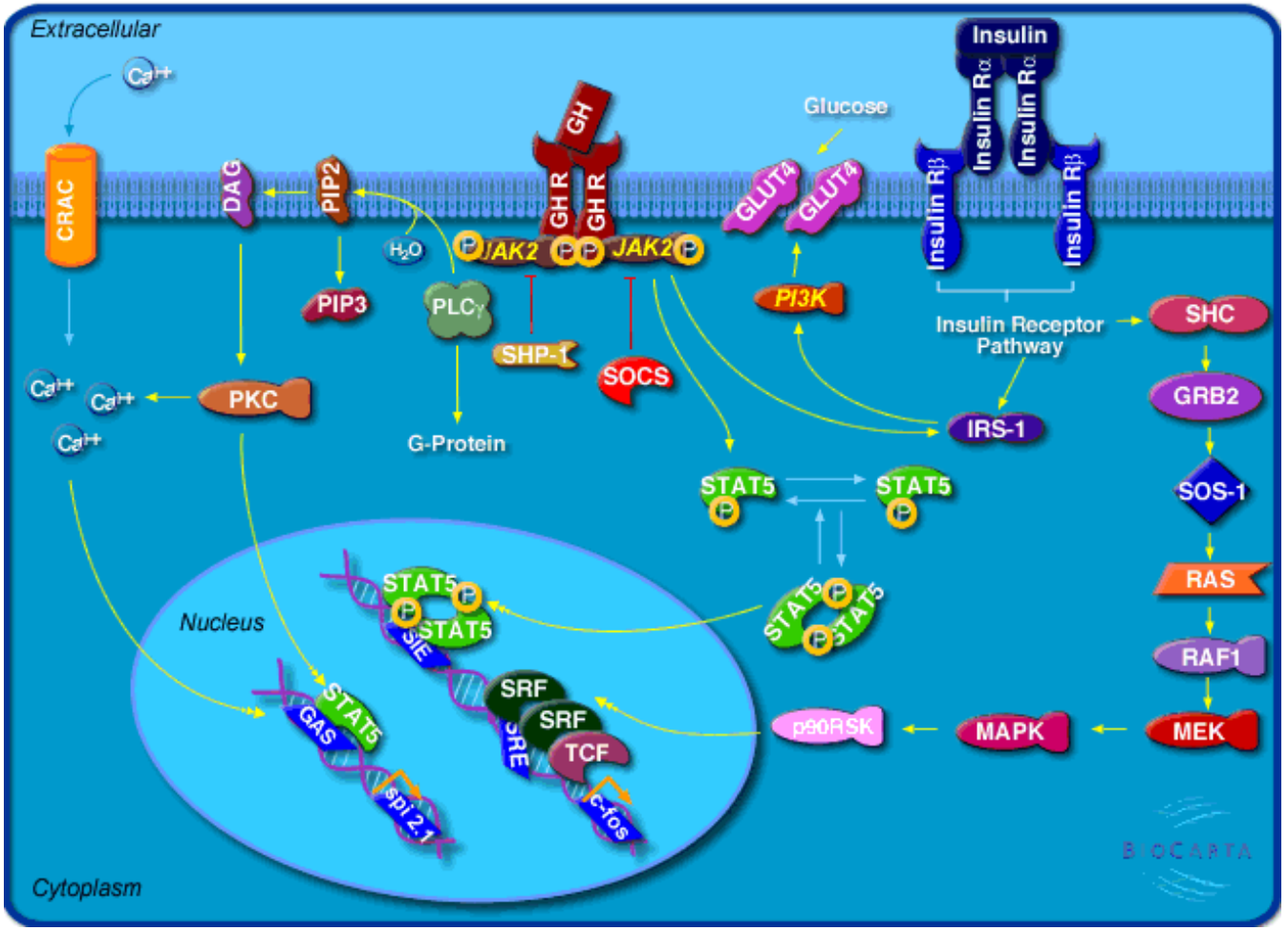
Workshop 4

In silico Promoter Detection and Characterization An Overview

Ulrike Gausmann
SFB604 Project B10
FLI Jena

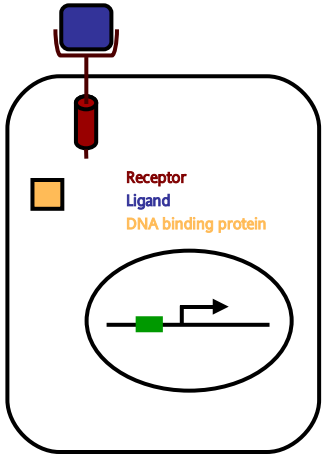
Model of Growth Hormone Pathway



Growth hormone signals a response in cells through the growth hormone receptor, a member of the cytokine receptor gene family. Growth hormone causes the receptor to dimerize, activating the JAK2 protein kinase. The activity of JAK2 mediates many of the downstream responses to growth hormone through phosphorylation of STAT transcription factors, MAP kinases, other kinase cascades and molecules involved in metabolism like IRS-1. Factors like SOCS and SHP-1 appear to play a role in the down regulation of signaling by growth hormone and cytokines.

Reduced Signalling Model

The reduced model is only dependent on **TFBS** (transcription factor binding sites) in promoter regions



Missing levels of complexity are

- cell cycle state
- other conditions as stress (e.g. hypoxia)
- metazoa: cell type & developmental stage

These biological states are characterized by

- availability of the sites
- factor cooperation
- interacting proteins
- receptor molecules
- ligand concentration

