







GeneGrid: Grid Service Based Virtual Bioinformatics Laboratory

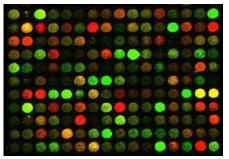
P.V. Jithesh

Bioinformatics - Data Driven

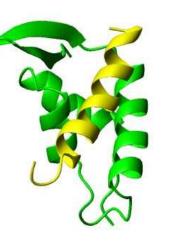


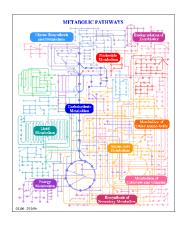
- Genome Sequencing Projects
 - 266 published complete genomes
 - 730 prokaryotic ongoing
 - 496 eukaryotic ongoing
 - http://www.genomesonline.org/
 - 21-06-2005
- Macromolecular
 Structure Elucidation

 Gene Expression Analysis



Metabolic pathways





Databases, Tools, Servers



- 719 databases (171 more than 2004 issue)
 - Nucleic Acids Research, 2005, Vol. 33
 (Database issue)
- Algorithms and tools for analysis plenty
- Most tools available through web servers
- 137 web servers
 - Nucleic Acids Research 2004, Vol. 32 (Web Server issue)

GeneGrid: Background



- Workflow Based Grid Computing project
- Initiated by Belfast e-Science Centre
- Commercial partners
- Antibody target discovery
- Genetic disease markers for New diagnostics
- Cancer and Immunology
- Potential Products from Molecular Mining
- Epilepsy







GeneGrid: Objectives



- Grid Based Framework for Bioinformatics Analysis
- Integration of Existing Technologies & Data Sets
- Production of a 'Virtual Bioinformatics Laboratory'
- Platform for scientists to access collective skills and experiences in a secure, reliable and scalable manner
- in silico knowledge discovery

GeneGrid: Components



- Application Integration & Management
- Data Access, Integration & Storage
- Resource Monitoring & Service Discovery
- Workflow Management
- Portal

Application Management



- Integrates with GeneGrid
 - Bioinformatics Applications
 - BLAST
 - TMHMM
 - SignalP
 - Primer3
 - HMMER
 - EMBOSS
 - ...
 - Utility Programs
- Highly extensible
- Two types of GT3 based Grid Services
 - Factory
 - · Persistent, Generic
 - Discoverable by other services through Registry service
 - Instance
 - Transient, Specific to task requested

Data Access, Integration and Storage



- Integrates with GeneGrid
 - Public biological databases
 - EMBL
 - SwissProt
 - ...
 - Private databases
- Manages GeneGrid specific databases
 - GeneGrid Workflow Definition Database (GWDD)
 - GeneGrid Status Tracking, Result & Input Parameter Database (GSTRIP)
- Based on OGSA-DAI
 - Replicates Data Manager Service Factory and Data Manager Service
 - Extended to support flat files

Resource Monitoring & Service Discovery



- GeneGrid Application & Resources Registry (GARR)
 - Central registry service GT3 based
 - Receives data about resources & services, Stores in database
 - Provides interface to query the data
- Node Monitors
 - Present on all resources
 - Transmits resource status & service availability to GARR

Workflow Management



- GeneGrid Workflow Manager roles
 - Processing of workflows
 - Resource identification
 - Task dispatch
 - Task status update
- GT3 based services
 - Factory
 - Persistent
 - Discoverable
 - Instance
 - Transient
 - Specific to one workflow

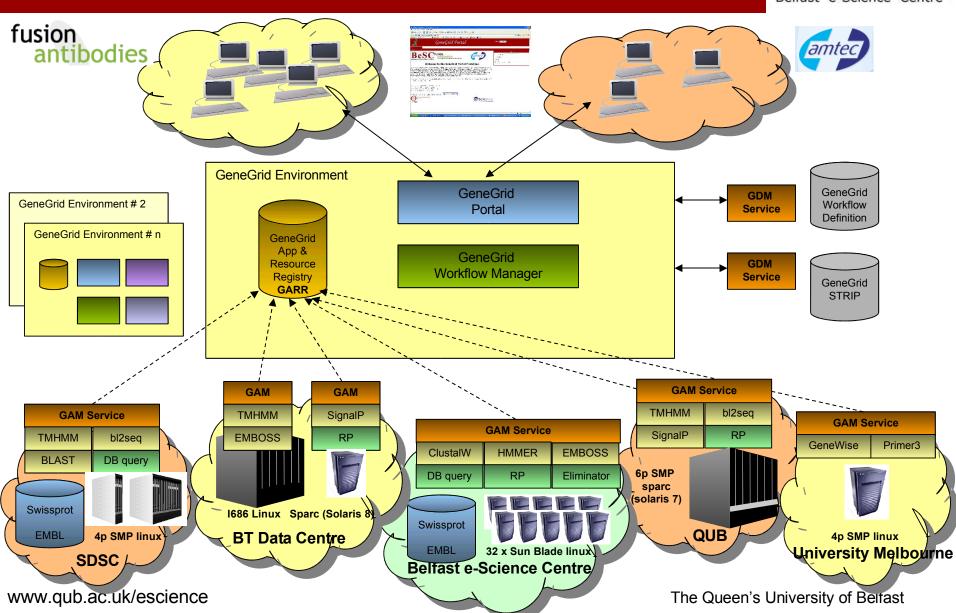
Portal



- User interface
- Creation and validation of workflows
- Query and display of results
- Conceals the complexity of Grid from the user
- Relies on data from 2 databases
 - GeneGrid Workflow Definition Database (GWDD)
 - Master Workflow Definition XML
 - GeneGrid Status Tracking, Results & Input Parameters Database (GSTRIP)
 - Input files and parameters
 - · Results and metadata
- Based on GridSphere
 - JSR 168 Compliant Portlets
 - Creation & Submission of workflows
 - Querying workflow status
 - Display of results
 - Administration

