A Computational Framework for the Identification, Cataloguing, and Classification of Evolutionarily Conserved Genomic DNA

by

Sunil K. Saluja

M.D.
University of Pennsylvania 1995

BS Microbiology, Medieval and Renaissance Studies
University of Michigan 1990

Submitted to the Harvard/MIT Division of Health Sciences and Technology in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Biomedical Informatics

at the

Massachusetts Institute of Technology

June 2004

This work has been funded by the National Library of Medicine Fellowship in Medical Informatics.

Signature of Author .....
Division of Health Sciences and Technology
May 7, 2004

Certified by .....
Isaac S. Kohane, M.D., Ph.D
Henderson Professor of Health Sciences and Technology
Thesis Supervisor

Accepted by .....
Martha L. Gray, Ph.D.
Edward Hood Taplin Professor of Medical and Electrical Engineering
Co-Director, Harvard-MIT Division of Health, Sciences and Technology

Abstract

Evolutionarily conserved genomic regions (*ecores*) are understudied, and yet comprise a very large percentage of the Human Genome. Highly conserved human-mouse non-coding *ecores*, for example, are more abundant within the Human Genome than those regions, which are currently estimated to encode for proteins. Subsets of these *ecores* also exhibit conservation that extends across several species. These genomic regions have managed to survive millions of years of evolution despite the fact that they do not appear to directly encode for proteins. The survival of these regions compels us to investigate their potential function. Development of a computational framework for the classification and clustering of these regions may be the first step in understanding their function. The need for a standardized framework is underscored by the explosive growth in the number of publicly available, fully sequenced genomes, and the diverse set of methodologies used to generate cross-species alignments.

This project describes the design and implementation of a system for the identification, classification and cataloguing of ecores across multiple species. A key feature of this system is its ability to quickly incorporate new genomes and assemblies as they become available. Additionally, this system provides investigators with a feature rich user interface, which facilitates the retrieval of ecores based on a wide range of parameters. The system returns a dynamically annotated list of evolutionarily conserved regions, which is used as input to several classification schemes, aimed at identifying families of ecores that share similar features, including depth of evolutionary conservation, position relative to known genes, sequence similarity, and content of transcription factor binding sites. Families of ecores have already been retrieved by the system and clustered using this feature space, and are currently awaiting biological validation.

Background

In the short period of time extending from the release of the first draft of the Human Genome in 2001[1] to the publication of the final draft in April 2003, whole genome assemblies of other species have been generated at an unprecedented rate. Prior to the initial sequencing and analysis of the Human Genome, *C. elegans[2]*, and *Drosophila menalogaster[3]* were the only other two metazoan genomes available. Since that time, *Mus. musculis[4]*, *Fugu rubripes[5]*, *Anopheles gambiae[6]*, and others have been added to the list. *Rattus norvegicus* soon followed, with large portions of the rat genome already available in public databases.[7] While initial sequencing targeted those species, which were most amenable to experimental manipulation in the laboratory, more recent efforts will also provide a robust substrate for studies in comparative genomics. The complete sequencing of *Caenorhabditis briggsae*, and *Pan troglodytes* (chimpanzee) genomes, for example, will soon be providing a genome-wide mechanism for comparing *C. elegans* and *Homo sapiens* to their very close evolutionary relatives.[8] The addition of bovine, canine, feline, and avian assemblies will extend the range of metazoans to include a wide variety of branches on the evolutionary tree.

This availability of multiple completed genomes will contribute to the study of anthropology, evolutionary biology, and cellular/systems biology through different mechanisms. Evolutionary biologists and anthropologists use the assumption that all species ultimately diverged from a common ancestor to develop evolutionary clocks, aimed at re-constructing the tree of life.[9] Comparative genomics serves as the chronometer for this clock by enabling the measurement of sequence divergence in coding and non-coding sequences of two or more species over

evolutionary time. This ratio of synonymous or neutral substitutions to non-synonymous substitutions has become the basis for modern genomic correlates to evolutionary theory.[10] While there are several benefits to using a genomic approach to re-constructing a tree of life, limitations to an exclusively genomic approach have also become apparent. The rate of mutation is not constant throughout evolution, and selective pressure is likely to have an inconsistent effect on the rate of base pair changes between two different genomes.[11] More recent studies suggest that a combination of genomic approaches along with the traditional fossil record will help to most accurately define the branches of evolution.[12]

The genomic equivalent to a fossil, is an *ecore*, or evolutionarily conserved region. Since *ecores* have persisted over evolutionary time, and are by definition present in two or more species, they were also present at the points at which those species diverged from a common ancestor. Unlike fossils, however, *ecores* have yet to be catalogued and classified using a formalized set of metrics. Additionally, while fossils are non-functional petrified remnants of extinct species, ecores are living replicating remnants, which are likely to retain some functional characteristics. Extensive cataloguing of *ecores*, along with the subsequent measurement of *ecore* loss and gain between species may be provide a powerful tool in the reconstruction of phylogenetic trees.

Comparative genomics may provide a mechanism for the identification of conserved non-coding genomic sequences with important molecular functions.

While emphasizing the differences between species may aid in the re-construction of more accurate phylogenetic trees, emphasizing the conspicuous similarities between widely divergent species may provide a wealth of data for those who study cellular function. The potential

biological significance of *ecores* is based on the premise that functionally important genomic regions should be more highly conserved across evolution than non-functional regions. Additionally, while some *ecores* have been conserved across a much wider span of evolution, other regions are only shared between closely related species.[13] This depth of conservation is one of many potential features, which may help define the functional characteristics of particular conserved regions.[14] While *ecores* with greater evolutionary depth may be involved in basic cellular functions that are shared by a wide variety of species, *ecores* with less depth may be responsible for the subtle differences in gene regulation, transcription, translation or splicing that account for the divergence of one metazoan species from another.

