

BLAST Application on the GPE/UnicoreGS Grid

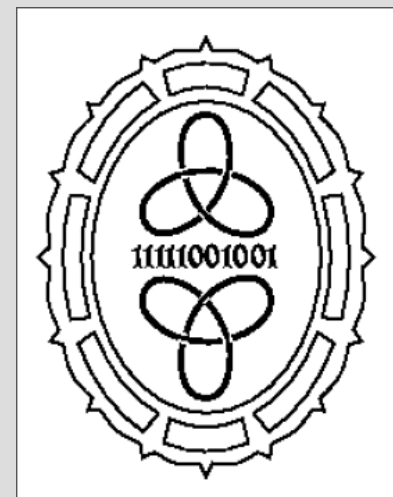
Rafał Kluszczyński

klusi@mat.uni.torun.pl

(joint work with: Marcelina Borcz
and Piotr Bała)

Faculty of Mathematics and Computer Science,
Nicolaus Copernicus University,
Toruń, Poland.

UNICORE Summit
30-31 August 2006

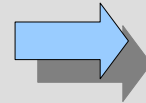


Outline

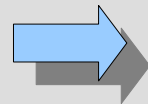
- Introduction
- BLAST Software
- Grid Programming Environment
- GridBean Approach
- BLAST GridBean
- Conclusions
- Future Work

Introduction

- Weather forecast
- Molecular simulations
- 3D rendering
- etc.



Demand for
Computing Power



The Concept of
Grid Computing

Grid Computing

- Checklist definition of the Grid technology:
 - No centralized control,
 - Using open and standard protocols,
 - Combined distributed resources are much more worthwhile.

(Suggested by Ian Foster)

BLAST

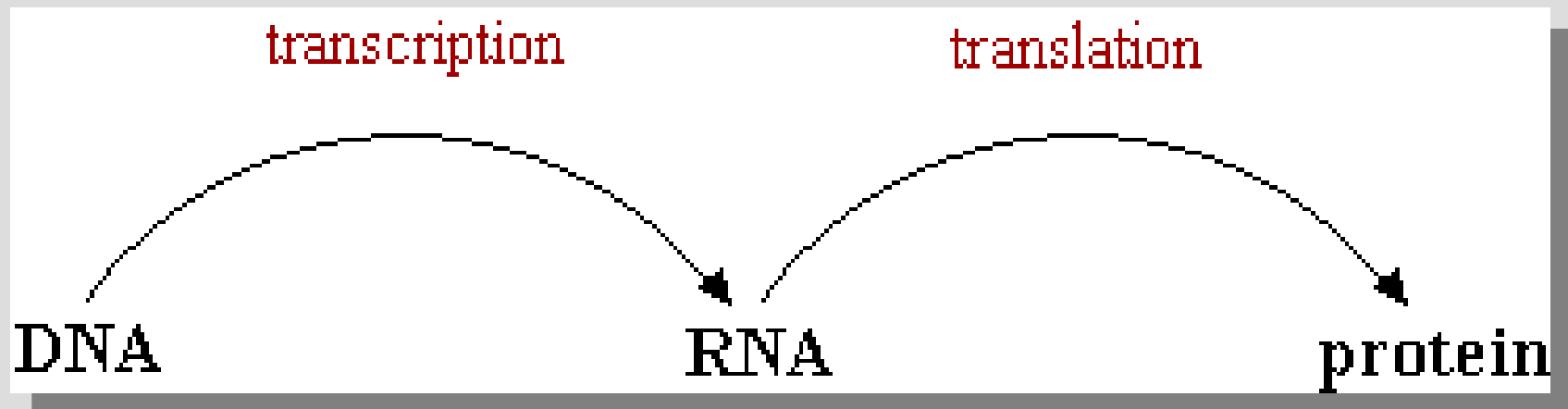
- BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) provides a method for quick searching of nucleotide and protein databases.

Altschul, S., Gish, W., Miller, W., Myers, E.W., Lipman, D.: "A Basic Local Alignment Search Tool". Journal of Molecular Biology 215:403-410 (1990)

- BLAST is the most popular and the most accepted sequence analysis tool nowadays.

BLAST motivation

- The Central Dogma of Biology:



BLAST

- BLAST algorithm finds statistically significant **local** similarities between pairs: user-defined (protein or DNA) sequence and sequences from databases.
- **Local** means that we align sequence segments, rather than align the entire sequence.

Why BLAST ?

- Sequence alignments provide a powerful way to compare novel sequences with previously characterized genes.
- Both functional and evolutionary information can be predicted from well designed queries and alignments.

