TITLE

Analysis of regulatory sequences controlling the expression of gene networks

MOTIVATION

Functional genomics techniques are defining sets of genes likely to act in concert. From expression profiles, chromatin analysis, proteomics interaction data or phenotypic screens, we emerge with clusters of connected genes or gene networks. In many cases, there is a reasonable anticipation that the members of gene networks will be expressed in concert through the action of one or more mediating transcription factors. By analyzing the regulatory sequences from the network members, it is often possible to define the common regulatory motifs shared by the genes and suggest the identity of the mediating factor(s).

OBJECTIVES

Participants should emerge with the following:

- The underlying biochemistry governing coordinated gene regulation
- An understanding of the methods and resources for the analysis of transcription factor binding sites in eukaryotic regulatory sequences
- An overview of pattern detection methods for the discovery of motifs in regulatory sequences from co-regulated genes, including the relative merits of different approaches
- The methods and motivation for the incorporation of phylogenetic footprinting into pattern discovery and sequence analysis
- The use of the TFBS programming tools for the analysis of regulatory sequences

AUDIENCE

The audience should have prior experience in Bioinformatics. Participants should have a rudimentary understanding of the flow of information in cells (DNA>RNA>proteins) and the detection of motifs in biosequences. Prior experience with BioPerl or an object-oriented programming language would be highly beneficial for the final hour of the tutorial.

COURSE INSTRUCTORS

Wyeth W. Wasserman, Ph.D.
Associate Professor, University of British Columbia, Canada
Dr. Wasserman develops novel bioinformatics algorithms for the analysis of regulatory sequences in eukaryotic genomes. His long-standing interest in unraveling the mechanisms underlying the coordinated regulation of sets of genes has motivated new approaches based on biochemistry and sequence evolution. He has served as an instructor in gene regulation bioinformatics at multiple universities and in the Canadian Bioinformatics Workshop series.

Boris Lenhard, Ph.D.
Assistant Professor, Karolinska Institutet, Sweden

Dr. Lenhard has developed bioinformatics systems and programming tools that dramatically accelerate the analysis of regulatory sequences. His GeneLynx system facilitates immediate access to valuable annotation information for human genes. The TFBS Perl module has emerged as a valuable tool for investigators interested in high-throughput analysis of non-coding sequences. Dr. Lenhard is been an award winning teacher, with extensive experience in the instruction of biochemistry and bioinformatics.

Albin Sandelin, M.Sc.
Senior Graduate Student, Karolinska Institutet, Sweden
Mr. Sandelin develops novel algorithms, internet resources and databases for the analysis of regulatory sequences. His recent Yeast Regulatory System Analysis system provides an integrated environment for the analysis of co-regulated genes. Mr. Sandelin coordinates an annual course in graduate-level bioinformatics course and is co-instructing a Perl for Biologists course.

DETAILED OUTLINE

I.INTRODUCTION TO THE ANALYSIS OF TRANSCRIPTION FACTOR BINDING SITES (W.Wasserman)

A. Introduction to Eurkaryotic Transcription
1. Promoters, enhancers, repressors
2. DNA binding proteins
3. Cooperative interactions between transcription factors
4. Chromatin

B. Modeling the binding properties of transcription factors
1. Regular expressions
2. Matrix models
a. construction and pseudocounts
b. application in sequence analysis
3. Advanced models
4. Specificity challenges

C. Resources for the analysis of regulatory sequences
1. Databases with binding profiles
2. Online tools for the detection of TF binding sites
3. Online databases and analysis tools for eukaryotic promoters

D. Phylogenetic footprinting
1. Methods for cross-species sequence alignment
2. Online resources for pre-computed genome alignments
3. Detecting TF binding sites with phylogenetic footprinting

E. Clusters of Transcription Factor Binding Sites
1. Examples of

II.PATTERN DETECTION METHODS FOR THE DISCOVERY OF TRANSCRIPTION FACTOR BINDING SITES (A.Sandelin)

A.Definition of Gene Networks in Genomics Data
1.Well-Defined Networks/Systems
2.Clustering Large Data Sets (Introduction and Issues)

B.Selection of Promoter Sequences for Analysis
1.Considerations in selecting sequence for analysis
2.Resources for sequence selection and manipulation

C.Exhaustive pattern detection algorithms
1.Overview of approach
2.Over-represented strings

3.Corrections for Background Properties
4.Matrix-based approaches
5.Resources

D.Gibbs Sampling / EM algorithms for pattern detection
1.Motivation
2.Greedy algorithm
3.Gibbs sampling
4.Corrections for background properties
5.Resources

E.Enhanced pattern detection methods
1.Cross-species comparisons
2.Information segmentation
3.Familial binding models as priors
4.Resources

F.Characterization and evaluation of patterns
1.Motivation for pattern comparison
2.Motif comparison algorithms
3.Resources

III.PROGRAMMING RESOURCES FOR THE ANALYSIS OF REGULATORY SEQUENCES (B.Lenhard)

A.Introduction to BioPerl (Brief)
1.Motivation
2.Objects in BioPerl
3.Resources

B.Introduction to the TFBS Perl Modules
1.Relationship to BioPerl
2.Installation Issues
3.Associated Software

C.Sequence Features

D.Matrix searches

E.Pattern discovery

F.Matrix comparison

G.Phylogenetic footprints

H. Working Example #1: Discovery of HNF1 binding sites in the UDPGT1 gene

I. Working Example #2: Detection of clusters of Hox binding sites

J. Working Example #3: Pattern discovery with promoters from a yeast gene network

K. Working Example #3: Analysis of binding sites in the promoters of all yeast genes