BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 8 2005, pages 1693–1694
doi:10.1093/bioinformatics/bti161

Genome analysis

Procom: a web-based tool to compare multiple eukaryotic proteomes

Jin Billy Li[1,*,†], Miao Zhang[1,2,†], Susan K. Dutcher[1] and Gary D. Stormo[1]

[1]Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA and
[2]Department of Biomedical Engineering, Washington University, St. Louis, MO 63130, USA

Received on October 7, 2004; revised on November 9, 2004; accepted on November 17, 2004
Advance Access publication November 25, 2004

ABSTRACT

Summary: Each organism has traits that are shared with some, but not all, organisms. Identification of genes needed for a particular trait can be accomplished by a comparative genomics approach using three or more organisms. Genes that occur in organisms without the trait are removed from the set of genes in common among organisms with the trait. To facilitate these comparisons, a web-based server, Procom, was developed to identify the subset of genes that may be needed for a trait.

Availability: The Procom program is freely available with documentation and examples at http://ural.wustl.edu/~billy/Procom/

Contact: billy@ural.wustl.edu

Comparative genomics has proven extremely powerful in several aspects of genomic sciences that include gene prediction and regulatory element identification (Ureta-Vidal *et al*., 2003). Most comparative genomics studies focus on finding features in common among diverse organisms. Comparisons of closely related organisms often reveal too many candidates to narrow down the gene list of interest. Identification of genes that are retained in some organisms, but are lost in others, also provides important additional information. This comparison is useful when the genes encode proteins associated with a trait that is specific to only a subset of the organisms. Recently, a collection of genes enriched for ciliary and basal body proteins was obtained by comparing proteomes of ciliated organisms, which include *Caenorhabditis elegans, Chlamydomonas reinhardtii, Ciona intestinalis, Drosophila melanogaster, Homo sapiens* and *Mus musculus*, and non-ciliated organisms, which include *Arabidopsis thaliana* and *Saccharomyces cerevisiae* (Avidor-Reiss *et al*., 2004; Li *et al*., 2004). This set of proteins greatly facilitated positional cloning of a new human disease gene associated with ciliary or basal body defects by narrowing down 230 candidate genes to only 2 (Li *et al*., 2004). This accomplishment prompted the development of Procom (Proteome comparison), a generalized web-based tool to compare eukaryotic proteomes; this tool will facilitate other comparisons by this method.

All predicted proteins from the genomes of 30 completely, or nearly completely, sequenced eukaryotic organisms are used for comparison (see Fig. 1 for list). More organisms will be added as the sequences are completed. The proteomes were pair-wise compared with each other using WU-BLASTP (W. Gish, http://blast.wustl.edu) with the threshold E-value = 1 to produce 870 BLASTP output files. To accelerate the subsequent proteome–proteome comparisons, each of these files was parsed to retrieve the query and subject names, and the E-values.

The user specifies three classes of organisms for comparisons; they are the anchor, intersection and subtraction organisms (Fig. 1). The anchor organism serves as the query. The intersection organisms (0 or more) are used to identify the set of shared proteins, and the subtraction organisms (0 or more) are used to exclude proteins that are shared among the three classes of organisms. The user specifies the BLASTP E-values, which can be different for the intersection and subtraction classes. The low intersection E-value and high subtraction E-value will generate a stringent list of output proteins. In contrast, the high intersection E-value and low subtraction E-value will impose loose criteria.

