields for selecting sequences (see "Importing sequences" below). Below that are four rows of labels that display (1) the status of the program, (2, 3) the zoomlevel of the two sequences, and (4) a description of the selected feature(s), if any. Most of the window is taken up by the display area which shows a graphic representation of the sequences and any features associated with these.

Importing sequences

There are several ways of importing sequences into Alfresco. Currently sequence files must be in **fasta** format.

- The "Select Sequence" buttons at the top of the main window brings up a file dialog so that you can choose the fasta file with the sequence you want to use.
- You can also type the name of the fasta file containing the sequence directly into the textfield and hit return to load the sequence.
- Finally, you can select "Read .cmp file" from the "File" menu, which will automatically load the right fasta files if they are present.

Display area

The display shows the pair of sequences (or *entries*) as lines in the middle. At the top and the bottom of the display there are scale bars for the two entries. *Features* are drawn as boxes above or below the line depending on if they are on the upper or lower strand. Features can be on other levels (tiers) above (or below) the sequences as well. Features will have a colour dependent on their type, e.g. coding exons are green, UTRs are yellow etc.

Genes are slightly different as they are composed of other features such as Exons, UTRs, promoters and polyA signals. A gene is drawn as a blue border outlining the features it contains and a name tag with ">" and "<" indicating the direction of transcription.

A special kind of feature is the *Reciprocal*. This is drawn as a grey line between two features on either entry, and indicates that there is a homologous relationship between the two features.

Features can be **selected** by clicking with the left mouse button, and additional features can be selected by clicking while pressing the shift key. Ranges of the entries can also be selected by click-dragging with the left mouse button, and additional ranges can be selected by click-dragging with the shift key. (***Note to old users!*** *The middle button clicking of pre-0.9 versions has been disabled.*) Selecting a feature or range will make the label just above the display area show what type of feature(s) has been selected and the range it covers. Some functions will only operate on a selected feature(s) or range(s) (see the Edit, Function, and View menu descriptions below).

There are scrollbars to manipulate the view of the sequences:

- The scrollbars *above* and *below* the display are for **scrolling** the sequences individually, and the slightly darker grey scrollbar at the bottom is for scrolling them synchronously.
- The scrollbars to the *right* of the display are for **zooming** the sequences individually, and the slightly darker grey scrollbar at the far right is for zooming them synchronously.

Menus

File

Open...

Brings up a dialog to choose a previously saved Entry pair file.

Save...

Saves an Entry pair to a file in an application specific format.

Close...

Closes the current Entry pair. Asks if the Entry pair should be saved before closing.

Read .cmp file

Reads in a .cmp file. If two entries are loaded the features specified in the .cmp file will be added to those.
If **no** entries are loaded Alfresco will look for fasta files with the names specified on the ID line of the .cmp file (excluding any colon delimited prefix, e.g. *em:*), in the *current working directory*. If no fasta files are found Alfresco will try to call efetch to get local copies of the sequence files, otherwise it will fail.

Read .gff file

Reads in a .gff file and adds the features specified in the file to the appropriate entry/entries.

Read .cgff file

Reads in a .cgff file and adds reciprocals for the features specified in the file.

Write .gff file...

Writes information about the features of the current Entry pair to a file in .gff format.

Write .cgff file...

Writes information about reciprocal features to a file in format + a .gff

file (see above).

Print..

Sends the contents of the sequence canvas to a printer.

Open Remote EntryPair

Opens a dialog for selecting a pre-analysed pair of sequences from a CORBA server. (Might disapear in the near future).

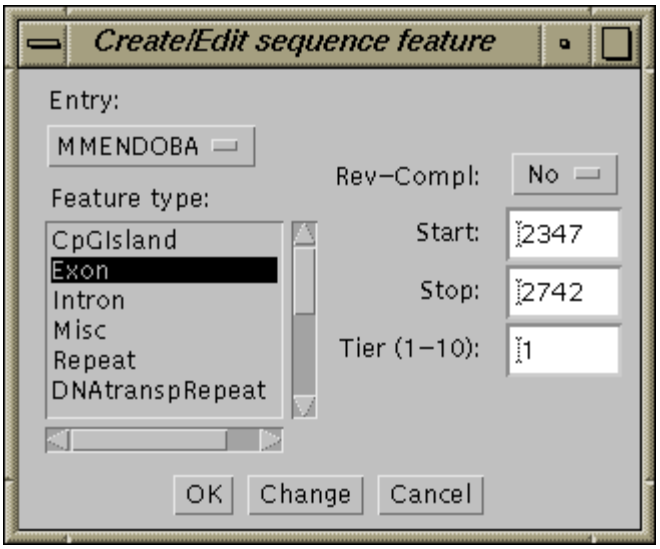
Quit

Quits Alfresco

Edit

Edit/Create feature

This brings up a dialog window for editing an existing feature, or for creating a new one.