Several studies are already using evolutionary conservation as a means identifying putative regulatory elements and transcription factor binding sites in chromosomal DNA.[15-17] A much smaller number of regions have been biologically verified as important in the regulation of gene transcription.[18] In these circumstances, the primary functional role of these regions has been transcriptional regulation through binding to a peptide ligand. In addition to the having a direct role in protein binding, conserved non-coding sequences have been identified, which encode for microRNA's, small RNA oligomers, which do not directly encode for proteins, but appear to play a significant role in translation control.[19] These entities have been identified in species extending from humans to yeast, and a significant percentage of the chromosomal regions encoding for these have demonstrated tight cross-species conservation.[20, 21]

In addition to being enriched with transcription-factor binding sites or encoding for microRNA, evolutionarily conserved non protein-coding regions may play an important functional role in tissue-specific coordination of gene expression, stability of the molecule during mitosis, and a myriad of other undiscovered and/or poorly understood functions, which are governed within the

vast non-coding genomic space of mammals. Functional elements may be present anywhere within this vast space. Without the proper cataloguing and classification of *ecore*s a subset of these elements, their functional properties may remain undiscovered.

Several computational methods have been developed for performing whole genome alignments, and identifying evolutionarily conserved regions (ecores)

A rapidly emerging interest in comparative genomics has stimulated the development of a number of tools aimed at performing inter-species alignments of large genomic sequences. Since the availability of a nearly complete mouse genome, three groups have performed and published genome-wide cross-species alignments between the human and mouse, each using a slightly different tool kits, and methods for visualization. The University of California Santa Cruz (UCSC) Genome Browser/Genome Project, in collaboration with Penn State, has performed genome-wide mouse-human alignments using BLAT/BASTZ methodology.[22, 23] The Ensembl browser uses Phusion/blastn to calculate human/mouse/rat alignments.[24, 25] The Berkeley Genome pipeline uses BLAT/AVID.[26] Additional species are being added to each of these resources, but inconsistently, and with variability in terms of which species is being used as a reference. Figure 1 displays a schematic, published by by Ureta-Vidal et. al, which represents currently available pre-computed alignments, and the species which they include.[27]

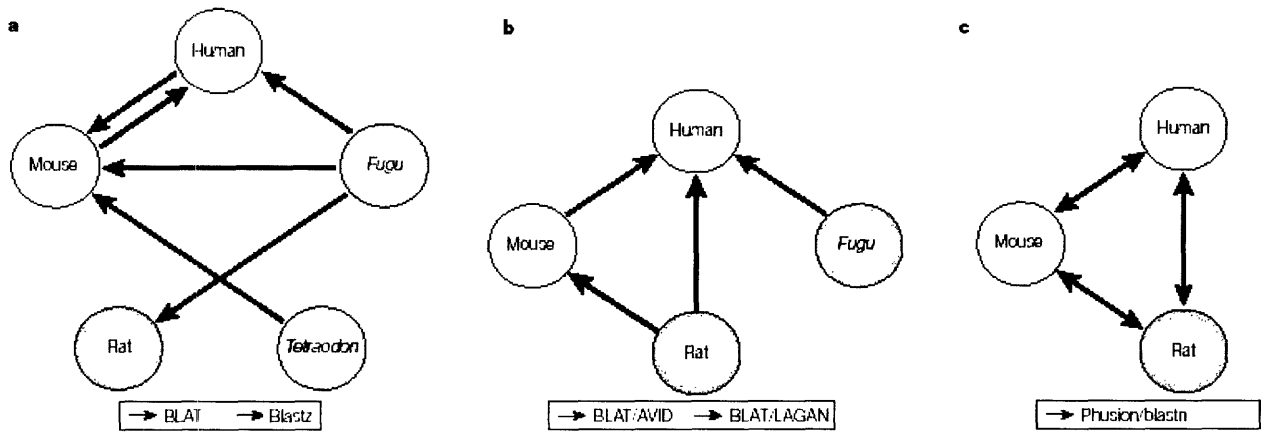


Figure 1: This schematic diagram, published by Ureta-Vidal et. al in 2003, displays three different pre-computed cross species alignments, which are publicly available through the world wide web. In all cases the direction of the arrow indicates the reference genome upon which alignments are mapped. The colors of the arrows correspond to different methods used based on the species being aligned. The following groups use the corresponding methodologies: (a) UCSC Genome Project and Penn State Bioinformatics group. (b) Godzilla, the Berkeley Genome Pipeline, and (c) Ensembl Genome Browser.[27]

A consistent method for pre-computing, identifying, and cataloguing of ecores has not yet been implemented.

As figure 1 indicates, in the very short period of time since the emergence of complete or nearly complete genome assemblies, several approaches to whole-genome alignment have been implemented, with varying degrees of success depending on the outcome desired.

An essential step in the computational alignment process involves the creation of a matrix using the dynamic programming principle. This method involves the representation of nucleotide sequences to be aligned along the horizontal and vertical axes of an extended matrix. All possible alignments between the two sequences are then considered, with penalties being given for mismatches and gaps in the sequence alignment. The highest scoring alignment is returned. Two widely used algorithms for implementing dynamic programming principles in sequence alignment are the Needleman-Wunsch algorithm, a global alignment method, and the Smith-Waterman algorithm, which is a local alignment method.[28, 29] While global methods

return the optimal alignment between two strings, for example two syntenic Mouse-Human Chromosomal regions, local alignment methods return local regions, or substrings which have the highest pairwise alignment scores. Local methods will tend to be more sensitive, but less specific. In syntenic regions, deletions, insertions, and shifts in sequence can influence the global alignment, whereas a local alignment will identify the substrings, which are tightly aligned. Conversely, local aligners are less constrained by global similarities and differences, and hence may identify tightly aligned regions, which are very different in terms of their relative position. Each method may therefore generate a slightly different subset of *ecores*.

