

Computational Biology Applications Suite for High Performance Computing (BioHPC)

Jaroslav Pillardy

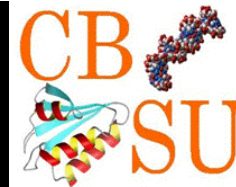
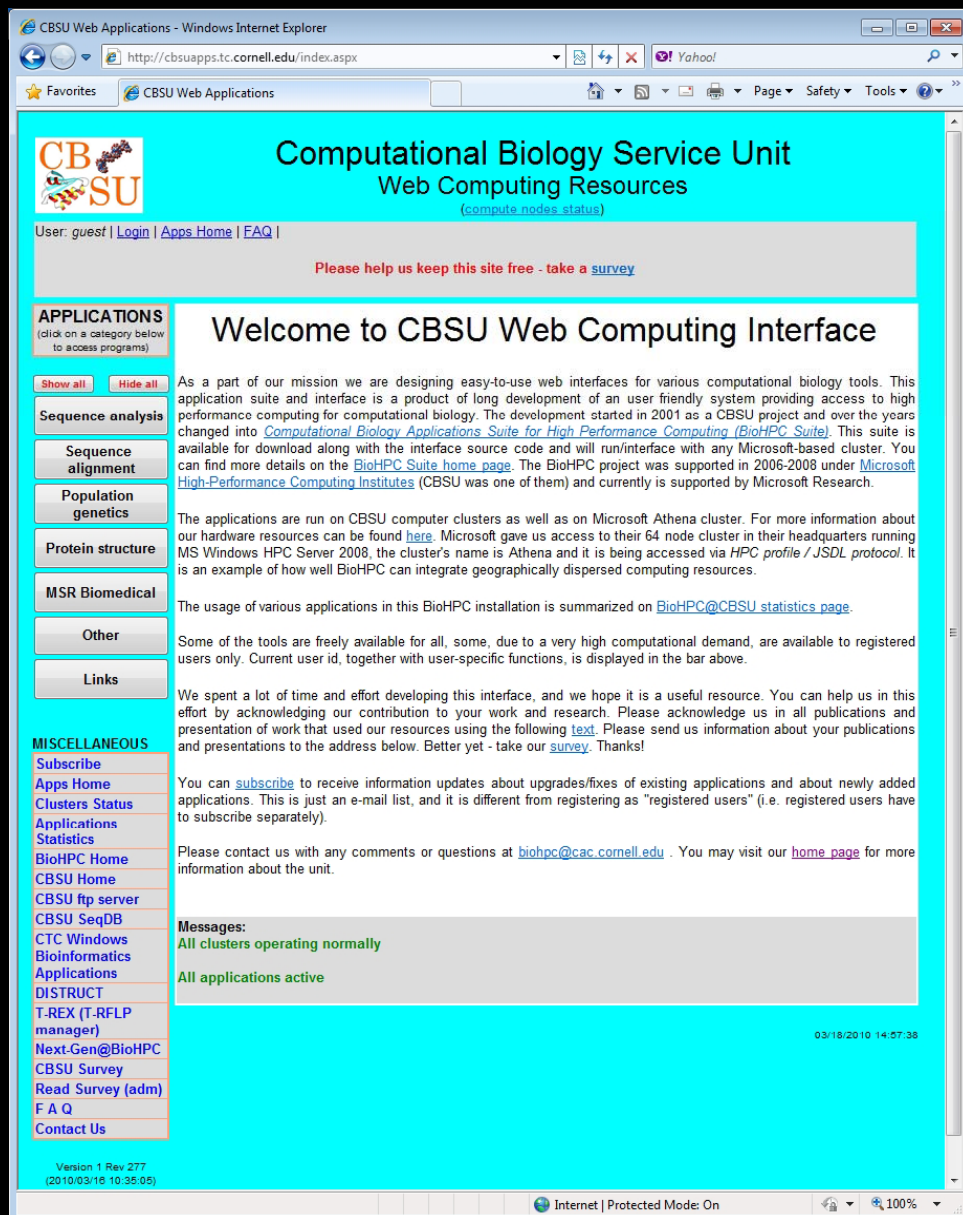
CBSU,
Life Sciences Core Laboratories Center
Cornell University

Microsoft External Research Symposium

April 6-7, 2010
Redmond, Washington



Computational Biology Applications Suite for High Performance Computing (BioHPC)

The screenshot shows a web browser window titled "CBSU Web Applications - Windows Internet Explorer". The address bar shows "http://cbsuapps.tc.cornell.edu/index.aspx". The page content includes a header with the CBSU logo and the text "Computational Biology Service Unit Web Computing Resources". Below the header, there are navigation links for "User: guest", "Login", "Apps Home", and "FAQ". A survey request is displayed: "Please help us keep this site free - take a survey". The main content area is titled "Welcome to CBSU Web Computing Interface" and contains a paragraph of introductory text. To the left, there is a sidebar with "APPLICATIONS" and "MISCELLANEOUS" sections. The "APPLICATIONS" section lists categories like "Sequence analysis", "Protein structure", and "MSR Biomedical". The "MISCELLANEOUS" section lists various links such as "Subscribe", "Apps Home", "Clusters Status", and "BioHPC Home". At the bottom of the page, there is a "Messages" section indicating "All clusters operating normally" and "All applications active". The footer shows "Version 1 Rev 277 (2010/03/16 10:35:05)".

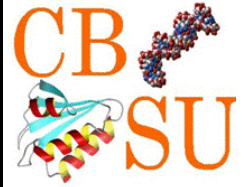
- User accessibility is major problem in HPC and bio-computing
- Using a parallel cluster or remote resources required knowledge of the operating system, queuing system and parallel programming
- BioHPC suite provides easy access to standardized applications on Windows platform
- BioHPC suite provides easy way manage and integrate distributed computational resources
- Written in C# / ASP.NET



Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

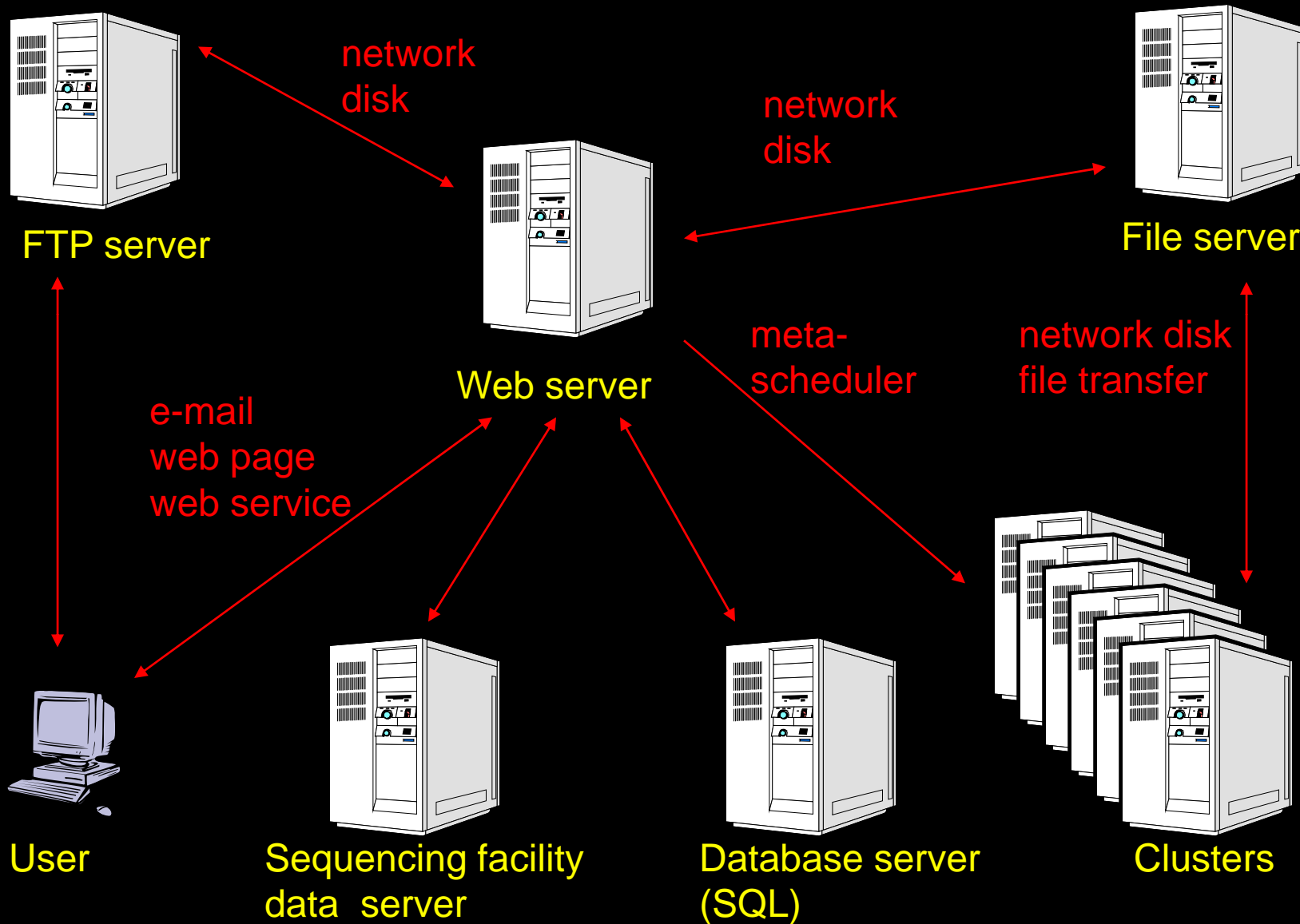
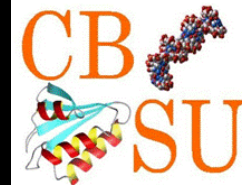
Computational Biology Applications Suite for High Performance Computing (BioHPC)



What it does:

- Web-based, point-and-click access to a variety of bioinformatics applications with the underlying structure of computational platform transparent to the user
- Enhancement of standard applications through parallelization, transparent to the user
- Integration and simplified access to geographically dispersed hardware resources
- Web-based administration of users, jobs, applications, and clusters within the suite
- Standardized access to and maintenance of bioinformatics databases
- Next generation data management and distribution – from sequencing facility to users.
- Next generation sequencing pipelines and applications – with internal data management or optional external data upload

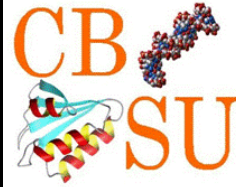
Computational Biology Applications Suite for High Performance Computing (BioHPC)