If a feature or a range is selected the values of that selection will be filled in the different textfields and pop-up menus of the dialog, otherwise they will be empty.

Entry
A pop-up menu to select the Entry the feature belongs to.
Feature type
A list of feature types.
Rev-Compl
Whether the feature is on the reverse strand, or not.
Start
Start nucleotide position of the feature.

Note! If the feature is on the reverse strand this will be the stop position, i.e. the Start field value must always be less than or equal to the Stop field value.

Stop

Stop nucleotide position of the feature.
Note! If the feature is on the reverse strand this will be the start position, i.e. the Stop field value must always be greater than or equal to the Start field value.

Tier

The tier on which the feature should be drawn.

Change Tier

This brings up a dialog to specify the tier on which a selected feature should be drawn.

Remove feature

Removes selected feature(s).

Make Gene

Creates a new gene from selected Exons and UTRs. Brings up a dialog to specify a gene name and the tier the gene should be drawn on.

Edit Gene Attributes

Brings up the same dialog as the Make Gene menu item, to specify a gene name and the tier the gene should be drawn on.

Add Feature to Gene

Adds a selected Exon or UTR to a selected gene.

Remove Feature from Gene

Removes a selected Exon or UTR from the containg gene.

Confirm Exons

Tags the exon as a real exon.

Unconfirm Exons

Tags the exon as an exon that is not confirmed to be real.

Hide Unconfirmed Exons

Toggles between hiding and showing unconfirmed exons. The default is to show unconfirmed exons

Make Reciprocal

Creates a new Reciprocal between two *selected* features.
Note! There is **no** checking to make sure that the features are similar.

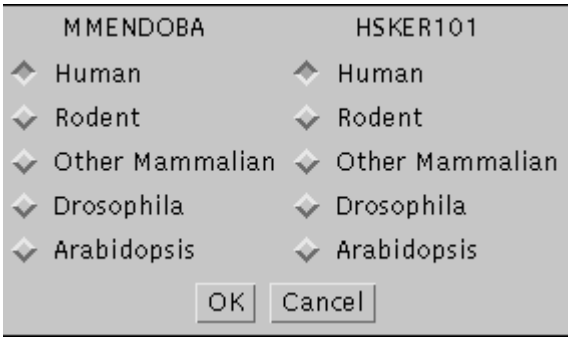
Function

CpG

Calls the cpg program for both entries, and draws CpG islands on the display area. Settings for the cpg program can be altered through the settings menu.

RepeatMasker

Calls the RepeatMasker program for both entries. Brings up a dialog window for selecting the species that each sequence belongs to. Settings for the RepeatMasker program can be altered through the settings menu.



Repeats are drawn as boxes couloured depending on the type of repeat. LINEs - light green; SINEs - blue; LTR repeats - magenta; DNA transposons - orange; Low complexity repeats - brown; Simple repeats - pink; other repeats - black.

Genscan

Calls the Genscan program for both entries. Genscan will not search for exons in regions labled as repeats. Settings for the genscan can be altered through the settings menu.

Blastn alignment

Finds regions conserved in the two entries using blastn. Regions are added to the sequences as Similarity features with Reciprocals between them. If you want to see an alignment use the NW alignment from the View menu.

BlastWise

Does a blastx search against a protein sequence database with both sequences. Significant non-overlapping matching protein sequences are aligned to the genomic sequences by the genewise program. Gene structures predicted by genewise are added to the sequences as Genes. Settings for the blastwise program, such as the database to be searched, can be altered through the settings menu.

