Getting around genomes

Bi 188

Ali Mortazavi

3/26/07

Genomics

- Biology is evolving from a data poor science to a data-rich one
- Microarrays and high-throughput methods are generate 100k to millions of data points per experiment
- Analysis has become the bottleneck to new discovery
- Much of this is due to the rise of genomics in the last fifteen years (30 if you are Barbara)

Exploring vertebrate genomes

- All living organisms have genomes
- Vertebrates have relatively large genomes:
  - 0.4 Gb to 4 Gb
  - About 25k genes
- Only a handful of these genes have been studied intensively, and typically only in one or two model organisms
- Genomics deals with genome-scale properties of genes (and other functional elements)

What is a gene ?

- Not as trivial as it sounds….and what does that make the rest of the genome ?
- For our purposes, genes are a region of the genome to which we can attach a set of attributes:
  - On chromosome 4 (positional information)
  - Nuclear (localization)
  - Transcriptional Repressor Activity(function)
  - Not expressed in the brain (expression)
  - Highly conserved in vertebrates (comparative)

Positional Information (1)

- Because of technology “limitations”, genomes are sequenced in small chunks of several hundred bp and assembled into virtual chromosomes (“pseudo-molecules”) using sophisticated algorithms
- Genomes as software - assembly are periodically redone with additional data and improved algorithms, which results in new versions (“builds”)
- Absolute positions of genes can move by 10+ Mb between builds

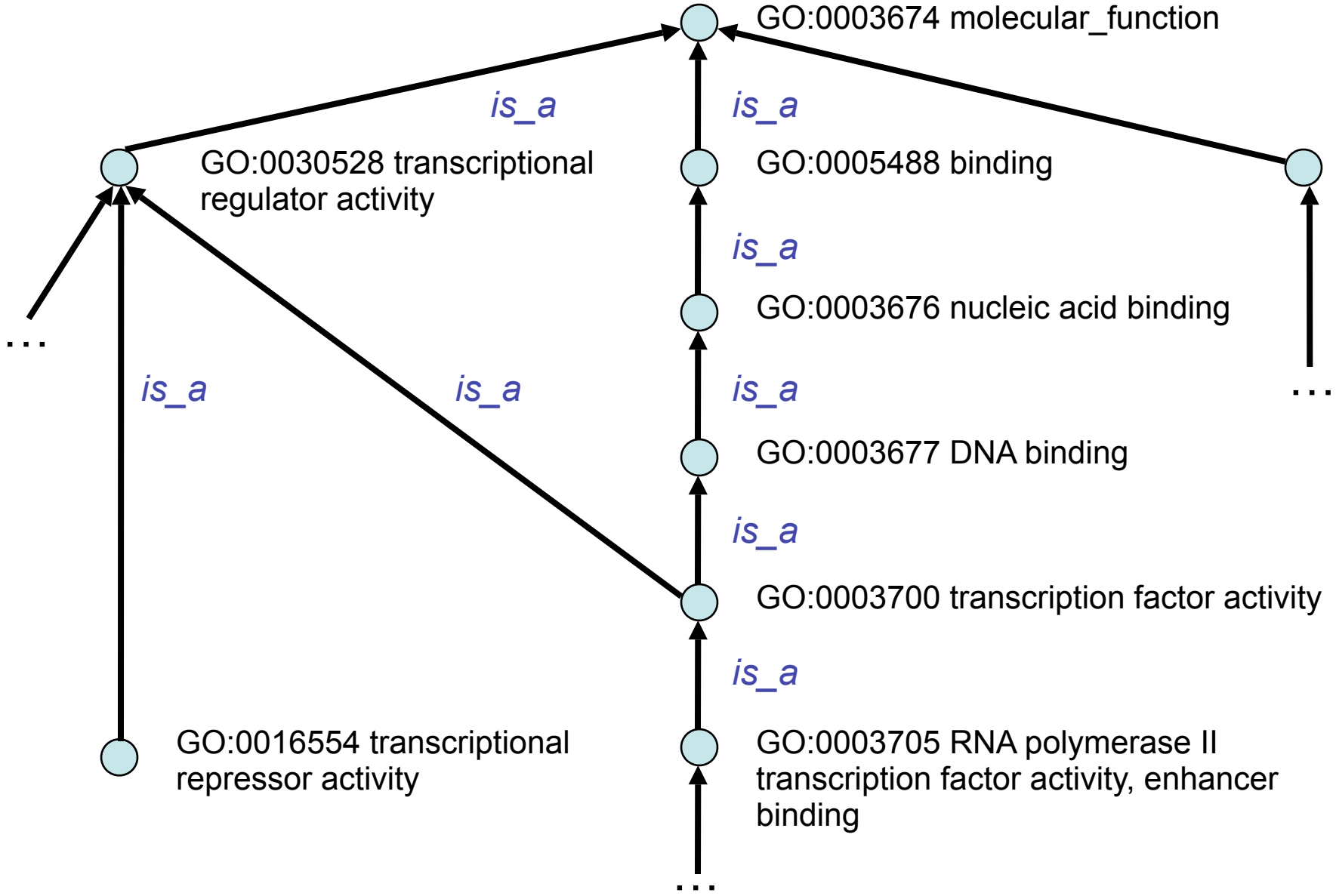
Positional Information (2)

