

INTRODUCTION

One of the challenges of High Performance Computing (HPC) is the user accessibility. At the Cornell University Computational Biology Service Unit, which is also a Microsoft HPC institute, we have developed a suit of computational biology applications for HPC (BioHPC) that allows researchers from biological laboratories to submit their jobs to the parallel cluster through an easy-to-use web interface. They don't need to deal with parallel job submission, queues, clusters – knowing the application, parameters and input is all that is required.

TAIR, the major database of the plant model organism Arabidopsis, and ${\rm SGN},$ the international tomato genome database, are both using our system for data analysis

ARCHITECHTURE

The system consists of a **web server** running the interface (ASP.NET C#), **Microsoft SQL server** (ADO.NET), **compute clusters** running Microsoft Windows, **ftp server** and **file server**. Two local compute cluster schedulers are supported (CCS and HPC Server 2008), remote clusters can be used via JSDL/HPC Profile. **JSDL connection** is now implemented and linked to Athena cluster at Microsoft (Redmond, WA) and biosim (Cornell).

Users interact with their jobs and data primarily by a **web browser** and **e-mail**. Jobs are submitted through our active web pages, fully compatible with all the popular web browsers supporting DOM and Javascipt. Notification e-mails sent to users upon job submission, start, and completion contain links for job progress monitoring, job cancellation and restart, and results retrieval (by http or ftp). Job and data control functions can also be performed via a recently developed **web service interface** which enables users to build custom clients independent of the web browser. The number of applications offered through the web service (currently 10) is growing.

Computational Biology Application Suite for High Performance Computing

Jaroslaw Pillardy, Robert Bukowski, Qi Sun, Mary Howard Computational Biology Service Unit, Microsoft HPC Institute, Cornell University, Ithaca, N.Y., 14853

http://BioHPC.net/



Besides the job control interface, BioHPC also features a built in user and data management system which can limit software and/or database access to specified users. Data access between different users is restricted – users can access only their own jobs and data. There is also administrative interface allowing for easy management of jobs, clusters and applications with automatic e-mail notification of possible problems.

APPLICATIONS

Through this system, we are providing users with popular bioinformatics tools covering various aspects of computational biology:

- <u>Data mining / sequence analysis (BLAST, HMMER, InterProScan, GIMSAN, SLIM)</u>
- <u>Protein structure prediction and modeling (LOOPP, Modeller)</u>
- <u>Population genetics</u> (BEAST, BEST, CLUMPP, IM, IMa, InStruct, MDIV, Migrate, MKPRF, MSVAR, OmegaMap, Parentage, SFS_CODE, Structurama, Structure, TESS)
- <u>Phylogenetics</u> (MrBayes, ClustalW, Stretcher, T-COFFEE)
- <u>Association analysis / statistics (PLINK, R)</u>
- <u>MSR Biomedical applications</u> (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. It is also scalable, currently it processes over 40,000 job submissions a year, many of them parallel. BioHPC integrates distributed cluster resources in a user-transparent way.