Calculating optimal whole genome alignments is further complicated by the size of the regions to be aligned. Dynamic programming principles cannot be used in their pure form since alignments of this size quickly become computationally intractable. Hence, all current methodologies must implement some form of a heuristically guided search in order to predict alignments. This heuristic is based on the assumption that tightly aligned sequences are likely to have significant stretches of un-gapped matches. The strategy deployed first identifies a *seed* sequence with a certain number of un-gapped matches. Once appropriate seed sequences are identified, they are extended, and used as anchor points upon which to begin implementation of the chosen alignment method. This approach is tantamount to implementing an extensive heuristically guided tree-pruning algorithm, which reduces the size of the search space dramatically prior to performing a simple depth first search. *BLAST* (Basic Local Alignment Search Tool), the most widely used tool to search large genomic databases, uses 11 consecutive identical base pairs as its seed sequence.[30] Blastz, which is used by the UCSC Human Genome project to perform Mouse-Human genome-wide alignments, modifies the Blast heuristic, by extending gapped seed sequences, in order to improve both sensitivity and specificity, and adjust for the larger task of performing whole-genome alignments.[31]

These alignment methods are further confounded by the presence of a large number of repetitive elements within many genomes, including *homo sapiens*. Whole genome alignment methods be capable of filtering out low and intermediate complexity repetitive DNA, prior to performing alignments. Once alignments are completed, repeats may re-introduced into the genomes in order to maintain positional integrity.

In addition to the choice of a global vs. local alignment, seed sequence structure, and a method for handling repeats, alignment algorithms must also include specific scoring parameters, and have a well developed rationale for the parameters which are chosen. The amount to penalize gaps vs. mismatches may significantly affect the alignment score returned. The initial scoring matrix used for generating human-mouse alignments using Blastz, for example, was determined by reverse engineering the optimal matrix to align well known homologous regions between human and mouse. The result was a matrix that returned sequences corresponding to the top 40% of regions within the human genome demonstrating significant sequence similarity to the mouse. Adjustment of this matrix by providing greater penalties for mismatches and gaps increased the stringency of this algorithm, with the resulting dataset corresponding to the 6^{th} percentile of the human genome with the highest degree of similarity to the mouse. In this example, altering the matrix resulted in significantly different datasets. Since the algorithm was otherwise the same, the *ecores* present in the tightest set of alignments using blastz, constitute a subset of the *ecores* present in the larger set.

While blastz has been used to illustrate an example of the parameters necessary to perform genome-wide alignments, this specific algorithm is only one of several which are currently being used in the informatics community. The basic principles however are the same. Blastz performs local alignments, AVID and LAGAN perform global alignments. AVID and LAGAN algorithms thus achieve specificity by using a global alignment framework, but must work under the premise that the two sub-genomic sequences being aligned are homologous or syntenic.

Translated BLAT is used to identify syntenic egions prior to performing the alignments. The algorithms also differ from each other and blastz in terms of the heuristics used to identify anchor sequences, and the level of degeneracy they allow prior to performing an alignment. While AVID has been designed to be better suited for the alignment of close evolutionary relatives, for example, LAGAN allows for more divergence between sequences, and may be better suited for aligning more distantly related species.[27]

Ecores generated using different methods are not catalogued or referenced in a single database.

Despite the availability of a diverse set of algorithms for the characterization and identification of *ecores*, no currently available resource exploits the differences and similarities between these algorithms to construct a single unified database, which represents *ecores* in a species independent fashion. While several investigators have exploited data derived from genome-wide alignments to identify putative regulatory elements, the reason for selection of one dataset over another is often not clear. The identified *ecores*, however, may be present in one dataset and absent in another. Additionally, there is not one unified source of conservation information from where to retrieve and study *ecores* derived using a diverse set of methods. Current data models represent *ecores* within flat cross-species alignment files, which contain alignment scores, positions, and the putative *ecore* sequences of the two species being aligned. The addition of a third or a fourth species necessarily complicates this representation, and hence is excluded. Determining conservation across more than two species therefore involves consecutive lookups between databases.

An alternative, and potentially much more powerful approach to the representation of *ecores* is one which is defined by the *ecore* itself, and not by the methodology used, or the species being

aligned. The key to this approach is the premise that *ecores* are distinct entities, much like fossils in a fossil record. While they may be found in one species, they may be absent in another. While one *ecore* may be have been identified using particular methodology, it may be absent when a different methodology is employed. *Ecores* must therefore not only be represented in terms of the species in which they are found, but also under what circumstances they are found, and using what methodology.

Preliminary Studies

Dynamically evolving whole genome assemblies significantly influence the identification and characterization of ecores.

Initial attempts to identify and characterize ecores in this laboratory were performed using the November 2002 draft of *Homo sapiens* (hg13), and the February 2002 draft of *mus musculis* (mm2). Conserved regions were classified as exonic, intronic, intergenic, flanking, and untranslated. Flanking regions spanned across an exon boundary. Un-translated regions were transcribed, but not translated. This nomenclature is visually displayed in figure 2.

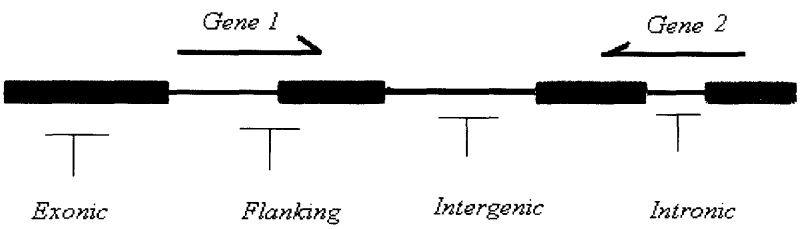


Figure 2. Identified ecores were initially classified as Exonic, Flanking, Intergenic, or Intronic.

(a) Homo Sapiens (hg13) Vs. *Ms Musculis* (mm2)

Ecores	Exonic	Flanking	Intronic	Intergenic
Number of conserved Regions	5138	10362	112,475	653,291
Percent Represented	0.66	1.3	16.9	81.1
Average Size (bp)	212	379	192	206

(b) Homo Sapiens (hg16) Vs. *Ms Musculis* (mm4)

Ecores	Exonic	Flanking	Intronic	Extragenic
Number	11523	128546	212787	440985
Percent represented	1.5	16.2	26.8	55.5
Average Size (bp)	212	379	192	206

(Table 1) Average Size and Geographic Distribution of Ecores identified through the alignment of (a) *Homo Sapiens* November 2002 Release (hg13) and *Mus Musculis* February 2002 release (mm2) compared with (b) *Homo Sapiens* July 2003 Release (hg16) and *Mus Musculis* October 2003 release (mm4)

The most remarkable finding in this initial characterization of *ecores* was the sparse representation of protein –coding sequences within the set of tightly conserved regions, comprising less than 3 percent of all *ecores*. (Table 1a.)