BlEst_genome

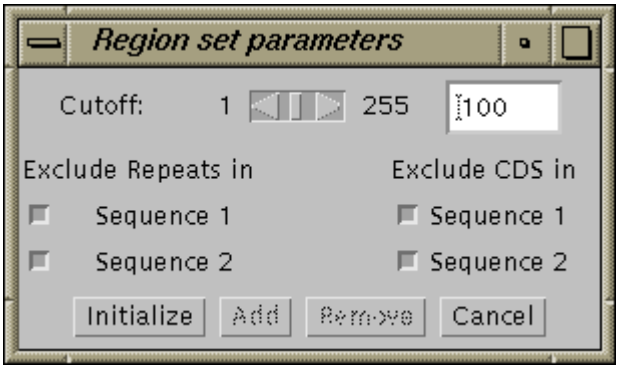
Does a blastn search against a DNA sequence database (preferably a RNA database, such as the EST subsection of EMBL). Significant non-overlapping matching sequences are aligned to the genomic sequences by the est_genome program. Gene structures predicted by est_genome are added to the sequences as Genes. Settings for the est_genome program, such as the database to be searched, can be altered through the settings menu. The blEst_genome allows for a specified number of overlapping hits to be aligned to the genomic sequence to be able to detect alternatively spliced gene structures.

est_genome

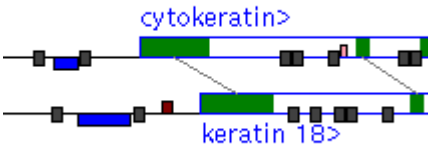
Brings up a dialog for selecting a fasta file of RNA sequences that will be aligned to both enties by the est_genome program. Gene structures predicted by est_genome are added to the sequences as Genes. At the moment no settings dialog has been implemented for est_genome.

Region Set

The Region set method is a way to find regions similar to any other region, be that in the same or the other entry. It is intended to be used to find regions outside of the coding and repeat regions, i.e. potential regulatory regions. Brings up a dialog window:

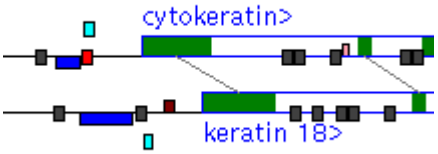


Before clicking the Initialize button, exclusion (or not) of repeats and coding exons should be selected using the checkboxes. Clicking initalize will start some external processes that will *take some time*, especially if the sequences are long. When it is done, regions having similarity to any other region will be covered by a small dark box, something like this:



The stringency of similarity can be varied using the Cutoff scrollbar or text field in the Region set parameter dialog.

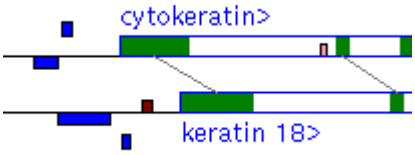
Selecting a region will display with boxes, on a higher tier, all regions similar to the selected one:



A selected Region set can be added to the display area by clicking the Add button of the Region set parameter dialog. This will bring up yet another dialog where the colour of the Region set can be choosen:

A Region set that has been added to the display can be removed using the Remove button of the Region set parameter dialog.

Clicking the Cancel button will close the Region set parameter dialog, and hide all the Region sets except those that have been added to the display area:



Region sets can be examined in various ways by selecting "Sequence", Dotter", or "(new)dba alignment" from the View menu.

View

Sequence

Displays the DNA sequence of a selected feature or range in a separate window:

The sequence can be saved to file either as displayed in the window or in fasta format by clicking on the appropriate button. A gff file of *relative* positions of the features covering the sub-sequence can also be saved.

Amino acid sequence translation

Displays the amino acid translation of a *selected* Gene. Note that no checking is done to see if the start of the first exon is in the right frame. The sequence can be saved to file either as displayed in the window or in fasta format by clicking on the appropriate button.

