Genome Annotation

Phil McClean
September 2005

The most time consuming and costliest aspect of the early stages of a genome project is the collecting the DNA sequence of a genome. This is a linear collection of all the sequences that define the species. But as a dataset, this sequence itself is devoid of content. The genome must be annotated, or described, in a manner that can be of use to biologists of all types. Given the wealth of information that is found within the sequence, it is not too much of a stretch to consider the possibility that in the years following the publication of the sequence much more time and money will be spent on deciphering all the nuances and subtleties.

When genome annotation is first mentioned, what typically comes to mind is the description of the genes that are distributed throughout the genome. The genes themselves contain a wealth of information that helps to describe the species. It is these genes whose collective expression define what the species will look like through out its life cycle, how it will reproduce, and the manner it will respond to its environment. Individually, the coding region of a gene contains the information that defines the nature of an expressed protein or a functional RNA molecule. In the controlling regions of a gene, sequences can be discovered that define where, when, and the degree to which the gene will be expressed. As you can imagine defining genes and their control regions is one aspect of the genome sequence that is of interest to many researchers. But this is not the only information that is important.

Complex eukaryotes also contain a repetitive class of sequences. The best know and well described repetitive sequences are the transposable elements. These include the DNA elements similar to those first described in corn by McClintock. But these are not the most abundant transposable element class. The distinction is held by the retrotransposons class of elements that move via an RNA intermediate. The shear fact that nearly half of the human genome consists of transposable elements makes them a significant research topic.

Simple sequence repeats, or SSRs, are another major repetitive class found in genomes. These repeats, that consist of localized repetitions of di- or tri-nucleotides, have been well described in many species and currently are widely used as markers linked to other genes in the genome. Using *in silico* approaches to uncover the proximity relationship of SSRs to all genes has tremendous potential for diagnostic purposes.

Genomes have long histories involving many different types of events that lead to the construction of the current genomic landscape. We are now beginning to appreciate the extent to which genomes evolved by large scale, or segmental duplications. Take the case of *Arabidopsis*. It was chosen as a model organism because it was considered to be essentially devoid of gene duplications. Early studies suggested that the genome was essentially single copy. But the final sequence of the *Arabidopsis* genome revealed that nearly all of the genome had undergone duplications on a large scale. Similar patterns of local (intrachromosomal) and distal

(interchromosomal) duplications were also noted for the human, but to a lesser degree in the mouse genome.

A final non-gene description of a genome characterizes single nucleotide polymorphisms (SNPs). Allelic differences between genes underlie the variation in gene expression. These can be large scale deletions that render a gene essentially useless, or a triplet deletion that causes the loss of a key amino acid and a subsequent loss of function. A famous example is the triplet nucleotide deletion in the cystic fibrosis gene that leads to the loss of an important phenylalanine amino acid. Other differences, though are much more subtle such as the single nucleotide difference in the β-hemaglobin gene that generates the sickle cell anemia allele. Such differences between the two alleles is called a SNP. If a genome project uses multiple individuals as there DNA source, SNPs can be uncovered. These can then be used in association mapping studies to determine if a particular allelic difference is responsible for an unusual phenotype. The public human genome project uncovered 1.4 million SNPs, and other projects are searching for more. The current estimate is that the human genome contains 10 million SNPs.

Annotation Approaches

Nucleotide annotation. The first step of nucleotide annotation is to find a sequence that has the features of a gene. Many eukaryotic genes contain specific features, such as introns that separate exons, that can serve are markers for the discovery process. Therefore, it is important to develop a software program that properly recognizes such features. A number of programs are available that perform these searches. A key feature of each of these programs are ***sensor*** algorithms that identifies the key structural features. For genes, these would include, for example, introns that are defined by the consensus splice site junctions (GT…AG). The program might also include other sensors that detect a transcriptional start site or recognize specific GC content. Collectively, potential genes are discovered by scanning the DNA sequence in all six possible reading frames to ensure all possible genes are recognized for further analysis.

Once a sequence has been defined as a gene, the next step is to name it. The naming of genes relies upon the significant amount of research that predated genome projects. This research was historically done on a gene-by-gene approach to clone and characterize individual genes that were of interest to a specific research group. For example, many of the proteins involved in the housekeeping processes of a cell have been characterized at the nucleotide and protein levels. This information is stored in large databases such as GenBank and Swiss-Prot. Therefore with a specific sequence highlighted as a potential gene, the next step is to determine if that sequence indeed is like some other gene or protein.