BLAST tasks

- Identifying protein family,
- Prediction of biological function and structure of new sequences,
- Exploring of evolutionary relationship between organisms.

BLAST features

- Reliability,
- Speed,
- Flexibility.

BLAST uses heuristic method to find optimal alignments. In such a way the program is 50-100 times faster than using only dynamic programming.

BLAST package

- BLAST package includes many specialized programs.
- **blastn**: nucleotide query vs. nucleotide database,

```
Score = 38.2 bits (19), Expect = 1.8
Identities = 19/19 (100%), Gaps = 0/19 (0%)
Strand=Plus/Minus

Query 19      GACCAATGACCCAGTAGGG 37
              ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 56720   GACCAATGACCCAGTAGGG 56702
```

BLAST package

- **blastp**: protein query vs. protein database,

```
Score = 75.1 bits (183), Expect = 9e-13, Method: Composition-based stats.  
Identities = 39/40 (97%), Positives = 40/40 (100%), Gaps = 0/40 (0%)
```

```
Query 1  RGYISTNRSKHNLKAHLILVCKYRKKLLQGDLNNFIKSVI 40  
         RGYISTNRSKHNLKAHLILVCKYRKKLLQGD LN+FIKSVI  
Sbjct 15 RGYISTNRSKHNLKAHLILVCKYRKKLLQGD LNDFIKSVI 54
```

- **blastx**: translated query vs. protein database,
- **tblastn**: protein query vs. translated database,
- **tblastx**: translated query vs translated database.

BLAST software

- BLAST is being developed by many institutes.
- The most popular versions are:
 - NCBI-BLAST
(at: *National Center for Biotechnology Information*)
 - WU-BLAST
(at: *Washington University in Saint Louis*)

NCBI BLAST

The image shows a screenshot of the NCBI BLAST website. At the top, there is a browser window with the address bar showing "http://www.ncbi.nlm.nih.gov/BLAST/". The website header includes the NCBI logo and the text "BLAST" on the left, and "Latest news: 7 May 2006 : BLAST 2.2.14 released" on the right. The main content area is divided into several sections. On the left, there is a navigation menu with categories: "About", "More info", "Software", and "Other resources". The main content area is a grid of boxes. The top box is a general description of BLAST. Below it are four boxes: "Nucleotide", "Protein", "Translated", and "Genomes". Each box contains a list of search options. At the bottom, there are two boxes labeled "Special" and "Meta".

NCBI →
BLAST

Latest news: 7 May 2006 : BLAST 2.2.14 released

About

- Getting started
- News
- FAQs

More info

- NAR 2004
- NCBI Handbook
- The Statistics of Sequence Similarity Scores

Software

- Downloads
- Developer info

Other resources

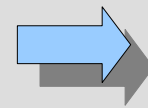
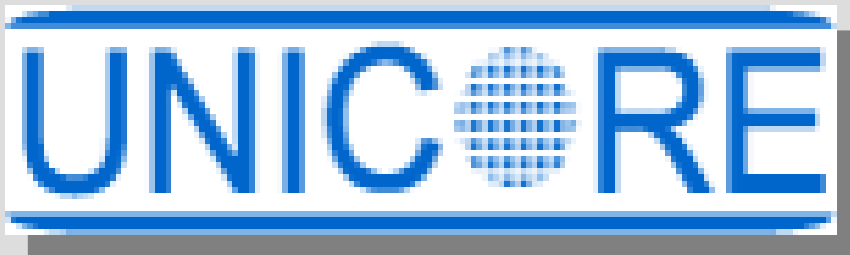
- References
- NCBI Contributors
- Mailing list
- Contact us

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Nucleotide <ul style="list-style-type: none">• Quickly search for highly similar sequences (megablast)• Quickly search for divergent sequences (discontiguous megablast)• Nucleotide-nucleotide BLAST (blastn)• Search for short, nearly exact matches• Search trace archives with megablast or discontiguous megablast	Protein <ul style="list-style-type: none">• Protein-protein BLAST (blastp)• Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)• Search for short, nearly exact matches• Search the conserved domain database (rpsblast)• Protein homology by domain architecture (cdart)
Translated <ul style="list-style-type: none">• Translated query vs. protein database (blastx)• Protein query vs. translated database (tblastn)• Translated query vs. translated database (tblastx)	Genomes <ul style="list-style-type: none">• Human, mouse, rat, chimp, cow, pig, dog, sheep, cat• Chicken, puffer fish, zebrafish• Fly, honey bee, other insects• Microbes, environmental samples• Plants, nematodes• Fungi, protozoa, other eukaryotes
Special	Meta

Grid Environments

- Globus Toolkit
- UNICORE



Grid
Programming
Environment
(**GPE**)

GPE

- Based on the experience from UNICORE implementation,
- Establishing a stable interface between different Grid middlewares,
- Designing more flexible and user-friendly client framework.