ATGs

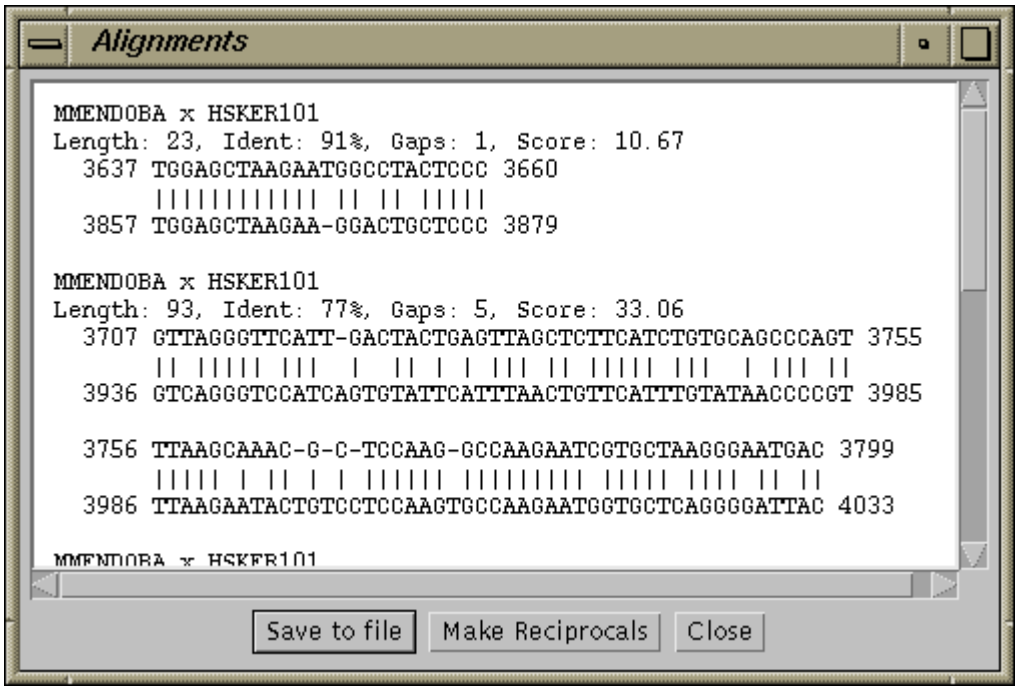
Shows the locations of ATG codons on the upper and lower strands. Note that ATGs are only visible at higher zoom levels.

Dotter

Calls Dotter in interactive mode with a pair of selected features or ranges.

dba alignment

Calls dba on a pair of selected features or ranges. The resulting alignments are displayed in a separate window:



The alignments can be saved to file by clicking the Save to file button. The aligned ranges and reciprocals can be added to the display area by clicking the Make Reciprocals button.

The settings for the dba algorithm can be altered using the "dba settings" from the Settings menu.

NW alignment

Does a Needleman-Wunsh (global) type of alignment on two selected features or ranges. The resulting alignment is displayed in a separate alignment window, as for dba alignments.

SW alignment

Does a Smith-Waterman (local) type of alignment on two selected features or ranges. The resulting alignment is displayed in a separate alignment window, as for dba alignments.

save exons to file

Experimental feature. Saves all exons as separate sequences in one fasta file for each entry. The files are called "exons1" and "exons2". (Yes, it's ugly! ;)

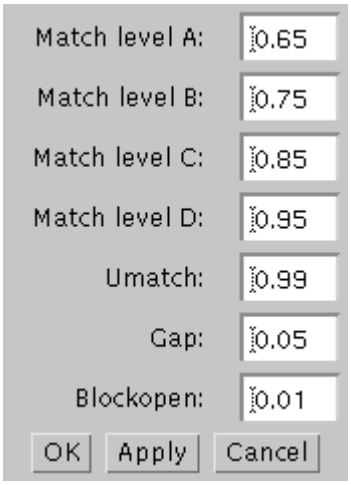
Settings

Show Introns

Toggles between showing and hiding introns (if there are any introns defined). Default is to hide introns.

dba settings

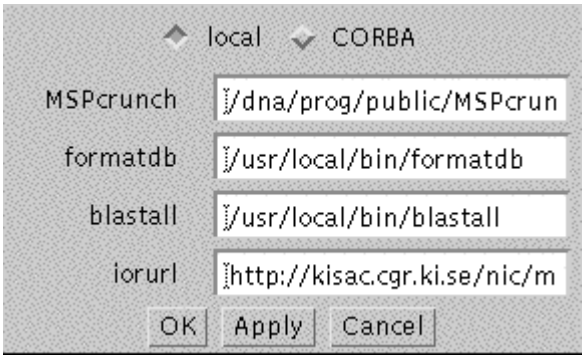
Brings up a dialog window for changing the parameters for dba:



The values are probability values, and the only parameters that there would be any reason to change are the match and gap parameters.