Architecture





Use Cases

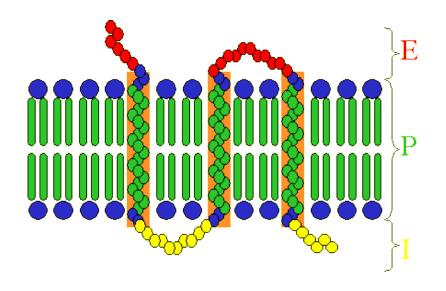


- A Identification of Novel Protein Family Members
- B Automated Antigenic Region Detection

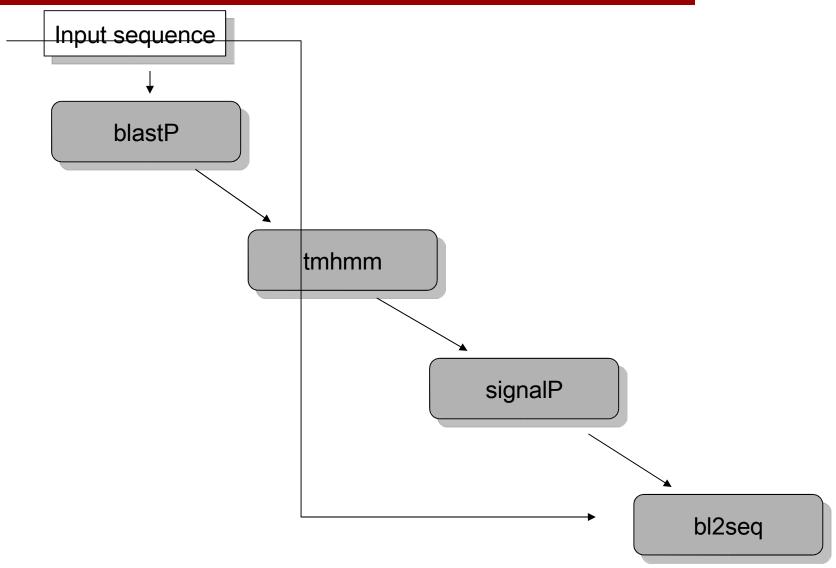
A - Identification of Novel Protein Family Members



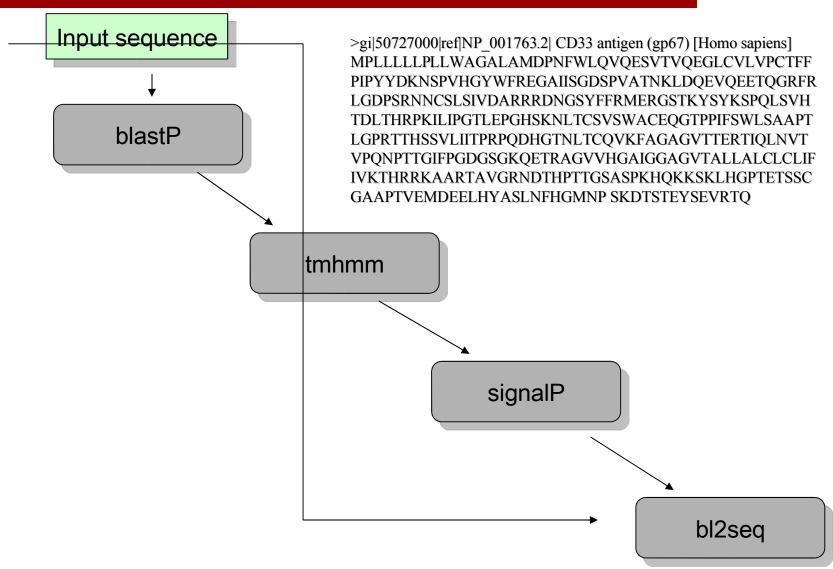
- Identify novel proteins of a family
- Cell surface proteins usually targets for the action of drugs
- Sialic acid binding Immunoglobulin-like lectins (Siglec) family



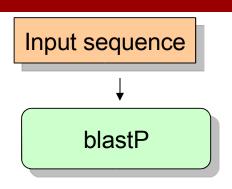












BLASTP 2.2.9 [May-01-2004]

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

Query= gi|50727000|ref|NP_001763.2| CD33 antigen (gp67) [Homo sapiens] (364 letters)