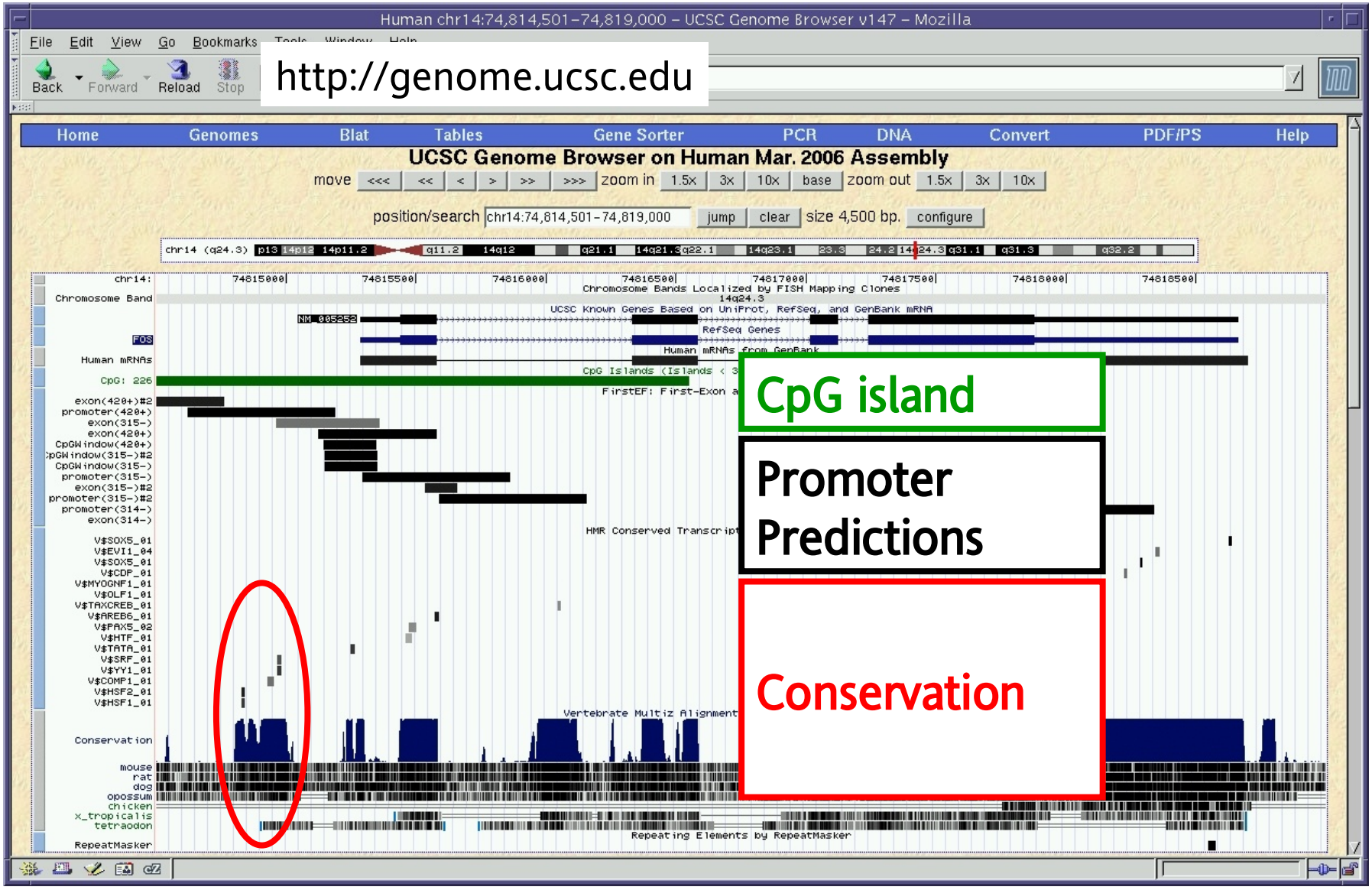
TFBS Prediction

Name	URL (http://)
Dragon ERE Finder	sdmc.lit.org.sg/ERE-V2/index
ECR browser	ecrbrowser.dcode.org
MatInspector	www.genomatix.de/products/MatInspector
MotifViz	biowulf.bu.edu/MotifViz
RSAT	rsat.ulb.ac.be/rsat/
SiteSeer	rocky.bms.umist.ac.uk/SiteSeer/
