Metazome

http://cigbrowser.berkeley.edu/annotation.html

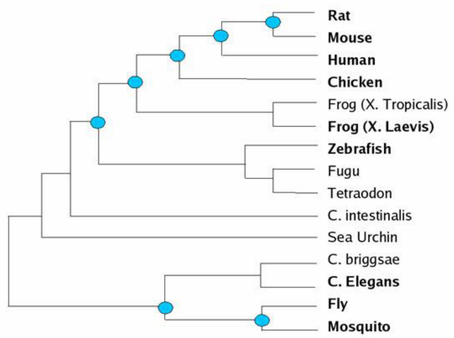
Bill Dirks[1], David Goodstein[2], Uffe Hellsten[2], Dan Rokhsar[1,2]
1. Center for Integrative Genomics; UC Berkeley; Berkeley, CA 2. DOE Joint Genome Institute; Walnut Creek, CA

*This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-765F00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.

Abstract

Metazome is a database and graphical user interface enabling comparative genomic studies within the Metazoa. This system allows uniform access to a high quality set of protein-coding genes from all annotated metazoan genomes. Each gene has been assigned with PFAM, KOG, and PANTHER annotations, and publicly available annotations from RefSeq, SwissProt, Ensembl, and JGI are hyper-linked and searchable. In addition, these genes are organized into orthologous gene families each representing the modern descendents of the ancestral gene at a key phylogenetic node such as the ancestral mammal, tetrapod, vertebrate, etc.

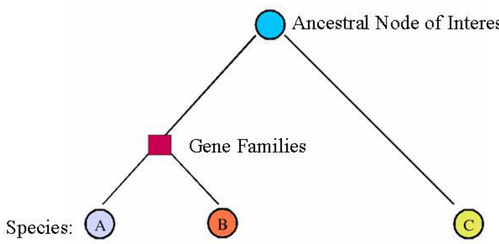
Current Genomes and Gene Families



Available Gene Families

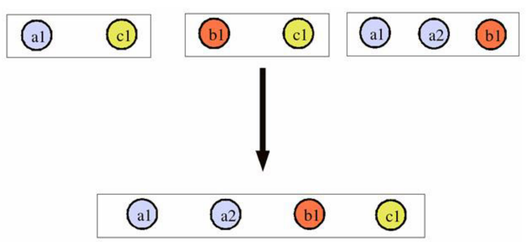
Gene Family Creation Algorithm

Step 1 - Making a Backbone of high quality putative orthologs



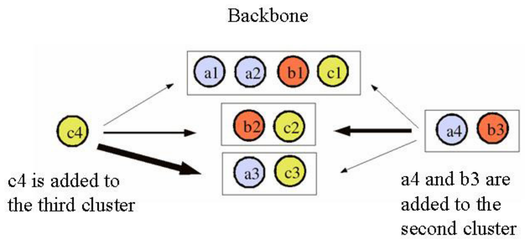
First, we find mutual best hits across the ancestral node of interest. This is done using BLAST, stringing together the highest scoring set of nonoverlapping hsps. This gives us pseudo-global alignments.

Next, we combine all the pairwise best hits and previous families if they share common genes.



This becomes a backbone element of orthologous genes.

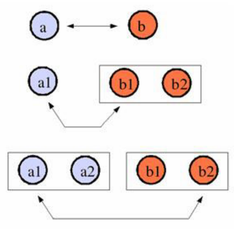
Step 2 - The backbones are grown out into gene families



Genes and previously made gene families are added to a backbone element if we believe they are paralogs to this backbone element at this depth of evolution. We assert a gene or gene family is paralogous to its closest backbone element if the distance is better than a pre-computed empirical threshold (see below).

Distances

Our distances respect the variable mutability of different gene families.



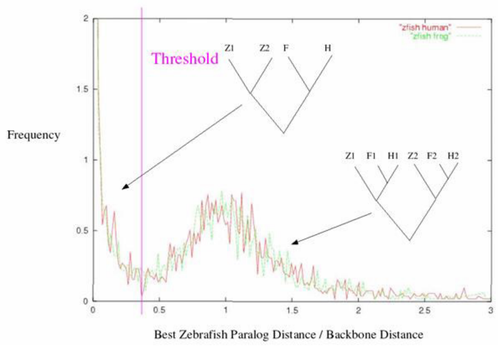
d(a,b) = peptide percent difference

d(a,B) = min { d(a,b)/Diversity(B) | b is an element of B}

d(A,B) = max { d(a,B) | a is an element of A }

diversity(B) = min { d(b1,b2) | b1, b2 is a mutual best hit backbone element of B }

Empirical Paralogy Threshold for Constructing Gene Families



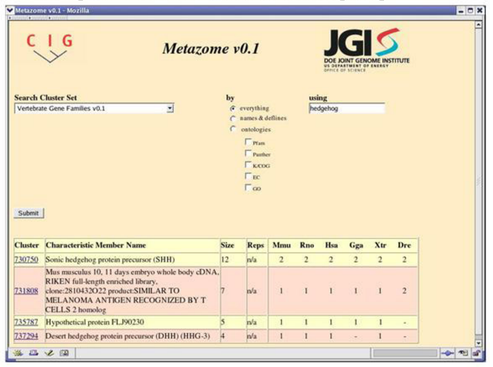
When choosing a threshold for constructing gene families we look at triplets of best hit gene pairs within a species and between species. This lets us distinguish old paralogs from new ones.

Gene Family Statistics

Evolutionary Depth	N. of Organisms	N. of Gene Families	Average Family Size	Family Size Standard Deviation	N. of Genes in Families	N. of Genes Total
Rodent	2	16569	2.051	0.003	33893	47127
Mammal	3	15645	3.082	0.009	48225	69216
Amniote	4	11369	4.398	0.003	49435	86611
Tetrapod	5	10787	5.539	0.034	76570	143484
Vertebrate	6	10684	7.167	0.050	76570	143484
Fly-Mosquito	2	7297	2.088	0.001	15235	27418
Ecdysozoa	3	4926	3.095	0.016	15247	47238

Screenshots

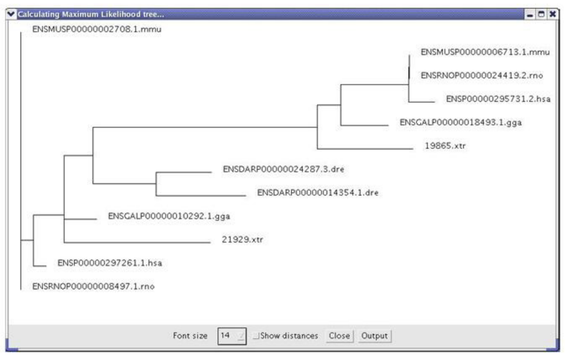
Searching the vertebrate families for hedgehog



Cluster Detail Information

Examining a pre-computed multiple sequence alignment with JalView

Retrieving a tree precomputed using Tree-Puzzle



Coming Features

1) additional organisms and gene families
2) Searching with sequence (BLAST)
2) Retrieving nearby genomic sequence to facilitate cis-regulation studies
3) Manual curation of family names for core families at key nodes
4) Incorporation of expression data and protein-protein interactions

*This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.