Different Usage Scenarios

- The concept of GPE introduce different client applications operating on the Grid for different types of users:
 - Expert Users,
 - Application Users,
 - Unaware Users.

Expert Client

- For users with some knowledge about the Grid.
- It is a successor of UNICORE Client:
 - Workflow Editor,
 - Allow loading multiple plug-ins,
 - Managing multiple identities.

Application Client

- For users which need only one application at the time.
- It has limitation of loading **ONLY** one plug-in at the time (no workflows).
- By limiting the functionality it is possible to run on the mobile devices.

Unaware Users

- For users which likes Internet browsers.
- There is no need to install a Client Application.
- Thanks to *portal* solution, one can access the Grid from any Internet cafe.
- Good for mobile users.

Plug-ins for Clients

- During UNICORE development it occurs that plug-ins are a very good solution.
- Among others it also occurs that they:
 - should not depend on the Client application,
 - should be easy do develop.

GridBean Approach

- Easy distribution and update,
- Overview of supported applications,
- Flexibility of using GridBeans,
- Easy implementation.

GridBean Service

The screenshot shows the GPE Client - BLAST interface. The main window has a menu bar with 'File' and 'Tools'. Below the menu bar, there are tabs for 'Target Systems', 'BLAST', 'Outcome: DNAsShortSequenceJob', and 'Files'. The 'BLAST' tab is active, displaying a table of jobs. A 'Select GridBean' dialog box is open in the foreground, listing various applications with 'BLAST' selected. The status bar at the bottom shows 'No certificate', a message about getting a list of GridBeans, and 'Running Threads: 0'.

Job Name	Application	State	Termination Time
SayHiJob1	HI 1.0	SUCCESSFUL	27.07.06 18:37
WhatDateJob	DATE 1.0	SUCCESSFUL	27.07.06 18:40
DNAsShortSequenceJob	BLAST 2.2.13	SUCCESSFUL	28.07.06 18:45
EcoliBigSequenceJob1	BLAST 2.2.13	RUNNING	29.07.06 18:47

Select GridBean ...