Database: swissprot

154,145 sequences; 56,721,989 total letters

Searching......done

Score E

Sequences producing significant alignments: (bits) Value

sp|P20138|CD33_HUMAN Myeloid cell surface antigen CD33 precursor... 675 0.0 sp|O43699|SIL6_HUMAN Sialic acid binding Ig-like lectin 6 precur... 313 4e-85 sp|Q9NYZ4|SIL8_HUMAN Sialic acid binding Ig-like lectin 8 precur... 295 1e-79 sp|Q95LH0|SILL_PANTR Sialic acid binding Ig-like lectin-like 1 p... 287 3e-77 sp|Q9Y336|SIL9_HUMAN Sialic acid binding Ig-like lectin 9 precur... 286 4e-77 sp|Q9Y286|SIL7_HUMAN Sialic acid binding Ig-like lectin 7 precur... 286 5e-77 sp|Q96PQ1|SILL_HUMAN Sialic acid binding Ig-like lectin-like 1 p... 285 1e-76 sp|Q63994|CD33_MOUSE Myeloid cell surface antigen CD33 precursor... 266 8e-71 sp|Q920G3|SILF_MOUSE Sialic acid binding Ig-like lectin-F precur... 253 4e-67 sp|O15389|SIL5_HUMAN Sialic acid binding Ig-like lectin 5 precur... 248 2e-65

.....

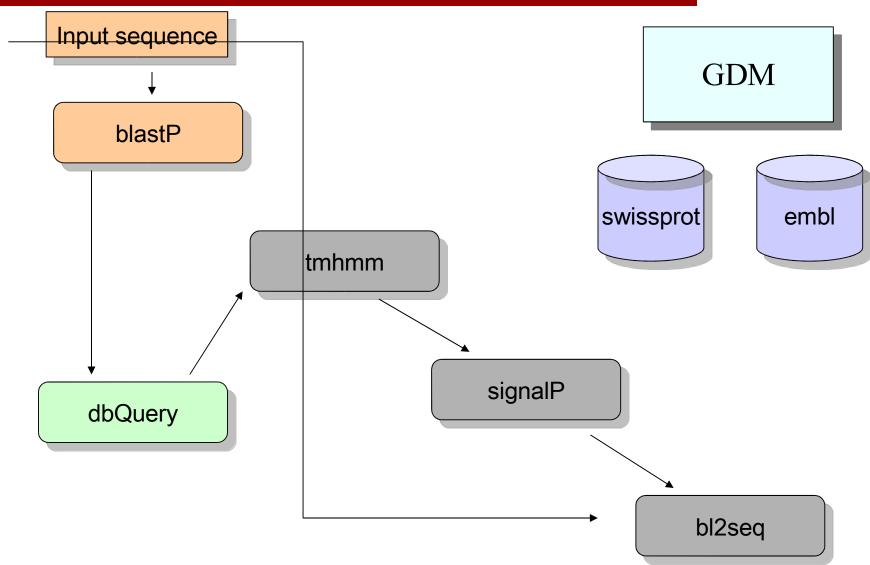
>sp|P20138|CD33_HUMAN Myeloid cell surface antigen CD33 precursor (gp67) (Siglec-3) Length = 364

Score = 675 bits (1742), Expect = 0.0 Identities = 328/354 (92%), Positives = 328/354 (92%)

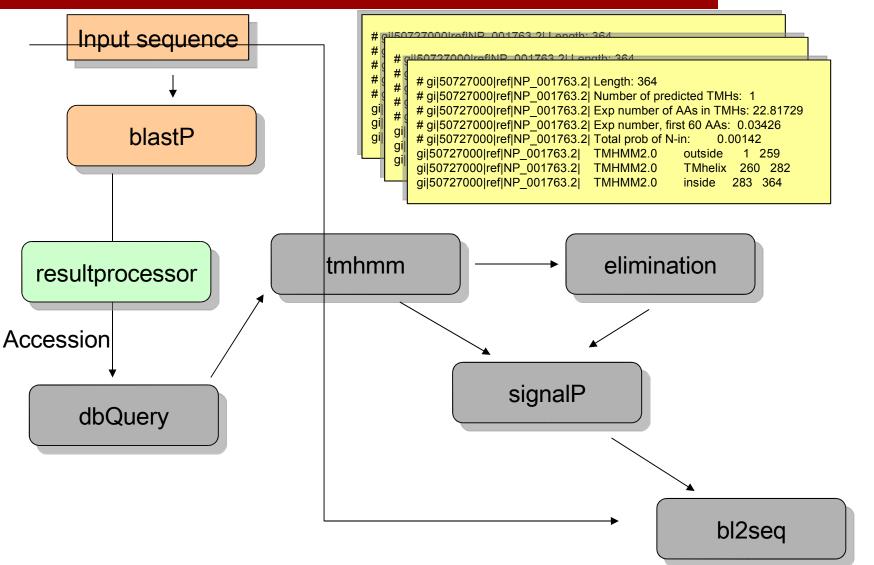
Query: 11 WAGALAMDPNFWLQVQESVTVQEGLCVLVPCTFFHPIPYYDKNSPVHGYWFREGAIISGD 70 WAGALAMDPNFWLQVQESVTVQEGLCVLVPCTFFHPIPYYDKNSPVHGYWFREGAIISGD Sbjct: 11 WAGALAMDPNFWLQVQESVTVQEGLCVLVPCTFFHPIPYYDKNSPVHGYWFREGAIISGD 70