Computational Biology
Service Unit

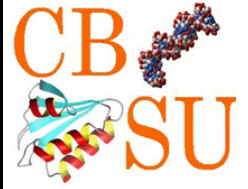
Computational Biology Applications Suite for High Performance Computing (BioHPC)



ARCHITECTURE

- Web server running the interface (ASP.NET C#)
- Microsoft SQL server (ADO.NET)
- Compute clusters running Microsoft Windows
- Ftp server / file server
- Two local compute cluster schedulers are supported (CCS and HPC Server 2008)
- Remote clusters can be used via JSDL/HPC Profile

Computational Biology Applications Suite for High Performance Computing (BioHPC)

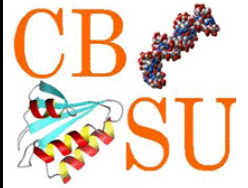


37 applications available

- **Data mining / sequence analysis** (BLAST, HMMER, InterProScan, GIMSAN, SLIM)
- **Protein structure prediction and modeling** (LOOPP, Modeller)
- **Population genetics** (BEAST, BEST, Clumpp, IM, IMa, IMa2, InStruct, LAMARC, MDIV, Migrate, MKPRF, MSVAR, OmegaMap, Parentage, SFS_CODE, Structurama, Structure, TESS)
- **Phylogenetics** (MrBayes, ClustalW, Stretcher, T-COFFEE)
- **Association analysis / statistics** (PLINK, R)
- **MSR Biomedical** (CreateEpitome, Epipred, FalseDiscoveryRate, HlaAssignment, HlaCompletion, PhyloD)

The system is flexible and can be easily customized to include other software. The interface to each application is standardized, users can choose the cluster, number of nodes or allow the interface to determine it based on the best load balance and node availability

Computational Biology Applications Suite for High Performance Computing (BioHPC)



The most popular applications

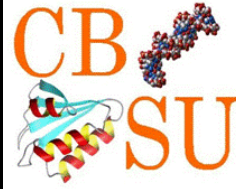
Job submission from 6/13/2003 to 3/18/2010

LOOPP	20,385	protein structure prediction
MDIV	20,965	population genetics
P-BLAST	4,504	sequence analysis / data mining
MrBayes	18,799	population genetics
IM/IMa/IMa2	22,567	population genetics
STRUCTURE	17,968	population genetics

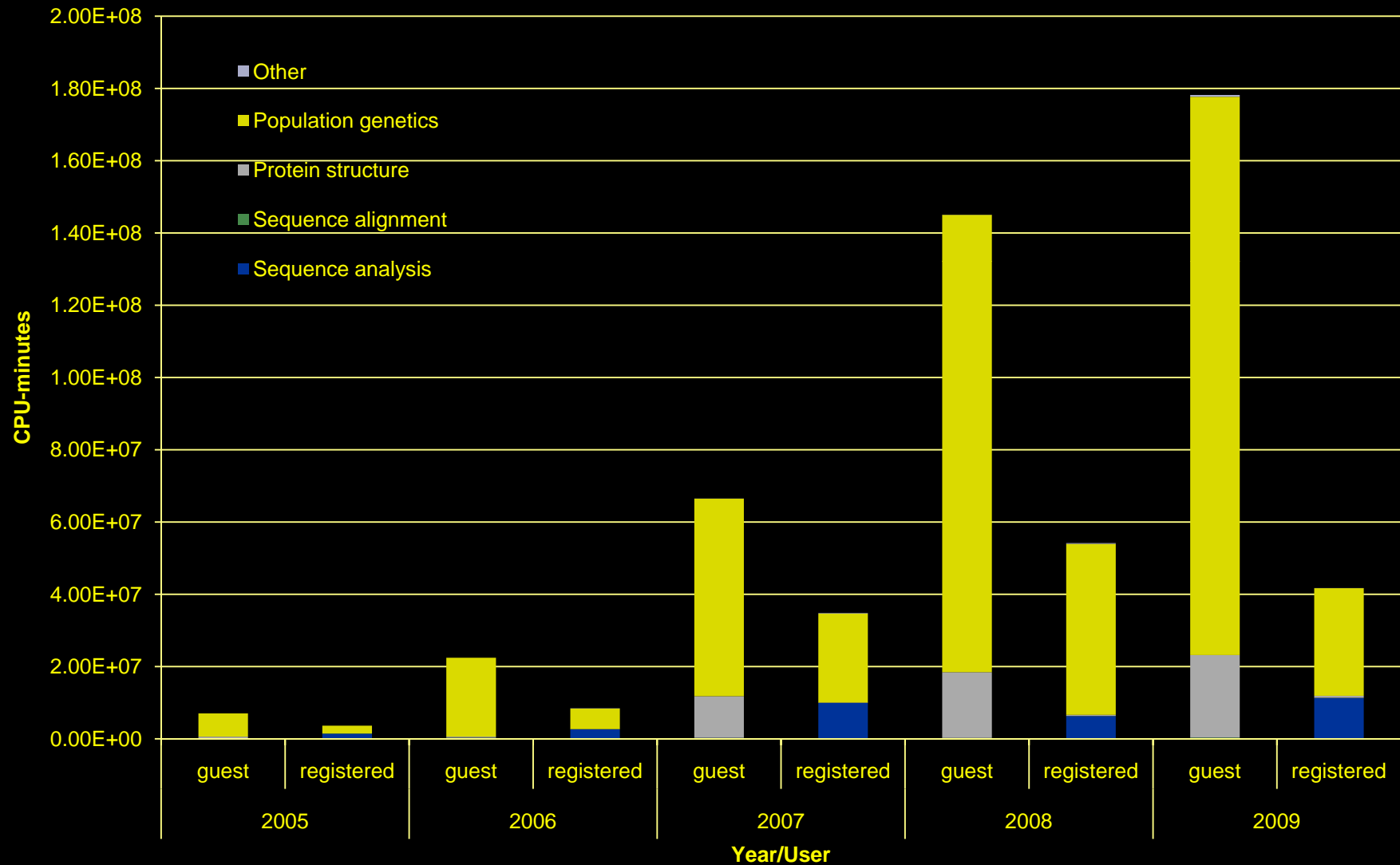
All applications **140,141** (average **20,090** per year, **49,738** last year)

LOOPP	parallel, uses 5-20 cores for 3-10 hours
MDIV	serial, uses 1 core from few hours to two weeks (average: 2-5 days)
P-BLAST	parallel, restricted resource, uses 10 – 100 cores for a few days to a week (average: few days)
MrBayes	parallel, uses 8-20 cores for a few hours to two weeks (average: a week)

Computational Biology Applications Suite for High Performance Computing (BioHPC)



Dynamics of BioHPC Utilization

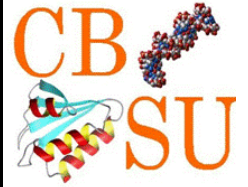




Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



The jobs were submitted by 11,471 unique users from 83 countries

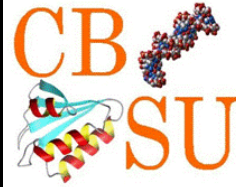
The majority (57% by CPU time used) coming from the USA

52% of the USA utilized CPU time coming from New York

Among them there are

- 257 unique Cornell users,
- 2,580 users from *.edu* domains
- 426 unique *.edu* institutions
- 4,813 users from *.com* domains
- 4,191 users with Yahoo, Gmail and Hotmail e-mail addresses

Computational Biology Applications Suite for High Performance Computing (BioHPC)



CBSU Web Computing Resources
Microsoft High-Performance Computing Institute
([compute nodes status](#))

User: [jpb6@cornell.edu](#) | [Apps Home](#) | [Logout](#) | [Change password](#) | [My jobs](#) | [Manage jobs](#) | [Administration](#)

APPLICATIONS
(click on a category below to access programs)

Show all Hide all

Sequence analysis

- Genome SeqAlign
- P-BLAST**
- P-HMMER
- P-IPRSCAN
- RepeatFinder
- RetrieveSeq

Sequence alignment

Population genetics

Protein structure

Other

Links

MISCELLANEOUS

- [Apps Home](#)
- [Clusters Status](#)
- [Applications](#)
- [Statistics](#)
- [CBSU Home](#)
- [CBSU ftp server](#)
- [CBSU SeqDB](#)
- [CTC Windows](#)
- [Bioinformatics](#)
- [Applications](#)
- [Contact Us](#)

P-BLAST @ CBSU

You will receive an e-mail once the blast search is finished. The blast calculations will be carried out at the CBSU compute nodes cluster at the Cornell Theory Center.

NOTE: Do not use "Back" button in your browser to reaccess this page after submitting. Use menu on the left instead.

Job name: (please, no spaces, special characters etc., underscore is OK)

Query file (Required): upload copy paste

Upload your FASTA file. This is an http upload, and it may be too slow for large files (10MB or more). Server will not accept http upload for files larger than 40MB. For large files please consider using "copy" option.

Choose database for BLAST: from list copy upload

Choose database from a list below for the following category: Common

Available	Chosen
arabidopsis_genome	
chimpanzee_WIBR	
chimpanzee_WUGSC	
est_others	
nt	
pdbnt	
RefSeq_mammalian.ma	
rice_cds TIGR release 3	
rice_con TIGR release 3	
rice_seq TIGR release 3	

BLAST program: Run: CPUs: Cluster:

([Show timeout info](#))

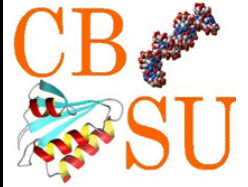
Options:

Output File Format: Low Complexity Filter:

Cutoff E-value: Maximum Targets:

- Each application interface is standardized as much as practical
- Some applications can be used only by registered users
- Users can upload their data files via http, place them on our ftp server, or use their local network drive
- In addition to application-specific options, users can choose number of nodes, scheduler, and cluster

Computational Biology Applications Suite for High Performance Computing (BioHPC)



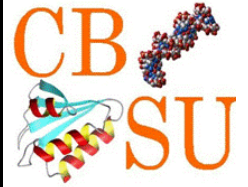
- For trivially parallelized programs extensive control over task performance is provided, preventing waste of computational resources in a case of errors in input
- Some application-specific options are available on the interface, some rarely used ones can be entered as string



Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



The screenshot shows a web browser window at the URL `http://cbsuapps.tc.cornell.edu/jarekp/blast_s.aspx`. The page title is "P-BLAST @ CBSU". The main content area displays the following text:

```

Submitting your job ... please wait
registering job info ...
* user => jp86@cornell.edu
* cluster => cbsu1
* scheduler => vsched
* CPUs => 10
* shared binaries from => H:\CBSU\bin_x32\
* parallel binaries from => H:\CBSU\pTools\mpipro\
* job id => 26123
... done
creating directories ... done
creating files ...
query file "pblast_job" done
options file done
script file 1 done
script file 2 done
... done
submitting job to cluster ...

* cluster job id => 71578
... done
finalizing db entries ... done
Your P-BLAST job pblast_job (26123) has been SUBMITTED

You will receive another email once your job is finished.

You can view the current results here
Your final result file will be available for download via http here or via ftp here
Your final result file will be available in compressed form (gz) for download via http here or via ftp here
You may follow program's progress by viewing here
Timeout information and the current job status can be found here
If you want to CANCEL the job, please click here

Messages:
All clusters operating normally
All applications active
  
```

The left sidebar contains a navigation menu with categories like "Sequence analysis", "Protein structure", and "Miscellaneous". The "P-BLAST" link under "Sequence analysis" is highlighted.