TESS	www.cbil.upenn.edu/tess
TFbind	tfbind.ims.u-tokyo.ac.jp
TFSEARCH	www.cbrc.jp/research/db/TFSEARCH.html
Toucan	homes.esat.kuleuven.be/~saerts/software/toucan.php
TRED	rulai.cshl.edu/TRED

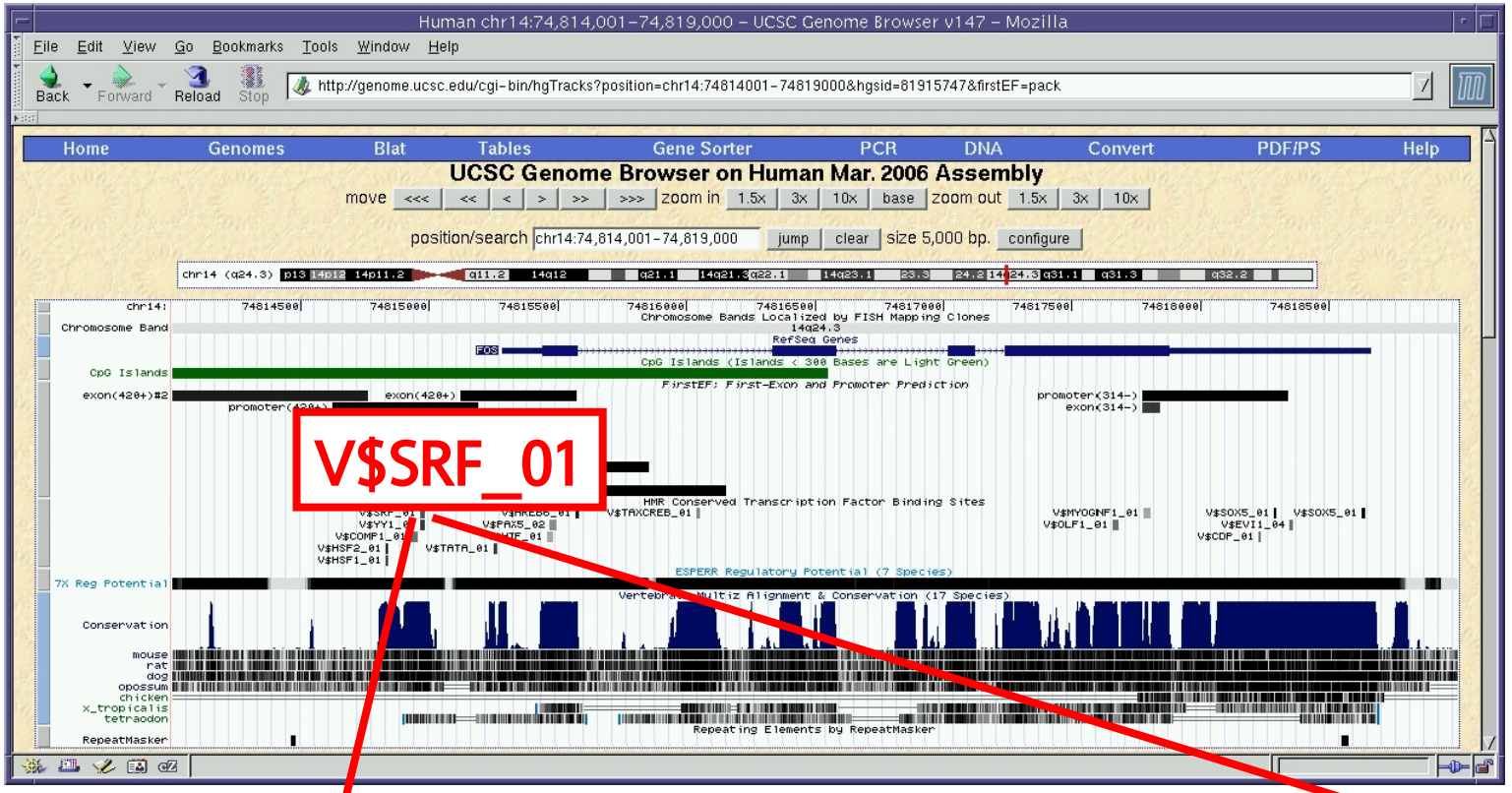
Selection of Web based tools for TFBS prediction in your sequence

