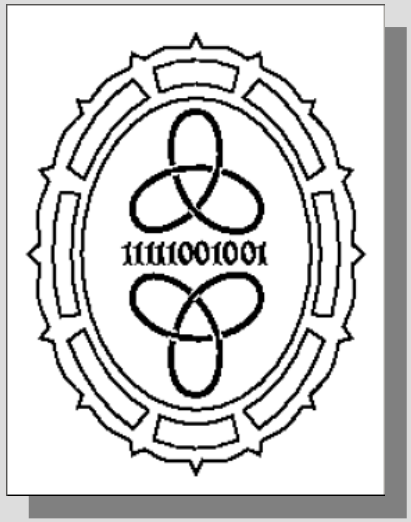
BLAST Application on the GPE/UnicoreGS Grid

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.



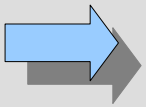
UNICORE Summit
30-31 August 2006

Outline

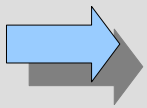
- Introduction
- BLAST Software
- Grid Programming Environment
- GridBean Approach
- BLAST GridBean
- Conclusions
- Future Work

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

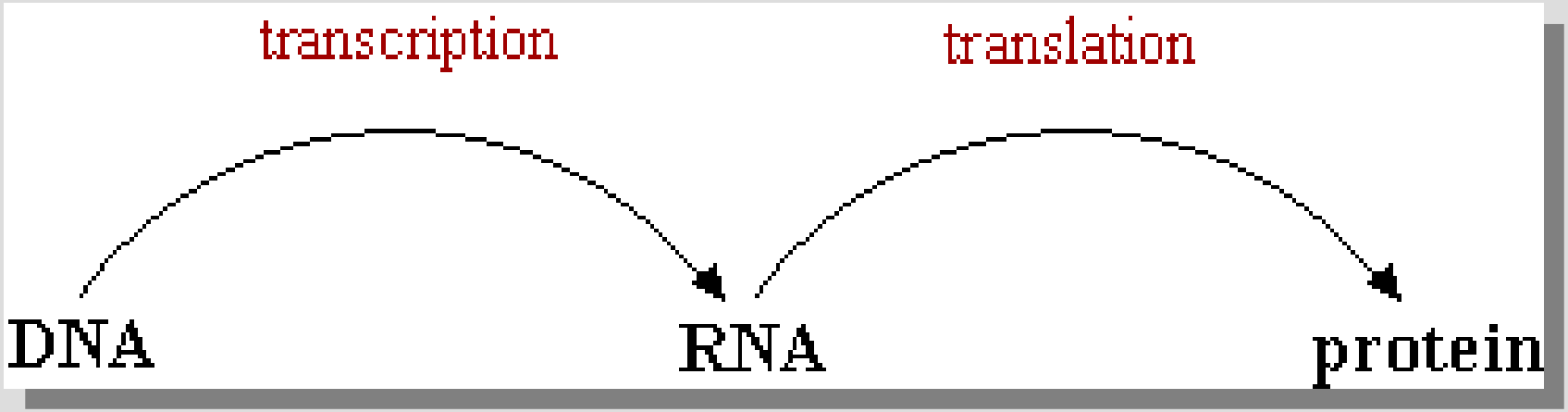
- BLAST (***B**asic **L**ocal **A**lignment **S**earch **T**ool)* provides a method for quick searching of nucleotide and protein databases.

  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 then 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,

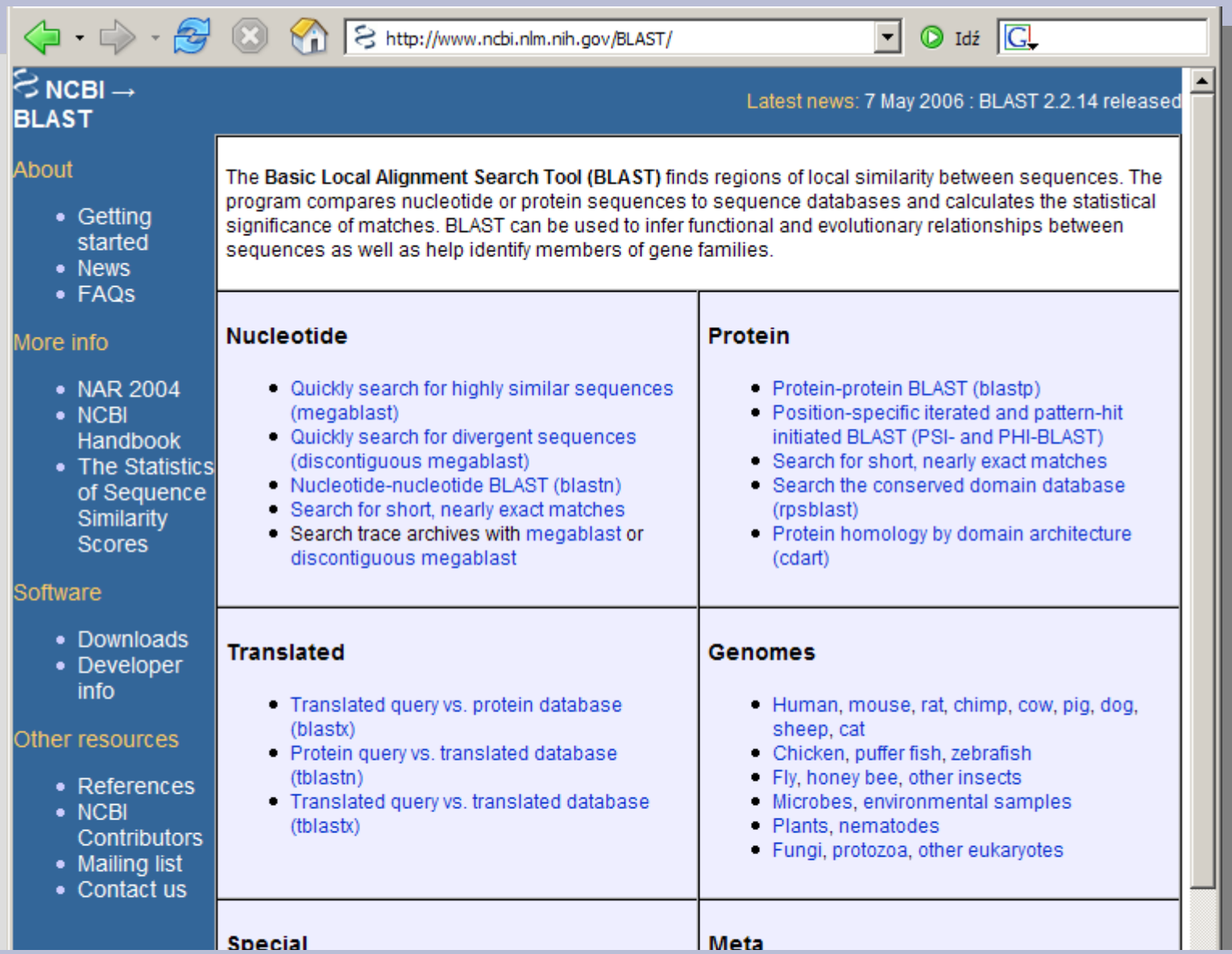
```
Score = 75.1 bits (183),  Expect = 9e-13, Method: Composition-based stats.
Identities = 39/40 (97%), Positives = 40/40 (100%), Gaps = 0/40 (0%)

Query  1  RGYISTNRSKHNLKAHLILVCKYRKKLLQGDLNNFIKSVI  40
          RGYISTNRSKHNLKAHLILVCKYRKKLLQGDLN+FIKSVI
Sbjct  15 RGYISTNRSKHNLKAHLILVCKYRKKLLQGDLNDFIKSVI  54
```

