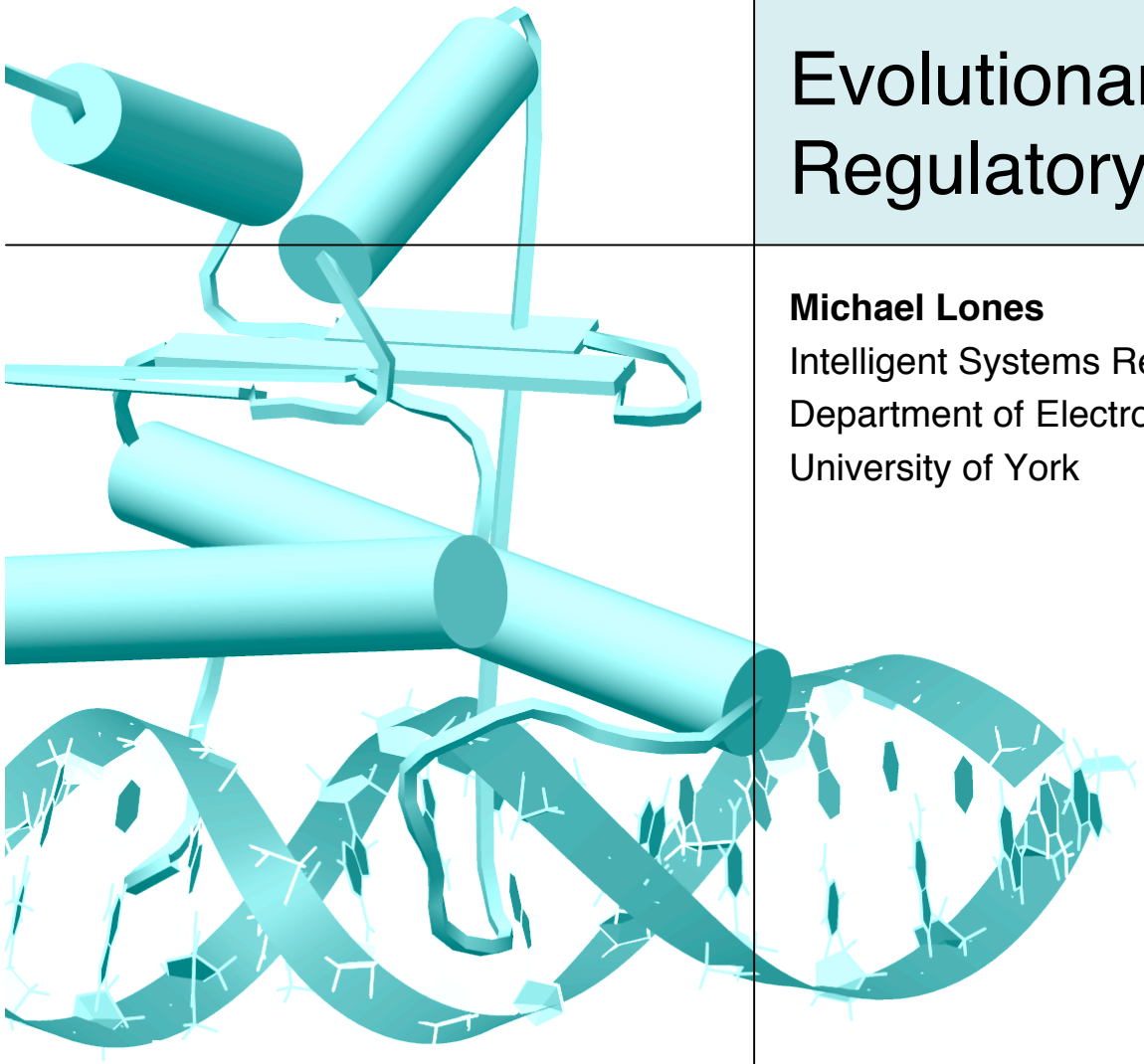
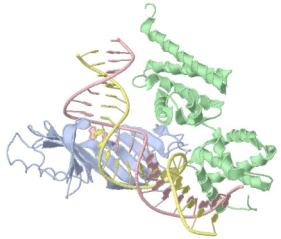


Evolutionary Algorithms for Regulatory Motif Discovery

Michael Lones
Intelligent Systems Research Group
Department of Electronics
University of York

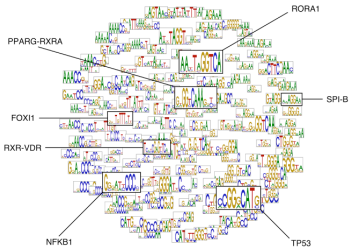
Overview

- What are regulatory motifs?

