University of Manitoba

BioLegato: A Programmable, Object-Oriented Graphic User Interface

Dr. Brian Fristensky
Department of Plant Science
University of Manitoba
Winnipeg, Canada

GenomeCanada

MGCB2 Microbial Genomics for Biofuels and Co-Products from Biorefining Processes

GenomePrairie

Bioinformatics Team

UNIVERSITY OF MANITOBA

Intellectual critical mass

Continuity of expertise

Synergies

Faculty
Brian Fristensky

Bioinformaticists
Graham Alvare
Justin Zhang

PhD Student
Abiel Roche

Research Associate
Natalie Bjorklund
Biostatistician TBA

Undergraduate Students
Dale Hamel

GenomeCanada

MGCB2 Microbial Genomics for Biofuels and Co-Products from Biorefining Processes

GenomePrairie

BioLegato - Outline

The Object-Oriented World View

The BioLegato Concept
- End user applications as Objects
- A programmable GUI that calls external programs
- Point and click data pipelining

PCD - Pythonesque Command Definition
- Definition
- Implementation

Pluggable Canvases

BioLegato

Why a Duck?

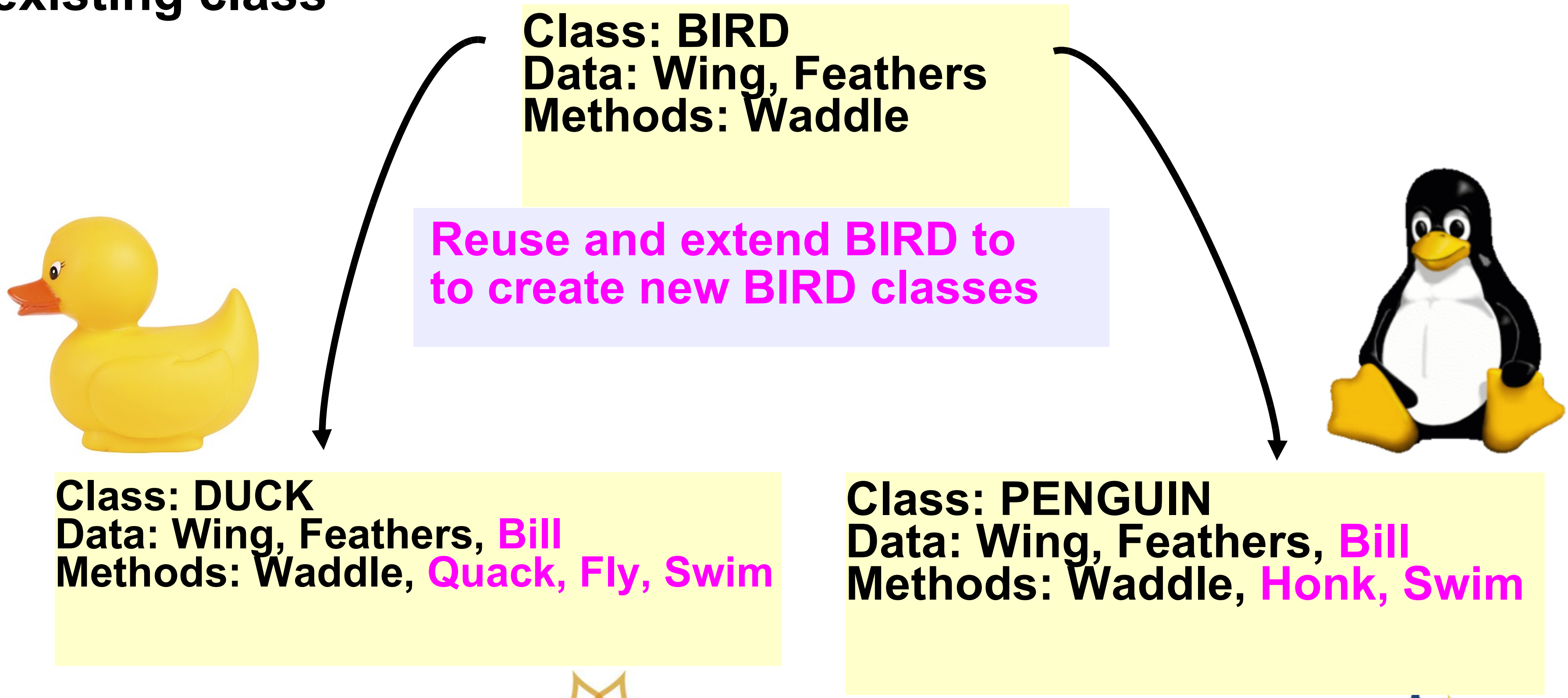
The Object-Oriented World View

Common wisdom: "If it looks like a duck, and waddles like a duck, and quacks like a duck, then it's probably a duck."

OO: If you want a duck, then design it to look like a duck, and waddle like a duck, and quack like a duck.

BioLegato - The OO World View

1. A class is a formula for creating an object
2. Everything is an Object
3. Objects have Data and Methods
4. You can make new classes by reusing and extending an existing class



Class: BIRD
Data: Wing, Feathers
Methods: Waddle

Reuse and extend BIRD to
to create new BIRD classes

Class: DUCK
Data: Wing, Feathers, Bill
Methods: Waddle, Quack, Fly, Swim

Class: PENGUIN
Data: Wing, Feathers, Bill
Methods: Waddle, Honk, Swim

BioLegato - Outline

The Object-Oriented World View

The BioLegato Concept
- End user applications as Objects
- A programmable GUI that calls external programs
- Point and click data pipelining

PCD - Pythonesque Command Definition
- Definition
- Implementation

Pluggable Canvases

GenomeCanada

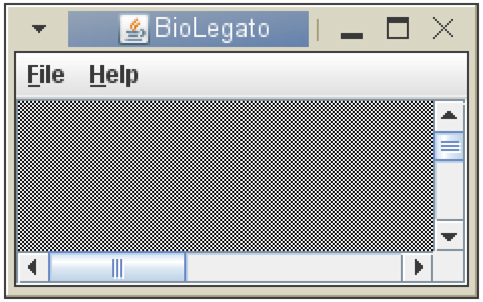
MGCB2 Microbial Genomics for Biofuels and Co-Products from Biorefining Processes

GenomePrairie

The BioLegato Concept

- Unique concept in software design.*
- Almost no functionality is written into the program itself.
- Menus, and a choice of canvas to display data, are read when the BioLegato launches.
- ANY program can be run from BioLegato:
  - Locally-installed programs or scripts in any language
  - Remote web services called by a program or script

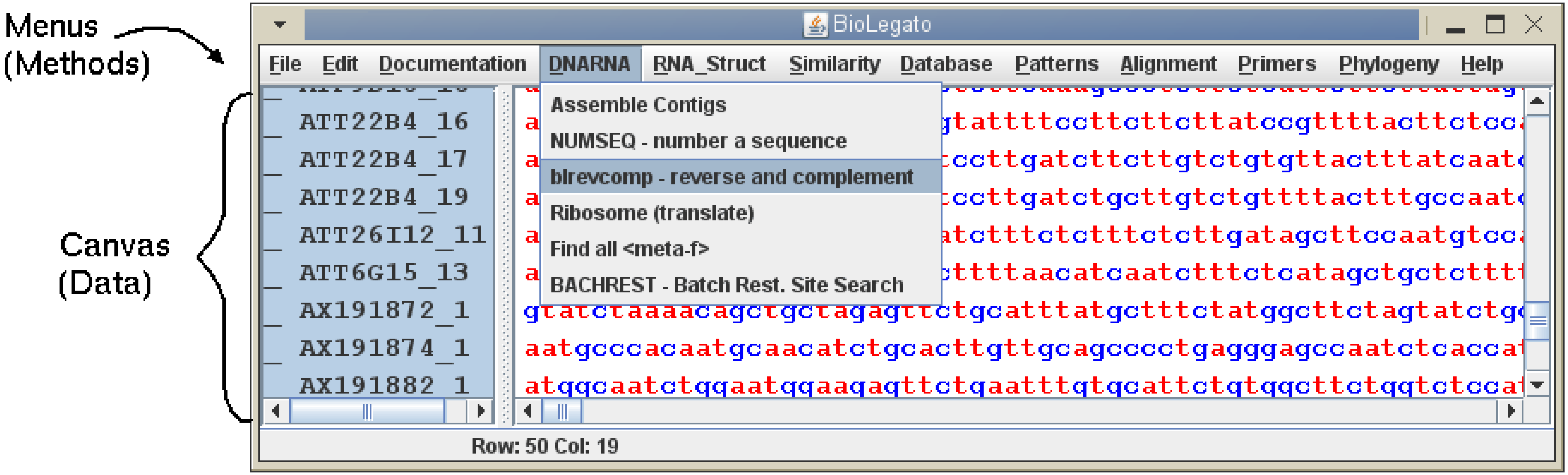
Empty BioLegato interface: null menus, empty canvas



* Inspired by GDE/SeqLab of Steven Smith

BioLegato - Objects have data and methods

Menus read at runtime as either PCD code, or Java



```
a                    gtattttccttcttcttatccgttttacttctcc
a                    ccttgatcttcttgtctgtgttactttatcaat
a                    ccttgatctgcttgtctgttttactttgccaat
a                    atctttctctttctcttgatagcttccaatgtcc
a                    cttttaacatcaatctttctcatagctgctcttt
gtatctaaaacagctgctagagttctgcatttatgctttctatggcttctagtatctg
aatgcccacaatgcaacatctgcacttgttgcagcccctgagggagccaatctcacca
atggcaatctggaatggaagagttctgaatttgtgcattctgtggcttctggtctcca
```

Row: 50 Col: 19

Canvas read at runtime as a Java plugin

GenomeCanada

MGCB² Microbial Genomics for Biofuels and Co-Products from Biorefining Processes

GenomePrairie

BioLegato - Outline

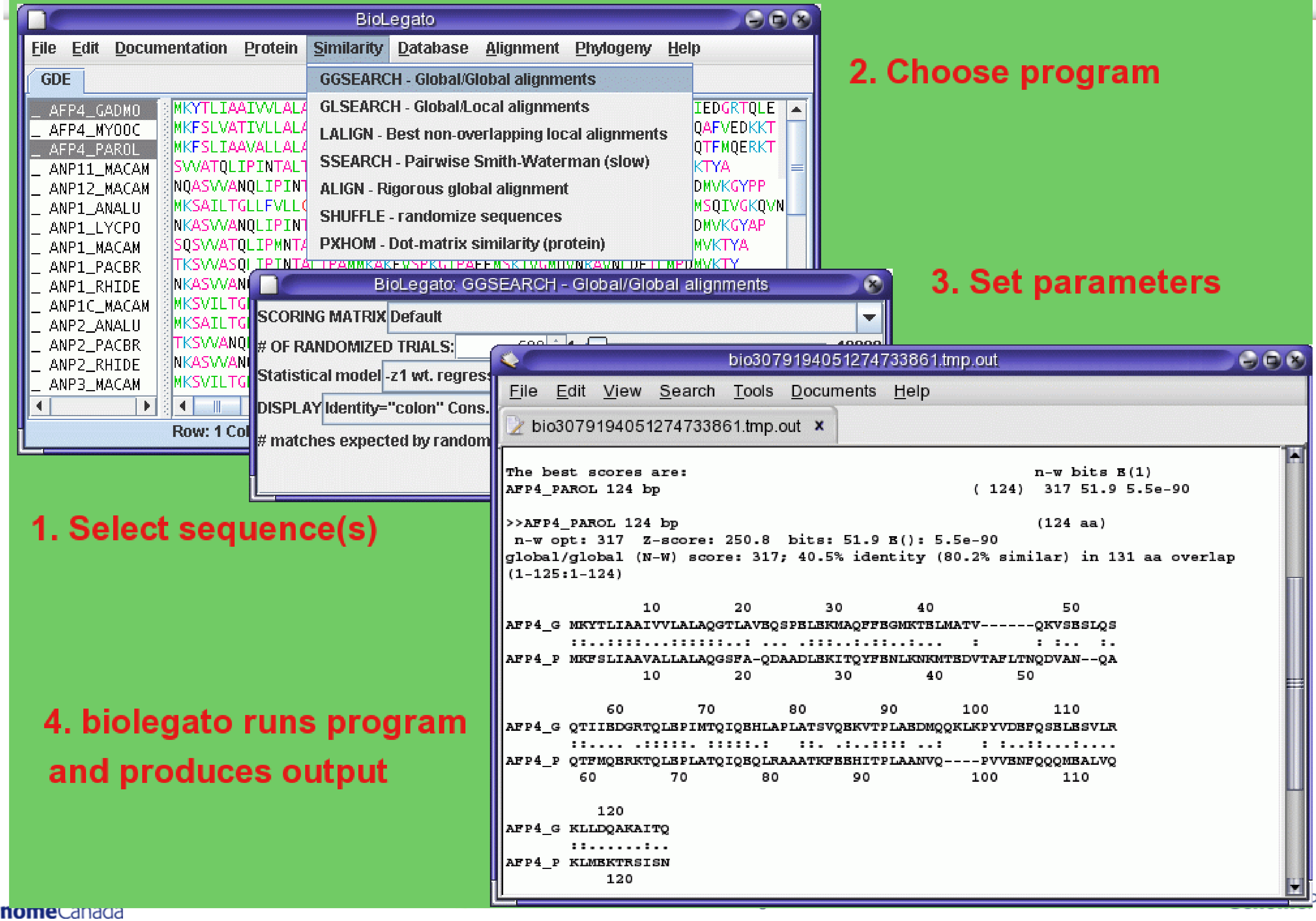
The Object-Oriented World View

The BioLegato Concept
- End user applications as Objects
- A programmable GUI that calls external programs
- Point and click data pipelining

