Biosphere: the interoperation of web services in microarray cluster analysis

Kei-Hoi Cheung[1,2,*], Remko de Knikker[1], Youjun Guo[1], Guoneng Zhong[1], Janet Hager[3,4], Kevin Y. Yip[5], Albert K.H. Kwan[5], Peter Li[6], and David W. Cheung[5]

[1]Center for Medical Informatics, [2]Department of Genetics, [3]Keck Biotechnology Resource Laboratory, [4]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, USA, [5]Department of Computer and Information Systems, Pokfulam Road, University of Hong Kong, Hong Kong, [6]School of Computing Science, University of Newcastle, NE1 7RU, UK.

ABSTRACT

Summary: The growing use of DNA microarrays in biomedical research has led to the proliferation of analysis tools. These software programs address different aspects of analysis (e.g., normalization and clustering within and across individual arrays) as well as extended analysis methods (e.g., clustering, annotation and mining of multiple data sets). Microarray data analysis therefore typically requires the interoperability of multiple software programs involving different analysis types and methods. Such interoperation is often hampered by the heterogeneity involved in the software tools (which may function by implementing different interfaces and using different programming languages). To address this problem, we employ the SOAP-based web service approach that provides a uniform programmatic interface to these heterogeneous software components. To demonstrate this approach in the microarray context, we have created a web server application, **Biosphere**, which interoperates a number of web services that are geographically widely distributed. These web services include a **clustering web service**, that is a suite of different clustering algorithms for analyzing microarray data, **XEMBL** developed at EBI for retrieving EMBL sequence data, and three gene annotation web services: **getGO**, **getHAPI**, and **getUMLS**. While **getGO** allows retrieval of Gene Ontology (GO) annotation, the other two web services retrieve annotation from the biomedical literature that is indexed based on the Medical Subject Headings (MeSH) terms. With these web services, the application allows the users to do the following: i) cluster gene expression data using seven different algorithms, ii) visualize the clustering results statistically grouped in color, and iii) retrieve sequence, annotation and citation data for the genes of interest.

Availability: The web server application and its web services described in WSDL can be accessed at: http://rook.cecid.hku.hk:8280/BiosphereServer or https://rook.cecid.hku.hk:8443/BiosphereServer (this works for browsers such as Internet Explorer 6 or higher that supports secure data encryption)

Contact: kei.cheung@yale.edu

INTRODUCTION

Interpretation of microarray gene expression data utilizes many different types of computational and statistical analyses methods (e.g., normalization, clustering, etc). Despite the fact that a large repository of software tools have been developed and made

[*] To whom correspondence should be addressed.

available to the microarray community (Dudoit et al. 2003), researchers are confronted with the following problems.

1. Identifying and locating available tools may be difficult since they are widely scattered and developed by geographically disperse groups.
2. Some types of microarray analysis may require large amounts of memory or computational power especially when large quantities of data are involved. Researchers may not have adequate local computational resources to perform such analyses.
3. Tools available by download are often written in different languages and may not be cross-platform-compatible.
4. Some tools provide a web server interface that allows users to submit their data over the Internet for processing. While this type of interface is friendly to biological users, it is difficult for bioinformatics developers to access programmatically.
5. It is often difficult to orchestrate a specific set of bioinformatics tools from diverse sources for a particular application due to their incompatibility and heterogeneity.

To address these problems, we employ web services, which are implemented based on the XML-based Simple Object Access Protocol (SOAP), as a uniform model to provide a common programmatic interface to heterogeneous software applications in the bioinformatics context. There have been a growing number of SOAP web services available for bioinformatics use. This includes XEMBL (Wang et al. 2002), Soaplab (http://industry.ebi.ac.uk/soaplab), DDBJ Soap Server (http://xml.ddbj.nig.ac.jp/soapp.html), and so on. As pointed out by Stein (Stein 2002), web services have the potential of unifying the “bioinformatics nation”. We demonstrate how web services can be integrated to create a single-entry portal that is broadly accessible to researchers for data analysis involving multiple programs and databases.