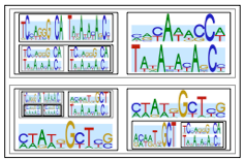


- Regulatory motif discovery

- Diversity in evolutionary algorithms

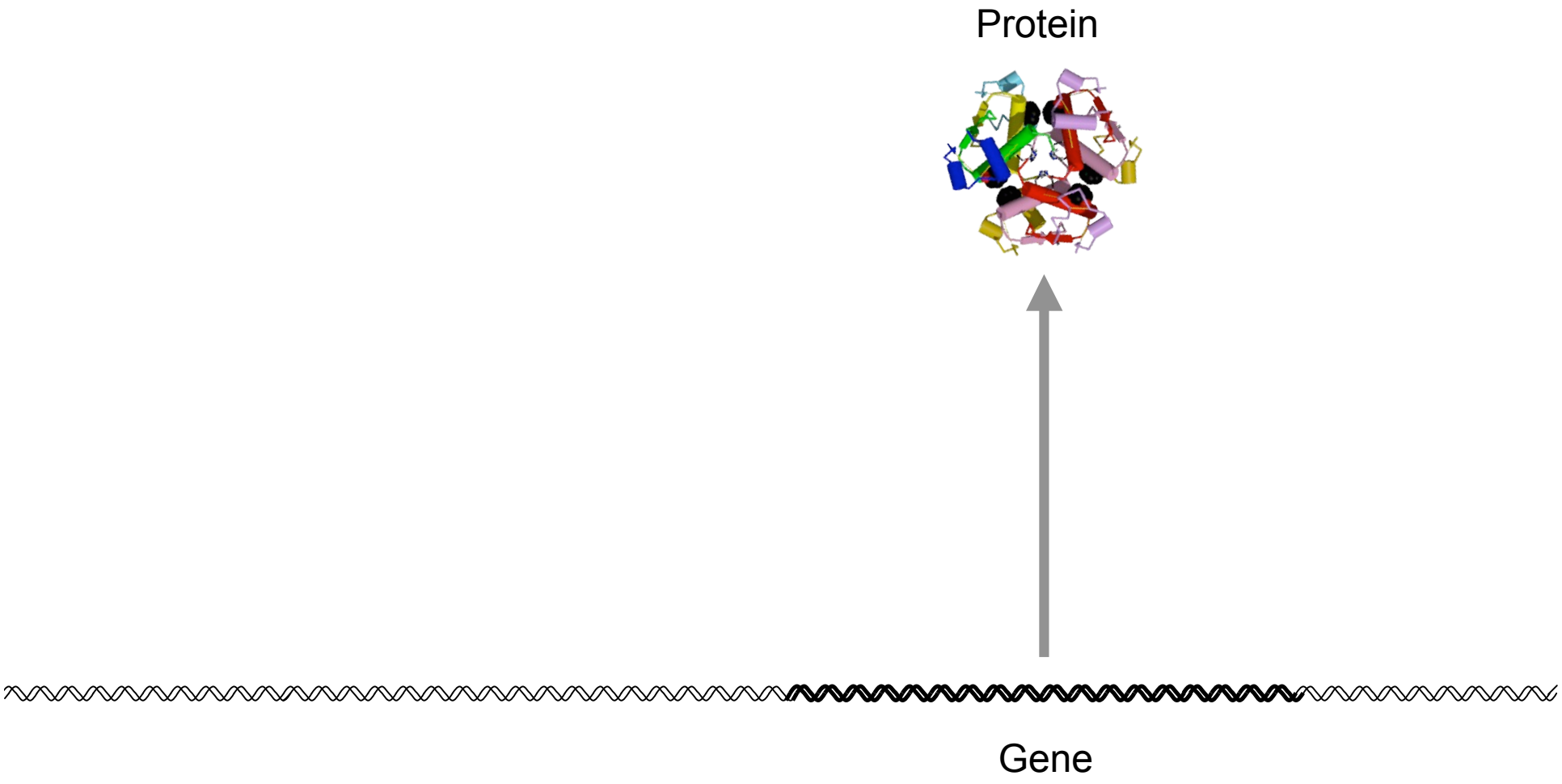


- Population clustering
  - Discovering regulatory motifs

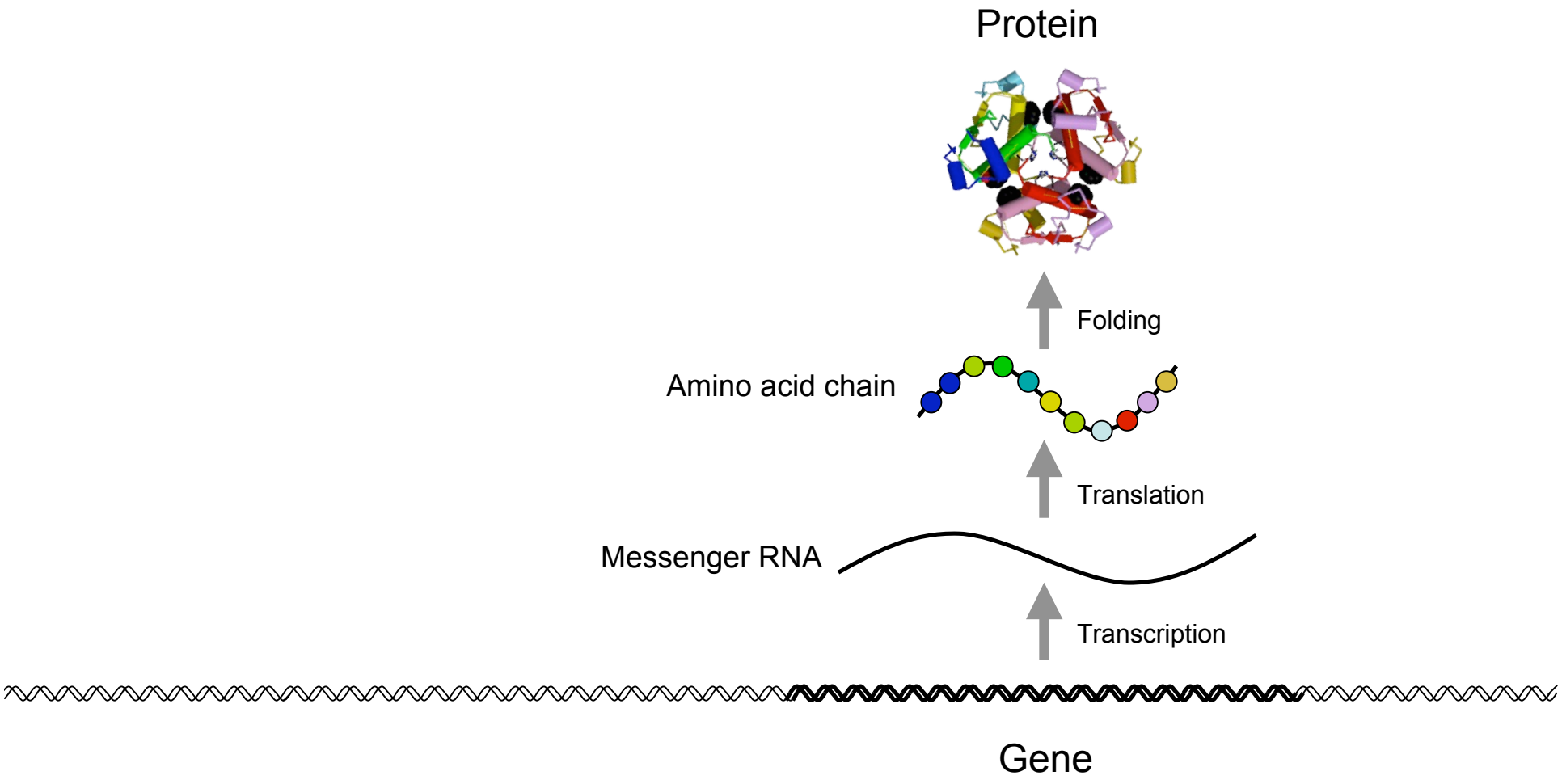


- Motif-rule co-evolution
  - Discovering higher-order motifs

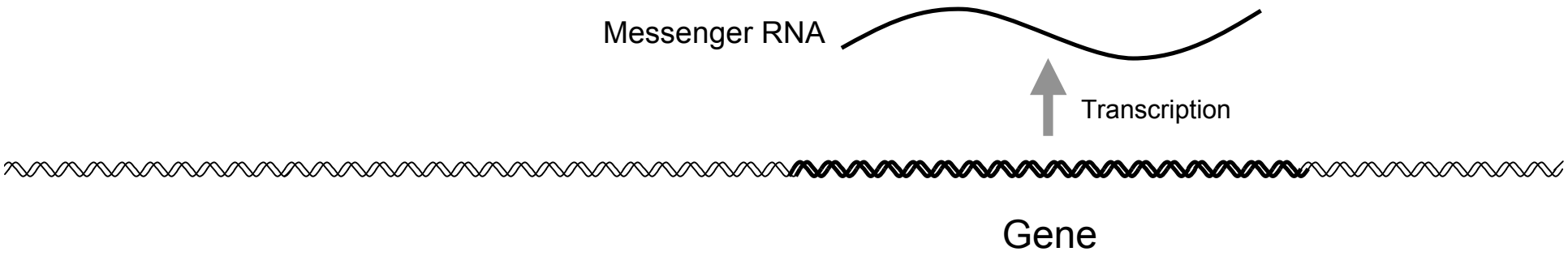
Gene Regulation



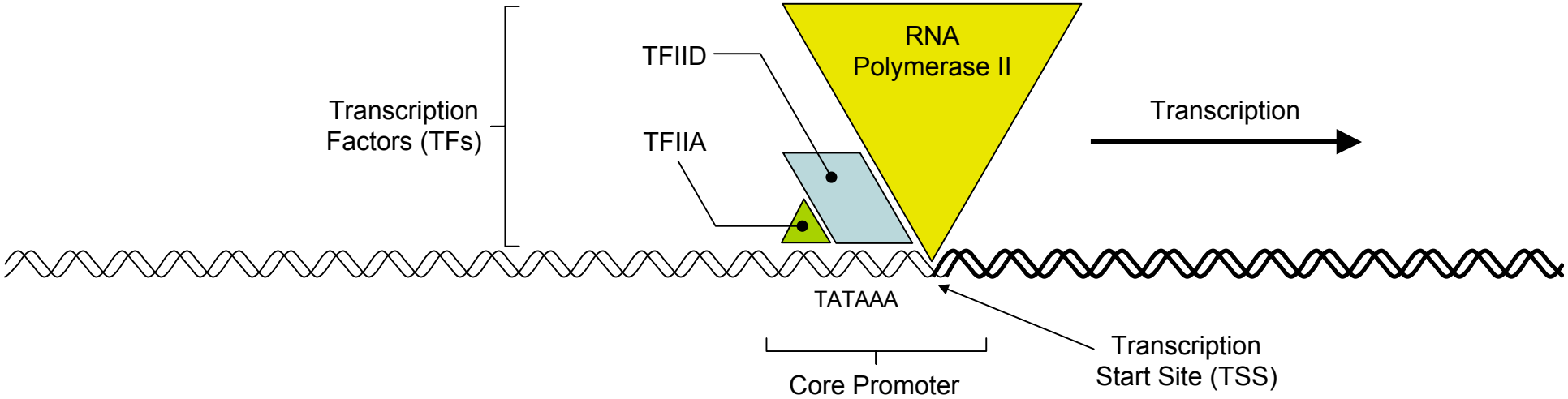
Gene Regulation



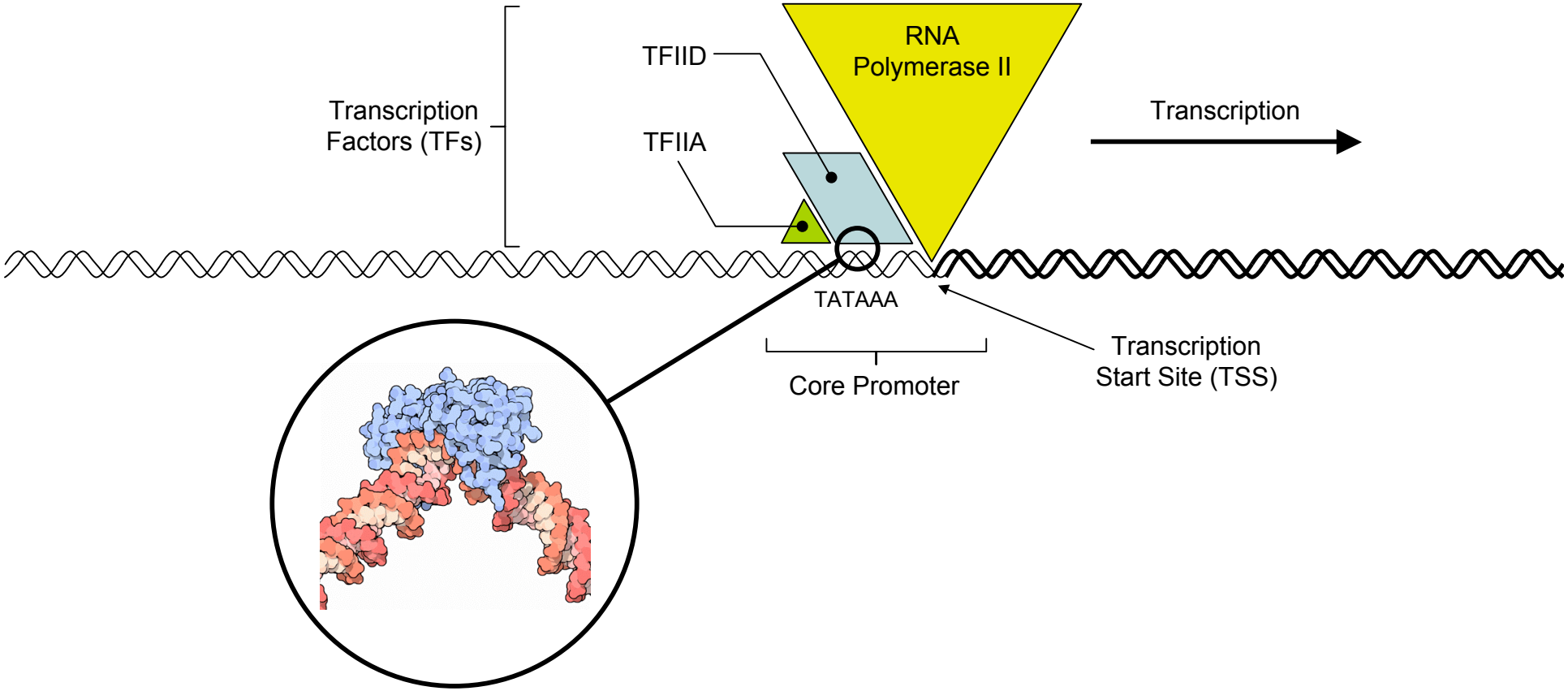
Transcription Regulation



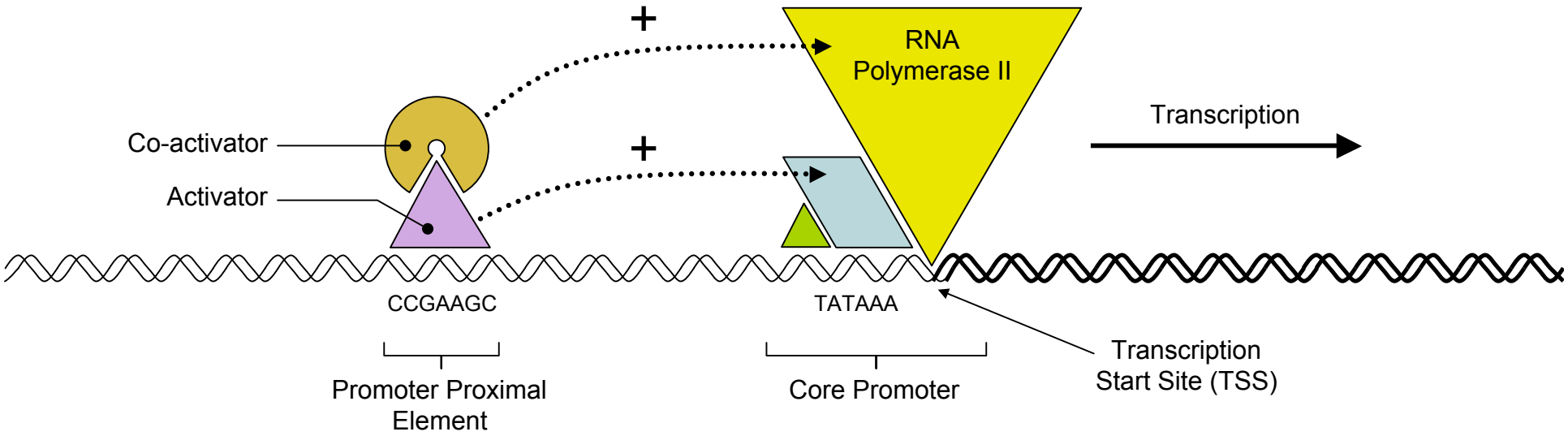
Basal Transcription Complex



Basal Transcription Complex

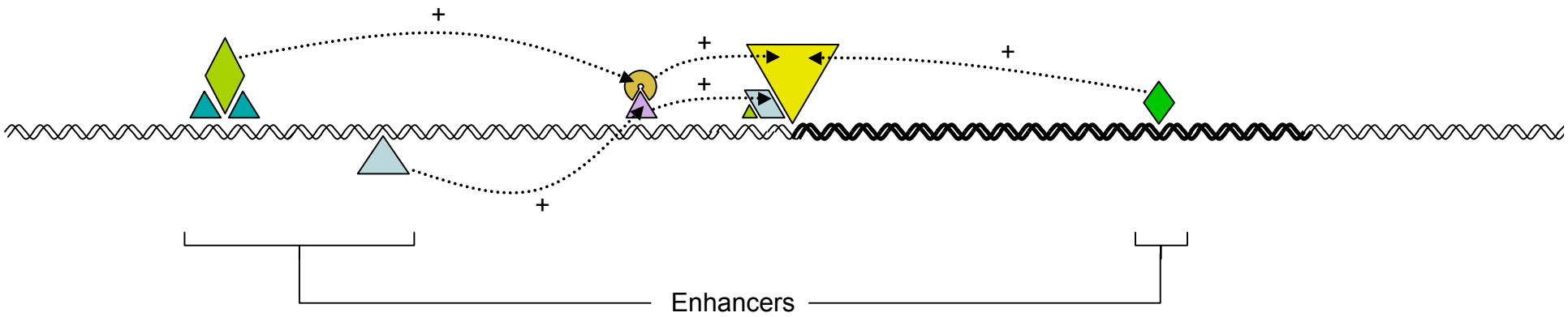


Promoter Proximal Elements



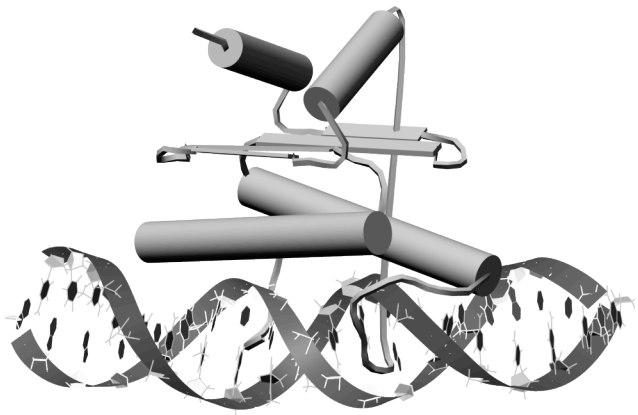
Enhancers

- Enhance the binding of the transcription complex or activators
- May occur up to ~10kb up/down stream
- Can occur in either orientation



Transcription Factor Binding Sites

- ~5-8bp in contact with TF
  - Requires certain nucleotides
- ~10-20bp occluded by TF
  - Constrains nucleotides

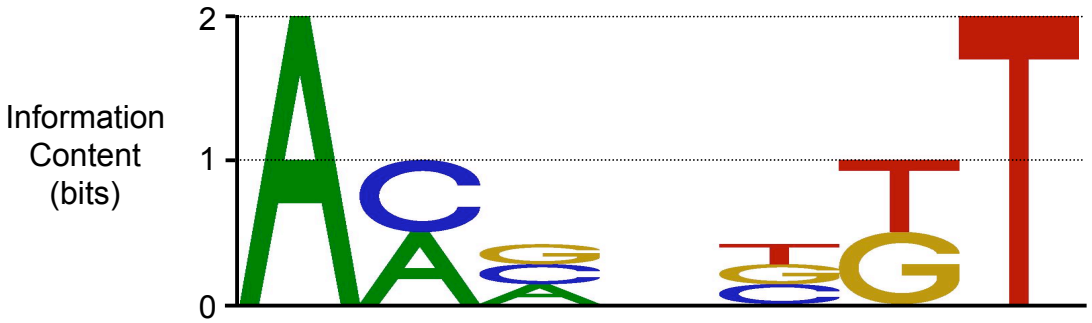


- Regulatory motifs describe TF binding sites (TFBSs)
  - Consensus sequences for conserved motifs, e.g. TATAAAA
  - Regular expressions also used, e.g. TATA[AT]A[AT]
  - Frequency matrices often used to capture variation
  - Hidden Markov models also sometimes used