A job is now submitted

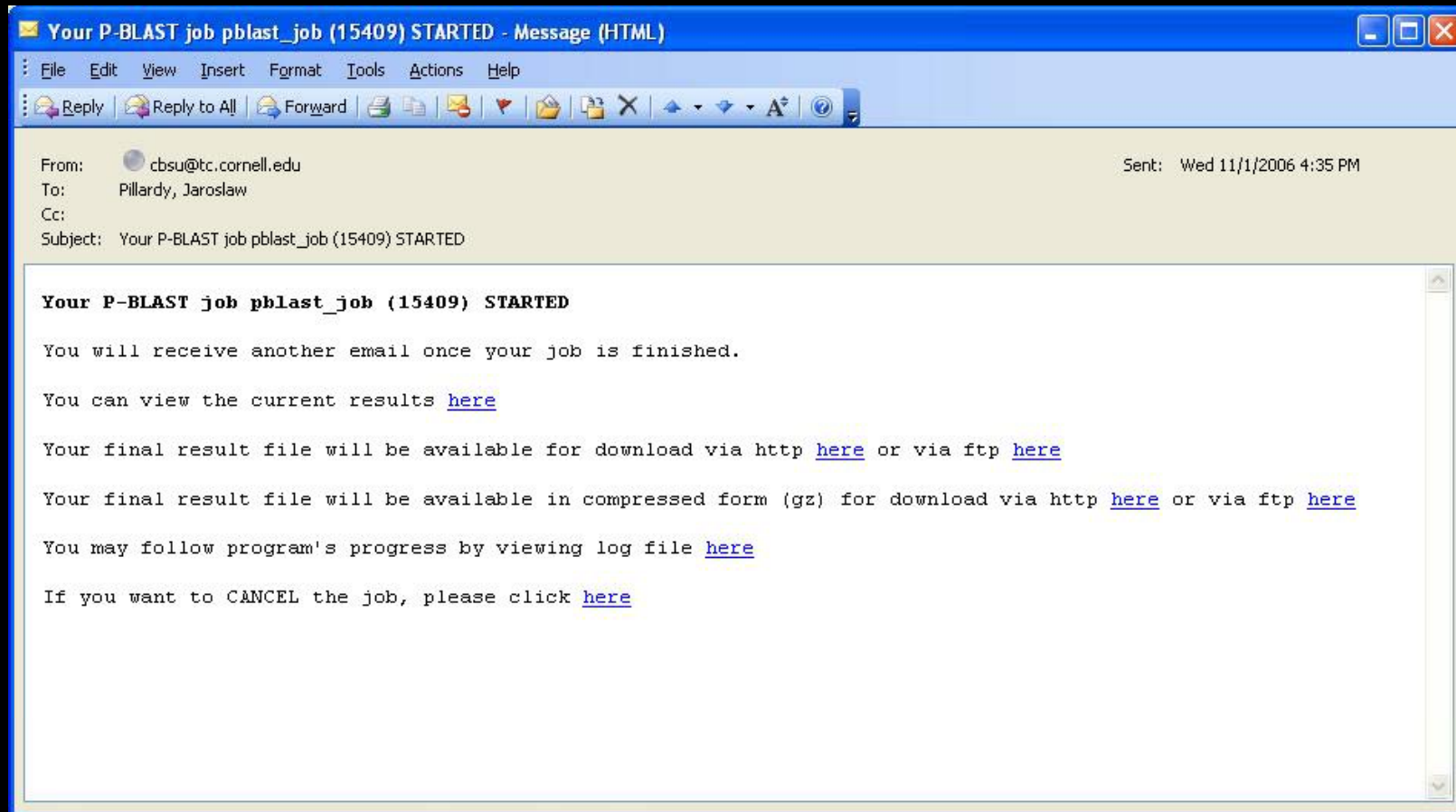
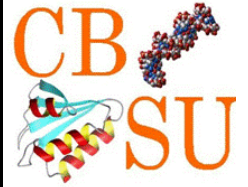
An email has been dispatched with links to output files and job control functions. These links are also available on this page along with submit log.



Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



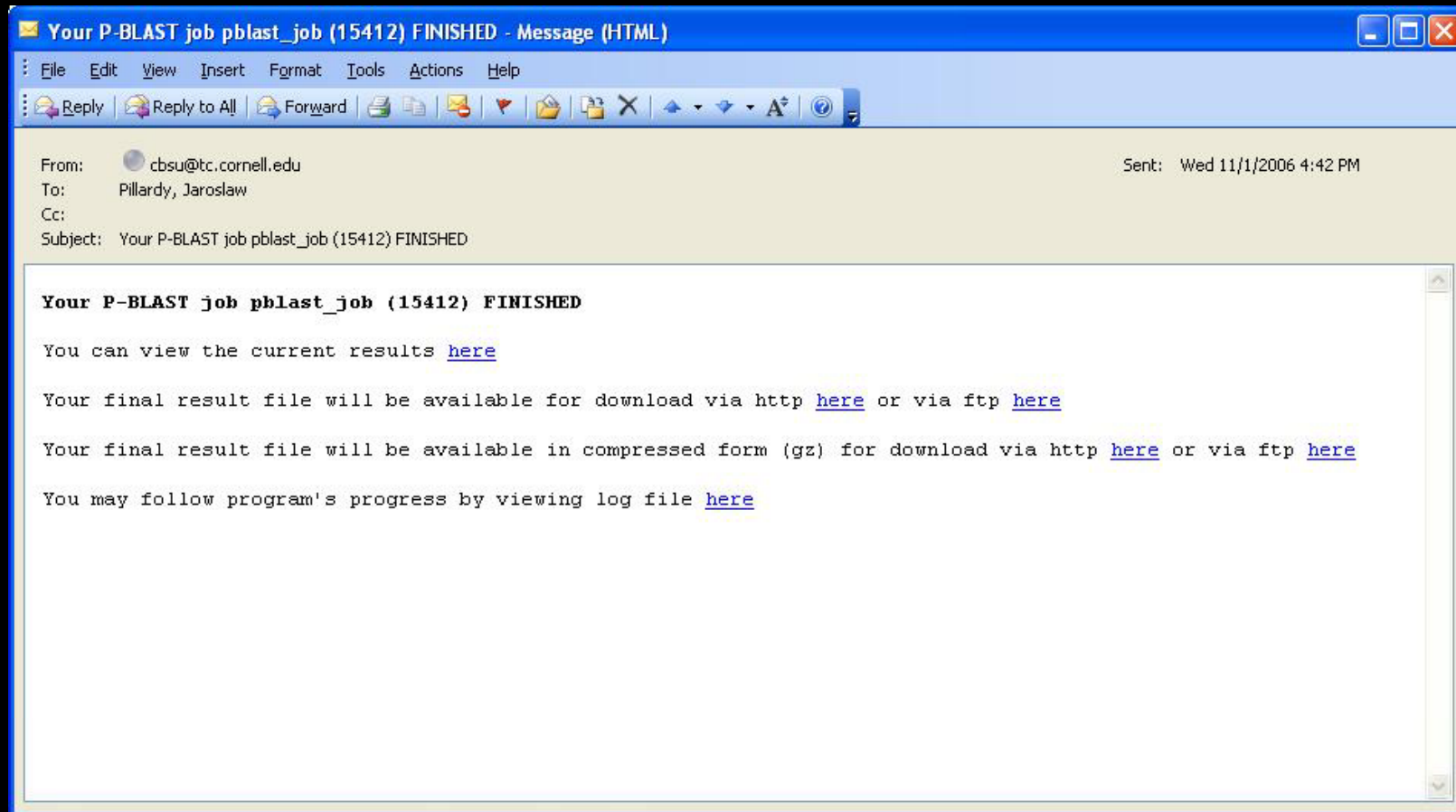
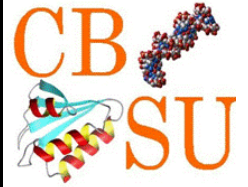
Jobs interact with user via e-mails. Links in the e-mail allow for viewing current results, computations progress (log) as well as cancelling the job if necessary.



Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

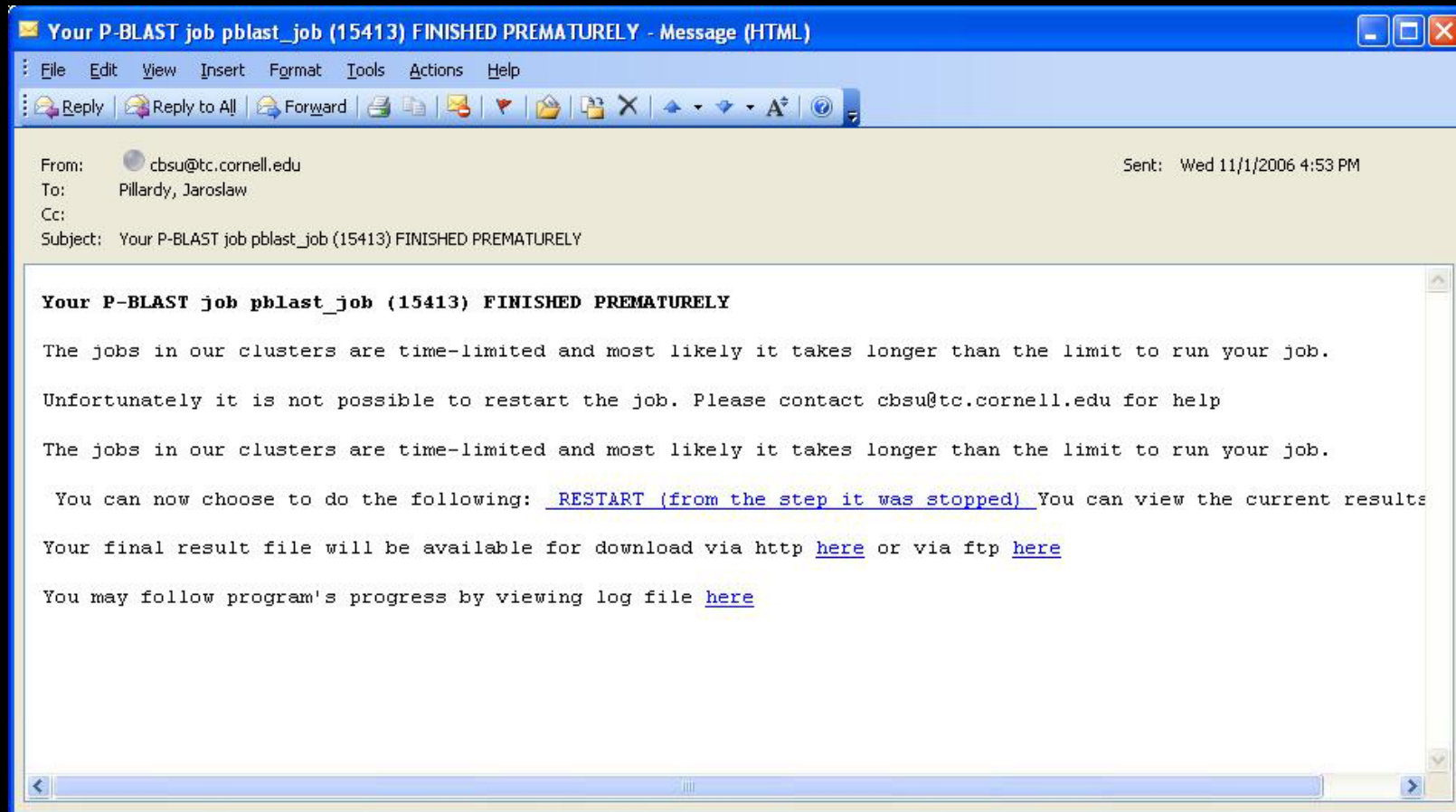
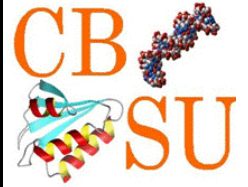


When job finishes, another e-mail is sent.



Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



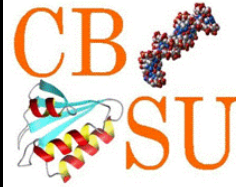
Sometimes a job finishes prematurely. Usually it happens for a very long jobs run on relatively small number of nodes. Many applications can be restarted and continued from the stopping point via a link.



Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



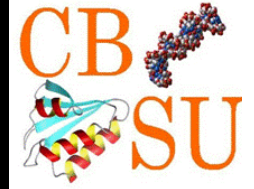
The screenshot shows a web browser window displaying the 'General Administration' page of the Computational Biology Service Unit. The page has a light blue header with the CB SU logo and the title 'Computational Biology Service Unit Web Computing Resources'. Below the header, there is a navigation bar with links for 'User: jp86@cornell.edu', 'Apps Home', 'FAQ', 'Logout', 'Change password', 'My jobs', 'Manage jobs', and 'Administration'. A message asks users to help keep the site free by taking a survey. The main content area is divided into two sections: 'APPLICATIONS' on the left and 'ADMINISTRATIVE LINKS' on the right. The 'APPLICATIONS' section includes buttons for 'Show all', 'Hide all', 'Sequence analysis', 'Sequence alignment', 'Population genetics', 'Protein structure', 'MSR Biomedical', 'Other', and 'Links'. The 'ADMINISTRATIVE LINKS' section includes buttons for 'Manage jobs', 'Manage users', 'Manage applications', 'Manage clusters', 'Fix job', 'E-mail check', 'Remove old jobs data', 'Remove old JSDL jobs data', 'Manage BLAST databases', 'Post message', 'Bulk mail', and 'Site Setup'.

Links to administration pages.



Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



Job Monitoring - Windows Internet Explorer
http://cbsuapps.tc.cornell.edu/checkinfo.aspx

Computational Biology Service Unit
Web Computing Resources
(compute nodes status)

User: jp86@cornell.edu | [Apps Home](#) | [FAQ](#) | [Logout](#) | [Change password](#) | [My jobs](#) | [Manage jobs](#) | [Administration](#)

Please help us keep this site free - take a [survey](#)

Job admin

On this page: total 198 jobs, 9 queued, 133 running, 41 finished, 9 failed, 6 other
cache last updated between 10/28/2009 4:01:00 PM and 10/28/2009 4:14:40 PM cache hits 99.3% over last 1000 calls

Display Display & update up to 600 jobs (300 per page). Last refreshed: 10/28/2009 4:19:58 PM Autorefresh in 14min

S	No	ID	Appname	Submitted	User	Cluster	Nodes/ (CPU)	Cluster jobid	Timeout	Started	Ending	Job name	Status	Active?	Files	Action
		>0		after 12/31/2004 before <999999	*	*	>0	*	>0	after 12/31/2004 before 12/31/2010	12/31/2004 12/31/2010	*	ANY	Yes		
	1	102826	MrBayes	10/28/2009 4:09:57 PM	dlanner@biology.usu.edu	Athena (Auto)	(4)	16088	10080	10/28/2009 4:09:57 PM	10/28/2009 4:16:52 PM	odonnell_its1	FINISHED CORRECTLY	Yes	input: show/htmlfile output: show/htmlfile log file: show DIR	ARCHIVE DEL RESTART NOTIFY: S O I E P STAT: SC FL EV PM
	2	102825	MrBayes	10/28/2009 4:00:59 PM	chris.veston@ucsc.edu	biosim (Auto)	(4)	24920	10080	10/28/2009 4:02:10 PM	11/4/2009 4:02:10 PM	nemertesiaalignmentmega_tmb_mim_constrained	RUNNING	Yes	input: show/htmlfile output: show/htmlfile log file: show DIR	STOP RESTART NOTIFY: S O I E P STAT: SC FL EV PM
	3	102824	InStruct	10/28/2009 3:48:47 PM	alvarosolopezarzon@gmail.com	Athena (Auto)	(1)	16087	7200	10/28/2009 3:51:50 PM	11/2/2009 3:51:50 PM	Silvestres_MATRIX_INSTRUCT.txt	RUNNING	Yes	input: show/htmlfile output: show/htmlfile log file: show DIR	STOP RESTART NOTIFY: S O I E P STAT: SC FL EV PM
	4	102823	BEAST	10/28/2009 2:57:02 PM	chuan2@unlv.nevada.edu	obsum2 (obsum2)	(2)	4016	7200			Flooides_Run_4	QUEUED	Yes	input: show/htmlfile output: show/htmlfile log file: show DIR	STOP RESTART NOTIFY: S O I E P STAT: SC FL EV PM
	5	102822	P-CLUSTALW	10/28/2009 2:49:51 PM	bukowski@ccc.cornell.edu	obsum2k8 (obsum2k8)	(1)	1623	4320	10/28/2009 2:49:56 PM	10/28/2009 2:50:04 PM	clustalw_job	FINISHED CORRECTLY	Yes	input: show/htmlfile output: show/htmlfile log file: show DIR	ARCHIVE DEL RESTART NOTIFY: S O I E P STAT: SC FL EV PM
	6	102821	MrBayes	10/28/2009 2:44:38 PM	wsvape@ucdavis.edu	biosim (Auto)	(4)	24919	14400	10/28/2009 2:47:10 PM	11/7/2009 2:47:10 PM	adelphe_efa	RUNNING	Yes	input: show/htmlfile output: show/htmlfile log file: show DIR	STOP RESTART NOTIFY: S O I E P STAT: SC FL EV PM
	7	102820	MrBayes	10/28/2009 2:44:28 PM	wsvape@ucdavis.edu	biosim (Auto)	(4)	24918	14400	10/28/2009 2:47:09 PM	11/7/2009 2:47:09 PM	adelphe_cool	RUNNING	Yes	input: show/htmlfile output: show/htmlfile log file: show DIR	STOP RESTART NOTIFY: S O I E P STAT: SC FL EV PM
	8	102819	P-CLUSTALW	10/28/2009 2:22:45 PM	bukowski@ccc.cornell.edu	obsumv05 (Auto)	(2)	886	180	10/28/2009 2:22:50 PM	10/28/2009 2:22:52 PM	clustalw_job	FINISHED CORRECTLY	Yes	input: show/htmlfile output: show/htmlfile log file: show DIR	ARCHIVE DEL RESTART NOTIFY: S O I E P STAT: SC FL EV PM
	9	102818	IMa	10/28/2009 2:22:26 PM	forister@gmail.com	obsum1 (Auto)	(1)	16239	4320	10/28/2009 2:22:51 PM	10/31/2009 2:22:51 PM	kbbIM_3_H_numfixed_4.txt	RUNNING	Yes	input: show/htmlfile output: show/htmlfile log file: show DIR	STOP RESTART NOTIFY: S O I E P STAT: SC FL EV PM
	10	102817	IMa	10/28/2009 2:13:17 PM	forister@gmail.com	obsum1 (Auto)	(1)	16238	4320	10/28/2009 2:13:42 PM	10/28/2009 2:18:06 PM	kbbIM_3_H_numfixed_4.txt	FINISHED CORRECTLY	Yes	input: show/htmlfile output: show/htmlfile log file: show DIR	ARCHIVE DEL RESTART NOTIFY: S O I E P STAT: SC FL EV PM
	11	102816	MrBayes	10/28/2009 2:00:46 PM	wsvape@ucdavis.edu	obsum1 (Auto)	(4)	16237	14400	10/28/2009 2:01:10 PM	11/7/2009 2:01:10 PM	adelphe_by_gene	RUNNING	Yes	input: show/htmlfile output: show/htmlfile log file: show DIR	STOP RESTART NOTIFY: S O I E P STAT: SC FL EV PM
	12	102815	STRUCTURE	10/28/2009 1:59:01 PM	gc2mudrim@uoc.es	obsum2k8 (Auto)	(1)	1622	1440	10/28/2009 2:10:49 PM	10/28/2009 2:10:49 PM	LOC.str	RUNNING	Yes	input: show/htmlfile output: show/htmlfile log file: show DIR	STOP RESTART NOTIFY: S O I E P STAT: SC FL EV PM
	13	102814	IM	10/28/2009 1:57:31 PM	felipemartins@hotmail.com	obsum1 (Auto)	(1)	16236	7200	10/28/2009 1:57:54 PM	11/2/2009 1:57:54 PM	desmodulM.txt	RUNNING	Yes	input: show/htmlfile output: show/htmlfile log file: show DIR	STOP RESTART NOTIFY: S O I E P STAT: SC FL EV PM
	14	102813	IM	10/28/2009 1:56:27 PM	felipemartins@hotmail.com	obsum1 (Auto)	(1)	16235	7200	10/28/2009 1:56:51 PM	11/2/2009 1:56:51 PM	desmodulM.txt	RUNNING	Yes	input: show/htmlfile output: show/htmlfile log file: show DIR	STOP RESTART NOTIFY: S O I E P STAT: SC FL EV PM

Version 1.9au.228

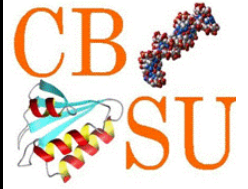
Job administration.



Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



BioHPC as a web service

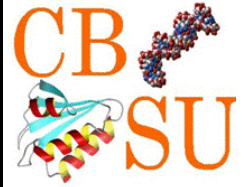
- Convenient user interfaces other than web forms, e.g. Excel
 - Job submission with immediate results visualization / analysis
- Incorporate HPC applications in automated analysis pipelines
 - Especially important in the context of Next Generation Sequencing pipelines
- BioHPC resources available through Microsoft Biology Foundation for command-line utilization



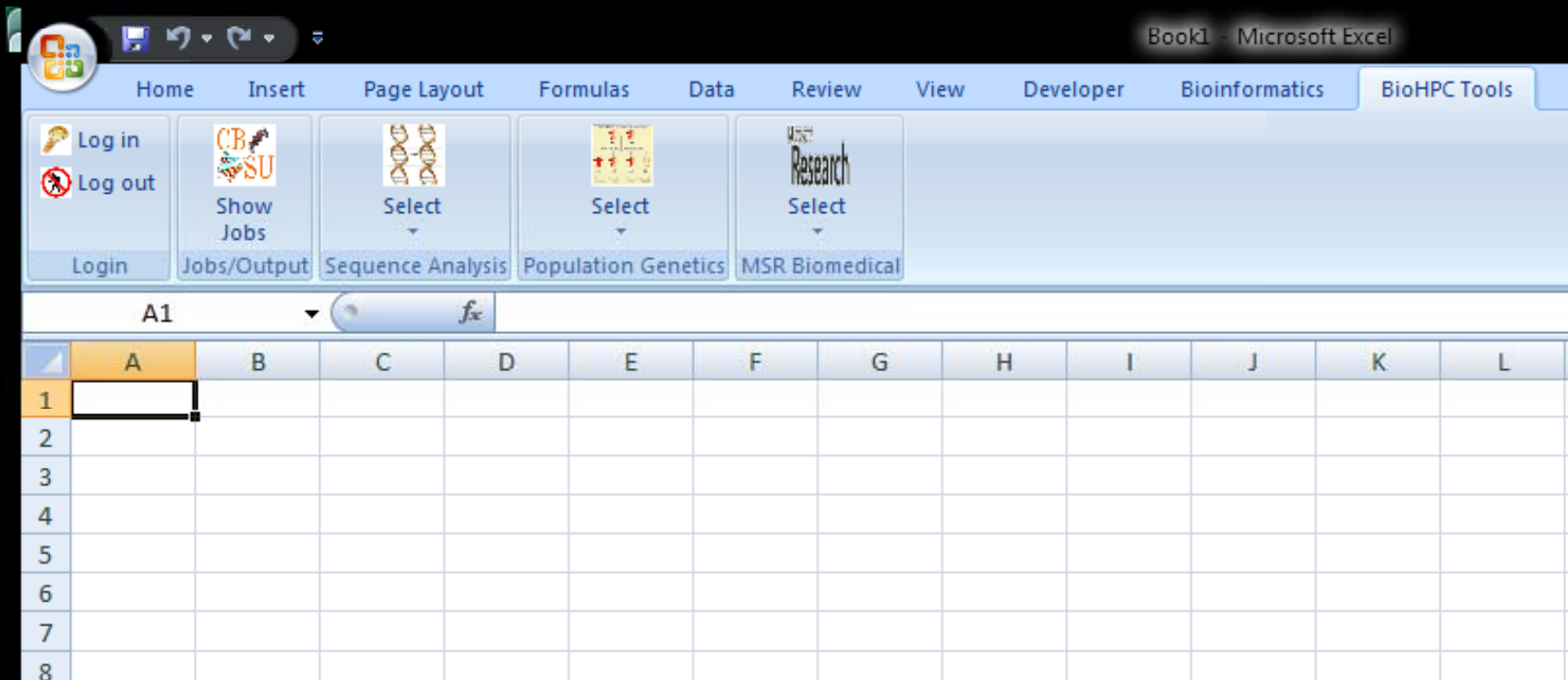
Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



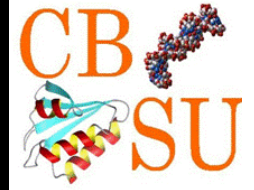
Web service application example:
BioHPC Excel add-in





Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



Book1 - Microsoft Excel

Home Insert Page Layout Formulas Data Review View Developer Bioinformatics BioHPC Tools HPC Services

Log in Log out Show Jobs Select Select

Login Jobs/Output Sequence Analysis Population Genetics MSR Biomedical

K22

SubmitMdiv

Welcome to MDIV submission form

Job Name:
my_MDIV_job

Input File:
C:\Users\vb299\Documents\tst\MDIV_websvc\infileC

Model:
Infinite Sites Model

Length of Markov Chain Max Scaled Migration Rate
2000000 10

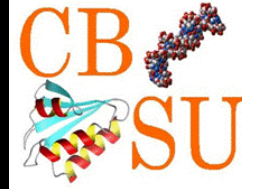
Burn-in Steps Max Scaled Divergence Time
500000 5

Max Theta
(0 for automatic initialization) Random Number Seed
0 123456

Sheet1 Sheet2 Sheet3

Ready 100% 1:37 PM

Computational Biology Applications Suite for High Performance Computing (BioHPC)



Book1 - Microsoft Excel

Home Insert Page Layout Formulas Data Review View Developer Bioinformatics BioHPC Tools HPC Services

Log in Log out Show Jobs Select Select Select
Login Jobs/Output Sequence Analysis Population Genetics MSR Biomedical

A1

My Jobs

BioHPC Jobs Summary

Show 10 most recent jobs running application MDIV

submitted after Saturday, January 01, 2000 and before Tuesday, October 13, 2009

Show/Refresh Jobs

Select job from list below:

Job ID	Job Name	Status	Submitted	Started	Ending/Ended	Timeout
93910	mdiv_job	FINISHED	8/20/2009 11:0...	8/20/2009 11:1...	8/20/2009 11:1...	1440
88951	mdiv_job	FINISHED	7/7/2009 6:05:...	7/7/2009 6:05:...	7/9/2009 4:25:...	1440
88950	mdiv_job	MAINTENANCE	7/7/2009 5:39:...	7/7/2009 5:39:...	7/9/2009 4:25:...	1440
88948	mdiv_job	CANCELED	7/7/2009 5:10:...	7/7/2009 5:34:...	7/7/2009 5:39:...	1440
81650	BBBB	CANCELED	5/11/2009 6:19:...	5/11/2009 6:19:...	5/11/2009 6:20:...	7200
81648	mdiv_job	CANCELED	5/11/2009 6:03:...	5/11/2009 6:03:...	5/11/2009 6:30:...	7200
81647	AAAAA	CANCELED	5/11/2009 6:00:...	5/11/2009 6:00:...	5/11/2009 6:07:...	7200

Show output files for selected job Import output into Excel Cancel selected job Download Selected Files

Select files from list below:

File Name	Size
infile.txt	761
out	889
outfile.txt	28546

Messages:
Selected job: 93910
Control number: 846700561
Application: MDIV

Exit

Browse For Folder

Select Output Directory

- SQL Server Management Studio Express
- tst
 - Rin_NFT_tst
 - BLAST_websvc
 - Epitome_websvc
 - heattest
 - HMMER_websvc
 - IPRSCAN_websvc
 - MDIV_websvc
- SFS_tst
- squashed

Make New Folder OK Cancel

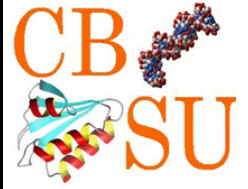
Sheet1 Sheet2 Sheet3

Ready

100%

1:31 PM

Computational Biology Applications Suite for High Performance Computing (BioHPC)



Book1 - Microsoft Excel

Home Insert Page Layout Formulas Data Review View Developer Bioinformatics BioHPC Tools HPC Services

Clipboard Font Alignment Number Styles Cells Editing

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AE
1	M		P(M)	T																						
2		0.02	0.000048	0.01	0.002057																					
3		0.04	0.00021	0.02	0.001818																					
4		0.06	0.000115	0.03	0.001757																					
5		0.08	0.000243	0.04	0.003302																					
6		0.1	0.000253	0.05	0.001808																					
7		0.12	0.000133	0.06	0.002577																					
8		0.14	0.000262	0.07	0.00228																					
9		0.16	0.00022	0.08	0.002302																					
10		0.18	0.00034	0.09	0.001948																					
11		0.2	0.000258	0.1	0.00201																					
12		0.22	0.0004	0.11	0.002195																					
13		0.24	0.000393	0.12	0.002098																					
14		0.26	0.000345	0.13	0.002053																					
15		0.28	0.000407	0.14	0.002175																					
16		0.3	0.000467	0.15	0.002045																					
17		0.32	0.000433	0.16	0.002155																					
18		0.34	0.000395	0.17	0.00212																					
19		0.36	0.000483	0.18	0.001868																					
20		0.38	0.000515	0.19	0.002065																					
21		0.4	0.00044	0.2	0.002603																					
22		0.42	0.00071	0.21	0.002585																					
23		0.44	0.000562	0.22	0.001835																					
24		0.46	0.000645	0.23	0.00195																					
25		0.48	0.00055	0.24	0.002253																					
26		0.5	0.000673	0.25	0.002188																					
27		0.52	0.000543	0.26	0.00192																					
28		0.54	0.000652	0.27	0.002265																					
29		0.56	0.000737	0.28	0.002087																					
30		0.58	0.00078	0.29	0.001928																					
31		0.6	0.000678	0.3	0.001845																					
32		0.62	0.000855	0.31	0.002478																					
33		0.64	0.000953	0.32	0.002227																					
34		0.66	0.000963	0.33	0.001915																					
35		0.68	0.0009	0.34	0.002067																					
36		0.7	0.000785	0.35	0.002475																					
37		0.72	0.000825	0.36	0.002215																					
38		0.74	0.000888	0.37	0.002023																					
39		0.76	0.00095	0.38	0.001958																					
40		0.78	0.000925	0.39	0.002413																					