Naming the genes. The software tool most often used to annotate (or name) a gene is BLAST. This stands for **B**asic **L**ocal **A**lignment **S**earch **T**ool. This series of computer programs looks for sequence similarities. Basically, it consists of a query (the sequence to which you are looking for a match) and a database. Typically, a database such as GenBank or Swiss-Prot is used to uncover sequences that are similar to the query. The query can be either a protein or nucleotide sequence. The database can be either a nucleotide or protein database. It is also possible to use a translated version of a nucleotide query, or to search a translated version of a

database when the input query is an amino acid sequence. The more recent versions of BLAST can incorporate gaps to discover homologies.

A critical database used to determine if a sequence is indeed a gene contains EST (**E**xpressed **S**equence **T**ags) sequences. ESTs are DNA sequences of expressed genes that are represented in a cDNA library. The data is collected by end sequencing (usually the 3' end) a large collection of clones representing transcripts expressed under a specific developmental or environmental condition. Because they are expressed, then they clearly were transcribed from functioning gene. Therefore, the predicted genes are also used as a BLAST query against an EST database

Non-gene RNA sequences. Programs are also available that searce for non-gene RNA sequences that are important components of the genome. These sequences include the ribosomal RNAs and tRNAs that are essential for protein translation. In addition, the small nuclear RNAs important to processes such as RNA splicing are necessarily components of the gneome. These sequences exhibit a high degree of conservation, and therefore, are easily recognizable. Finally, the recent discoveyr of microRNAs has added another sequence class. These RNAs act as suppressors of gene expression by binding to the mRNA of specific genes.

The abundance of research that targets the discovery of regulatory regions in the genome has described many short sequences (motifs) that are target sites to which proteins that regulate gene expression bind. Similarity searches of these short motifs are relatively straightforward. The success of this process, though, depends on experiments that define these and other motifs. Using the concept of conserved orthology among species, scanning sequences in the promoter region upstream of the transcriptional start site for several related species may uncover previously unknown conserved motifs that can later be tested experimentally.

Repetitive elements can populate genomes to large degrees. Because of the conserved nature of these sequences, it is fairly straightforward to discover these. For example, full (or nearly full) retrotransposon elements encode a reverse transcriptase protein. Similarity searches for these are fairly powerful. Once these are discovered, they can be used for subsequent nucleotide searchs for other elements that have diverged over time. Cataloging the repetitive elements is a critical first annotation step because it greatly reduces the amount of sequence that must be searched during the gene discovery process described above.

To discover segmental gene duplications, the repetitive class of sequences is first removed by a process called repeat masking. Then each gene is compared against the genome sequence itself to discover if it represented again in the genome. In this manner, blocks of genes that have undergone duplication can be uncovered.

When looking for SNPs, researchers have a number of resources. The basic reference gene can be from the individual used in the initial sequencing. That gene sequence can then be compared to other databases representing sequences derived from other individuals. A good comparison would involve the reference gene with the sequences found in an EST database. The source mRNA for EST projects is typically not from the original source used for genomic sequencing. Therefore variability between the reference and EST sequence is a SNP.

A novel search for controlling element motifs. All genes are controlled by sequences upstream of the transcriptional start site. A number of the sequences are important because they represent the site to which transcription factor, proteins that control gene expression, bind. A major goal of annotation would be to describe those sequences, and eventually determine how universal those sequences are in the promoter of specific genes. The first step is to describe such sequences in a reference species and use that information for further comparative analyses. A recent report [Science (2003) 301: 71] describes a sequence-based approach to uncovering these sequence motifs.

The yeast (*Saccharomyces cerevisiae*) genome has been sequenced and many members of the total gene array (6331 genes) have been named. Each of these genes contains an upstream controlling region. These controlling sequences, by their very nature, must reside in the "intergenic" region that lies between each gene. For yeast, these regions are rather small and average about 500 nucleotides. Over evolutionary time, these regions tend to diverge to a greater extent than the coding regions. This divergence actually offers a means to uncover important sequences because these would be under selection pressure and would maintain sequence continuity overtime to ensure the gene can still be properly expressed in the evolved species.