Transcription Factor Binding Sites

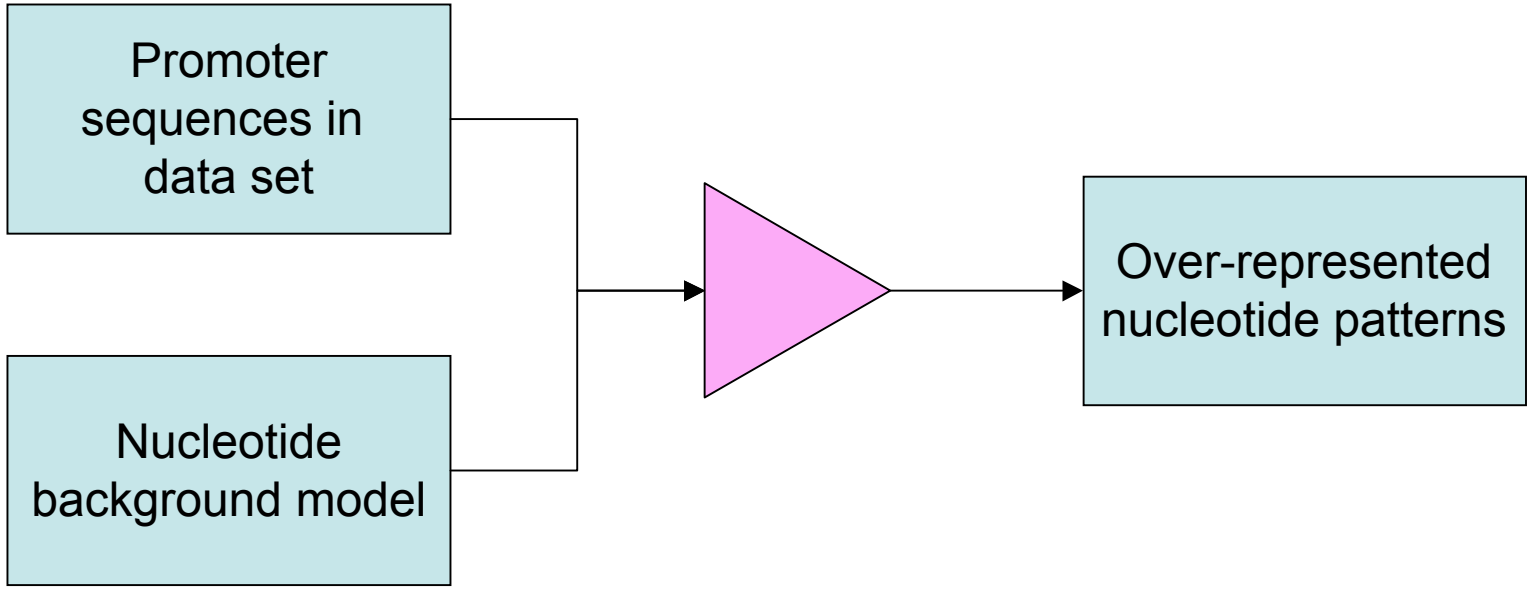
$$
\begin{array}{c}
A \\ C \\ G \\ T
\end{array}
\begin{pmatrix}
1.00 & 0.50 & 0.33 & 0.25 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.50 & 0.33 & 0.25 & 0.33 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.33 & 0.25 & 0.33 & 0.50 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.25 & 0.33 & 0.50 & 1.00
\end{pmatrix}
$$

Position frequency matrix (PFM)



Sequence logo

Regulatory Motif Discovery



- Discovery of patterns of DNA bases which are over-represented in the data set relative to the nucleotide background

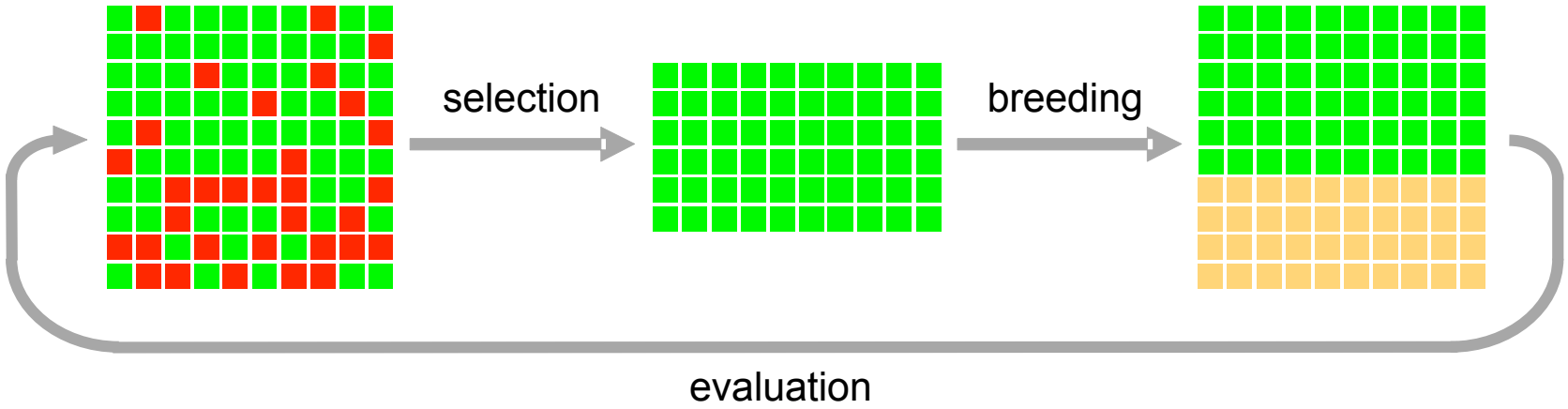
Motif Discovery Techniques

- Enumerative approaches
  - Enumerate every motif up to a certain length
  - Generally limited to short motifs and discrete models
  - Although always find most significant motifs where applicable
- Statistical approaches
  - Typically expectation-maximisation and/or Gibbs sampling
  - Iteratively refine initial estimate of motif model's parameters
- Other approaches
  - Bayesian modelling, neural networks, dynamic programming
  - Evolutionary computation

Limitations

- Sequence length is the main limiting factor
  - Generally limited to promoter sequences less than ~1kb
- Often discover biologically meaningless motifs
  - Most significant motifs are not necessarily meaningful
- Sensitivity to motif length
  - Poor performance with inappropriate model sizes
- Poor performance on metazoan data sets
  - Background models are biased towards yeast

Evolutionary Algorithms

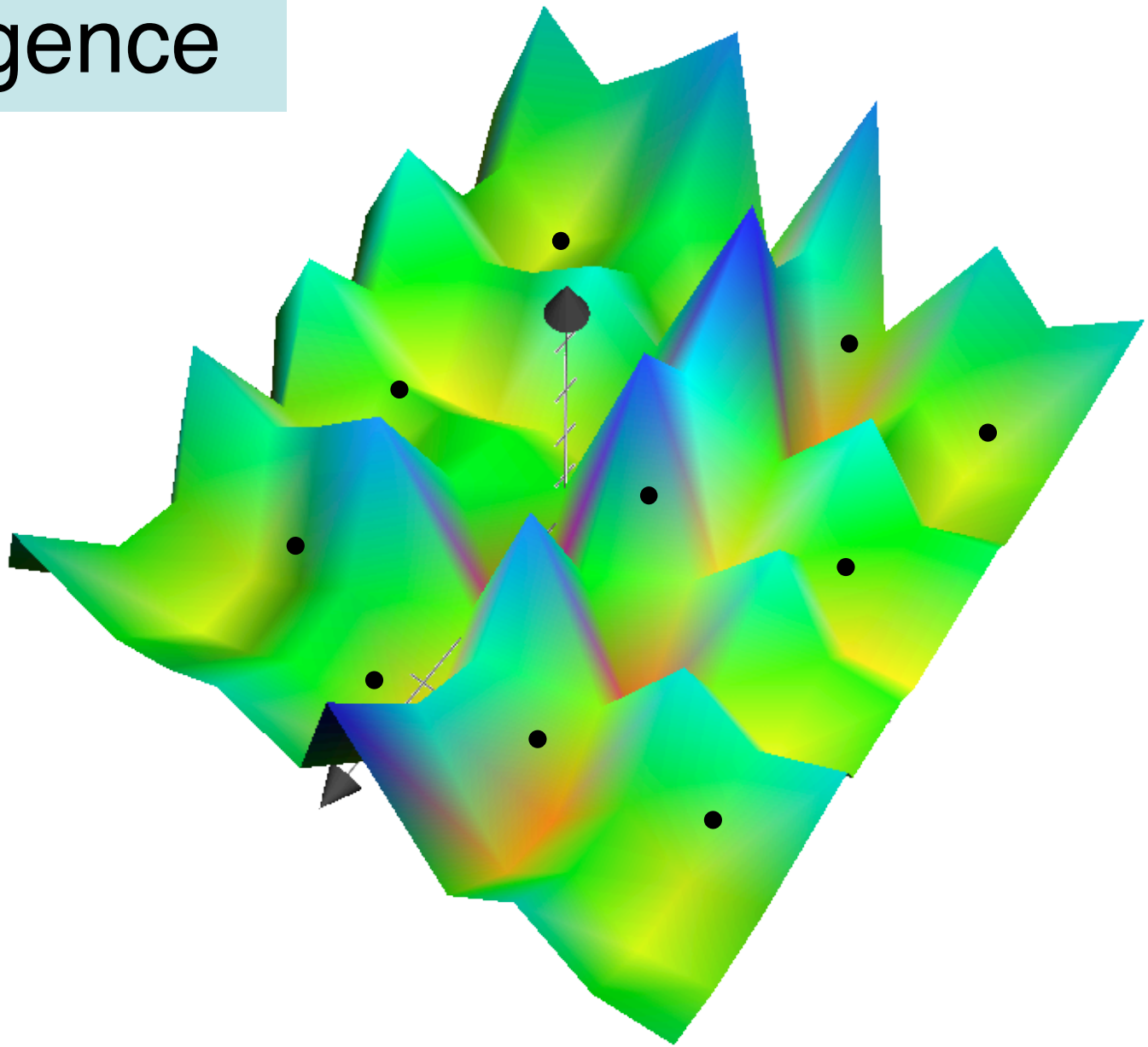


- Generate population of random solutions
- Repeat
  - Remove relatively poor solutions
  - Derive new solutions from relatively fit solutions
- until optimal solution found

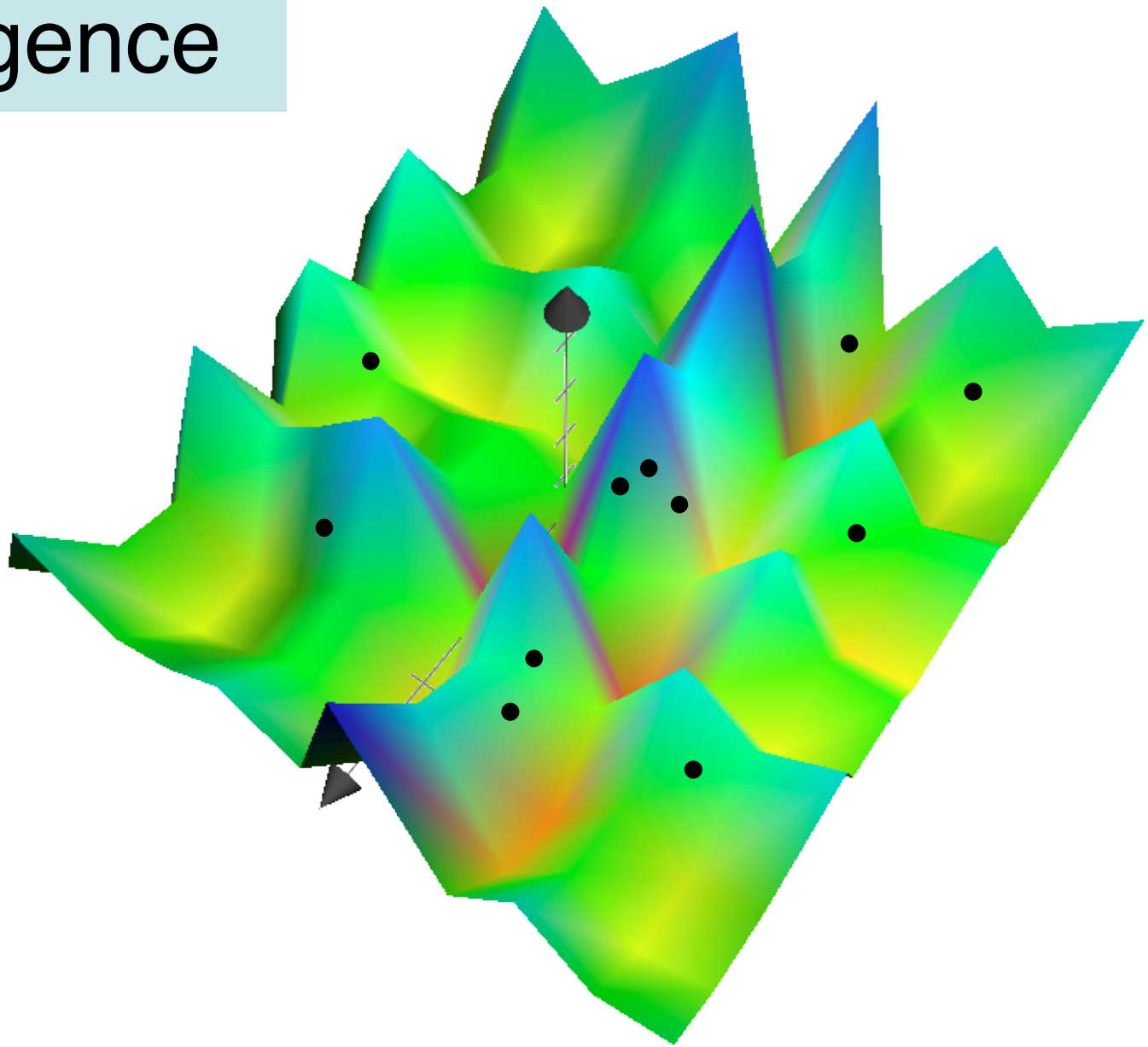
Evolutionary Algorithms

- Potential benefits for motif discovery
  - Global, non-exhaustive search with no specific heuristics
  - Representational flexibility
  - No dependence between solution derivation and scoring
  - Multiple solutions