E[TMRCR | data]: 2.697569

BioHPC Jobs Summary

Show 10 most recent jobs running application MDIV

submitted after Saturday, January 01, 2000 and before Tuesday, October 13, 2009

Show/Refresh Jobs

Select job from list below:

Job ID	Job Name	Status	Submitted	Started	Ending/Ended	Timeout
93910	mdiv_job	FINISHED	8/20/2009 11:0...	8/20/2009 11:1...	8/20/2009 11:1...	1440
88951	mdiv_job	FINISHED	7/7/2009 6:05:...	7/7/2009 6:05:...	7/9/2009 4:25:...	1440
88950	mdiv_job	MAINTENANCE	7/7/2009 5:39:...	7/7/2009 5:39:...	7/9/2009 4:25:...	1440
88948	mdiv_job	CANCELED	7/7/2009 5:10:...	7/7/2009 5:34:...	7/7/2009 5:39:...	1440
81650	BBBB	CANCELED	5/11/2009 6:19:...	5/11/2009 6:19:...	5/11/2009 6:20:...	7200
81648	mdiv_job	CANCELED	5/11/2009 6:03:...	5/11/2009 6:03:...	5/11/2009 6:30:...	7200
81647	AAAAA	CANCELED	5/11/2009 6:00:...	5/11/2009 6:00:...	5/11/2009 6:07:...	7200

Show output files for selected job Import output into Excel Cancel selected job

Messages:
Selected job: 93910
Control number: 846700561
Application: MDIV

Exit



Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



BioHPC BLAST web service in Microsoft Biology Foundation client

The screenshot shows the Microsoft Excel interface with the Bioinformatics ribbon selected. The ribbon includes tabs for Home, Insert, Page Layout, Formulas, Data, Review, View, Developer, Bioinformatics, BioHPC Tools, and HPC Services. The Bioinformatics ribbon contains buttons for 'Select BLAST service', 'Charts', 'Operate on BED files', '3D Molecular Viewer', 'Configure', 'Microsoft Biology Foundation', and 'Specify bounds'. A dropdown menu is open under 'Select BLAST service', listing four options: 'Search NCBI QBLAST database', 'Search EBI WU-BLAST database', 'Search Azure BLAST database', and 'Search BioHPC BLAST database'. The 'Search BioHPC BLAST database' option is highlighted. Below the menu, an 'ExcelWorkbench' tooltip is visible with the text 'Press F1 for more help.' The spreadsheet area shows a sequence alignment table with columns labeled A through CM and rows containing amino acid sequences. The status bar at the bottom indicates 'Count: 247' and '70%' zoom.



Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



BioHPC BLAST web service in Microsoft Biology Foundation client

The screenshot shows a Microsoft Excel window titled "Book1 - Microsoft Excel" with a custom ribbon for Bioinformatics. The ribbon includes tabs for Home, Insert, Page Layout, Formulas, Data, Review, View, Developer, Bioinformatics, BioHPC Tools, and HPC Services. The Bioinformatics tab is active, showing options like Import From, Export To, Select Aligners, Select BLAST service, Charts, Operate on BED files, 3D Molecular Viewer, Configure, Microsoft Biology Foundation, and Specify bounds. The BioHPC Tools tab shows a "ReportTools" button. The HPC Services tab is currently empty. The spreadsheet contains a sequence of amino acids in row 3, starting with "Sequence Data" in cell A3. A dialog box titled "BLAST WebService" is open, prompting for general parameters. The dialog has three columns: Program, Database, and MatrixName. The Program is set to "blastp", the Database is "arabidopsis_genovnr", and the MatrixName is "BLOSUM62". Other parameters include Filter: "T", Alignments: "500", Database: "arabidopsis_genovnr", Email: "bukowski@tc.cornell.edu", Expect: "10.0", Password: "*****", and MinQueryLength: "15". Buttons for "Select sequences and submit" and "Cancel" are at the bottom of the dialog.

Program	Filter	Alignments
blastp	T	500

Database	Email	Expect
arabidopsis_genovnr	bukowski@tc.cornell.edu	10.0

Database	Password	MinQueryLength
arabidopsis_genovnr	*****	15

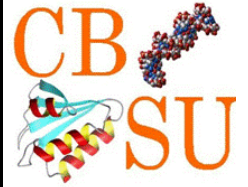
Database	MatrixName
arabidopsis_genovnr	BLOSUM62



Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



Next Generation Sequencing and BioHPC

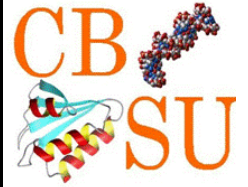
There are a few LIMS-like systems available for next generation sequencing, but none has HPC bio computing implemented and none uses Windows



Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



Next Generation Sequencing @ BioHPC

- Data management

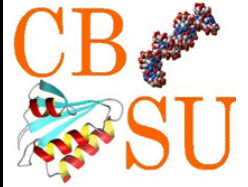
Run Manager: connects to the sequencing facility and automatically detects finished sequencing runs for which base calling has been completed. It then configures the run in BioHPC database and sends an invitation to the facility manager to approve the results for distribution to users. Once approved, the results (read files) are asynchronously transferred to BioHPC file server and catalogued there for further use. Once the transfer is complete, all users assigned to distributed lanes are automatically notified by an e-mail message containing download links.

Lane Browser: allows users to browse their sequencing read files (Illumina lanes) catalogued at BioHPC. The browser displays lane annotation information and allows the file owner to grant additional users access to a file. Read files obtained outside of the Cornell sequencing facility can also be uploaded and catalogued at BioHPC.



Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



Next Generation Sequencing @ BioHPC

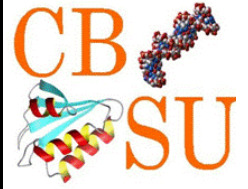
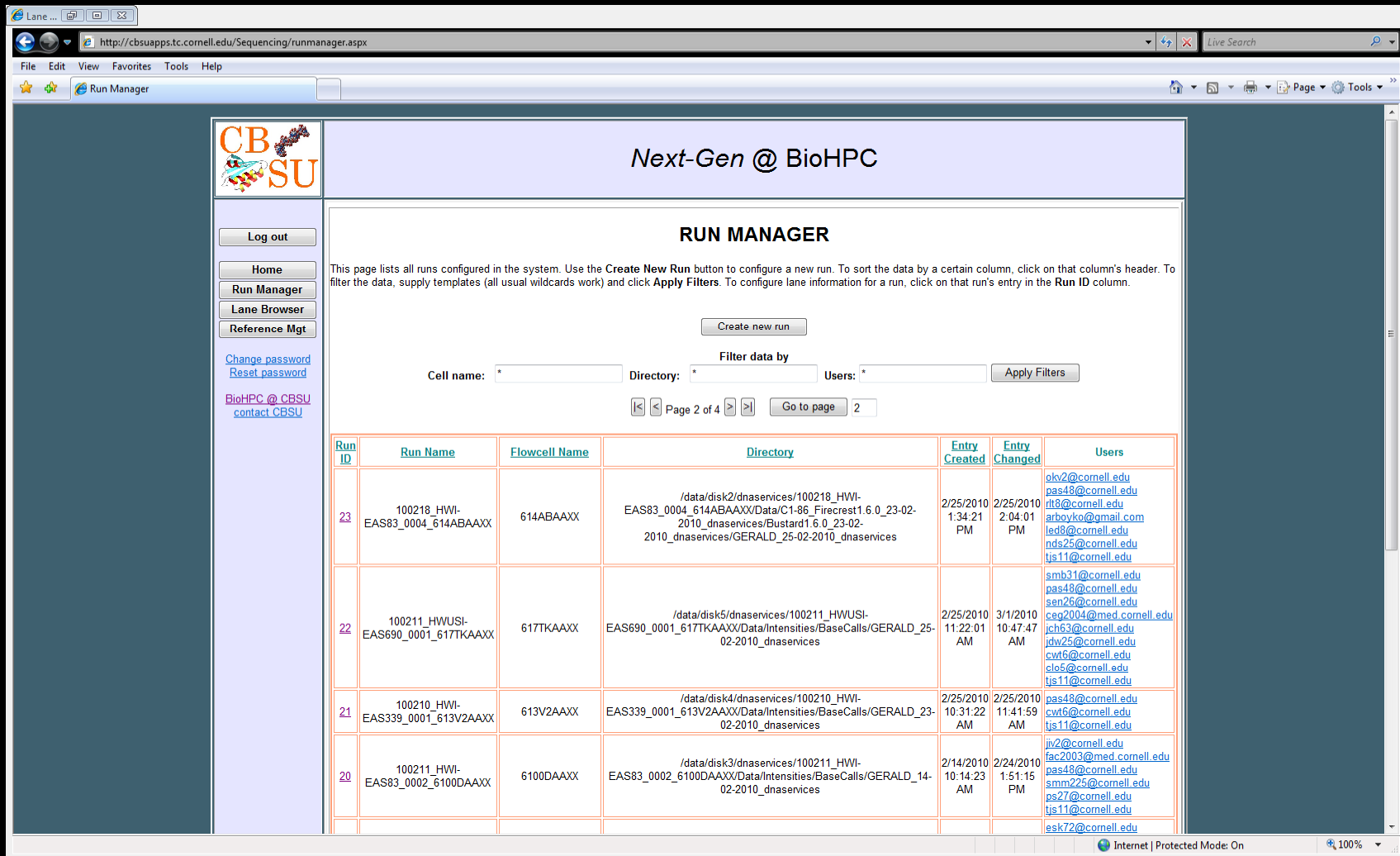
- Data analysis

Reference Manager: allows users to upload and catalogue reference genome files and annotation files needed in downstream data analysis.

Pipeline Manager (under development): allows users to construct and run various analysis pipelines using sequencing reads and reference files stored at BioHPC as input. While default parameters are provided, steps of each pipeline will be individually configurable by a user. Users interface with pipeline manager using our specially constructed web interface or using a web service layer. Computationally intensive steps run on clusters linked to BioHPC.