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

BLAST software

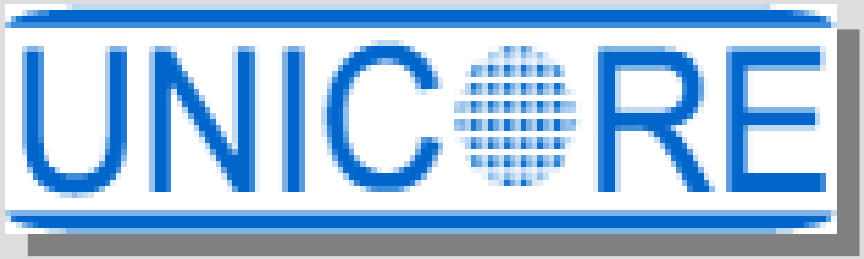
- BLAST is being developed by many institutes.
- The most popular versions are:
  - NCBI-BLAST
    (at: *National Center for Biotechnology Information)*
  - WU-BLAST
    (at: *Washington University in Saint Louis)*

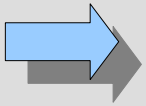
NCBI BLAST


http://www.ncbi.nlm.nih.gov/BLAST/
Idź
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
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
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
Translated query vs. protein database (blastx)
Protein query vs. translated database (tblastn)
Translated query vs. translated database (tblastx)
Genomes
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 indentities.

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

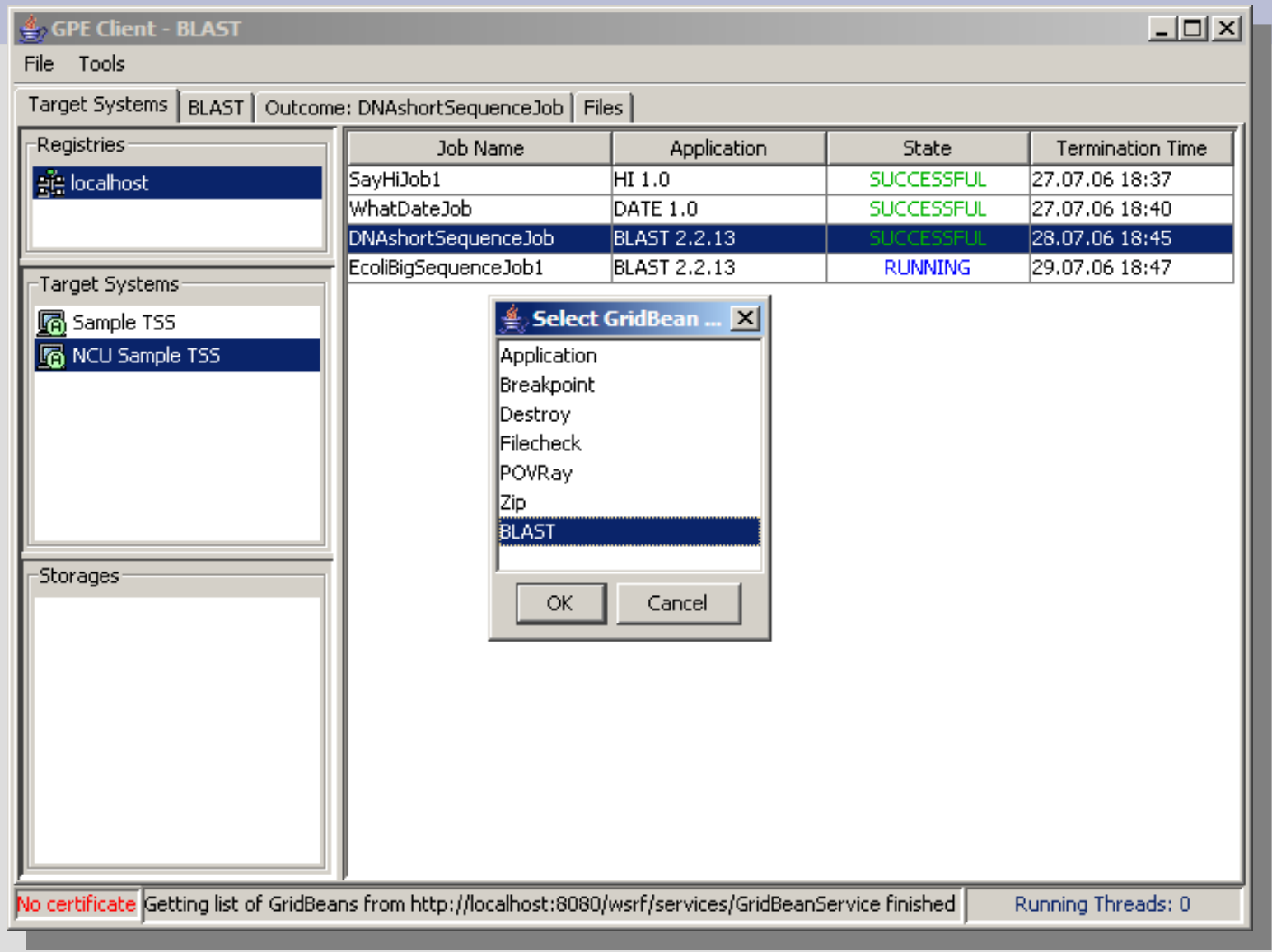
- During UNICORE development it occurs that plug-ins are a very good solution.

