Middleware Framework for HG2C Project

Dec 13, 2006

Jaeyoung Choi

Soongsil University, Seoul Korea

choi@ssu.ac.kr

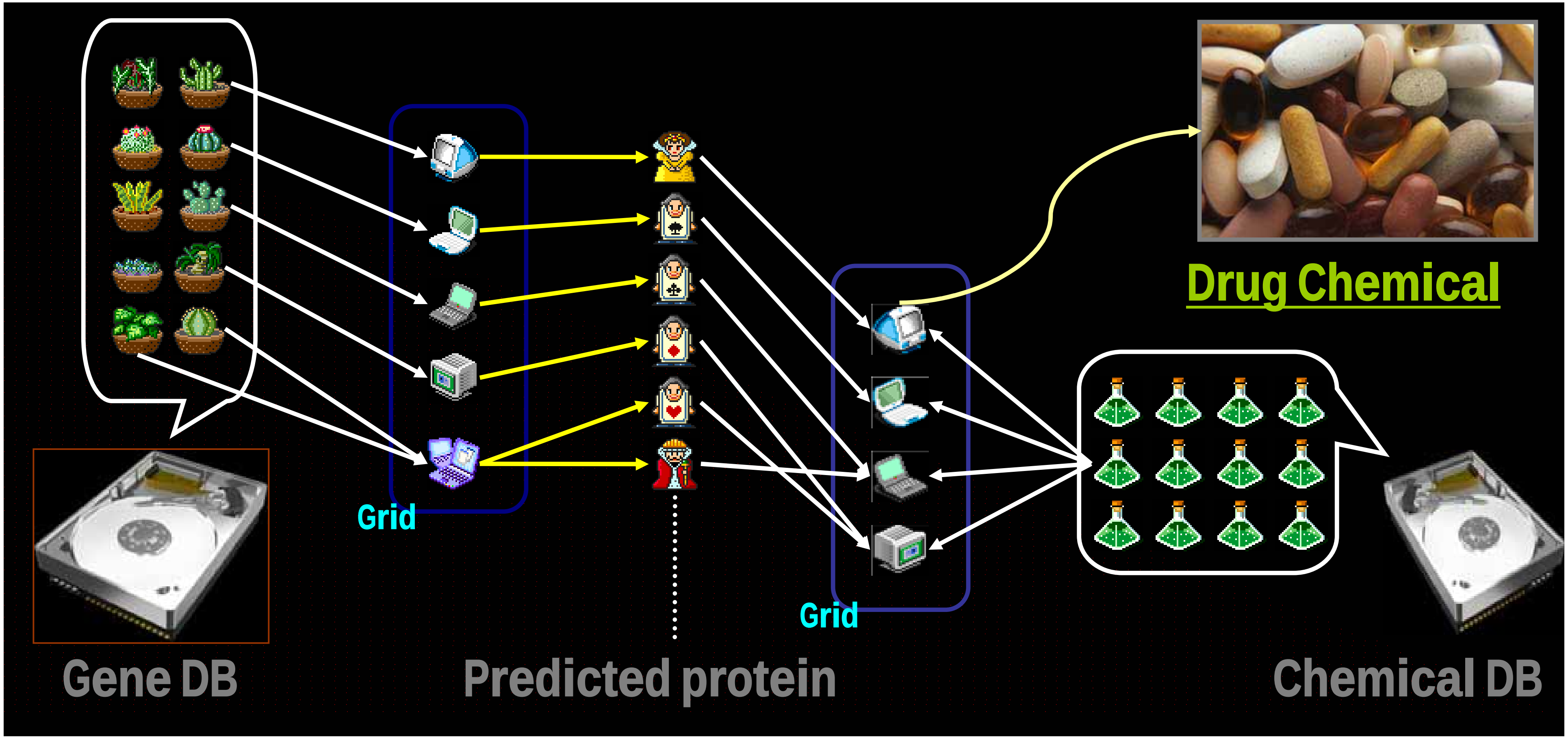
Middleware Framework for HG2C Project

December 2006

Jaeyoung Choi
Soongsil University, Seoul Korea
choi@ssu.ac.kr

HG2C (Human Genomes to Chemicals)

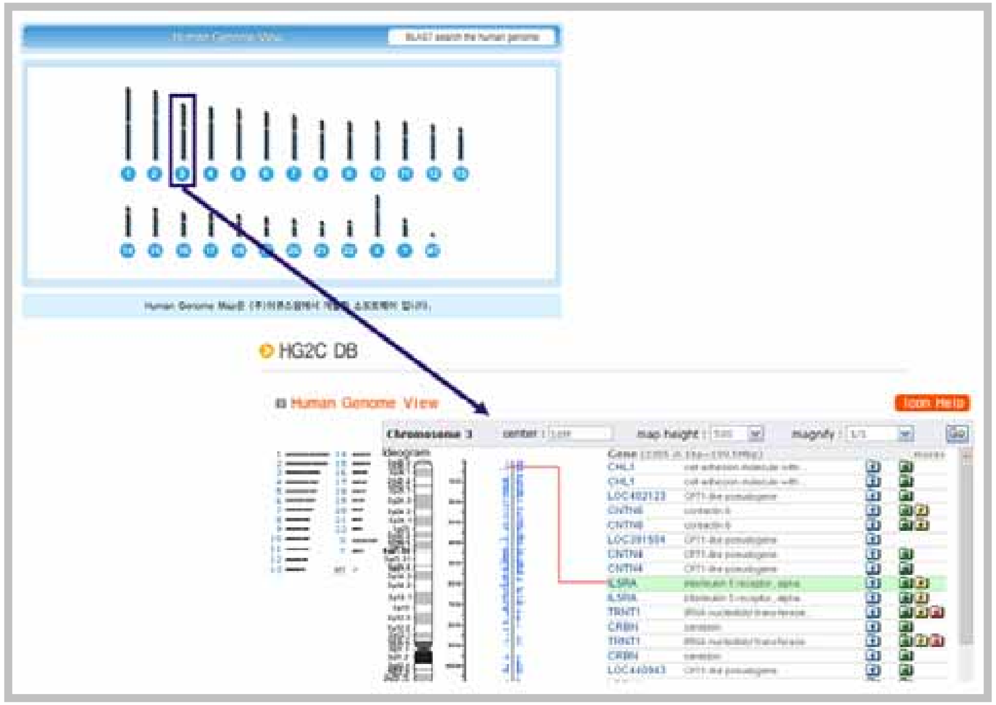
Dream of HG2C



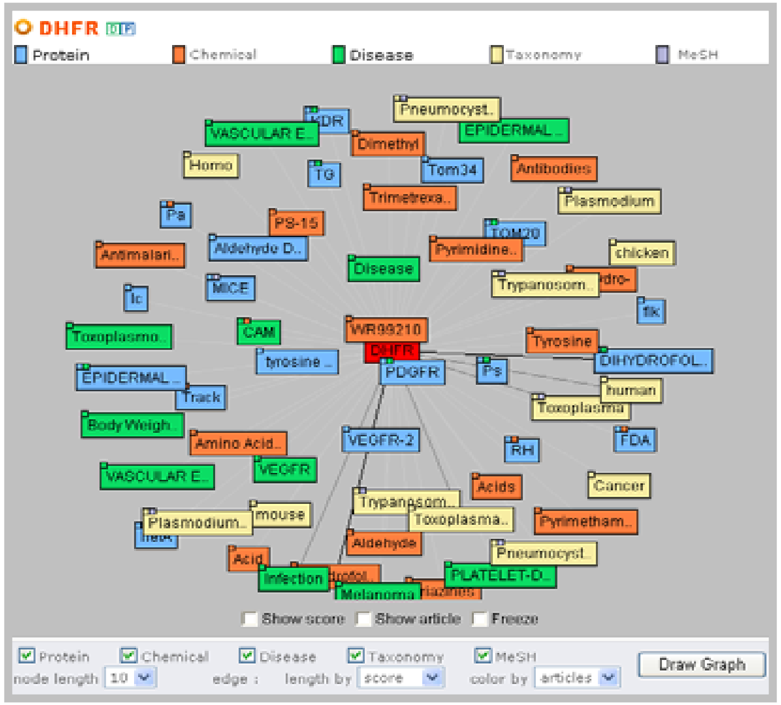
HG2C (1)

- Genome
  - Literature Information DB system (PubLink^{TM})

    Genes - Proteins - Chemicals
  - Chromosome/Gene/DNA sequence relation



CHL1	cell adhesion molecule with...
CHL1	cell adhesion molecule with...
LOC402123	CFT1-like pseudogene
CNTN6	contactin 6
CNTN6	contactin 6
LOC391504	CFT1-like pseudogene
CNTN4	CFT1-like pseudogene
CNTN4	CFT1-like pseudogene
IL5RA	interleukin 5 receptor, alpha
IL5RA	interleukin 5 receptor, alpha
TRNT1	tRNA nucleotidyl transferase
CRBN	cereblon
TRNT1	tRNA nucleotidyl transferase
CRBN	cereblon
LOC440943	CFT1-like pseudogene



□ Show score □ Show article □ Freeze

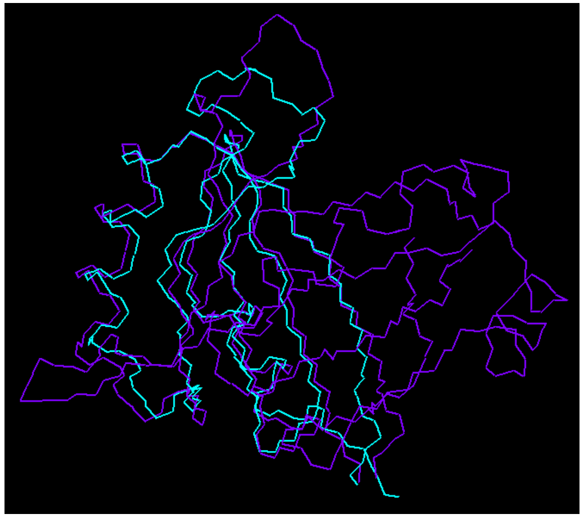
☑ Protein ☑ Chemical ☑ Disease ☑ Taxonomy ☑ MeSH

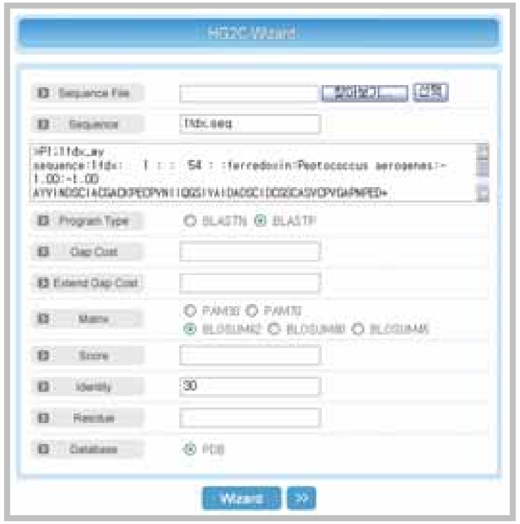
node length 10 edge : length by score color by articles