Cliften et al. (2003) applied a ***phylogenetic footprinting*** method to uncover these conserved motifs. They chose five additional *Saccharomyces* species to analyze. Three of these are closely related to the standard yeast speices (*S. mikatae*, *S. kundriavzevii*, and *S. bayanus*). and two more distant species (*S. castelli* and *S. kluyveri*). The closely related species average 59 to 67% sequence identity, whereas the distant species are 33.9% identical. Each of these five species were sequenced to a 2-3X genome coverage using the whole genome shotgun approach. This provided an 85-95% coverage of the genome sequence of each species. With five species in hand, the group reannotated the genomes. This lead to a reduction in the number of "real" gene to a value of 5773.

With each gene annotated, the researchers took the shared intergenic sequences upstream of each gene and aligned them using CLUSTALW, a program that aligns multiple sequences. They aligned the four related sequences and then performed the alignment with all six species. With the four species, alignment over 50% of the intergenic regions were shared, whereas with all species this number was reduced to 40%. Sequence identity among the four species averaged 37.1% over all of the promoters. Peak identity was located at 125-250 nt upstream of the ATG translational start site. This suggests 1) this region is enriched for regulatory sequences, 2) regulatory sequences are close to ATG site in yeast, and 3) phylogenomics might uncover controlling elements by looking for conserved sequence elements.

These experiments discovered the following. First, only 15.7% of the integenic regions contain one of the seven sequence motifs that define a TATA box. This is a striking result since TATA boxes are considered essential elements for transcription initiation. Several motifs were shared among genes with similar biological functions. Other motifs were common to genes expressed during similar expression conditions such as stress, cell cycling or meiosis. This provides a new catalog of potential sequences that can later be studied experimentally to determine their biological significance and to uncover the factors which interact with these sequence to control gene expression.

Genome Annotation

Genome Sequencing

- Costliest aspect of sequencing the genome
  - But
    - Devoid of content
- Genome must be annotated
  - Annotation definition
    - Analyzing the raw sequence of a genome and describing relevant genetic and genomic features such as genes, mobile elements, repetitive elements, duplications, and polymorphisms
- Annotation costs originally low
  - May eventually exceed the sequencing cost
    - Why??
      - Continued reanalysis is required to find all of the meaning

What Does Annotation Describe???

- Genes
- Mobile genetic elements
- Small repeats
- Genome duplications
- Genetic diversity

Genes

- Genes define the species itself
  - Development
  - Reproduction
  - Response to environment
    - Biotic
    - Abiotic
- What defines the gene???
  - Coding region
    - Contains the information that defines the nature of an expressed protein or a functional RNA molecule
  - Controlling regions
    - Sequences that define where, when, and the degree to which the gene will be expressed
- Major goal of annotation
  - Defining genes and their controlling regions

Mobile Genetic Elements

- Also called transposable elements
- A major component of some genomes
- Classes
  - DNA elements
    - McClintock discovery
  - Retrotransposons
    - Most abundant class of repeats
    - Abundance
      - Human
        - 50% of genome is mobile elements
      - Arabidopsis
        - 10 % of total DNA
        - 20 % of gene rich-region

Other Repeat Elements

- Simple Sequence Repeats
  - SSRs
  - Major repetitive class found in genomes
  - Defined as
    - Localized repetitions of di- or tri-nucleotides
  - Well described in many species
  - Widely used as molecular markers

Duplications

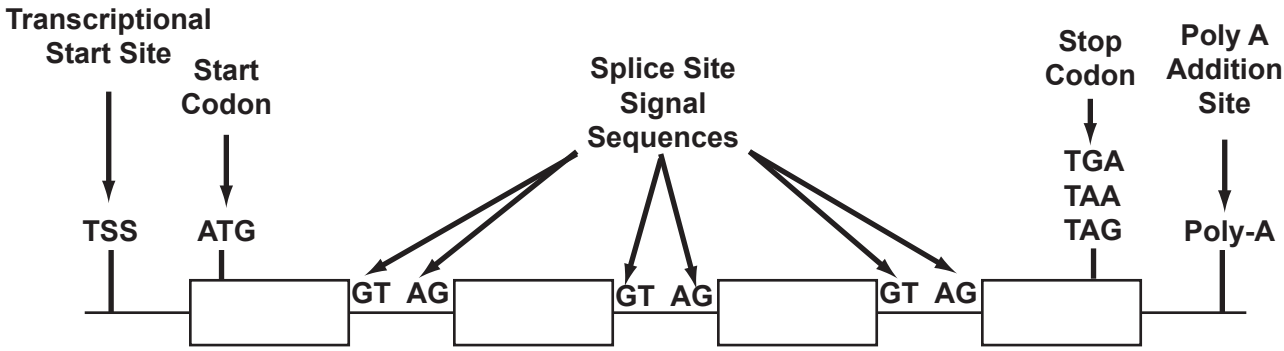
- Are their large blocks of duplications in genomes?
  - *Arabidopsis* genome
    - Model organism
      - Originally considered devoid of gene duplications
    - Sequencing discovery
      - Large blocks of segmental duplication
        - Local duplication
          - Intrachromosomal
        - Distal duplications
          - Interchromosomal
  - Human genome
    - Duplication pattern similar to *Arabidopsis*
  - Mouse genome
    - Lesser degree of duplication