- 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

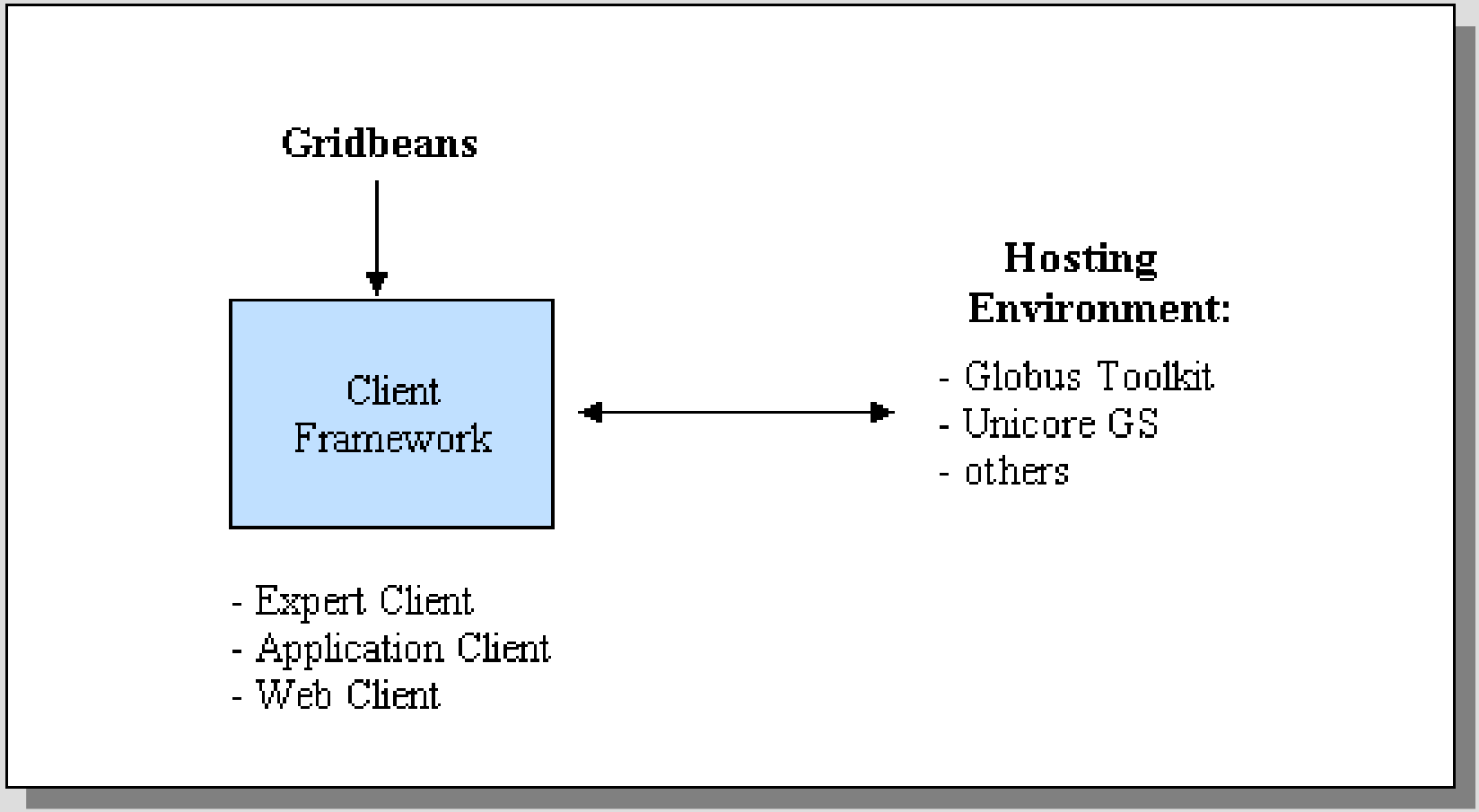


Interoperability

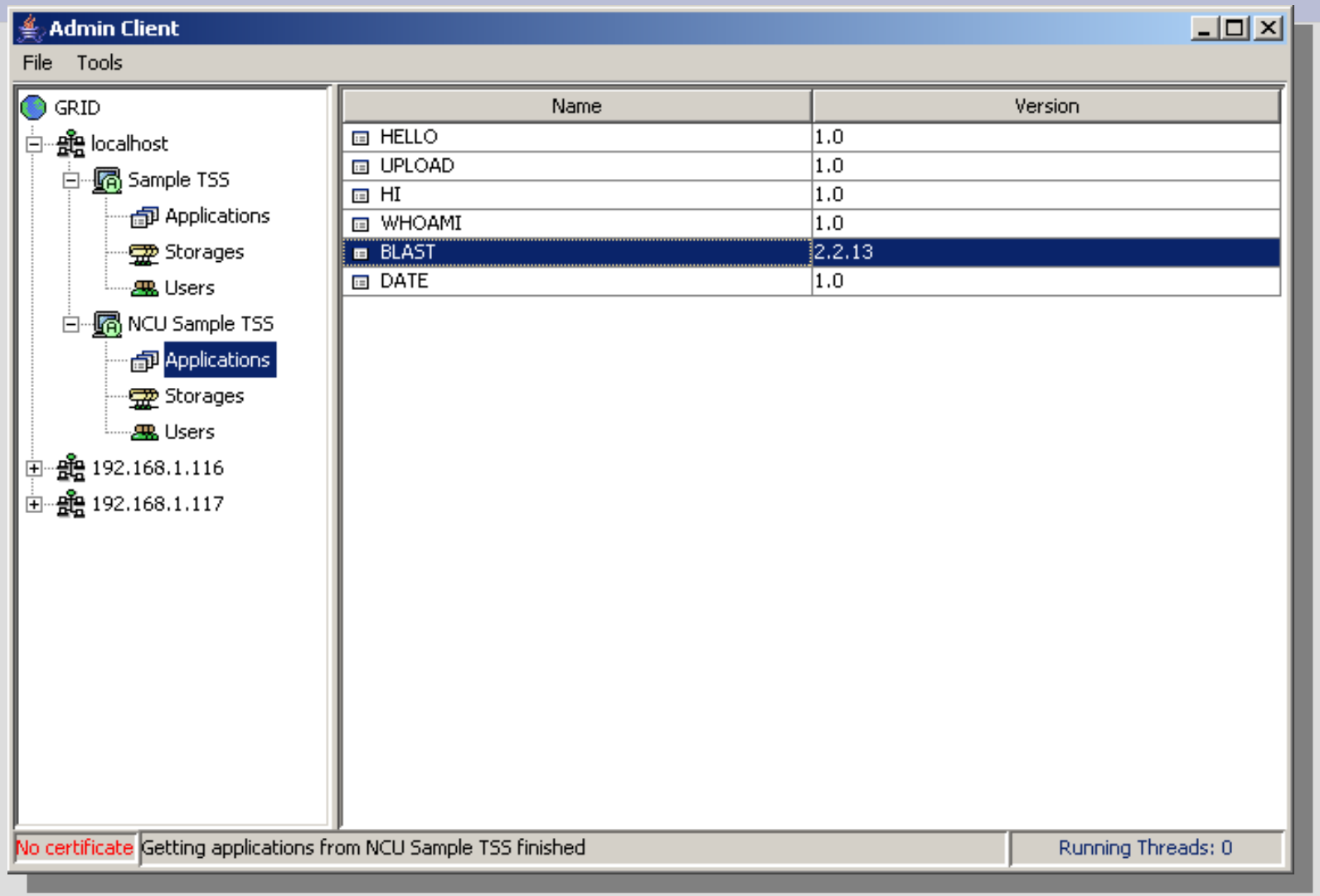
- The GPE Clients can contact to TSSs available on different hosting environments like:
  - Globus,
  - UnicoreGS,
  - others.

- Once implemented GridBean can be used with different Grid middlewares.