Draw Graph

HG2C (2)

- Gene
  - From DNA sequence
  - BLAST homology search
  - Comparative modelling from sequence
  - Protein structure prediction





Sequence File

Sequence: 1fdx.seq

```
>P1;1fdx_my
sequence:1fdx:   1 : :  54 : :ferredoxin:Peptococcus aerogenes:-
1.00:-1.00
AYVINDSCIACGACKPECPVNIIQGSIYAIDADSCIDCGSCASVCPVGAPNPED*
```

Program Type: ○ BLASTN ● BLASTP

Gap Cost:

Extend Gap Cost:

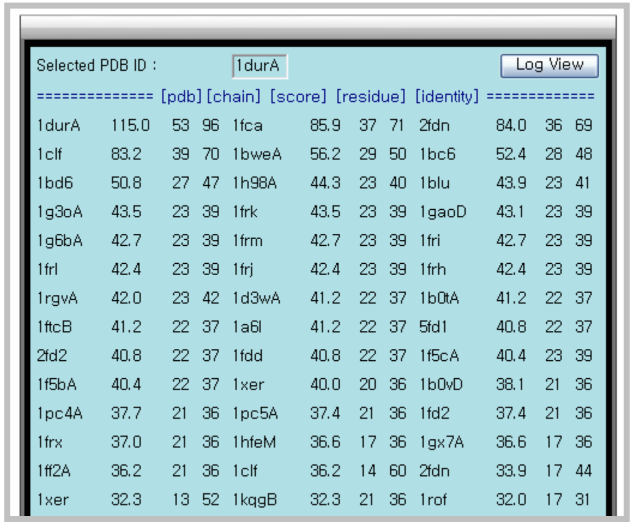
Matrix: ○ PAM30 ○ PAM70 ● BLOSUM62 ○ BLOSUM80 ○ BLOSUM45

Score:

Identity: 30

Residue:

Database: ● PDB

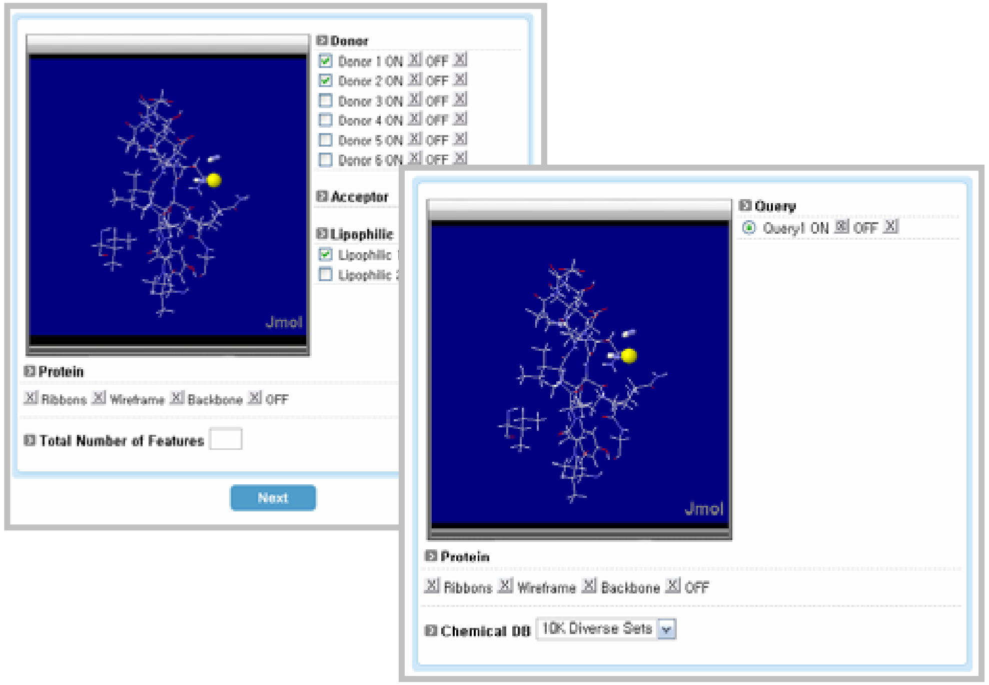


Selected PDB ID : 1durA

```
============= [pdb] [chain] [score] [residue] [identity] ============
1durA   115.0  53  96  1fca    85.9  37  71  2fdn   84.0  36  69
1clf     83.2  39  70  1bweA   56.2  29  50  1bc6   52.4  28  48
1bd6     50.8  27  47  1h98A   44.3  23  40  1blu   43.9  23  41
1g3oA    43.5  23  39  1frk    43.5  23  39  1gaoD  43.1  23  39
1g6bA    42.7  23  39  1frm    42.7  23  39  1fri   42.7  23  39
1frl     42.4  23  39  1frj    42.4  23  39  1frh   42.4  23  39
1rgvA    42.0  23  42  1d3wA   41.2  22  37  1b0tA  41.2  22  37
1ftcB    41.2  22  37  1a6l    41.2  22  37  5fd1   40.8  22  37
2fd2     40.8  22  37  1fdd    40.8  22  37  1f5cA  40.4  23  39
1f5bA    40.4  22  37  1xer    40.0  20  36  1b0vD  38.1  21  36
1pc4A    37.7  21  36  1pc5A   37.4  21  36  1fd2   37.4  21  36
1frx     37.0  21  36  1hfeM   36.6  17  36  1gx7A  36.6  17  36
1ff2A    36.2  21  36  1clf    36.2  14  60  2fdn   33.9  17  44
1xer     32.3  13  52  1kqgB   32.3  21  36  1rof   32.0  17  31
```

HG2C (3)

- Protein (IDPro™, PharmoMap™)
  - Active site searching from protein structure
  - Interaction feature searching from active sites



Donor
☑ Donor 1 ON ☒ OFF ☒
☑ Donor 2 ON ☒ OFF ☒
☐ Donor 3 ON ☒ OFF ☒
☐ Donor 4 ON ☒ OFF ☒
☐ Donor 5 ON ☒ OFF ☒
☐ Donor 6 ON ☒ OFF ☒

Acceptor

Lipophilic
☑ Lipophilic
☐ Lipophilic

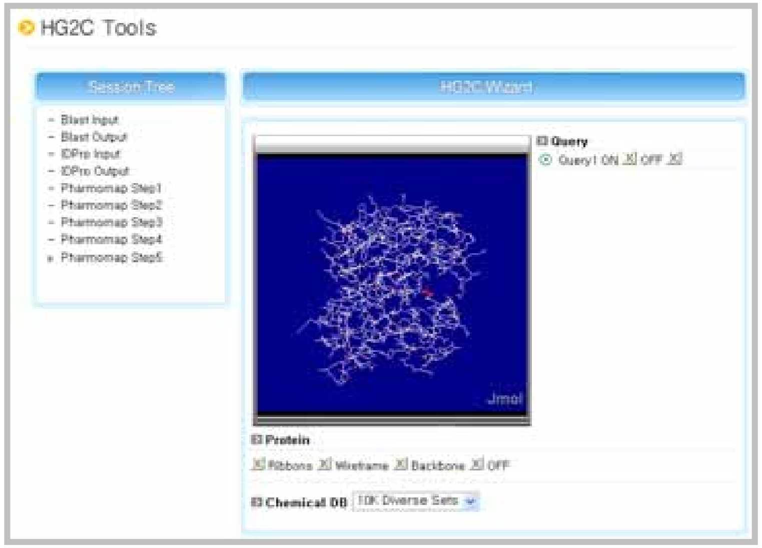
Protein
☒ Ribbons ☒ Wireframe ☒ Backbone ☒ OFF

Total Number of Features

Query
○ Query1 ON ☒ OFF ☒

Protein
☒ Ribbons ☒ Wireframe ☒ Backbone ☒ OFF

Chemical DB: 10K Diverse Sets



HG2C Tools

Session Tree
- Blast Input
- Blast Output
- IDPro Input
- IDPro Output
- Pharmomap Step1
- Pharmomap Step2
- Pharmomap Step3
- Pharmomap Step4
- Pharmomap Step5

HG2C Wizard

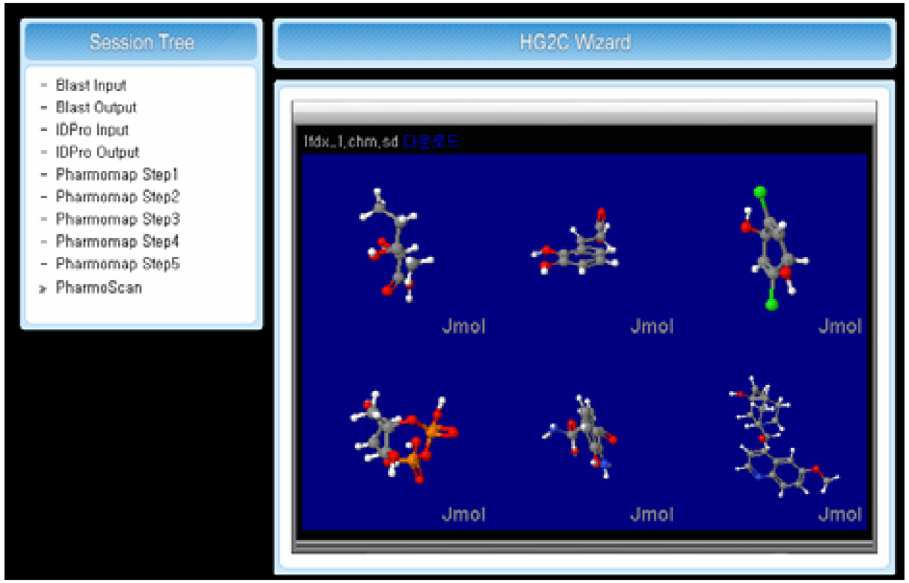
Query
○ Query1 ON ☒ OFF ☒

Protein
☒ Ribbons ☒ Wireframe ☒ Backbone ☒ OFF

Chemical DB: 10K Diverse Sets

HG2C (4)

- Chemicals (PharmoScan^{TM})
  - Combinations of interaction features
  - Chemical DB scanning : interaction combination as scanning query
  - Scanned results : virtual / known compounds



HG2C (5)