WEB SERVICES APPLICATION

A wide variety of clustering algorithms have been used to cluster genome-wide expression patterns (Eisen et al. 1998). In addition to traditional methods, new approaches to clustering gene expression data have been developed (e.g., HARP (Yip et al. 2004)). It has become important to evaluate and compare these methods. In (Datta et al. 2003), a number of algorithms (including hierarchical, K-means, Diana, fanny, and model-based methods) were compared in terms of their performance in gene clustering. To access and use these clustering programs may not be straightforward since they are scattered and available in different formats (some of them are platform-specific). For example, Eisen’s cluster program (http://rana.lbl.gov/EisenSoftware.htm) can only be run on Windows. EPCLUST (http://ep.ebi.ac.uk/EP/EPCLUST/) is a web server application allowing clustering, visualization and analysis of gene expression data, but it lacks a programmatic interface, making automatic interoperation with other programs difficult. To demonstrate how these issues are addressed, we have built a web server application, **Biosphere**, which interoperates a number of web services. **Biosphere** allows different algorithms to be used to cluster gene expression datasets uploaded by individual users via the Web. The algorithms include both hierarchical and non-hierarchical methods, which are programmatically accessible through a “cluster” web service. In addition, we use the following web services to retrieve annotation and sequences of the clustered genes.

1. **getHAPI:** This is the web service counterpart of HAPI (Masys et al. 2001), which extracts MeSH terms represented by the Unified Medical Language System (UMLS) concept IDs from the biomedical literature.
2. **getUMLS:** This web service builds a hierarchy of UMLS concepts for a set of concept IDs.
3. **getGO:** This web service retrieves genome annotation (human only) from the Gene Ontology database (http://www.geneontology.org/).
4. **XEMBL:** This web service was developed at EBI (http://www.ebi.ac.uk), which allows retrieval of sequence data from EMBL.

Extensibility is an important property of web-service based analyzing systems. By defining standard service interfaces through Web Service Description Language (WSDL), new algorithms and modules (e.g. data preprocessing and charting) can easily be implemented by any parties and added to an existing system. Additional data mining algorithm implementations can be plugged into the system easily. The clustering and annotation web services are implemented and deployed using Axis that is the Java implementation of SOAP on Apache. The underlying programs are written in Java. To increase performance, the clustering programs run on a two-node cluster (with each node being an Intel Pentium 4 processor with 1 GB memory and 40 GB hard disk space). In addition, the system notifies the users by e-mail when clustering is done. This feature is especially useful for clustering large datasets. For security, Biosphere provides the option of using the Secure Sockets Layer (SSL) to encrypt data transmitted over the Internet.

Fig. 1 shows the colored cluster output display generated based on the Diana algorithm asking for genes (not samples) to be clustered into six groups based on expression ratio (together with the GO and HAPI annotation). This represents a subset of data generated from a human microarray experiment using 2 μgmRNAfrom HeLa and HL60 cell lines that was amino-allyl indirect cy-dye labeled using oligo dT primer and reverse transcription. The array data input for Biosphere are three technical hybridization replicates (ITIK_HUCOMP3V2C-319U, ITIK_HUCOMP3V2C-319L, ITIK_HUCOMP3V2C-320U). The array used (HUCOMP3V2C-300-400) is a composite array of both cDNA and oligonucleotide targets. The color difference of a gene across different samples is totally dependent on the expression values of the gene in the input data file. A perfect cluster is a group of genes that have zero distances from each other, i.e., in each sample, all the genes have the same expression value.

ACKNOWLEDGEMENTS

This work was supported in part by NIH grants K25 HG02378 from the National Human Genome Research Institute, NIH 5 U24 DK58776 from the National Institute of Diabetes and Digestive and Kidney Diseases and Anna Hull Foundation, and NSF grant BDI-0135442. The work of the authors in the University of Hong Kong was supported in part by a grant from the University Grant Council of Hong Kong. The authors would like to thank Tom Oinn at EBI for testing and using our web services and providing feedback.

REFERENCES

Datta, S. and S. Datta (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics **19**(4):459-466.

Dudoit, S., R. Gentleman and J. Quackenbush (2003). Open source software for the analysis of microarray data. Biotechniques **Suppl**:45-51.

