

# Outline to TF binding sites bioinformatics lecture

- Introduction
- TFBS predictors
- TFBS databases

# Transcription factors

- Sequence-specific DNA binding factors
- Anticipated number in human 2 – 3 000, predicted 1962
- DNA binding domains
- Co-factors and chromatin help regulate transcription

# TF binding sites

- Response elements
- 5 – 20 bp
- Palindromic and gapped motifs
- Not exact
  
- Consensus sequences
- Position Weight Matrixes
  
- DNA accessibility
- Co-factors

# Combinatorics

- MyoD predicted 1/500 bp
  - ~100 000 sites predicted
  - ~1 000 sites believed functional
- TATA-like sequence elements every 250 bp
- Futility theorem – essentially all predicted TF binding sites (1000 fold excess) will have no functional role

# Example TBP (TATA binding protein)

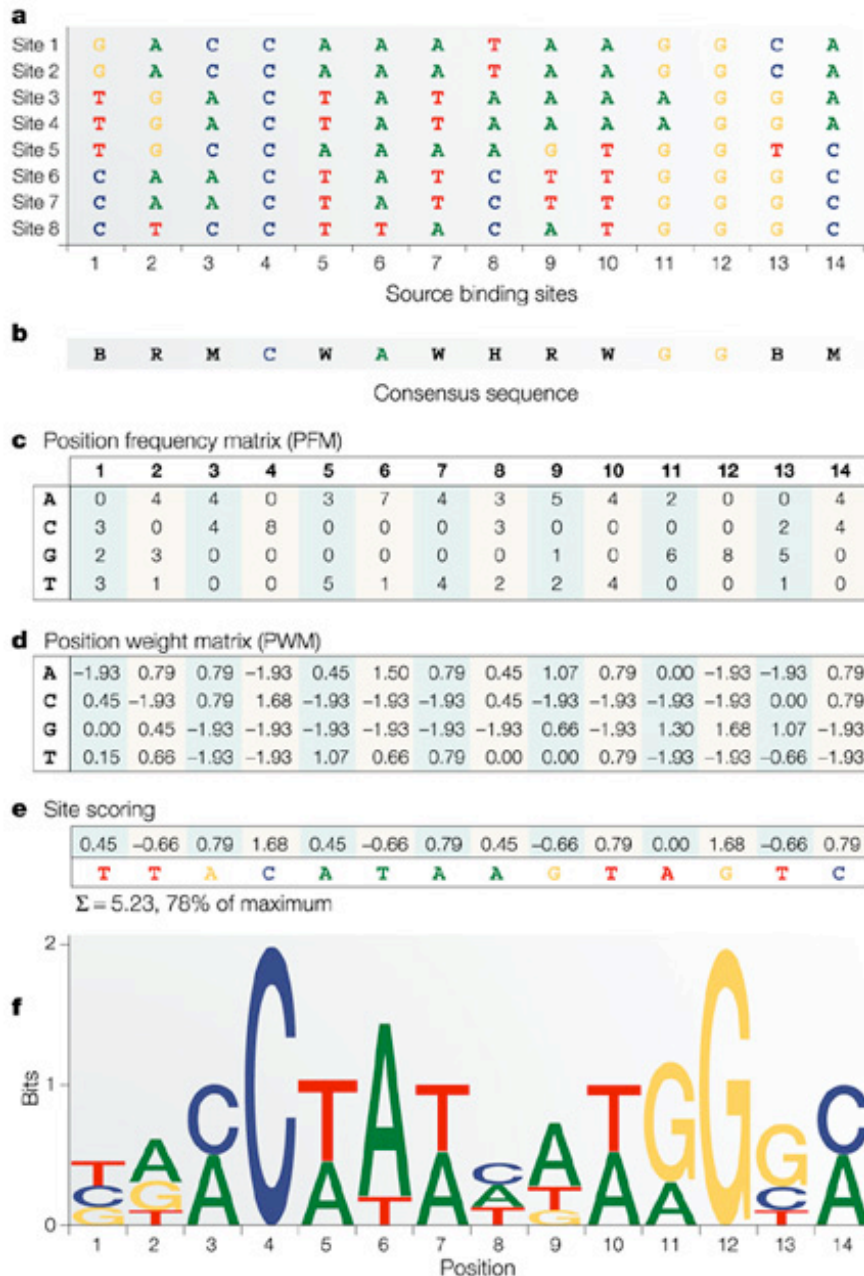
Consensus TATAAAA

But TBP also binds TATATAT and TATATAA

# Position Weight Matrixes

- Represent information content
- Assuming positions are independent
- Use validated TFBS
- Position frequency matrix
- Divide nucleotide probabilities with expected background probabilities and convert to log scale -> Position Weight Matrix
- Position-Specific Scoring Matrix, PSSM, 'possum'
- Sequence logos gives information content in a intuitive way