PCD - Pythonesque Command Definition
- Definition
- Implementation

Pluggable Canvases

BioLegato executes external programs


BioLegato
File Edit Documentation Protein Similarity Database Alignment Phylogeny Help
GDE
AFP4_GADMO
AFP4_MYOOC
AFP4_PAROL
ANP11_MACAM
ANP12_MACAM
ANP1_ANALU
ANP1_LYCPO
ANP1_MACAM
ANP1_PACBR
ANP1_RHIDE
ANP1C_MACAM
ANP2_ANALU
ANP2_PACBR
ANP2_RHIDE
ANP3_MACAM
Row: 1 Col
GGSEARCH - Global/Global alignments
GLSEARCH - Global/Local alignments
LALIGN - Best non-overlapping local alignments
SSEARCH - Pairwise Smith-Waterman (slow)
ALIGN - Rigorous global alignment
SHUFFLE - randomize sequences
PXHOM - Dot-matrix similarity (protein)
BioLegato: GGSEARCH - Global/Global alignments
bio3079194051274733861.tmp.out
File Edit View Search Tools Documents Help

SCORING MATRIX Default

OF RANDOMIZED TRIALS:

Statistical model -z1 wt. regress

DISPLAY Identity="colon" Cons.

matches expected by random

```
The best scores are:                                      n-w bits E(1)
AFP4_PAROL 124 bp                                  ( 124)  317 51.9 5.5e-90

>>AFP4_PAROL 124 bp                                       (124 aa)
 n-w opt: 317  Z-score: 250.8  bits: 51.9 E(): 5.5e-90
global/global (N-W) score: 317; 40.5% identity (80.2% similar) in 131 aa overlap
(1-125:1-124)

               10        20        30        40              50
AFP4_G MKYTLIAAIVVLALAQGTLAVEQSPELEKMAQFFEGMKTELMATV------QKVSESLQS
       ::..::::...::::::..: ... .:::..:.::..:...   :      : :..  :.
AFP4_P MKFSLIAAVALLALAQGSFA-QDAADLEKITQYFENLKNKMTEDVTAFLTNQDVAN--QA
               10        20         30        40        50

           60        70        80        90       100       110
AFP4_G QTIIEDGRTQLEPIMTQIQEHLAPLATSVQEKVTPLAEDMQQKLKPYVDEFQSELESVLR
       ::.... .:::::. :::::.:   ::. .:..:::: ..:    : :..::...:....
AFP4_P QTFMQERKTQLEPLATQIQEQLRAAATKFEEHITPLAANVQ----PVVENFQQQMEALVQ
        60        70        80        90           100       110

          120
AFP4_G KLLDQAKAITQ
       ::......:..
AFP4_P KLMEKTRSISN
           120
```

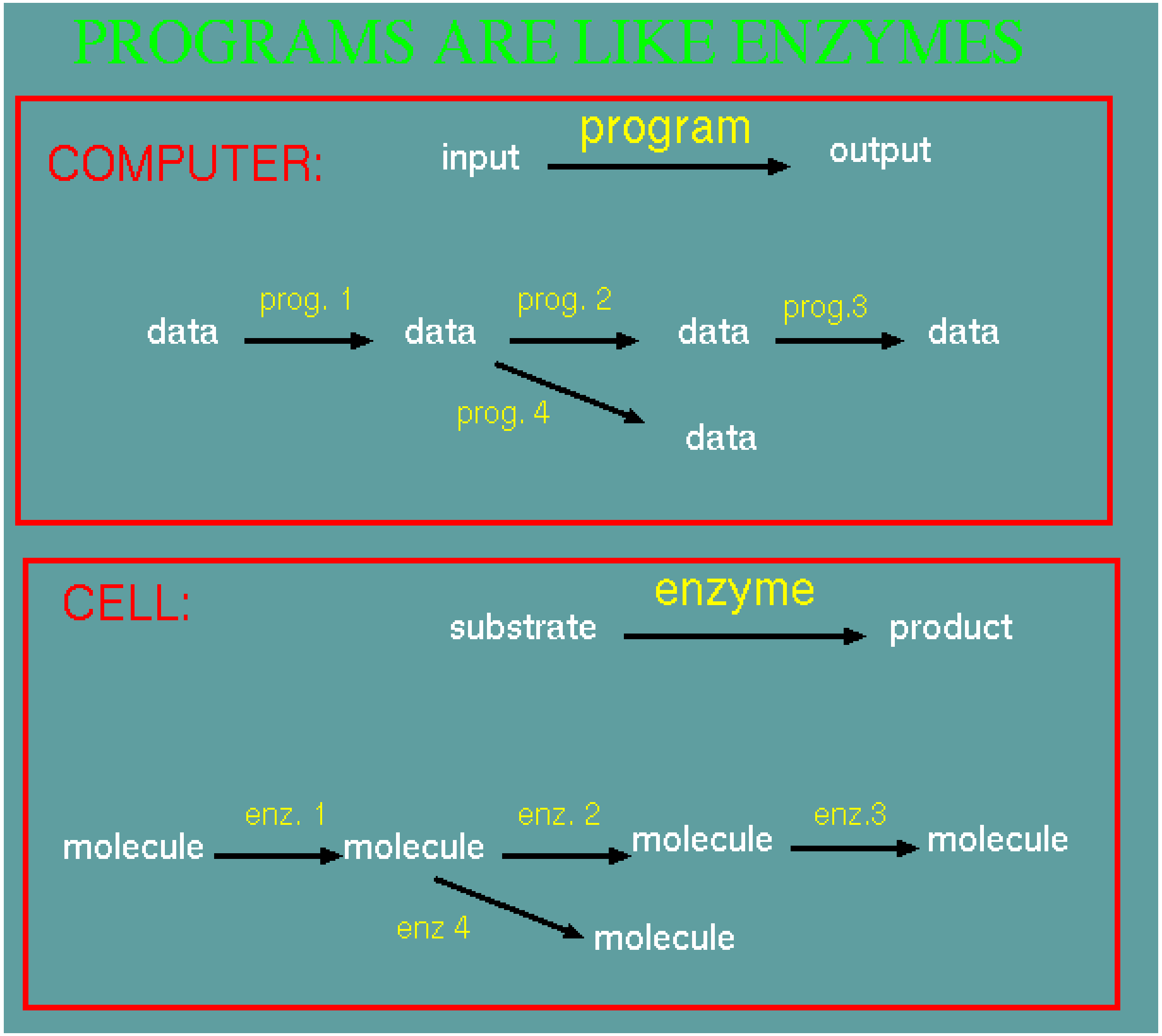
1. Select sequence(s)

2. Choose program

3. Set parameters

4. biolegato runs program and produces output

Complex behavior from a simple rule



PROGRAMS ARE LIKE ENZYMES

COMPUTER:

input program output

data prog. 1 data prog. 2 data prog.3 data

prog. 4 data

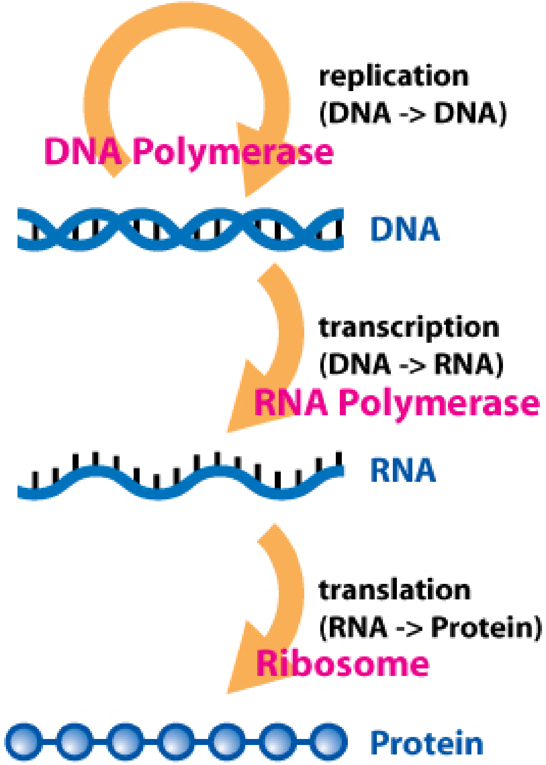
CELL:

substrate enzyme product

molecule enz. 1 molecule enz. 2 molecule enz.3 molecule

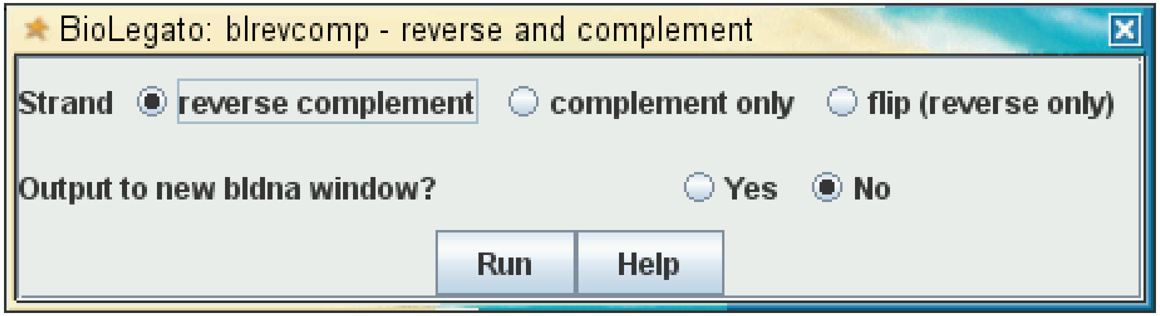
enz 4 molecule

DNA --> RNA --> protein; disk-->RAM->bytecode



PCD - GUI scripting language

```
name          "blrevcomp - reverse and complement"
var "in1"
    type        tempfile
    direction   in
    format      flat
var "strand"
    type        chooser
    label       "Strand"
    default     0
    choices
        "reverse complement"   "-r"
        "complement only"      "-c"
        "flip (reverse only)"  "-f"
var "out1"
    type        tempfile
    direction   out
    format      flat
var "gdeoutput"
    type        chooser
    label       "Output to new bldna window?"
    default     1
    choices
        "Yes"       "(echo '' > %out1%; bldna %in1%.blrevcomp; $RM_CMD -f %in1%.blrevcomp)&"
        "No"        "mv %in1%.blrevcomp %out1%"
panel
    var "Run"
        type        button
        label       "Run"
        shell       "blrevcomp %STRAND% %in1% %in1%.blrevcomp; $RM_CMD %in1%; %GDEOUTPUT%"
        close       true
    var "Help"
        type        button
        label       "Help"
        shell       "$BIRCH/script/gde_help_viewer.csh null/$BIRCH/doc/BioLegato/blrevcomp.html"
        close       false
```