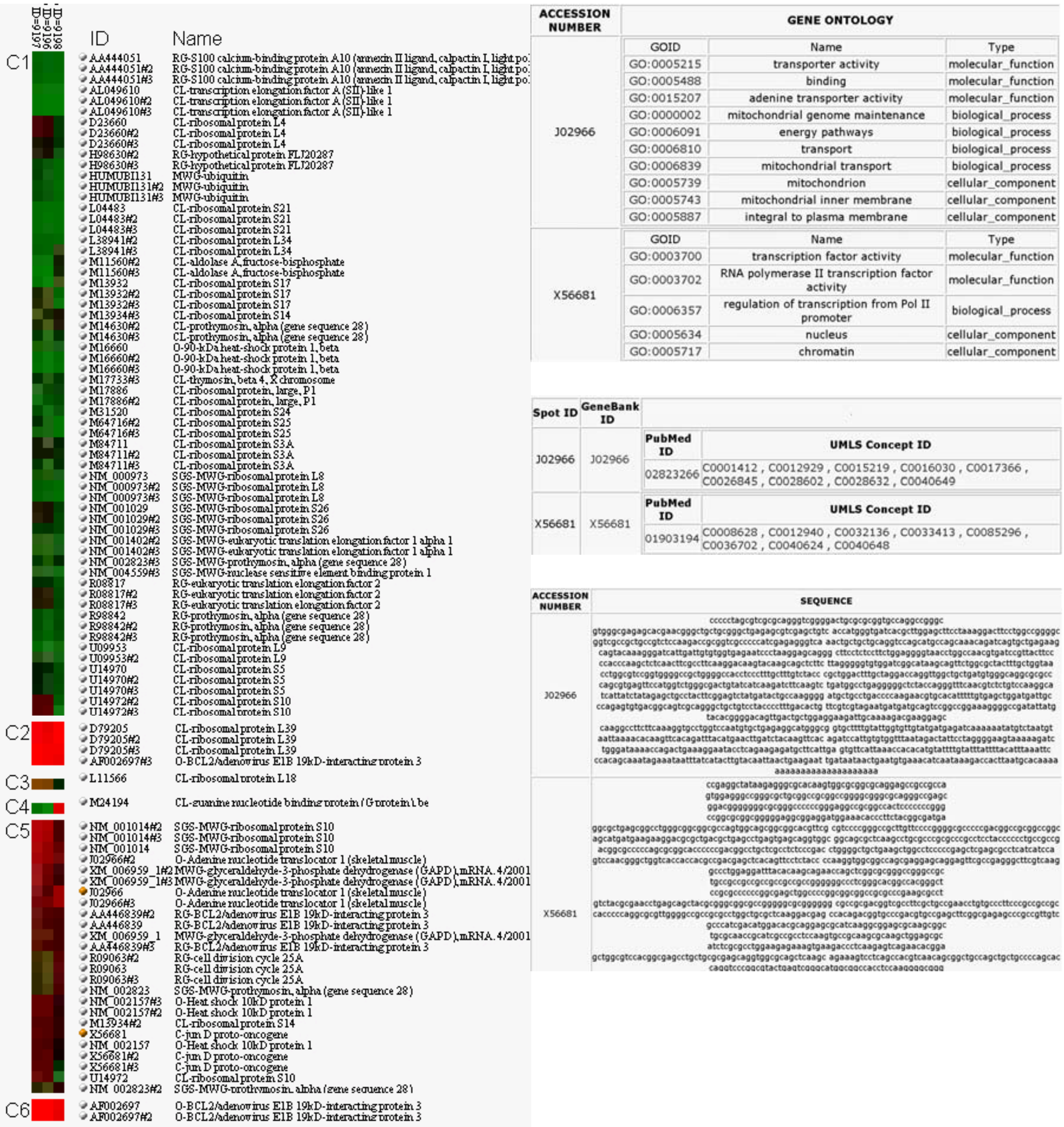
Eisen, M., P. Spellman, P. Brown and D. Botstein (1998). Cluster analysis and display of genome-wide expression patterns. PNAS **95**(25):14863-8.

Masys, D., J. Welsh, J. L. Fink, M. Gribskov, I. Klacansky and J. Corbeil (2001). Use of keyword hierarchies to interpret gene expression patterns. Bioinformatics **17**(4):319-26.

Stein, L. (2002). Creating a bioinformatics nation. Nature **417**:119-20.

Wang, L., J. Riethoven and A. Robinson (2002). XEMBL: distributing EMBL data in XML format. Bioinformatics **18**(8):1147-8.