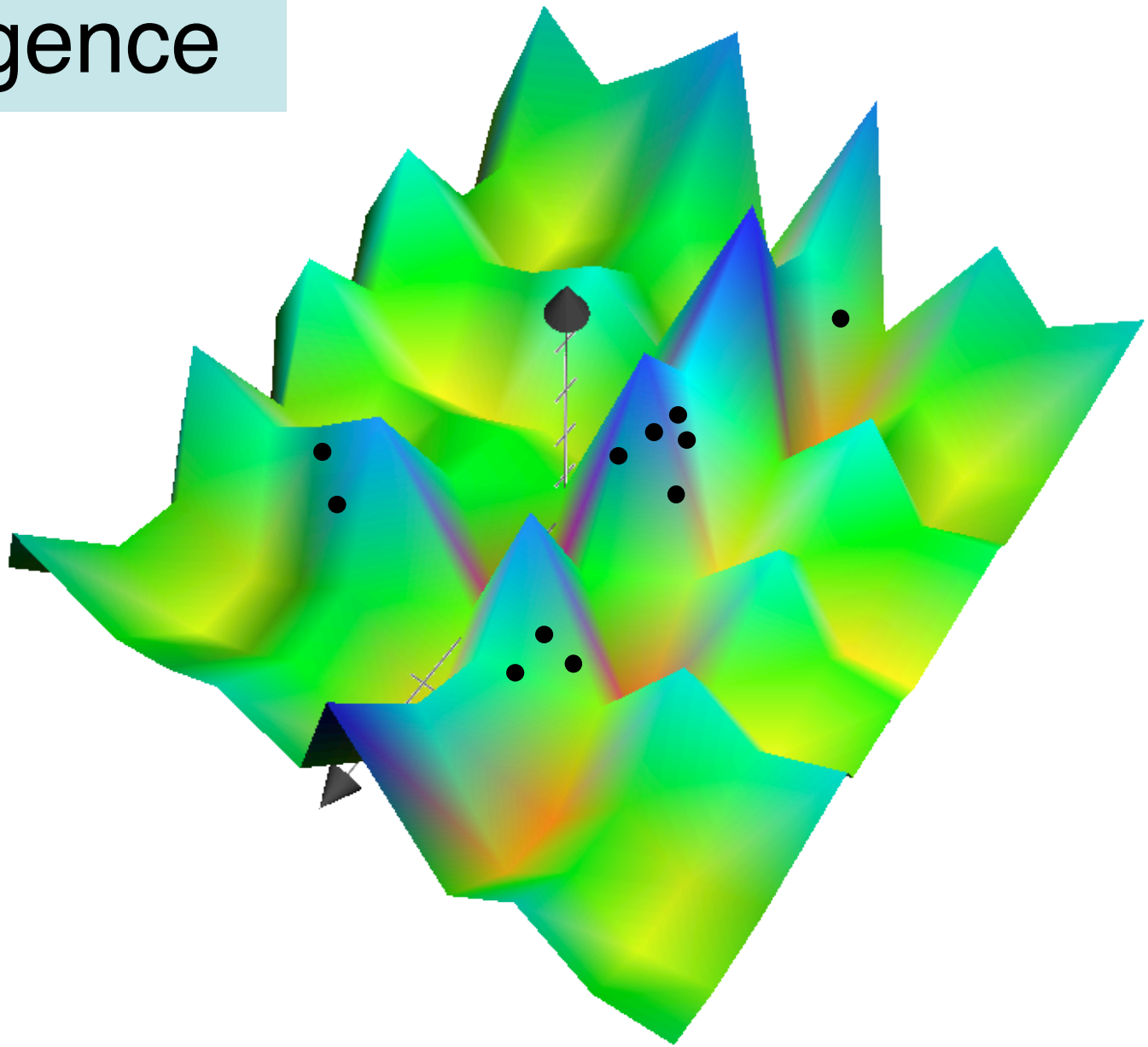
Convergence



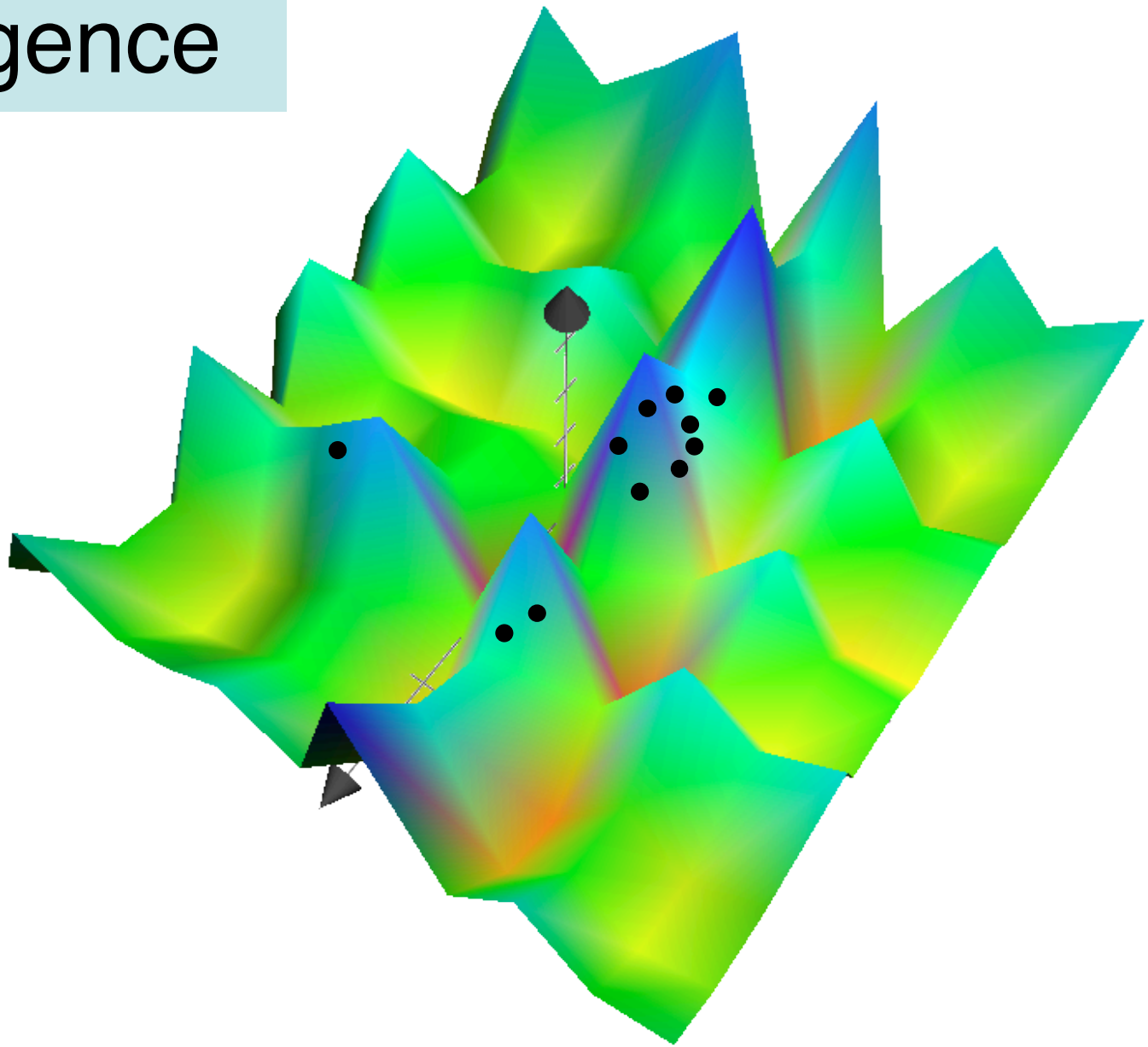
Convergence



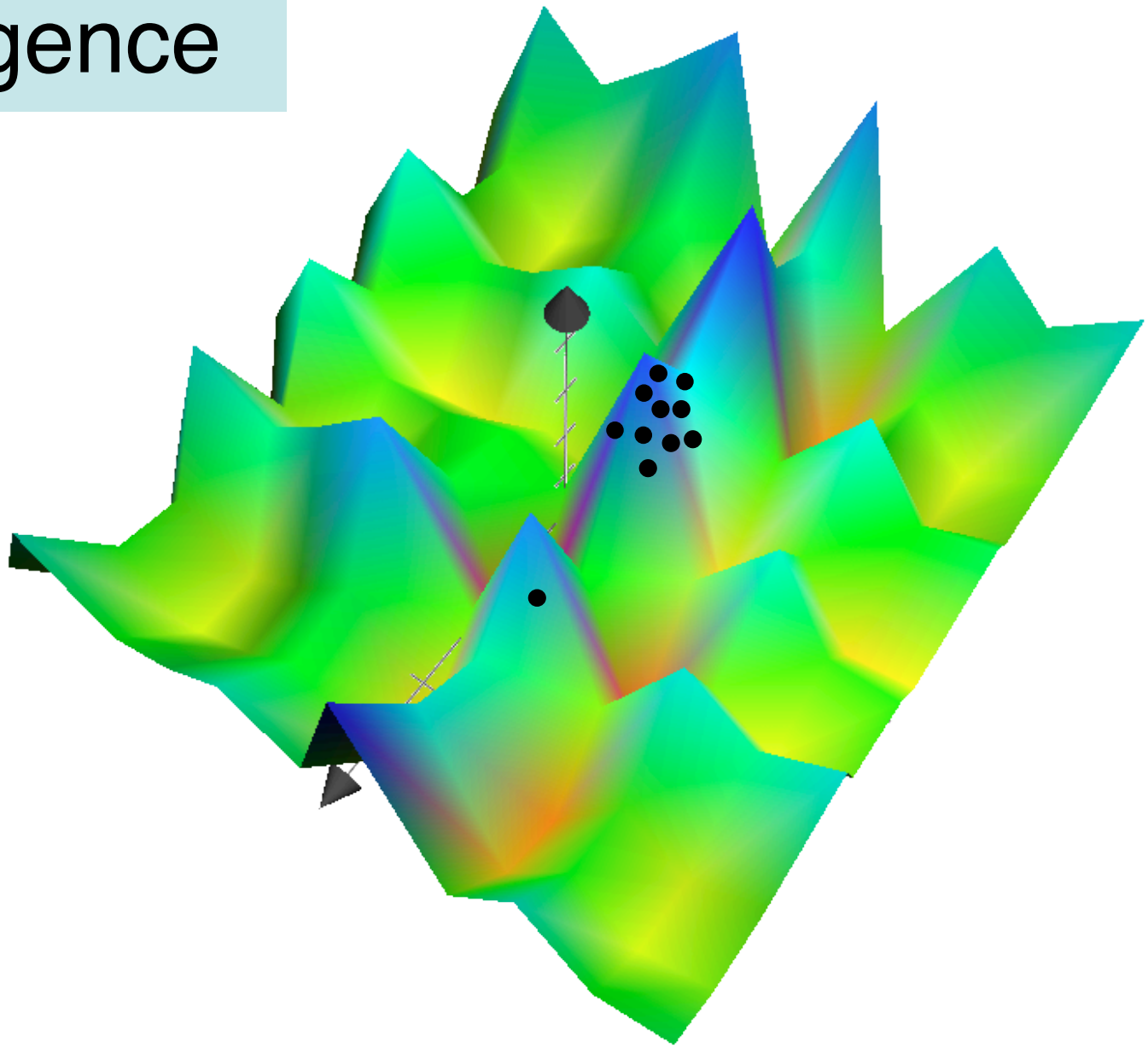
Convergence



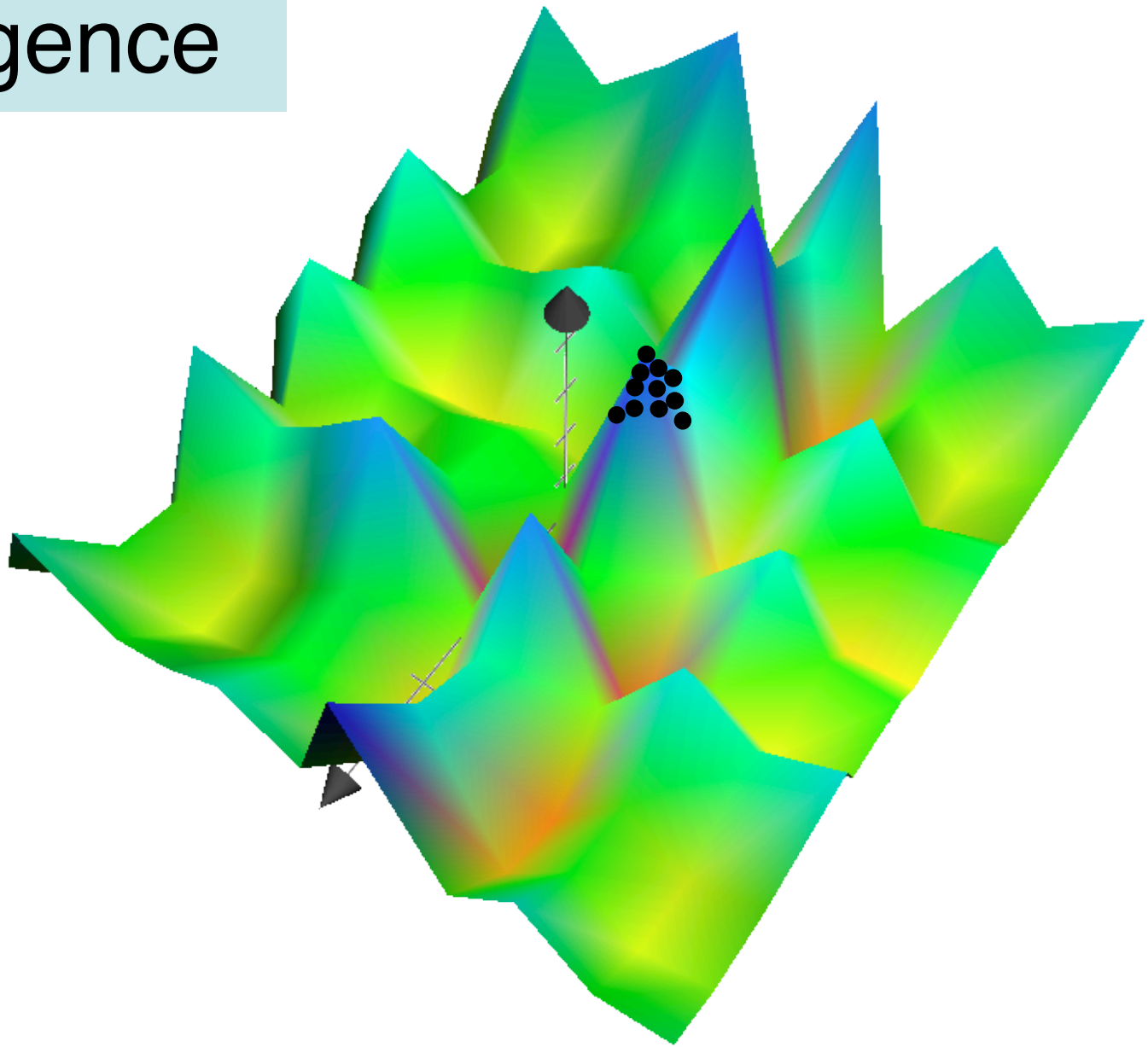
Convergence



Convergence



Convergence

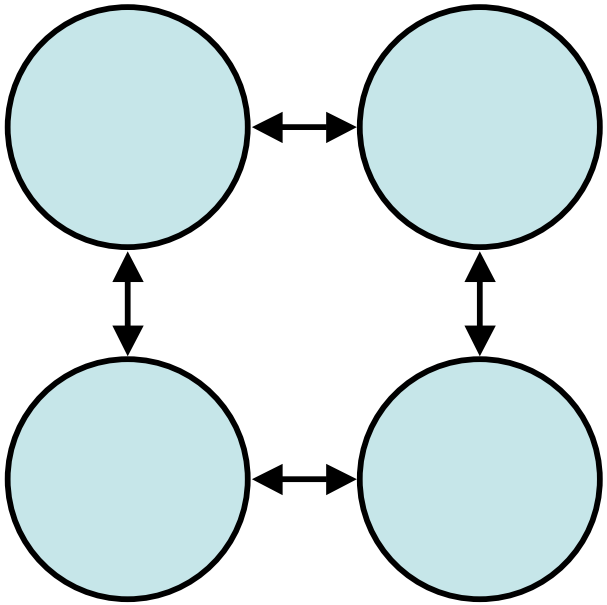


Niching Methods

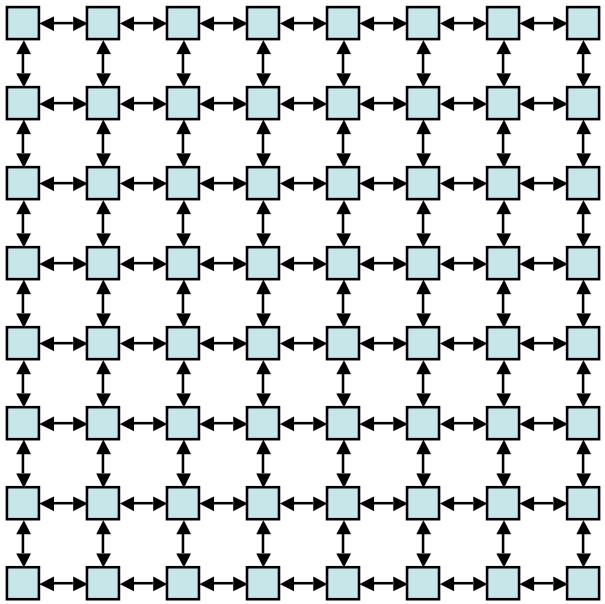
- Fitness sharing
  - Reduce the fitness of over-represented solutions
- Crowding
  - Replace over-represented solutions with new ones
- Sexual selection
  - Limit crossover to similar solutions
- Distributed populations
  - Split the population into multiple sub-populations
  - or spatially-distribute the population

Distributed Populations

Island model



Spatially-distributed



Niching Methods

- Fitness sharing
  - Reduce the fitness of over-represented solutions
- Crowding
  - Replace over-represented solutions with new ones
- Sexual selection
  - Limit crossover to similar solutions

Indirect

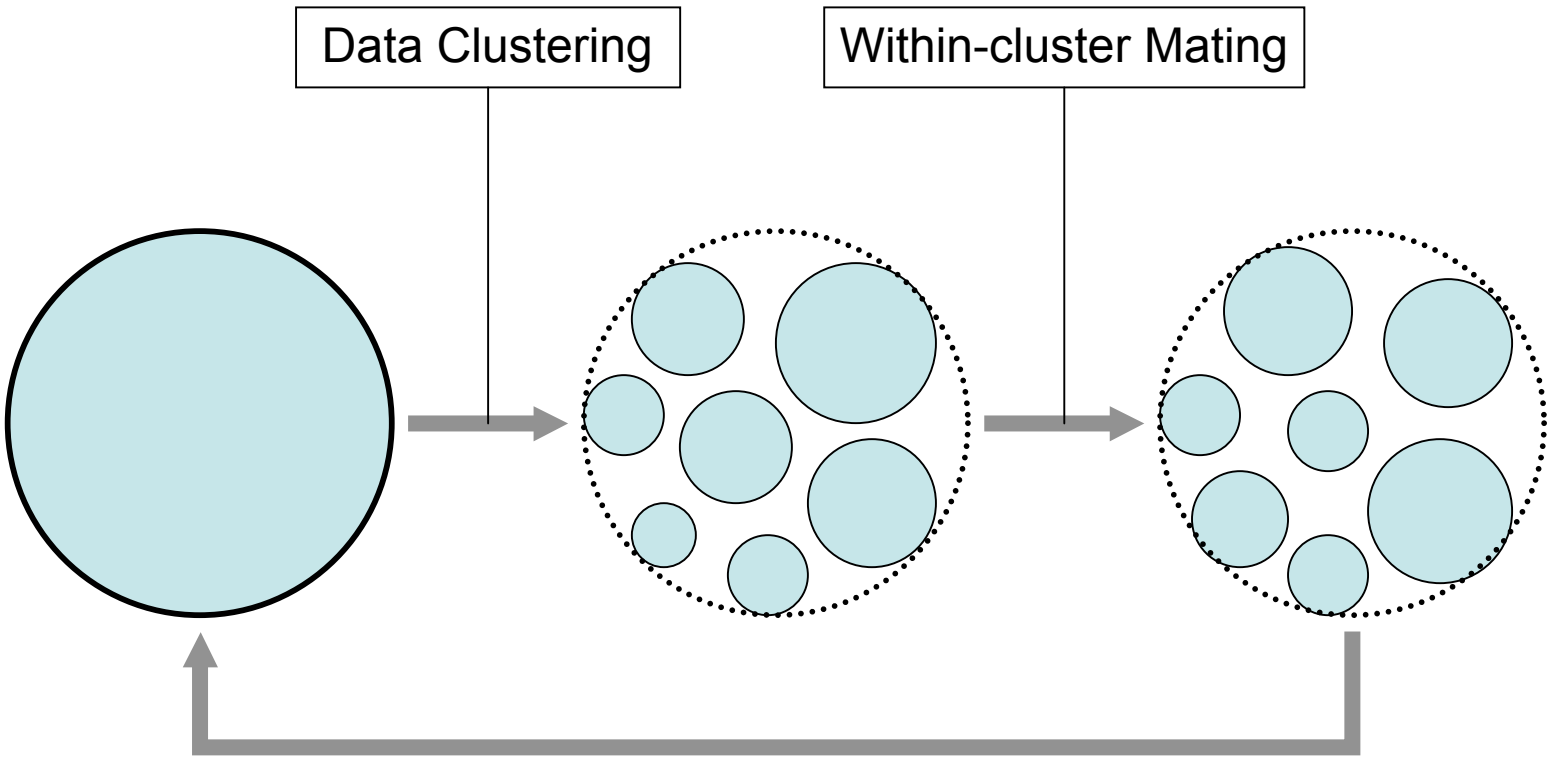
- Distributed populations
  - Split the population into multiple sub-populations
  - or spatially-distribute the population

Direct

