Databases and ontologies

SPEED: a molecular-evolution-based database of mammalian orthologous groups

Eric J. Vallender¹, Justin E. Paschall², Christine M. Malcom¹,³, Bruce T. Lahn¹ and Gerald J. Wyckoff²,*

¹Department of Human Genetics and Committee on Genetics, Howard Hughes Medical Institute, University of Chicago, 920 East 58th Street, Chicago, IL 60637, USA, ²Division of Molecular Biology and Biochemistry, University of Missouri-Kansas City, 5007 Rockhill Road, Kansas City, MO 64110, USA and ³Department of Anthropology, University of Chicago, 920 East 58th Street, Chicago, IL 60637, USA

Received on June 1, 2006; revised on August 29, 2006; accepted on August 31, 2006
Advance Access publication September 11, 2006
Associate Editor: Chris Stoeckert

ABSTRACT
Motivation: The abundance of nucleotide sequence information available has expanded horizons of inquiry for molecular evolution; however, the full potential of whole-genome analysis has not been realized because of inadequate tools. Here, we present one of the first toolkits to aid multidisciplinary high-throughput analysis.
Summary: SPEED was created to integrate molecular evolutionary data with existing genetic resources and provide a straightforward user interface to 17,352 orthologous gene groups, containing representatives from eight mammalian species and an avian outgroup.
Availability: See http://bioinfobase.umkc.edu/speed/ for access
Contact: wyckoffg@umkc.edu
Supplementary information: A larger version of the data model and a site map are available online.

1 INTRODUCTION
Mammalian genomes are being sequenced at an increasingly rapid rate. High quality analyses of five mammalian genomes have already been released (human (Lander et al., 2001; Venter et al., 2001), mouse (Waterston et al., 2002), rat (Gibbs et al., 2004), chimpanzee (The Chimpanzee Sequencing and Analysis Consortium, 2005) and dog (Lindblad-Toh et al., 2005), with two more (rhesus macaque and cow) in publicly available draft versions. These diverse species open up countless avenues of exploration for comparative evolution; studies of changes in mutation rates, chromosomal rearrangements and non-coding regions are all prevalent.

The accessibility of whole-genome assemblies is changing the face of protein studies. Largely, these studies have focused on deciphering patterns of gene divergence across chromosomes (Malcom et al., 2003; Vallender and Lahn, 2004; Webster et al., 2004) and detecting positive selection (Bustamante et al., 2005; Clark et al., 2003; Nielsen et al., 2005). Although, the methodology of whole-genome analysis has broad applications, studies to date have often used different datasets that are difficult or impossible to recapitulate and compare. The format and size of these datasets often create prohibitive hurdles for researchers without the expertise or tools necessary to interpret and manipulate them.

Several platforms, such as The National Center for Biotechnology Information (NCBI) Homologene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene) (Wheeler et al., 2006) and Ensembl (http://www.ensembl.org) (Birney et al., 2006) have integrated evolutionary analysis of orthologs into their presentations of gene information. Although both platforms serve as the lingua franca of the genetics community, neither of them gives more than the basic evolutionary information. Molecular biologists still need to enlist evolutionists to probe beyond basic summary statistics and contextualize information.

Here, we present a database that houses a large set of mammalian orthologs, plus an array of evolutionary information for each gene. The Searchable Prototype Experimental Evolutionary Database (http://bioinfobase.umkc.edu/speed) (SPEED) integrates information currently scattered throughout disparate sources (e.g. domain structure, expression, disease phenotype and function), and marries these preexisting data with molecular evolutionary parameters. SPEED offers the benefit of a web interface that places this information easily within the grasp of researchers, enabling them to devise and conduct complete evolutionary studies of mammalian protein evolution.

2 METHODS
SPEED is constructed as a MySQL relational database with a PHP-based web front end. Genes from each species are assigned unique sequence ids. These are grouped into orthologs and then assigned a unique orthologous group id (SPEED id). Links to outside databases and information unique to species-specific sequences are constructed based on sequence ids, whereas information related to the orthologous groups, such as the evolutionary information, is linked to the SPEED ids. This robust structure lends itself to expansion, both in terms of the amount of data stored (i.e. orthologous sequences and species included) and the queryable parameters available (i.e. evolutionary, physiological and other characteristics).

The primary interface for most users is the web-based front end. Designed to be user friendly without limiting access to all data, this can be partitioned between gene search and data display sections via the gene search portals, to be user friendly without limiting access to all data, this can be partitioned between gene search and data display sections via the gene search portals, the queryable parameters available (i.e. evolutionary, physiological and other characteristics).