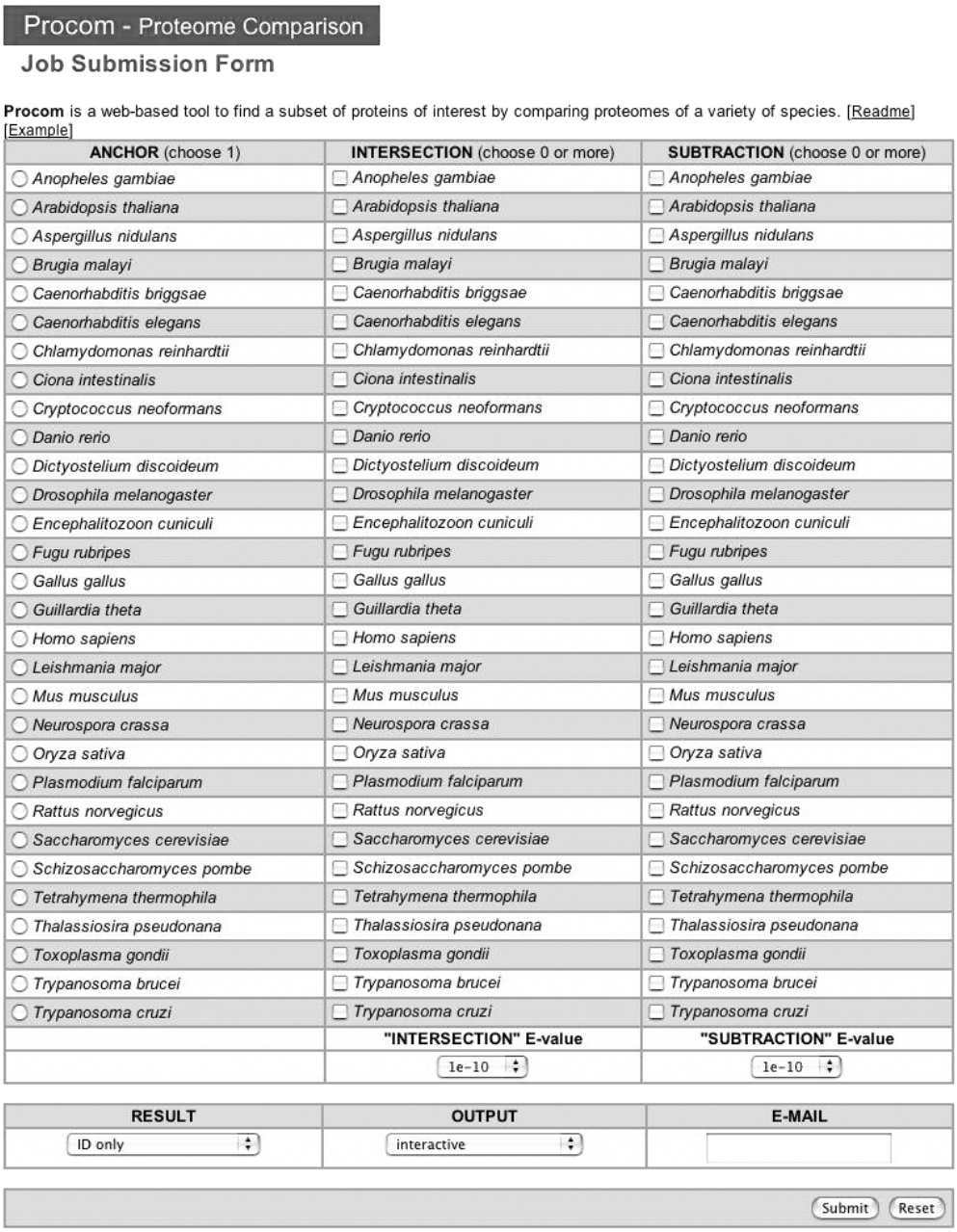
For each of the selected intersection and subtraction organisms, the previously parsed pair-wise BLASTP file with the anchor as query is retrieved to obtain the protein IDs with an E-value less than specified by the user. The collections of the protein IDs are compared with each other to obtain the overlap for intersection organisms and remove the overlap for subtraction organisms. The output of Procom is the protein IDs of the anchor organism. The user can request both the protein IDs and the corresponding sequences. A link for each protein is provided to relevant databases.

The Procom program is written in Perl using the CGI module. It is user friendly and takes no more than 1 min for any combination of comparisons.

Procom should allow users to identify a set of proteins that may be associated with a trait of interest. The proteins associated with the trait must be conserved among organisms retaining the trait, but must be missing in organisms lacking the trait. Procom is by no means a comprehensive tool for comparative genomics analysis, but it provides a novel strategy to compartmentalize candidate genes by the disparity among the proteomes of similar and dissimilar organisms. For example, one could identify photosynthesis candidate genes by searching for *Arabidopsis* proteins that have homologs in rice and the green alga *Chlamydomonas*, but not in animals. Fungal specific genes could be enriched in a set of proteins that are shared by *S.cerevisiae*, *Schizosaccharomyces pombe*, *Aspergillus nidulans*, *Cryptococcus neoformans*, *Encephalitozoon*

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.