Ick Monitoreg - 7				a charlenda ana											84×	The line	610
Ferrites @								-		-							- Safety - Tools - 6
	40000		_					_									
CB SU								Co	mput V	Veb	Computing	ogy Ser	vice Unit es				
er ptt@cone	el ete le	loga Ho	CR 154	G Logout Cham	ge. password 1 My p	ita 1 Managar julia 1 Admin	intration 1		Second back		op this site free . t						
									1000			and a second					
APPLICATIO	talin Ni	this pa Display		198 jobs, 9 queu Clear Cache & D		finished, 9 failed, 6 other lighty & update up to						28/2009 4 14 40 P	M cache hts 99.3% over last 1000 calls	-			
equence analy		ha lo	1	CHEF CALLS & D	handhad .	libert	Custer	Notes Jona		of refes	Sec. 10/28/2009 4	TESEPM IC A	Johnson Unlock Admin	The last	Autor?	Time	Autor
Sequence alignment				•	atar 12/31/2004			• -1		14	eter 13:01:3054 Instan	after 12/31/2004 before		anty	• 7m •		
Population genetics		ge 1			12/21/2018			_			12012010	12912010					
rotein structu	_	TT	0838	WEaper	10-26-2009 4-09-87 PM	deconditiona.ala	Athena JAuna	(4)	19944	10000	10:00:0000 4:09:07 PM	10.08.0008 4 16.50 PM	adormali_N ¹	FixesHED COMMECTLY	Vai	input stantinging manuf manipaths ing the class 22	NOTIFY DELLER NOTIFY DELLER STAT SCILLOUTS
ASR Biomodic	cal .	2 19	aka i	hidayee	10/28/2009 4 00 19	to motion and	biosim (Ayb)	141	2+920	10000	10-28-3009 + 82-18 PM	11/42008 4 82 18 PM	hanabalah provinsipation providence	RUNING .	-	HARA EXECUTES INDEAL EXECUTES Har Rev Car 20	HOT#Y SEISE BTAT BOLLTUTH
Other	-	3 10	082+	rifered	10:05:0000 3:48:47 PM	el susteavandenel an	Ameria (Auto)	iti	14247	7296	10.20.2009 2.81.50	11/2/2009 3:51:55 PM	Silverine_MATRIX_INSTRUCT or	#jhnend	144	Here and an and a second secon	STOP RESTART WORKS SILLS STAT DO DU DU STOP RESTART
Links	-		0423	BEAS7	19-28-2009-2-57-62 PM	mon-Sillion receils, sile	daurd (Smjath)	121	4018	7200			Preside_Run_4	0.6.40	28	indext standards	NOTIFY BBIEF STAT DO FLOU PM
SCELLANEOU	rs. E		0822	P-CLUSTALH	19/05/2009/2 +9/01 PM	hateret distances in the last	(8%-04) (8%-04)	19	1920	4325	15/26/2020 2 48 58 PM	10:28:2009.2:50:0+ PM	ەدراولەراد	FINISHED CORRECTLY	10	input stantmarts mained scoretarts tog the start 20	AND HIS LEL HELTAR NOTIFY 2212 F
discribe ga Home	E		1980	todayee	10-28-2309 2 44 38 Ph2	other standing to the state	Silain (Auto)	14	2+919	14400	13-28-2029-2 47-19 PM	11/7/2000 2 47:18 PM	edelaria_effie	RUNNID	140	input insettanta indevi insettanta ing fre itan 20	NOTFYSSIEC STAT 2015 DUTY
untern Statue plications atistics	-	7 10	9626	N/Sayes	10:28:2009 2 44 28 Phi	and sumBarrier, hada	kinder (Auto)	(4)	2+918	14400	10 20 2000 2 47 38 PM	11/10000 2 47 38 PM	etelpha_mi	Ruberts	-	Have standtautta mobil standtautta hag the stan 22	NOTIFY BEIEC STAT BOELDED
SU Home	-		G818	P-CLUSTALW	10-08-0009-2-02-48 PM	helenet Aproximited, whe	staurold (Ayla)	a,		188	10-28-2009-2-22-50 PM	10-28-2009 2-22-52 PM	ور خدد	FINEHED CORRECTLY	14	input stantisatta pulaut stantisatta ing the star 20	ADDITIVE DEL RESTAR NOTIFICE STATE BEAT DE LA SUIDI
ISU Rp server ISU SegDB	۲ I	• •	G818	Rie	19-08-0009-2-00-2-00 PM	transformationer	sipeumi (Auto)	(1)	16239	4320	16 (36 (36 (4 2 (2 2 4 1 PM	10:51:0008:2:22:81 PM	4004L3_4_netBat_4.64	15.000kg	14	input stantisarts subsit stantisarts ing the start 20	NOTIFY B 2 1 5 5 STAT 35 5 50 FM
IC Windows olidormatics oplications) . E	10 10	2917	ille.	15/25/2008/2 13:17 PM	transfarrat.aug	1894/111 (Fe/10)	(9)	18238	4329	10.28.2009.2.13.42 PM	10/28/2009 2:18 68 PM	and the second second	PHILIPED CORRECTLY	786	Hand stantistics and an extension ing the star bit	NOTIFY BELSE WATER LESS C
STRUCT REX (T.RFLP	-	11 10	C816	hida,m	10.05.000 2.00 48 PM	standallation ada	illevent (Avit)	(4)	16237	14400	1928-3018-2-51.19 PM	11/120062-01.18 PM	alalpha, 31, gave	Runnens	-	Ment Explicitly todayl startcasts reg file star 20	NOTFY 22150 STAT. 25 (2 TU CM
anager) ISU Survey	1000	12 10	GB1E	attracture.	15/08/2008 1 69/01 PM	distribute a	86-48 (81)	(11)	1822	1440	10:03:0009:0 10:48 PM	10:29:0009:2:10:49 PM	LOC#	RUNNING .	1.00	input abartophia puted standtophia tog the glass 20	NOTIFY A DIEC STAT SEEL TO PM
ad Survey (a A Q	den) E	13 19	0814	and in the second se	15 (34 (30 (34 (17 (34) PM	MantetralitonLan	staum? (Auto)	195	18239	7355	10.05.0008 1.87.84 PM2	1133000 187.84 PM	democratical lat	RUNING	744	Indext standtautha indext standtautha ing the start litt	ACCEPTED 110
antact Us	-	14 10	CRI IS	96	10:08:0000 1 55:27 PM	Nisenstrathunal.arp	(864/m) (Auto)	(9)	18218	7200	10/28/2018 1 54 51	11/2/2008 1 88 81 PM	Germonicaliti bat	PLANING	786	Input stastituita subul stastituita topite star 20	ADDER BERGET HOTEFFEELED BTAT DE FLEU PM