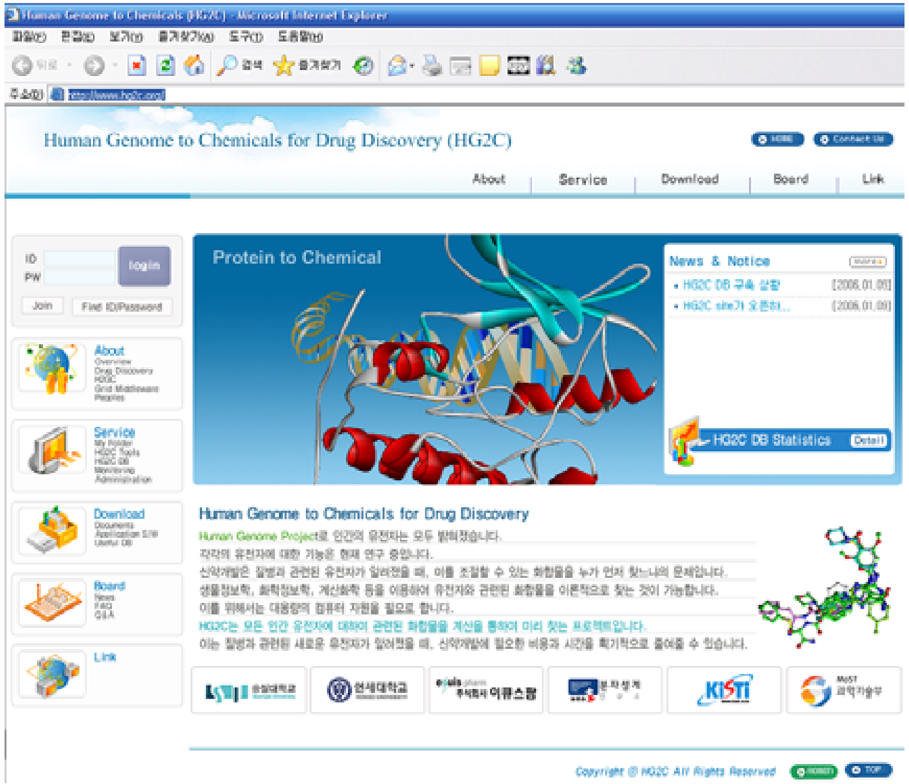
- Database
  - Iterative calculations on all the human genes
  - Browsing result DB :
    - Gene-Protein-Chemical relations :
      - ✓ key information for drug discovery
  - Result chemical analysis for drug utility and chemical attributes :
    - IDBManage^{TM}
    - JDDW (IDPro^{TM}, IDPharmo^{TM}, IDChemo^{TM})
    - PubLink^{TM}

HG2C (6)

- Massive Calculation
  - For all the genomes with a huge chemical DB
  - Choices of :
    - Protein model candidates
    - Active site candidates
    - Virtual screening query candidates
  - Huge size of computing requirement :
    - Grid computing: dynamic computing resource management with grid application utility MAGE support

HG2C Portal (1)

- Completion of Human Genome Project
- Theoretical prediction of chemical relations among gene functions and chemicals
- Complete independence of the each simulation
- Based on virtual computing:
  - Grid computing technology : MSF, MAGE
- http://www.hg2c.org



Human Genome to Chemicals (HG2C) - Microsoft Internet Explorer
파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)
뒤로 검색 즐겨찾기
주소(D) http://www.hg2c.org/
HOME
Contact Us
About
Service
Download
Board
Link
ID
PW
login
Join
Find ID/Password
About
Service
Download
Board
Link
Protein to Chemical
News & Notice
HG2C DB Statistics
Detail
Copyright © HG2C All Rights Reserved
TOP

Human Genome to Chemicals for Drug Discovery (HG2C)

Human Genome to Chemicals for Drug Discovery

Human Genome Project로 인간의 유전자는 모두 밝혀졌습니다.

각각의 유전자에 대한 기능은 현재 연구 중입니다.

신약개발은 질병과 관련된 유전자가 알려졌을 때, 이를 조절할 수 있는 화합물을 누가 언제 찾느냐의 문제입니다.

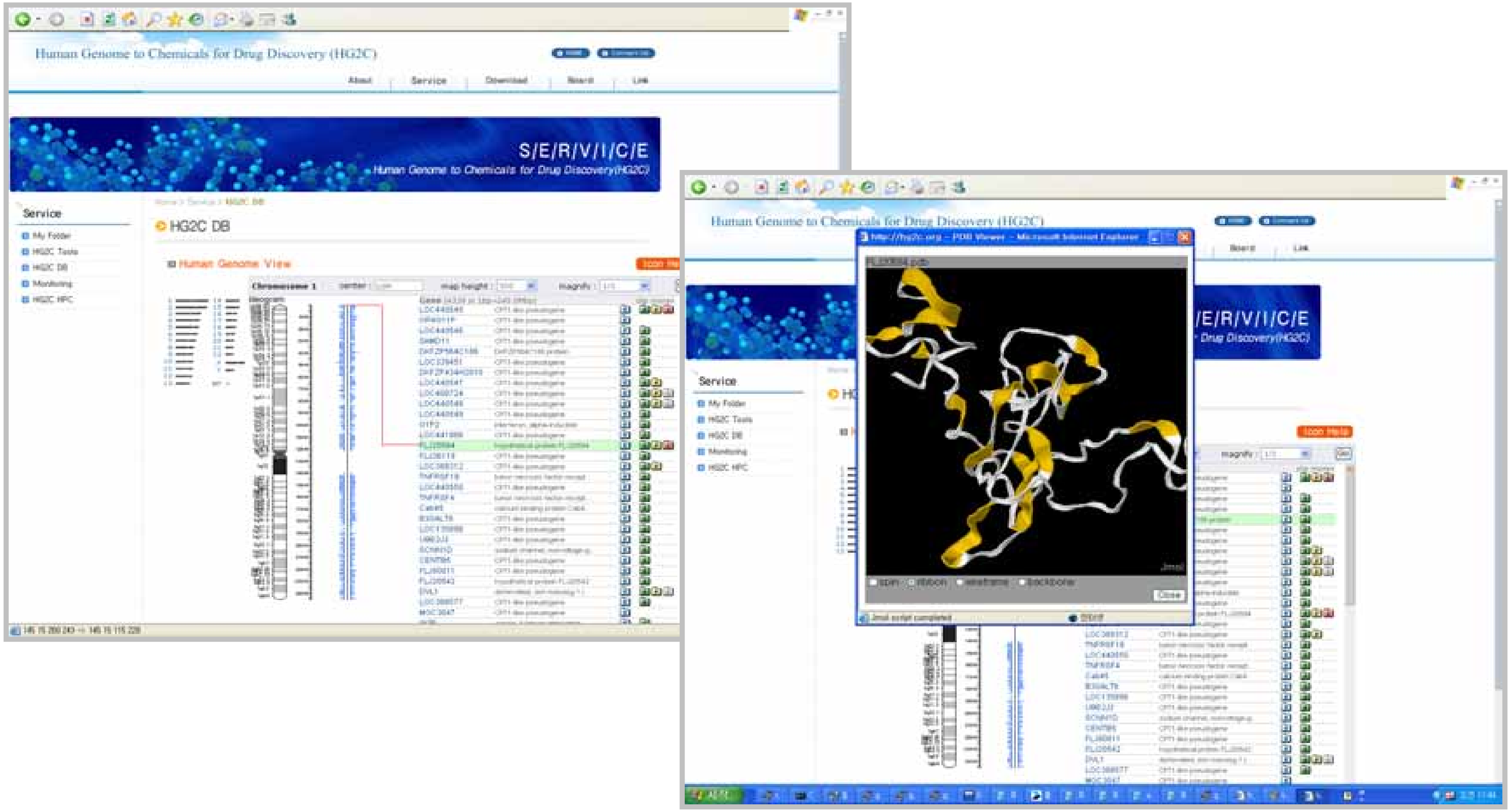
생물정보학, 화학정보학, 계산화학 등을 이용하여 유전자와 관련된 화합물을 이론적으로 찾는 것이 가능합니다.

이를 위해서는 대용량의 컴퓨터 자원을 필요로 합니다.

HG2C는 모든 인간 유전자에 대하여 관련된 화합물을 계산을 통하여 미리 찾는 프로젝트입니다.

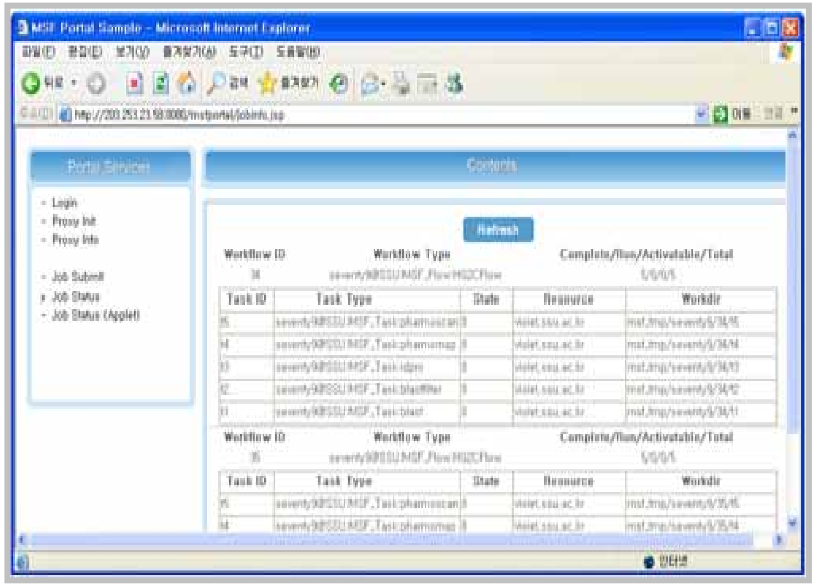
이는 질병과 관련된 새로운 유전자가 알려졌을 때, 신약개발에 필요한 비용과 시간을 획기적으로 줄여줄 수 있습니다.