Query: 71 SPVATNKLDQEVQEETQGRFRLLGDPSRNNCSLSIVDARRRDNGSYFFRMERGSTKYSYK 130 SPVATNKLDQEVQEETQGRFRLLGDPSRNNCSLSIVDARRRDNGSYFFRMERGSTKYSYK Sbjct: 71 SPVATNKLDQEVQEETQGRFRLLGDPSRNNCSLSIVDARRRDNGSYFFRMERGSTKYSYK 130

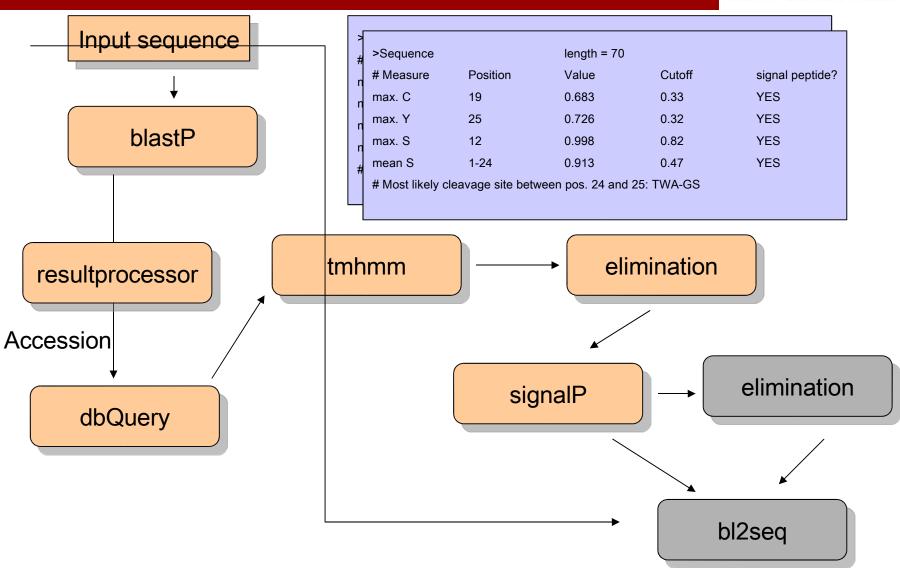




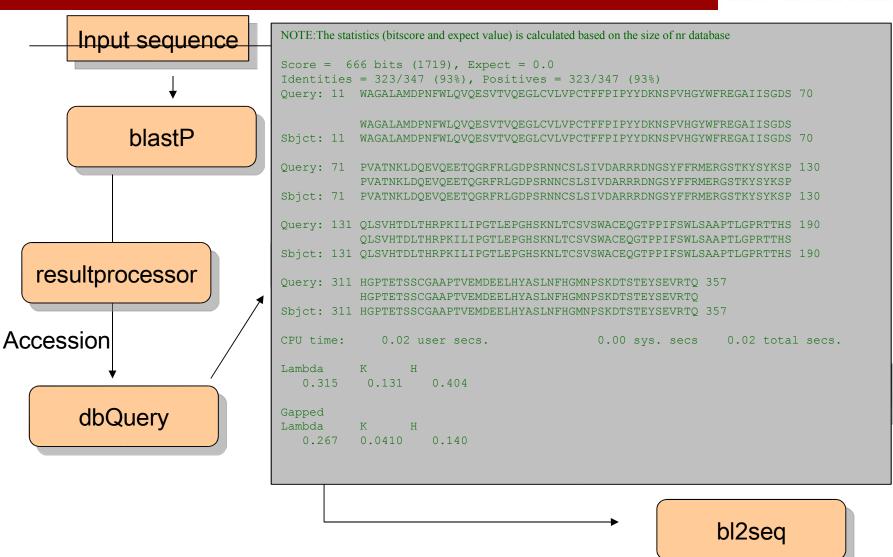












Use Case A - Results



- 6 Uncharacterised and potentially new siglecs
- Current experiment execution time: 1 day
- GeneGrid 20 mins
- Different applications were accessed from different resources
 - BLAST Linux Cluster at BeSC
 - TMHMM Linux Cluster at SDSC
 - SignalP Sun SMP machine at QUB

Use Case A - Results



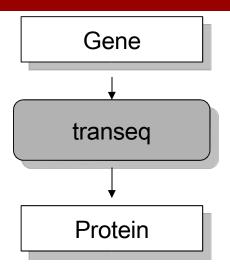
- Extended workflow involves beginning with a number of characterised sequences from a family
- Multiple sequence alignment and Profile generation (clustalW, hmmer etc.)
- Profile search against databases for sensitivity
- Finding whether the selected genes are actually transcribed (est database etc.)
- Phylogenetic analysis by dendrogram generation (Pileup etc.)
- Looking for characteristic domains of the family (rpsblast x CDD)

B - Automated Antigenic Region Detection



- Identification of Antigenic regions in proteins starting from the genes
- Routine Bioinformatics procedure in partner company for clients & in-house
- More than 100 genes at a time to be examined using a number of tools
 - 30-60 mins per gene
- GeneGrid allows automated detection of antigenic regions from genes