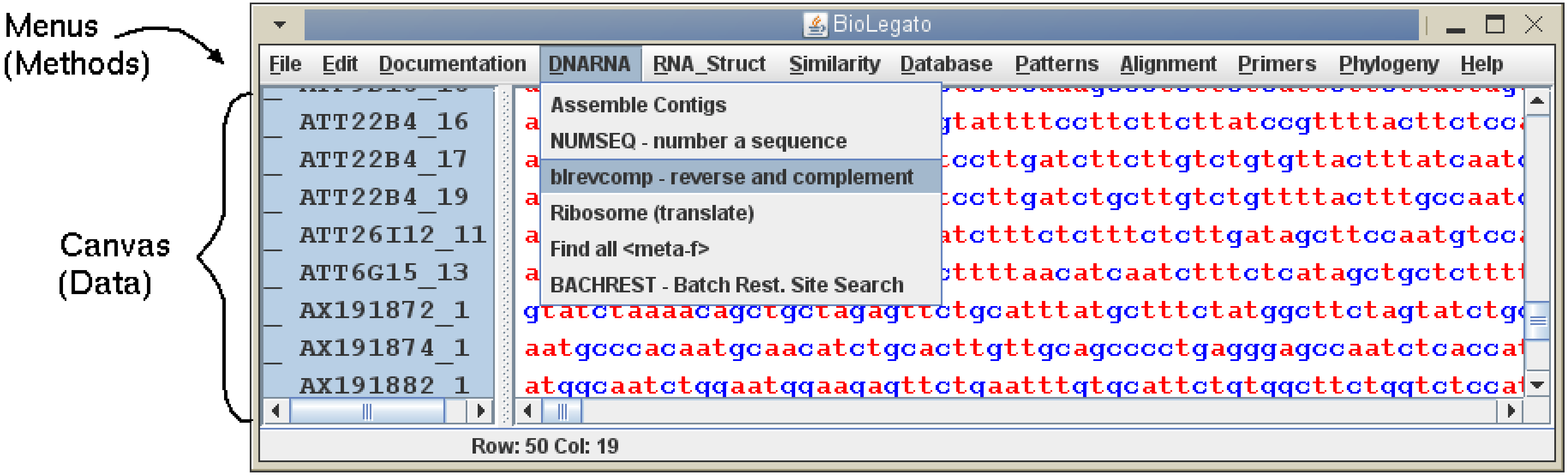
BioLegato: blrevcomp - reverse and complement

Strand ● reverse complement ○ complement only ○ flip (reverse only)

Output to new bldna window? ○ Yes ● No

BioLegato - Objects have data and methods

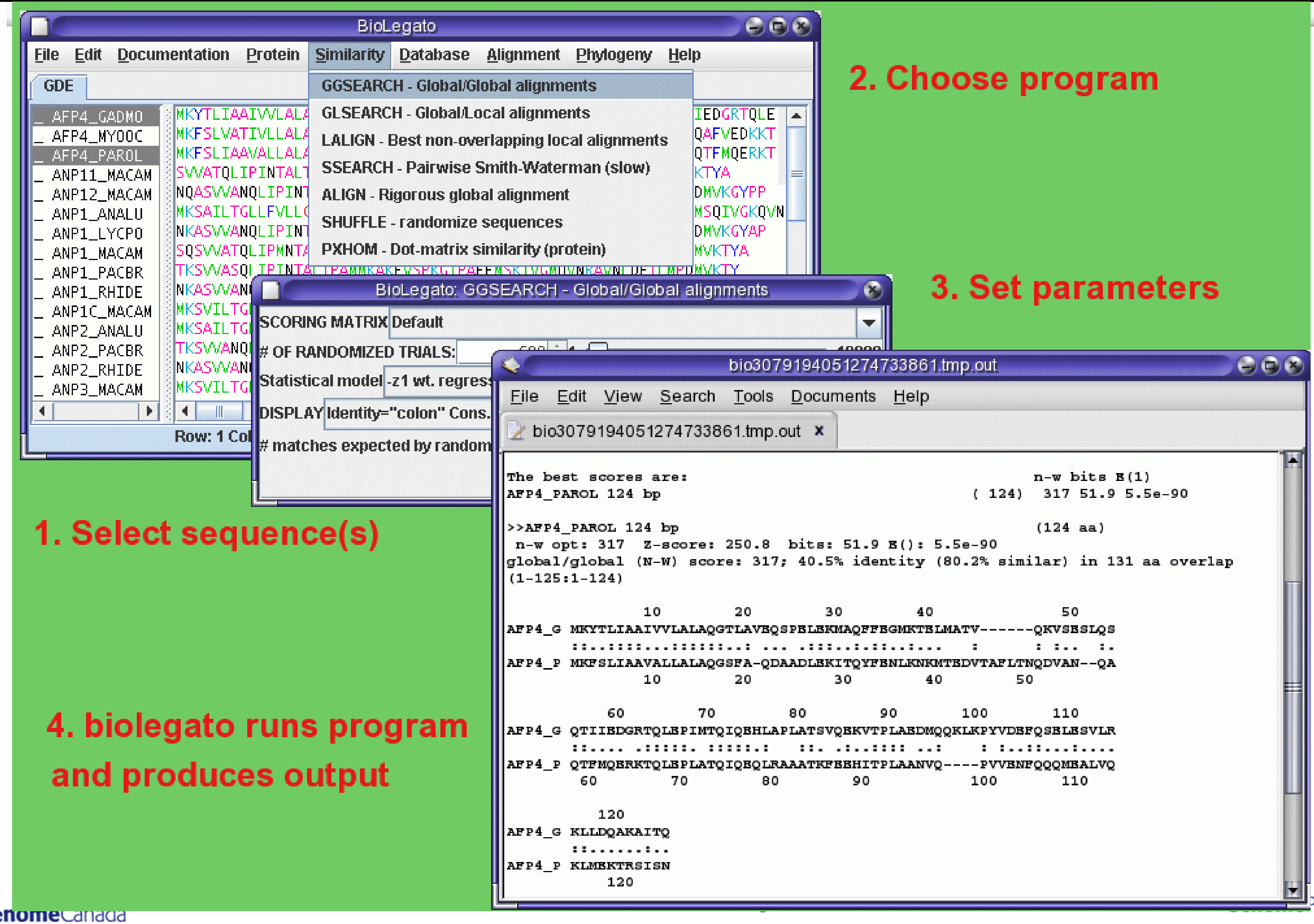
Menus read at runtime as either PCD code, or Java



```
ATT22B4_16     a  gtattttccttcttcttatccgttttacttctcc
ATT22B4_17     a  ccttgatcttcttgtctgtgttactttatcaat
ATT22B4_19     a  ccttgatctgcttgtctgttttactttgccaat
ATT26I12_11    a  atctttctctttctcttgatagcttccaatgtcc
ATT6G15_13     a  cttttaacatcaatctttctcatagctgctttt
AX191872_1     gtatctaaaacagctgctagagttctgcatttatgctttctatggcttctagtatctg
AX191874_1     aatgcccacaatgcaacatctgcacttgttgcagcccctgagggagccaatctcacca
AX191882_1     atggcaatctggaatggaagagttctgaatttgtgcattctgtggcttctggtctcca
Row: 50 Col: 19
```

Canvas read at runtime as a Java plugin

BioLegato executes external programs


BioLegato
File Edit Documentation Protein Similarity Database Alignment Phylogeny Help
GDE
GGSEARCH - Global/Global alignments
GLSEARCH - Global/Local alignments
LALIGN - Best non-overlapping local alignments
SSEARCH - Pairwise Smith-Waterman (slow)
ALIGN - Rigorous global alignment
SHUFFLE - randomize sequences
PXHOM - Dot-matrix similarity (protein)
_ AFP4_GADMO
_ AFP4_MYOOC
_ AFP4_PAROL
_ ANP11_MACAM
_ ANP12_MACAM
_ ANP1_ANALU
_ ANP1_LYCPO
_ ANP1_MACAM
_ ANP1_PACBR
_ ANP1_RHIDE
_ ANP1C_MACAM
_ ANP2_ANALU
_ ANP2_PACBR
_ ANP2_RHIDE
_ ANP3_MACAM
Row: 1 Col
BioLegato: GGSEARCH - Global/Global alignments
bio3079194051274733861.tmp.out
File Edit View Search Tools Documents Help
bio3079194051274733861.tmp.out

SCORING MATRIX Default

OF RANDOMIZED TRIALS:

Statistical model -z1 wt. regress

DISPLAY Identity="colon" Cons.

matches expected by random

```
The best scores are:                                      n-w bits E(1)
AFP4_PAROL 124 bp                                   ( 124)  317 51.9 5.5e-90

>>AFP4_PAROL 124 bp                                       (124 aa)
 n-w opt: 317  Z-score: 250.8  bits: 51.9 E(): 5.5e-90
global/global (N-W) score: 317; 40.5% identity (80.2% similar) in 131 aa overlap
(1-125:1-124)

               10        20        30        40              50
AFP4_G MKYTLIAAIVVLALAQGTLAVEQSPELEKMAQFFEGMKTELMATV------QKVSESLQS
       ::..::::..::::::::..: ... .::::..:::::.:....   :      : :..  :.
AFP4_P MKFSLIAAVALLALAQGSFA-QDAADLEKITQYFENLKNKMTEDVTAFLTNQDVAN--QA
               10        20          30        40        50

           60        70        80        90        100       110
AFP4_G QTIIEDGRTQLEPIMTQIQEHLAPLATSVQEKVTPLAEDMQQKLKPYVDEFQSELESVLR
       ::.... .:::::. ::::::.:    ::. .:..:::: ..:     : :..::...:....
AFP4_P QTFMQERKTQLEPLATQIQEQLRAAATKFEEHITPLAANVQ----PVVENFQQQMEALVQ
        60        70        80        90             100       110

          120
AFP4_G KLLDQAKAITQ
       ::.......:..
AFP4_P KLMEKTRSISN
           120
```

1. Select sequence(s)

2. Choose program

3. Set parameters

4. biolegato runs program and produces output

BioLegato - Outline

The Object-Oriented World View

The BioLegato Concept

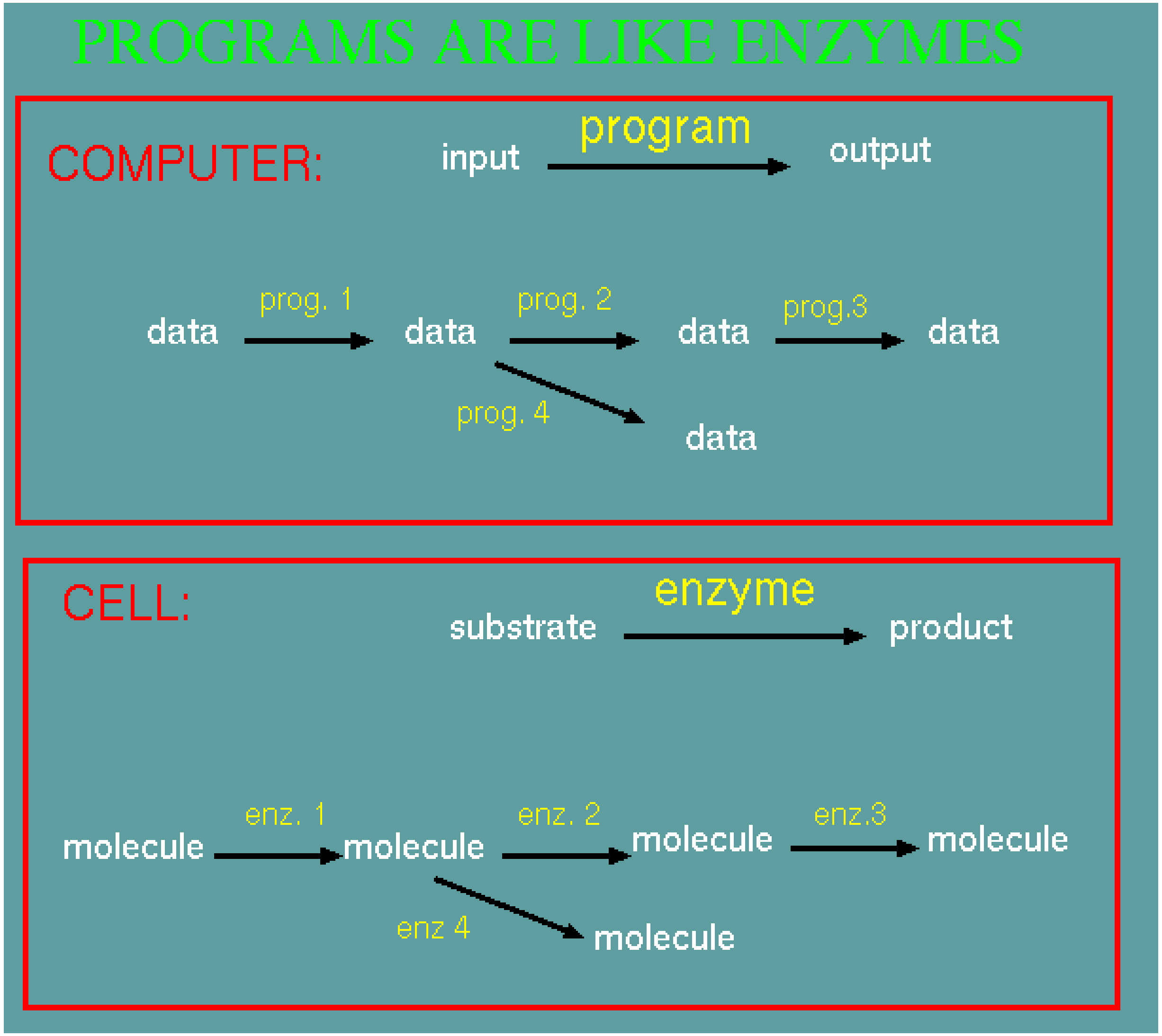
- End user applications as Objects
- A programmable GUI that calls external programs
- Point and click data pipelining