- Application
- Breakpoint
- Destroy
- Filecheck
- POVRay
- Zip
- BLAST**

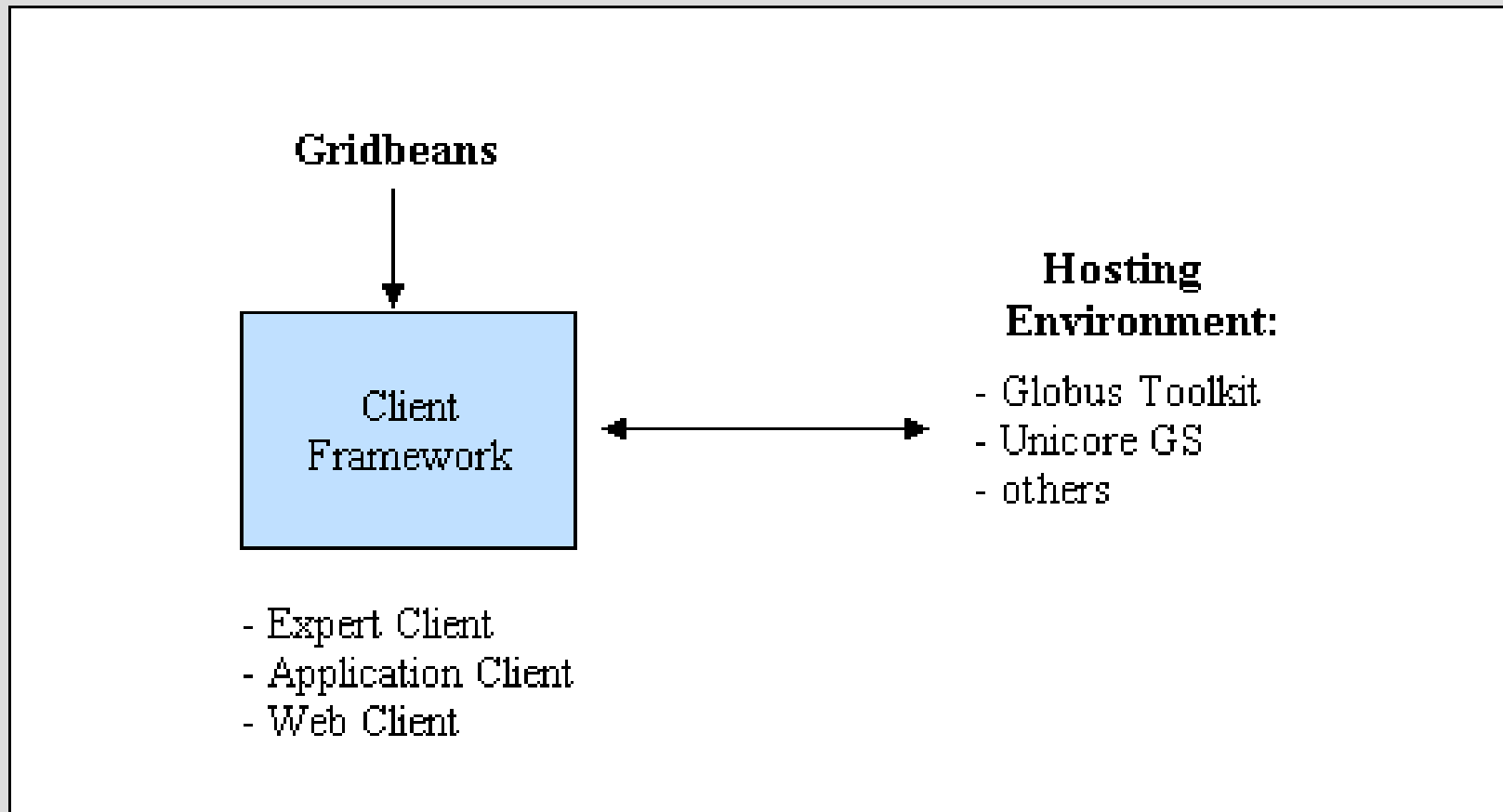
OK Cancel

No certificate Getting list of GridBeans from http://localhost:8080/wsrf/services/GridBeanService finished Running Threads: 0

Interoperability

- The GPE Clients can contact to TSSs available on different hosting environments like:
 - Globus,
 - UnicoreGS,
 - others.
- Once implemented GridBean can be used with different Grid middlewares.

Interoperability



GPE Admin Client

The screenshot shows the GPE Admin Client interface. The left pane displays a tree view of the GRID structure, including 'localhost', 'Sample TSS', 'NCU Sample TSS', and two remote hosts at 192.168.1.116 and 192.168.1.117. The right pane shows a table of applications with columns for Name and Version. The 'BLAST' application is highlighted.

Name	Version
HELLO	1.0
UPLOAD	1.0
HI	1.0
WHOAMI	1.0
BLAST	2.2.13
DATE	1.0

At the bottom of the window, there is a status bar with the following text: "No certificate" (in red), "Getting applications from NCU Sample TSS finished", and "Running Threads: 0".

GridBean Implementation

- In order to create a GridBean the developer have to :
 - Design the graphical layout,
 - Implement job description method,
 - Implement job reconstructing method.

Job Reconstructing

The screenshot shows the GPE Client - BLAST interface. The main window displays a table of jobs with columns for Job Name, Application, State, and Termination Time. A context menu is open over the 'DNAShortSequenceJob' row, with 'Reconstruct Input' selected. The interface also includes panels for Registries, Target Systems, and Storages, and a status bar at the bottom.

Job Name	Application	State	Termination Time
SayHiJob1	HI 1.0	SUCCESSFUL	27.07.06 18:37
WhatDateJob	DATE 1.0	SUCCESSFUL	27.07.06 18:40
DNAShortSequenceJob	BLAST 2.2.13	SUCCESSFUL	28.07.06 18:45
EcoliBigSequenceJob1		RUNNING	29.07.06 18:47

Context Menu Options:

- Start
- Abort
- Hold
- Resume
- Destroy
- Refresh
- Fetch Outcome
- Reconstruct Input**
- Properties
- Change Termination Time...