# MEF-2



# Answers – or at least help

- Promoter prediction
  - CpG islands
- Comparative genomics
- Cluster detection

# Phylogenetic foot-printing

- Compare genomes of species of suitable distance to find regions under evolutionary constraint
- First step – ortholog prediction
- Choose species at a medium evolutionary distance
- Multiple species comparison increase the specificity of the results
- Human – mouse – rat alignments increase the specificity of TFBS predictions 44-fold

# Cluster predictions

- TFs often work co-operatively
- Binding in close physical proximity
- Predict co-occurring motifs
- *cis*-Regulatory Modules – CRMs
- Machine learning or statistical likelihood
  - Specificity increased to ~50%
- Predictions of 118 000 modules

# Bioinformatic resources

- Pattern matching
- Pattern discovery
- Cluster detection

# Pattern Matching

- MatInspector
- ConSite
- SiteSeer
- Etc...

# Example: Consite

CONSITE - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://asp.il.uib.no:8090/cgi-bin/CONSITE/consite/>

*ConSite*

*Explore transcription factor binding sites shared by two genomic sequences*

Home

GeneLynx

*Analyze orthologous pairs of genomic sequences* **Go**

*Analyze an existing alignment of genomic sequences* **Go**

*Fetch genomic sequences of human:mouse orthologs by ID or keyword - TEMPORARILY DISABLED* **Go**

*Analyze single sequence* **Go**

Start | Mess... | CON... | snp\_... | tfbs\_... | Internet | 10:47 AM

# Analyze orthologous pair

**ConSite**

## Sequence Entry

Enter two orthologous genomic sequences and a cDNA (optional)

**NOTE:** Due to current hardware limitations, the lengths of sequences are limited to 10000 nucleotides. Longer sequences will be truncated to that size. For longer pairs of sequences you can [submit an existing alignment file](#)

In the test period, the input form is preloaded with test sequences. If you wish to enter your own sequences, clear the individual ones or