blastalign

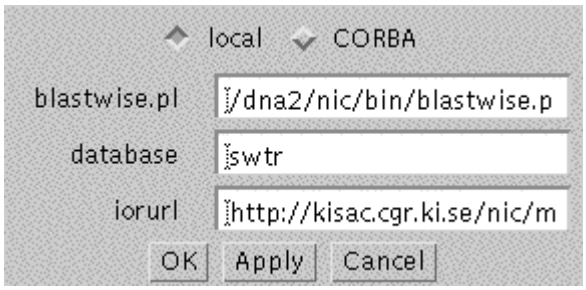
Brings up a dialog window for changing the parameters for blastwise:



MSPcrunch: path to the MSPcrunch executable
formatdb: path to the formatdb executable
blastall: path to the blastall executable
iorurl: url to the CORBA server IOR file

blastwise

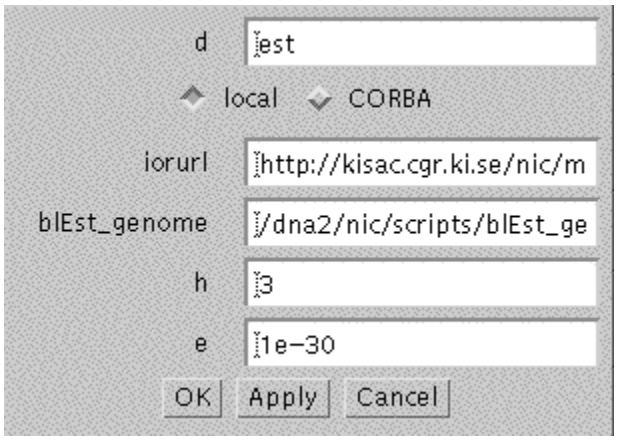
Brings up a dialog window for changing the parameters for blastwise:



Use the radio buttons to choose if the analysis should be run locally or through CORBA.
blastwise.pl: path to the blastwise script
database: name of blast database to be searched
iorurl: url to the CORBA server IOR file

blEst_genome

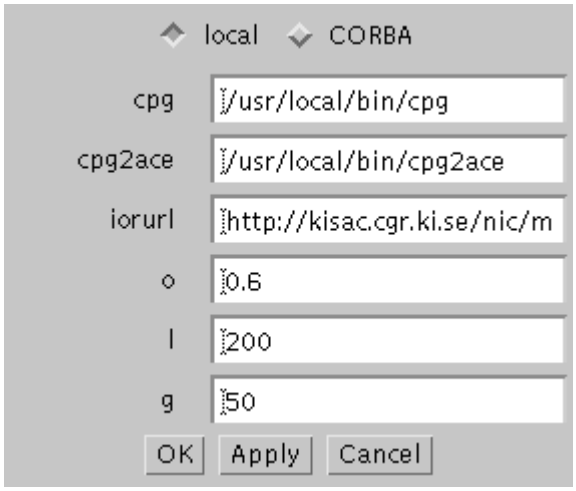
Brings up a dialog window for changing the parameters for blEst_genome:



Use the radio buttons to choose if the analysis should be run locally or through CORBA.
d: name of blast database to be searched
iorurl: url to the CORBA server IOR file
blEst_genome: path to the blEst_genome script
h: number of overlapping hits to align
e: e-value cutoff (matches with higher values are excluded)

CpG

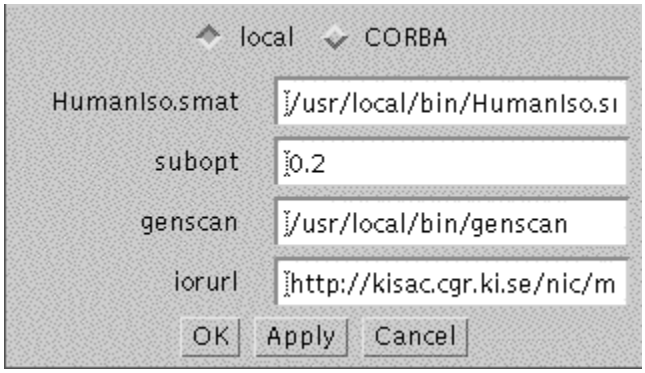
Brings up a dialog window for changing the parameters for cpg:



Use the radio buttons to choose if the analysis should be run locally or through CORBA.
cpg: path to the cpg executable
cpg2ace: path to the cpg2ace script
iorurl: url to the CORBA server IOR file
o: obs/exp CpG frequency
l: min length
g: min GC content

Genscan

Brings up a dialog window for changing the parameters for genscan:



Use the radio buttons to choose if the analysis should be run locally or through CORBA.