Population Clustering

- Uses a data clustering algorithm
  - Applied to the population prior to reproduction
- Mating takes place solely within clusters
  - Maintaining each cluster's genetic identity
- Number of children proportional to cluster fitness
  - More exploration within fit clusters
- All clusters generate children
  - Maintains overall genetic diversity of population

Population Clustering



Population Clustering

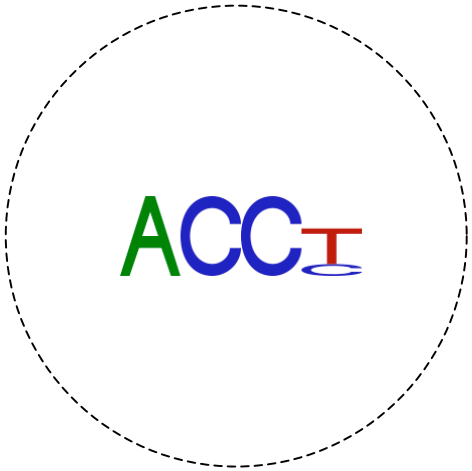
- Local selection and mating
  - Unlike indirect methods
  - Different selective pressures within and between clusters
- Population is distributed logically
  - Rather than via arbitrary evolutionary history
  - More likely to cover the search space

Population Clustering

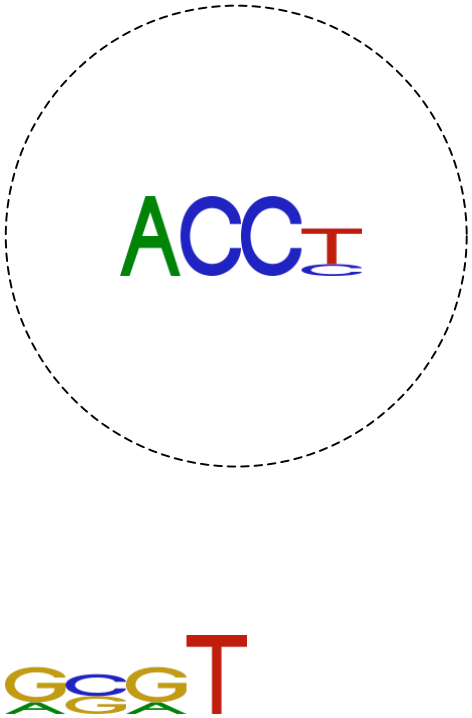
- Most data clustering algorithms are iterative
  - k-means, ISODATA, Kohonen neural networks
- Incremental clustering algorithms
  - Leader, ART, cobweb, genIC
  - Lower time complexity
- Leader sequential clustering
  - Single pass through data
    - Assign each data item to nearest cluster centroid
    - Or, if no nearby cluster, create new cluster and insert item