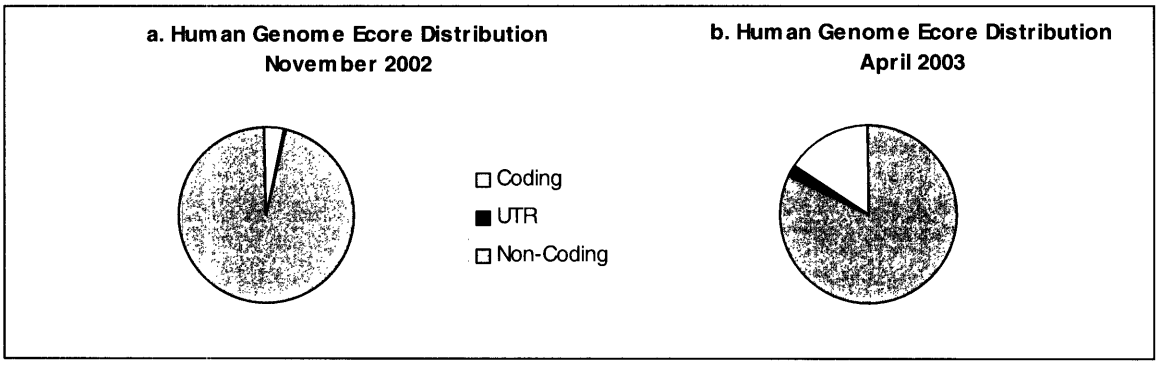


Figure 3. Comparison of Human Genome ecore Disribution between Nov 2002, and April 2003

In an effort to re-compute *ecore* distribution based on the changing profiles of whole genome assemblies, whole genome alignments using Blastz, the April 2003 final draft of *homo sapiens*, and the most recent draft of *mus musculis*, were used to query the NCBI April 2003 reference sequence database of gene products. The April 2003 refSeq annotation contained 19,174 entries as opposed to 18,214 entries in the November 2002 release. Additionally, the mouse-human alignment based on the most recent draft genomes was comprised of 793,841 unique *ecores*, as opposed to 781,266 *ecores* in the previous alignment, indicating an increase in the number of *ecores* by 1.6 percent over a six-month period. (Table 1b) Despite these relatively modest changes between drafts, the mapping of *ecores* to regions within the genome demonstrated a significant increase in the number of protein-coding regions represented. (Figure 3) The vast majority of these regions also contained peri-exonic non-coding sequence (Table 1b).

This marked change in the annotation of *ecores* following the release of the final draft of the human genome suggests that dynamically evolving whole genome assemblies can significantly influence the identification and characterization of *ecores*. This change appears to be validated by the increased representation of protein-coding regions in the superset of *ecores*. While protein-coding regions make-up less than 3 percent of the human genome, over 17 percent of all *ecores* now appeared to contain at least some protein-coding sequence.

While the assemblies initially used to identify ecores were near completion, the differences in identified ecores between draft and final assemblies were striking. This phenomenon underscores an emerging challenge in comparative genomics. While there is significant scientific motivation to identify ecores across multiple genomes, this must be tempered by a realization that early comparisons may not reflect the true biology of the organisms being

studied. Computational methods for the dynamic incorporation of new assemblies and ecores and retirement of older assemblies and ecores will be of paramount importance to future investigators, as the pool of whole genome assemblies continues to grow. This computational challenge is amplified by the need to identify ecores across several species, wherein the introduction of a single new assembly could potentially change the ecore profile across all other species. This project attempts to overcome these obstacles through the implementation of a system that is designed to identify and catalogue ecores in a dynamic fashion in response to changing assemblies. In addition, this schema will also allow the incorporation of ecores that have been generated using different methodologies.

System Design

Database Design

The first major step in the design of a system to classify ecores was the identification of a standard vocabulary through which to define all ecores. This vocabulary was then subsequently used to construct a relational database, which is more capable of dynamically incorporating new alignments than the currently deployed flat file system. The major features of this database are as follows.

- Ecore specific schema

Since a standard vocabulary is a prerequisite for the construction of a relational database, the key paradigm in database design is that *ecore*s should be treated as species independent entities. While current data repositories often use different vocabularies depending on the investigator, methodology, and the species being compared, these differences can be

integrated into a common vocabulary. In fact, the use of a standard vocabulary may more accurately reflect the phenomenon of evolutionary conservation. Since *ecores* are generated through the genome-wide alignment of at least two different species, they are by definition older in evolutionary terms, than either of the species from which they were derived. Hence, using a species-specific approach to referencing ecores unnecessarily biases the data. This principle becomes more important as *ecores* are expected to meet additional criterion for validity, such as being identified using more than one alignment method, or being found within more than two species.

- Incorporation of multiple alignment algorithms.

An important feature of an ecore is the alignment method used to generate it. While the use of competing alignment methods will return overlapping sets of ecores, this level of redundancy can be easily overcome though the use of a well-constructed SQL statement. In fact, the incorporation of competing alignments within a single relational database allows the investigator to limit the set of retrieved cores to only those that have been derived using more than one methodology. This methodologically agnostic approach is achieved by maintaining the positional integrity of the ecore, relative to the reference species, regardless of the alignment method deployed.