[see supplement of Elnitski et al. (2006) Genome Res 16:1455 for more]

Putative Promoter Regions - *FOS*



Comparative Analysis - *FOS*




```
hsFOS   ... TCC-----CCCCTTACACAGGATGTCCATATTAGGACATCTGCGTCAGCAGGTTTCCACGGCCTTTCCCTGTAGCC--CTGG ...
ptFOS   ... TCC-----CCCCTTACACAGGATGTCCATATTAGGACATCTGCGTCAGCAGGTTTCCACGGCCTTTCCCTGTAGCC--CTGG ...
paFOS   ... T-C-----CCCCTTACACAGGATGTCCATATTAGGACATCTGCGTCAGCAGGTTTCCACGGCCTTTCCCTGTAGCC--CTGG ...
ssFOS   ... TCCCTCC--CGTTTACACAGGATGTCCATATTAGGACATCTGCGTCAGCAGGTTTCCACGGCCGTTCCCTGAAGTT--GTGG ...
btFOS   ... TCCCTCC--TCTTTACACAGGATGTCCATATTAGGACATCTGCGTCAGCAGGTTTCCACGGCTGTCCGCTGCAGTC--GTGG ...
fcFOS   ... TCCCCCCTCCCCTTACACAGGATGTCCATATTAGGACATCTGCGTCAGCAGGTTTCCACGGCCGTTCCCTGCGGTC--GTGG ...
cfFOS   ... TCC-----CCCCTTACACAGGATGTCCATATTAGGACATCTGCGTCAGCAGGTTTCCACGGCCGTTCCCTGCGGTC--GTGG ...
mmFOS   ... TCCCTCCT----TTACACAGGATGTCCATATTAGGACATCTGCGTCAGCAGGTTTCCACGGCCGGTCCCTGTTGTTCTGGGG ...
rnFOS   ... TCCCTCCT----TTACACAGGATGTCCATATTAGGACATCTGCGTCAGCAGGTTTCCACGGCCGGTCCCTGTTGTCCT--GG ...
```