Interoperability



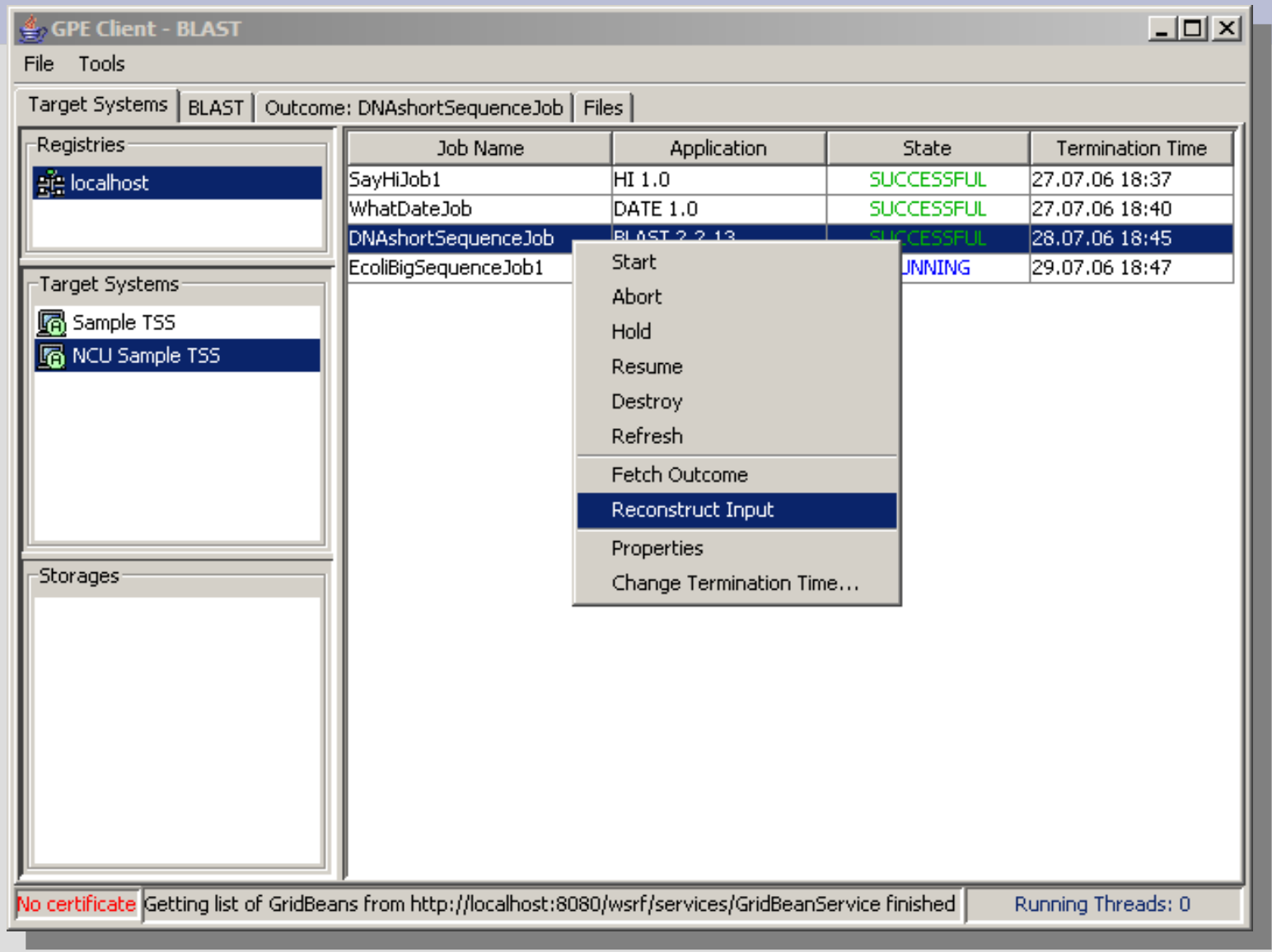
GPE Admin Client



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



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

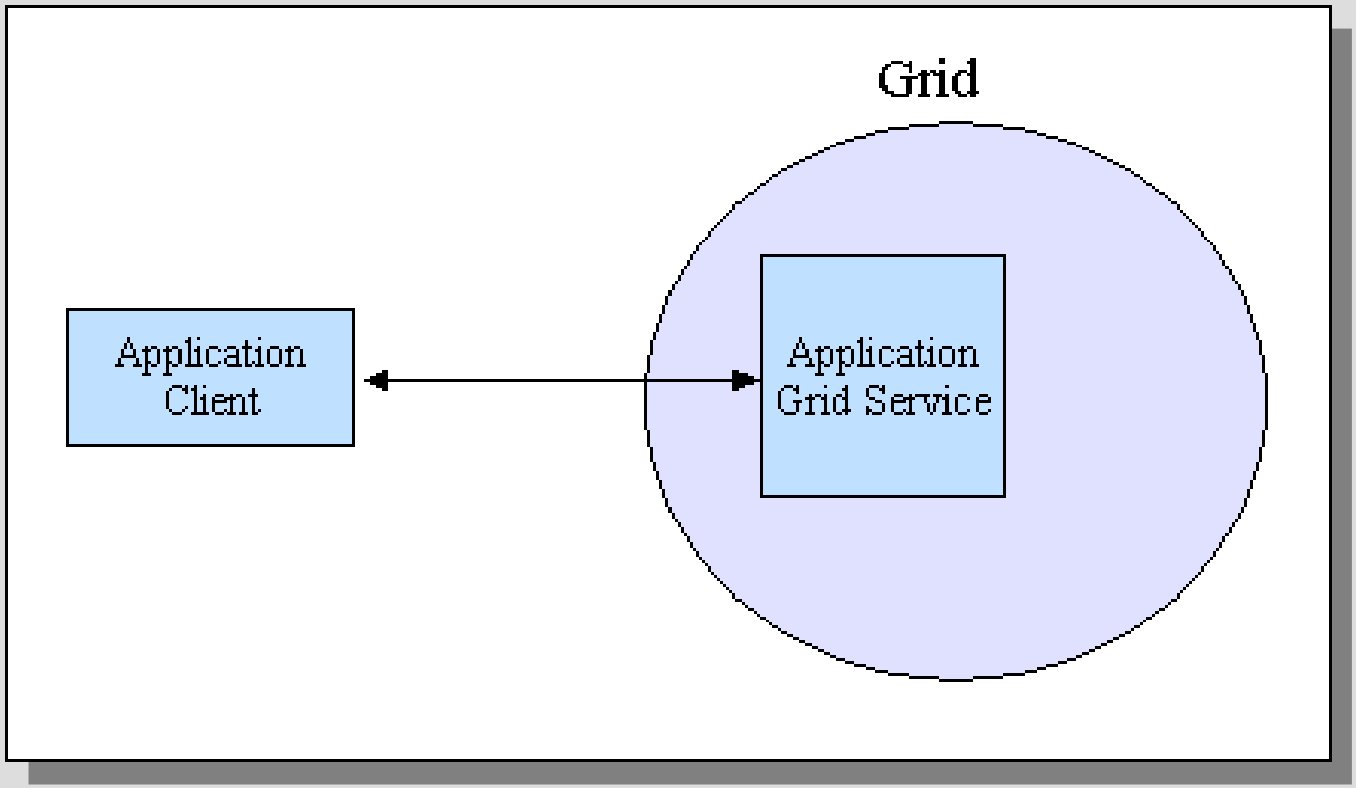
```
public class HiGridBean extends DefaultGridBeanModel implements IGridBean {

    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>
```