- Positional Consistency

While ecores were ultimately be treated as species independent entities, the positional integrity of ecores could only be maintained using species-specific positional information. This was accomplished by defining ecores based on reference genome position, and cross-species position. When searches extend across several genomes, the system was designed to keep track of ecore positions within the original reference genome and all other genomes used for cross-species comparison. This allows for the incorporation of new assemblies into the data structure without losing any positional information. (Fig 3)

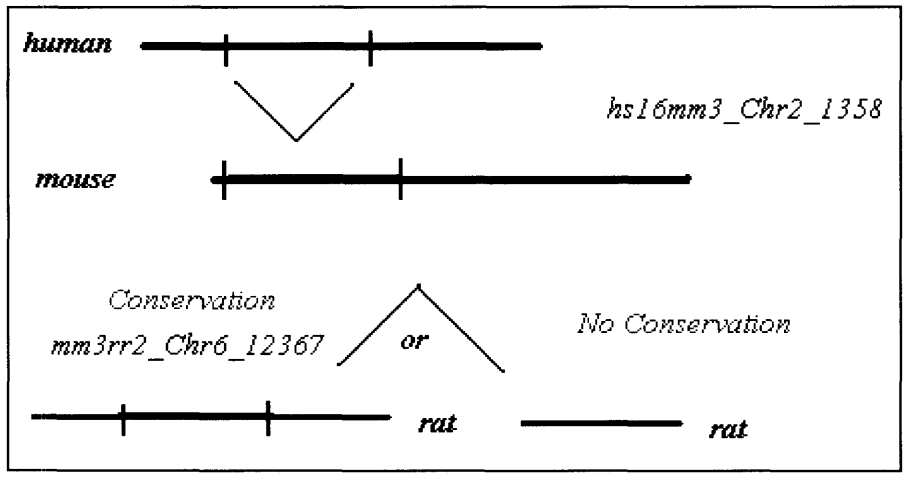


Figure 3: Schematic demonstrating mechanism to map ecores across multiple assemblies by maintaining positional reference. First, an ecore identified through the alignment of human and mouse genomes is stored within the database with the identifier *hs16mm3_Chr2_1358*. This record has coordinates in the reference genome (human) and the cross genome (mouse). In order determine penetration of this ecore into the rat, the database of mouse-rat conservation is queried for ecores that overlap the corresponding mouse coordinates of *hs16mm3_Chr2_1358.*. This method can thereby propagate across several species, until the ecore is no longer found.

- Dynamic incorporation of new species and *ecores*

The rapid emergence of completed genome assemblies is providing a growing source of input to the newly implemented system. The database framework must be able to withstand significant growth in size. Based on the premise that new species may be added to the database at any time, integration of this information into the data structure should be performed in a somewhat automated fashion. In this new system, the addition of a new cross-species alignment, results in the addition of a single table to the database. Existing tables are only updated if there is a change in genome version. Any upgrade in a whole genome assembly results in propagation of changes to all member tables using on a simplified unique identifier schema. Unlike identifiers in current data repositories, this new identifier provides an alphanumeric reference to information about the draft genomes used to provide the alignment. An example identifier was demonstrated in figure 3, and is represented below:

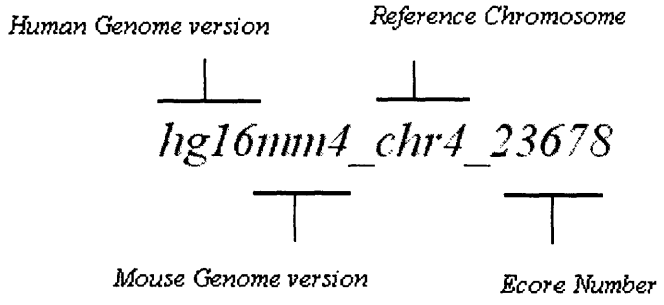


Figure 4. Construction of a unique ecore identifier

The identifier will contain the two-letter abbreviation of the reference genome as well as the version number, followed by the two-letter reference and version number of the cross-species genome. This is followed by the chromosome location on the reference genome and then the ecore number on that chromosome. When mouse version 5 is made available, for example, new alignments will be generated between hg16 and mm5. All alignments generated using mouse version 4 will be retired through a simple query referenced against the ecore identifier. In practice, since ecores are annotated with the assemblies and methodology used, they expire if they cease to be identified when the identical methodology is used to compare more mature assemblies with one another.

- Table Types and SQL Statements

The following table types were used to generate the database of evolutionary conservation.

1. Ecore Table, which stores all ecores within each reference genome, in this case, human.

```
CREATE TABLE human (
ecoreID varchar(22) NOT NULL default '',
crossSpecies varchar(30) default NULL,
chromosome varchar(10) default NULL,
start int(11) default NULL,
stop int(11) default NULL,
crossChromosome varchar(10) default NULL,
crossStart int(11) default NULL,
crossStop int(11) default NULL,
```

```
  alignScore int(11) default NULL,
  strand tinyint(1) default NULL,
  sequence text,
  crossSequence text,
  alignMethod text,
  PRIMARY KEY  (ecoreID),
  KEY chromosome (chromosome),
  KEY start (start)
  )
```

2. Gene Table, which maintains the assembly specific positional integrity of genes within each genome.

```
CREATE TABLE humanGenes (
  id int(11) NOT NULL default '0',
  org char(2) default NULL,
  name varchar(30) default NULL,
  chrom varchar(5) default NULL,
  strand char(1) default NULL,
  txstart int(11) default NULL,
  txend int(11) default NULL,
  cdsstart int(11) default NULL,
  cdsend int(11) default NULL,
  product varchar(255) default NULL,
  mrnaAcc varchar(20) default NULL,
  protAcc varchar(20) default NULL,
  refseqAcc varchar(20) default NULL,
  PRIMARY KEY  (id),
  KEY genes_ll_idx (locusLinkId),
  KEY genes_bypos_idx (chrom,txstart),
  KEY genes_bymrna_idx (mrnaAcc)
)
```

3. Exon Table, which maintains positional integrity of exons within genes.

```
CREATE TABLE exons (
  id int(11) default NULL,
  cnt int(11) default NULL,
  exstart int(11) default NULL,
  exend int(11) default NULL,
  KEY exons_idx (id,cnt)
)
```

4. Source Genome Assemblies, made available to the system as flat files, in fasta format.

Gene and exon tables are standard derivations of ncbi refSeq tables available through public data repositories, and were not developed for the purposes of this comparative genomics system, but easily incorporated into the system without additional modification. In addition to the representation of alignment data on a MySql server, source genome assemblies in fasta format, completed the database. A system update would therefore entail the consecutive update of source genome assemblies used to generate alignments, followed by an update of gene and exon tables for those assemblies, followed by an update of the ecore database. Corrections propagate through the system following the introduction of a new whole-genome assembly. Automation of the system was managed using perl, and the perl-DBI interface.