**Sequence#1**

Sequence name

Paste the sequence

OR enter valid accession number

OR upload a sequence file

**Sequence#2**

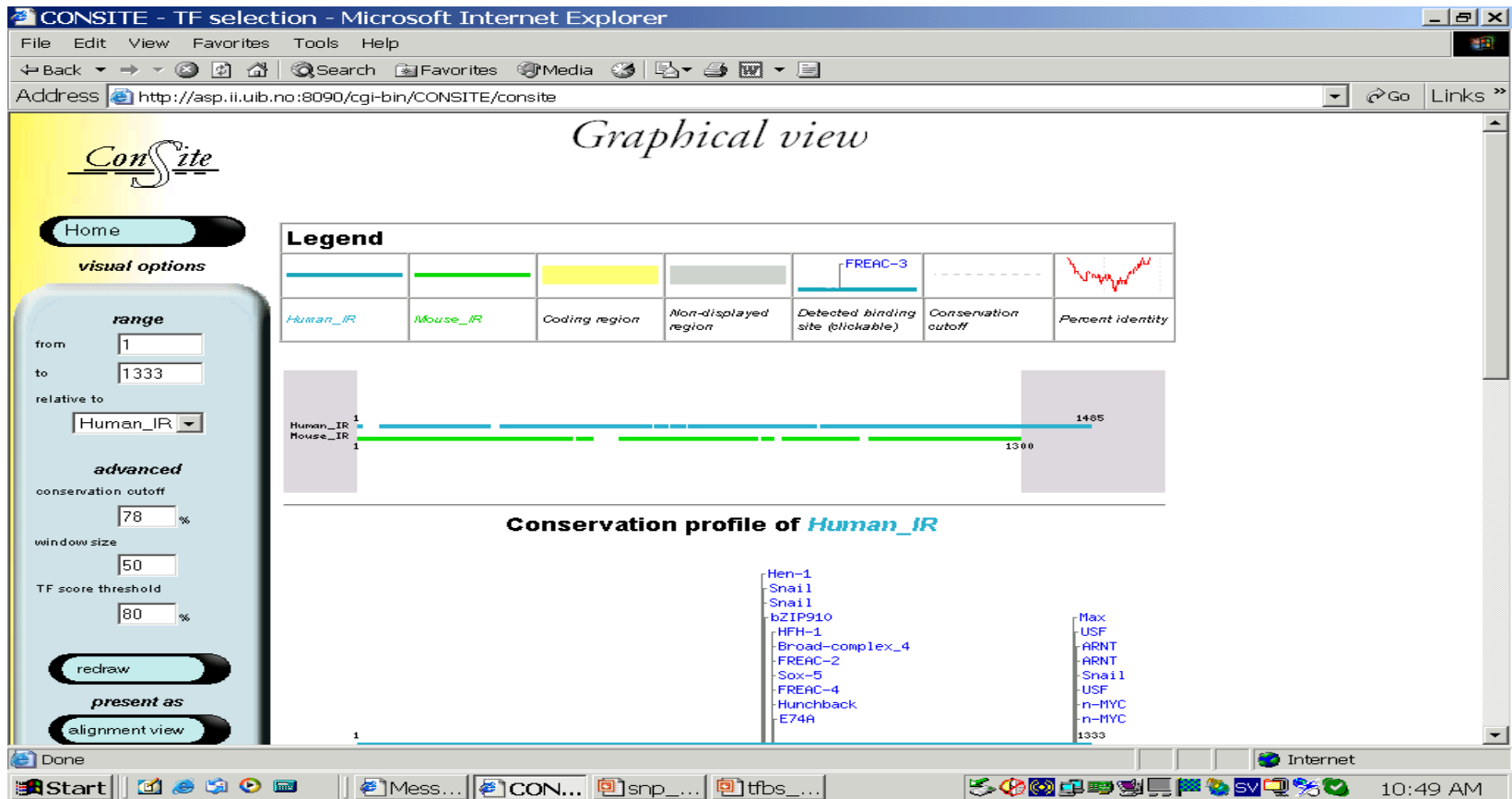
Sequence name

Paste the sequence

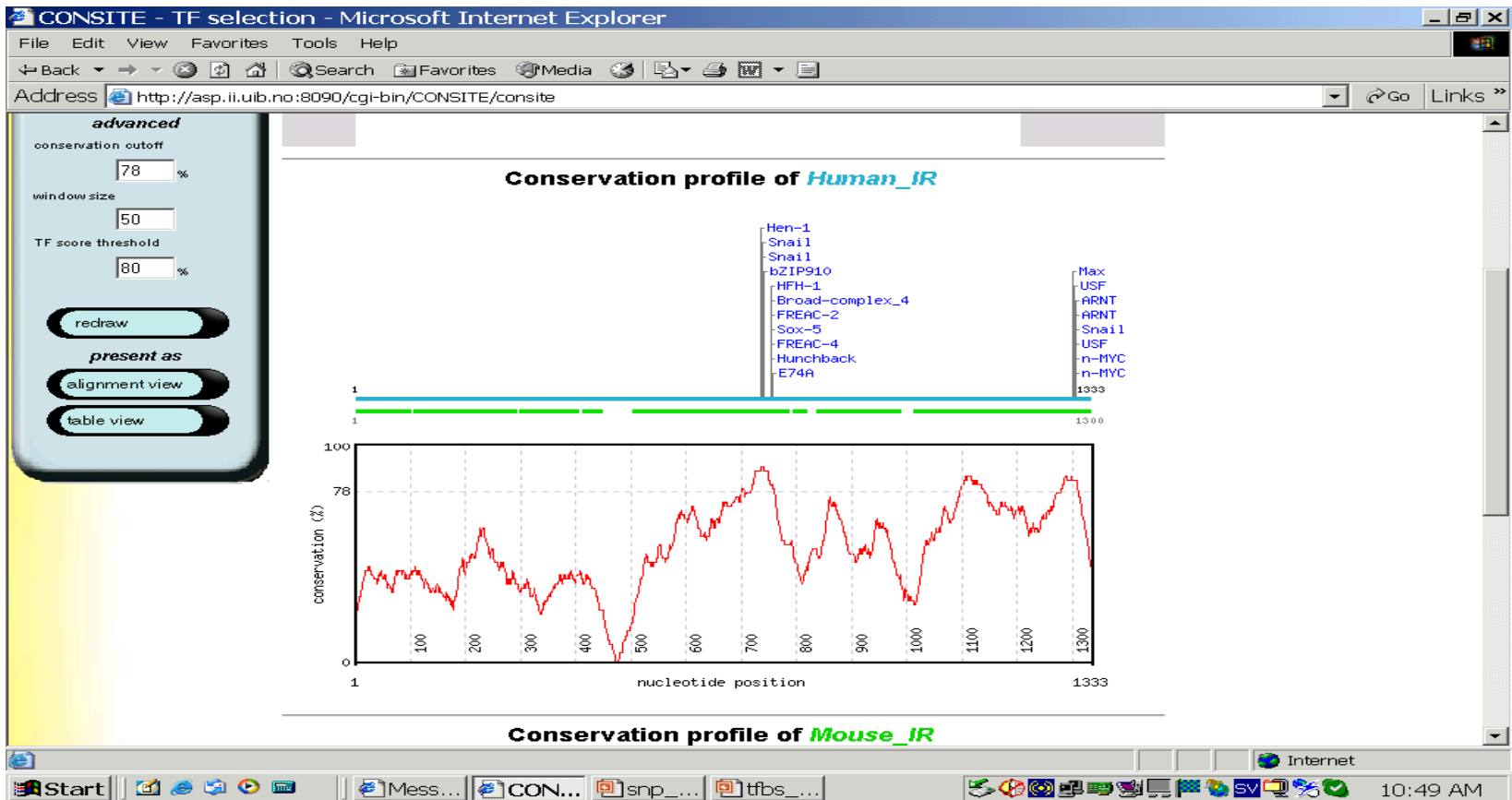
Done

Start | Mess... | CON... | snp... | tfbs... | Internet | 10:48 AM

# Results



# Results



# Pattern Discovery

- Consensus, PhyloCon
- FootPrinter
- Gibbs Motif Sampler
- MotifSampler
- Etc...

# Example: Gibbs Motif Sampler

- Mathematical algorithm to approximate the joint qualities of two samples (sequences)
- Four versions, from finding one motif to a defined set of motifs to any number of motifs
- Uses phylogenetic data – PhyloGibbs

# Cluster predictions

- Cluster buster
- MSCAN
- MEME, OrthoMEME
- OTFBS Tool
- ModuleMiner
- Etc...

# Example: ModuleMiner

ModuleMiner - Computational Detection of cis-regulatory modules - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://auihe8.esat.kuleuven.be/moduleminer/>

## ModuleMiner

KATHOLIEKE UNIVERSITEIT LEUVEN Bio@Sista VIB