Funded in part by Microsoft Research

There are both parallel and serial applications available through the interface. LOOPP and MrBayes are examples of genuine parallel applications. P-BLAST, P-HMMER and P-IPRSCAN are parallelized through input sequence distribution (trivial parallelization). MPI is used for communication.

Selected popular applications Job submission from 6/13/2003 to 1/24/2010

LOOPP	19,726	protein structure prediction
MDIV	20,754	population genetics
IM/IMa	21,635	population genetics
P-BLAST	4,422	sequence analysis / data mining
MrBayes	17,308	population genetics
STRUCTURE	15,339	population genetics
All applications	total 130 585	(average 20,090 per year)

- since 1/24/2009 40,581 (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 1-4 cores for a few hours to two weeks
 - (average: a week)

The BioHPC jobs have been submitted by 11 471 unique users from 83 countries, the majority (57% by CPU time used) coming from the USA, with 52% of the utilized CPU time from the USA coming from New York State. These users include 257 unique users from Cornell, 2,580 users from *.edu* domains representing 426 unique *.edu* institutions, and 4,813 users from *.com* domains (including 4,191 users with Yahoo, Gmail and Hotmail e-mail addresses).



FUTURE

The interface source code is freely available. The "packaged" version of the interface can be downloaded from BioHPC.net and installed locally with any Microsoft CCS or HPC 2008 cluster.

We are planning to use our web service interface to integrate BioHPC with the **Microsoft Biology Foundation**. The pool of applications available through the web service will be expanded and a client will be developed as MS Excel add-in. We currently working on support for 2nd generation sequencing pipelines. The suit will be optimized to take better advantage of multi-core nodes.

ABOUT CBSU

Computational Biology Service Unit (CBSU) of the Cornell University Life Sciences Core Laboratories Center was initiated by the Tri-Institutional collaboration among Cornell University, Weill Cornell Medical College, Rockefeller University, and Memorial Sloan-Kettering Cancer Center. In February 2006 CBSU became Microsoft HPC Institute charter member. CBSU is Cornell core facility for computational biology. BioHPC development is now partially funded by Microsoft Research and CBSU is a **Microsoft Biology Initiative** partner.

The BioHPC installation at CBSU is currently using 5 Microsoft Windows based local compute clusters totaling 976 cores. The local nodes use Microsoft Server 2003 with CCS and Microsoft Server 2008 with HPC Server 2008. 80 CPU cores of the remote cluster Athena (located in Redmond, WA) are also available via JSDL, courtesy of Microsoft.