PCD - Pythonesque Command Definition

- Definition
- Implementation

Pluggable Canvases

Complex behavior from a simple rule



PROGRAMS ARE LIKE ENZYMES

COMPUTER:

input $\xrightarrow{\text{program}}$ output

data $\xrightarrow{\text{prog. 1}}$ data $\xrightarrow{\text{prog. 2}}$ data $\xrightarrow{\text{prog.3}}$ data

data $\xrightarrow{\text{prog. 4}}$ data

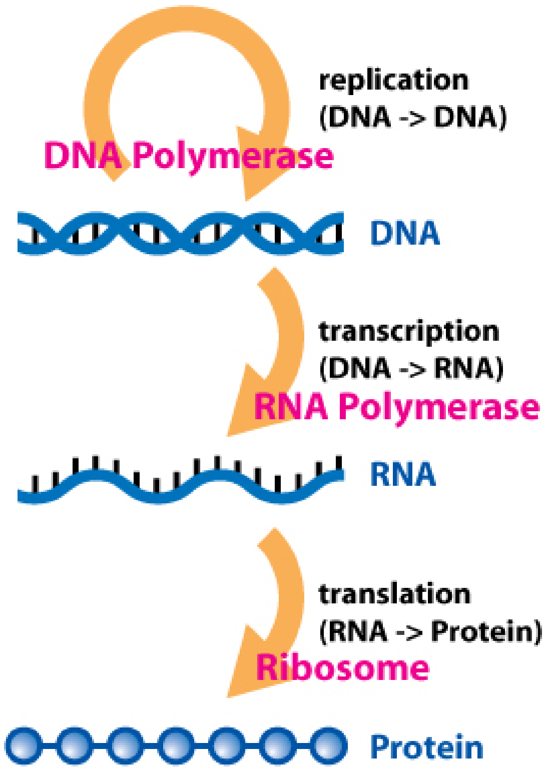
CELL:

substrate $\xrightarrow{\text{enzyme}}$ product

molecule $\xrightarrow{\text{enz. 1}}$ molecule $\xrightarrow{\text{enz. 2}}$ molecule $\xrightarrow{\text{enz.3}}$ molecule

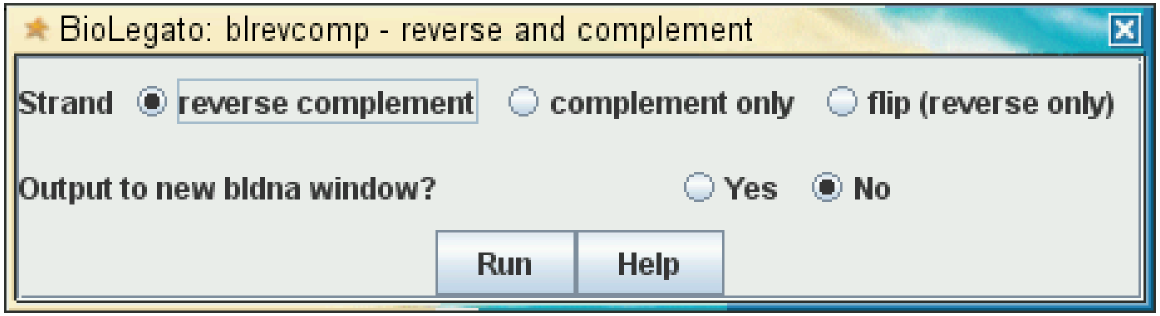
molecule $\xrightarrow{\text{enz 4}}$ molecule

DNA --> RNA --> protein; disk-->RAM->bytecode



PCD - GUI scripting language

```
name        "blrevcomp - reverse and complement"
var "in1"
    type        tempfile
    direction   in
    format      flat
var "strand"
    type        chooser
    label       "Strand"
    default     0
    choices
        "reverse complement"  "-r"
        "complement only"     "-c"
        "flip (reverse only)" "-f"
var "out1"
    type        tempfile
    direction   out
    format      flat
var "gdeoutput"
    type        chooser
    label       "Output to new bldna window?"
    default     1
    choices
        "Yes"       "(echo '' > %out1%; bldna %in1%.blrevcomp; $RM_CMD -f %in1%.blrevcomp)&"
        "No"        "mv %in1%.blrevcomp %out1%"
panel
    var "Run"
        type        button
        label       "Run"
        shell       "blrevcomp %STRAND% %in1% %in1%.blrevcomp; $RM_CMD %in1%; %GDEOUTPUT%"
        close       true
    var "Help"
        type        button
        label       "Help"
        shell       "$BIRCH/script/gde_help_viewer.csh null/$BIRCH/doc/BioLegato/blrevcomp.html"
        close       false
```

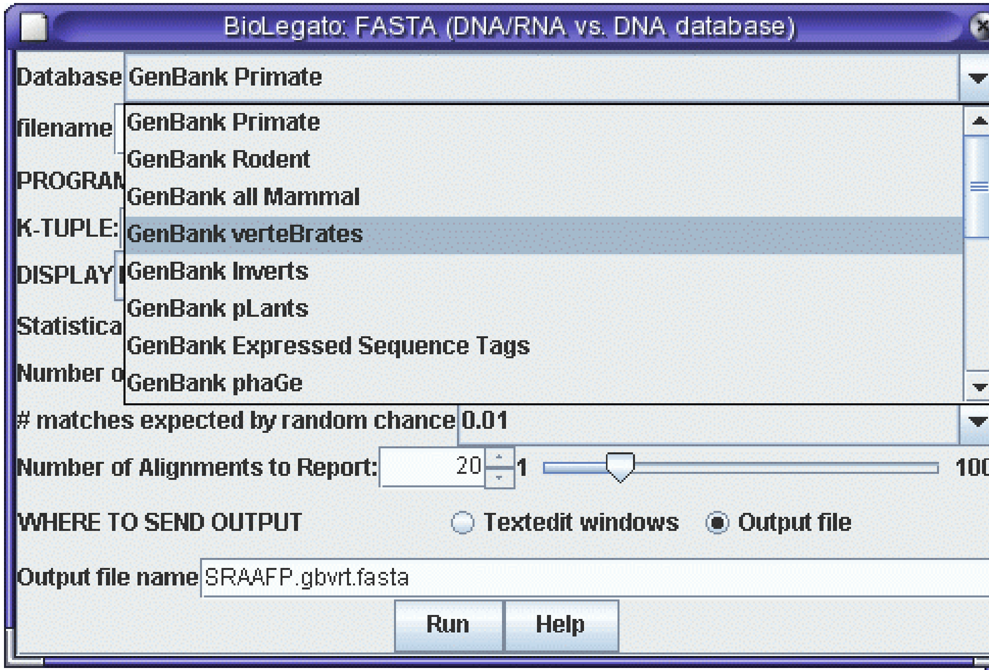


BioLegato: blrevcomp - reverse and complement

Strand ● reverse complement ○ complement only ○ flip (reverse only)

Output to new bldna window? ○ Yes ● No

Examples of menus



BioLegato: FASTA (DNA/RNA vs. DNA database)

Database: GenBank Primate

- GenBank Primate
- GenBank Rodent
- GenBank all Mammal
- GenBank verteBrates
- GenBank Inverts
- GenBank pLants
- GenBank Expressed Sequence Tags
- GenBank phaGe

filename

PROGRAM

K-TUPLE:

DISPLAY

Statistica

Number o

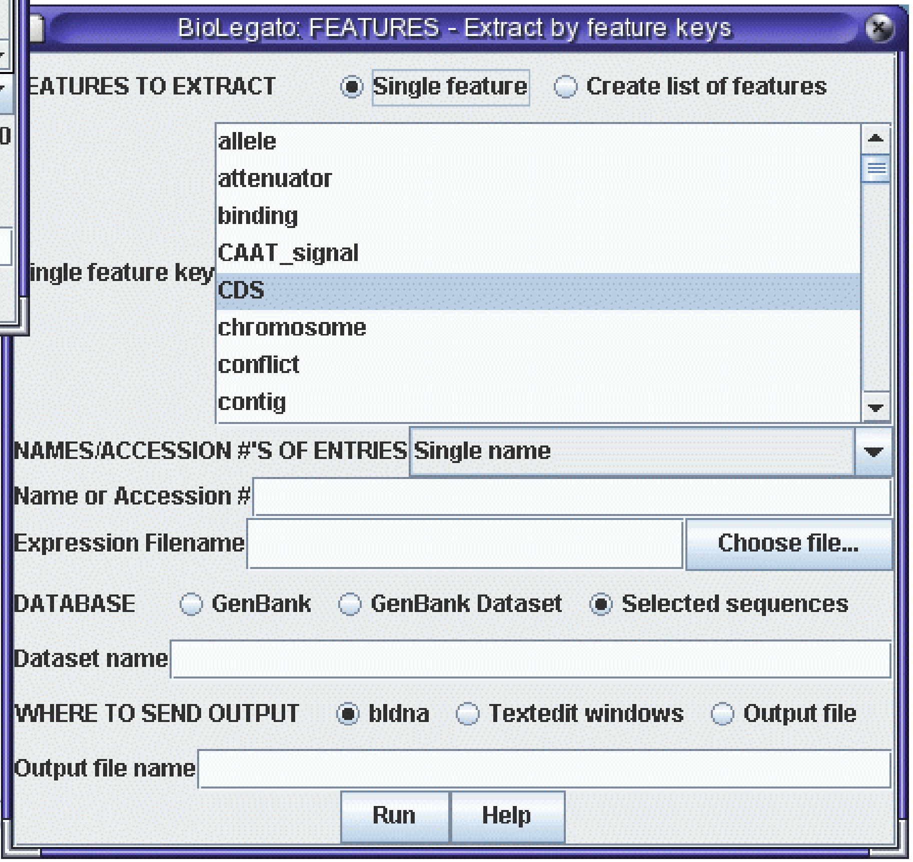
matches expected by random chance: 0.01

Number of Alignments to Report: 20 (slider 1 – 100)

WHERE TO SEND OUTPUT: ○ Textedit windows ● Output file

Output file name: SRAAFP.gbvrt.fasta

Run | Help



BioLegato: FEATURES - Extract by feature keys

EATURES TO EXTRACT: ● Single feature ○ Create list of features

ingle feature key:

- allele
- attenuator
- binding
- CAAT_signal
- CDS
- chromosome
- conflict
- contig

NAMES/ACCESSION #'S OF ENTRIES: Single name

Name or Accession #:

Expression Filename: Choose file...

DATABASE: ○ GenBank ○ GenBank Dataset ● Selected sequences

Dataset name:

WHERE TO SEND OUTPUT: ● bldna ○ Textedit windows ○ Output file

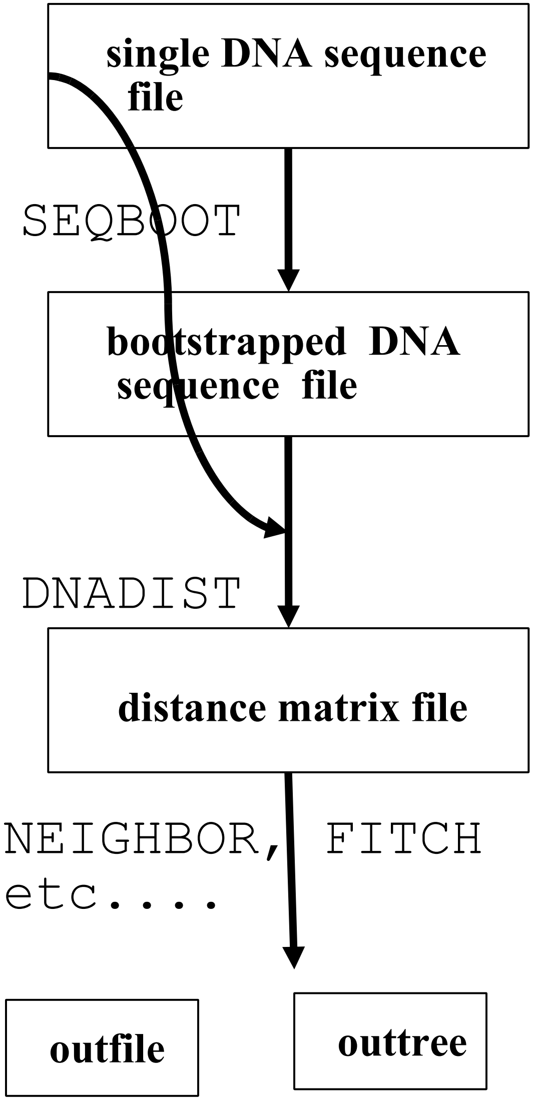
Output file name:

Run | Help

GenomeCanada

MGCB² Microbial Genor Co-Products fro

Setting parameters for a workflow


single DNA sequence file

SEQBOOT

bootstrapped DNA sequence file

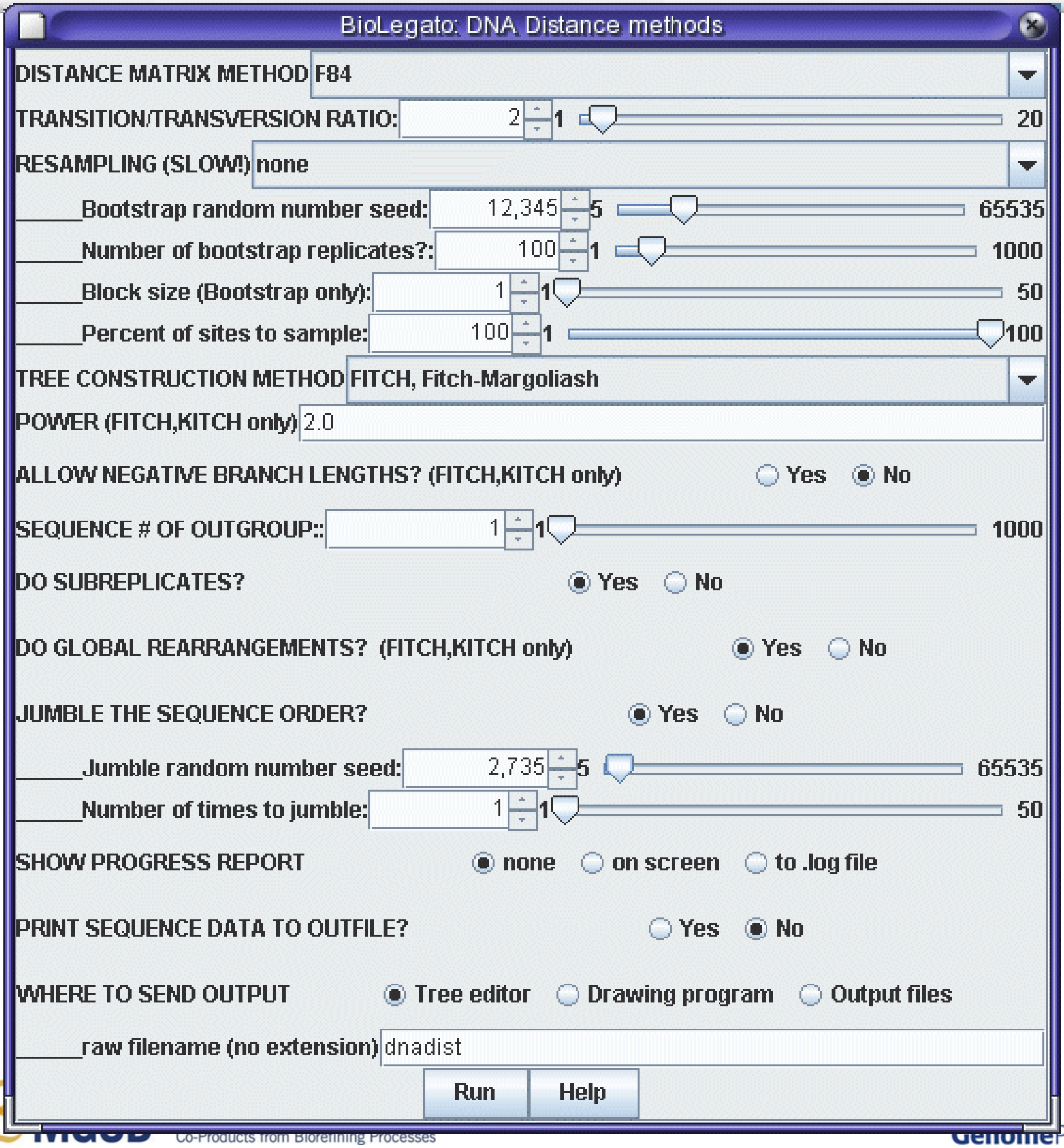
DNADIST

distance matrix file

NEIGHBOR, FITCH etc....

outfile

outtree



DISTANCE MATRIX METHOD F84

TRANSITION/TRANSVERSION RATIO: 2 (1–20)

RESAMPLING (SLOW!) none

_____Bootstrap random number seed: 12,345 (5–65535)

_____Number of bootstrap replicates?: 100 (1–1000)

_____Block size (Bootstrap only): 1 (1–50)

_____Percent of sites to sample: 100 (1–100)

TREE CONSTRUCTION METHOD FITCH, Fitch-Margoliash

POWER (FITCH,KITCH only) 2.0

ALLOW NEGATIVE BRANCH LENGTHS? (FITCH,KITCH only) ○ Yes ● No

SEQUENCE # OF OUTGROUP:: 1 (1–1000)

DO SUBREPLICATES? ● Yes ○ No

DO GLOBAL REARRANGEMENTS? (FITCH,KITCH only) ● Yes ○ No

JUMBLE THE SEQUENCE ORDER? ● Yes ○ No

_____Jumble random number seed: 2,735 (5–65535)

_____Number of times to jumble: 1 (1–50)

SHOW PROGRESS REPORT ● none ○ on screen ○ to .log file

PRINT SEQUENCE DATA TO OUTFILE? ○ Yes ● No

WHERE TO SEND OUTPUT ● Tree editor ○ Drawing program ○ Output files

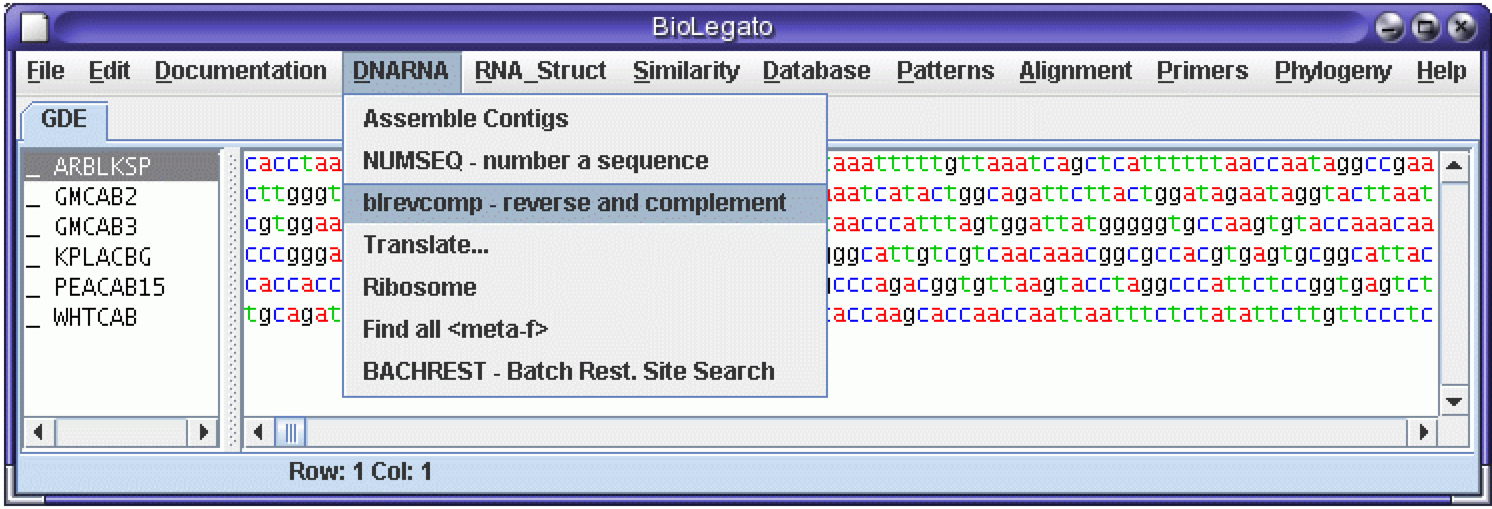
_____raw filename (no extension) dnadist

BioLegato - menus + canvas = new GUI

A GUI that will run just about any program for any kind of data can easily be made by writing new menus and choosing a canvas.

bldna -

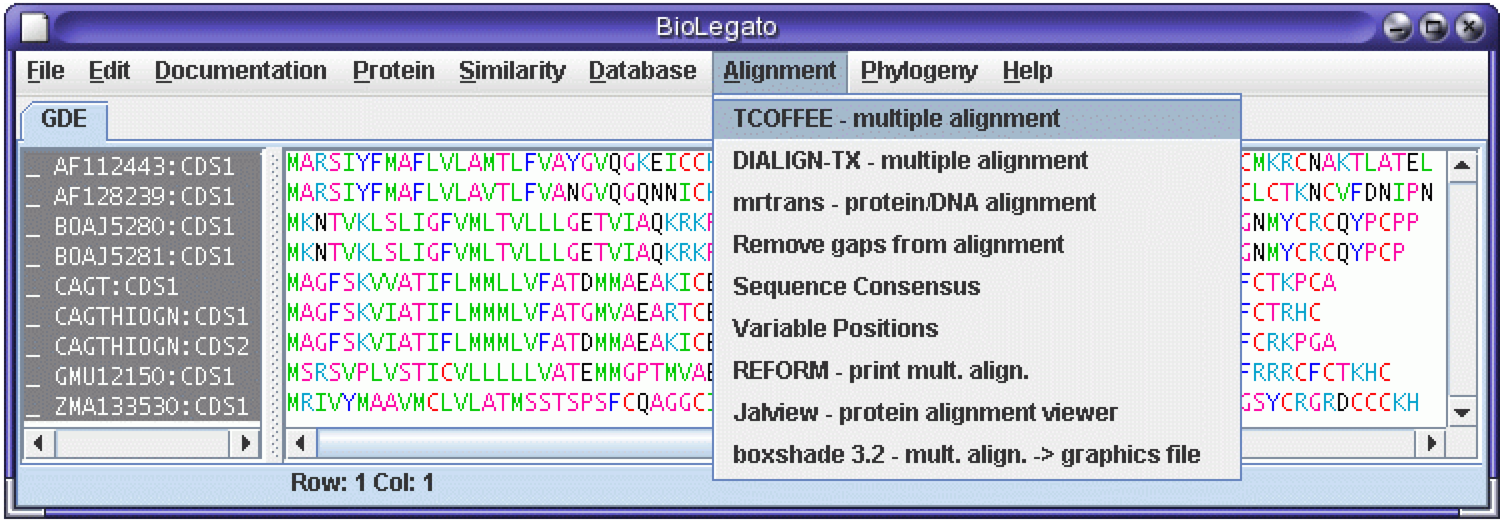
- **menus for DNA programs**
- **multiple alignment canvas**



```
_ ARBLKSP    cacctaa   aaattttgttaaatcagctcatttttaaccaataggccgaa
_ GMCAB2     cttgggt   aatcatactggcagattcttactggatagaataggtacttaat
_ GMCAB3     cgtggaa   aacccatttagtggattatgggggtgccaagtgtaccaaacaa
_ KPLACBG    cccggga   ggcattgtcgtcaacaaacggcgccacgtgagtgcggcattac
_ PEACAB15   caccacc   cccagacggtgttaagtacctaggcccattctccggtgagtct
_ WHTCAB     tgcagat   accaagcaccaaccaattaatttctctatattcttgttccctc
```

blprotein -

- **menus for protein programs**
- **multiple alignment canvas**


BioLegato
File
Edit
Documentation
Protein
Similarity
Database
Alignment
Phylogeny
Help
GDE
TCOFFEE - multiple alignment
DIALIGN-TX - multiple alignment
mrtrans - protein/DNA alignment
Remove gaps from alignment
Sequence Consensus
Variable Positions
REFORM - print mult. align.
Jalview - protein alignment viewer
boxshade 3.2 - mult. align. -> graphics file
Row: 1 Col: 1