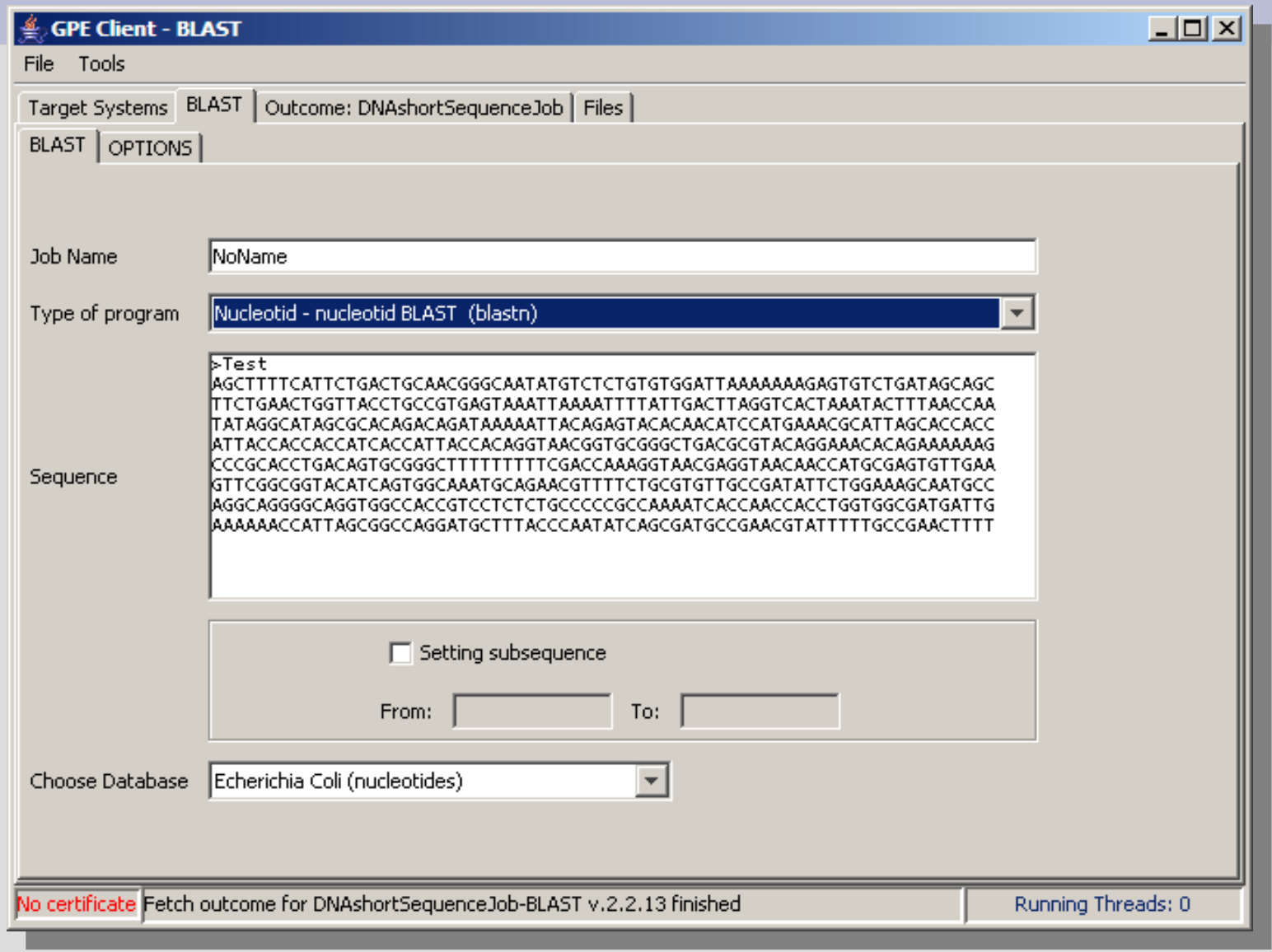
Application Services



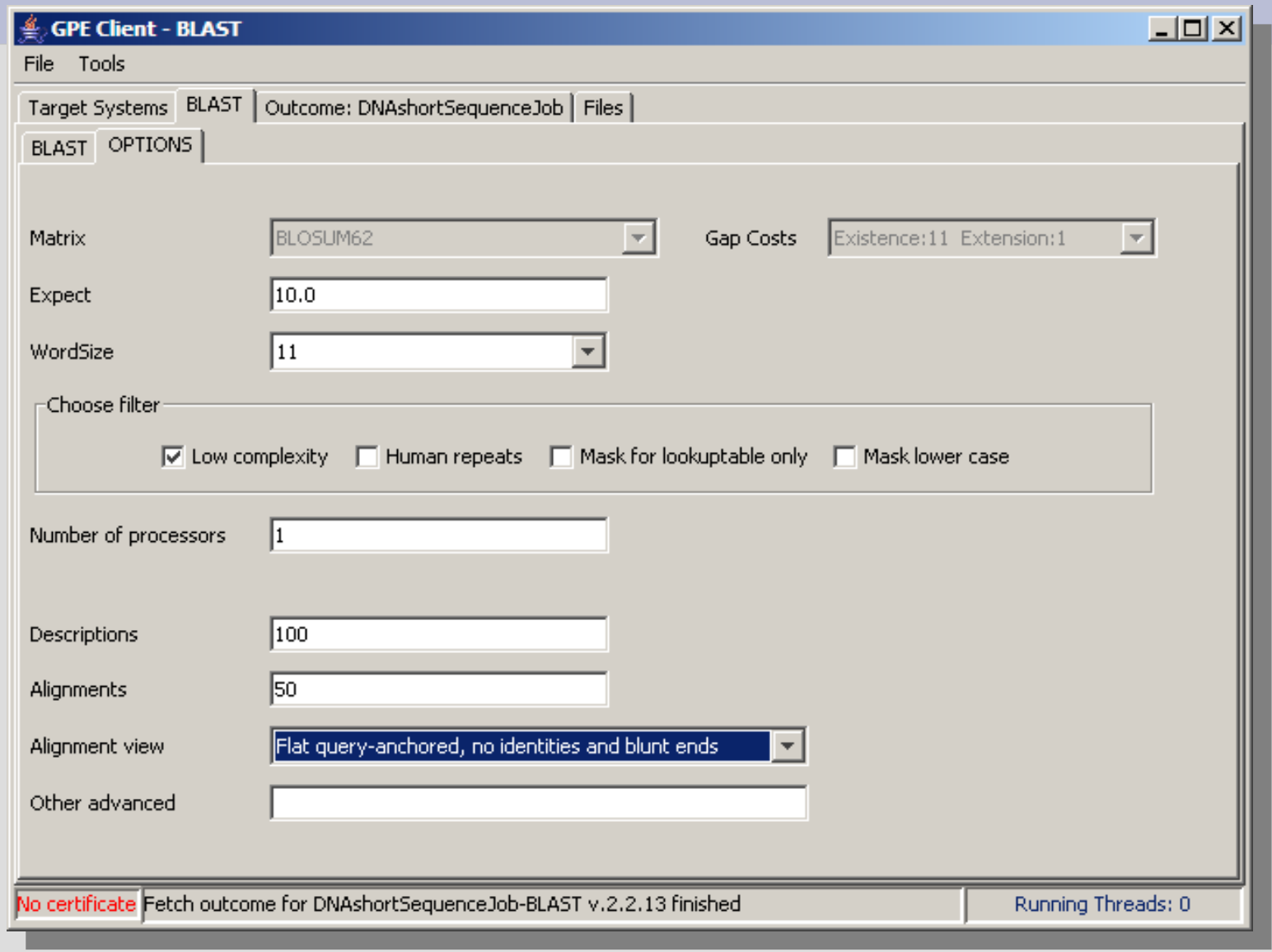
BLAST GridBean

- 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.