Status Bar: No certificate Getting list of GridBeans from http://localhost:8080/wsrf/services/GridBeanService finished Running Threads: 0

GridBean GUI

- Designing GridBean graphical layout means organizing visual area for parameter controls.

```
package pl.torun.uni.mat.gpe.gridbeans.hi.plugin;

public class HiInputPanel extends GridBeanPanel {
    public HiInputPanel(com.intel.gpe.client2.Client parent) {
    }

    public void updateValues(com.intel.gpe.client2.Client client) {
    }
}
```

Job Description

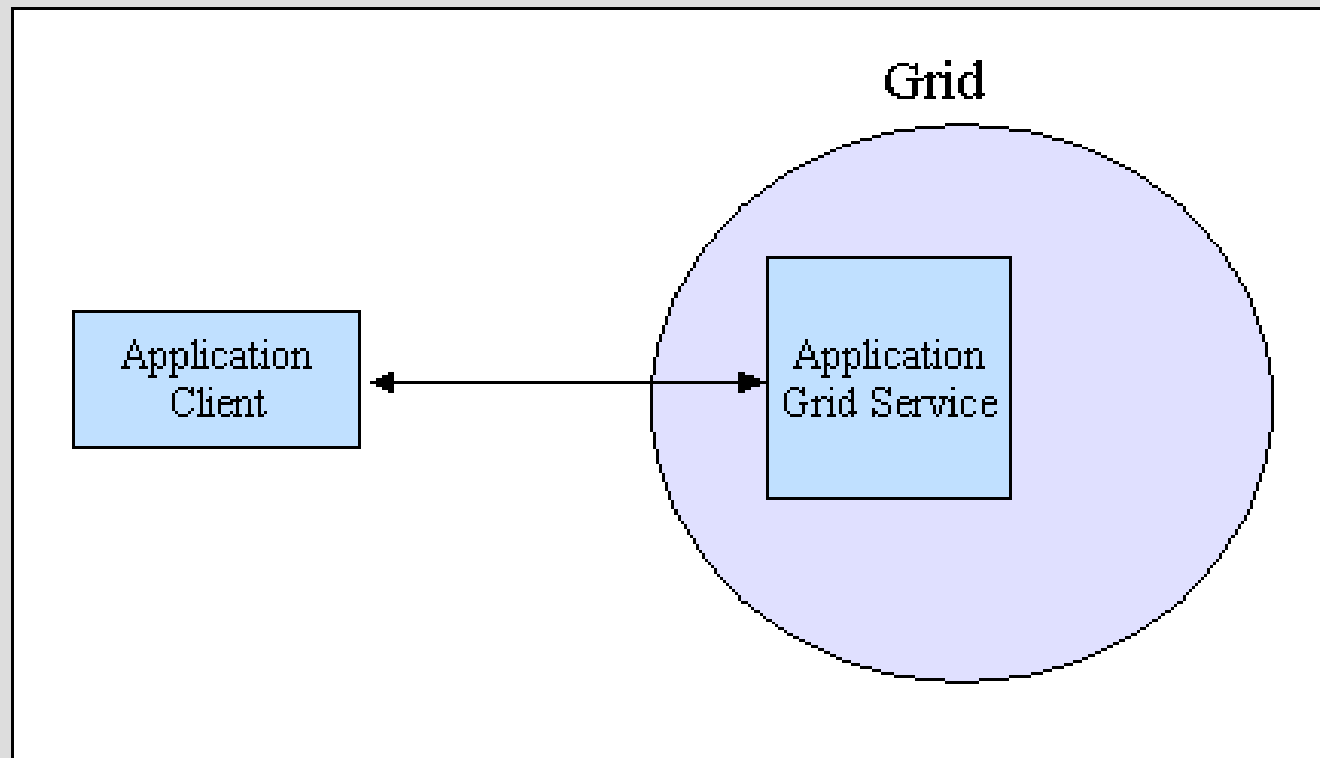
```
public class HiGridBean extends DefaultGridBeanModel implements IGridBean {  
  
    public HiGridBean() {}  
  
    public List<GridBeanParameter> getInputParameters() {}  
  
    public List<GridBeanParameter> getOutputParameters() {}  
  
    public void setupJobDefinition(Job job) throws GridBeanException {}  
    public void parseJobDefinition(Job job) throws GridBeanException {}  
  
    public String getName() {}  
  
}
```

Application on the Grid

- Another thing is to bringing application on the Grid, which means we have to design script template to run.

```
<?xml version="1.0" ?>
- <idb:Template name="HI" isPublic="true"
  xmlns:idb="http://gpe.intel.com/idb">
- <idb:Invocation name="">
  - <idb:Body>
    <![CDATA[ echo Hi at NCU, <USER_NAME> <Exclamations> ]]>
    </idb:Body>
  </idb:Invocation>
- <idb:Field name="Exclamations">
  <idb:Tag name="true">!!!</idb:Tag>
  <idb:Tag name="false">!</idb:Tag>
  </idb:Field>
</idb:Template>
```

Application Services



BLAST GridBean

- BLAST became one of the fundamental tool in sequence analysis.
- BLAST uses heuristic algorithm, but its computation time increases with the size of database.
- BLAST needs a lot of computing power.

