Promoter analysis

CSC, Otaniemi, February 10-11, 2004

Martti Tolvanen
IMT Bioinformatics
University of Tampere

IMT Bioinformatics

2. Finding promoter sequences

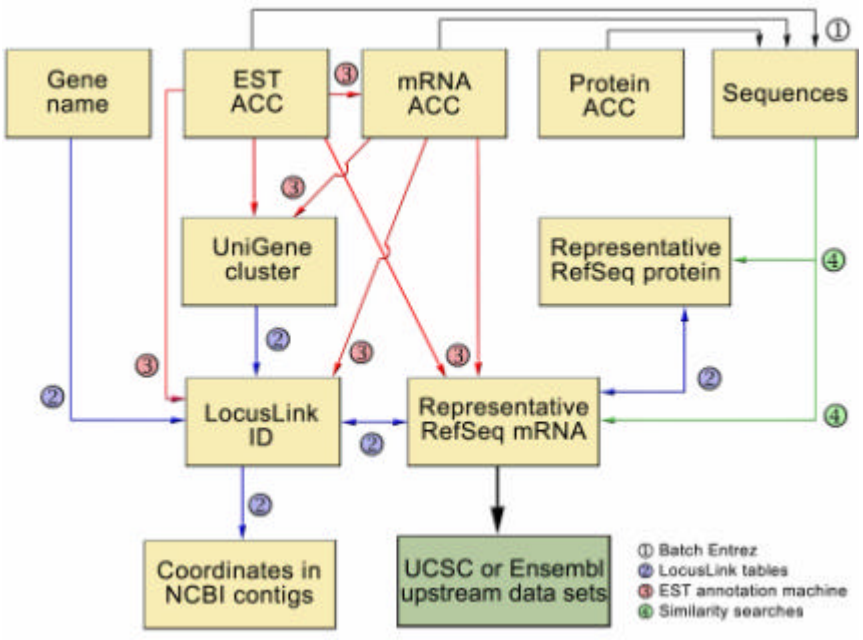
Three frequent problems:

- which genes do you need?
- where is the actual transcription start site (TSS)?
- does the upstream sequence overlap another promoter?

Knowing your genes

You may have data about varying codes:

- proprietary codes from chip manufacturer
- gene names (from when?)
- UniGene clusters (changing between releases)
- GenBank (EST), RefSeq or LocusLink codes



Where is the TSS?

- one gene may have several alternative transcripts in RefSeq
  - if first exon has alternatives, you may have two TSSs – which one do you want?
- even RefSeq mRNAs may miss part of the 5'-sequence (20-25 % of cases?)
  - RNase degradation in mRNA isolation
  - incomplete copying by reverse transcriptase

Finding full-length mRNAs

- resources for full-length sequences and true TSSs:
  - http://dbtss.hgc.jp/index.html
  - http://biowulf.bu.edu/zlab/PromoSer/
  - http://www.epd.isb-sib.ch/
    - only experimentally verified TSS! – not many
  - http://sdmc.lit.org.sg/FIE2.0/

Do you need the True TSS?

- exact TSS location may not be a problem if:
  - you do not compare TFBS positions relative to TSS
  - you take some excess sequence (like –800 to +100)
- but: extra sequences add noise
- and: if the 5'-UTR + first intron is several kB, you cannot know if you actually have the promoter or not
  - garbage in – garbage out

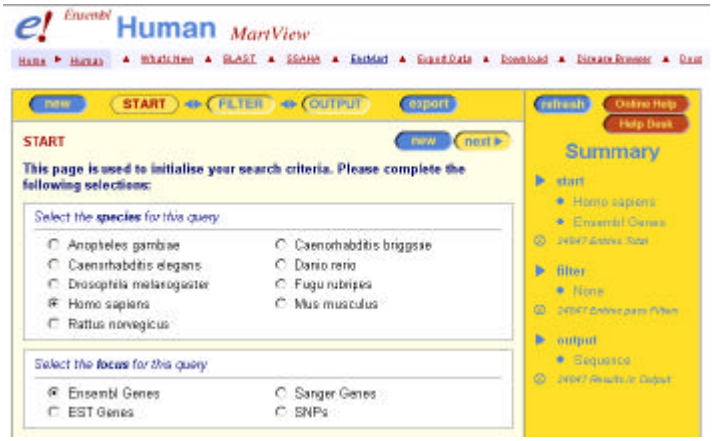
Easy resources for 5'-sequences

- http://genome.ucsc.edu/downloads.html
  - upstream1000.zip etc. under Full data set
  - based on TSS as given in RefSeq entries
- http://www.ensembl.org/EnsMart/
  - customizable for retrieving any length of sequence around TSS
  - relies on Ensembl gene definitions
  - searchable with many different kinds of codes!
- http://rsat.ulb.ac.be/rsat/ ??