Population Clustering

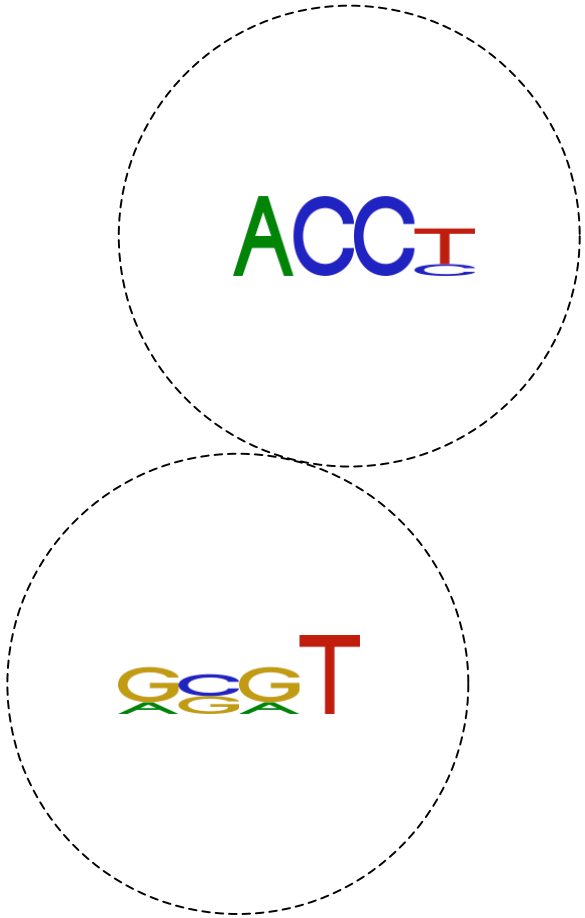
Population Clustering



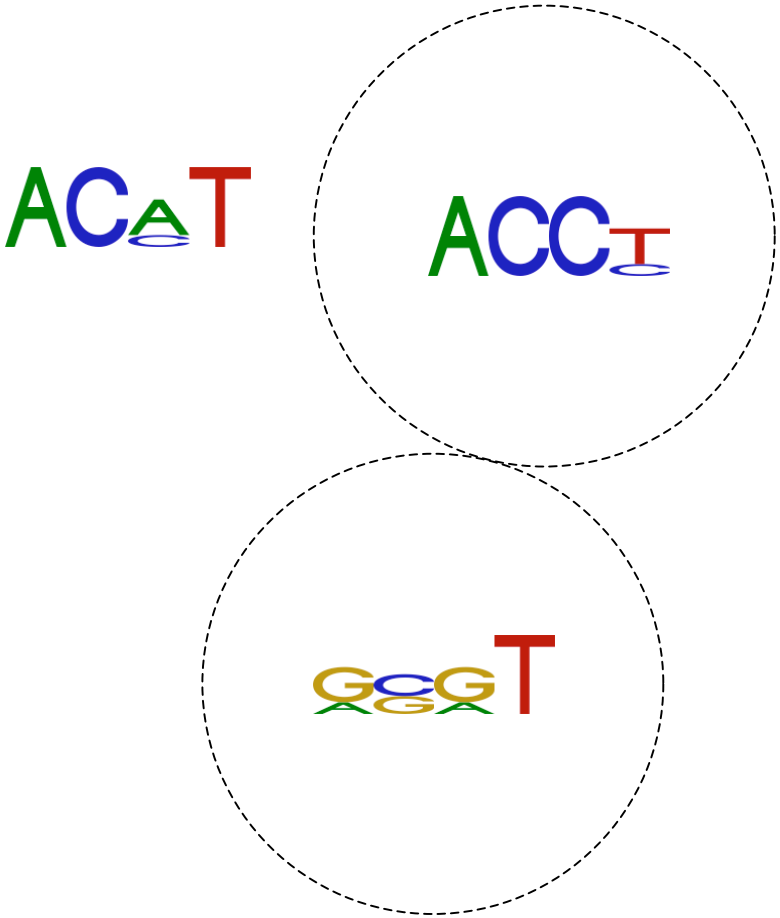
Population Clustering



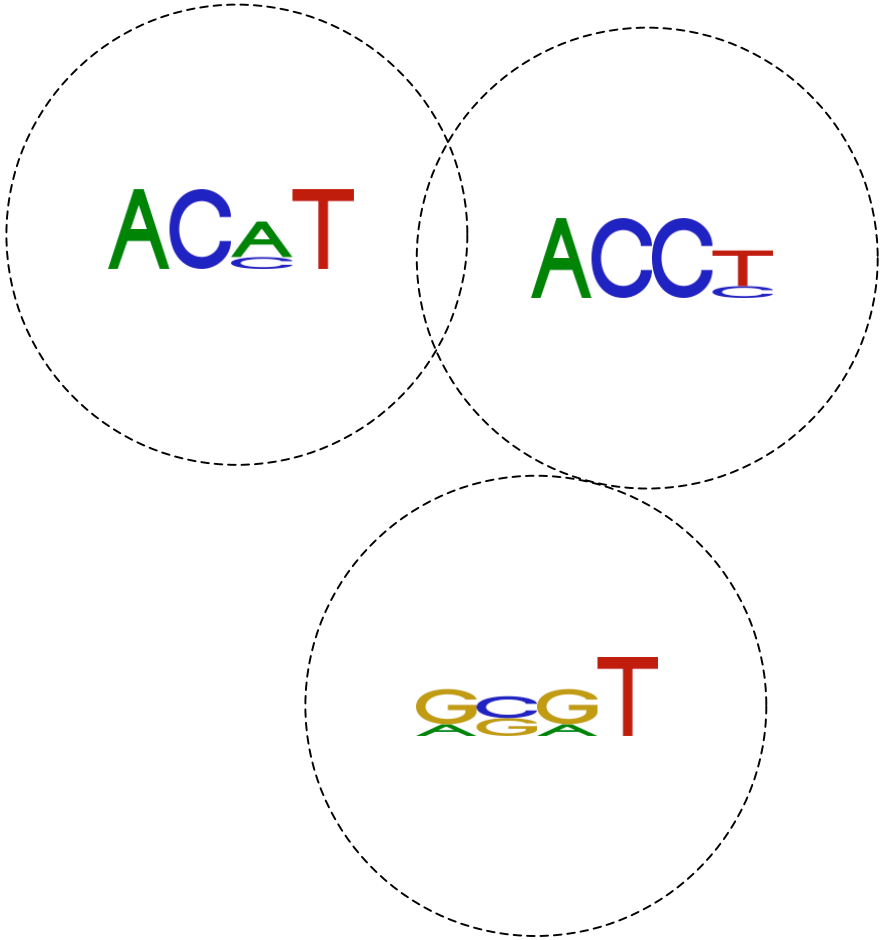
Population Clustering



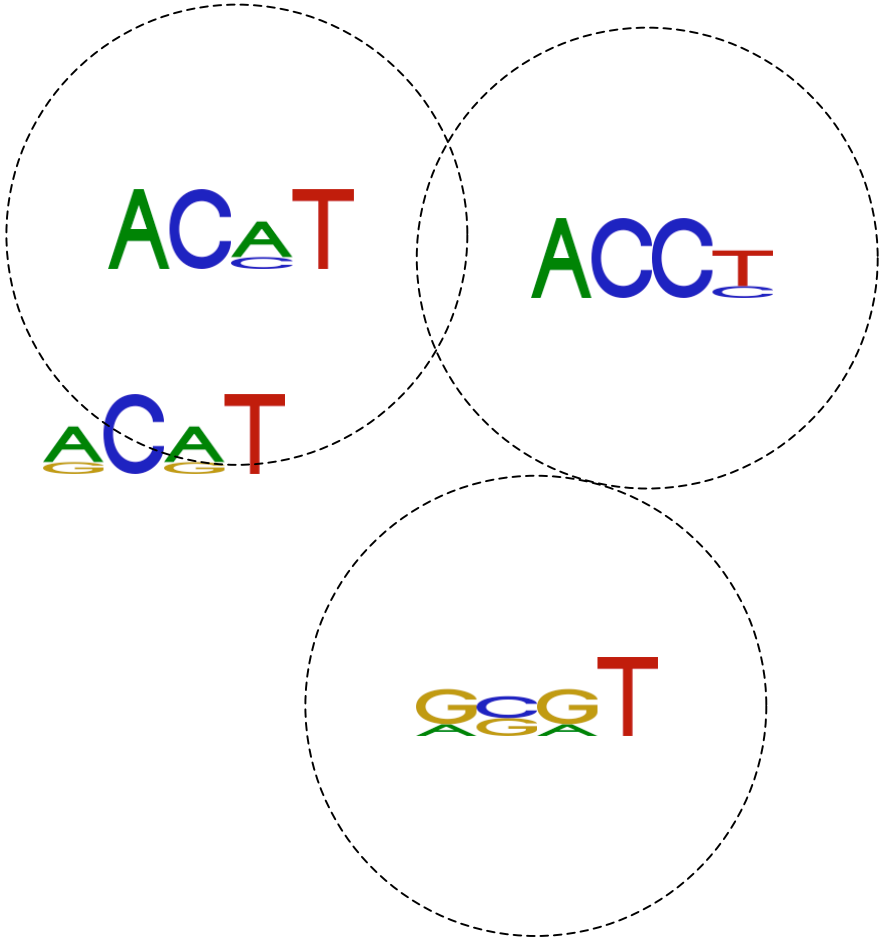
Population Clustering



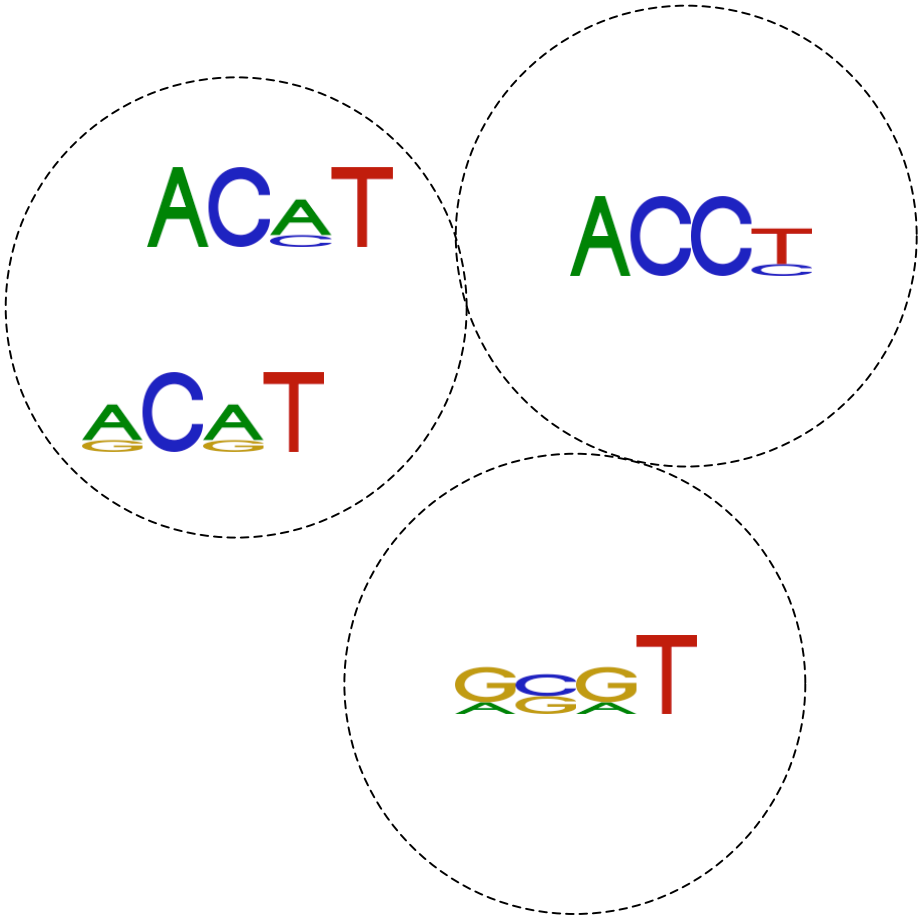
Population Clustering



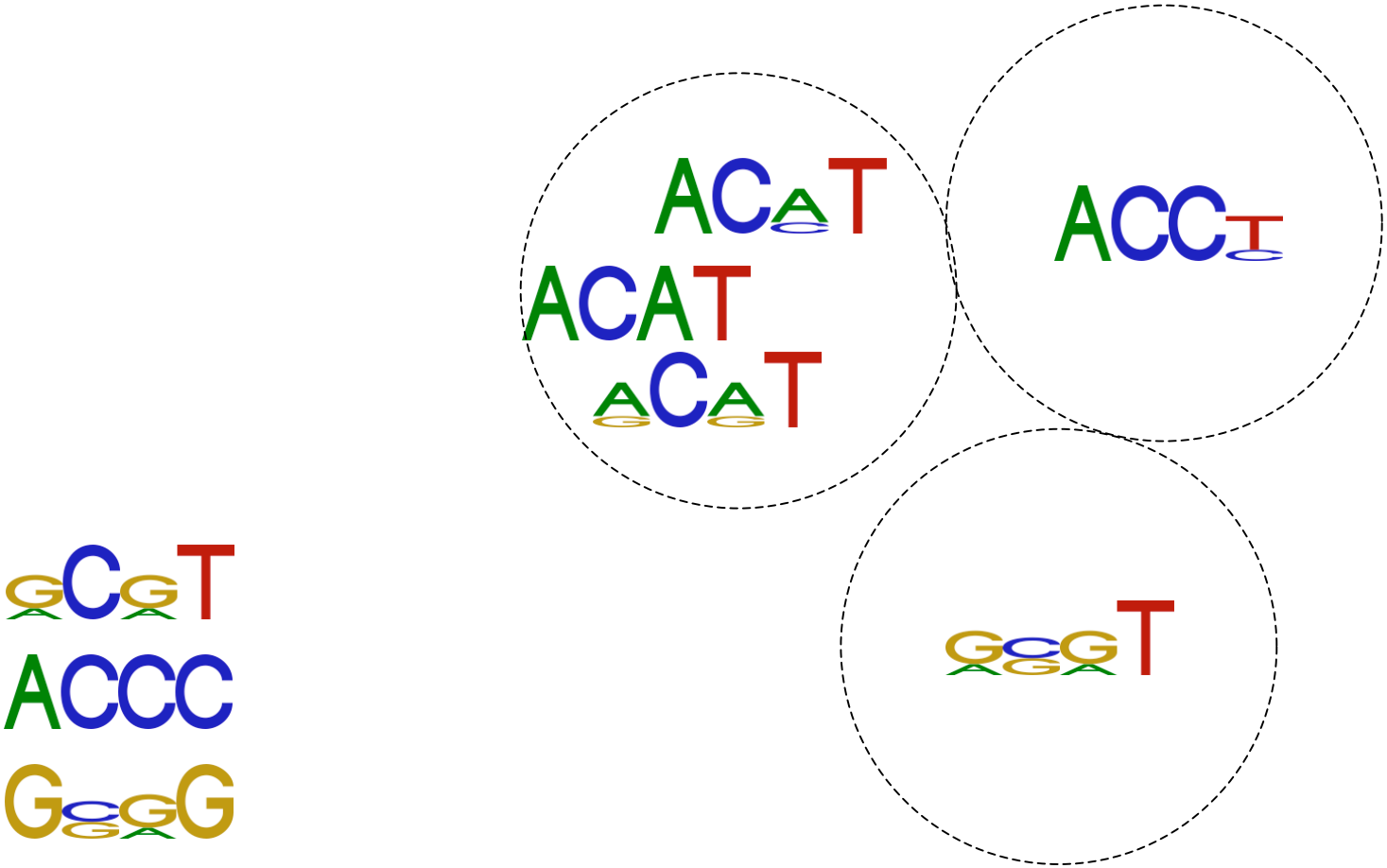
Population Clustering



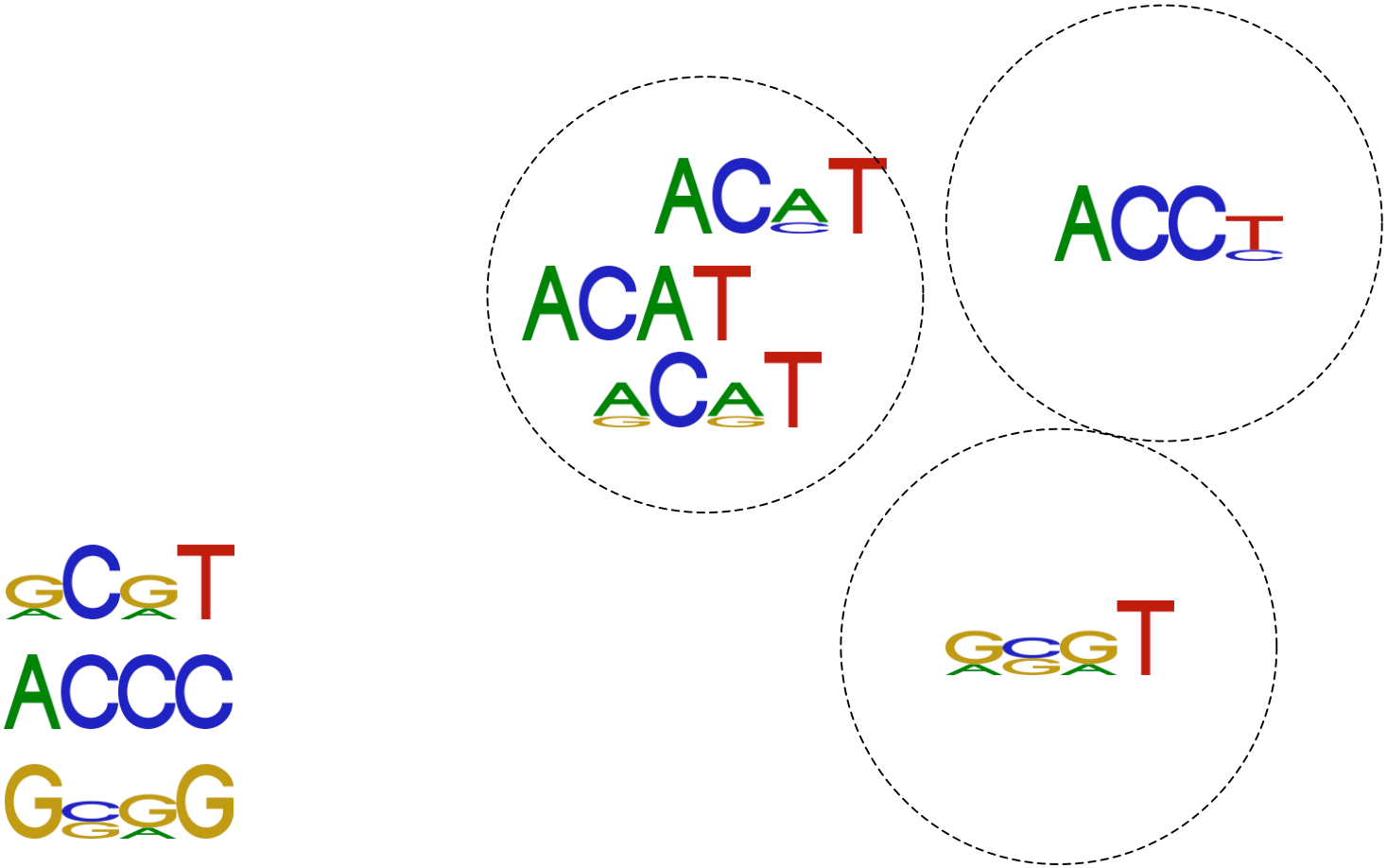
Population Clustering



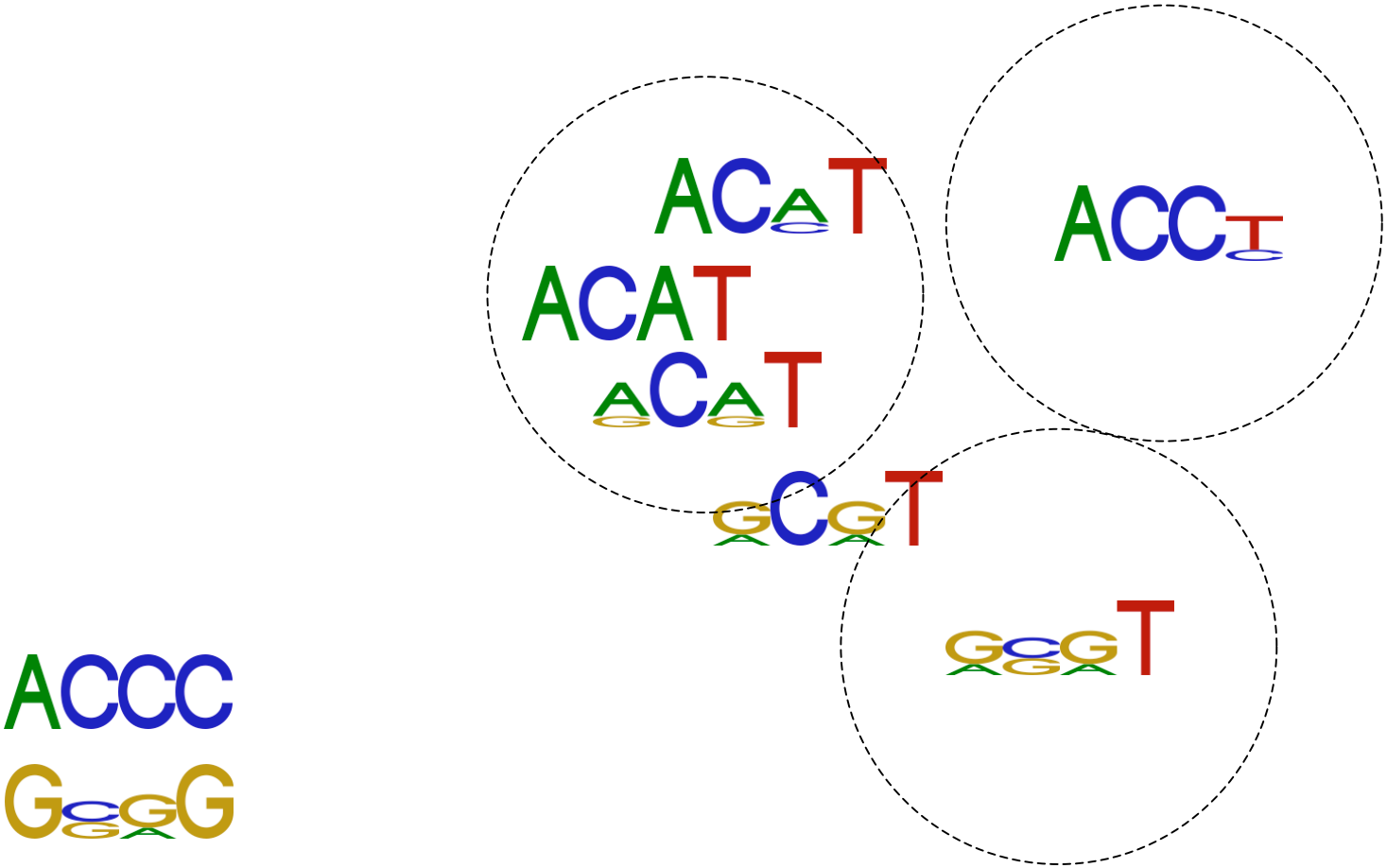
