Comparative genomics tools for biological discovery

Lior Pachter
Dept. Mathematics, U.C. Berkeley
and LBNL
lpachter@math.berkeley.edu

Inna Dubchak, LBNL

Outline

What is comparative genomics?

VISTA tools developed for comparative genomics.

Related biological stories

Large scale VISTA applications including automatic computational system for comparing the human and mouse genomes
The Human genome

From the Nature paper:

The next steps:

Developing the IGI (integrated gene index) and IPI (integrated protein index)
Large-scale identification of regulatory regions
Sequencing of additional large genomes
Completing the catalogue of human variation
From sequence to function
Distant Non-Coding Sequences Causing Disease

<table>
<thead>
<tr>
<th>Disease</th>
<th>Gene</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campomelic displasia</td>
<td>SOX9</td>
<td>850kb</td>
</tr>
<tr>
<td>Aniridia</td>
<td>PAX6</td>
<td>125kb</td>
</tr>
<tr>
<td>X-Linked Deafness</td>
<td>POU3F4</td>
<td>900kb</td>
</tr>
<tr>
<td>Saethre-Chotzen syndrome</td>
<td>TWIST</td>
<td>250kb</td>
</tr>
<tr>
<td>Rieger syndrome</td>
<td>PITX2</td>
<td>90kb</td>
</tr>
<tr>
<td>Split hand/split foot malformation</td>
<td>SHFM1</td>
<td>450kb</td>
</tr>
</tbody>
</table>
**Background**

Evolution can help!

In general, functionally important sequences are conserved

Conserved sequences are functionally important

Raw sequence can help in finding biological function

---

**Comparison of 1196 orthologous genes**  
(*Makalowski et al., 1996*)

- Sequence identity:
  - exons: 84.6%
  - protein: 85.4%
  - introns: 35%
  - 5' UTRs: 67%
  - 3' UTRs: 69%

- 27 proteins were 100% identical

*Integrating data into more powerful gene prediction models than with human genomic sequence alone.*
Comparing sequences of different organisms

- Helps in gene predictions
- Helps in understanding evolution
- Conserved between species non-coding sequences are reliable guides to regulatory elements
- Differences between evolutionary closely related sequences help to discover gene functions

Challenges

- Sequence at different stages of completion, difficult to compare
  - Whole genome shotgun
  - Finished BACs
- Fast and accurate analysis
- Scaling up to the size of whole genomes
http://www-gsd.lbl.gov/vista

Processed ~ 11000 queries on-line, distributed > 560 copies of the program in 34 countries

Modules of VISTA:

• Program for global alignment of DNA fragments of any length

• Visualization of alignment and various sequence features for any number of species

• Evaluation and retrieval of all regions with predefined levels of conservation
Sequence comparisons. How?

Three variations:

Find the best *OVERALL* alignment.

*Global alignment*

Find *ALL* regions of similarity.

*Local alignment*

Find the *BEST* region of similarity.

*Optimal local alignment*

Aligning large genomic regions

- Long sequences lead to memory problems
- Speed becomes an issue
- Long alignments are very sensitive to parameters
- Draft sequences present a nontrivial problem
- Accuracy is difficult to measure and to achieve

References for some existing programs:

*Glass:*
Domino Tiling, Gene Recognition, and Mice.

*Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction.*

*MUMmer*

*PipMaker*

*Scan2*
Dbscan/Scan2: Fast alignment of mega-sequences.

Seledtsov I.A., Solovyev V.V. To Appear. Web site  http://softberry.com/
Local alignment algorithms are designed to search for highly similar regions in two sequences that may not be highly similar in their entirety. The algorithm works by first finding very short common segments between the input sequence and database sequences, and then expanding out the matching regions as far as possible.

For cross-species comparison one needs to accurately align two complete sequences. It is insufficient to find common similar regions in the two sequences, rather, what is needed is a global map specifying how the two sequences fit together, much like understanding how the pieces in a puzzle connect up with each other.