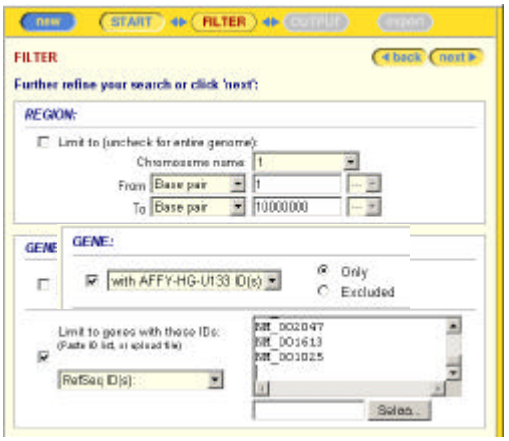
Even easier?

- Genomatix offers ready-made promoter sequence sets for several microarray chips, including annotation of TFBS that Genomatix data and software can detect
- academic price: € 1400 per average-size set, or 5x that in for-profit use

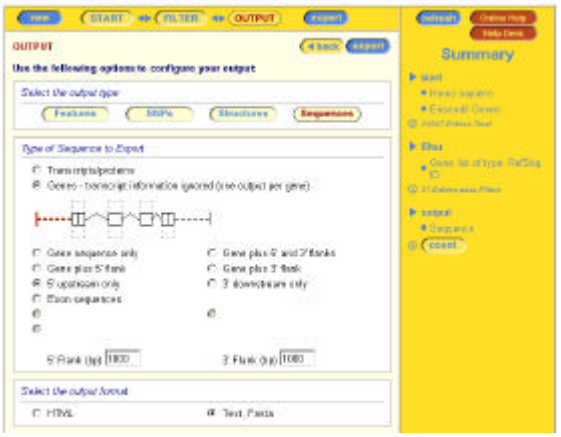
EnsMART start



Defining what you search



Selecting output type



Exercise

Find the 5'-sequence (-1000 to 0) which corresponds to human RefSeq mRNA NM_001675

Comparison of results from various services:

- how many sequences were found?
- are the sequences identical?

Exercise (continued)

Services to be tested:

- Ensembl/Ensmart
- PromoSer
- RSAT
- FIE2
- other?

Results of the exercise

When the searches were carried out on Dec 1st, 2003, we got confusing results:

- in PromoSer, small changes in options gave you widely different presumed TSS locations, even with over 20 alternative sites for one gene, over 70 kB of sequence
- in random picks, some sequences given by RSAT had no match in the sequence sets from EnsEMBL or PromoSer
- in RSAT results, it would have been hard work to connect the sequences to the codes used in the query

Sequence retrieval - conclusions

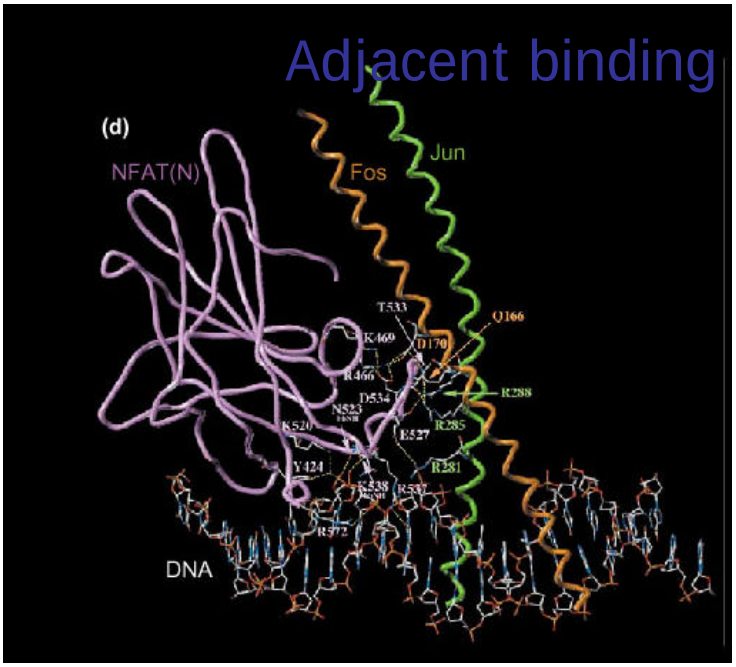
- for well-documented genomes, such as yeast, it is straightforward to get the sequences you need for analysis
- for human genes, reliable large-scale analyses are nearly impossible at the present because of unavailability of data and/or lack of quality indicators therein
  - (in the Genomatix package, however, a grading of promoter reliability is shown)
- with appropriate caution, your best guess would be in the EnsMart, DBTSS, UCSC sequences