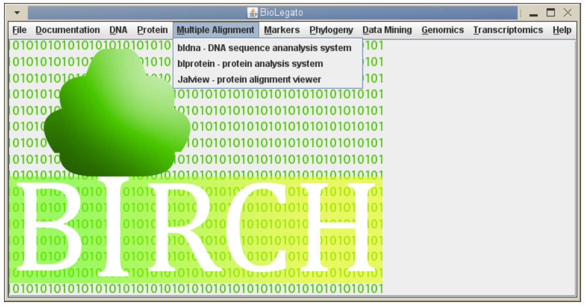
```
_ AF112443:CDS1     MARSIYFMAFLVLAMTLFVAYGVQGKEICC   CMKRCNAKTLATEL
_ AF128239:CDS1     MARSIYFMAFLVLAVTLFVANGVQGQNNIC   CLCTKNCVFDNIPN
_ B0AJ5280:CDS1     MKNTVKLSLIGFVMLTVLLLGETVIAQKRK   GNMYCRCQYPCPP
_ B0AJ5281:CDS1     MKNTVKLSLIGFVMLTVLLLGETVIAQKRK   GNMYCRCQYPCP
_ CAGT:CDS1         MAGFSKVVATIFLMMLLVFATDMMAEAKIC   FCTKPCA
_ CAGTHIOGN:CDS1    MAGFSKVIATIFLMMMLVFATGMVAEARTC   FCTRHC
_ CAGTHIOGN:CDS2    MAGFSKVIATIFLMMMLVFATDMMAEAKIC   FCRKPGA
_ GMU12150:CDS1     MSRSVPLVSTICVLLLLLVATEMMGPTMVA   FRRRCFCTKHC
_ ZMA133530:CDS1    MRIVYMAAVMCLVLATMSSTSPSFCQAGGC   GSYCRGRDCCCKH
```

GenomeCanada

BioLegato - rapid deployment of new GUIs

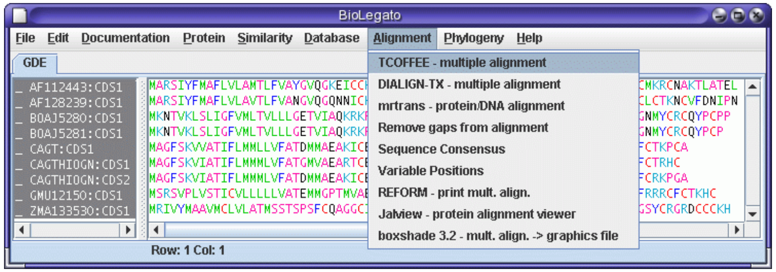
GENOGRAPHER
WEIGHBOR
MRBAYES
ARTEMIS
DIALIGN
GLSEARCH
CN3D
FASTXY
BACHREST

birch - application launcher



DRAWTREE
TCOFFEE
ATV
SEQUIN
NUMSEQ
BLREVCOMP
RIBOSOME
TMEV
DIGEST

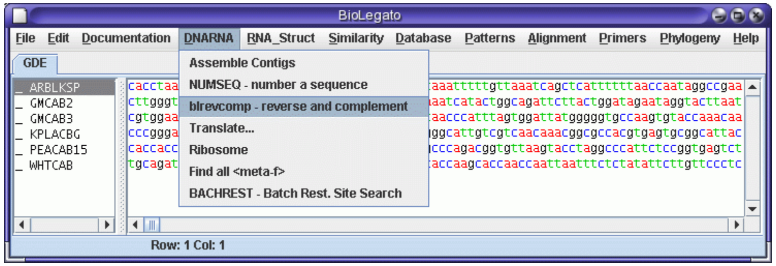
blprotein - protein sequence analysis


BioLegato
File
Edit
Documentation
Protein
Similarity
Database
Alignment
Phylogeny
Help
GDE
TCOFFEE - multiple alignment
DIALIGN-TX - multiple alignment
mrtrans - protein/DNA alignment
Remove gaps from alignment
Sequence Consensus
Variable Positions
REFORM - print mult. align.
Jalview - protein alignment viewer
boxshade 3.2 - mult. align. -> graphics file

```
_ AF112443:CDS1    MARSIYFMAFLVLAMTLFVAYGVQGKEICC    CMKRCNAKTLATEL
_ AF128239:CDS1    MARSIYFMAFLVLAVTLFVANGVQGQNNIC    CLCTKNCVFDNIPN
_ B0AJ5280:CDS1    MKNTVKLSLIGFVMLTVLLLGETVIAQKRK    GNMYCRCQYPCPP
_ B0AJ5281:CDS1    MKNTVKLSLIGFVMLTVLLLGETVIAQKRK    GNMYCRCQYPCP
_ CAGT:CDS1        MAGFSKVVATIFLMMLLVFATDMMAEAKIC    FCTKPCA
_ CAGTHIOGN:CDS1   MAGFSKVIATIFLMMMLVFATGMVAEARTC    FCTRHC
_ CAGTHIOGN:CDS2   MAGFSKVIATIFLMMMLVFATDMMAEAKIC    FCRKPGA
_ GMU12150:CDS1    MSRSVPLVSTICVLLLLLVATEMMGPTMVA    FRRRCFCTKHC
_ ZMA133530:CDS1   MRIVYMAAVMCLVLATMSSTSPSFCQAGGC    GSYCRGRDCCCKH
```

Row: 1 Col: 1

bldna - DNA sequence analysis

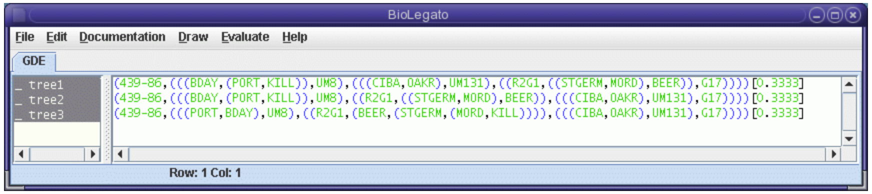


```
_ ARBLKSP    cacctaa    aaatttttgttaaatcagctcatttttaaccaataggccgaa
_ GMCAB2     cttgggt    aatcatactggcagattcttactggatagaataggtacttaat
_ GMCAB3     cgtggaa    aacccatttagtggattatgggggtgccaagtgtaccaaacaa
_ KPLACBG    cccggga    ggcattgtcgtcaacaaacggcgccacgtgagtgcggcattac
_ PEACAB15   caccacc    cccagacggtgttaagtacctaggcccattctccggtgagtct
_ WHTCAB     tgcagat    accaagcaccaaccaattaatttctctatattcttgttccctc
```

Row: 1 Col: 1

PRIMER3
SHUFFLE

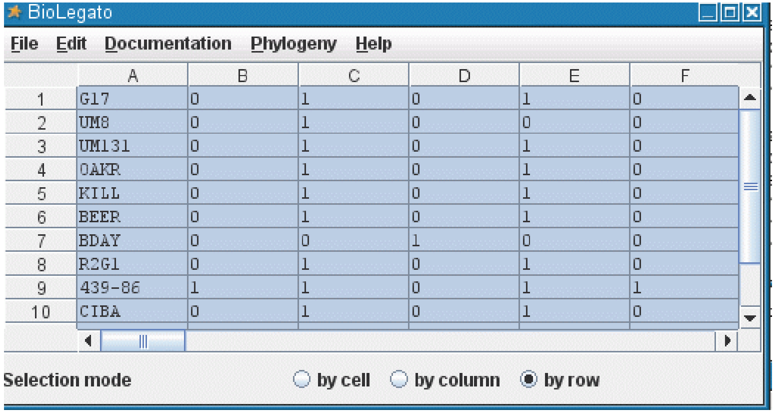
bltree - draw and evaluate trees



```
_ tree1   (439-86,(((BDAY,(PORT,KILL)),UM8),(((CIBA,OAKR),UM131),((R2G1,((STGERM,MORD),BEER)),G17)))) [0.3333]
_ tree2   (439-86,(((BDAY,(PORT,KILL)),UM8),((R2G1,((STGERM,MORD),BEER)),(((CIBA,OAKR),UM131),G17)))) [0.3333]
_ tree3   (439-86,(((PORT,BDAY),UM8),((R2G1,(BEER,(STGERM,(MORD,KILL)))),(((CIBA,OAKR),UM131),G17)))) [0.3333]
```

Row: 1 Col: 1

blmarker - molecular markers



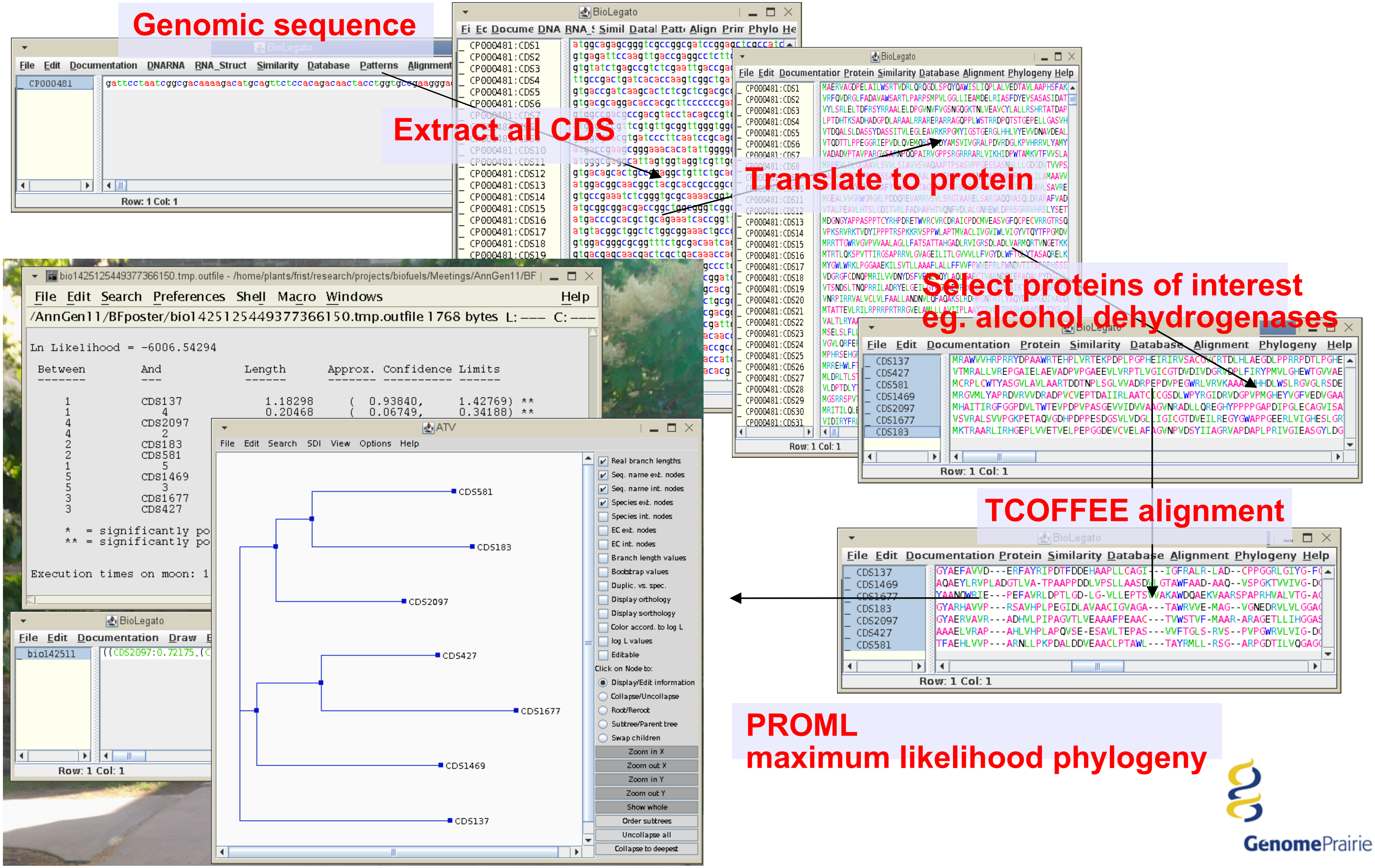
	A	B	C	D	E	F
1	G17	0	1	0	1	0
2	UM8	0	1	0	0	0
3	UM131	0	1	0	1	0
4	OAKR	0	1	0	1	0
5	KILL	0	1	0	1	0
6	BEER	0	1	0	1	0
7	BDAY	0	0	1	0	0
8	R2G1	0	1	0	1	0
9	439-86	1	1	0	1	1
10	CIBA	0	1	0	1	0

Selection mode ○ by cell ○ by column ● by row

DOLLOP
CLUSTALW

OntarioGenomicsInstitute

BioLegato - Point and click data pipelining



Genomic sequence

Extract all CDS

Translate to protein

Select proteins of interest
eg. alcohol dehydrogenases

TCOFFEE alignment

PROML
maximum likelihood phylogeny

```
/AnnGen11/BFposter/bio1425125449377366150.tmp.outfile 1768 bytes L: --- C: ---

Ln Likelihood = -6006.54294

 Between        And            Length      Approx. Confidence Limits
 -------        ---            ------      ------- ---------- ------

      1          CDS137           1.18298     (   0.93840,     1.42769) **
      1             4             0.20468     (   0.06749,     0.34188) **
      4          CDS2097
      4             2
      2          CDS183
      2          CDS581
      1             5
      5          CDS1469
      5             3
      3          CDS1677
      3          CDS427

      *  = significantly po
      ** = significantly po

Execution times on moon: 1
```