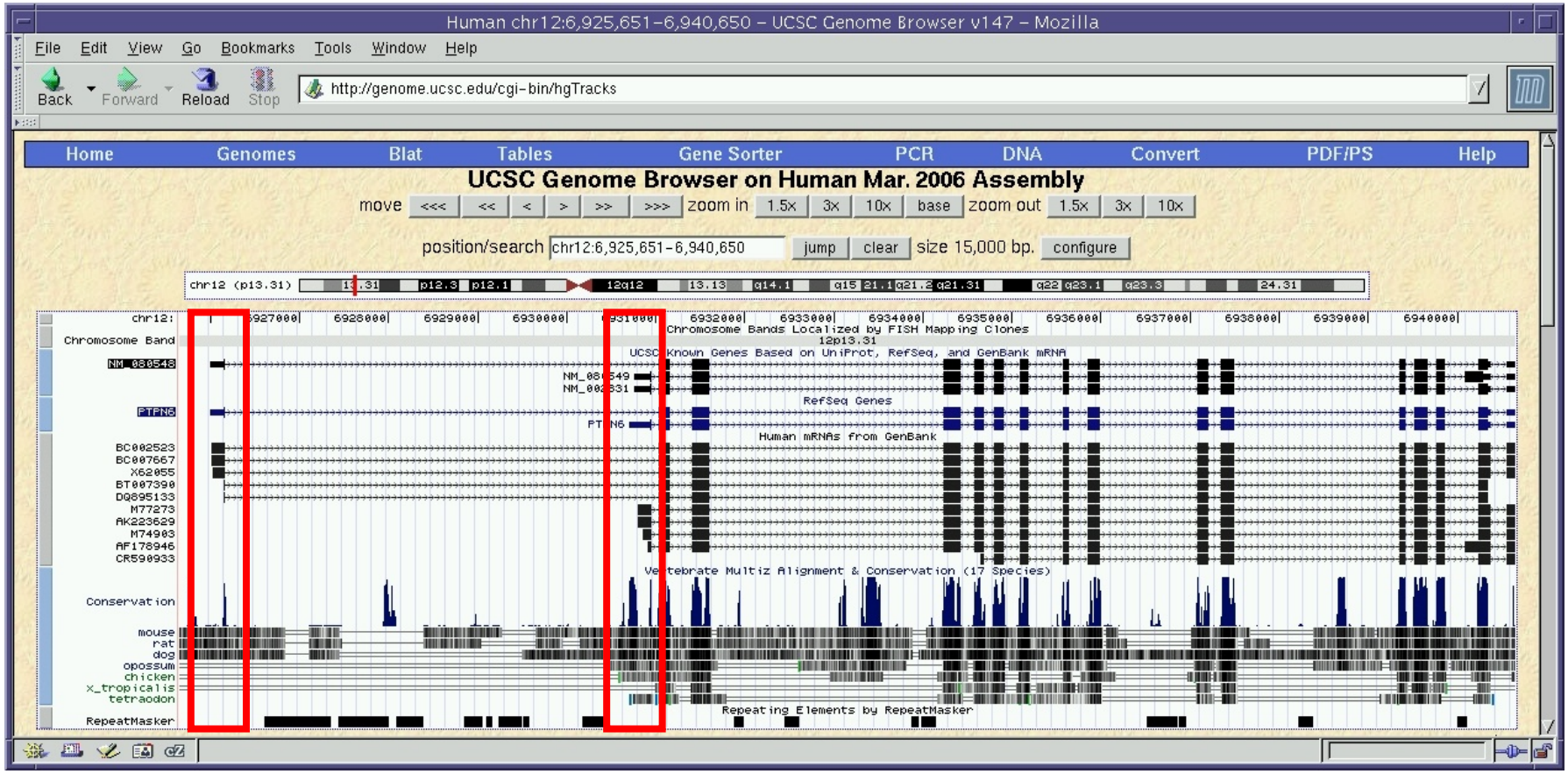
Identification of TFBS

- **Problem of false positive hits**

  First attempts to reveal regulatory mechanisms dealt with single cell organisms ($\rightarrow$ minus development and celltype)

- **Prepare sequences as short as possible**
  - check for putative promoter regions (CpG islands, first exon localization ?)
  - reduce by comparative analysis/phylogenetic footprinting
  - be sure about the first exon of your gene of interest (alternative transcription start sites (TSS), non-coding)

First Exon – *PTPN6*



Two alternative first exons → two promoter regions

Identification of TFBS (2)

- **Problem of false positive hits**

  First attempts to reveal regulatory mechanisms dealt with single cell organisms (→ minus development and celltype)

- **Prepare sequences as short as possible**
  - check for putative promoter regions (CpG islands, first exon localization ?)
  - reduce by comparative analysis/phylogenetic footprinting
  - be sure about the first exon of your gene of interest (alternative transcription start sites (TSS), non-coding)
    - ➢ genome wide analysis for human promoters was done

Application: Genome Wide Analysis

see Xie et al. (2005) Nature 434:338

- Collection of human promoter regions based upon genome wide alignments of human, dog, mouse & rat
- Identification of overrepresented motifs
- Comparison to TransFac database PWMs

Results:

- 69 motifs with TransFac sites identified
  - strongest signals: binding sites for ELK-1, Myc, NRF-1
- 105 new candidate motifs identified
- testing for tissue specificity (75 human tissues)
  - in 59 of the 69 motifs found
  - positional bias relative to TSS in the range of –100 to –50nt