Sequence retrieval - conclusions (2)

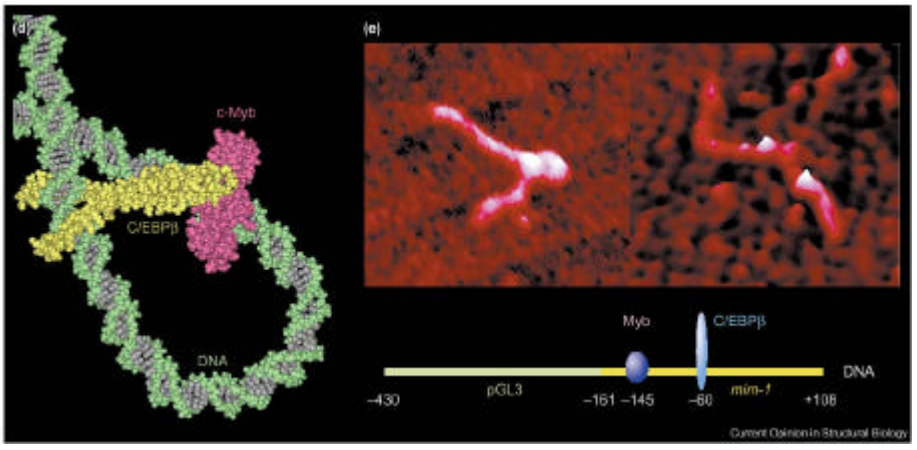
- for analysis of single genes, interspecies comparison may add confidence to your data
- find conservation of 5' elements in aligned human vs. rodent genome sequences
  - this option is available in EnsMART

Interlude: structural basis of TF interactions

- several structures of transcription factors bound to their target DNAs are available in the PDB
- I present two examples of interactions between two different TFs



Binding from a distance



Phylogenetic shadowing

- in closely related species it is difficult to distinguish functional from passive conservation
- however, the additive collective divergence in a group of higher primates as a group is comparable to that of humans and mice
- therefore, a comparison of numerous primate species can be used to identify regulatory regions in genes which do not exist in rodents

Strategies for microarray data based promoter analysis

- Find known TFBS, look for significant enrichment and/or clustering
- Pattern recognition
  - Compare to known TFBS
  - Test found sequences in lab
- Do comparative genomics first (e.g. find only mouse/human conserved sites), then analyze enrichment of sites of clusters

5. Pattern recognition

two principal types of promoter analysis:

- pattern matching (known patterns)
  - finds only what is defined in your pattern library
- pattern recognition (no knowledge of what will be found)
  - you have to evaluate the biological significance for all new findings

Pattern recognition

- does not require:
  - previous sequence alignment
  - knowledge of patterns to be searched
- in many programs, you do not need to know:
  - length of pattern
  - how many of your sequences contain the pattern

Pattern recognition

- produces short ”local multiple alignments”
- is capable of detecting more subtle patterns than standard methods (such as Blast)
- most pattern recognition programs use statistical methods to assess the significance of the findings

Pattern significance

- statistically significant patterns are not always biologically significant!
  - many TFBS are short and frequently found by chance
  - low-complexity regions in upstream sequences (CpG repeats, microsatellites) may result to ”strong” patterns
    - check your sequences for repeats etc.!

Pattern search strategies

- found patterns are eliminated from further search
  - produces non-overlapping patterns
- same sequences may participate in several patterns
  - produces large numbers of patterns
  - pattern clustering can reduce the number
- gapped vs. non-gapped patterns
- multiple hits in one sequence?
- hits needed in all sequences/part of submitted

Control data? !!

