



PatSearch in GENIUS

Flavio Licciulli

CNR – Institute of Biomedical Technologies (ITB), Bari

BIOINFOGRID Initial Course Training

Bari, 10 March 2006



PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences.

Giorgio Grillo¹, Flavio Licciulli¹, Sabino Liuni¹, Elisabetta Sbisà¹ and Graziano Pesole²

¹ Sezione di Bioinformatica e Genomica di Bari, Istituto Tecnologie Biomediche - CNR, via Amendola 168/5, 70125 Bari, Italy

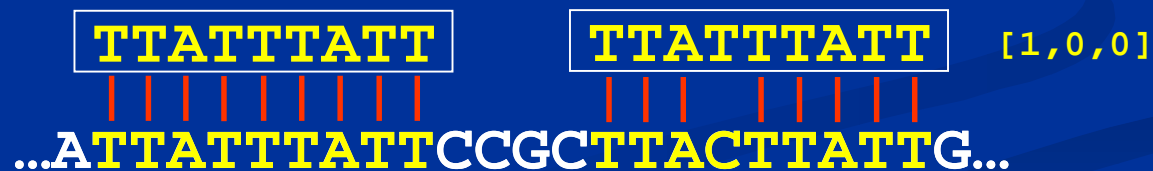
² Dipartimento di Fisiologia e Biochimica Generali, Università di Milano, via Celoria 26, 20133 Milano, Italy

WHY ?

- Regulation of gene expression at transcriptional and post-transcriptional level involves the interaction between short DNA or RNA tracts and the corresponding trans-acting protein factors. Detection of such elements in genome-wide screenings may significantly contribute to genome annotation and comparative analysis as well as to target functional characterization experiments.

WHAT IS ?

- A flexible and fast pattern matcher able to search specific combinations of oligonucleotide consensi, secondary structure elements and position-weight matrices also allowing for mismatches/mispairings below a user fixed threshold.





Pattern Syntax

■ String pattern unit

Two examples of string pattern units are **AAUAAA**, the polyadenylation site; or **TTTSSCGS** (S = C or G) the consensus site for E2F transcription factor (TRANSFAC site ID E2F\$CONS_01).

For example, **TTATTTATT[1,0,0]** would match any sequence with up to one mismatch with the sequence TTATTTATT.

■ Range pattern unit

1...200 match any subsequence from 1 to 200 characters.

■ Hairpin loop pattern unit

The pattern **p1=6...7 3...6 ~p1** matches a stem-loop structure with a stem six to seven nt long and a loop three to six nt long.

■ Position weight matrix (PWM) pattern unit

An example of PWM with similarity threshold 0.70 and core 0.90

{						
01	12	22	17	199	T	
02	210	12	9	19	*	A
03	46	32	36	136	*	T
04	159	33	29	29	*	A
05	158	37	31	24	*	A
06	11	21	9	209	T	
}	> 0.70, 0.90					

■ Repeat pattern unit

For example the syntax: **20 > frepeat (p1=NNN) 0...0 > 10** defines a sequence string of 10 to 20 identical tandem trinucleotides.

■ Length constraints

For example the pattern: **AUG p1=0...300 ((UAA|UAG)|UGA) length(p1) mod 3** can be used to search open reading frames up to 303 nt long in genomic sequences

■ Post-processing

It may be very convenient to be able to reprocess a section of a sequence that has been already matched



PatSearch is available on the web at
<http://www.ba.itb.cnr.it/BIG/PatSearch>

- developed in *C* using EMBOSS, SRS, NCBI libraries
- Unix (Compaq TRU64 Unix), Linux (RH 9)
- CGI (cgi_lite) interface through perl script
- user authentication and job management software (handmade, thesis work)
- outputs storage in a scratch area
- notification of the result to the user by email

BIG
Bioinformatics and Genomic Group - C.N.R.

PatSearch

[Help](#)

The PatSearch program is a pattern matching tool, that can find a well defined pattern against a given sequence(s) or database (primary or specialized) divisions.

The first time you use this program you should complete the form
[Registration Form](#)

Input

Sequence Type: DNA Protein

Enter sequence(s) below in [FASTA/EMBL/Genbank/...](#) format

or load it from disk:

or choose [Database](#) to search:

Query Pattern

Enter a query [pattern](#):

or load it from disk

Parameters

[Complement](#): [Overlap](#):

[Output Format](#):

[Maxmatch \(1..1000\)](#):

[Advanced Parameters](#):

Results by e-mail to:



Characteristics

- Low computational complexity (in general)
- Performance and execution time depends on:
 - pattern complexity
 - number of sequences to search (database size)
- Access to flat-file databases through:
 - direct access (Fasta, Embl, Genbank format)
 - EMBOSS, SRS, BLAST indices (ex: human division, accession number list,...)