Identification of TFBS (3)

- **Problem of false positive hits**

  First attempts to reveal regulatory mechanisms dealt with single cell organisms (→ minus development and celltype)

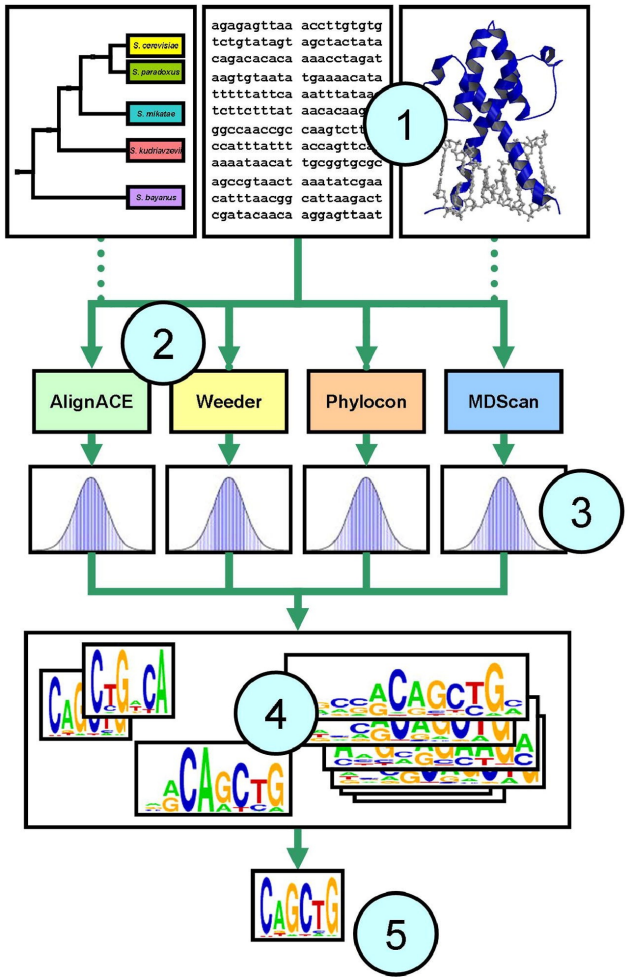
- **Prepare sequences as short as possible**
  - check for putative promoter regions (CpG islands, first exon localization ?)
  - reduce by comparative analysis/phylogenetic footprinting
  - be sure about the first exon of your gene of interest (alternative transcription start sites (TSS), non-coding)
- **Programs are working with different algorithms and are based upon different TFBS representations**
- **Good to know: none of the methods and progams gives "the" best results – results from different sources should be combined instead**

[Huber & Bulyk (2006) BMC Bioinformatics 7:229]

Identification of Regulatory Motifs

- comparative analysis/phylogenetic footprinting
- ➢ i.e. collection of orthologous sequences
- ➢ get sequences from (hypothetically) co-regulated genes
- ➢ search a set of sequences for common motifs
- ➢ several web tools available for the biologist now

Motif Discovery



1. **Assemble input data.** Results may be improved by restricting the input to high-confidence sequences. Some algorithms achieve improved performance by using phylogenetic conservation information from orthologous sequences or information about protein DNA-binding domains.

2. **Choose several motif discovery programs for the analysis.** For recommended programs see Figure 3.

3. **Test the statistical significance of the resulting motifs.** Use control calculations to estimate the empirical distribution of scores produced by each program on random data.

4. **Clustering and post-processing the motifs.** Motif discovery analyses often produce many similar motifs, which may be combined using clustering. Phylogenetic conservation information may be used to filter out statistically significant, but non-conserved motifs that are more likely to correspond to spurious sequence patterns.

5. **Interpretation of motifs.** Algorithms exist for linking motifs to transcription factors and for combining motif discovery with expression data.