HG2C Portal (2)



How to explore? - Sequentially

- Sequential processing: gene1, gene2, gene3…
- Automatic batch processing on Grid environment
- Using work flow engine MSF features of batch running, data file transferring and storing



MSF Portal Sample - Microsoft Internet Explorer

http://203.253.23.58:8080/msfportal/jobinfo.jsp

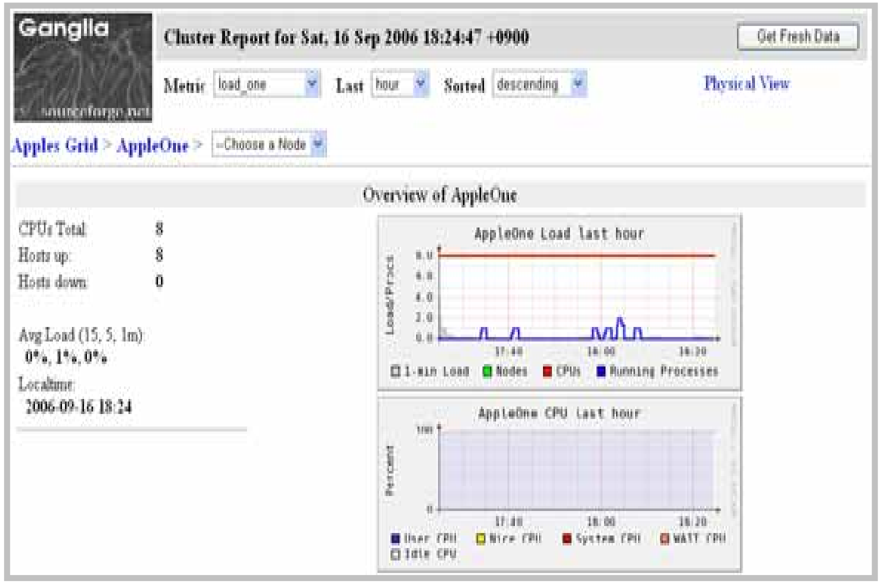
- Login
- Proxy Init
- Proxy Info
- Job Submit
- Job Status
- Job Status (Applet)

Workflow ID	Workflow Type	Complete/Run/Activatable/Total
34	seventy9@SSU.MSF_Flow.HG2CFlow	5/0/0/5

Task ID	Task Type	State	Resource	Workdir
t5	seventy9@SSU.MSF_Task.pharmoscan	0	violet.ssu.ac.kr	msf_tmp/seventy9/34/t5
t4	seventy9@SSU.MSF_Task.pharmomap	0	violet.ssu.ac.kr	msf_tmp/seventy9/34/t4
t3	seventy9@SSU.MSF_Task.idpro	0	violet.ssu.ac.kr	msf_tmp/seventy9/34/t3
t2	seventy9@SSU.MSF_Task.blastfilter	0	violet.ssu.ac.kr	msf_tmp/seventy9/34/t2
t1	seventy9@SSU.MSF_Task.blast	0	violet.ssu.ac.kr	msf_tmp/seventy9/34/t1

Workflow ID	Workflow Type	Complete/Run/Activatable/Total
35	seventy9@SSU.MSF_Flow.HG2CFlow	5/0/0/5

Task ID	Task Type	State	Resource	Workdir
t5	seventy9@SSU.MSF_Task.pharmoscan	0	violet.ssu.ac.kr	msf_tmp/seventy9/35/t5
t4	seventy9@SSU.MSF_Task.pharmomap	0	violet.ssu.ac.kr	msf_tmp/seventy9/35/t4



Cluster Report for Sat, 16 Sep 2006 18:24:47 +0900

Metric: load_one Last: hour Sorted: descending

Apples Grid > AppleOne > --Choose a Node

Overview of AppleOne

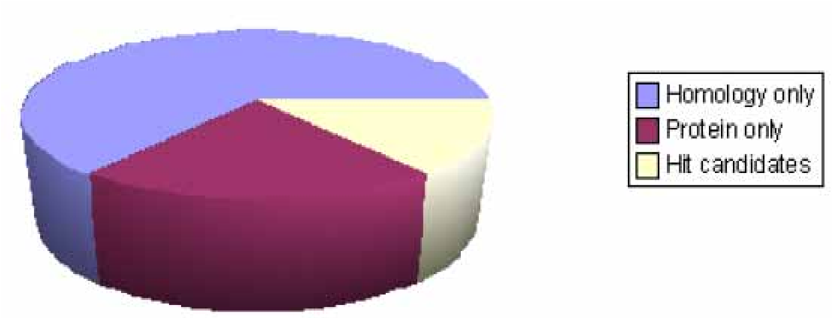
CPUs Total:	8
Hosts up:	8
Hosts down:	0

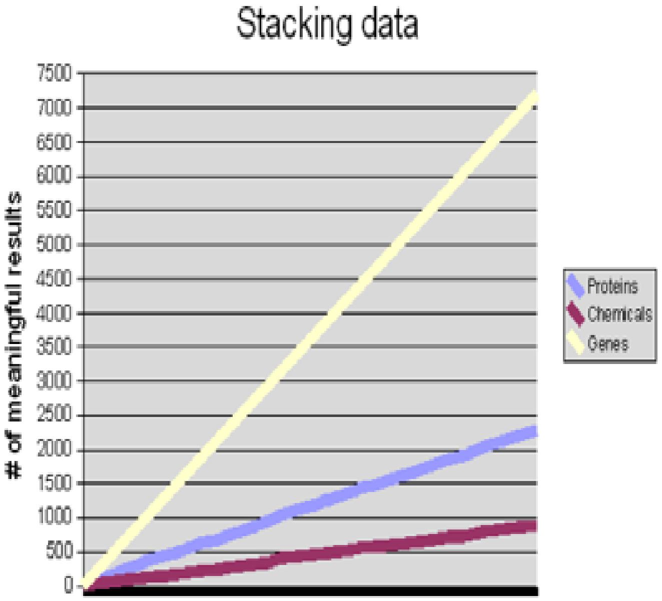
Avg Load (15, 5, 1m):
0%, 1%, 0%

Localtime:
2006-09-16 18:24

We have explored … (1)

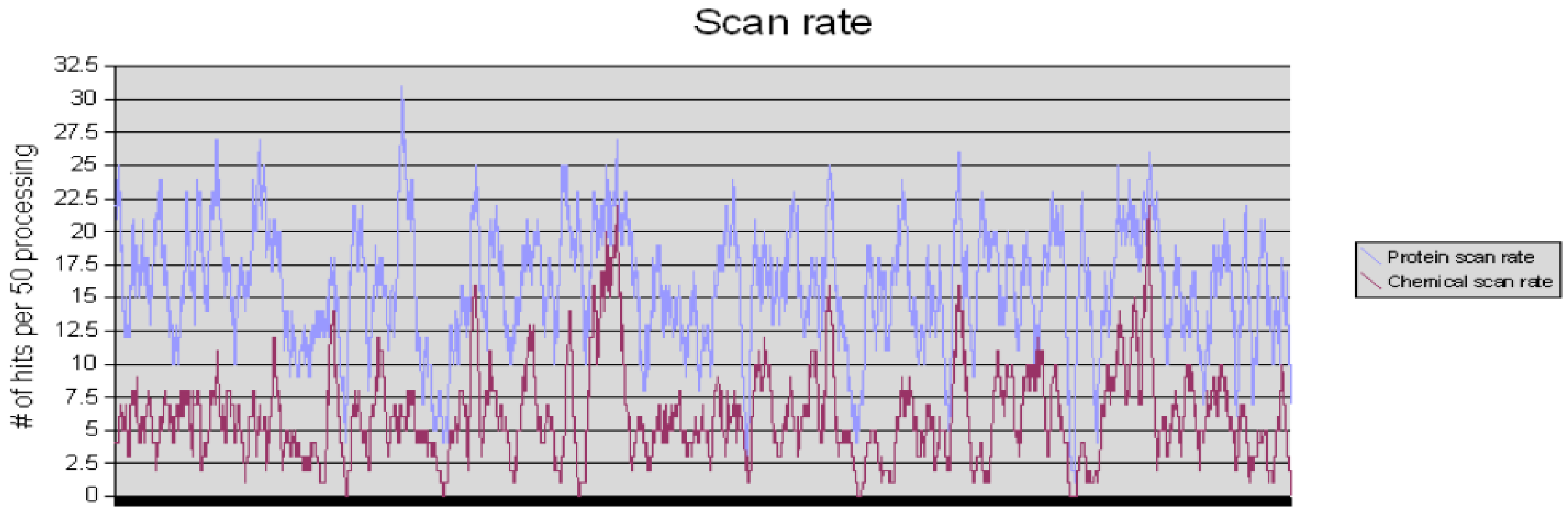
- Using 1 ~ 21 CPUs (Sep. 1st. 2006)
  - 7,191 genes calculated
  - 2,264 meaningful protein models achieved (31%)
  - 874 meaningful chemical sets achieved (12%)





We have explored … (2)

- Chromosome dimension vs. Structural significance
  - Independent protein homology patterns along the distribution on chromosome (StdDev: 9%)
  - Independent structural complexity patterns along the distribution on chromosome (StdDev: 5%)



Implementation Features

- HG2C DB opened http://www.hg2c.org
  - Human chromosome : Gene – Protein – Chemical
- Interactive GUI opened
- MSF integrated in modular usage level
- MAGE being integrated
- MSF package released http://sourceforge.net/
- MAGE package released http://sourceforge.net/