This problem is called global alignment.
**AVID- the alignment engine behind VISTA**

- **Very fast** global alignment of megabases of sequence.
- **Provides details** about ordered and oriented contigs, and accurate placement in the finished sequence.
- **Full integration** with repeat masking

• ORDER and ORIENT
• FIND all common k-long words (k-mers)
• ALIGN k-mers scoring by local homology
• FIX k-mers with good local homology
• RECURSE with smaller k (shorter words)

**Visualization**

`tgattacattcaattatg------ttctcaaaagtgagcatgaca-accttttttccatgg`  
`tgatgacatctatttggcctttttagaaactgcatgagacctggctaggg`

Window of length L is centered at a particular nucleotide in the base sequence.

Percent of identical nucleotides in L positions of the alignment is calculated and plotted.

Move to the next nucleotide.
Finding conserved regions with percentage and length cutoffs

Conserved segments with percent identity X and length Y - regions in which every contiguous subsegment of length Y was at least X% identical to its paired sequence. These segments are merged to define the conserved regions.

Output:

11054 - 11156 = 103bp at 77.670%        NONCODING
13241 - 13453 = 213bp at 87.793%        EXON
14698 - 14822 = 125bp at 84.800%        EXON

VISTA plot

KIF Gene

Conserved Non Coding Sequences

% Identity Between Humans/Mice (Vertical Axis)

Human Sequence (horizontal axis)
Nuclear Hormone Receptor: LXR-Alpha

Human/Mouse

Human/Rabbit

Human/Opossum
**Multi-Species Comparative Analysis (VISTA)**

- **Apolipoprotein AI gene**
  - human/macaque: 100%
  - human/pig: 75%
  - human/rabbit: 50/100%
  - human/mouse: 75%
  - human/rat: 50/100%
  - human/chicken: 75%

**Liver enhancer**
VISTA input files

• Sequences of two or more organisms
• Annotation file for a base sequence if available

VISTA output files

• All pair wise global alignments of the sequences
• VISTA plot
• The list of conserved regions at predefined by the user length and conservation cutoffs

VISTA flavors

• VISTA - comparing DNA of multiple organisms
• for 3 species - analyzing cutoffs to define actively conserved non-coding sequences
• cVISTA - comparing two closely related species
• rVISTA - regulatory VISTA
Active conservation of noncoding sequences - present in more than two mammals

% Cutoff
sum of three pair wise Intersection/Union values is maximal

Over 120 basepairs:
- H/D > 92%
- H/M > 80%
- D/M > 77%

Example: Dubchak et al., 2000, Genome Research, 10: 1304-1306.
Identifying non-coding sequences (CNSs) involved in transcriptional regulation

rVISTA - prediction of transcription factor binding sites

- Simultaneous searches of the major transcription factor binding site database (Transfac) and the use of global sequence alignment to sieve through the data

- Combination of database searches with comparative sequence analysis reduces the number of predicted transcription factor binding sites by several orders of magnitude
Regulatory VISTA (rVISTA)

1. Identify potential transcription factor binding sites for each sequence using library of matrices (TRANSFAC)
2. Identify aligned sites using VISTA
3. Identify conserved sites using dynamic shifting window

Percentage of conserved sites of the total 3-5%

~1 Meg region, 5q31

<table>
<thead>
<tr>
<th>Coding</th>
<th>Noncoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human interval Transfac predictions for GATA sites</td>
<td>839 20654</td>
</tr>
<tr>
<td>Aligned with the same predicted site in the mouse seq.</td>
<td>450 2618</td>
</tr>
<tr>
<td>Alligned sites conserved at 80% / 24 bp dynamic window</td>
<td>303 731</td>
</tr>
<tr>
<td>Random DNA sequence of the same length</td>
<td>29280</td>
</tr>
</tbody>
</table>
2 Exp. Verified GATA-3 Sites

GATA-3 (28) → IL 5

GATA-3 Conserved (4)

Choose Families to Visualize

Select Picture Properties, Clustering, and Type of Visualization
Sequence motif recognition + multiple sequence alignment of syntenic regions, a high throughput strategy for filtering and prioritizing putative DNA binding sites genomically informed starting place for globally investigating detailed regulation.
Main features of VISTA

- Clear, configurable output
- Ability to visualize several global alignments on the same scale
- Alignments up to several megabases
- Working with finished and draft sequences
- Available source code and WEB site

Related publications

- **ONE**

- **TWO**
What if you don’t have a sequence of other species for the region of your interest?