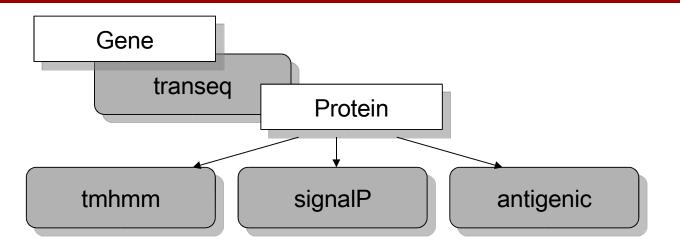




1	atggccgtca	tggcgccccg	aaccctcctc	ctgctactct	cgggggccct	ggccctgacc
61	cagacctggg	cgggctccca	ctccatgagg	tatttcttca	catccgtgtc	ccggcccggc
121	cgcggggagc	cccgcttcat	cgccgtgggc	tacgtggacg	acacgcagtt	cgtgcggttc
181	gacagcgacg	ccgcgagcca	gaggatggag	ccgcgggcgc	cgtggataga	gcaggagggg
241	ccggagtatt	gggaccagga	gacacggaat	gtgaaggccc	agtcacagac	tgaccgagtg
301	gacctgggga	ccctgcgcgg	ctactacaac	cagagcgagg	ccggttctca	caccatccag
361	ataatgtatg	gctgcgacgt	ggggtcggac	gggcgcttcc	tccgcgggta	ccggcaggac
421	gcctacgacg	gcaaggatta	catcgccctg	aacgaggacc	tgcgctcttg	gaccgcggcg
481	gacatggcgg	ctcagatcac	caagcgcaag	tgggaggcgg	cccatgaggc	ggagcagttg
541	agagcctacc	tggatggcac	gtgcgtggag	tggctccgca	gatacctgga	gaacgggaag
601	gagacgctgc	agcgcacgga	ccccccaag	acacatatga	cccaccaccc	catctctgac
661	catgaggcca	ccctgaggtg	ctgggccctg	ggcttctacc	ctgcggagat	cacactgacc
721	tggcagcggg	atggggagga	ccagacccag	gacacggagc	tcgtggagac	caggcctgca
781	ggggatggaa	ccttccagaa	gtgggcggct	gtggtggtgc	cttctggaga	ggagcagaga
841	tacacctgcc	atgtgcagca	tgagggtctg	cccaagcccc	tcaccctgag	atgggagctg
901	tcttcccagc	ccaccatccc	catcgtgggc	atcattgctg	gcctggttct	ccttggagct
961	gtgatcactg	gagctgtggt	cgctgccgtg	atgtggagga	ggaagagctc	agatagaaaa
1021	ggagggagtt	acactcaggc	tgcaagcagt	gacagtgccc	agggctctga	tgtgtccctc
1081	acagettgta	aaqtqtqa				

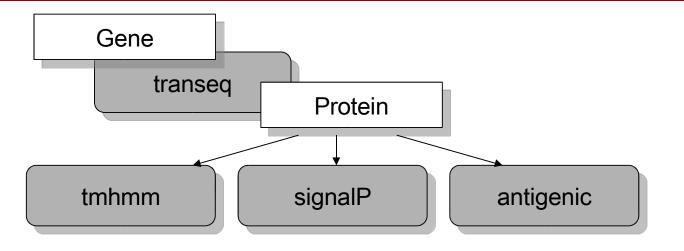
MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFFTSVSRPGRGEPRFIAVGYVDDTQFVRF
DSDAASQRMEPRAPWIEQEGPEYWDQETRNVKAQSQTDRVDLGTLRGYYNQSEAGSHTIQ
IMYGCDVGSDGRFLRGYRQDAYDGKDYIALNEDLRSWTAADMAAQITKRKWEAAHEAEQL
RAYLDGTCVEWLRRYLENGKETLQRTDPPKTHMTHHPISDHEATLRCWALGFYPAEITLT
WQRDGEDQTQDTELVETRPAGDGTFQKWAAVVVPSGEEQRYTCHVQHEGLPKPLTLRWEL
SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSDRKGGSYTQAASSDSAQGSDVSL
TACKV





- # Sequence Length: 365
- # Sequence Number of predicted TMHs: 1
- # Sequence Exp number of AAs in TMHs: 30.43917
- # Sequence Exp number, first 60 AAs: 7.38298
- # Sequence Total prob of N-in: 0.37875
- Sequence TMHMM2.0 outside 1 307
- Sequence TMHMM2.0 TMhelix 308 330
- Sequence TMHMM2.0 inside 331 365