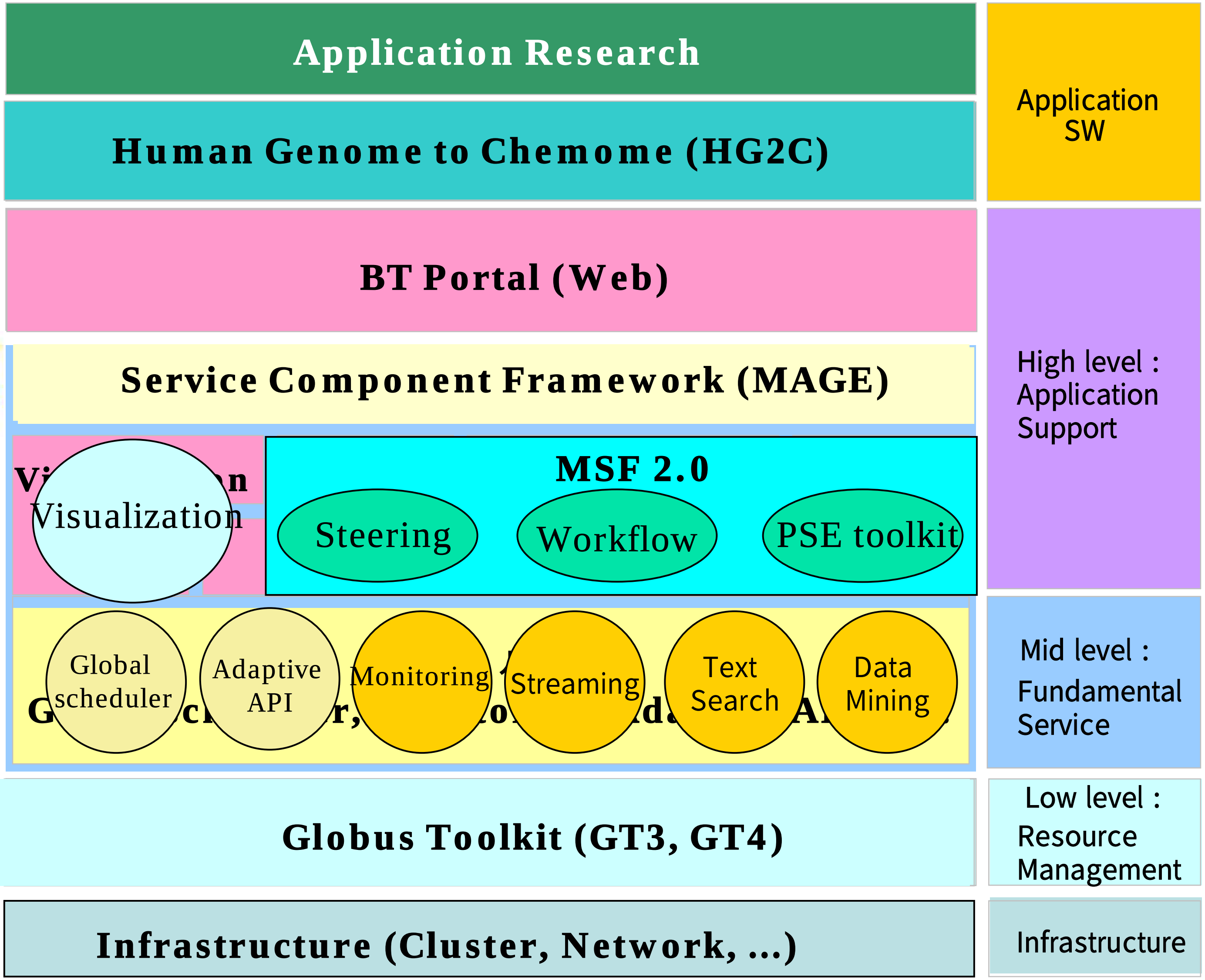
How to explore? - Directionally

- Directional processing
  - Interactive GUI for researcher decision :
    - Protein models, active sites VS queries, etc.
  - Conditional decision module :
    - Additional decision program

- MSF utility
  - User workspace management for multiple researcher
  - Batch processing management for various procedure

MSF and MAGE

HG2C System Architecture



Application Research

Human Genome to Chemome (HG2C)

BT Portal (Web)

Service Component Framework (MAGE)

Visualization

MSF 2.0

Steering

Workflow

PSE toolkit

Global scheduler

Adaptive API

Monitoring

Streaming

Text Search

Data Mining

Globus Toolkit (GT3, GT4)

Infrastructure (Cluster, Network, ...)

Application SW

High level : Application Support

Mid level : Fundamental Service

Low level : Resource Management

Infrastructure

Meta Services

- Provide reusable and adaptable workflow environments
- Define a part of a workflow as a new service
  - Workflow instance can be declared as a workflow unit in the service description
  - By overriding some attributes of a workflow unit,
    - Pass parameters of a service to the workflow's attributes
    - Setup service specific information
  - The new service can be wrapped to a Web service or a Grid service, therefore it can be easily reused
- Manage service specific information
  - Restrict resources to allocate a specific service (user's preference and/or organization's policy)
  - Schedule jobs with priority

Meta Services Framework (1)

- Meta Services Framework (MSF)
  - MSF is a workflow system for Bio Grid portal
  - Users can easily compose a DAG-based workflow using legacy applications such as a BLAST
  - Schedule user's workflows on Grid environments
  - Provide reusable workflows using Meta Services
  - Users can compose services, flows, and tasks using XML
  - Can be easily installed and configured

Meta Services Framework (2)

- MSF Workflow model
  - Workflow divided into three layers

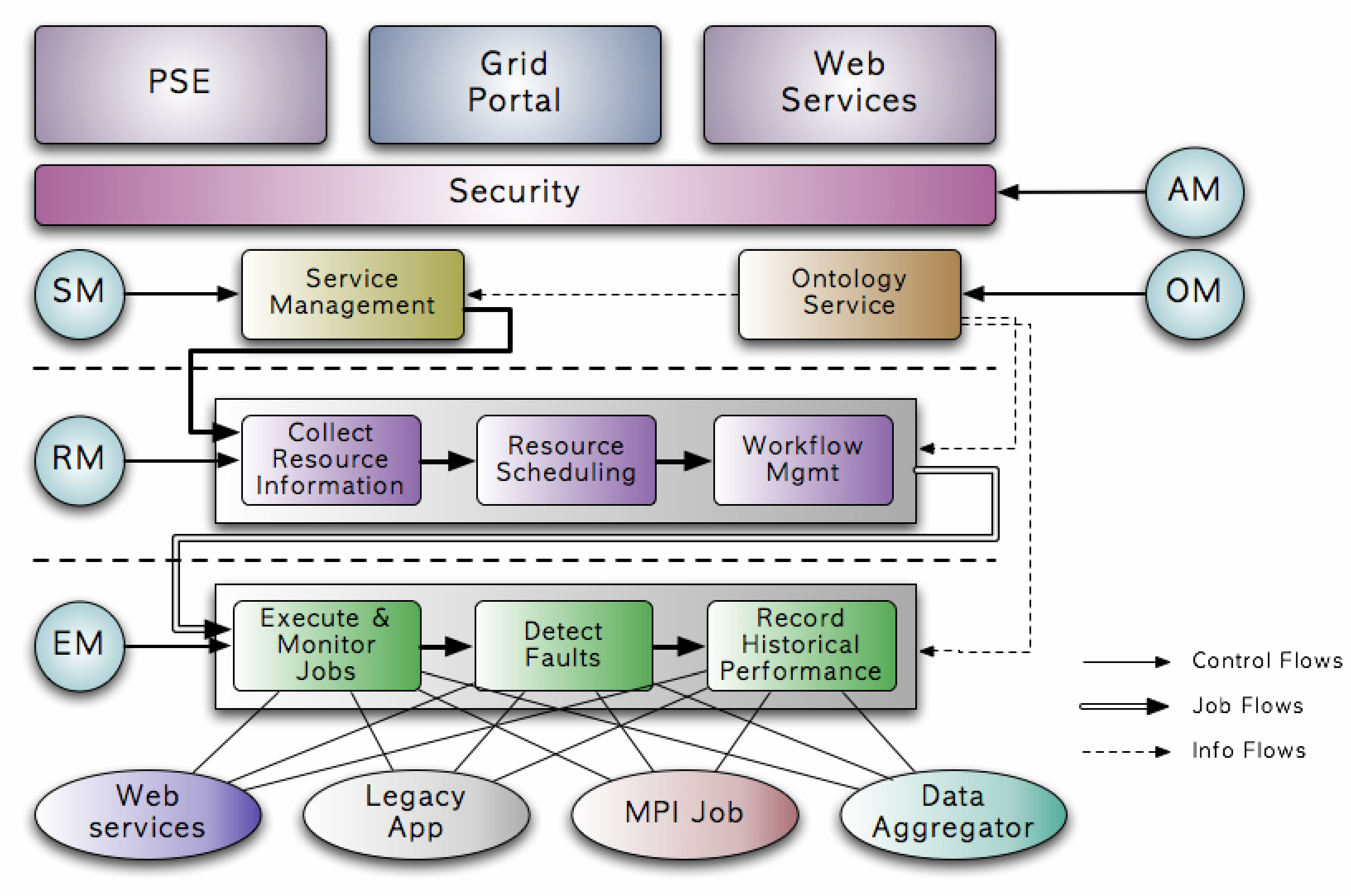
    - service layer, flow layer, and task layer
  - Increase reusability of workflow
  - Service layer is specially designed using Meta services concept

- MSF Middleware Architecture
  - Consist of five modules (SM, RM, EM, AM, OM)
  - Three agent modules process each layer of the workflow model (SM - service, RM - flow, EM - task)
  - AM manages authentication and access control
  - OM maintains XML description

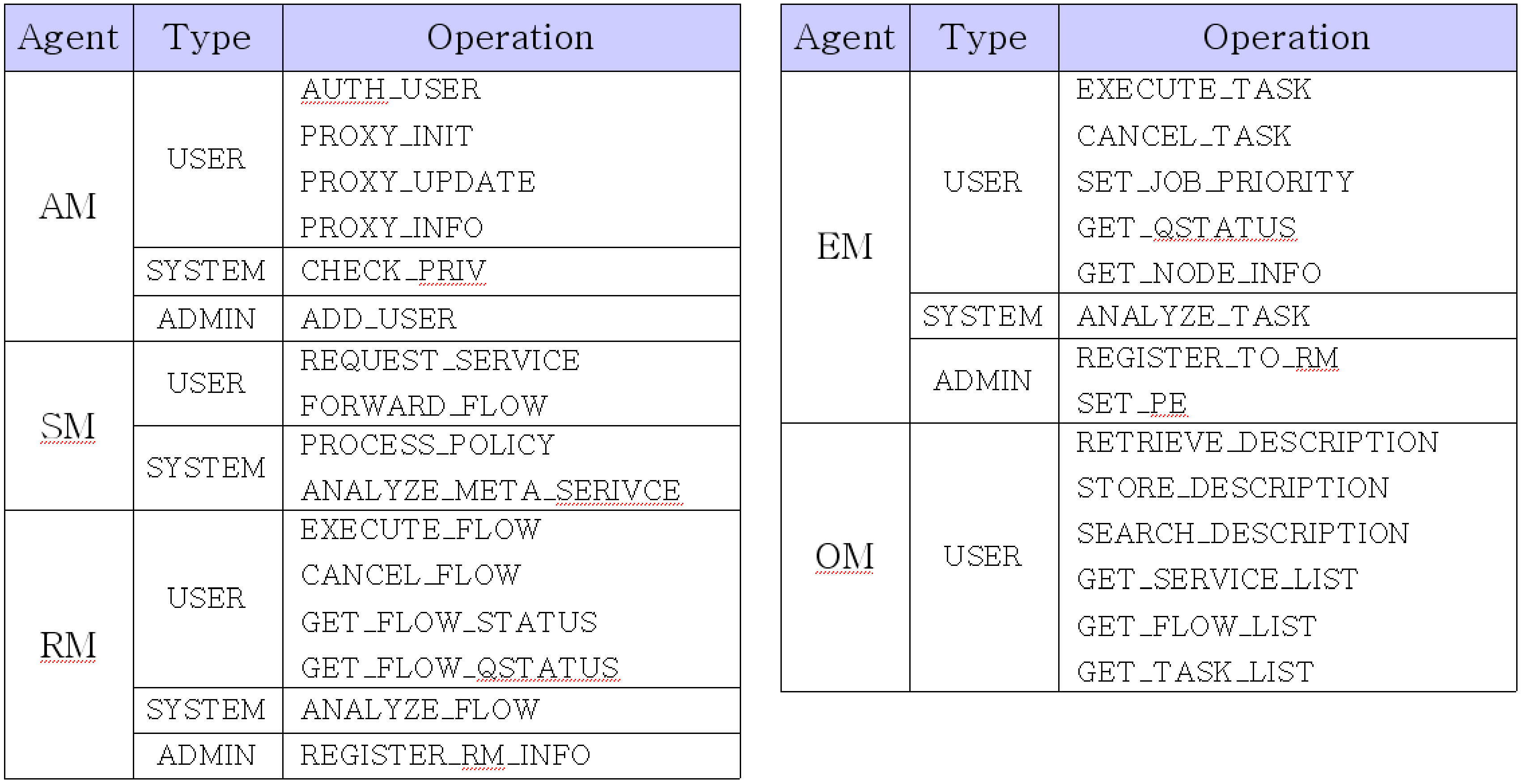
Five Agents in the Architecture

- SM (Service Manager) - service
  - Manage meta services and converts a meta service to a workflow
- RM (Resource Manager) - flow
  - Collect resource information, and allocating resources
- EM (Execution Manager) - task
  - Launch & monitor (workflow) jobs, detect faults, and collect results and performance data
- AM (Access Manager)
  - User authentication, environment setup, and a job submission service
- OM (Ontology Manager)
  - Manage ontology of service, flow, and task