Are there publicly available comparative genomics data?

Large scale VISTA applications:

Cardiovascular comparative genomics database
http://pga.lbl.gov

Godzilla – comparing the human and mouse genome
http://pipeline.lbl.gov

Godzilla - automatic computational system for comparative analysis of genomes


DATA
Base Human Genome - Golden Path Assembly

Mouse assemblies:
Arachne October 2001
Phusion November 2001
MGSC v3 April 2002
Main modules of the system

Mapping and alignment of mouse contigs against the human genome

Visualization

Analysis of conservation

Chromosome Comparison

Base pair alignment

247 GGTGAGGTACGGTGACCCTGCA CGGAGCTTATGGGAGGCA AGAGC
|:   ||  ||||:  |||| --:||  ||| |::|   |||---|||||
368 GAGTCGGGGGAGGGGGCTGCTGTTGGCTCTGGACAGCTTGCATTGAGAGG
Tandem Local/Global Alignment Approach

Sequence fragment anchoring (DNA and/or translated BLAT)
Multi-step verification of potential regions using global alignment (AVID)

Advantage of the tandem approach:
better sensitivity/specificity trade-off
fill-in effect
scoring longer alignments
Visualization – VistaBrowser & VistaTrack

Stand-alone Java applet for detailed comparison

Comparison combined with the human genome annotation on the UCSC Human Genome Browser
**MyGodzilla** - is an interactive web tool for comparing your favorite sequence against the human genome
MyGodzilla Tool

Submit a DNA sequence of ANY organism...
... or submit a whole chromosome and analyze another Genome

Your request is submitted and the results will be ready at this link: http://pipeline.lbl.gov/cache/24b/22599550.8711

to receive an automatic notification, type your email here: email me

Examples of Results

• Understanding the structure of conservation

• Identification of putative functional sites

• Discovery of new genes

• Detection of contamination and misassemblies
Two assemblies are better than one

Identification of a New Apo Gene on Human 11q23

Godzilla

UCSC Genome Browser on Human: Aug. 6, 2001 Freeze

Highly Conserved Region

Zoom In

ApoA4  ApoC3  ApoA1
Identification of a New Apo Gene on Human 11q23

Godzilla

Finding regulatory regions

Muscle Specific Regulatory Region: human beta enolase intronic enhancer
Comparative analysis of genomic intervals containing important cardiovascular genes

http://pga.lbl.gov

http://pga.lbl.gov/cvcgd.html
Example of CVCGD entry

Solute carrier family 22, organic cation transporter number 4 (SLC22A4, OCTN1)

- Category: ABC transporters
- Gene ID in the OMEM database: 606190
- Human cDNA accession: Sg3f
- GenBank accession number for human cDNA: NM_000359
- Mouse map location: 11
- GenBank accession number for mouse cDNA: NM_010677
- Annotation of the human sequence
- Human genome alignment

Note: If your browser hangs or crashes on the alignment page you can try this link instead.

Short annotation of the region

Annotation of the VA5q31 region *

94869 bp

- Assembly contains a deletion of 10022 bp after the 5′ exon of the BEK gene

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Identity/Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBEK-3</td>
<td>Identical to SBEK3</td>
</tr>
<tr>
<td>SBEK-4</td>
<td>Identical to SBEK3</td>
</tr>
<tr>
<td>SBEK-5</td>
<td>Identical to SBEK3</td>
</tr>
<tr>
<td>SBEK-6</td>
<td>Identical to SBEK3</td>
</tr>
<tr>
<td>SBEK-7</td>
<td>Identical to SBEK3</td>
</tr>
<tr>
<td>SBEK-8</td>
<td>Identical to SBEK3</td>
</tr>
<tr>
<td>SBEK-9</td>
<td>Identical to SBEK3</td>
</tr>
</tbody>
</table>
Detailed annotation in AceDB format

![Detailed annotation in AceDB format](image1)

VISTA plot of the region

![VISTA plot of the region](image2)
multiVISTA plot of the region

Genomic region containing Apolipoprotein A-I (APOA1)
This plot is not clickable. In order to view alignment regions please go back to the gene page and click on the alignment you are interested in.