Motif Discovery Workflow

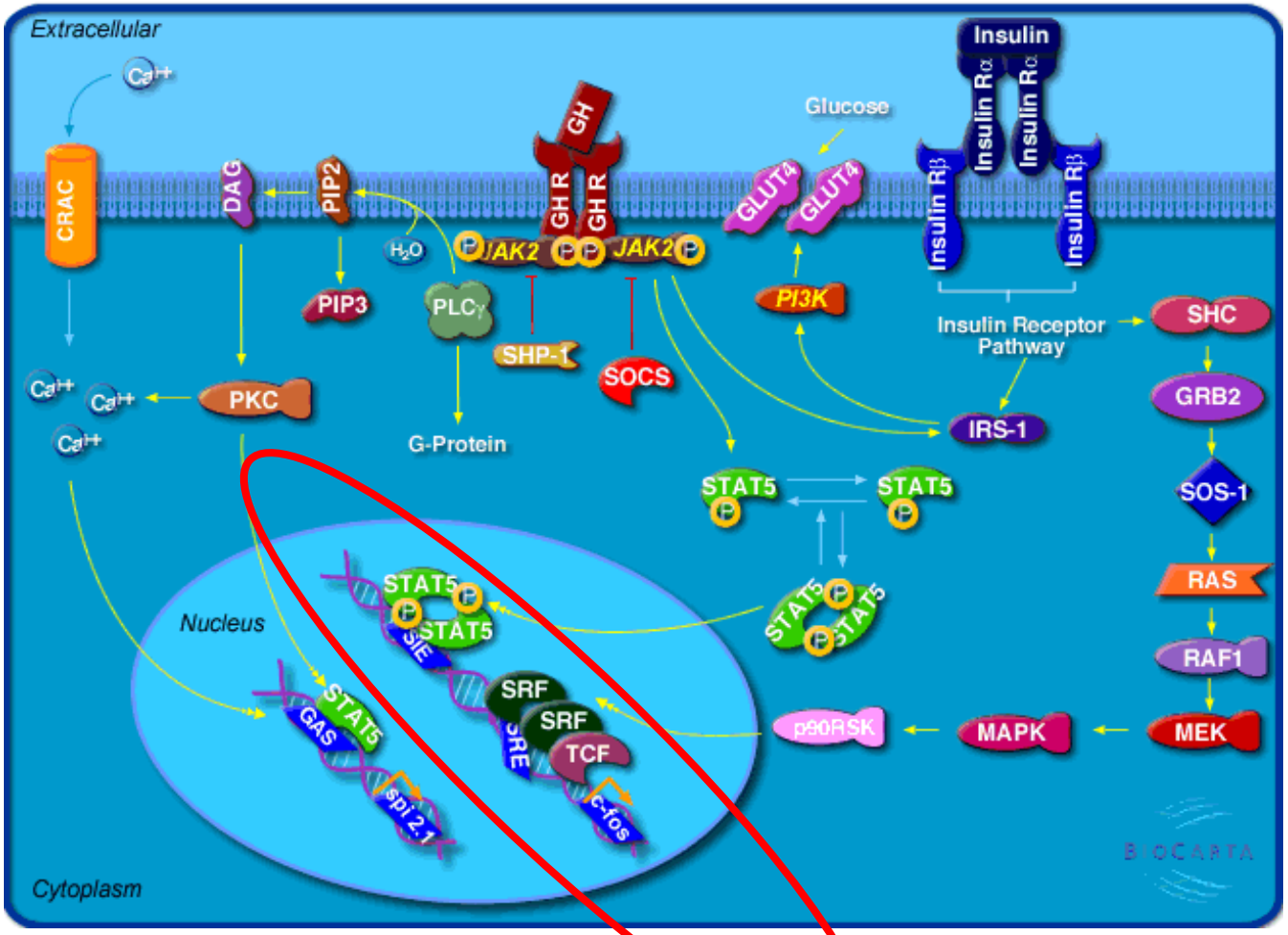
[MacIsaac & Fraenkel (2006) PLoS Comp Biol 2:e36]

MDScan **BioProspector** **Compare-Prospector**	MDScan uses ChIP-chip enrichment ratio data to help the motif search. BioProspector is a Gibb's sampling program. CompareProspector incorporates comparative genomics, biasing the search to regions of high conservation. http://seqmotifs.stanford.edu
Consensus **PhyloCon**	The Consensus program finds motifs in a set of unaligned sequences. PhyloCon builds on this framework by modeling conservation across orthologous genes from multiple species. http://ural.wustl.edu/
Weeder	An enumerative motif discovery program that performed well in a recent comparative analysis of fourteen algorithms. http://www.pesolelab.it/
MEME	The popular EM-based motif discovery program. Part of the MEME/MAST system for motif discovery and search. http://meme.sdsc.edu/meme/intro.html
AlignACE	A Gibbs sampling algorithm that can identify multiple motifs in a sequence set using an iterative masking procedure. http://atlas.med.harvard.edu/