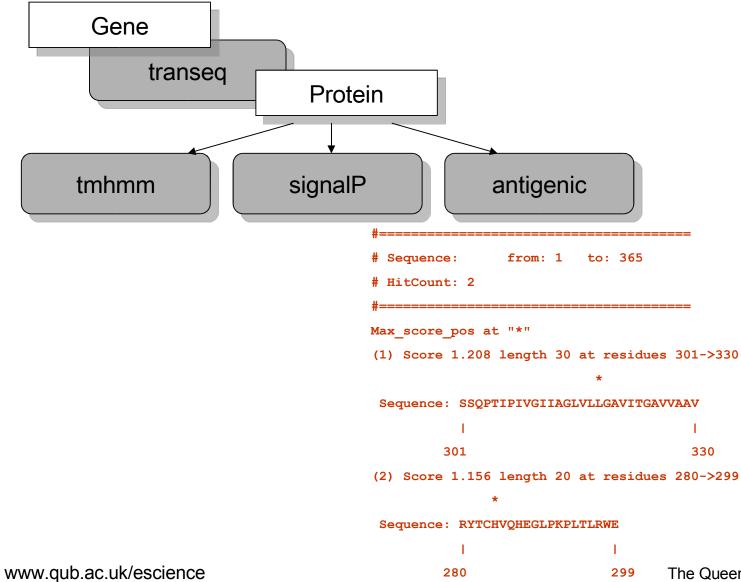




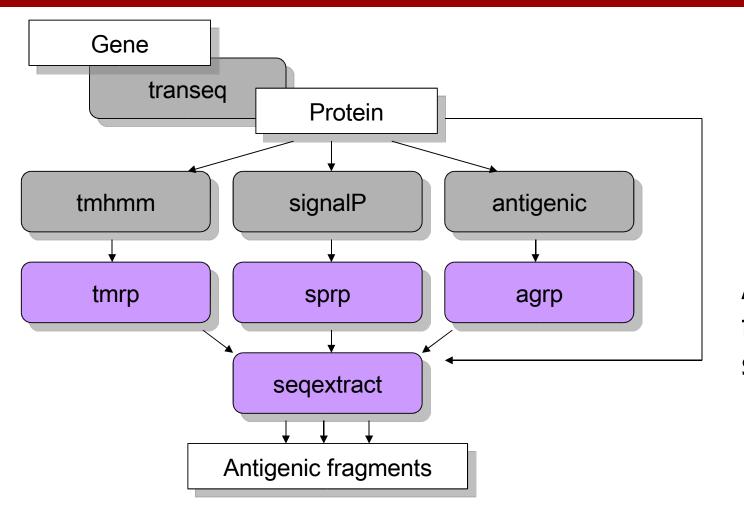
>Sequence		length = 70			
# Measure max. C	Position 19	Value 0.683	Cutoff 0.33	signal peptide? YES	
max. Y	25	0.726	0.32	YES	
max. S	12	0.998	0.82	YES	
mean S	1-24	0.913	0.47	YES	