BioLegato
File Edit Documentation DNARNA RNA_Struct Similarity Database Patterns Alignment
CP000481
gattcctaatcggcgacaaaagacatgcagttctccacagacaactacctggtgcggaaggga
Row: 1 Col: 1
Fi Ec Docume DNA RNA_ Simil Datal Patt Align Prir Phylo He
CP000481:CDS1
CP000481:CDS2
CP000481:CDS3
CP000481:CDS4
CP000481:CDS5
CP000481:CDS6
CP000481:CDS12
CP000481:CDS13
CP000481:CDS14
CP000481:CDS15
CP000481:CDS16
CP000481:CDS17
CP000481:CDS18
CP000481:CDS19
File Edit Documentation Protein Similarity Database Alignment Phylogeny Help
CDS137
CDS427
CDS581
CDS1469
CDS2097
CDS1677
CDS183
File Edit Search Preferences Shell Macro Windows Help
ATV
File Edit Search SDI View Options Help
CDS581
CDS183
CDS2097
CDS427
CDS1677
CDS1469
CDS137
☑ Real branch lengths
☑ Seq. name ext. nodes
☑ Seq. name int. nodes
☑ Species ext. nodes
☐ Species int. nodes
☐ EC ext. nodes
☐ EC int. nodes
☐ Branch length values
☐ Bootstrap values
☐ Duplic. vs. spec.
☐ Display orthology
☐ Display sorthology
☐ Color accord. to log L
☐ log L values
☐ Editable
Click on Node to:
● Display/Edit information
○ Collapse/Uncollapse
○ Root/Reroot
○ Subtree/Parent tree
○ Swap children
Zoom in X
Zoom out X
Zoom in Y
Zoom out Y
Show whole
Order subtrees
Uncollapse all
Collapse to deepest
File Edit Documentation Draw
bio142511
((CDS2097:0.72175,(
GYAEFAVVD---ERFAYRIPDTFDDEHAAPLLCAGI---IGFRALR-LAD--CPPGGRLGIYG-F
AQAEYLRVPLADGTLVA-TPAAPPDDLVPSLLAASDV-GTAWFAAD-AAQ--VSPGKTVVIVG-D
YAANQWRIE---PEFAVRLDPTLGD-LG-VLLEPTSVAKAWDQAEKVAARSPAPRHVALVTG-A
GYARHAVVP---RSAVHPLPEGIDLAVAACIGVAGA---TAWRVVE-MAG--VGNEDRVLVLGGA
GYAERVAVR---ADHVLPIPAGVTLVEAAAFPEAAC---TVWSTVF-MAAR-ARAGETLLIHGGAS
AAAELVRAP---AHLVHPLAPQVSE-ESAVLTEPAS---VVFTGLS-RVS--PVPGWRVLVIG-D
TFAEHLVVP---ARNLLPKPDALDDVEAACLPTAWL---TAYRMLL-RSG--ARPGDTILVQGAG

GenomePrairie

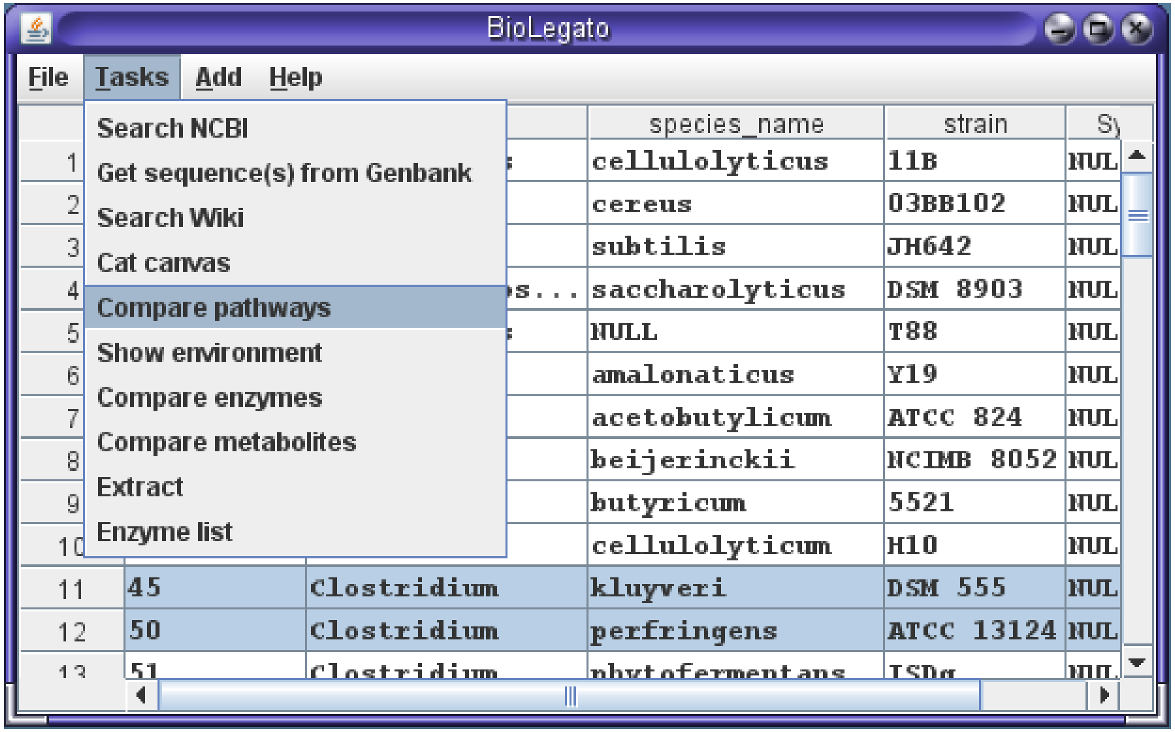
Bioinformatics - BioLegato Database Client

BioLegato lets us quickly build a database client tailored for the project.
Most databases just let you view and query the data.
BioLegato lets you select data, and run programs on it.

Custom Database Client for the MGCB2 project

bldb -

- menus for project-related programs
- table canvas

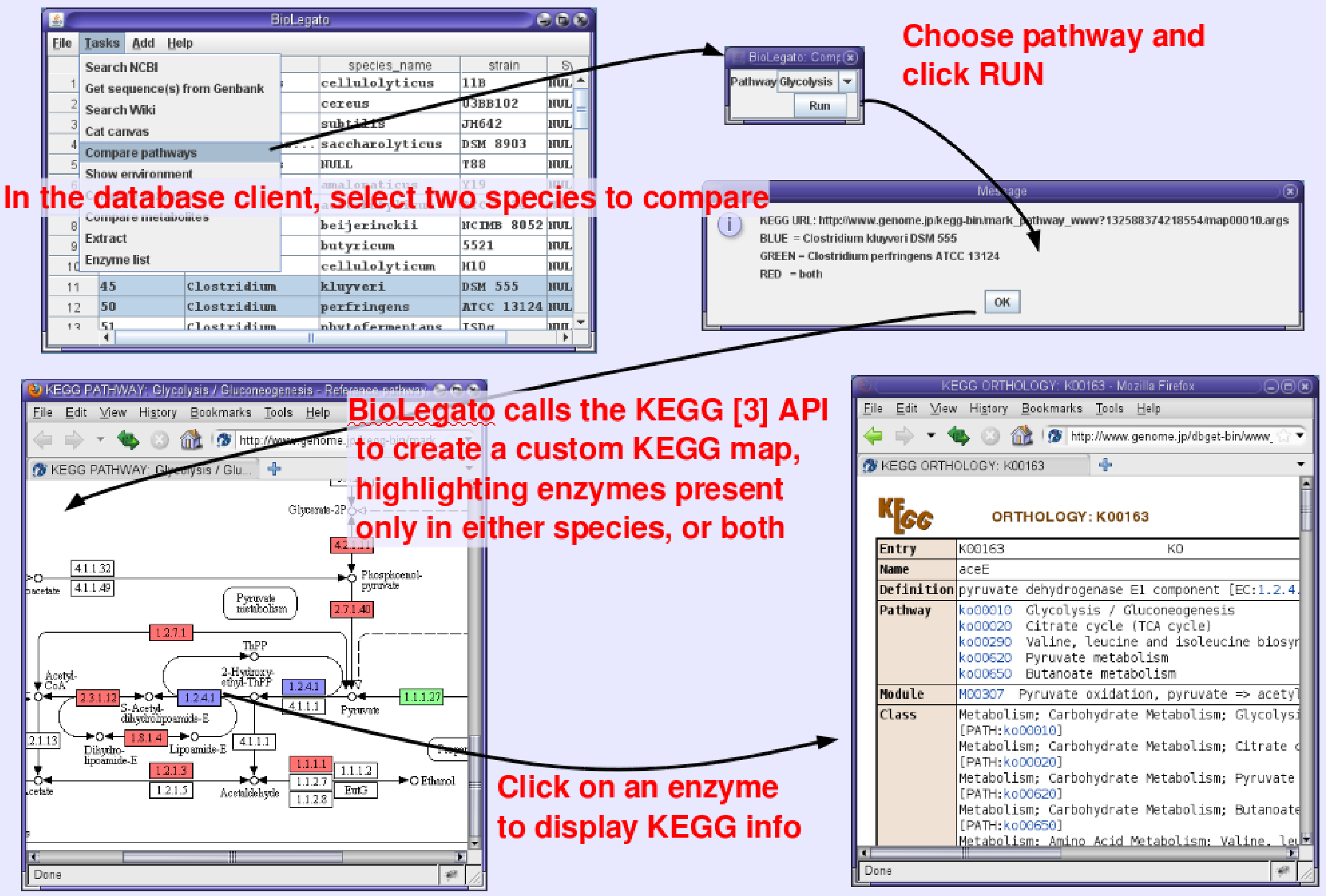


			species_name	strain	Sy
1			cellulolyticus	11B	NUL
2			cereus	03BB102	NUL
3			subtilis	JH642	NUL
4		s...	saccharolyticus	DSM 8903	NUL
5			NULL	T88	NUL
6			amalonaticus	Y19	NUL
7			acetobutylicum	ATCC 824	NUL
8			beijerinckii	NCIMB 8052	NUL
9			butyricum	5521	NUL
10			cellulolyticum	H10	NUL
11	45	Clostridium	kluyveri	DSM 555	NUL
12	50	Clostridium	perfringens	ATCC 13124	NUL
13	51	Clostridium	phytofermentans	ISDg	NUL

GenomeCanada

GenomePrairie

Example: Compare pathways between two species



			species_name	strain	S
1			cellulolyticus	11B	NUL
2			cereus	03BB102	NUL
3			subtilis	JH642	NUL
4			saccharolyticus	DSM 8903	NUL
5			NULL	T88	NUL
6			amalonaticus	Y19	NUL
7			[illegible]	[illegible]	[illegible]
8			beijerinckii	NCIMB 8052	NUL
9			butyricum	5521	NUL
10			cellulolyticum	H10	NUL
11	45	Clostridium	kluyveri	DSM 555	NUL
12	50	Clostridium	perfringens	ATCC 13124	NUL
13	51	Clostridium	phytofermentans	ISDg	NUL

In the database client, select two species to compare

Choose pathway and click RUN

```
KEGG URL: http://www.genome.jp/kegg-bin/mark_pathway_www?132588374218554/map00010.args
BLUE  = Clostridium kluyveri DSM 555
GREEN = Clostridium perfringens ATCC 13124
RED   = both
```

BioLegato calls the KEGG [3] API to create a custom KEGG map, highlighting enzymes present only in either species, or both