Middleware Architecture



Operations defined in MSF



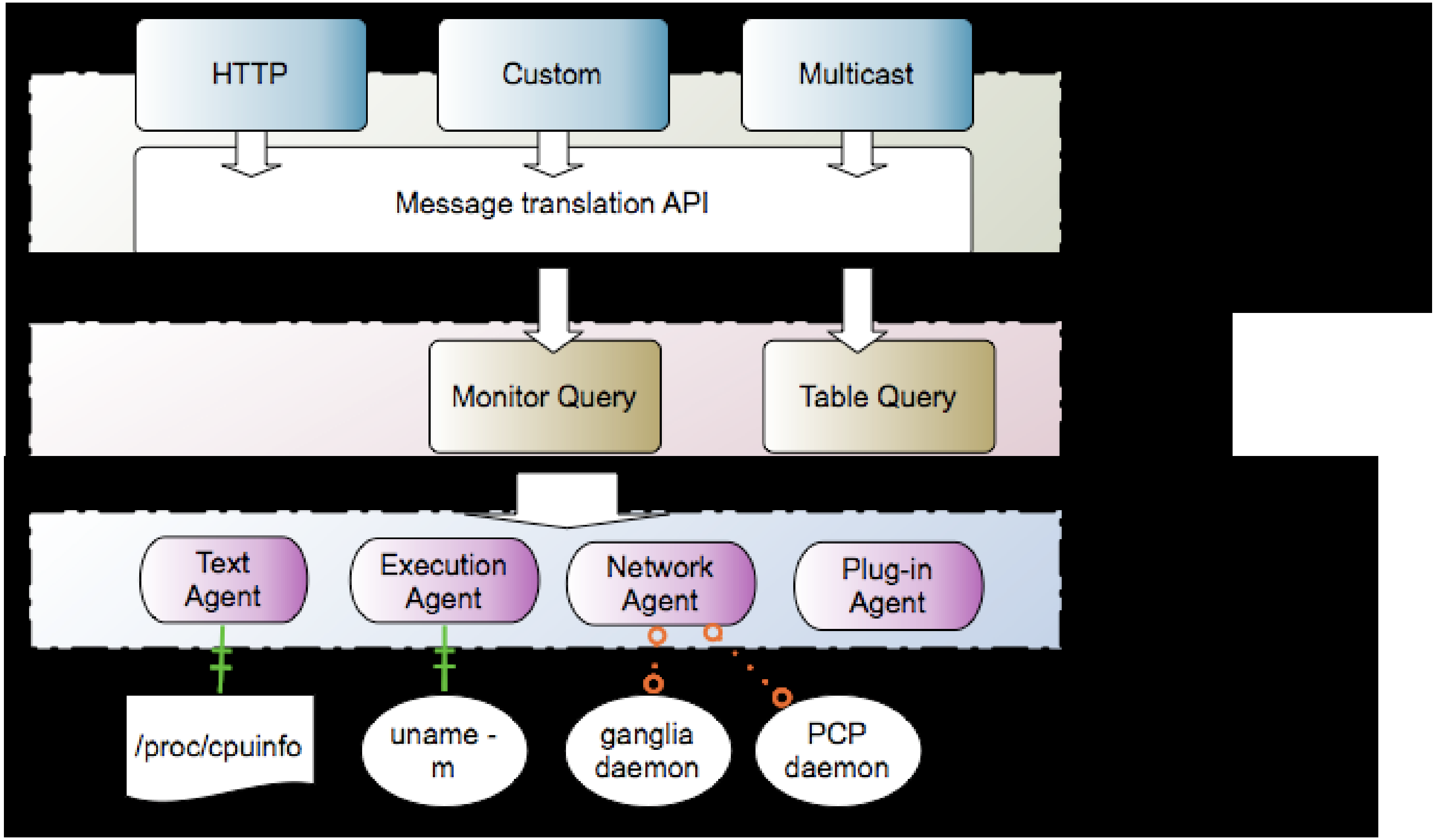
Agent	Type	Operation
AM	USER	AUTH_USER PROXY_INIT PROXY_UPDATE PROXY_INFO
	SYSTEM	CHECK_PRIV
	ADMIN	ADD_USER
SM	USER	REQUEST_SERVICE FORWARD_FLOW
	SYSTEM	PROCESS_POLICY ANALYZE_META_SERIVCE
RM	USER	EXECUTE_FLOW CANCEL_FLOW GET_FLOW_STATUS GET_FLOW_QSTATUS
	SYSTEM	ANALYZE_FLOW
	ADMIN	REGISTER_RM_INFO

Agent	Type	Operation
EM	USER	EXECUTE_TASK CANCEL_TASK SET_JOB_PRIORITY GET_QSTATUS GET_NODE_INFO
	SYSTEM	ANALYZE_TASK
	ADMIN	REGISTER_TO_RM SET_PE
OM	USER	RETRIEVE_DESCRIPTION STORE_DESCRIPTION SEARCH_DESCRIPTION GET_SERVICE_LIST GET_FLOW_LIST GET_TASK_LIST

MAGE features

- Provide API for easy development of Grid application
- Provide transparency to end-users and developers
  - Protocol transparency
  - Running location transparency
  - Message interpret transparency
- Provide layered architecture for easy replacement
- Mobility for each agent

MAGE architecture



Communication Layer

- To provide easy replacement of the communication protocol without affecting other layers
  - Administrator can select suitable protocol components before running application
  - Tasks and message interpretation does not affected by changing of communication protocol

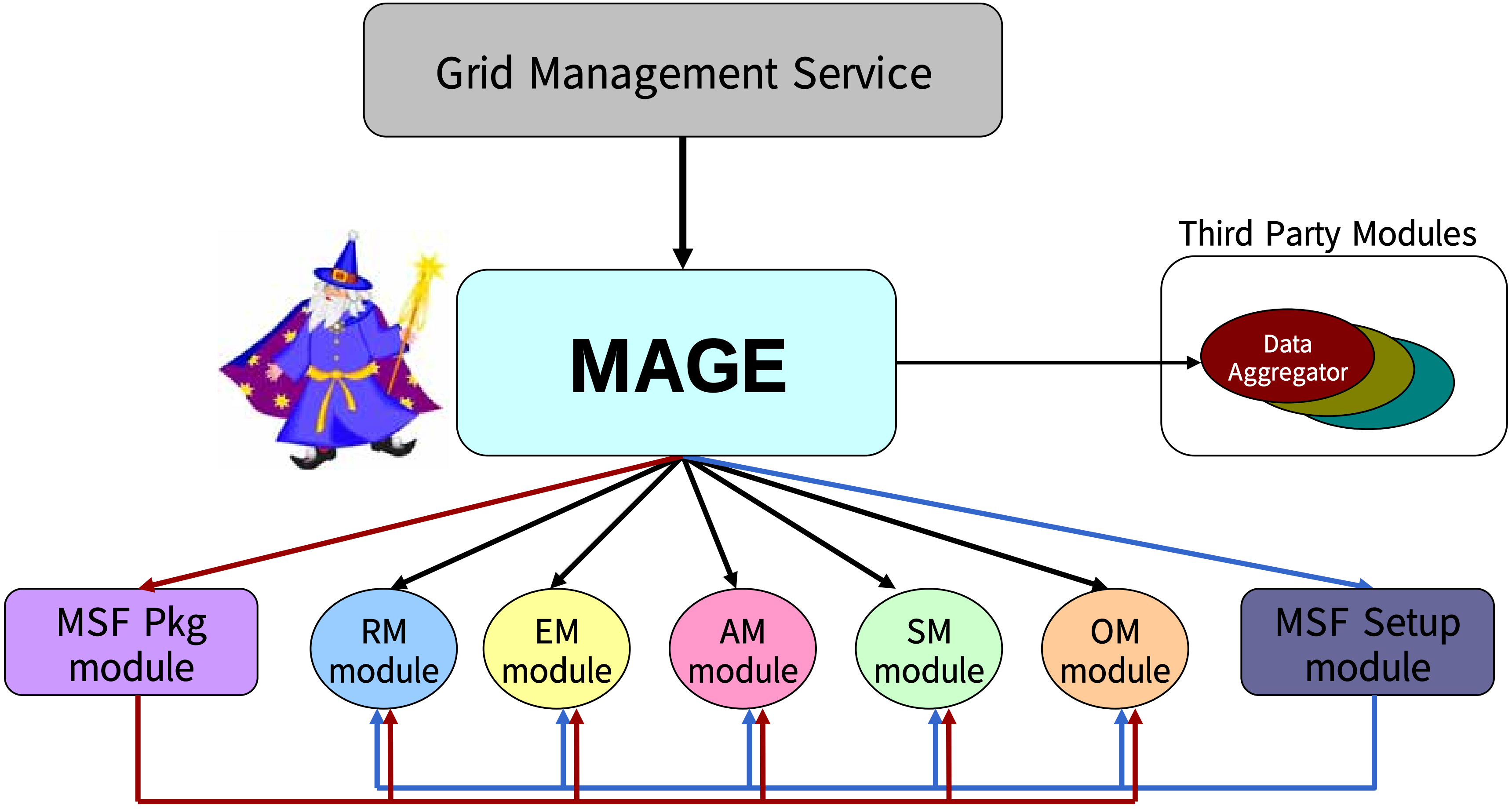
Interpreter Layer

- Interpret received messages and deliver to the appropriate task agents
- Two basic query components implemented
  - Monitor Query: use SQL’s SELECT-like statement
  - Table Query: use name=value pair table

Task Agent Management Layer

- Control the life cycle of task agents
- Provide feature of installation from remote
  - Base function for mobile agents
  - Provide function of stop the job and resume at another node