Most likely cleavage site between pos. 24 and 25: TWA-GS



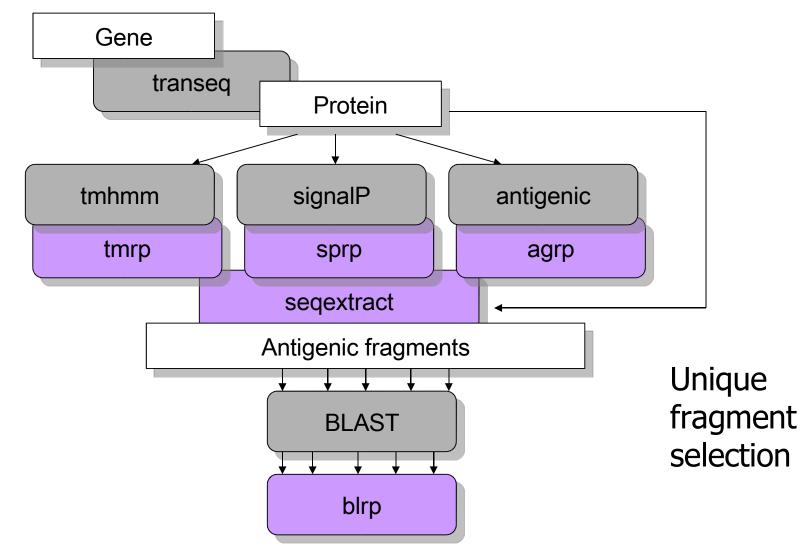




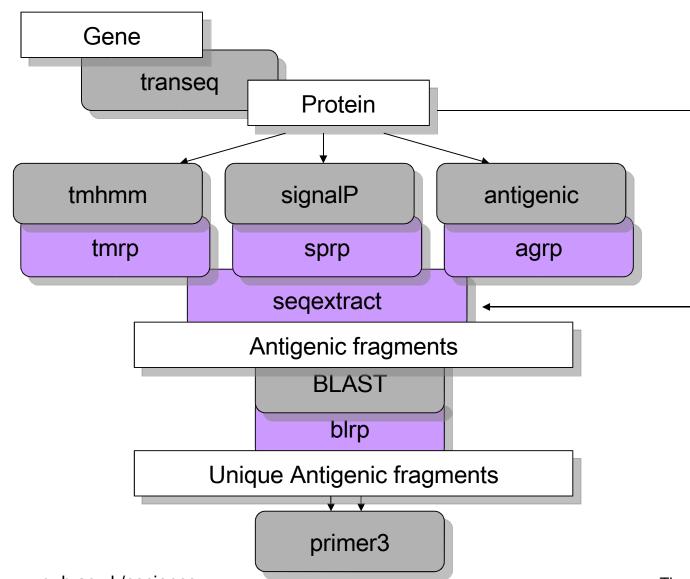


Antigenic fragment selection







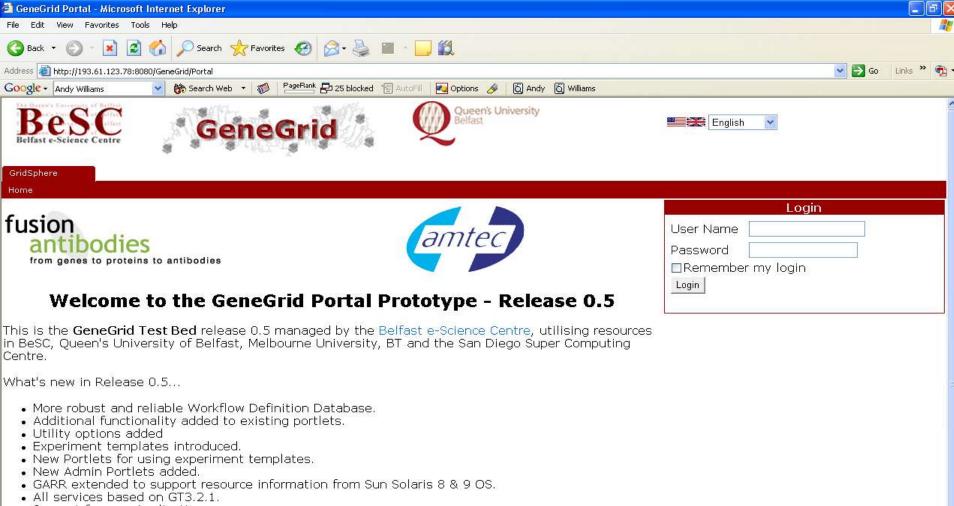


Select Primer sequences for PCR

Use Case B - Results



- Pre GeneGrid 30-60 min per gene
- GeneGrid 90 mins for 100 genes
- Resources used
 - BeSC, BT Datacentre, Uni Melbourne, SDSC
- Automation of time consuming routine bioinformatics tasks
- Individual task execution and overall experiment execution times reduced
- High throughput analysis of genes for potential antigenic regions



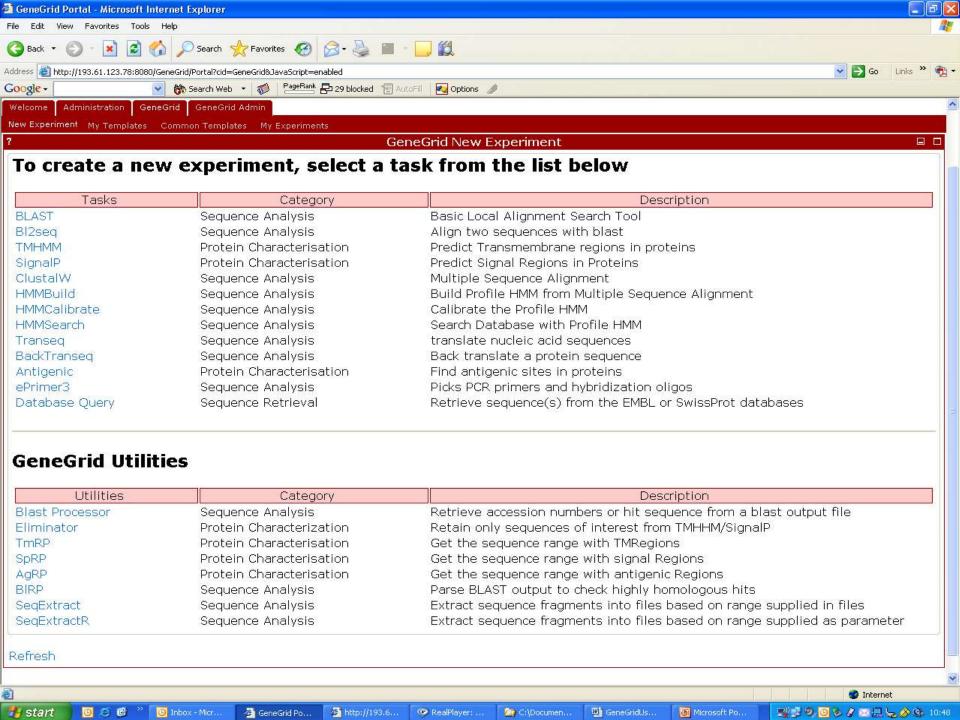
- Support for new Applications.
- Enhancements made to GSTRIP database