Database Implementation

User Input

Once designed, this database was implemented and tested by using a sample set of co-expressed genes as a query, with the aim of identifying putative cis-acting regulatory elements embedded within evolutionarily conserved non-coding regions. The user input form is in html and is displayed in figure 5. Input to the system had to be in standard vocabulary used to define gene transcripts across species. For this purpose NCBI MRNA Reference Sequence accession number was used to specify the genes of interest, and was the only acceptable form of user input into the system. Other accession numbers, such as Locus Link were not used since these references will soon be retired. Additionally, several of these identifiers point to more than one gene product, often including alternative splice variants of the same product.

ICE

Integration of **C**o-expression with **E**volutionary Conservation

The aim of this interface is to allow you to retreive sets of evolutionarily conserved genomic regions **(ecores)** that are associated with a set of co-expressed genes. Input data should be in the form of a list of genes identified by corresponding NCBI refSeq mrna accession number. This interface will then retreived evolutionarily conserved regions associated with these genes of interest based on the below specified parameters.

Please enter your email address:

Enter the gene list file to upload:

Select the reference genome from which this list has been derived:

◉ Human ○ Mouse ○ Rat

Next select which species you would like to choose for comparison. By selecting more than one species, you will limit your results to those regions which demonstrate evolutionary conservation across all selected species :

☐ mouse | ☐ rat | ☐ chicken | ☐ human

How far upstream of the start codon should we look for ecores?: 10000 bp.

How far downstream of the stop codon should we look for ecores?: 3000 bp.

Please choose a normalized bit score threshold for alignments: 3000 (3000 - 12000)

Please choose the minimum size for retrieved ecores: 50 bp.

Which types of regions would you like to retrieve?

◉ All Regions (Introns, exons and extragenic regions)

○ All noncoding regions (including introns)

○ Only extragenic non-coding regions

○ Only untranslated exonic regions (utrs)

Submit Query or Erase all

Since the aim of this interface is to identify putative cis-acting regulatory regions, reference genes by definition must originate from a single reference genome. This genome is designated using a radio button. Currently, only appropriate annotation information for human, rat, and mouse is accessible by the system. Following this initial implementation, however, other whole genome assemblies will be added, using the same schema.

While only a single reference genome must be specified, users have the option of selecting from more than one genome assembly for comparison. There are two possible approaches to the retrieval of ecores once these cross-species have been selected. First, the system can retrieve significant alignments between all species and the reference species. In this type of query, the number of returned regions increases as more species are selected. The alternative approach is to limit the set of ecores based on the number of cross-species selected, by restricting the results to only those ecores, which have been identified across all selected species. In this second approach, the set of retrieved ecores decreases as more species are selected, since the user is selecting ecores with increasing depth of evolutionary penetration. A commonly held assumption is that ecores with a greater depth of evolutionary penetration are more likely to be involved in basic biological functions that are ubiquitous across a larger cross-section of the living world. Conversely, ecores that are found only within certain branches of the evolutionary tree, are likely to confer more specialized functions.

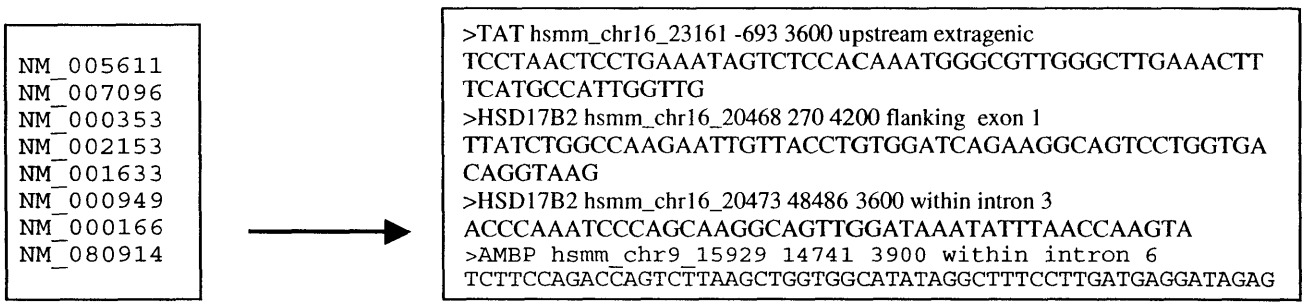
Once the cross-species are selected, the user is asked to choose the peri-genetic genomic regions region upon which to conduct the search. Users must specify upstream and downstream distances used to query the database. These have been set to the default values of 10KB upstream and 3kb downstream, however, these parameters are arbitrary, and should be adjusted by the investigator based on their preference. A considerable amount of flexibility is given to the investigator, since often very little is known about the regulatory mechanisms involved in the transcriptional activation of the genes in question. Additionally, in most vertebrates, regulator elements can be found upstream, downstream, or even in-stream of their target genes.

In addition to defining the search space, the investigator is also asked to provide a significance threshold for the retrieval of alignments. This threshold has been set using a normalized bit

score. The bit score is defined as the score determined by the match, mismatch, gap penalty matrix used to generate the alignment. While this score does returns a value that is entirely dependent on the matrix used to generate the alignment and is not reproducible across species or alignment methods, a normalized bit score does provide reproducibility. While other methods of computing significance are used in several sequence matching programs, these measures are less applicable to whole genome alignments. E values, for example provide the user with the probability that a given alignment has occurred by chance. When entire genomes are being aligned, however, E values do not provide the granularity necessary to determine which group of ecores should be studied further. Bit scores, however, do provide a measure of the relative stringency of one conserved region relative to all conserved regions generated using the same genome assembly and alignment methodology. Hence, normalization of this bit-score range across all generated alignments can provide a standard nomenclature for the retrieval of cross-species alignments. In the tightest subset of mouse-human alignments, for example, conserved regions with a bit-score over 3000 corresponded to the top 6^{th} percentile of human genomic sequence in terms of stringency in alignment to the mouse. Since the threshold value of 3000 was used previously, bit-score values in newly introduced alignments are normalized by the system, such that a score over 3000 will correspond to the top 6^{th} percentile of alignments. Hence specification of the bitscore value provides the necessary granularity needed to retrieve ecores based on alignment stringency. While the retrieval of ecores based on alignment score can be performed at the query level – if the default value of 3000 is selected, all alignments are returned by the system. In this case, filtering or clustering based on alignment stringency is still an option available to the investigator, since alignments score is a feature that is present in the returned fasta file output.