BLAST GridBean

The screenshot shows the 'GPE Client - BLAST' application window. The title bar includes standard window controls. The menu bar contains 'File' and 'Tools'. Below the menu bar, there are tabs for 'Target Systems' (selected), 'BLAST', 'Outcome: DNAshortSequenceJob', and 'Files'. Under the 'BLAST' tab, there are sub-tabs for 'BLAST' and 'OPTIONS'. The main area contains several input fields and a text area:

- Job Name:** A text box containing 'NoName'.
- Type of program:** A dropdown menu with 'Nucleotid - nucleotid BLAST (blastn)' selected.
- Sequence:** A large text area containing a DNA sequence starting with '>Test' and followed by 10 lines of nucleotide characters.
- Setting subsequence:** A checkbox that is currently unchecked, with 'From:' and 'To:' text boxes below it.
- Choose Database:** A dropdown menu with 'Echerichia Coli (nucleotides)' selected.

At the bottom of the window, there is a status bar with three sections: 'No certificate' (in red), 'Fetch outcome for DNAshortSequenceJob-BLAST v.2.2.13 finished', and 'Running Threads: 0'.

BLAST GridBean

The screenshot shows the 'GPE Client - BLAST' window. The title bar includes standard window controls. The menu bar has 'File' and 'Tools'. Below the menu bar, there are tabs for 'Target Systems' (selected), 'BLAST', 'Outcome: DNAShortSequenceJob', and 'Files'. Under the 'BLAST' tab, there are sub-tabs for 'BLAST' and 'OPTIONS'. The main area contains several configuration fields: 'Matrix' (BLOSUM62), 'Gap Costs' (Existence:11 Extension:1), 'Expect' (10.0), 'WordSize' (11), a 'Choose filter' section with checkboxes for 'Low complexity' (checked), 'Human repeats', 'Mask for lookuptable only', and 'Mask lower case', 'Number of processors' (1), 'Descriptions' (100), 'Alignments' (50), 'Alignment view' (Flat query-anchored, no identities and blunt ends), and 'Other advanced'. At the bottom, a status bar shows 'No certificate' in red, 'Fetch outcome for DNAShortSequenceJob-BLAST v.2.2.13 finished', and 'Running Threads: 0'.

GPE Client - BLAST

File Tools

Target Systems BLAST Outcome: DNAShortSequenceJob Files

BLAST OPTIONS

Matrix BLOSUM62 Gap Costs Existence:11 Extension:1

Expect 10.0

WordSize 11

Choose filter

Low complexity Human repeats Mask for lookuptable only Mask lower case

Number of processors 1

Descriptions 100

Alignments 50

Alignment view Flat query-anchored, no identities and blunt ends

Other advanced

No certificate Fetch outcome for DNAShortSequenceJob-BLAST v.2.2.13 finished Running Threads: 0

NCBI-BLAST Website

The screenshot shows the NCBI-BLAST website interface. At the top, there is a browser address bar with the URL `http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAY`. Below the address bar, there is a large empty text input field for the search query, with a [Search](#) link to its left. Below the search field, there are two input fields for "From:" and "To:" with a [Set subsequence](#) link to the left. Below these, there is a dropdown menu for "Choose database" with "nr" selected. At the bottom of this section, there are three buttons: "BLAST!", "Reset query", and "Reset all", with the word "Now:" to the left. A horizontal orange line separates this section from the "Options" section below. The "Options" section is titled "Options for advanced blasting". It contains a [Limit by entrez query](#) link, an input field, and a dropdown menu labeled "or select from:" with "All organisms" selected. Below this, there are four checkboxes: "Choose filter" (checked), "Low complexity" (unchecked), "Human repeats" (unchecked), "Mask for lookup table only" (unchecked), and "Mask lower case" (unchecked). Below the checkboxes, there is an "Expect" label and an input field containing "10". At the bottom, there is a "Word Size" label and a dropdown menu containing "11".

GridBean Panels

- During implementation of GridBean there should be at least one *Input Panel* responsible for application parameters.
- We can also add some *Output Panels* which can present our results in much more attractive (graphical) form.