Alignment

<p>| Genomic region containing Solute carrier family 22, organic cation transporter member |
|-----------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|</p>
<table>
<thead>
<tr>
<th>seq1: human</th>
<th>seq2: mouse</th>
<th>seq1: rat</th>
<th>seq2: <em>dan</em></th>
<th>seq1: <em>tak</em></th>
<th>seq2: <em>rif</em></th>
<th>seq1: <em>sau</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>CAGGCTCGCAAGAGCTACGAGAGG</td>
<td>CAGGCTCGCAAGAGCTACGAGAGG</td>
<td>CAGGCTCGCAAGAGCTACGAGAGG</td>
<td>CAGGCTCGCAAGAGCTACGAGAGG</td>
<td>CAGGCTCGCAAGAGCTACGAGAGG</td>
<td>CAGGCTCGCAAGAGCTACGAGAGG</td>
<td>CAGGCTCGCAAGAGCTACGAGAGG</td>
</tr>
<tr>
<td>6650</td>
<td>6650</td>
<td>6650</td>
<td>6650</td>
<td>6650</td>
<td>6650</td>
<td>6650</td>
</tr>
<tr>
<td>7020</td>
<td>7020</td>
<td>7020</td>
<td>7020</td>
<td>7020</td>
<td>7020</td>
<td>7020</td>
</tr>
<tr>
<td>seq1: <em>nov</em></td>
<td>seq2: <em>pis</em></td>
<td>seq1: <em>sau</em></td>
<td>seq2: <em>pis</em></td>
<td>seq1: <em>sau</em></td>
<td>seq2: <em>pis</em></td>
<td>seq1: <em>sau</em></td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>ACGGCTCGCAAGAGCTACGAGAGG</td>
<td>ACGGCTCGCAAGAGCTACGAGAGG</td>
<td>ACGGCTCGCAAGAGCTACGAGAGG</td>
<td>ACGGCTCGCAAGAGCTACGAGAGG</td>
<td>ACGGCTCGCAAGAGCTACGAGAGG</td>
<td>ACGGCTCGCAAGAGCTACGAGAGG</td>
<td>ACGGCTCGCAAGAGCTACGAGAGG</td>
</tr>
<tr>
<td>7100</td>
<td>7100</td>
<td>7100</td>
<td>7100</td>
<td>7100</td>
<td>7100</td>
<td>7100</td>
</tr>
<tr>
<td>seq1: <em>agr</em></td>
<td>seq2: <em>pis</em></td>
<td>seq1: <em>sau</em></td>
<td>seq2: <em>pis</em></td>
<td>seq1: <em>sau</em></td>
<td>seq2: <em>pis</em></td>
<td>seq1: <em>sau</em></td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>TACGCTCGCAAGAGCTACGAGAGG</td>
<td>TACGCTCGCAAGAGCTACGAGAGG</td>
<td>TACGCTCGCAAGAGCTACGAGAGG</td>
<td>TACGCTCGCAAGAGCTACGAGAGG</td>
<td>TACGCTCGCAAGAGCTACGAGAGG</td>
<td>TACGCTCGCAAGAGCTACGAGAGG</td>
<td>TACGCTCGCAAGAGCTACGAGAGG</td>
</tr>
<tr>
<td>8250</td>
<td>8250</td>
<td>8250</td>
<td>8250</td>
<td>8250</td>
<td>8250</td>
<td>8250</td>
</tr>
</tbody>
</table>
## Conserved regions