Data output

Once a list of genes is submitted, a fasta file containing ecore characteristics, as well as the ecore sequence on the reference genome is returned. An example of input and output files is specified below:



Input **Output**

The output file returned contains the conserved regions specified in fasta format, with a preliminary line identifier, which contains ecore features separated by whitespace. The current configuration returns gene symbol, unique ecoreID, distance of ecore from the translation start site, alignment score, and a verbal description of where the alignment is found. The nomenclature is standard and designates a region as intronic, extragenic, or exonic initially. Once this is determined, regions are further designated as being translated or untranslated or both. Regions are localized as belonging to a particular exon or intron, which are numbered in ascending order in the direction of transcription. If regions contain both transcribed and untranscribed sequence they are labeled as flanking. Flanking regions may or may not contain translated sequence, and are specified as such. The aim of this verbal classification system is to serve as the basis of clustering groups of ecores that share common descriptive features.

Initial application of classification Methods to System Output

Data that has been retrieved from the above interface has been used as input into a clustering algorithm, based on ncbi blast, which determines sequence similarity between all returned genes and clusters the nearest neighbors together based on stringency of alignment.[32] Input data was a set of 512 ecores, which were related to 40 co-expressed genes with unknown regulatory mechanism. The output included 6 independent clusters of ecores. The largest size cluster of 9 ecores, was entirely coding, and corresponded to translated protein domains within 3 paralogous genes. Three of the clusters corresponded to short, low complexity repeats, which were overlooked by the repeat masker program used by the alignment algorithm. The remaining 5 clusters correspond to conserved non-coding or partially coding regions, which may be involved in gene regulation. Over 90 percent of ecores did not fall into a cluster. The parameters of the clustering algorithm were set in order maximize the chances of finding sequence similarities between input genes.

Discussion and Future Directions

The implementation of this database enables investigators to retrieve evolutionary conserved genomic regions that are associated with genes of interest. The returned list of ecores is based on the most up to date whole genome sequence information that is publicly available at the time of the query. Additionally, investigators have the ability to filter and classify these evolutionarily conserved regions using a number of parameters that until this point been difficult to manipulate in a high-throughput fashion. Finally, the use of a single database of genomic evolution, coupled with the use of a standard vocabulary to define conservation across a wide variety of species provides a framework for emerging assemblies to be introduced into the database without much additional effort. This type of system design is an initial attempt at adapting to the rapid growth in the number of publicly available, fully sequenced genomes in a turnkey fashion.

In the next phase of this study, classification methods, which attempt to extract biological information from retrieved conservation data will be implemented and tested in order to clearly demonstrate the value of this system. Some of the features that will be used to classify these conserved regions will include the following:

- Depth of conservation

One potential advantage to the proposed ecore specific framework of representing conservation data, is that *ecore*s may be more easily classified based on depth of evolutionary conservation. *Ecore*s will be annotated in terms of the number of different species within which they are found to be present. They will also be annotated in terms of changing functional roles. For example, an *ecore* may have had some coding properties in primitive organisms, but may only persist as a non-coding entity throughout vertebrate evolution. These types of *ecore*s, based on their uniquely divergent functional role, would constitute a unique *ecore* family.

- Location relative to known genes

The identification of groups of *ecore*s at distinct positions from known genes is not limited simply to the groups that immediately flank exons. Annotation based on nearest genetic neighbors, position up-stream, down-stream or in-stream from these neighbors may provide a significant amount of data with which to classify these regions. In addition to known genes, chromosomal landmarks, such as the centromere, may be important reference points from which to map *ecore*s.[33, 34] Distance from these fixed landmarks may identify ecores that play an important role in functions such as mitosis, which are independent from transcription. *Ecores* will

therefore be annotated based on their proximity to chromosomal landmarks, as well as neighboring genes, and clustered using these metrics.

- Single Nucleotide Polymorphisms

As public databases of submitted and verified Human Single Nucleotide Polymorphisms become more complete, identifying the SNP density of *ecore*s or groups of *ecore*s will continue to be a potentially powerful measure of their functional importance. Future work will initially focus annotating human *ecore*s with SNP presence and location. As SNP databases from other organisms become more complete, this feature will be used to annotate *ecore*s outside of the homo-sapiens assembly as well.

- Gene-ontology

Gene-ontology terms, whenever applicable, may provide an important source of information for *ecore* family identification and classification.[35, 36] After *ecores* are localized relative to known genes, they will be linked to corresponding Gene-Ontology. This linkage will be dynamic in order to provide an up to date representation of current terminology and ontology assignment. This process will enable us to identify groups of *ecore*s, which are located relative to genes with similar functional characteristics.

- Motif identification

The primary focus of motif recognition will involve the identification of transcription factor binding motifs, using the trained hidden markov model developed by collaborators in our

laboratory. Following the implementation of this method, *ecores* belonging to the same family based on any set of features will be further classified based on the density of particular transcription factor binding sites. While the primary aim of motif identification will be on putative transcription factor binding sites, ecores demonstrating conspicuous and significant structural similarity will be further examined using a library of publicly available motif finding programs.[37-40]

Summary

While the vast non-coding space of vertebrate genomes remains a mystery, evolution provides a potentially powerful key to unlocking this mystery. The use of evolutionary theory to de-mystify the non-coding genome will, in part, be a function of the number of completely sequenced genomes available for comparison. Our ability to utilize this data, however, will be limited without the implementation of a standard vocabulary. This paper defines a computational framework for the representation of evolutionary conservation across multiple species using a standard vocabulary and a relational database. In this framework, an evolutionarily conserved region (ecore) is defined as a distinct genomic entity, which has sequence information, positional information, and can extend across two or more species. The use of this standard vocabulary enables the database to evolve in response to rapid growth of whole genomic databanks, thus maximizing the discovery potential of cross-species comparisons. Future studies will attempt to validate the use of this system through the characterization of ecores, which have been found to be associated with sets of co-expressed genes. Information about the availability of this resource will be posted at the Children's Hospital Informatics Program Website www.chip.org.

