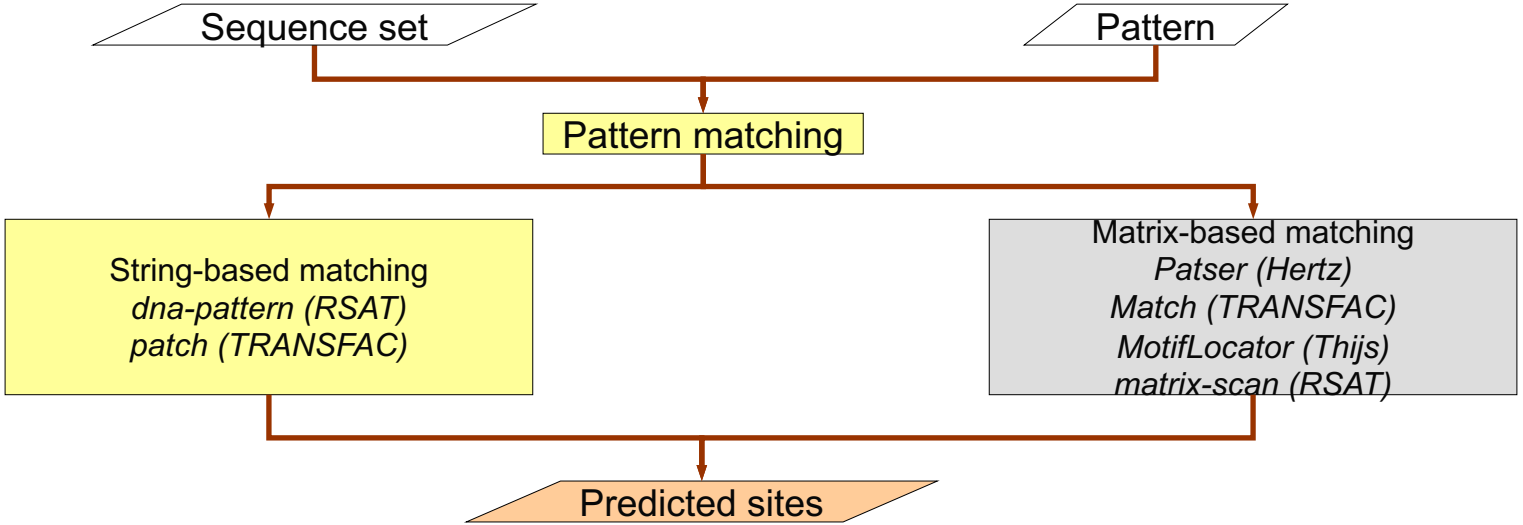
Pattern matching

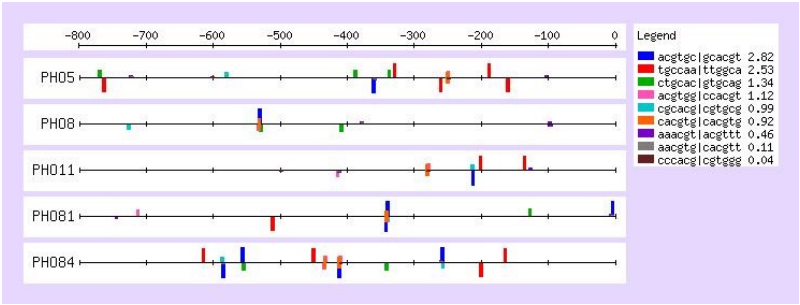
Jacques van Helden
https://orcid.org/0000-0002-8799-8584

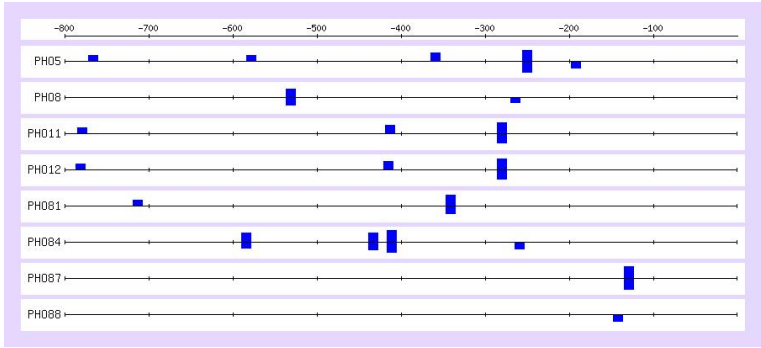
Aix-Marseille Université, France
Theory and Approaches of Genome Complexity (TAGC)

Institut Français de Bioinformatique (IFB)
http://www.france-bioinformatique.fr

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 matching a perfectly conserved hexanucleotide with strand-insensitive search (expectation: 1 occurrence every 2b) in non-coding sequences, depending to the genome size?

Organism	Genome size	Nb genes	Kb/gene	coding	non-coding	Non-coding size	Expected / genome	non-coding /gene	Expected / gene
	Mb		Kb	%	%	Mb		Kb	
Mycoplasma genitalium	0,6	481	1,25	90%	10%	0,1	28,85	0,12	0,03
Haemophilus influenzae	1,8	1.717	1,05	86%	14%	0,3	121,15	0,15	0,12
Escherichia coli	4,6	4.289	1,07	87%	13%	0,6	287,50	0,14	0,29
Saccharomyces cerevisiae	12	6.286	1,91	72%	28%	3,4	1.615,38	0,53	1,62
Arabidiopsis thaliana	120	27.000	4,44	30%	70%	84,0	40.384,62	3,11	40,38
Caenorhabditis elegans	97	19.000	5,11	27%	73%	70,8	34.043,27	3,73	34,04
Drosophila melanogaster	165	16.000	10,31	15%	85%	140,3	67.427,88	8,77	67,43
Homo sapiens	3.200	31.000	103,23	3%	97%	3.104,0	1.492.307,69	100,13	1.492,31

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