- Once a build is generated, it is annotated in a multitude of ways:
  - Mapping of mRNA of known genes
  - *de novo* predictions of genes
  - Repeats
  - microRNAs
  - Expression and other data types (typically by moving data from previous builds)
- The sequence and the annotations are the input material of most genomic analyses.

Gene Ontology (GO)

- A structured vocabulary to describe gene to describe:
  - Sub-cellular location
  - Biological process
  - Molecular function
- GO Annotations can derive from the literature, from automated annotations, or from GO terms for the same gene in another genome (an ortholog)

GO is a directed, acyclic graph



Gene Expression

- The expression pattern of a gene in relation to that of other genes is most informative
- Clustering of genes and conditions (e.g. tissues, conditions) relies on a multitude of machine learning techniques
- The meaning of these clusters is still up to the scientist to determine

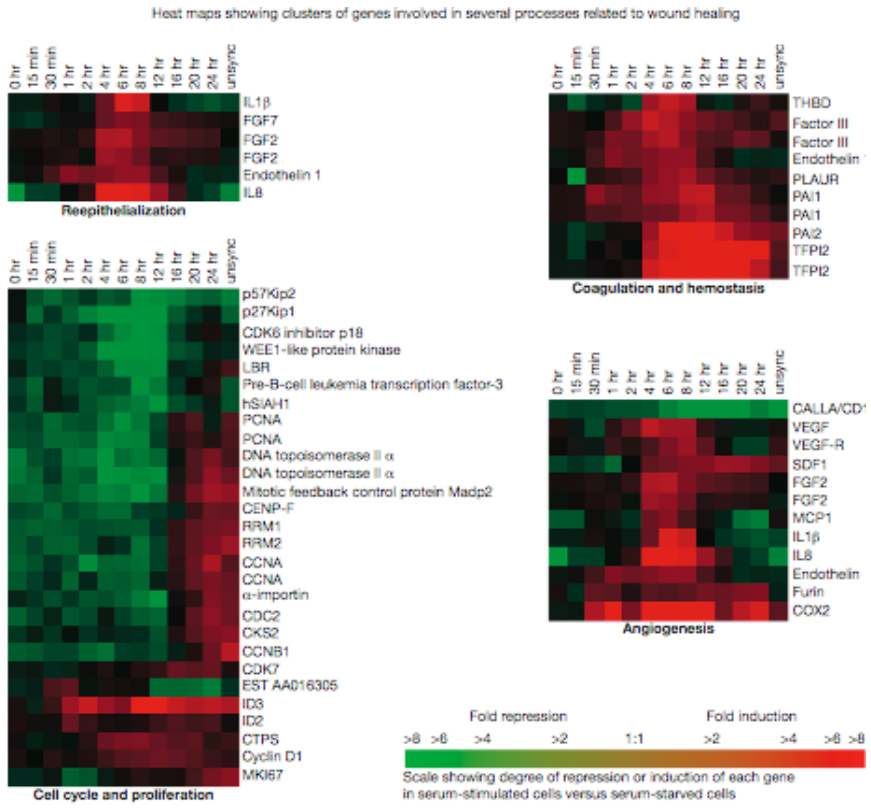


Fig 13-6

Comparative Genomics

- If it's conserved, it must be important!
- Tremendous amounts of effort have been invested in identifying regions of the genomes that are under purifying selection.
- Of course, some of the most interesting parts of human evolution would not show up as conserved:
  - Human Accelerated Regions (HARs)

Exploring the Genome

- All of the genomics data is readily available on the web.
- Multiple portals to this data exist:
  - NCBI
  - Ensembl
- In this class, we will focus primarily on one:
  - UCSC's Genome Bioinformatics Site (http://genome.ucsc.edu)
- Demo