Genetic Diversity

- How can gene sequences among individuals of a species vary?
    - Large deletions
        - Eliminate gene function
    - Small deletions
        - Gene functions but expression changed
            - Example
                - Cystic fibrosis gene
                    - Triplet lost
                    - Phenylalanine missing
                    - Mutant CF phenotype expressed
- Single nucleotide polymorphisms (SNPs)
    - A difference in a single nucleotide between two alleles
    - Sequencing multiple individuals during project
        - Uncovers SNP diversity
            - Private (Celera) human project
                - Source individuals
                    - African-American (1)
                    - Asian-Chinese (1)
                    - Hispanic-Mexican (1)
                    - Caucasian (2)
            - Human SNP project
                - Sequencing projects
                    - Uncovered 1.4 million SNPs
                    - Estimate of 10 million human SNPs

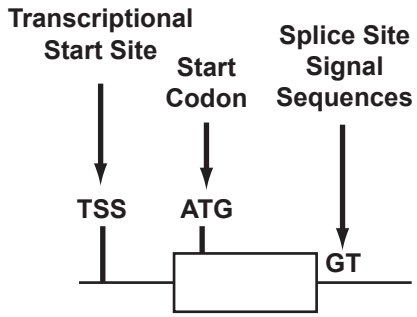
- Examples of SNPs
  - Sickle cell anemia
    - Adenine changed to thymine in sixth amino acid codon of β-globin gene
    - The change leads to the sickle cell phenotype
  - Mendel's plant height gene (Le)
    - Guanine to adenine change at nucleotide 685 of the mRNA
    - First nucleotide of 229 codon
    - Converts alanine to threonine
    - Changes function of gibberellin 3 beta-hydroxylase
    - Plant is short rather than tall

General Features of a Eukaryotic Gene

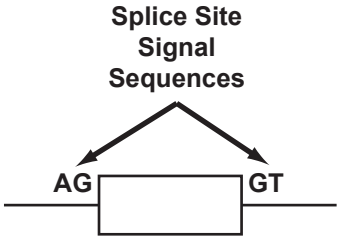


Specialized Gene Regions That Need Unique Gene Prediction Models

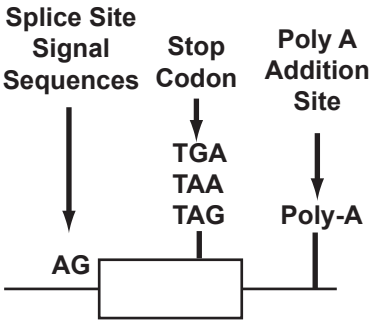
5' Exon



Internal Exon



3' Exon



Finding Genes In a Sea of Nucleotides

Extrinsic Content Detection

- Uses data from databases to discover genes
  - Search genomic sequence as a query against
    - Protein databases
    - Nucleotide databases
      - EST
        - Best information for gene structures because
        - Known to represent genes
        - Contain 3' sequences that are normally gene specific
- Problems
  - Exon-intron borders not always easy to predict
  - 5'-UTR sequences cannot be predicted

Intrinsic Content Detection

- ***Finding 5' exons***
  - Difficult process
    - 5' signal not fully defined
      - Transcriptional start site (TSS)
        - Few known
      - Promoter
        - Variation among promoters known
  - Algorithms search for:
    - CpG island
      - Normally gene rich
      - Some algorithms find TSS within the island
      - Some algorithms find TSS associated with TATA box in island
      - Identify ATG start site in island

- ***Finding 3' exons***
  - Identify polyadenylation addition site signal
    - AATAAA
  - Use stop codon as a 3' prediction signal
    - Essential for determining where one gene ends and another begins

- ***Finding internal exons***
  - Based on splice site features
    - Acceptor site
      - ***Exon*|*Intron***
      - AG|**GT**RAGT (R = A or G)
    - Donor site
      - ***Intron*|*Exon***
      - YYYYYYYYYYNC**AG**|G
        (Y = C or T; N = any nucleotide)

- Finding intronless exons
  - Difficult task
  - Must distinguish these from
    - Long internal exon
    - Pseudogenes that occurred by lost of intron in normal gene

What does a gene prediction program do?