Yip, K.Y., D.W. Cheung, M.K. Ng and K.H. Cheung (2004). Identifying projected clusters from gene expression profiles. IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE), pp. 259-66.



	ID	Name
C1	○ AA444051	RG-S100 calcium-binding protein A10 (annexin II ligand, calpactin I, light po
	○ AA444051#2	RG-S100 calcium-binding protein A10 (annexin II ligand, calpactin I, light po
	○ AA444051#3	RG-S100 calcium-binding protein A10 (annexin II ligand, calpactin I, light po
	○ AL049610	CL-transcription elongation factor A (SII)-like 1
	○ AL049610#2	CL-transcription elongation factor A (SII)-like 1
	○ AL049610#3	CL-transcription elongation factor A (SII)-like 1
	○ D23660	CL-ribosomal protein L4
	○ D23660#2	CL-ribosomal protein L4
	○ D23660#3	CL-ribosomal protein L4
	○ H98630#2	RG-hypothetical protein FLJ20287
	○ H98630#3	RG-hypothetical protein FLJ20287
	○ HUMUBI131	MWG-ubiquitin
	○ HUMUBI131#2	MWG-ubiquitin
	○ HUMUBI131#3	MWG-ubiquitin
	○ L04483	CL-ribosomal protein S21
	○ L04483#2	CL-ribosomal protein S21
	○ L04483#3	CL-ribosomal protein S21
	○ L38941#2	CL-ribosomal protein L34
	○ L38941#3	CL-ribosomal protein L34
	○ M11560#2	CL-aldolase A, fructose-bisphosphate
	○ M11560#3	CL-aldolase A, fructose-bisphosphate
	○ M13932	CL-ribosomal protein S17
	○ M13932#2	CL-ribosomal protein S17
	○ M13932#3	CL-ribosomal protein S17
	○ M13934#3	CL-ribosomal protein S14
	○ M14630#2	CL-prothymosin, alpha (gene sequence 28)
	○ M14630#3	CL-prothymosin, alpha (gene sequence 28)
	○ M16660	O-90-kDa heat-shock protein 1, beta
	○ M16660#2	O-90-kDa heat-shock protein 1, beta
	○ M16660#3	O-90-kDa heat-shock protein 1, beta
	○ M17733#3	CL-thymosin, beta 4, X chromosome
	○ M17886	CL-ribosomal protein, large, P1
	○ M17886#2	CL-ribosomal protein, large, P1
	○ M31520	CL-ribosomal protein S24
	○ M64716#2	CL-ribosomal protein S25
	○ M64716#3	CL-ribosomal protein S25
	○ M84711	CL-ribosomal protein S3A
	○ M84711#2	CL-ribosomal protein S3A
	○ M84711#3	CL-ribosomal protein S3A
	○ NM_000973	SGS-MWG-ribosomal protein L8
	○ NM_000973#2	SGS-MWG-ribosomal protein L8
	○ NM_000973#3	SGS-MWG-ribosomal protein L8
	○ NM_001029	SGS-MWG-ribosomal protein S26
	○ NM_001029#2	SGS-MWG-ribosomal protein S26
	○ NM_001029#3	SGS-MWG-ribosomal protein S26
	○ NM_001402#2	SGS-MWG-eukaryotic translation elongation factor 1 alpha 1
	○ NM_001402#3	SGS-MWG-eukaryotic translation elongation factor 1 alpha 1
	○ NM_002823#3	SGS-MWG-prothymosin, alpha (gene sequence 28)
	○ NM_004559#3	SGS-MWG-nuclease sensitive element binding protein 1
	○ R08817	RG-eukaryotic translation elongation factor 2
	○ R08817#2	RG-eukaryotic translation elongation factor 2
	○ R08817#3	RG-eukaryotic translation elongation factor 2
	○ R98842	RG-prothymosin, alpha (gene sequence 28)
	○ R98842#2	RG-prothymosin, alpha (gene sequence 28)
	○ R98842#3	RG-prothymosin, alpha (gene sequence 28)
	○ U09953	CL-ribosomal protein L9
	○ U09953#2	CL-ribosomal protein L9
	○ U14970	CL-ribosomal protein S5
	○ U14970#2	CL-ribosomal protein S5
	○ U14970#3	CL-ribosomal protein S5
	○ U14972#2	CL-ribosomal protein S10
	○ U14972#3	CL-ribosomal protein S10
C2	○ D79205	CL-ribosomal protein L39
	○ D79205#2	CL-ribosomal protein L39
	○ D79205#3	CL-ribosomal protein L39
	○ AF002697#3	O-BCL2/adenovirus E1B 19kD-interacting protein 3
C3	○ L11566	CL-ribosomal protein L18
C4	○ M24194	CL-guanine nucleotide binding protein (G protein) be
C5	○ NM_001014#2	SGS-MWG-ribosomal protein S10
	○ NM_001014#3	SGS-MWG-ribosomal protein S10
	○ NM_001014	SGS-MWG-ribosomal protein S10
	○ J02966#2	O-Adenine nucleotide translocator 1 (skeletal muscle)
	○ XM_006959_1#2	MWG-glyceraldehyde-3-phosphate dehydrogenase (GAPD), mRNA. 4/2001
	○ XM_006959_1#3	MWG-glyceraldehyde-3-phosphate dehydrogenase (GAPD), mRNA. 4/2001
	● J02966	O-Adenine nucleotide translocator 1 (skeletal muscle)
	○ J02966#3	O-Adenine nucleotide translocator 1 (skeletal muscle)
	○ AA446839#2	RG-BCL2/adenovirus E1B 19kD-interacting protein 3
	○ AA446839	RG-BCL2/adenovirus E1B 19kD-interacting protein 3
	○ XM_006959_1	MWG-glyceraldehyde-3-phosphate dehydrogenase (GAPD), mRNA. 4/2001
	○ AA446839#3	RG-BCL2/adenovirus E1B 19kD-interacting protein 3
	○ R09063#2	RG-cell division cycle 25A
	○ R09063	RG-cell division cycle 25A
	○ R09063#3	RG-cell division cycle 25A
	○ NM_002823	SGS-MWG-prothymosin, alpha (gene sequence 28)
	○ NM_002157#3	O-Heat shock 10kD protein 1
	○ NM_002157#2	O-Heat shock 10kD protein 1
	○ M13934#2	CL-ribosomal protein S14
	● X56681	C-jun D proto-oncogene
	○ NM_002157	O-Heat shock 10kD protein 1
	○ X56681#2	C-jun D proto-oncogene
	○ X56681#3	C-jun D proto-oncogene
	○ U14972	CL-ribosomal protein S10
	○ NM_002823#2	SGS-MWG-prothymosin, alpha (gene sequence 28)
C6	○ AF002697	O-BCL2/adenovirus E1B 19kD-interacting protein 3
	○ AF002697#2	O-BCL2/adenovirus E1B 19kD-interacting protein 3