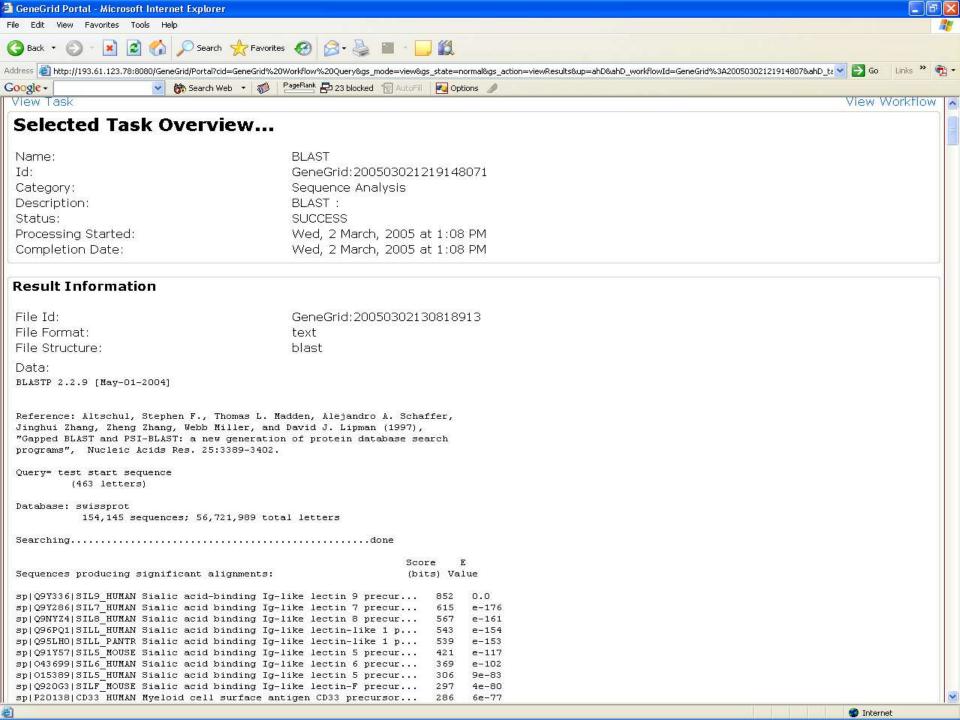
Users are limited to selected staff of both commercial partners - Fusion Antibodies, Amtec Medical - and the Belfast e-Science Centre. To obtain a user account, please contact the appropriate representative -Noel Kelly (BeSC), Mark McCurley (Fusion) or Dr. Shane McKee (Amtec). Authorized users will be provided with a username and password by BeSC.

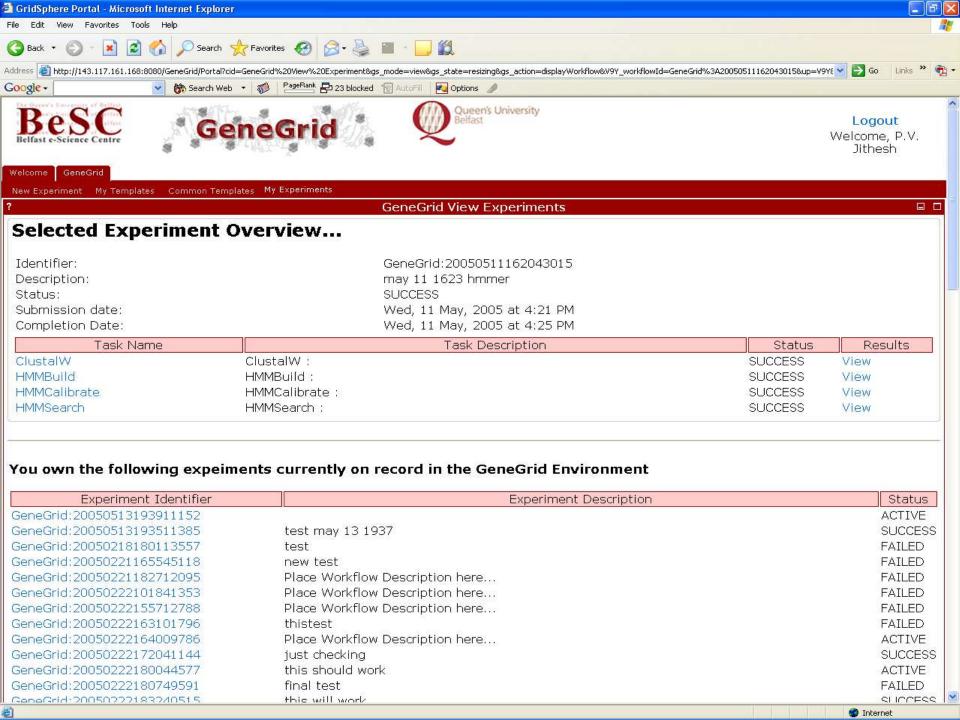
All users are requested to subscribe to the GeneGrid mailing list and to use it for directing queries etc. Mail GeneGrid, and place the word "subscribe" (without the quotes) in the message body.

For more on the GeneGrid project, please click here.









GeneGrid: Status



- 30 month project, started in August 2003
- Prototype Releases
 - 0.1 March 2004
 - Conceptual prototype
 - 0.2 August 2004
 - Functional prototype
 - 0.3 October 2004
 - First release for commercial partners' use
 - 0.4 January 2005
 - 0.5 June 2005

Thank You!



- Project Manager: Dr Paul Donachy
 - p.donachy@qub.ac.uk
- Senior Software Engineer: Noel Kelly
 - n.kelly@qub.ac.uk
- Grid Programmer: Sachin Wasnik
 - s.wasnik@qub.ac.uk
- Bioinformatician: P.V. Jithesh
 - p.jithesh@qub.ac.uk
- More information:

http://www.qub.ac.uk/escience/projects/genegrid/