- Calculates best scores for all gene features
    - Defines likelihood that neighboring coding features are really part of a gene
    - Likelihood is calculated as a
        - Weight
        - Probability
            - Hidden Markov Model (HMM) approach is currently preferred approach
                - DNA fragments (a few nucleotides at a time) are defined as a state
                - Probability that a neighboring state can be coupled with the first state to form a gene feature is calculated
                - This allows interdependencies between exons to be explored
                - Calculation based on a training set of genes
                    - Training set are genes from a similar taxonomic group with "putative" similar gene features
                - Probability that multiple states can define a gene is calculated

Predicting Multiple Genes

- HMM approaches easily extended to study both strands of a DNA sequence simultaneously
    - Value of modeling both strands
        - Prevents predicting two genes that overlap on the two strands
            - A rare eukaryotic event
- Need to understand features common across chromosomes
    - Insulator elements
    - Boundary elements
    - Matrix attachment regions
- Scaffold attachment regions

Comparative analysis

- One gene set can aid discovery in a related species
    - Gene order is conserved
    - Gene structure is conserved
    - Provides additional training set data for gene prediction
    - Example: Human gene models supporting mouse gene discovery

How does a popular software package do it??

FGENESH++C Pipeline (Softberry; quoted directly from company brochure)

1. RefSeq mRNA mapping by EST_MAP program – mapped genes are excluded from further gene prediction process.

2. *Ab initio* FGENESH gene prediction.

3. Search of all proeducts of predicted genes through NR database for protein homologs.

4. FGENESH+ gene prediction on sequences with found protein homology.

5. Second run of *ab initio* gene prediction in regions free from predictions made on stages 1 and 5.

6. Run of FGENESH gene predictions in large introns of known and predicted genes.

Naming The Genes

- Gene naming follows the discovery of potential genes
- Relies upon the significant amount of research that predated genome projects
    - Historically done on a gene-by-gene approach
    - Goals of gene-by-gene research goal is to clone and characterize an individual gene
    - Each gene is of interest to a specific research group
        - Housekeeping genes
            - Necessary for basic cellular biochemical processes of a cell
            - Nearly all are characterized at the nucleotide and protein levels
            - Sequence information is stored in large databases

The Naming Process

- BLAST
    - Software tool most often used to annotate (or name) a gene
    - **B**asic **L**ocal **A**lignment **S**earch **T**ool
    - Series of computer programs
        - Looks for sequence similarities between two sequences
    - Analysis consists of
        - Query
            - Sequence to which you are looking for a match
            - Nucleotide or amino acid sequences
        - Database
            - Set of sequences that may be like the query
                - Nucleotide sequences
                    - GenBank
                - Protein sequences
                    - Swiss-Prot is used to uncover sequences that are similar to the query

        - Translations possible
            - Nucleotide query sequence can be translated
            - Amino acid database sequences can be reverse translated
        - Recent BLAST innovations
            - Gaps can be incorporated to discover matches

EST Databases

- EST databases are critical in gene discovery
    - **E**xpressed **S**equence **T**ags
    - DNA sequences of expressed genes
    - Derived by end sequencing a cDNA library clone
    - Many libraries used
        - Represent specific
            - Dvelopmental or environmental conditions
    - EST represent genes because they are derived from RNA sequences
    - EST sequences used as a BLAST database

Non-gene RNA Sequences

- RNA molecules
    - Important components of the genome
        - Ribosomal RNAs and tRNAs
            - Both essential for protein translation
        - Small nuclear RNAs
            - Important for RNA splicing
            - Necessary component of the genome
            - Highly conserved
                - Easily recognizable
        - MicroRNAs
            - Short RNA sequences
            - 21-25 nt long
            - Negative regulators of gene expression
                - Bind target gene mRNAs and prevent their expression
- Programs that search specifically for these genes are available molecules