HumanIso.mat: path to the matrices file
subopt: cutoff score for sub-optimal exons
genscan: path to the genscan executable
iorurl: url to the CORBA server IOR file

RepeatMasker

Brings up a dialog window for changing the parameters for RepeatMasker:

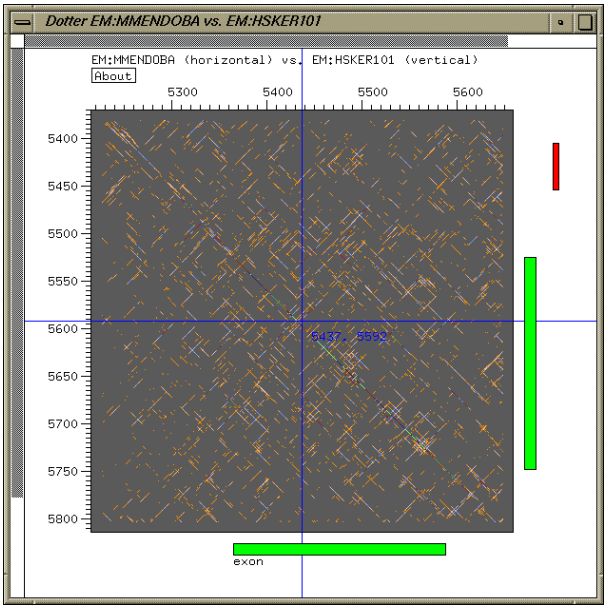
Use the radio buttons to choose if the analysis should be run locally or through CORBA.
RepeatMasker: path to the RepeatMasker script
iorurl: url to the CORBA server IOR file

External programs

Alfresco makes use of several external analysis programs. Here is a list of short descriptions of the programs currently used with links to further documentation if it is available. More programs will most certainly be incorporated in the future. Currently Alfresco makes system calls to these programs, and reads in the results directly or from files produced by the programs. Some programs will produce result files that might be left in your working directory.

Dotter

Dotter is a dotplot program for comparison of two sequences written by Erik Sonnhammer and Richard Durbin. One of the many nice features of dotter is the ability to import sequence feature data. When calling dotter from Alfresco sequence features associated with the sequences will be displayed on the sides of the dotplot:



(The colourmap of the dotpot area is screwed up)

Please see the documentation for more information on all the functionality of Dotter.

CpG

cpg is a CpG island finder. CpG islands are usually defined as being longer than 200 bp with a GC contant over 50% and an observed to expected ratio of CpG dinucleotides above 0.6 (Larsen et al. 1992 Genomics 13:1095-1107). The cpg program was written by Gos Micklem when he was at the Sanger Institute.

RepeatMasker

RepeatMasker, written by Arian Smit and Phil Green finds interspersed repeats known to exist in mamalian genomes, as well as simple repeats and low complexity regions.

Genscan

Genscan is a gene prediction program written by Chris Burge and Samuel Karlin. It is designed to predict complete gene structures in genomic sequences. Alfresco will display predicted genes, as well as "sub-optimal" exons predicted by Genscan.

DBA

dba is a **D**NA **b**lock **a**ligner, written by Ewan Birney, Richard Durbin, and, to a miniscule extent, myself. It finds co-linear blocks of high similarity in two DNA sequences. dba is now part of Ewan Birney's Wise2 package.

DBA can be called from Alfresco with two selected features or ranges. It is also currently used by the "Region Set" and "Find Reciprocals" methods.

The parameters for dba can be modified using the dba settings dialog. There are 7

parameters to change, but there should only be reason to change the **Match** and **Gap** parameters. The values are probability scores. Increasing the Match value will give alignments with higher similarity. Increasing the Gap value will give alignments with more gaps.