### Manuscript

Van Loo P., Aerts S., Thienpont B., De Moor B., Moreau Y. & Marynen P. (2008). **ModuleMiner: improved computational detection of cis-regulatory modules. Different modes of gene regulation in embryonic development and adult tissues?** *Genome Biology*, 9:R66.

### Supplementary data: ModuleMiner results

- [Smooth muscle marker genes](#)
- [ORegAnno Erythroid benchmark set](#)
- [Liver benchmark set](#)
- [Muscle benchmark set](#)
- [ORegAnno Stat1 benchmark set](#)
- [Microarray cluster 1: Protein synthesis](#)
- [Microarray cluster 2: Oocyte/fertilized egg](#)
- [Microarray cluster 3: Neural tissues](#)
- [Microarray cluster 4: Lymphocytes](#)
- [Microarray cluster 6: Liver](#)
- [Microarray cluster 7: Mitochondrion](#)
- [Microarray cluster 8: Extracellular matrix](#)
- [Microarray cluster 9: Cardiac muscle](#)
- [Microarray cluster 10: Energy metabolism](#)
- [Primary heart field](#)
- [Secondary heart field](#)
- [Neural crest cell migration](#)

Done Internet

Start Messen... Genom... Microso... Modul... 6:15 PM

# Future

- More experimental analysis
  - DNA Accessibility
  - More and better motifs
  - Protein binding microarray
- Combine all the different approaches
  - Conservation
  - Clustering