Regulatory Sequences

- Gene regulation a major area of research
  - Key to understanding gene expression
- Regulatory motifs discovered
  - Motifs
    - Short sequences that define a function
      - Nucleotide sequences
        - Sites where regulatory bind
          - Orthology searches
            - Scan promoter sequences
            - Search for conserved regulatory motifs
      - Amino acid sequences
        - Key sequences that bind DNA molecules
          - Orthology searches
            - Scan protein sequences
            - Search for DNA binding motifs
    - Discovered motifs must be tested experimentally
    - Functional genomics concern

Repetitive Elements

- Repetitive elements
    - May be major component of genome
    - Generally conserved
    - Fairly ease to discover
        - Example
            - Retrotransposons
                - Reverse transcriptase protein is conserved
                - Similarity searches for these proteins are fairly powerful
- Cataloging the repetitive elements
    - Critical first annotation step
        - Greatly reduces the amount of sequence that must be searched
    - Repeat masking
        - Procedure that removes the repetitive elements from the gene discovery process

Segmental Gene Duplications

- Segmental gene duplications: what are they???
    - Large gene blocks that are duplicated in the genome
    - Parts of chromosomes
        - Intrachromosomal duplication
            - Duplicated region moved to ***same*** chromosome
        - Interchromosomal duplication
            - Duplicated region moved to ***another*** chromosome

SNP Searches

- SNP
  - Single nucleotide polymorphisms
- Steps
  - Reference allele is defined
    - Usually the first allele that is sequenced
  - Gene sequence compared to other databases
    - Databases represent sequences derived from other individuals
      - Reference gene vs. EST database.
      - Source genotype for EST projects is typically not the same as the original source used for genomic sequencing
    - Is there variability between the reference and EST allele???
      - Yes???
        - A SNP is discovered

Phylogenetic Footprinting -A novel search for controlling element motifs.

- Principal of transcription regulation
  - Upstream sequences are important for transcription initiation
- Major goal of annotation
  - Describe these upstream sequences
  - Determine how universal these sequences are
- Critical steps
  - Describe such sequences in a reference species
  - Use that information for further comparative analyses.
- A recent report described such an approach
  - ***Phylogenetic footprinting***
    - Science (2003) 301: 71

Phylogenetic Footprinting in Yeast (*Saccharomyces cerevisiae*)

- Yeast genome
    - Total gene array
        - 6331 genes
        - Each gene contains an upstream controlling region
            - Controlling sequences must reside in the "intergenic" region between each gene
        - Intergenic regions in yeast are rather small
            - Average about 500 nt
            - These regions are more divergent than coding regions
            - Divergence offers a means to uncover important sequences
                - Essential, important sequences
                    - Under selection pressure
                    - Sequence is maintained overtime
                    - Ensures the gene will be properly expressed as species evolved

The Phylogenetic Footprinting Procedure

- Five additional *Saccharomyces* species studied
  - Three closely related to the standard yeast species
    - *S. mikatae*, *S. kundriavzevii*, and *S. bayanus*
    - 59 to 67% sequence identity
  - Two more distant species studied
    - *S. castelli* and *S. kluyveri*
    - These related species are 33.9% identical
- Draft sequence obtained for each species
  - 2-3X genome coverage
  - Whole genome shotgun approach
  - Provided 85-95% coverage of each species
- Yeast genome was reannotated
  - Used information of base species and the five other species
    - "Real" gene number
      - 5773 genes
- Alignment of upstream sequences
- Used CLUSTALW
  - Aligns multiple sequences at once
  - Aligned four related species first
    - Then aligned all six species
  - Four species alignment
    - Over 50% of the intergenic regions shared
  - Six species alignment
    - 40% of the intergenic regions shared
  - Peak identity
    - Located at 125-250 nt upstream of the ATG translational start site
    - Suggests
      - This region is enriched for regulatory sequences
      - Regulatory sequences are close to ATG site in yeast
      - Phylogenomics may uncover controlling elements by looking for conserved sequence elements

Major discoveries

- TATA box
    - Only 15.7% of the integenic regions contain one of the seven TATA box sequence motifs
        - Surprising result
            - TATA boxes are considered essential elements for transcription initiation.
- Other motifs
    - Several motifs were shared among genes with similar biological functions
        - Similar function ≈ Similar motifs
    - New catalog of potential sequences
        - These can be studied experimentally
            - Confirms/determines biological function