### Genomic region containing Solute carrier family 22, organic cation transporter member 4 (SLC22A4, OCTN1)

Criteria: 70% identity over 100 bp

<table>
<thead>
<tr>
<th>Conserved Regions</th>
<th>Human (mouse)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1469 to 1515 (426)</td>
<td>43bp at 85.1% exons</td>
</tr>
<tr>
<td>2608 to 2837 (2291)</td>
<td>155bp at 90.9% exons</td>
</tr>
<tr>
<td>4016 to 4370 (3543)</td>
<td>352bp at 100.0% exons</td>
</tr>
<tr>
<td>4044 to 4093 (490)</td>
<td>50bp at 87.4% exons</td>
</tr>
<tr>
<td>6717 to 7014 (2977)</td>
<td>297bp at 87.9% exons</td>
</tr>
<tr>
<td>10859 to 10927 (897)</td>
<td>68bp at 91.9% exons</td>
</tr>
<tr>
<td>13503 to 13753 (2511)</td>
<td>251bp at 91.9% exons</td>
</tr>
<tr>
<td>14508 to 14612 (1065)</td>
<td>113bp at 76.5% exons</td>
</tr>
<tr>
<td>14978 to 18000 (3022)</td>
<td>128bp at 76.6% exons</td>
</tr>
<tr>
<td>14784 to 14875 (901)</td>
<td>51bp at 89.6% exons</td>
</tr>
<tr>
<td>15779 to 15860 (882)</td>
<td>882bp at 83.9% exons</td>
</tr>
<tr>
<td>15975 to 16111 (1336)</td>
<td>1336bp at 90.9% exons</td>
</tr>
<tr>
<td>16165 to 16416 (2451)</td>
<td>2451bp at 91.7% exons</td>
</tr>
<tr>
<td>16437 to 16535 (108)</td>
<td>108bp at 100.0% exons</td>
</tr>
<tr>
<td>17524 to 17647 (1233)</td>
<td>1233bp at 97.9% exons</td>
</tr>
</tbody>
</table>

---

The SLAM server: submit pairs of syntenic sequences for gene annotation and alignment

![SLAM server](http://bio.math.berkeley.edu/slam/)

---

http://bio.math.berkeley.edu/slam/
SLAM components

- Splice site detector
  - VLMM
- Intron and intergenic regions
  - 2nd order Markov chain
  - independent geometric lengths
- Coding sequence
  - PHMM on protein level
  - generalized length distribution
- Conserved non-coding sequence
  - PHMM on DNA level

SLAM input and output

- Input:
  - Pair of syntenic sequences (FASTA).
- Output:
  - CDS and CNS predictions in both sequences.
  - Protein predictions.
  - Protein and CNS alignment.
Publications on our tools:


Related sites

• The Human Genome Browser & BLAT program http://genome.ucsc.edu/

• ENSEMBLE Project (Sanger Center) http://www.ensembl.org/

• AVID alignment program http://baboon.math.berkeley.edu/~syntenic/avid.html

• SLAM comparative gene prediction program http://bio.math.berkeley.edu/slam/mouse/

• PSU group’s MHC Human-Mouse comparison results http://bio.cse.psu.edu/mousegroup/MHC/

• PSU Pipmaker program http://bio.cse.psu.edu/pipmaker/
Summary

Suite of comparative genomics tools VISTA
http://www-gsd.lbl.gov

Godzilla comparing the human and mouse genome
http://pipeline.lbl.gov

Cardiovascular comparative genomics database
http://pga.lbl.gov

Questions? Write to vista@lbl.gov

Towards Better VISTAs

Information from a Single Sequence Alone

Multi-Organism High Quality Sequences
# Thanks

<table>
<thead>
<tr>
<th>Biology</th>
<th>Bioinformatics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kelly Frazer</td>
<td>Michael Brudno</td>
</tr>
<tr>
<td>Gaby Loots</td>
<td>Olivier Couronne</td>
</tr>
<tr>
<td>Len Pennacchio</td>
<td>Brian Klock</td>
</tr>
<tr>
<td></td>
<td>Chris Mayor</td>
</tr>
<tr>
<td></td>
<td>Ivan Ovcharenko</td>
</tr>
<tr>
<td></td>
<td>Alexander Poliakov</td>
</tr>
<tr>
<td></td>
<td>Jody Schwartz</td>
</tr>
<tr>
<td>Eddy Rubin</td>
<td>Inna Dubchak</td>
</tr>
</tbody>
</table>

Funding - Programs for Genomic Applications (PGA) by NHLBI