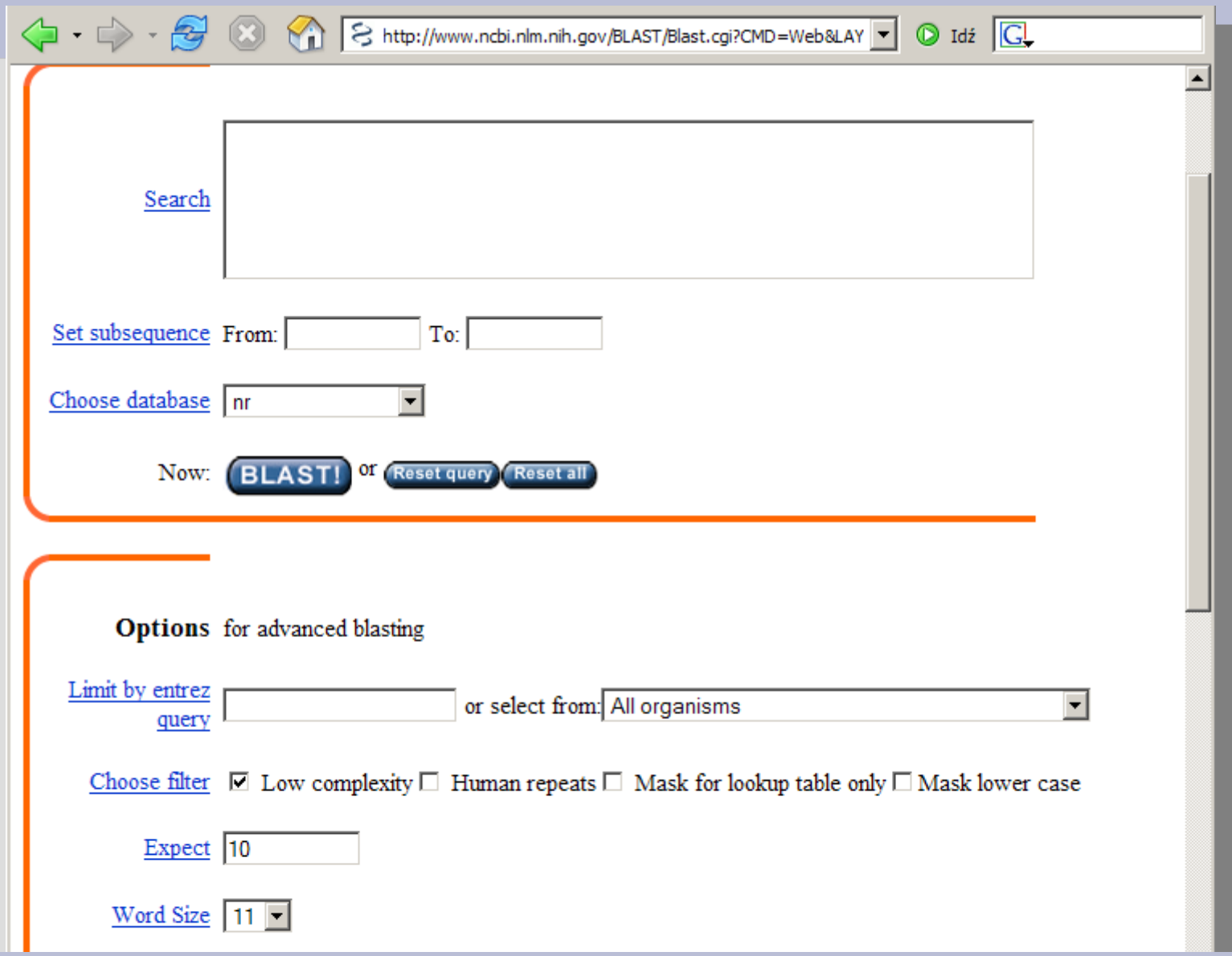
BLAST GridBean



BLAST GridBean



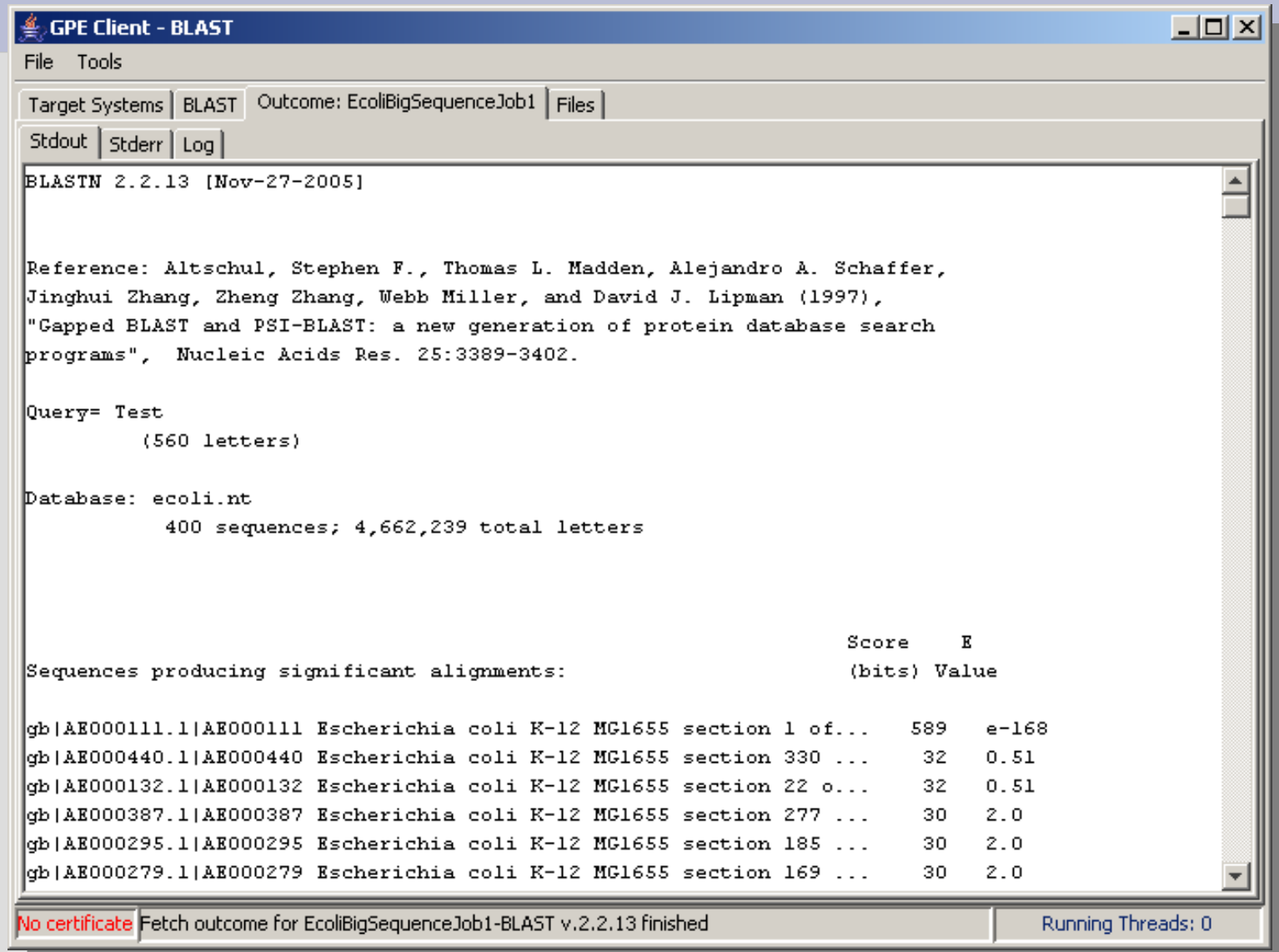
NCBI-BLAST Website



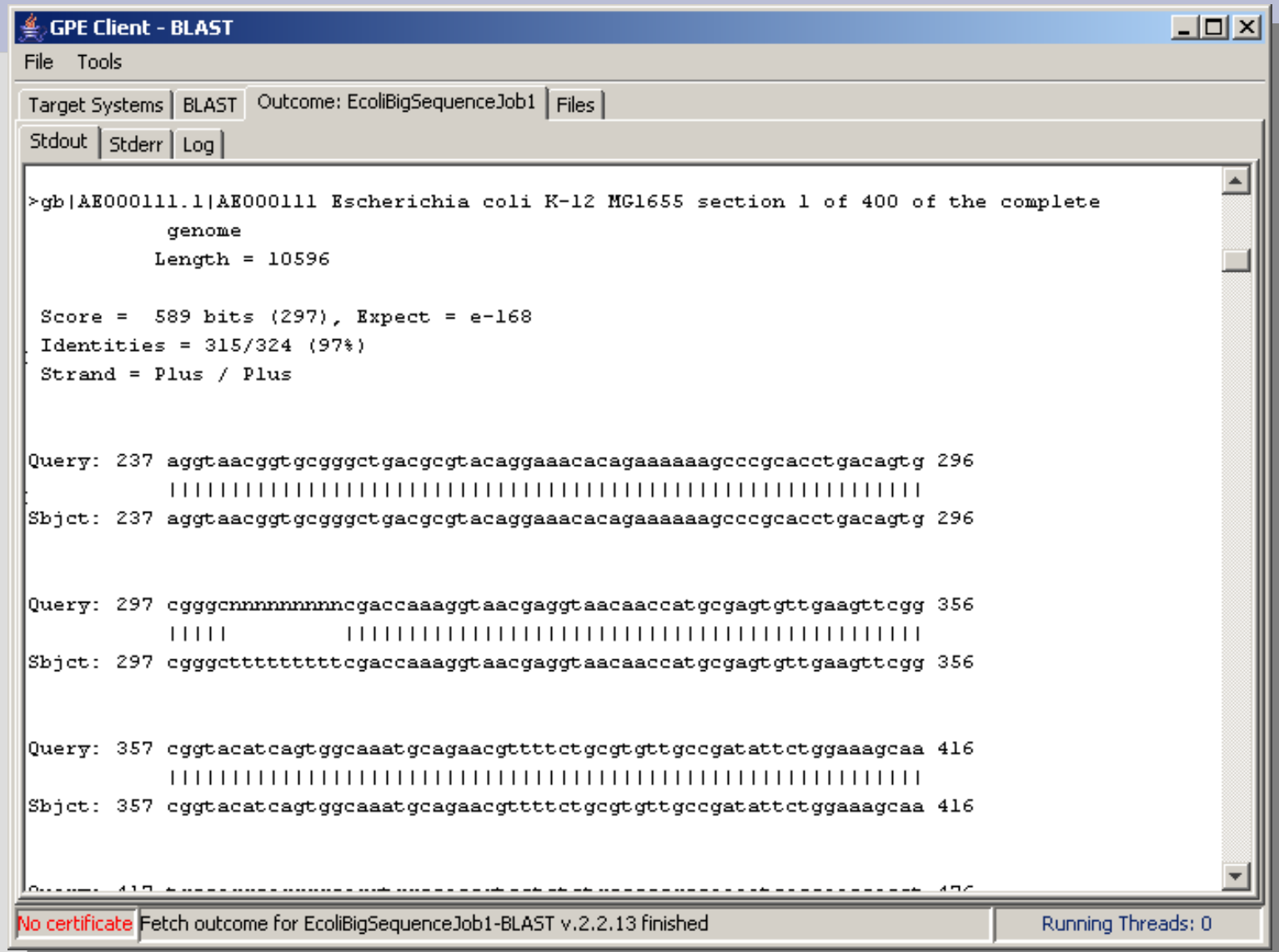
GridBean Panels

- 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



BLAST Alignments



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.