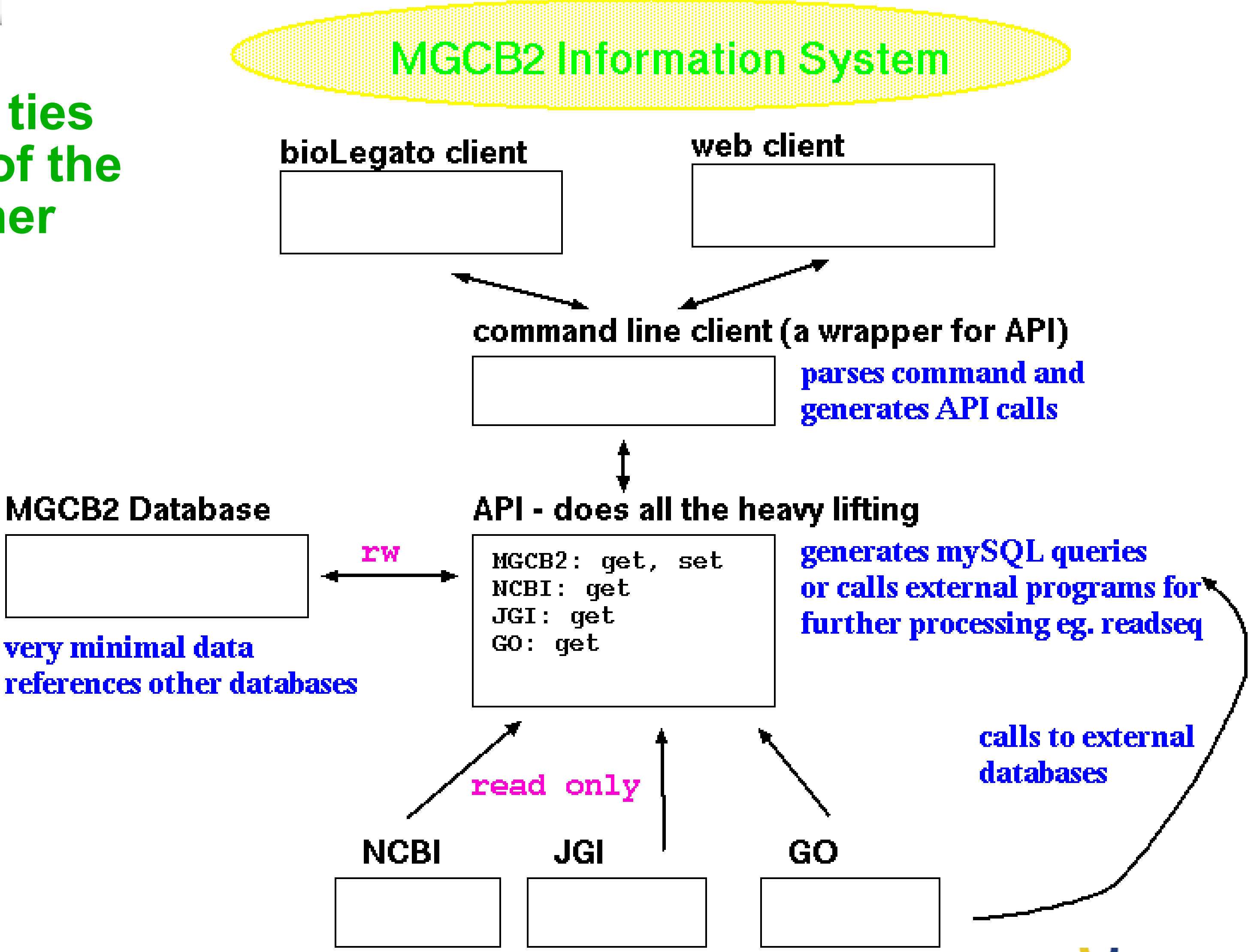
Click on an enzyme to display KEGG info

ORTHOLOGY: K00163

Entry	K00163 KO
Name	aceE
Definition	pyruvate dehydrogenase E1 component [EC:1.2.4.
Pathway	ko00010 Glycolysis / Gluconeogenesis ko00020 Citrate cycle (TCA cycle) ko00290 Valine, leucine and isoleucine biosyn ko00620 Pyruvate metabolism ko00650 Butanoate metabolism
Module	M00307 Pyruvate oxidation, pyruvate => acetyl
Class	Metabolism; Carbohydrate Metabolism; Glycolysi [PATH:ko00010] Metabolism; Carbohydrate Metabolism; Citrate c [PATH:ko00020] Metabolism; Carbohydrate Metabolism; Pyruvate [PATH:ko00620] Metabolism; Carbohydrate Metabolism; Butanoate [PATH:ko00650] Metabolism; Amino Acid Metabolism; Valine, leu

Bioinformatics - Database

The database ties components of the project together



MGCB2 Information System

bioLegato client

web client

command line client (a wrapper for API)

parses command and generates API calls

MGCB2 Database

rw

API - does all the heavy lifting

```
MGCB2: get, set
NCBI: get
JGI: get
GO: get
```

generates mySQL queries or calls external programs for further processing eg. readseq

very minimal data references other databases

read only

calls to external databases

NCBI

JGI

GO

BioLegato - Outline

The Object-Oriented World View

The BioLegato Concept

- End user applications as Objects
- A programmable GUI that calls external programs
- Point and click data pipelining

PCD - Pythonesque Command Definition

- Definition
- Implementation

Pluggable Canvases

PCD – Pythonesque Command Definition Language

- "Pythonesque" scope indentation scheme
- Context-free grammar with indentation control
- Visually relates data via. scope tabulation
- Avoids problems from Python by using fixed spacing rules
- Allows for data clustering and data structures
- Flexible as to what data can be stored
- Very easy to learn
- Human readable
- Any basic data structure can be represented in PCD (e.g. hashes, arrays – NOT JUST MENUS!)

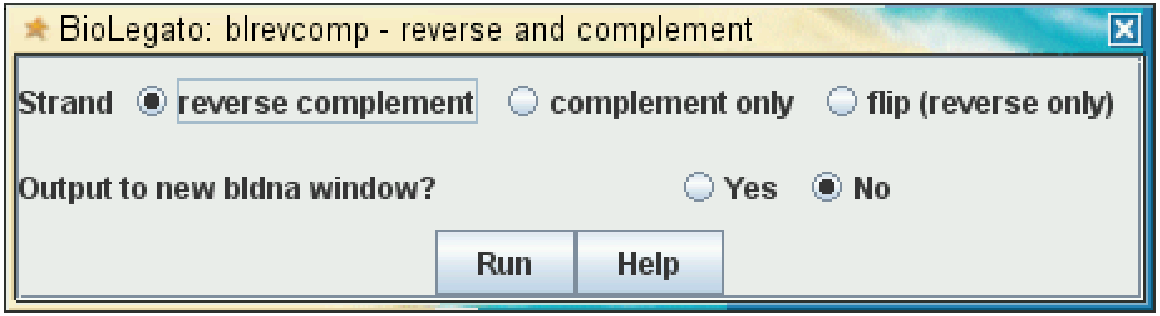
PCD – High level grammar

Non-terminal symbols

```
<sections>    ::= <block> <new_line> <sections>
                  ::= <block>

<block>       ::= <field> <new_line> <indent+1> <block>
                  ::= <value>
                  ::= <field> <value>

<field>       ::= <type_symbol> <statement>
                  ::= <statement>

<value>       ::= <statement>

<statement>   ::= " <quoted_string> "
                  ::= <number>
                  ::= <indentifier/keyword>
.....
<comments> ::= # <space> <comment_line>
```

PCD - GUI scripting language

```
name         "blrevcomp - reverse and complement"
var "in1"
    type         tempfile
    direction    in
    format       flat
var "strand"
    type         chooser
    label        "Strand"
    default      0
    choices
        "reverse complement"  "-r"
        "complement only"     "-c"
        "flip (reverse only)" "-f"
var "out1"
    type         tempfile
    direction    out
    format       flat
var "gdeoutput"
    type         chooser
    label        "Output to new bldna window?"
    default      1
    choices
        "Yes"        "(echo '' > %out1%; bldna %in1%.blrevcomp; $RM_CMD -f %in1%.blrevcomp)&"
        "No"         "mv %in1%.blrevcomp %out1%"
panel
    var "Run"
        type         button
        label        "Run"
        shell        "blrevcomp %STRAND% %in1% %in1%.blrevcomp; $RM_CMD %in1%; %GDEOUTPUT%"
        close        true
    var "Help"
        type         button
        label        "Help"
        shell        "$BIRCH/script/gde_help_viewer.csh null/$BIRCH/doc/BioLegato/blrevcomp.html"
        close        false
```



BioLegato: blrevcomp - reverse and complement

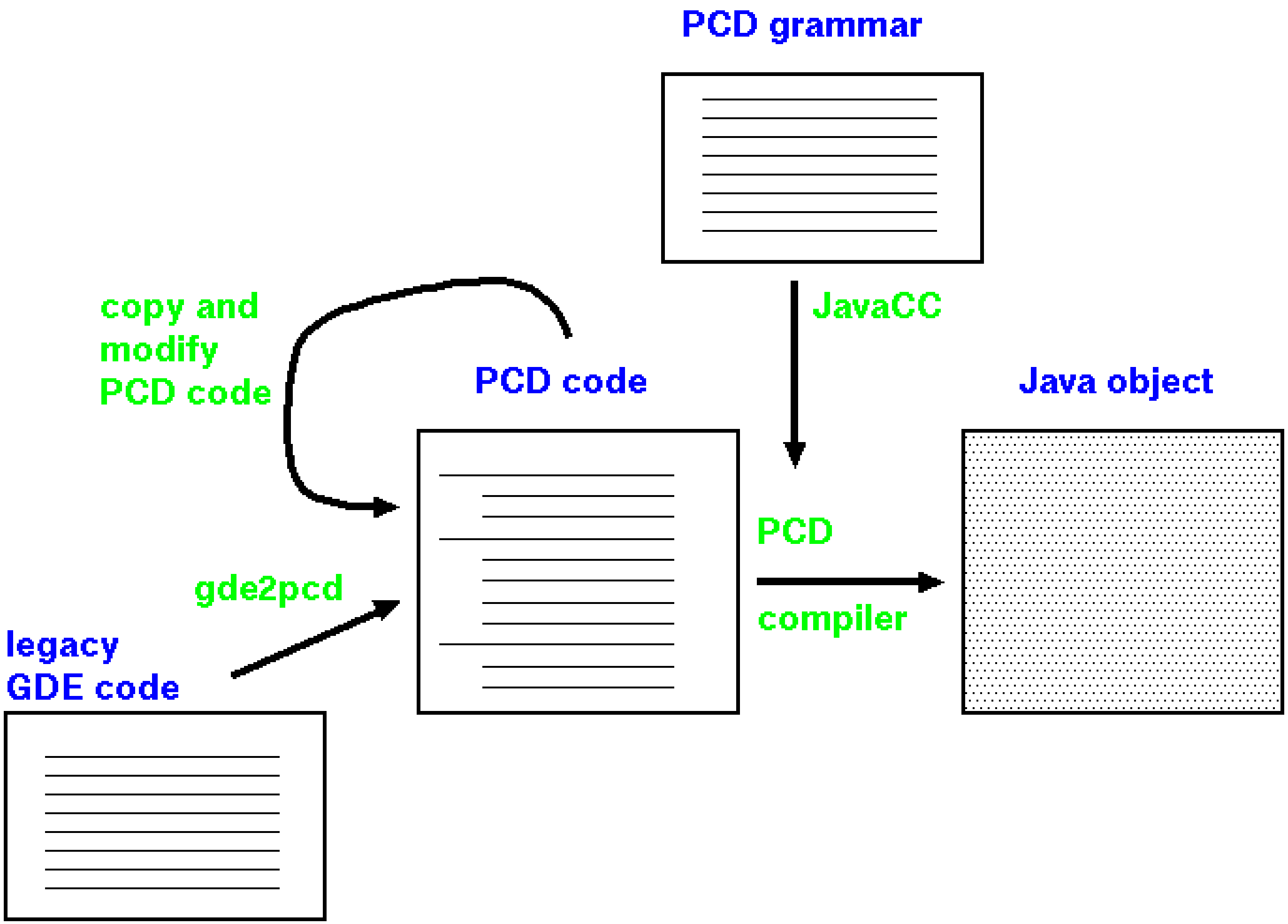
Strand ● reverse complement ○ complement only ○ flip (reverse only)

Output to new bldna window? ○ Yes ● No

PCD – Implementation and application

- Parser written in JavaCC using an LL(1) (Left recursion lookahead top-down) parser
- Allows Java override, where JMenuItem objects can be read in as plugins
- Takes advantage of keywords and scope to try to prevent parsed syntax errors.
- Very easy to modify menu items – requires almost no experience to understand already written menu items
- Can execute shell commands and pipe the output into the parser (dynamic PCD)
- BIRCHShell – cross-platform shell interpreter
- Everything above is written in Java

PCD – Implementation and application



BioLegato - Canvas

- Plugins written in Java (modular)
- The canvas stores and displays data
- Canvases can be put inside anything (because they are represented as JComponents), even other canvases
- Canvases will be able to interact with each other in a future version (currently in alpha stage)
- Simple low overhead plugin system can load plugins anywhere in any code

BioLegato - To Do List

- Inheritance
- Relations between classes
- More canvases
- More widgets eg. menu bars, buttons, side-panels
- Provenance
- Workflows
- Collaborations

• New BioLegato instances eg. geology, astronomy, imaging, agriculture, medicine, chemistry....

BioLegato/BIRCH

BioLegato is distributed with the BIRCH bioinformatics system

http:/home.cc.umanitoba.ca/~frist

- Solaris
- Linux
- Mac OSX
- Windows (2nd Quarter 2012)