General:

- Various process in execution contemporarily
- User space to manage and display results
- Statistical usage of the software



Test Case

`./PatSearch -INPUT=PRI.rna.gbff -OUTPUT=PRI.rna.gbff.out -COMMAND=vimentin.txt`

- **INPUT:** PRI.rna.gbff = 420Mb

RefSeq
(primates)
database

- **COMMAND:**

p1=yttrrrraa[2,0,0] 0...4

p2=cagcttcaagtgcctt[2,0,0] 0...2

p3=tscagtt[2,0,0] 6...7

p4=gagcg[2,0,0] 0...1 p5=aagatw[2,0,0]

p1/p2/p3/p4/p5:(p6=yttrrrraacagcttcaagtgccttscagttgagcgaagatw[2,0,0])

- **TIME:**

Real: 0m57.805s

User: 0m52.870s

Sys: 0m0.740s

Compaq/Digital AlphaServer
TRU64Unix (OSF)

- Preprocessing: database splitting (offline)
- 10 jobs => 1 x database chunk
- Postprocessing: output merging (online)



GENIUS implementation



Grid Enabled web eNvironment for site Independent User job Submission

RB: gilda-lcg

VO: gilda

Catalog: GILDA

Your Data

Logout

PatSearch is a flexible and fast pattern matcher able to search specific combinations of oligonucleotide consensi, secondary structure elements and position-weight matrices also allowing for mismatches/mispairing below a user fixed threshold.

It is able to find, in a given sequence(s), kinds of loop structures that characterize tRNAs, rRNAs (hairpin loop, stem loop with bulges or internal loops) and/or any kind of pattern in DNA and protein sequences.

Note: (*) You have to insert an unique production name.

Please, configure the inputs setting to start the production.

Production Name (*)

Database

Enter a query pattern

```
p1=yttrrrraa[2,0,0] 0...4
p2=cagcttcaagtgcctt[2,0,0] 0...2
p3=tscagtt[2,0,0] 6...7
p4=gagcg[2,0,0] 0...1 p5=aagatw[2,0,0]
p1/p2/p3/p4/p5:(p6=yttrrrraaacagcttcaagtgccttscagttgagcgaagatw[2,0,0])
```

load it from disk

or use the default one

Complement

Overlap

Maxmatch (1..1000)



GENIUS implementation

File Modifica Visualizza Vai Segnalibri Strumenti Fiestre Guida

Indietro Avanti Ricerca Stop <https://glite-tutor.ct.infn.it/> Cerca Stampa

Home Segnalibri mozilla.org mozillaitalia.org Forum di supporto Build di oggi

INFN
Istituto Nazionale di Fisica Nucleare


EnginFrame

genius

EGEE
Enabling Grids for E-science

Grid Enabled web eNvironment for site Independent User job Submission

RB: gilda-glite VO: gilda Catalog: GILDA Your Data Logout

 Production Name : test
Number of Events : 4
Last Submission Time : Mar 8 15:56:53
CET 2006

Production Status: Started

1) https://glite-rb.ct.infn.it:9000/smpg-_p2IGOfAU6YNZKnzA
2) <https://glite-rb.ct.infn.it:9000/Qi6WQ2djLHaN3ZEbPi4EhQ>
3) <https://glite-rb.ct.infn.it:9000/kzhK2qcZhafevx-11918yg>
4) <https://glite-rb.ct.infn.it:9000/SpYPGddmW3kTKJyLarsuMQ>

Ready Jobs = 0
Scheduled Jobs = 0
Done Jobs = 0
Aborted Jobs = 4
RUNNING Jobs = 0
PENDING Jobs = 0

powered by
[EnginFrame 3.2](#)
compliant with
[LCG-2 GRID-IT](#)
[gLite-1](#)



GENIUS implementation

Grid Enabled web eNvironment for site Independent User job Submission

RB: gilda-glite	VO: gilda	Catalog: GILDA	Your Data	Logout
<input type="button" value="Destroy"/>	Directory contents - prova_20060308_113703/flavioli_patsearch_prova_20060308113703			
Home				
Back				
flavioli_GIE0Daqq4eS5qOgTDrkeHQ	4,096	flavioli_SiYW7vfafFzmuUtFqQTIMQ	4,096	
flavioli_vfIdvL3- fh3LKqkoHWdg	4,096	final computation	4,096	
flavioli_-Xa81Tr0aBJ3GYFWAIXx2A	4,096			

powered by
EnginFrame 3.2
compliant with
LCG-2 GRID.IT
gLite-1



GENIUS implementation

The screenshot shows a web browser window displaying the GENIUS interface. The browser's address bar shows the URL <https://glite-tutor.ct.infn.it/>. The page features logos for INFN, EnginFrame, genius, and EGEE. Below the logos, the text reads "Grid Enabled web eNvironment for site Independent User job Submission".