BLAST Results

```
GPE Client - BLAST
File Tools
Target Systems | BLAST | Outcome: EcoliBigSequenceJob1 | Files
Stdout | Stderr | Log
BLASTN 2.2.13 [Nov-27-2005]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.

Query= Test
      (560 letters)

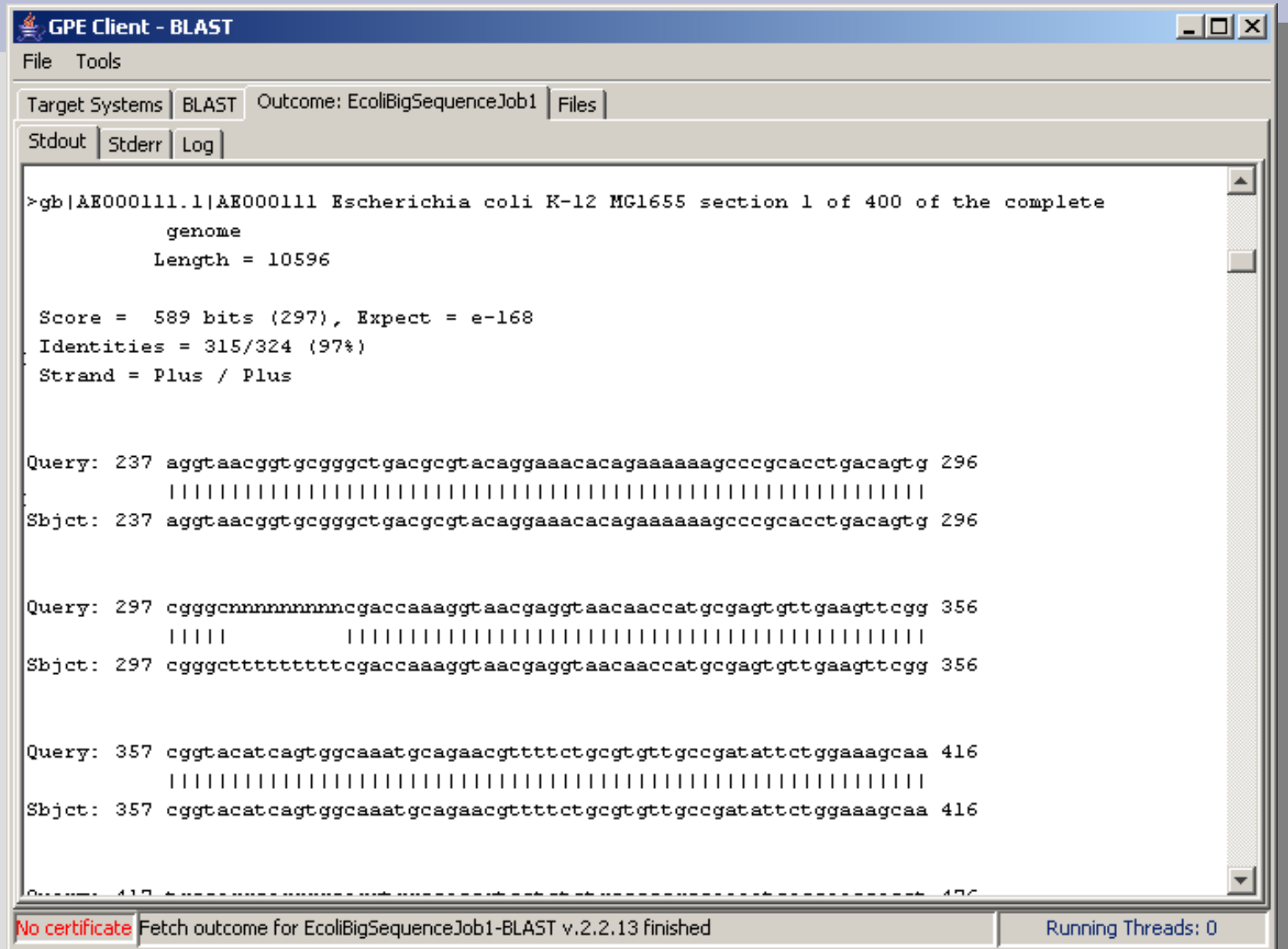
Database: ecoli.nt
      400 sequences; 4,662,239 total letters

Sequences producing significant alignments:

                                Score   E
                                (bits) Value
gb|AE000111.1|AE000111 Escherichia coli K-12 MG1655 section 1 of...   589   e-168
gb|AE000440.1|AE000440 Escherichia coli K-12 MG1655 section 330 ...    32   0.51
gb|AE000132.1|AE000132 Escherichia coli K-12 MG1655 section 22 o...    32   0.51
gb|AE000387.1|AE000387 Escherichia coli K-12 MG1655 section 277 ...    30   2.0
gb|AE000295.1|AE000295 Escherichia coli K-12 MG1655 section 185 ...    30   2.0
gb|AE000279.1|AE000279 Escherichia coli K-12 MG1655 section 169 ...    30   2.0

No certificate Fetch outcome for EcoliBigSequenceJob1-BLAST v.2.2.13 finished Running Threads: 0
```