MSF Management using MAGE



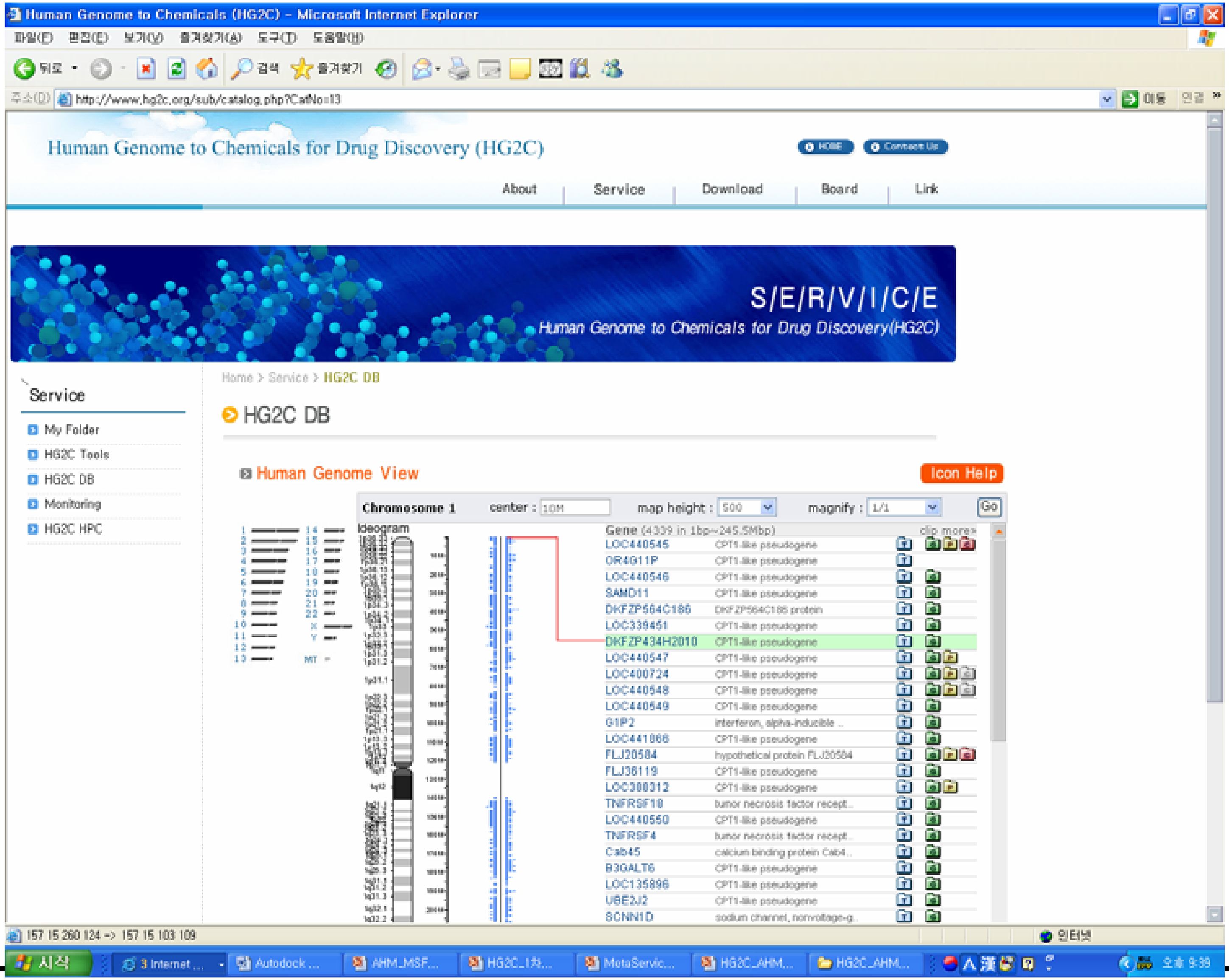
Distribute & Deploy MSF module Setup MSF Modules to cooperate

Conclusion

- MSF contains essential functionalities for Grid portals
  - workflow, service interface, job distribution, and parameter scheduling
- Distributing and deploying MSF modules using MAGE increases reconfigurability and adaptability to MSF modules
  - MSF modules can used as agents
  - MAGE environments can provides various services to MSF modules

    → Resource information, System environments, ..

HG2C Portal



http://www.hg2c.org/sub/catalog.php?CatNo=13

Human Genome to Chemicals for Drug Discovery (HG2C)

Home > Service > HG2C DB

HG2C DB

Human Genome View

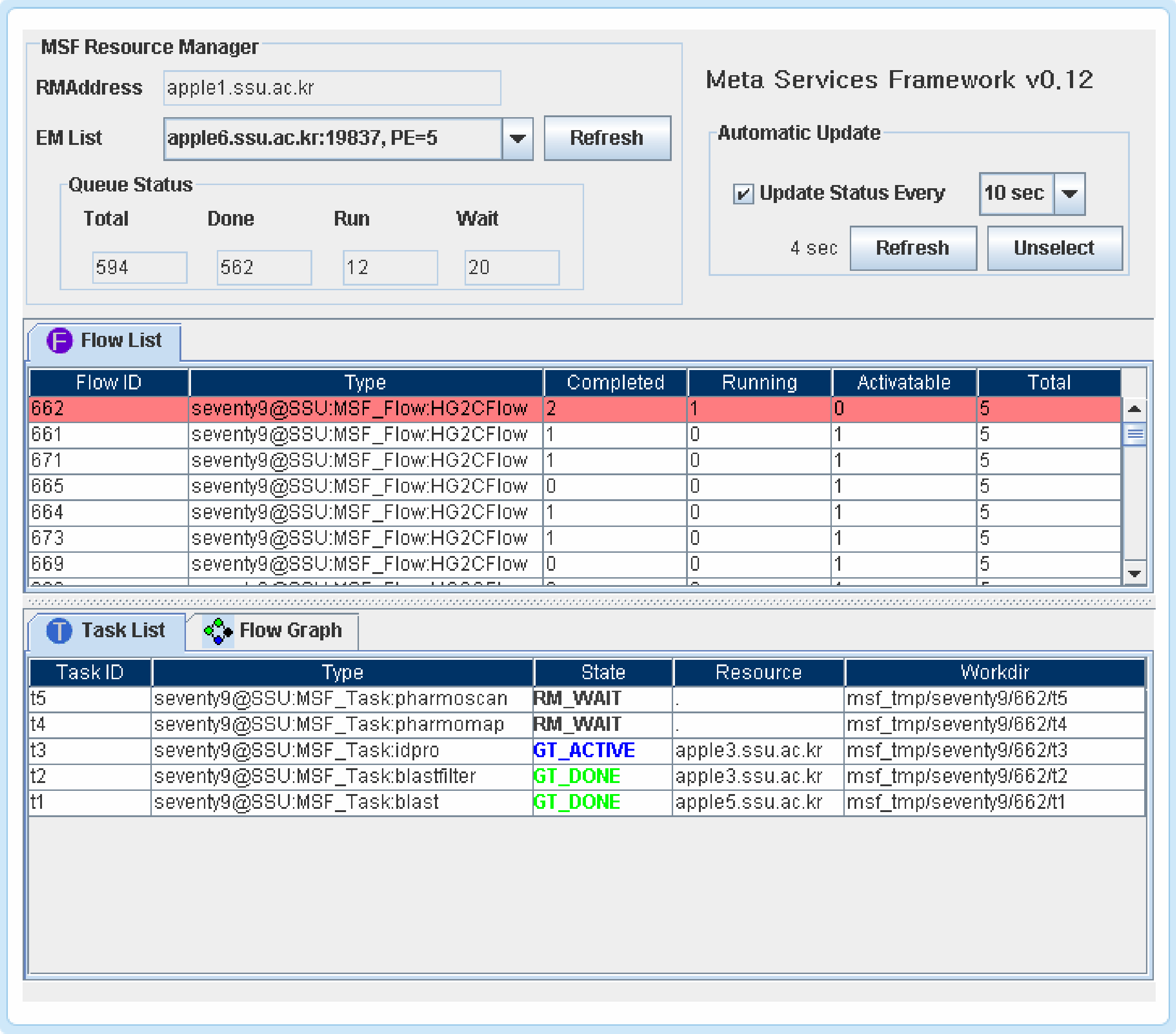
Chromosome 1 center : 10M map height : 500 magnify : 1/1

Gene (4339 in 1bp~245.5Mbp)

LOC440545	CPT1-like pseudogene
OR4G11P	CPT1-like pseudogene
LOC440546	CPT1-like pseudogene
SAMD11	CPT1-like pseudogene
DKFZP564C186	DKFZP564C186 protein
LOC339451	CPT1-like pseudogene
DKFZP434H2010	CPT1-like pseudogene
LOC440547	CPT1-like pseudogene
LOC400724	CPT1-like pseudogene
LOC440548	CPT1-like pseudogene
LOC440549	CPT1-like pseudogene
G1P2	interferon, alpha-inducible ..
LOC441866	CPT1-like pseudogene
FLJ20584	hypothetical protein FLJ20584
FLJ36119	CPT1-like pseudogene
LOC388312	CPT1-like pseudogene
TNFRSF18	tumor necrosis factor recept..
LOC440550	CPT1-like pseudogene
TNFRSF4	tumor necrosis factor recept..
Cab45	calcium binding protein Cab4..
B3GALT6	CPT1-like pseudogene
LOC135896	CPT1-like pseudogene
UBE2J2	CPT1-like pseudogene
SCNN1D	sodium channel, nonvoltage-g..