Bibliography

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
2. **Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium**. *Science* 1998, **282**(5396):2012-2018.
3. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al*: **The genome sequence of Drosophila melanogaster**. *Science* 2000, **287**(5461):2185-2195.
4. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P *et al*: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**(6915):520-562.
5. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A *et al*: **Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes**. *Science* 2002, **297**(5585):1301-1310.
6. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R *et al*: **The genome sequence of the malaria mosquito Anopheles gambiae**. *Science* 2002, **298**(5591):129-149.
7. Twigger S, Lu J, Shimoyama M, Chen D, Pasko D, Long H, Ginster J, Chen CF, Nigam R, Kwitek A *et al*: **Rat Genome Database (RGD): mapping disease onto the genome**. *Nucleic Acids Res* 2002, **30**(1):125-128.
8. Harris TW, Lee R, Schwarz E, Bradnam K, Lawson D, Chen W, Blasier D, Kenny E, Cunningham F, Kishore R *et al*: **WormBase: a cross-species database for comparative genomics**. *Nucleic Acids Res* 2003, **31**(1):133-137.
9. Kimura M: **Molecular evolutionary clock and the neutral theory**. *J Mol Evol* 1987, **26**(1-2):24-33.
10. Wildman DE, Uddin M, Liu G, Grossman LI, Goodman M: **Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus Homo**. *Proc Natl Acad Sci U S A* 2003, **100**(12):7181-7188.
11. Pamilo P, Bianchi NO: **Evolution of the Zfx and Zfy genes: rates and interdependence between the genes**. *Mol Biol Evol* 1993, **10**(2):271-281.
12. Benton MJ, Ayala FJ: **Dating the tree of life**. *Science* 2003, **300**(5626):1698-1700.
13. Gilligan P, Brenner S, Venkatesh B: **Fugu and human sequence comparison identifies novel human genes and conserved non-coding sequences**. *Gene* 2002, **294**(1-2):35-44.
14. Venkatesh B, Gilligan P, Brenner S: **Fugu: a compact vertebrate reference genome**. *FEBS Lett* 2000, **476**(1-2):3-7.
15. Levy S, Hannenhalli S, Workman C: **Enrichment of regulatory signals in conserved non-coding genomic sequence**. *Bioinformatics* 2001, **17**(10):871-877.
16. Levy S, Hannenhalli S: **Identification of transcription factor binding sites in the human genome sequence**. *Mamm Genome* 2002, **13**(9):510-514.

17. Dieterich C, Cusack B, Wang H, Rateitschak K, Krause A, Vingron M: **Annotating regulatory DNA based on man-mouse genomic comparison.** *Bioinformatics* 2002, **18 Suppl 2**:S84-90.
18. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE: **Human-mouse genome comparisons to locate regulatory sites.** *Nat Genet* 2000, **26**(2):225-228.
19. Moss EG, Tang L: **Conservation of the heterochronic regulator Lin-28, its developmental expression and microRNA complementary sites.** *Dev Biol* 2003, **258**(2):432-442.
20. Lai EC, Tomancak P, Williams RW, Rubin GM: **Computational identification of Drosophila microRNA genes.** *Genome Biol* 2003, **4**(7):R42.
21. Ambros V: **MicroRNA Pathways in Flies and Worms. Growth, Death, Fat, Stress, and Timing.** *Cell* 2003, **114**(2):269.
22. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ *et al*: **The UCSC Genome Browser Database.** *Nucleic Acids Res* 2003, **31**(1):51-54.
23. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12**(4):656-664.
24. Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V *et al*: **Ensembl 2002: accommodating comparative genomics.** *Nucleic Acids Res* 2003, **31**(1):38-42.
25. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T *et al*: **The Ensembl genome database project.** *Nucleic Acids Res* 2002, **30**(1):38-41.
26. Bray N, Dubchak I, Pachter L: **AVID: A global alignment program.** *Genome Res* 2003, **13**(1):97-102.
27. Ureta-Vidal A, Ettwiller L, Birney E: **Comparative genomics: genome-wide analysis in metazoan eukaryotes.** *Nat Rev Genet* 2003, **4**(4):251-262.
28. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *J Mol Biol* 1970, **48**(3):443-453.
29. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147**(1):195-197.
30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403-410.
31. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: **Human-mouse alignments with BLASTZ.** *Genome Res* 2003, **13**(1):103-107.
32. Ilya Dondoshansky YW: **Blastclust**. In., 1 edn. Washington DC: National Center for Biotechnology Information; 1998.
33. Malik HS, Vermaak D, Henikoff S: **Recurrent evolution of DNA-binding motifs in the Drosophila centromeric histone.** *Proc Natl Acad Sci U S A* 2002, **99**(3):1449-1454.
34. Abad JP, Villasante A: **Searching for a common centromeric structural motif: Drosophila centromeric satellite DNAs show propensity to form telomeric-like unusual DNA structures.** *Genetica* 2000, **109**(1-2):71-75.
35. Lord PW, Stevens RD, Brass A, Goble CA: **Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.** *Bioinformatics* 2003, **19**(10):1275-1283.

36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**(1):25-29.
37. Liu XS, Brutlag DL, Liu JS: **An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments.** *Nat Biotechnol* 2002, **20**(8):835-839.
38. Liu X, Brutlag DL, Liu JS: **BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.** *Pac Symp Biocomput* 2001:127-138.
39. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22**(3):281-285.
40. Rigoutsos I, Floratos A: **Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm.** *Bioinformatics* 1998, **14**(1):55-67.