BLAST Alignments



```
GPE Client - BLAST
File Tools
Target Systems | BLAST | Outcome: EcoliBigSequenceJob1 | Files
Stdout | Stderr | Log
>gb|AE000111.1|AE000111 Escherichia coli K-12 MG1655 section 1 of 400 of the complete
genome
Length = 10596

Score = 589 bits (297), Expect = e-168
Identities = 315/324 (97%)
Strand = Plus / Plus

Query: 237 aggtaacggtgctgggctgacgcgtacaggaaacacagaaaaagccccgcacctgacagtg 296
      |||
Sbjct: 237 aggtaacggtgctgggctgacgcgtacaggaaacacagaaaaagccccgcacctgacagtg 296

Query: 297 cgggcnmnmnmnncgaccaaaggtaacgaggtaaccaacctgagagtgtgaagttcgg 356
      |||||
Sbjct: 297 cgggcttttttttcgaccaaaggtaacgaggtaaccaacctgagagtgtgaagttcgg 356

Query: 357 cggtacatcagtgcaaatgcagaacgttttctgcgtgttgccgatattctgaaagcaa 416
      |||
Sbjct: 357 cggtacatcagtgcaaatgcagaacgttttctgcgtgttgccgatattctgaaagcaa 416

Query: 417
Sbjct: 417

No certificate Fetch outcome for EcoliBigSequenceJob1-BLAST v.2.2.13 finished Running Threads: 0
```

BLAST GridBean Tests

- BLAST GridBean was developed under GPE4GTK project.
- Successfully tested on binary distribution GPE-Lite (release 1.0.0).
- For testing have been used NCBI-BLAST package (ver. 2.2.13).

Conclusions

- Grid Computing is becoming more and more popular every day.
- GridBean approach is very promising, it simplifies using applications with different Grid middlewares.
- Computing centers offer more computing power.

Conclusions

- Major effort in bringing application on the Grid is to design and implement a GridBean.
- At the GPE4GTK project there are ready plug-ins for:
 - Povray,
 - PDBsearch,
 - PDB2POV, etc.

Conclusions

- What may need some effort is to implement some output panels with specific (usually some graphical view) presentation of the application results.
- Thanks to Grid Services configuration of some program parameters can be made on the Grid.

Future Work

- Add graphical presentation of BLAST results using BioJava package.
- Design GridBeans for other bioinformatic tools.
- Finish NAMDGridBean implementation.

NAMD GridBean

The screenshot shows the 'GPE Client - NAMD' window with the following configuration details:

- Target Systems: NAMD | Job Outcome | Files
- NAMD Config | PSFGEN Input | Input and Output Files | Basic Simulation | Additional Simulation
- JobName : NoNameJob_1156684260375
- Use psfgen program first
- numsteps : 100
- coordinates filename : D:\Inzynierka\namd-tutorial-files\common\ubq_ws.pdb [Browse ...]
- structure filename : D:\Inzynierka\namd-tutorial-files\common\ubq_ws.psf [Browse ...]
- parameters filename : [Browse ...]
- exclude : scaled1-4
- outputname filename : []
- temperature : 311
- velocities filename : [Browse ...]
- binvelocities filename : [Browse ...]

At the bottom, a status bar shows: **No certificate** Getting list of GridBeans from <http://192.168.1.116:8080/wsrf/services/GridBeanService> finished Running Threads: 0

NAMD GridBean

The screenshot shows the 'GPE Client - NAMD' window with the following configuration details:

- Target Systems:** NAMD (selected), Job Outcome, Files
- NAMD Config:** PSFGEN Input (selected), Input and Output Files, Basic Simulation, Additional Simulation
- Input Files:**
 - coordinates : ubq_ws.pdb
 - structure : ubq_ws.psf
 - parameters : (empty)
 - paraTypeXplor paraTypeCharmm
 - velocities : (empty)
 - binvelocities : (empty)
 - bincoordinates : (empty)
- Output Files:**
 - outputname : ubquitin_output
 - binaryoutput
 - restartname : (empty)
 - restartfreq : 0

At the bottom, a status bar displays: **No certificate** Getting list of GridBeans from <http://192.168.1.116:8080/wsrp/services/GridBeanService> finished Running Threads: 0

Thank You !

Any questions, comments, advices *etc.* ?