EUKARYOTIC FUNCTIONAL FAMILIES BASED ON PROKARYOTIC GENOMIC NEIGHBORHOODS

Terry Gaasterland and Bing Hai
Laboratory of Computational Genomics and Laboratory of Molecular and Cellular Neuroscience
The Rockefeller University, New York, New York, USA
{gaasterl,haib}@genomes.rockefeller.edu

Fully sequenced prokaryotic genomes exhibit novel putative operon structures in which functionally related genes occur closely together in the genome [1, 2, 3, 6, 5, 4]. In contrast, eukaryotic genes do not form operons. Moreover, clusters of functionally related genes in eukaryotic genomes are relatively rare. We have devised a strategy to take advantage of prokaryotic genome organization to find candidate functional relatives of eukaryotic genomes. We have applied this strategy to augment a set of cancer-related proteins, as determined by keyword searches of the SwissProt and EMBL nucleic acid databases. The augmented set of cancer-related proteins became the basis for target selection for a high throughput structural genomics initiative focused on solving structures related to cancer.

To execute the strategy, we built a system that uses two derived datasets to accept a eukaryotic query protein and return potential functional relatives. The first dataset contains pairwise alignments of protein sequences from available complete genomes. We used pairwise similarities generated with PSI-BLAST for 21 complete genomes. For the second dataset, the system derives intra-genome "ORF neighborhoods", according to user-configurable distance parameters. A neighborhood contains a set of same-stranded open reading frames (ORFs) each within a minimum distance of the next ORF. The system accepts one or a set of eukaryotic proteins as a query, identifies prokaryotic homologs of the query proteins, constructs a list of neighboring ORFs within the prokaryotic genomes, and then identifies their eukaryotic homologs in *C. elegans* and *S. cerevisiae*. This eukaryote-prokaryote-neighbor-eukaryote strategy enables us to augment any set of eukaryotic proteins with additional proteins that are likely to be functionally related.

To calibrate the degree to which prokaryotic proximity implies functional relationship, we computed ORF neighborhoods for 21 complete prokaryotic genomes at minimal distances ranging from 5 bases to 150 bases. For each genome with known operons, we identified the maximum distance threshold that associated known operons within the same ORF neighborhood. We also assessed the degree of genomic organization for each genome with neighborhoods computed at increments from 5 to 150 bases of minimum distance. We measured genomic organization by percentage of ORFs in neighborhoods, average number of ORFs per neighborhood, and neighborhoods per megabase. By these measures, the archaeal genomes are more organized than the bacterial genomes. Further, the archaeal genomes exhibit novel proximate pairs of genes, compared to bacterial genomes. Finally, we assessed the degree to which the functional categories assigned to ORFs within a prokaryotic-neighborhood were consistent.

For efficiency, we pre-computed prokaryotic-neighborhood-based families of proteins for *C. elegans* and yeast. From 2403 yeast proteins with *E.coli* homologs at a P-value better than 10^{-4}, we computed 684 prokaryote-neighborhood-based families. From an initial set of 5609 *C.elegans* proteins with *E.coli* homologs that met the similarity threshold, we computed 779 prokaryote-neighborhood-based families. These families are the basis for a prototype interactive query tool available at `http://genomes/cgi-bin/haib/search_nbhds`.

References

[1] Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., Shao, Y. (1997). The complete genome sequence of Escherichia coli K-12. Science Sep 5;277(5331):1453-74

[2] Deckert, G., Warren, P., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D., Overbeek, R., Snead, M., Keller, M., Aujay, M., Huber, R., Feldman, R., Short, J., Olsen, G., & Swanson, R. (1998). The *Aquifex aeolicus* genome. *Nature,* **392**, 352–358.

[3] Fitz-Gibbon S., Choi A.J., Miller J.H., Stetter K.O., Simon M.I., Swanson R., Kim U.J. (1997). A fosmid-based genomic map and identification of 474 genes of the hyperthermophilic archaeon Pyrobaculum aerophilum. *Extremophiles* **1**(1):36-51.

[4] Gaasterland, T. & Ragan, M. (1998). Phyletic and functional patterns of ORF distribution among prokaryotes. *Journal of Microbial and Comparative Genomics,* **3**, 199–217.

[5] Gaasterland, T. & Sensen, C. W. (1996). Fully automated genome analysis that reflects user needs and preferences — a detailed introduction to the MAGPIE system architecture. *Biochimie,* **78**, 302–310.

[6] Sensen, C., Klenk, H.-P., Singh, R.-K., Allard, G., Chan, C., Liu, Q., Penny, S., Young, F., Schenk, M., Gaasterland, T., Doolittle, W., Ragan, M., & Charlebois, R. (1996). Organizational characteristics and information content of an archaeal genome: 156 kbp of sequence from *Sulfolobus solfataricus* P2. *Journal of Molecular Microbiology,* **22**, 175–191.