The web service interface will allow pipelines to be controlled from any client application, such as the **MBF platform** or the **Illumina Genome Studio**, or **Trident** scientific workflow workbench.

Computational Biology Applications Suite for High Performance Computing (BioHPC)

The screenshot shows a web browser window displaying the 'Next-Gen @ BioHPC' Run Manager interface. The page title is 'Next-Gen @ BioHPC' and the main heading is 'RUN MANAGER'. A sidebar on the left contains navigation links: 'Log out', 'Home', 'Run Manager', 'Lane Browser', and 'Reference Mgt'. Below these are links for 'Change password', 'Reset password', and 'BioHPC @ CBSU contact CBSU'. The main content area includes a 'Create new run' button, a 'Filter data by' section with input fields for 'Cell name', 'Directory', and 'Users', and an 'Apply Filters' button. Below the filters is a table listing sequencing runs. The table has columns for 'Run ID', 'Run Name', 'Flowcell Name', 'Directory', 'Entry Created', 'Entry Changed', and 'Users'. The table contains four rows of data, each representing a different sequencing run with its associated details and user list.

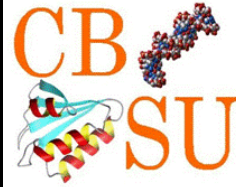
Run ID	Run Name	Flowcell Name	Directory	Entry Created	Entry Changed	Users
23	100218_HWI-EAS83_0004_614ABAAXX	614ABAAXX	/data/disk2/dnaservices/100218_HWI-EAS83_0004_614ABAAXX/Data/C1-86_Firecrest1.6.0_23-02-2010_dnaservices/Bustard1.6.0_23-02-2010_dnaservices/GERALD_25-02-2010_dnaservices	2/25/2010 1:34:21 PM	2/25/2010 2:04:01 PM	okv2@cornell.edu pas48@cornell.edu rt8@cornell.edu arbovko@gmail.com led8@cornell.edu nds25@cornell.edu tjs11@cornell.edu
22	100211_HWUSI-EAS690_0001_617TKAAXX	617TKAAXX	/data/disk5/dnaservices/100211_HWUSI-EAS690_0001_617TKAAXX/Data/Intensities/BaseCalls/GERALD_25-02-2010_dnaservices	2/25/2010 11:22:01 AM	3/1/2010 10:47:47 AM	smb31@cornell.edu pas48@cornell.edu sen26@cornell.edu ceg2004@med.cornell.edu jch63@cornell.edu jdw25@cornell.edu cwt6@cornell.edu clo5@cornell.edu tjs11@cornell.edu
21	100210_HWI-EAS339_0001_613V2AAXX	613V2AAXX	/data/disk4/dnaservices/100210_HWI-EAS339_0001_613V2AAXX/Data/Intensities/BaseCalls/GERALD_23-02-2010_dnaservices	2/25/2010 10:31:22 AM	2/25/2010 11:41:59 AM	pas48@cornell.edu cwt6@cornell.edu tjs11@cornell.edu
20	100211_HWI-EAS83_0002_6100DAAXX	6100DAAXX	/data/disk3/dnaservices/100211_HWI-EAS83_0002_6100DAAXX/Data/Intensities/BaseCalls/GERALD_14-02-2010_dnaservices	2/14/2010 10:14:23 AM	2/24/2010 1:51:15 PM	jiv2@cornell.edu fac2003@med.cornell.edu pas48@cornell.edu srm225@cornell.edu ps27@cornell.edu tjs11@cornell.edu ask72@cornell.edu

Intercept finished sequencing runs and configure them in BioHPC data manager.



Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



New Illumina run 100312_HWI-EAS339_0008_61GJ3AAXX configured with RunID 33 - Message (HTML)

Message Developer

Reply Reply Forward Delete Move to Create Other Block Safe Lists Categorize Follow Mark as Send to
to All to All Folder Rule Actions Sender Lists Up Unread OneNote
Respond Actions Junk E-mail Options Find Find OneNote

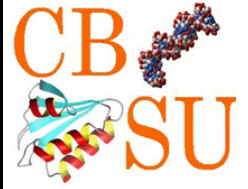
From: biohpc@cac.cornell.edu Sent: Wed 3/17/2010 9:42 AM
To: James Van Ee; pas48@cornell.edu; tjs11@cornell.edu; Robert Bukowski; Robert Bukowski
Cc:
Subject: New Illumina run 100312_HWI-EAS339_0008_61GJ3AAXX configured with RunID 33

New Illumina run 100312_HWI-EAS339_0008_61GJ3AAXX has been configured with BioHPC RunID 33.

[Verify all information and schedule the run for distribution](#)

Notify sequencing facility administrators about the new results to be approved for distribution to users.

Computational Biology Applications Suite for High Performance Computing (BioHPC)



EDIT RUN 33

Extract flowcell name and machine run name from directory string

Run ID	Run Name	Flowcell Name	Control Lane	Entry Created	Entry Changed
33	100312_HWI-EAS339_0008_61GJ3AAXX	61GJ3AAXX	8	3/17/2010 9:41:47 AM	3/18/2010 9:28:52 AM

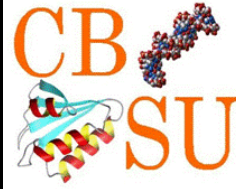
Location: /data/disk3/dnaservices/100312_HWI-EAS339_0008_61GJ3AAXX/Data/Intensities/BaseCalls/GERALD_17-03-2010_dnaservices

Lane info:

Lane	Type	Sample Name	Prefilter	Status	Users (check box to remove upon submission)	Lab	Order#	Clear upon submission
1	Standard	Par	No	ready	<input type="checkbox"/> jr286@cornell.edu set as owner --Add user from list-- Add users by e-mail:	N/A	10214774	<input type="checkbox"/>
2	Standard	Var	No	ready	<input type="checkbox"/> jr286@cornell.edu set as owner --Add user from list-- Add users by e-mail:	N/A	10214774	<input type="checkbox"/>
3	Standard	End	No	ready	<input type="checkbox"/> jr286@cornell.edu set as owner --Add user from list-- Add users by e-mail:	N/A	10214774	<input type="checkbox"/>
4	Standard	V-1	No	ready	<input type="checkbox"/> vnm1@cornell.edu set as owner --Add user from list-- Add users by e-mail:	N/A	10216342	<input type="checkbox"/>
5	Standard	mel2 sim2 yak2 ana2	No	processing(zip)	<input type="checkbox"/> j12434@cornell.edu set as owner --Add user from list-- Add users by e-mail:	N/A	10216484	<input type="checkbox"/>
6	Standard	Pooled samples	No	approved	<input type="checkbox"/> samantha.brooks@cornell.edu set as owner --Add user from list-- Add users by e-mail:	N/A	10216429	<input type="checkbox"/>
7	Standard	Pooled samples	No	approved	<input type="checkbox"/> ac347@cornell.edu set as owner --Add user from list-- Add users by e-mail:	N/A	10216609	<input type="checkbox"/>

Approval page for sequencing facility. Transfer (asynchronous) to BioHPC will start after a lane status is changed to approved.

Computational Biology Applications Suite for High Performance Computing (BioHPC)



http://cbsuapps.tc.cornell.edu/Sequencing/lanebrowser.aspx

File Edit View Favorites Tools Help

Lane Browser

BROWSE ALL LANES

This is a list of all lanes configured in the system. To sort results, click on a column header. To filter results, supply templates (all usual wildcards apply) and click **Apply Filters**. Clicking on an entry in **Run ID** column will open the run manager utility, where the lane information can be configured. To download files for a lane, click on the link **(files)** underneath the LaneID. For assistance with download of multiple files in batch mode, use the **Make download script** button.

[Log out](#)

[Home](#)
[Run Manager](#)
[Lane Browser](#)
[Reference Mgt](#)

[Change password](#)
[Reset password](#)

[BioHPC @ CBSU](#)
[contact CBSU](#)

[Make download script](#) [Register new lane](#)

("check" all lanes you want to include in a download script and click button above) (use this option only if you want to manually upload a lane from outside of the CLC sequencing facility)

Filter data by
Status: Sample name: Users: [Apply Filters](#)

Page 1 of 10 [Go to page](#)

Lane ID	Run ID	Run Name	Lane#	Type	Sample Name	Status	Annotations	Users	Lab	Order#												
191 (files)	33	100312_HWI-EAS339_0008_61GJ3AAXX	1	Standard	Par	ready	<table border="1"> <tr> <td>Parameter</td> <td>This Lane</td> <td>Ctrl Lane</td> </tr> <tr> <td>Length</td> <td>unknown</td> <td>unknown</td> </tr> <tr> <td>Clusters_raw</td> <td>18.8M</td> <td>27.6M</td> </tr> <tr> <td>Clusters_PF</td> <td>11.9M</td> <td>16.8M</td> </tr> </table>	Parameter	This Lane	Ctrl Lane	Length	unknown	unknown	Clusters_raw	18.8M	27.6M	Clusters_PF	11.9M	16.8M	jr286@cornell.edu (owner)	N/A	10214774
Parameter	This Lane	Ctrl Lane																				
Length	unknown	unknown																				
Clusters_raw	18.8M	27.6M																				
Clusters_PF	11.9M	16.8M																				
192 (files)	33	100312_HWI-EAS339_0008_61GJ3AAXX	2	Standard	Var	ready	<table border="1"> <tr> <td>Parameter</td> <td>This Lane</td> <td>Ctrl Lane</td> </tr> <tr> <td>Length</td> <td>unknown</td> <td>unknown</td> </tr> <tr> <td>Clusters_raw</td> <td>18.7M</td> <td>27.6M</td> </tr> <tr> <td>Clusters_PF</td> <td>14.4M</td> <td>16.8M</td> </tr> </table>	Parameter	This Lane	Ctrl Lane	Length	unknown	unknown	Clusters_raw	18.7M	27.6M	Clusters_PF	14.4M	16.8M	jr286@cornell.edu (owner)	N/A	10214774
Parameter	This Lane	Ctrl Lane																				
Length	unknown	unknown																				
Clusters_raw	18.7M	27.6M																				
Clusters_PF	14.4M	16.8M																				
193 (files)	33	100312_HWI-EAS339_0008_61GJ3AAXX	3	Standard	End	ready	<table border="1"> <tr> <td>Parameter</td> <td>This Lane</td> <td>Ctrl Lane</td> </tr> <tr> <td>Length</td> <td>unknown</td> <td>unknown</td> </tr> <tr> <td>Clusters_raw</td> <td>25.2M</td> <td>27.6M</td> </tr> <tr> <td>Clusters_PF</td> <td>20.0M</td> <td>16.8M</td> </tr> </table>	Parameter	This Lane	Ctrl Lane	Length	unknown	unknown	Clusters_raw	25.2M	27.6M	Clusters_PF	20.0M	16.8M	jr286@cornell.edu (owner)	N/A	10214774
Parameter	This Lane	Ctrl Lane																				
Length	unknown	unknown																				
Clusters_raw	25.2M	27.6M																				
Clusters_PF	20.0M	16.8M																				
194	33	100312_HWI-EAS339_0008_61GJ3AAXX	4	Standard	V-1	processing (transit)	--- NO ANNOTATIONS ---	ynm1@cornell.edu (owner)	N/A	10216342												
195	33	100312_HWI-EAS339_0008_61GJ3AAXX	5	Standard	meI2 sim2 yak2 ana2	processing (zip)	--- NO ANNOTATIONS ---	jl2434@cornell.edu (owner)	N/A	10216484												
196	33	100312_HWI-EAS339_0008_61GJ3AAXX	6	Standard	Pooled samples	approved	--- NO ANNOTATIONS ---	samantha.brooks@cornell.edu (owner)	N/A	10216429												
197	33	100312_HWI-EAS339_0008_61GJ3AAXX	7	Standard	Pooled samples	approved	--- NO ANNOTATIONS ---	ac347@cornell.edu (owner)	N/A	10216609												
198	33	100312_HWI-EAS339_0008_61GJ3AAXX	8	PhiX	PhiX	pending	--- NO ANNOTATIONS ---	tjs11@cornell.edu (owner)	N/A	NA												
183	32	100312_HWI-EAS339_0008_61GJ3AAXX	1	Standard	Par	pending	--- NO ANNOTATIONS ---	jr286@cornell.edu (owner)	N/A	10214774												
184	32	100312_HWI-EAS339_0008_61GJ3AAXX	2	Standard	Var	pending	--- NO ANNOTATIONS ---	jr286@cornell.edu (owner)	N/A	10214774												