Procom - Proteome Comparison

Job Submission Form

Procom is a web-based tool to find a subset of proteins of interest by comparing proteomes of a variety of species. [Readme] [Example]

ANCHOR (choose 1)	**INTERSECTION** (choose 0 or more)	**SUBTRACTION** (choose 0 or more)
○ *Anopheles gambiae*	☐ *Anopheles gambiae*	☐ *Anopheles gambiae*
○ *Arabidopsis thaliana*	☐ *Arabidopsis thaliana*	☐ *Arabidopsis thaliana*
○ *Aspergillus nidulans*	☐ *Aspergillus nidulans*	☐ *Aspergillus nidulans*
○ *Brugia malayi*	☐ *Brugia malayi*	☐ *Brugia malayi*
○ *Caenorhabditis briggsae*	☐ *Caenorhabditis briggsae*	☐ *Caenorhabditis briggsae*
○ *Caenorhabditis elegans*	☐ *Caenorhabditis elegans*	☐ *Caenorhabditis elegans*
○ *Chlamydomonas reinhardtii*	☐ *Chlamydomonas reinhardtii*	☐ *Chlamydomonas reinhardtii*
○ *Ciona intestinalis*	☐ *Ciona intestinalis*	☐ *Ciona intestinalis*
○ *Cryptococcus neoformans*	☐ *Cryptococcus neoformans*	☐ *Cryptococcus neoformans*
○ *Danio rerio*	☐ *Danio rerio*	☐ *Danio rerio*
○ *Dictyostelium discoideum*	☐ *Dictyostelium discoideum*	☐ *Dictyostelium discoideum*
○ *Drosophila melanogaster*	☐ *Drosophila melanogaster*	☐ *Drosophila melanogaster*
○ *Encephalitozoon cuniculi*	☐ *Encephalitozoon cuniculi*	☐ *Encephalitozoon cuniculi*
○ *Fugu rubripes*	☐ *Fugu rubripes*	☐ *Fugu rubripes*
○ *Gallus gallus*	☐ *Gallus gallus*	☐ *Gallus gallus*
○ *Guillardia theta*	☐ *Guillardia theta*	☐ *Guillardia theta*
○ *Homo sapiens*	☐ *Homo sapiens*	☐ *Homo sapiens*
○ *Leishmania major*	☐ *Leishmania major*	☐ *Leishmania major*
○ *Mus musculus*	☐ *Mus musculus*	☐ *Mus musculus*
○ *Neurospora crassa*	☐ *Neurospora crassa*	☐ *Neurospora crassa*
○ *Oryza sativa*	☐ *Oryza sativa*	☐ *Oryza sativa*
○ *Plasmodium falciparum*	☐ *Plasmodium falciparum*	☐ *Plasmodium falciparum*
○ *Rattus norvegicus*	☐ *Rattus norvegicus*	☐ *Rattus norvegicus*
○ *Saccharomyces cerevisiae*	☐ *Saccharomyces cerevisiae*	☐ *Saccharomyces cerevisiae*
○ *Schizosaccharomyces pombe*	☐ *Schizosaccharomyces pombe*	☐ *Schizosaccharomyces pombe*
○ *Tetrahymena thermophila*	☐ *Tetrahymena thermophila*	☐ *Tetrahymena thermophila*
○ *Thalassiosira pseudonana*	☐ *Thalassiosira pseudonana*	☐ *Thalassiosira pseudonana*
○ *Toxoplasma gondii*	☐ *Toxoplasma gondii*	☐ *Toxoplasma gondii*
○ *Trypanosoma brucei*	☐ *Trypanosoma brucei*	☐ *Trypanosoma brucei*
○ *Trypanosoma cruzi*	☐ *Trypanosoma cruzi*	☐ *Trypanosoma cruzi*
	"INTERSECTION" E-value	**"SUBTRACTION" E-value**
	1e-10	1e-10

RESULT	**OUTPUT**	**E-MAIL**
ID only	interactive	

Submit Reset

Fig. 1. Snapshot of the Procom interface. The output is the ANCHOR proteins that have matches in all of the chosen INTERSECTION organisms, but do not have matches in any of the chosen SUBTRACTION organisms.

cuniculi and *Neurospora crassa*, but are missing in the remaining organisms.

ACKNOWLEDGEMENTS

We thank various genome sequencing consortia for access of the sequences and Dr Warren Gish for providing some resources used in this work. This work was supported by National Institutes of Health HG-00249 (to GDS) and GM-32843 (to SKD). JBL was supported in part by the Monsanto Fellowship at Washington University.

REFERENCES

Avidor-Reiss,T., Maer,A.M., Koundakjian,E., Polyanovsky,A., Keil,T., Subramaniam,S. and Zuker,C.S. (2004) Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. *Cell*, **117**, 527–539.

Li,J.B., Gerdes,J.M., Haycraft,C.J., Fan,Y., Teslovich,T.M., May-Simera,H., Li,H., Blacque,O.E., Li,L., Leitch,C.C. *et al.* (2004) Comparative genomics identifies a flagellar and basal body proteome that includes the *BBS5* human disease gene. *Cell*, **117**, 541–552.

Ureta-Vidal,A., Ettwiller,L. and Birney,E. (2003) Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.*, **4**, 251–262.