The main content area displays a table with the following structure:

RB: gilda-glite	VO: gilda	Catalog: GILDA	Your Data	Logout
Directory contents -				
Destroy	prova_20060308_113703/flavioli_patsearch_prova_20060308113703/flavioli_GIE0Daqq4eS5qOgTDrkeHQ			
Home				
Back				
std.out	835	PRI.rna.gbff.out	382,439	
std.err	0			

A red arrow points to the [std.out](#) link in the table. The left sidebar contains navigation links for PATSEARCH, including "up", "Configure Inputs Setting", "Inspect Status", "Clean PATSEARCH Queue", and "PATSEARCH Data Spooler". At the bottom, it states "powered by EnginFrame 3.2 compliant with LCG-2 GRID-IT gLite-1".



GENIUS implementation

File Modifica Visualizza Vg Segnalibri Strumenti Piastre Guida

Indietro Avanti Ricarica Stop <https://glite-tutor.ct.infn.it/> Cerca Stampa

Home Segnalibri mozilla.org mozillaitalia.org Forum di supporto Build di oggi

INFN
Istituto Nazionale
di Fisica Nucleare

EnginFrame

genius

EGEE
Enabling Grids
for E-science

Grid Enabled web eNvironment for site Independent User job Submission

dgt08.ui.savba.sk
Using LFC Catalog type.

PatSearch

PatSearch ...

Author: GIORGIO GRILLO

Pre-Processing...

Processing Commands...

Processing...

Post-Processing...

Match Total = 5000 Matched Sequences = 1973

Sequences = 5000

powered by
[EnginFrame 3.2](#)
compliant with
[LCG-2 GRID-IT](#)
[gLite-1](#)

PATSEARCH

up

- ▶ PATSEARCH
- ▶ Configure Inputs
- ▶ Setting
- ▶ Inspect Status
- ▶ Clean PATSEARCH
- ▶ Queue
- ▶ PATSEARCH Data
- ▶ Spooler



GENIUS implementation

Grid Enabled web eNvironment for site Independent User job Submission

RB: gilda-glite	VO: gilda	Catalog: GILDA	Your Data	Logout
Destroy	Directory contents - prova_20060308_113703/flavioli_patsearch_prova_20060308113703			
Home				
Back				
flavioli GIE0Daqq4eS5qOgTDrkeHQ	4,096	flavioli SiYW7vfafFzmuUtFqQTIMQ	4,096	
flavioli vfldyL3- fh3LKqkoHWdg	4,096	flavioli final computation	4,096	
flavioli -Xa81Tr0aBJ3GYFWAIXx2A	4,096			

powered by
EnginFrame 3.2
compliant with
LCG-2 GRID.IT
gLite-1



GENIUS implementation

powered by
EnginFrame 3.2
compliant with
LCG-2 GRID-IT
gLite-1

PatSearch

COMMAND :
p1=rrrcwggyyy[3,0,0] 0...13 p2=rrrcwggyyy[3,0,0]
0...13 p3=rrrcwggyyy[3,0,0]
p1/p2/p3: (p4=rrrcwggyyyrrrcwggyyy[3,0,0])
p1/p2/p3: (p5=nnnnnnnnnnnnnnnnnnnn)

XM_526565 : [201,246] : gagcaattec cgattggtagaa gaacttgctg taac acacaogtac
XM_526565 : [253,292] : gaccccgccc caa cgactagccc ctctct gtgcaacccc
XM_526565 : [1764,1805] : ggggatctgc tgtgat gggcaaatc cattca aggtttact
XM_526567 : [208,238] : aaaagaattt c atactggct caactgect
XM_526570 : [969,1005] : aagtatgect c agactcctcc otttto agcaatgttt
XM_526573 : [690,742] : aaagatatcc aggtcacttg gggotagtgt ttattgtgttggg gtcottgatt
XM_526573 : [822,857] : aggaataatt aattct atactagctc agtcatgtac
XM_526574 : [235,268] : gaacaagccc ct aaggaaatto tt cagtctgtct
XM_526574 : [470,503] : gaatatgtaa ctt cgtcaagtct a agacaggct
XM_526574 : [871,923] : ggacaatgt gcttcaggag ggacaagcca aaaacacctatg gaacaggttg
XM_526578 : [315,352] : taacaatctg ttaag ggttttgcct tat gaacttgctc
XM_526578 : [878,918] : agtcaogtgc caa aatcatgtct gagcagga gagaaagcat
XM_526581 : [33,72] : caacaagggc cacaaggt gaccaagat gt gagcaagccc
XM_526582 : [205,235] : aatggagttc aaacatgtct a aaccagatt
XM_526583 : [1234,1275] : gaactctcc a gaactcact cactataaac acacaagat



GENIUS implementation

WHISHES:

- More database to search, bigger database/dataset
 - usage of Emboss, SRS, Blast indices
- User defined set of sequences
- Dynamic/automatic splitting of the searched database



GENIUS implementation

Thanks to

Giuseppe La Rocca (INFN Catania)

and the

GILDA's team