Motif Discovery Programs

[MacIsaac *et al.* (2006) PLoS Comp Biol 2:e36]

TF Cluster

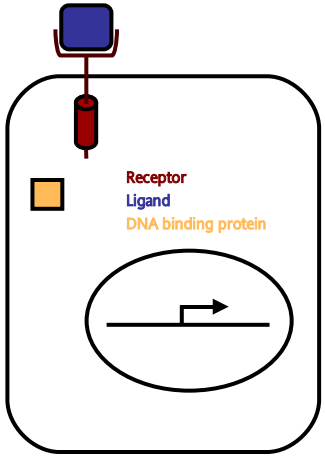


TFBS Cluster

TF Cluster

Name	URL (http://)
Cluster Buster	zlab.bu.edu/cluster-buster
CoMoDis	hscl.cimr.cam.ac.uk/CiMoDis_portal.html
CRÈME	creme.dcode.org
CRSD	biocip.nchu.edu.tw/crsd1/
Improbizer	www.cse.ucsc.edu/~kent/improbizer/improbizer.html
ModuleSearcher	homes.esat.kuleuven.be/~saerts/software/modulesearcher.html
MEME	meme.sdsc.edu/meme/
Over-represented Transcription Factor Binding Site Prediction Tool (OTFBS)	www.bioinfo.tsinghua.edu.cn/%7Ezhengjsh/OTFBS/index.html
TraFaC	trafac.cchmc.org/trafac/index.jsp
TFBind	tfbind.ims.u-tokyo.ac.jp
TRANSCompel	www.gene-regulation.com/pub/databases.html#transcompel

Identification of TFBS (3)



Developments since the SFB Workshop in 2002
"How to find transcription factor binding sites"

- More sequences available → data basis increased
  - binding site modelling
  - comparative analyses
  - statistical methods
- Support from experimental improvements
  - gene expression data (microarray analyses)
    - ➢ extraction of co-regulated gene sets
  - ChIP (-chip)
    - ➢ enrichment of transcriptionally active regions
    - ➢ for known transcription factors only
    - ➢ regions containg sites of interest
    - ➢ statistical follow up needed

Archives/Tools for ChIP (-chip) Analyses

Name	URL (http://)
ArrayExpress	www.ebi.ac.uk/arrayexpress
ChIPOTle	www.bio.unc.edu/faculty/lieb/labpages/ChIPOTle/home.html
GALAXY2	www.bx.psu.edu
GEO	www.ncbi.nlm.nih.gov/geo
MPEAK	www.stat.ucla.edu/~zmdl/mpeak/
PeakFinder	research.stowers-institute.org/jeg/2004/cohesin/peakfinder

- **Computational follow up**
  - at the moment no "standart technique" for the differentiation of signifant vs. non-significant hits recommended

    example for an integrated analysis (yeast)
    Kato et al. (2004) Genome Biol.5:R56
  - data analyses integrating different methods
    - Cheng et al. (2006) Mol Cell 21:393
    - Jin et al. (2006) Genome Res 1455:1585

Search for Higher Order Regulation

- **CRMs = *cis* -regulatory modules of TFBS**
- **homotypic**
- **heterotypic**
  - Zhu et al (2006) Genome Res. 15:848
    - human-mouse comparison: pairs of TFBS
    - newly fidentified:
      motif for G2-M phase harbouring sites for NF-Y/CDE/CHR
    - genes *CDC25*, *CDC2*, *CCNA2*, *PLK* and others

Search for Higher Order Regulation

- **CRMs = *cis* -regulatory modules of TFBS**
- **homotypic**
- **heterotypic**
  - Zhu et al (2006) Genome Res. 15:848
    - human-mouse comparison: pairs of TFBS
    - newly fidentified:
      motif for G2-M phase harbouring sites for NF-Y/CDE/CHR
    - genes *CDC25*, *CDC2*, *CCNA2*, *PLK* and others

- **Role of miRNA in gene expression**
  - Wu et al (2006) Genome Biol 7:R85

    REST, CREB & miRNA network in neuronal cells