- for estimating the significance of your findings, you need some controls
- most programs offer various choices for background sequence models
  - pay attention to the species you analyze
  - non-promoter intergenic sequences
  - promoters from all ORFs in your species
  - random sequence with selected C+G %
  - other modelled sequences

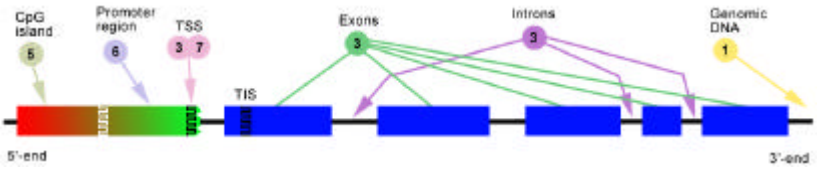
Control data? (2)

- even when the program gives statistic significance to patterns, that is always relative to the chosen background model
- **you have to know the background model and be sure it is meaningful for your data**

What to expect?

- if you analyze your microarray data, you would like to believe that coexpression means coregulation
- however, this is not likely to be 100 % true in most cases
- (digression to levels of genetic control)
- so, you should be happy if you get some signals even if it is only in part of your sequences in some single gene clusters

Basics



Levels of biological control

- gene dosage
- **chromatin structure**
- **transcription initiation rate**
- (rate of splicing)
- (rate of nuclear export)
- RNA stability
  - up to this point: effect to RNA expression results
- Translational control & rate
- protein stability
  - up to this point: effect to proteomics results
- post-translational control (phosphorylation etc.)
  - Different forms of protein seen in proteomics

Levels of biological control (2)

- because of many different mechanisms which affect the final level of mRNAs, it is naive to assume that the rate of transcription is directly proportional to the observed RNA levels
- in addition, several TF systems may operate concurrently, giving simultaneous changes in several unrelated sets of genes

Sequence pattern regognition programs

- MEME (http://meme.sdsc.edu/meme/website/intro.html)
- AlignACE (http://atlas.med.harvard.edu/)
- Gibbs Motif Sampler (http://bayesweb.wadsworth.org/gibbs/gibbs.html)
- Gibbs Recursive Sampler (as above)
  - finds multiple occurrences of a pattern
- (SPEXS (http://ep.ebi.ac.uk/EP/SPEXS/))

Alternative approach

Kimono (http://www.fruitfly.org/~ihh/kimono/)

- combines pattern finding and expression data clustering in one step
- open question: is this scientifically valid?
  - is it legal to change clustering to arrive at more significant patterns?

Pattern regognition program details (1)

MEME

- algorithm: multiple expectation maximization
- non-overlapping patterns
- user may presume that the pattern is found in all sequences or in part of them, which will affect the results
- no gaps – but gapped motifs will come up as separate patterns
- shows information content at each site in the pattern
  - low variation = high information content

Pattern regognition program details (2)

Gibbs sampling based programs:

- Gibbs Motif Sampler (&recursive)
- AlignAce – based on Gibbs Motif Sampler
- Kimono – identical to AlignAce, if expression data is weighted to zero

Pattern regognition program details (3)

AlignACE differences to original GMS:

- optimized for finding multiple motifs
  - iterative masking of found patterns
- considers automatically both strands of nucleic acid

Pattern regognition program details (4)

SPEXS

- exhaustive pattern search = lots of results
- no public tools for grouping and joining the patterns into consensus

Pattern recognition exercise

- Retrieve the following data set, bacterial sequences in Fasta format:
  - http://bioinf.uta.fi/courses/yhteiset/lexA.html
- Use MEME to find patterns in these sequences
  - study the results of a ready-made MEME run at http://bioinf.uta.fi/courses/yhteiset/meme-summary.html and http://bioinf.uta.fi/courses/yhteiset/meme-alignments.html
- optionally, use other services for the same task

Algorithm examples: Timothy Bailey

Gibbs sampling and EM are well explained in two lecture slide series from Tim Bailey:

http://www.acmc.uq.edu.au/DETYA/biol3014/pattern-discovery.ppt
http://www.acmc.uq.edu.au/DETYA/biol3014/meme-mcast.ppt

All the remaining slides are directly from these sources