http://cbsuapps.tc.cornell.edu/resetpass.aspx

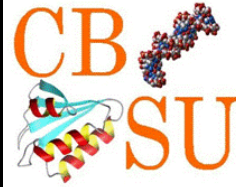
Internet | Protected Mode: On 100%

Main administration page for lanes. Users can only manage their own data.



Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



[BioHPC]: CLC Illumina Genome Analyzer Results: 100312_HWI-EAS339_0008_61GJ3AAXX / Order: 10214774 / Sample: Par

Message Developer

From: biohpc@cac.cornell.edu
To: jr286@cornell.edu
Cc: James Van Ee; pas48@cornell.edu; tjs11@cornell.edu; Robert Bukowski
Subject: [BioHPC]: CLC Illumina Genome Analyzer Results: 100312_HWI-EAS339_0008_61GJ3AAXX / Order: 10214774 / Sample: Par

Sent: Thu 3/18/2010 9:45 AM

The sequencing result for sample "Par" (lane 1 from run 100312_HWI-EAS339_0008_61GJ3AAXX) has been registered with the CBSU file manager as Lane 191. The files are ready for pickup using the following link:

<http://cbsuapps.tc.cornell.edu/Sequencing/showseqfile.aspx?mode=outlist&cntrl=1758691306&laneid=191>

This and other sequencing results may also be retrieved using the following methods:

- From the Order Status/Results section of the CLC web site (<https://cores.lifesciences.cornell.edu/userdev>)
- From the BioHPC Next Generation Data Analysis site at Computational Biology Service Unit (CBSU) (<http://cbsuapps.tc.cornell.edu/Sequencing/seqmain.aspx>). To learn how to access the BioHPC site please see the explanation below.

About BioHPC:
The BioHPC Next Generation Sequencing Data Analysis site allows users to access the results of Illumina sequencing runs performed on their behalf by the Sequencing Facility at Cornell CLC. Read files obtained outside of this facility can also be catalogued and stored here. In the future, users will also be able to upload and manage reference genome and annotation files and run various analysis pipelines involving read and reference files stored and catalogued at CBSU. To access the site, navigate to

<http://cbsuapps.tc.cornell.edu/Sequencing/seqmain.aspx>

and log in using your e-mail address jr286@cornell.edu as login ID and your BioHPC password. Please note that your account and password at BioHPC are separate from your account at the CLC sequencing facility. If you do not yet know your BioHPC password or if you need to reset it, go to

<http://cbsuapps.tc.cornell.edu/resetpass.aspx?userid=jr286@cornell.edu>

If you would like to use the BioHPC site but are not yet registered, contact CBSU at <http://cbsuapps.tc.cornell.edu/contactus.aspx>.

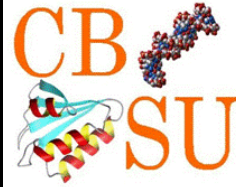
File Formats and Data:
Data are distributed as gzipped "fastq" formatted files generated from the Illumina Genome Analyzer Analysis Pipeline. Unfiltered data in Illumina "qseq" format or other intermediate data from the pipeline and data collection process are available by special arrangement with the CLC (dna_services@cornell.edu). Special data handling may result in additional processing fees.

Data Retention Policy:
Data distributed through the BioHPC web site is available for 30 days (once analysis pipelines become operational, retention time may be extended for lanes involved in calculations). After 30 days, the data may be reposted subject to a service fee. The CLC maintains an archive of all sequencing data, however, we cannot guarantee the availability of data after the initial distribution. Customers are responsible for retrieving and storing their data safely. CLC and BioHPC cannot serve as a backup or archive.

More Information:
Life Science Core Laboratories Center (CLC): <http://cores.lifesciences.cornell.edu>

Once data files are transferred users obtain links to download them.

Computational Biology Applications Suite for High Performance Computing (BioHPC)



showSeqfile - Internet Explorer provided by Dell

http://cbsuapps.tc.cornell.edu/Sequencing/showseqfile.aspx?mode=outlist&cntrl=1251963929&laneid=192

File Edit View Favorites Tools Help

showSeqfile

Sequencing results for sample "Var"

Parameter	Value												
Run Name:	100312_HWI-EAS339_0008_61GJ3AAXX												
Lane#:	2												
Analysis Software:	RTA 1.6.32.0												
Sample Name:	Var												
Lane Annotations:	<table border="1"> <thead> <tr> <th>Parameter</th> <th>This Lane</th> <th>Ctrl Lane</th> </tr> </thead> <tbody> <tr> <td>Length</td> <td>unknown</td> <td>unknown</td> </tr> <tr> <td>Clusters_raw</td> <td>18.7M</td> <td>27.6M</td> </tr> <tr> <td>Clusters_PF</td> <td>14.4M</td> <td>16.8M</td> </tr> </tbody> </table>	Parameter	This Lane	Ctrl Lane	Length	unknown	unknown	Clusters_raw	18.7M	27.6M	Clusters_PF	14.4M	16.8M
	Parameter	This Lane	Ctrl Lane										
	Length	unknown	unknown										
Clusters_raw	18.7M	27.6M											
Clusters_PF	14.4M	16.8M											
Order#:	10214774												
Expiration Date:	4/18/2010												

Files will be available for download until **4/18/2010 (30 days left)**

File (click to download)	Size [bytes]	MD5 sum
10214774_61GJ3AAXX_s_2_sequence.txt.gz	868,121,294	a454bec5ee3eb258c73f0ec023e35a78

Prefer to download multiple files in batch mode?
Use Lane Browser at [BioHPC Next Generation Data Analysis site](#) to generate a download script.

Internet | Protected Mode: On 100%

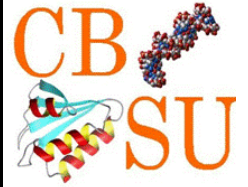
User data download page.



Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



Problems

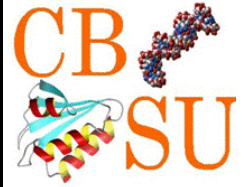
- **HPC basic profile / JSDL limitations in Windows Server 2008 HPC**
Only limited subset of commands and controls is implemented in the native HPC Basic Profile service. Need to develop BioHPC web service to control HPC scheduler.
- **SUA and porting**
Porting applications to Windows environment and convincing original authors to keep them updated on Windows is still a challenge. SUA only supported 32 bit development, severely limiting memory usage – same with Cygwin. Direct native porting not an efficient choice for rapidly changing, not yet established software. Experimenting with MINGW 64 bit environment.



Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



Progress to date summary

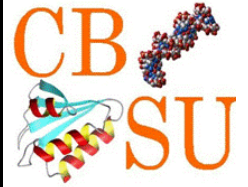
- Fully functional HPC computational biology application suite offering 37 applications, resource integration and management available as open source.
- Very popular service – massively utilized by Cornell and external users
- Web service access implemented for several applications, will be available for all suitable applications by the end of this year
- Integration with MBF via web services and Excel client implementation is in progress for several applications. Extensive participation in MBF development and testing.
- Fully implemented next generation sequencing data manager with asynchronous data transfer from sequencing facility and data distribution to users



Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



Future effort and directions

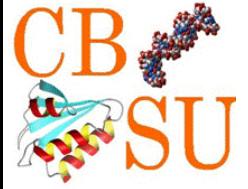
- Keeping up to date with new developments in bio computing: applications and algorithm/software updates, especially in next generation sequencing
- Better integration with MBF, especially regarding data management
- Integrating with commercial applications (Illumina, Real Time Genomics)
- Full implementation of next generation sequencing pipeline manager as web form, web services integrated into MBF, and available as Trident workflows
- Support for Azure cloud – will allow users to install and use BioHPC locally and utilize Azure as a remote HPC resource
- Improved internal maintenance tools (external authentication, better user group management, improved asynchronous data transfer)



Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



Computational Biology Service Unit Cornell core facility for computational biology and bioinformatics

Part of Cornell Life Sciences Core Facilities Center

Provides bioinformatics and computational support for biological research at Cornell and beyond by means of research collaborations, consultations, software development and more.

Genomics/Proteomics

Qi Sun
Lalit Ponnala
Stefan Stefanov

HPC / Computing

Jaroslav Pillardy (director)
Robert Bukowski
Mary Howard

System Biology

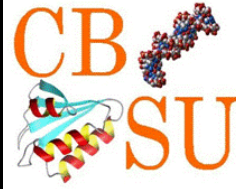
Chris Myers



Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



**Many thanks for Microsoft Research
for support that allowed us to
to develop our own local computing solutions
into a tool that can help others.**

Without MSR BioHPC would be just a set of unorganized
interface and admin tools useful only for us.

