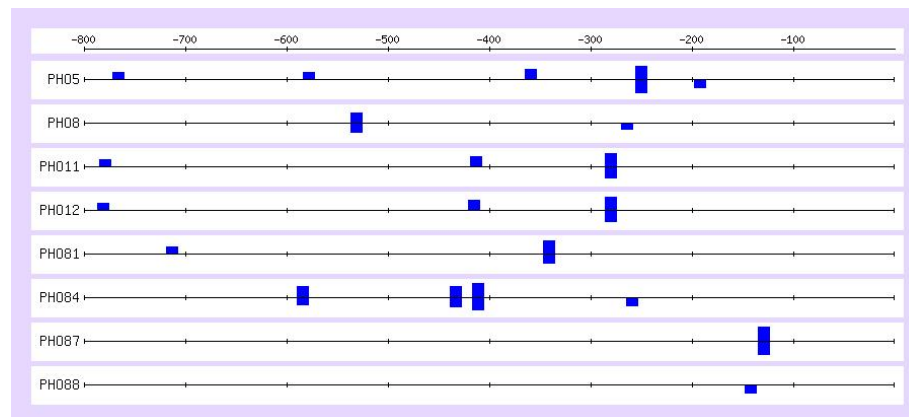
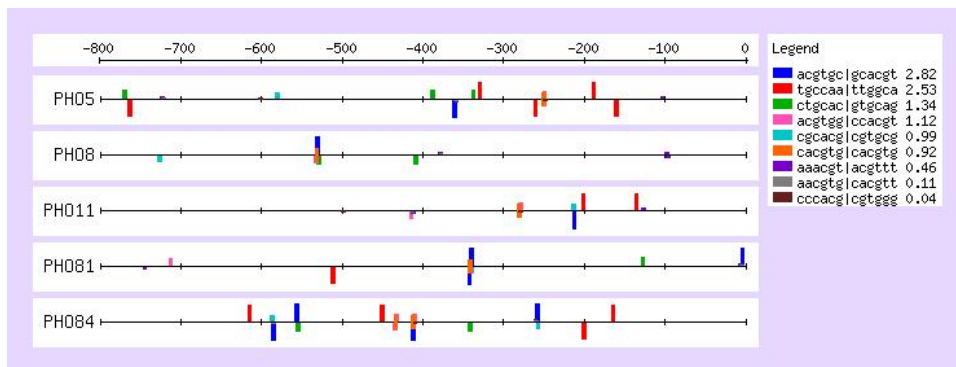
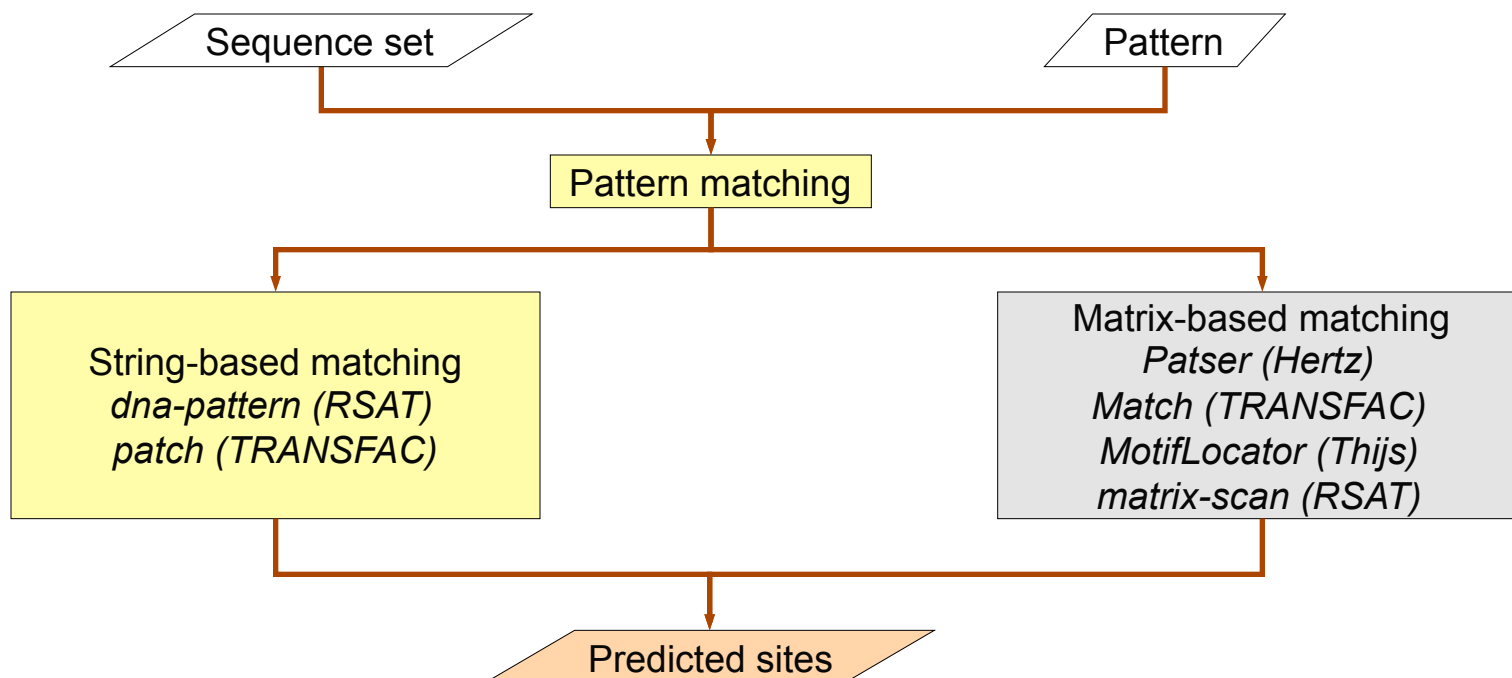