ACCESSION NUMBER	GENE ONTOLOGY		
	GOID	Name	Type
J02966	GO:0005215	transporter activity	molecular_function
	GO:0005488	binding	molecular_function
	GO:0015207	adenine transporter activity	molecular_function
	GO:0000002	mitochondrial genome maintenance	biological_process
	GO:0006091	energy pathways	biological_process
	GO:0006810	transport	biological_process
	GO:0006839	mitochondrial transport	biological_process
	GO:0005739	mitochondrion	cellular_component
	GO:0005743	mitochondrial inner membrane	cellular_component
	GO:0005887	integral to plasma membrane	cellular_component
	GOID	Name	Type
X56681	GO:0003700	transcription factor activity	molecular_function
	GO:0003702	RNA polymerase II transcription factor activity	molecular_function
	GO:0006357	regulation of transcription from Pol II promoter	biological_process
	GO:0005634	nucleus	cellular_component
	GO:0005717	chromatin	cellular_component

Spot ID	GeneBank ID		
J02966	J02966	PubMed ID	UMLS Concept ID
		02823266	C0001412 , C0012929 , C0015219 , C0016030 , C0017366 , C0026845 , C0028602 , C0028632 , C0040649
X56681	X56681	PubMed ID	UMLS Concept ID
		01903194	C0008628 , C0012940 , C0032136 , C0033413 , C0085296 , C0036702 , C0040624 , C0040648