157 15 260 124 -> 157 15 103 109

Meta Services Framework Demo (1)



MSF Resource Manager

RMAddress: apple1.ssu.ac.kr

EM List: apple6.ssu.ac.kr:19837, PE=5

Queue Status

Total	Done	Run	Wait
594	562	12	20

Meta Services Framework v0.12

Automatic Update

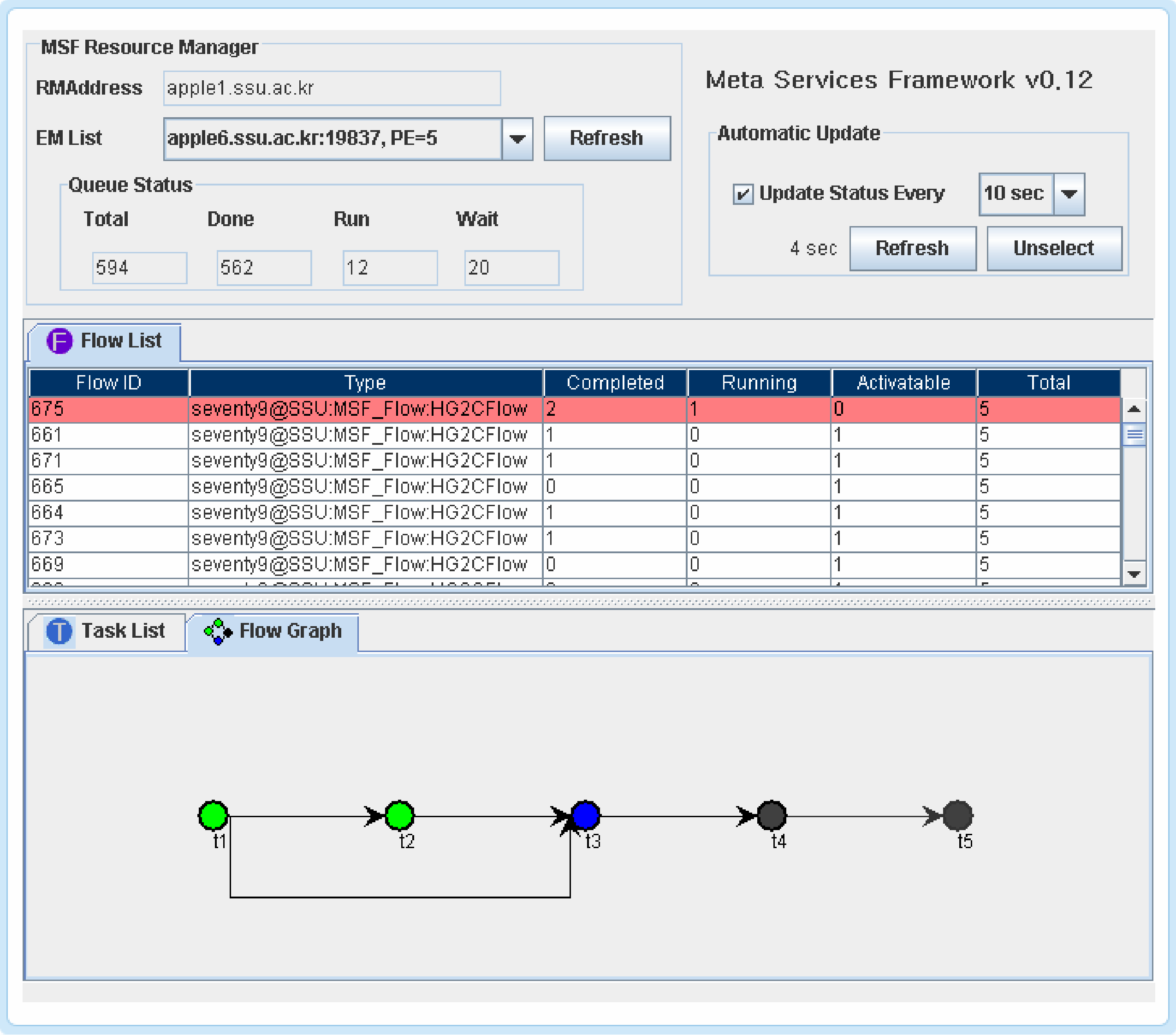
☑ Update Status Every: 10 sec

4 sec

Flow ID	Type	Completed	Running	Activatable	Total
662	seventy9@SSU:MSF_Flow:HG2CFlow	2	1	0	5
661	seventy9@SSU:MSF_Flow:HG2CFlow	1	0	1	5
671	seventy9@SSU:MSF_Flow:HG2CFlow	1	0	1	5
665	seventy9@SSU:MSF_Flow:HG2CFlow	0	0	1	5
664	seventy9@SSU:MSF_Flow:HG2CFlow	1	0	1	5
673	seventy9@SSU:MSF_Flow:HG2CFlow	1	0	1	5
669	seventy9@SSU:MSF_Flow:HG2CFlow	0	0	1	5

Task ID	Type	State	Resource	Workdir
t5	seventy9@SSU:MSF_Task:pharmoscan	RM_WAIT	.	msf_tmp/seventy9/662/t5
t4	seventy9@SSU:MSF_Task:pharmomap	RM_WAIT	.	msf_tmp/seventy9/662/t4
t3	seventy9@SSU:MSF_Task:idpro	GT_ACTIVE	apple3.ssu.ac.kr	msf_tmp/seventy9/662/t3
t2	seventy9@SSU:MSF_Task:blastfilter	GT_DONE	apple3.ssu.ac.kr	msf_tmp/seventy9/662/t2
t1	seventy9@SSU:MSF_Task:blast	GT_DONE	apple5.ssu.ac.kr	msf_tmp/seventy9/662/t1

Meta Services Framework Demo (2)



MSF Resource Manager

RMAddress: apple1.ssu.ac.kr

EM List: apple6.ssu.ac.kr:19837, PE=5

Queue Status

Total	Done	Run	Wait
594	562	12	20

Meta Services Framework v0.12

Automatic Update

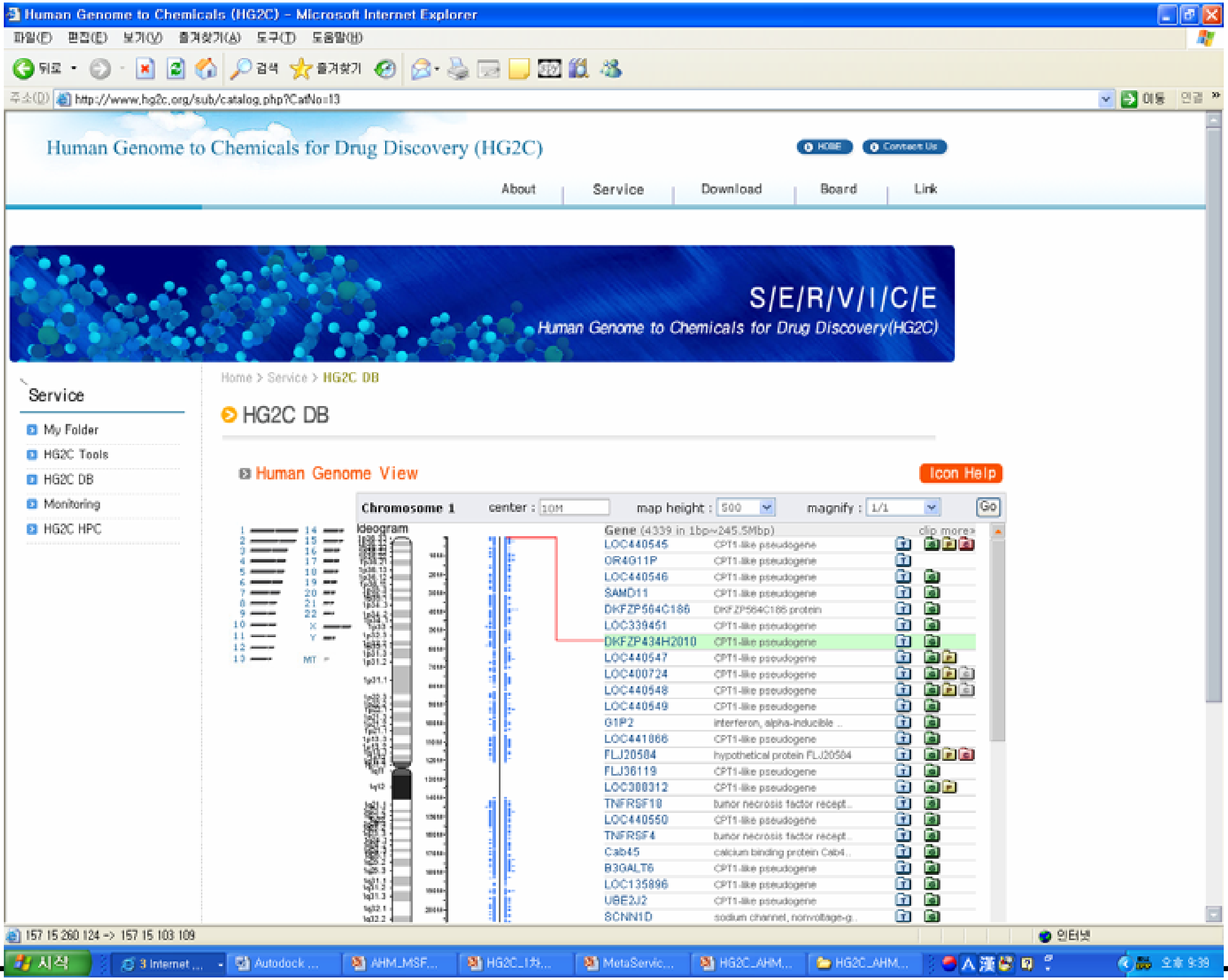
☑ Update Status Every 10 sec

4 sec

Flow List

Flow ID	Type	Completed	Running	Activatable	Total
675	seventy9@SSU:MSF_Flow:HG2CFlow	2	1	0	5
661	seventy9@SSU:MSF_Flow:HG2CFlow	1	0	1	5
671	seventy9@SSU:MSF_Flow:HG2CFlow	1	0	1	5
665	seventy9@SSU:MSF_Flow:HG2CFlow	0	0	1	5
664	seventy9@SSU:MSF_Flow:HG2CFlow	1	0	1	5
673	seventy9@SSU:MSF_Flow:HG2CFlow	1	0	1	5
669	seventy9@SSU:MSF_Flow:HG2CFlow	0	0	1	5

HG2C Pubmap



Human Genome to Chemicals for Drug Discovery (HG2C)

Human Genome to Chemicals for Drug Discovery(HG2C)

HG2C DB

Human Genome View

Chromosome 1 center : 10M map height : 500 magnify : 1/1

Gene (4339 in 1bp~245.5Mbp)	
LOC440545	CPT1-like pseudogene
OR4G11P	CPT1-like pseudogene
LOC440546	CPT1-like pseudogene
SAMD11	CPT1-like pseudogene
DKFZP564C186	DKFZP564C186 protein
LOC339451	CPT1-like pseudogene
DKFZP434H2010	CPT1-like pseudogene
LOC440547	CPT1-like pseudogene
LOC400724	CPT1-like pseudogene
LOC440548	CPT1-like pseudogene
LOC440549	CPT1-like pseudogene
G1P2	interferon, alpha-inducible ..
LOC441866	CPT1-like pseudogene
FLJ20504	hypothetical protein FLJ20504
FLJ36119	CPT1-like pseudogene
LOC388312	CPT1-like pseudogene
TNFRSF18	tumor necrosis factor recept..
LOC440550	CPT1-like pseudogene
TNFRSF4	tumor necrosis factor recept..
Cab45	calcium binding protein Cab4..
B3GALT6	CPT1-like pseudogene
LOC135896	CPT1-like pseudogene
UBE2J2	CPT1-like pseudogene
SCNN1D	sodium channel, nonvoltage-g..

HG2C References

- HG2C Homepage
  - http://www.hg2c.org

- Open sources of MSF & MAGE
  - http:// www.sourceforge.net/projects/mage4ubi
  - http:// www.sourceforge.net/projects/msf

Q & A