# Recent advances

- MORPH - Using alignment and TFBS prediction simultaneously to increase alignment quality and binding site prediction (Sinha S, He X (2007) MORPH: Probabilistic alignment combined with hidden Markov models of *cis*-regulatory modules. PLOS Comput Biol 3(11))
- Predicting the transcriptional output mediated by a CRM and vice versa - predicting a CRM as the driver for the expression in a specific tissue or situation (Bauer DC, Bailey TL (2008) Studying the functional conservation of *cis*-regulatory modules and their transcriptional output. BMC Bioinformatics 9(1):220)

# TFBS Databases

- Eucaryotic Promoter Db (EPD)
- ooTFD
- PReMod
  
- JASPAR
- TRANSFAC

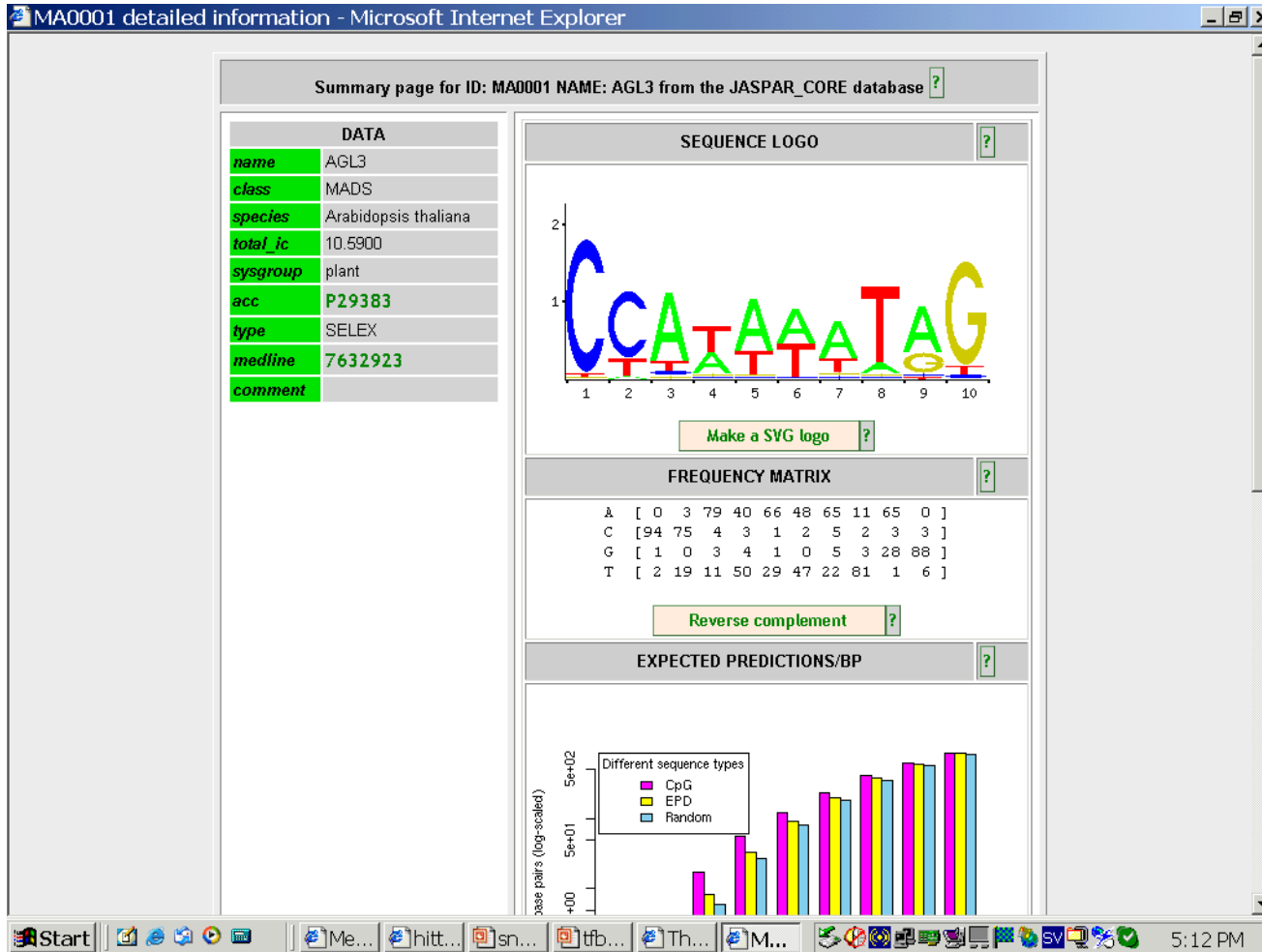
# Example: JASPAR

The screenshot shows a Microsoft Internet Explorer browser window displaying the JASPAR database homepage. The address bar shows <http://jaspar.genereg.net/>. The page content includes:

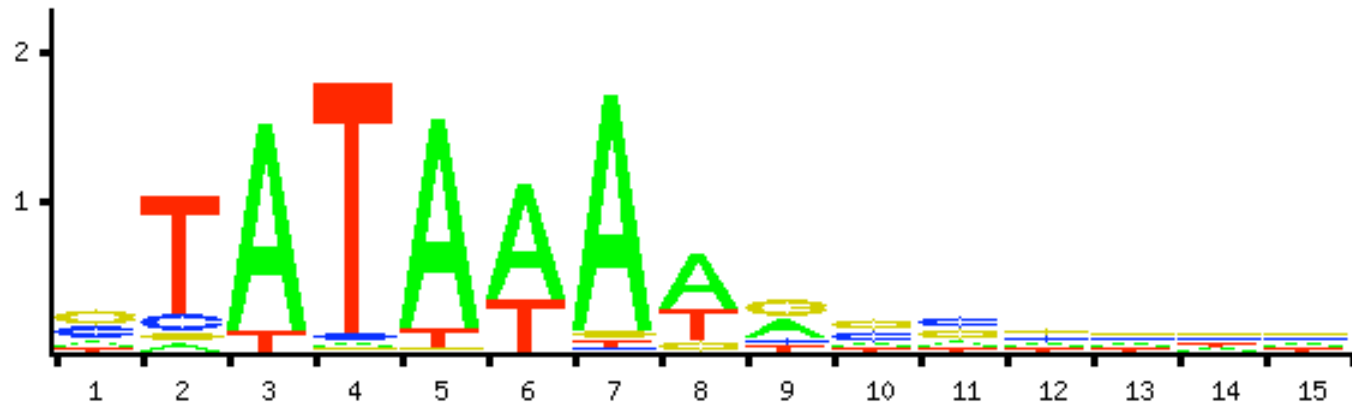
- Version: 3.0 STABLE
- Information: You are using the JASPAR stable server: [jaspar.genereg.net](http://jaspar.genereg.net) and Visit the developmental JASPAR server: [jaspardev.genereg.net](http://jaspardev.genereg.net)
- A 3D visualization of the word "JASPAR" in colorful letters, with "AG" and "AC" on the left and "GT" and "CT" on the right, positioned over a DNA sequence.
- Text: "The high-quality transcription factor binding profile database"
- Button: "Browse the JASPAR\_CORE database right away!"
- Navigation buttons: "DOCUMENTATION", "DOWNLOAD", and "CONTACT"
- A section titled "Select a JASPAR database" with a dropdown menu showing "JASPAR\_CORE" selected. Below it, a list of "JASPAR Collections" is shown: JASPAR\_CORE, JASPAR\_PHYLOFACTS, JASPAR\_FAM, JASPAR\_POLII, JASPAR\_CNE, and JASPAR\_SPLICE.
- Text describing the JASPAR CORE database: "The JASPAR CORE database contains a curated, non-redundant set of 123 profiles, derived from published collections of experimentally defined transcription factor binding sites for multicellular eukaryotes. The prime difference to similar resources (TRANSFAC, TESS etc) consist of the open data access, non-redundancy and quality: JASPAR CORE is a smaller set that is non-redundant and curated." and "When should it be used? When seeking models for specific factors or structural classes, or if experimental evidence is paramount"

The browser's taskbar at the bottom shows the Start button, several open applications (Me..., Bio..., sn..., tfb..., Th...), and the system tray with the time 5:07 PM.

# Example: JASPAR



# TATA



# ENCODE

- Encyclopedia of DNA Elements
- High throughput methods to identify and catalogue the functional elements
- 1% of human genome (30Mb)
- 44 regions
- Findings:
  - Most of the genome is transcribed
  - Regulatory sequences symmetrically distributed
  - 5% of the human genome under evolutionary constraints, of which 60% (as yet found) has evidence of function
- ENCODEdb