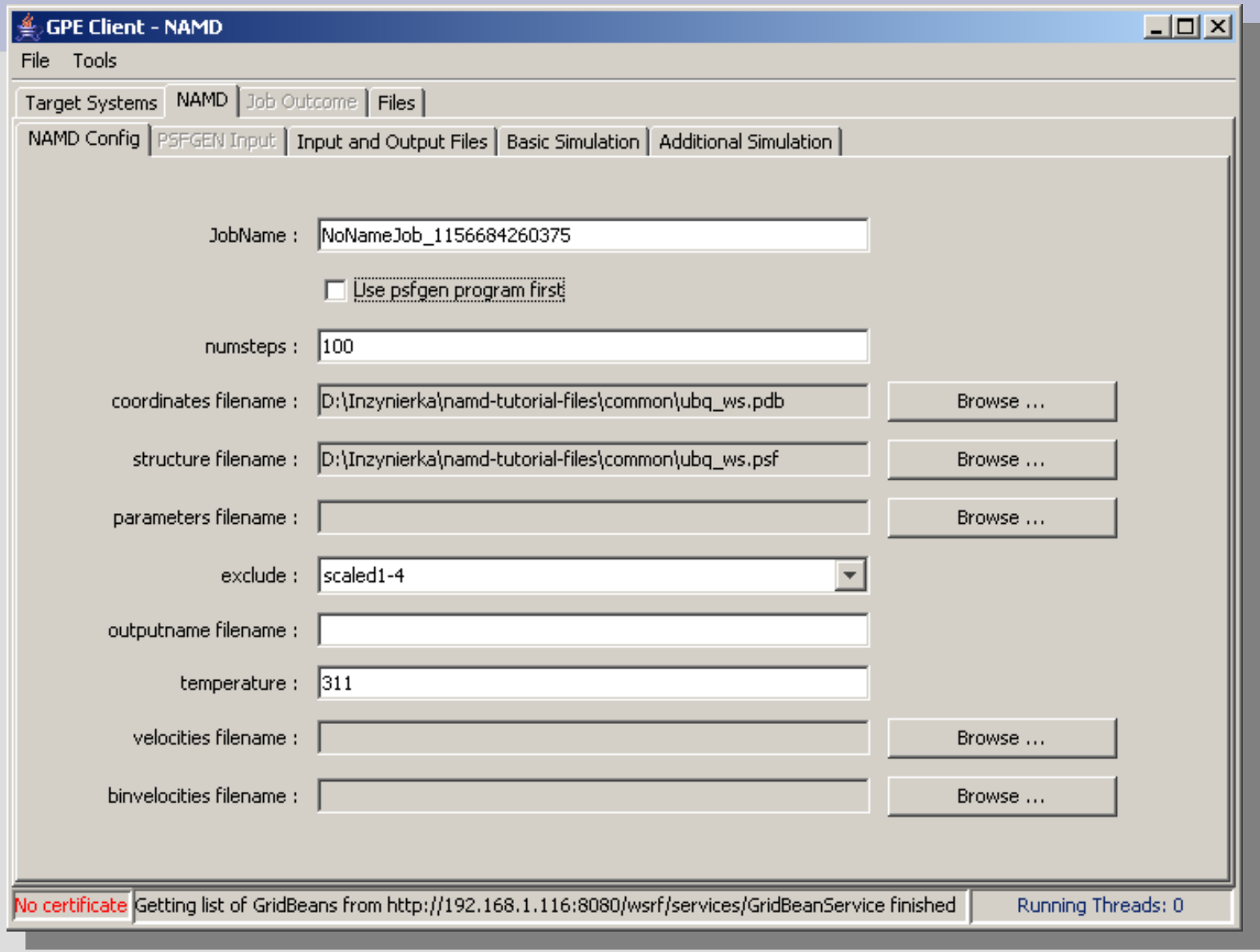
Conclusions

- 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.

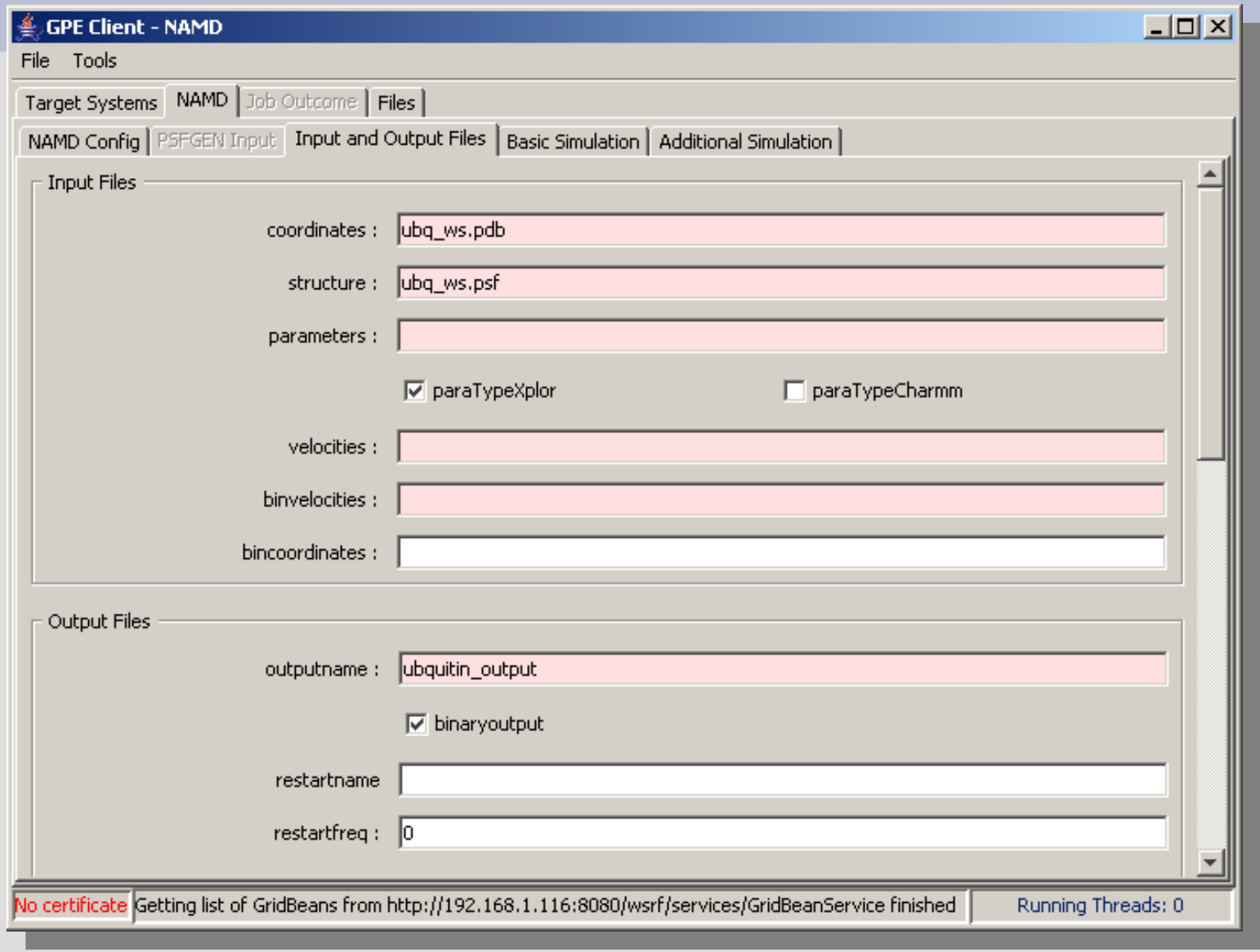
Future Work

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

NAMD GridBean



NAMD GridBean



Thank You !

Any questions, comments, advices *etc.* ?