Blast/MSPcrunch

Blast (NCBI's version 2) is used by Alfresco for database searching from the blastwise and blEst_genome scripts. Blast is also used to identify conserved regions in the two entries. One of the sequences is formated into a blast searchable database. The other sequence is then used as a query against this on-sequence-database. The blast output is run through Erik Sonnhammer's MSPcrunch program to parse the result and filter out low-scoring matches.

Genewise

Genewise is part of Ewan Birney's Wise2 package. It aligns a protein sequences to a genomic sequence allowing for introns and frameshifts.

Est_genome

Est_genome aligns an RNA sequence to a genomic sequence allowing for introns and frameshifts. It was written by Richard Mott when he was at the Sanger Institute. There is a paper describing it in Bioinformatics/CABIOS.

Example session

Here is a proposal for a working order for analysing a pair of sequences.

- Start by importing your sequences.
- Run CpG and RepeatMasker to find CpG islands and repeats.
- Run Blastn alignment to determine where similar regions are.
- Run BlastWise and blEst_genome to predict gene structures from homologous sequences. These steps are quite time consuming.
- Run Genscan to find potential genes and exons. Genscan will not search for exons in regions labled as repeats.
- Examine the results, and manually edit or remove exons using the functions of the Edit menu.
- For finding conserved regions outside of the coding regions run DBA (and/or Region Set).

Alfresco and ACEDB

If you have done analysis of your sequences in ACEDB you can get the features from ACEDB to Alfresco by exporting them in gff format: from a fmap window in ACEDB select Export Features from the right button pull-down menu. Then select Read .gff file from the File menu in Alfresco. In the future Alfresco might be able to talk to ACEDB

directly, so you don't need to go through this step of an intermediate file.

Appendix

cmp file format

This format was developed to be able to import data available from the EMBL database. The columns on each line are tab delimited. The first line defines the species. The second and third lines define the EMBL id and accession number, respectively. The fourth line defines the gene name(s). For each feature pair there is a line defining feature type, start and stop in first entry, start and stop in second entry, and optionally, lengths of features, and difference in length of features. The 'Sums:...' line shows sums of coding exons and exon length differences.

```
OS      Mus musculus    Homo sapiens
ID      em:MMENDOBA     em:HSKER101
AC      M22832  M24842
GENE    cytokeratin, NCBI gi: 532610    keratin 18
CDS     2347,2742       2580,2996       395     416     -21
CDS     3535,3617       3738,3820       82      82      0
CDS     3892,4048       4158,4314       156     156     0
CDS     4688,4852       4887,5051       164     164     0
CDS     4964,5089       5137,5262       125     125     0
CDS     5365,5588       5525,5748       223     223     0
CDS     5910,6030       6128,6248       120     120     0
Sums:                   1265    1286    -21
```

gff file format

gff is an exchange format for gene finding features proposed by Richard Durbin and David Haussler. The definition is:

```
<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [group]
```

Here are some example records:

```
SEQ1    EMBL    atg     103     105     0.9     +       0
SEQ1    EMBL    exon    103     172     6.42    +       0
SEQ1    EMBL    splice5 172     173     0       +       .
SEQ1    netgene splice5 172     173     0.94    +       .
SEQ1    genie   sp5-20  163     182     2.3     +       .
SEQ1    genie   sp5-10  168     177     2.1     +       .
SEQ2    grail   ATG     17      19      2.1     -       0
```

There is now a version 2 of the gff format. The main change conserns the structure of the last group field. Alfresco currently only deals with version 1 of the format, but this will probably change in the future.

cgff file format

This format was devised to be able to specify pairs (or sets) of sequence features that

are homologous, and to be used together with feature definition files in gff format. The first line defines a fileformat number (should there be any future changes to the format), followed by a date line. The third line defines the gff file(s) describing the entries, and the fourth line defines the sequence fasta file names. For each pair of features is defined the feature type, start and stop. The columns on each line are tab delimited.

```
##cgff-version  1
##date  1998-03-23
##gff   ker.gff
##seq   MMENDOBA        HSKER101
Exon    2347    2742    Exon    2580    2996
Exon    3535    3617    Exon    3738    3820
Exon    3892    4048    Exon    4158    4314
Exon    4688    4852    Exon    4887    5051
Exon    4964    5089    Exon    5137    5262
Exon    5365    5588    Exon    5525    5748
Exon    5910    6030    Exon    6128    6248
```