## *Regulatory Sequence Analysis*

# *Pattern matching*

# Pattern matching



## *Pattern matching in a small set of sequences*

- Goal: knowing the pattern, find the matching positions in the sequence set of interest
- Assign a score to each position
  - Indicate quality of the match
    - Substitutions for string-based pattern matching
    - Weight scores for matrix-based pattern matching
  - Indicate a priori importance of each pattern
    - e.g. significance from pattern discovery

## Expected matches for a consensus in whole genomes

- How many matches would we expect from a genome-scale pattern matching
  - Assuming a perfectly conserved hexanucleotide, with strand-insensitive activity
  - Expected matching rate: 1 occ / 2 kb

Organism	Size	Genes	Kb/gene	Non-coding size	non-coding /gene	exp_occ / reg_seq
	Mb		Kb	Mb	Kb	
<i>Mycoplasma genitalium</i>	0.6	481	1.25	0.1	0.12	0.06
<i>Haemophilus influenzae</i>	1.8	1 717	1.05	0.3	0.15	0.07
<i>Escherichia coli</i>	4.6	4 289	1.07	0.6	0.14	0.07
<i>Saccharomyces cerevisiae</i>	12	6 286	1.91	3.4	0.53	0.26
<i>Arabidopsis thaliana</i>	120	27 000	4.44	84.0	3.11	1.50
<i>Caenorhabditis elegans</i>	97	19 000	5.11	70.8	3.73	1.79
<i>Drosophila melanogaster</i>	165	16 000	10.31	140.3	8.77	4.21
<i>Homo sapiens</i>	3 200	31 000	103.23	3 104.0	100.13	48.14

# Genome-scale pattern matching

- Goal : given a pattern, find matches in the whole genome
  - → identify genes potentially regulated by a given transcription factor
- In general, a search based on a single signal returns many false positive
- Improvements
  - search for a repeated signal (e.g. GATA boxes)
  - search for combinations of signals
  - constraints on positions
  - combination of coding sequence information