ACCESSION NUMBER	SEQUENCE
J02966	ccccctagcgtcgcgcagggtcggggactgcgcggtgccaggccgggc gtgggcgagagcacgaacgggctgctgcgggctgagagcgtcgagctgtc accatgggtgatcacgcttggagcttcctaaaggacttcctggccggggc ggtcgccgctgccgtctccaagaccgcggtcgcccccatcgagagggtca aactgctgctgcaggtccagcatgccagcaaacagatcagtgctgagaag cagtacaaagggatcattgattgtgtggtgagaatccctaaggagcaggg cttcctctccttctggaggggtaacctggccaacgtgatccgttacttcc ccacccaagctctcaacttcgccttcaaggacaagtacaagcagctcttc ttagggggtgtggatcggcataagcagttctggcgctactttgctggtaa cctggcgtccggtggggccgctggggccacctccctttgctttgtctacc cgctggactttgctaggaccaggttggctgctgatgtgggcaggcgcgcc cagcgtgagttccatggtctgggcgactgtatcatcaagatcttcaagtc tgatggcctgagggggctctaccagggtttcaacgtctctgtccaaggca tcattatctatagagctgcctacttcggagtctatgatactgccaagggg atgctgcctgaccccaagaacgtgcacatttttgtgagctggatgattgc ccagagtgtgacggcagtcgcagggctgctgtcctacccctttgacactg ttcgtcgtagaatgatgatgcagtccggccggaaaggggccgatattatg tacacggggacagttgactgctggaggaagattgcaaaagacgaaggagc caaggccttcttcaaaggtgcctggtccaatgtgctgagaggcatgggcg gtgcttttgtattggtgttgtatgatgagatcaaaaaatatgtctaatgt aattaaaacacaagttcacagatttacatgaacttgatctacaagttcac agatccattgtgtggtttaatagactattcctaggggaagtaaaaagatc tgggataaaaccagactgaaaggaatacctcagaagagatgcttcattga gtgttcattaaaccacacatgtattttgtatttattttacatttaaattc ccacagcaaatagaaataatttatcatacttgtacaattaactgaagaat tgataataactgaatgtgaaacatcaataaagaccacttaatgcacaaaa aaaaaaaaaaaaaaaaaaaa
X56681	ccgaggctataagagggcgcacaagtggcgcggcgcaggagccgccgca gtggagggccgggcgctgcggccgcggccggggcgggcgcagggccgagc ggacgggggggcgcgggcccccgggaggccgcggccactcccccccggg ccggcgcggcgggggaggcggaggatggaaacacccttctacggcgatga ggcgctgagcggcctgggcggcggcgccagtggcagcggcggcacgttcg cgtccccgggccgcttgttccccggggcgcccccgacggccgcggccggc agcatgatgaagaaggacgcgctgacgctgagcctgagtgagcaggtggc ggcagcgctcaagcctgcgcccgcgcccgcctcctaccccctgccgccg acggcgcccccagcgcggcacccccgacggcctgctcgcctctcccgac ctggggctgctgaagctggcctcccccgagctcgagcgcctcatcatcca gtccaacgggctggtcaccaccacgccgacgagctcacagttcctctacc ccaaggtggcggccagcgaggagcaggagttcgccgagggcttcgtcaag gccctggaggatttacacaagcagaaccagctcggcgcgggccgggccgc tgccgccgccgccgccgccgggggccctcgggcacggccacgggct ccgcgcccccggcgagctggccccggcggcggccgcgcccgaagcgcct gtctacgcgaacctgagcagctacgcgggcggcgccgggggcgcggggg cgccgcgacggtcgccttcgctgccgaacctgtgcccttcccgccgccgc cacccccaggcgcgttggggccgccgcgcctggctgcgctcaaggacgag ccacagacggtgcccgacgtgccgagcttcggcgagagcccgccgttgtc gcccatcgacatggacacgcaggagcgcatcaaggcggagcgcaagcggc tgcgcaaccgcatcgccgcctccaagtgccgcaagcgcaagctggagcgc atctcgcgcctggaagagaaagtgaagaccctcaagagtcagaacacgga gctggcgtccacggcgagcctgctgcgcgagcaggtggcgcagctcaagc agaaagtcctcagccacgtcaacagcggctgccagctgctgccccagcac caggtcccggcgtactgagtcgggcatggcggccacctccaaggggcggg

Fig. 1. Colored display of Diana cluster output together with GO annotation (top right), HAPI results (middle right), and XEMBL sequences (bottom right) for the selected genes indicated by the yellow radio buttons next to them.