The primary interface for most users is the web-based front end. Designed to be user friendly without limiting access to all data, this can be partitioned between gene search and data display sections via the gene search portals, orthologous groups of interest are identified and displayed. A complete data mode is included as Supplementary Figure 1 and a site map for the SPEED homepage, is included as Supplementary Figure 2.
Datasets used were obtained from Ensembl v36 (Binney et al., 2006). A total of seven mammalian datasets were used: human, *Homo sapiens* (NCBI 35); chimp, *Pan troglodytes* (PanTro 1.0); rhesus macaque, *Macaca mulatta* (Mmul 0.1); mouse, *Mus musculus* (NCBI m34); rat, *Rattus norvegicus* (RGSC 3.4); dog, *Canis familiaris* (CanFam 1.0); cow, *Bos taurus* (Btau 2.0). Two additional species were included where possible: opossum, *Monodelphis domestica* (source: MonDom 2.0) as a basal mammalian representative and chicken, *Gallus gallus* (source: WASHUC1) as a true outgroup.

Orthologous groups were identified using BLAST reciprocal best hits (RBH) (Tatusov et al., 1997, 2003). Clusters of RBHs were generated by first requiring that all genes show RBH with all other genes in the cluster. Additional genes were then added by requiring RBH to at least three members of an orthologous group. To verify putative orthologous relationships, crude syteny maps were created using orthologs as anchors. These maps were in agreement with previously published syteny maps (Brudno et al., 2004). A total of 17 352 orthologous groups were identified. Among these, 5370 have orthologs identified in each of the nine species interrogated and 9401 represent groups containing orthologs for all seven mammalian species. A total of 13 506 groups contain orthologs in all three sequenced primate species, with 10 929 of these also containing mouse and rat orthologs.

Alignments were accomplished using CLUSTAL (Chenna et al., 2003) on translated sequences frame-aligned with nucleotide sequences, preserving gaps in the amino acid alignment. Divergence data were calculated using the modified Li method (Li, 1993). Sliding-window analyses were conducted using a window size of 33 codons and a step size of 10 codons. Expression information was drawn from the Cancer Genome Anatomy Project (CGAP) (Riggins and Strausberg, 2001) and information on metabolic pathways was culled from Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2006).

3 IMPLEMENTATION

Six primary portals for gene information retrieval are provided: (1) Gene name, symbol or description; (2) associated disease; (3) Pfam conserved domain; (4) tissue of expression; (5) functional pathway or (6) protein amino acid motif. Upon selection of a single orthologous group, both summary information on individual species and data on evolutionary comparisons of orthologs are provided. This summary information includes gene names and descriptions for each ortholog as well as links out to additional databases, (such as ReiSeq and EntrezGene). Also provided are the sequence length, cytenygetic position, a global multi-way alignment (both of the nucleotide and amino acid sequences), expression and pathway information. For each orthologous group, every pair-wise comparison is made and values are presented for Ka, Ks, Ka/Ks, as well as numbers of amino acid and synonymous changes (raw and corrected for multiple hits). A breakdown of amino acid changes (with residue-specific information) and a sliding-window display of sequence evolution are also available. Aligned sequences can be downloaded for further local analysis.

4 DISCUSSION

Additional genomes and the large-scale data analysis projects that accompany them are producing valuable information that is broadly applicable to the life sciences. SPEED is a conduit through which molecular evolutionary data can be made available to researchers regardless of discipline.

Earlier versions of SPEED have formed the backbone for a number of large-scale evolutionary studies (Choi et al., 2005; Malcom et al., 2003; Vallender and Lahn 2004; Wyckoff et al., 2005) as well as studies on specific genes of interest (Dorus et al., 2004; Gilbert et al., 2005). SPEED can be used to determine orthologous genes across species and display the similarities, differences and rates of evolution between species. It allows for consideration of variation in these factors within genes whether by focusing on positional heterogeneity, evolutionary parameters of individual domains or even properties of mutations between individual amino acids. The orthologous groups are tied into many traditionally useful databases, allowing genes and their evolutionary signatures to be considered in context. We believe SPEED will greatly aid inter-disciplinary studies in the post-genomic era.

ACKNOWLEDGEMENTS

This work was supported by a University of Chicago William Rainey Harper Dissertation Fellowship (to E.J.V.), a University of Missouri Research Board grant (to G.J.W) and a Searle Scholarship and Burroughs Wellcome Career Award (to B.T.L).

Conflict of Interest: none declared.

REFERENCES