Population Clustering



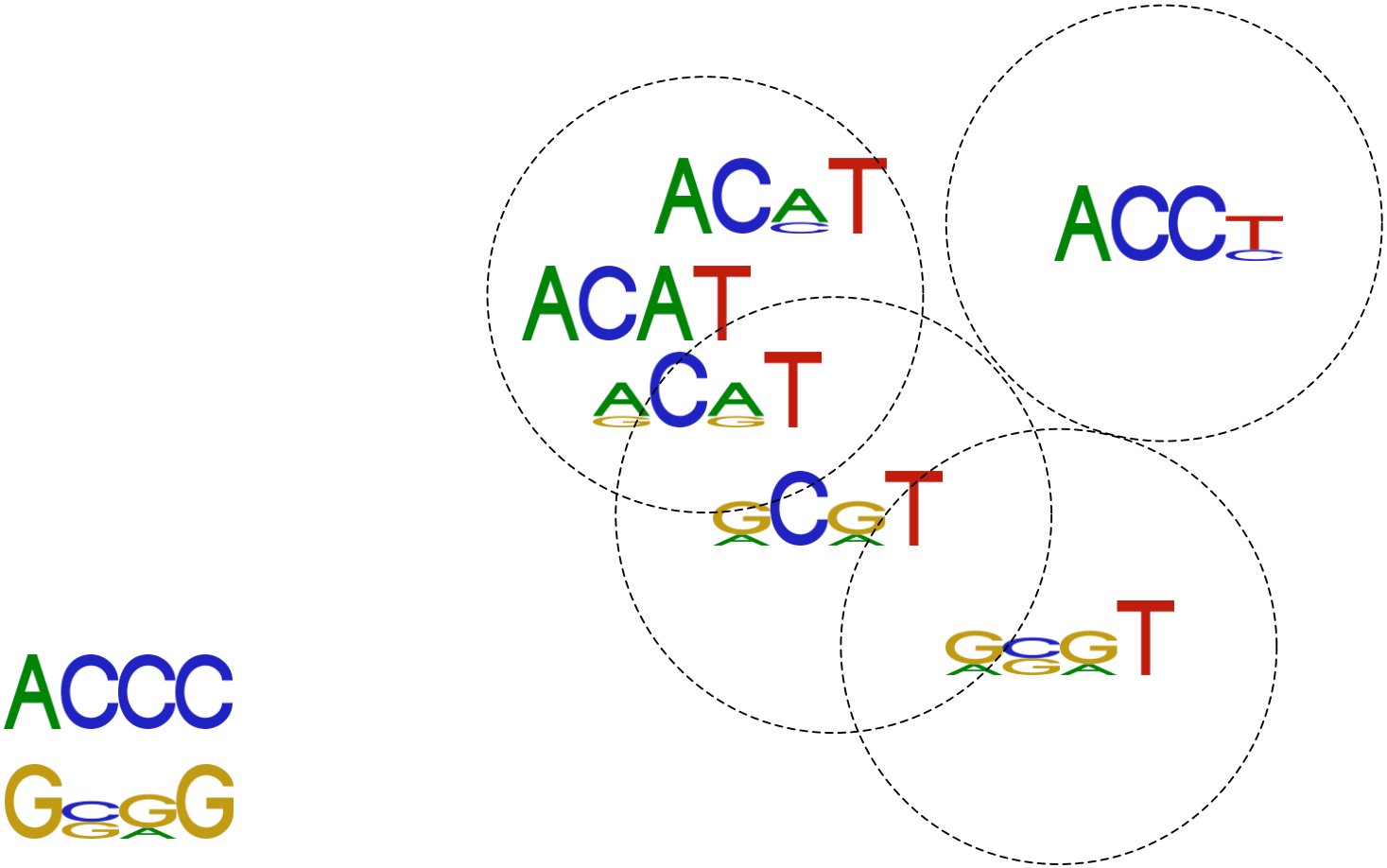
Population Clustering



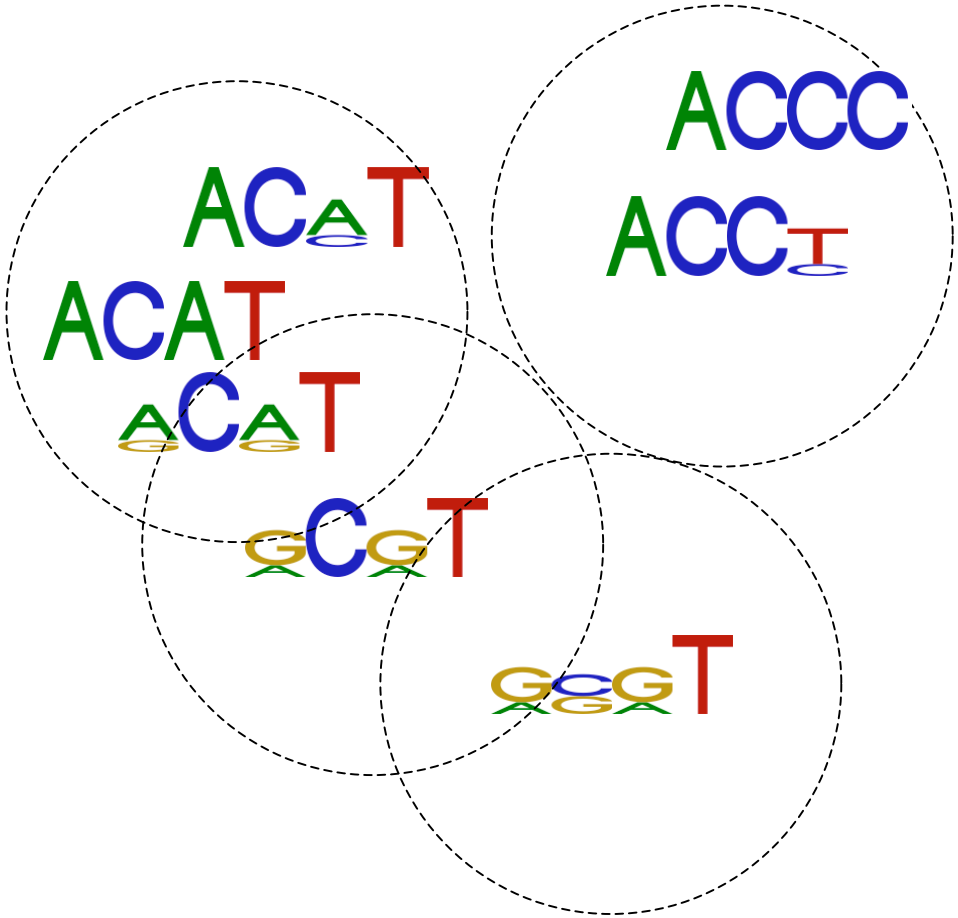
Population Clustering



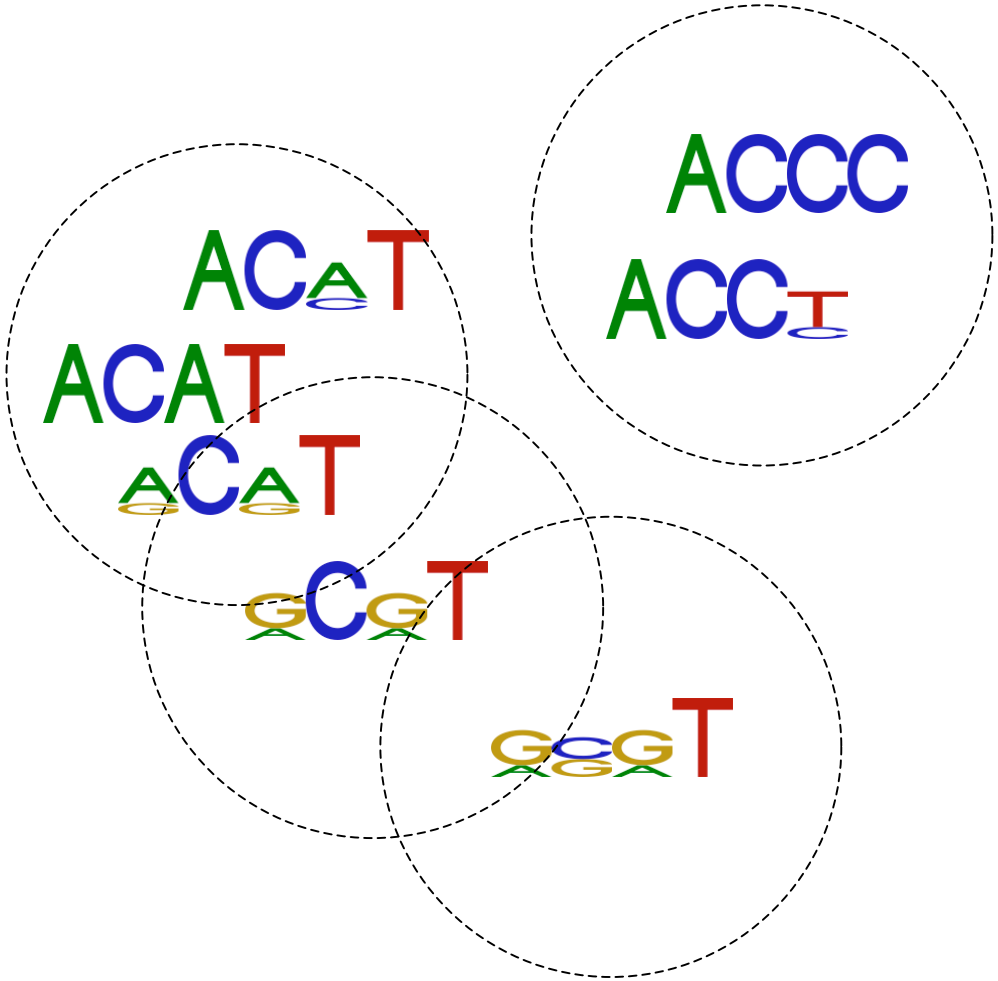
Population Clustering



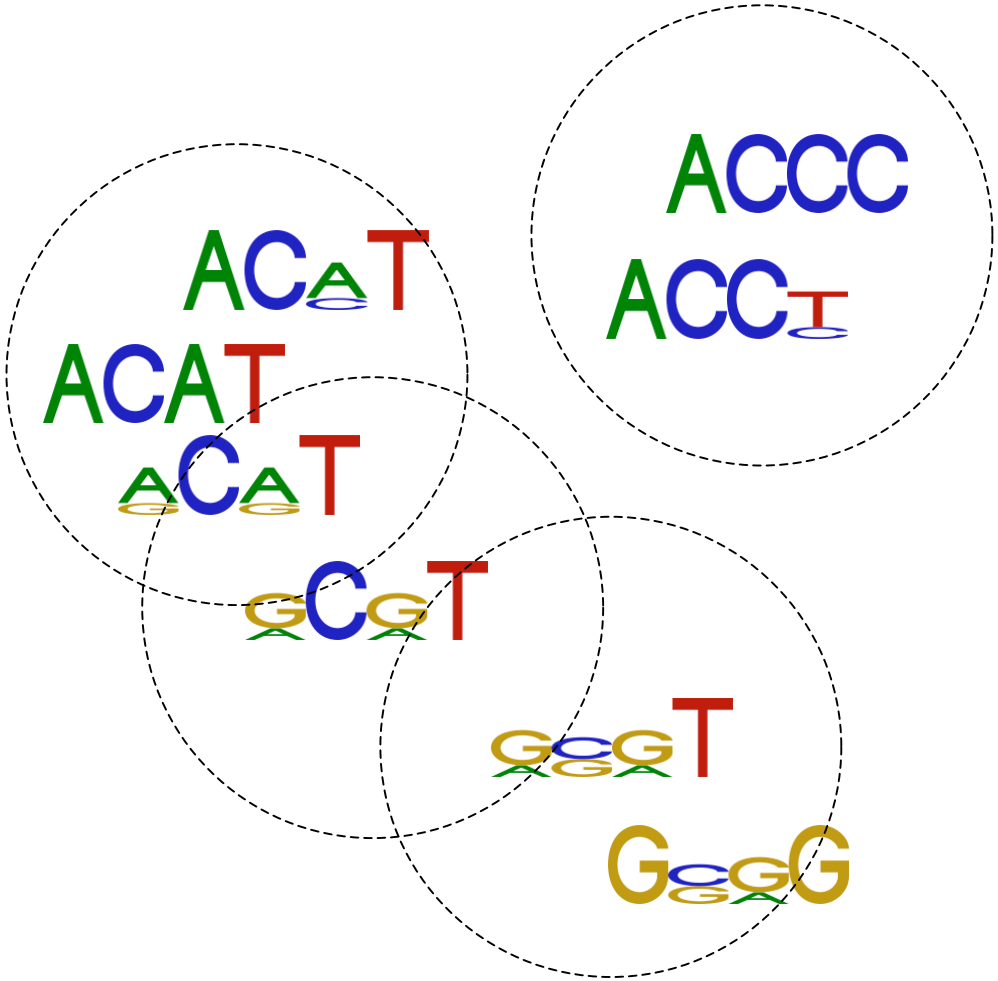
Population Clustering



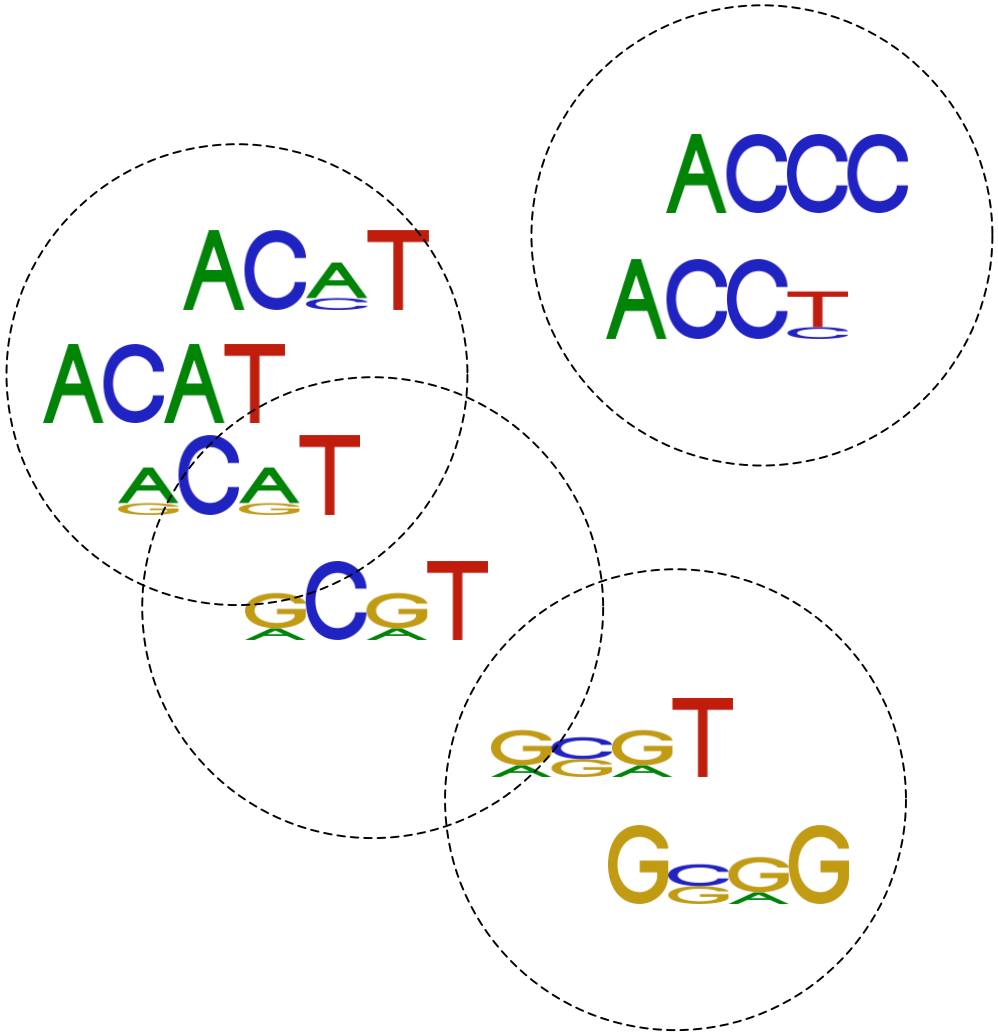
Population Clustering



Population Clustering

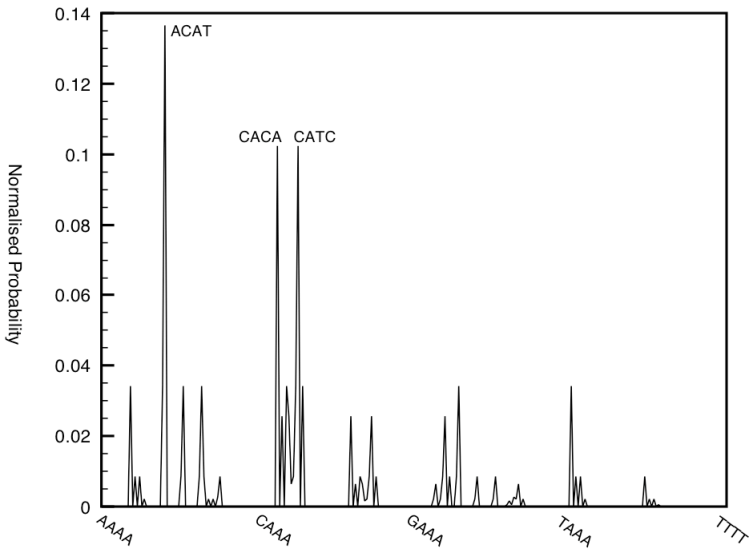


Population Clustering



Clustering Metric

- Distance between tetra-nucleotide distributions
  - Sum of probabilities of each 4-tuple of bases
  - Normalised by length of profile
  - Used to identify TFBS families in Transfac [Grote et al, 1999]



Evaluation

- Promoter sequences can be pretty long (≤ ~10kb)
  - Determine maximum searchable sequence sizes
- They usually contain multiple motifs
  - Look for multiple motifs at once
- Promoter regions are not generally well understood
  - Use synthetic data containing known motifs
  - generated by embedding JASPAR motifs into EPD sequences

Single Motif

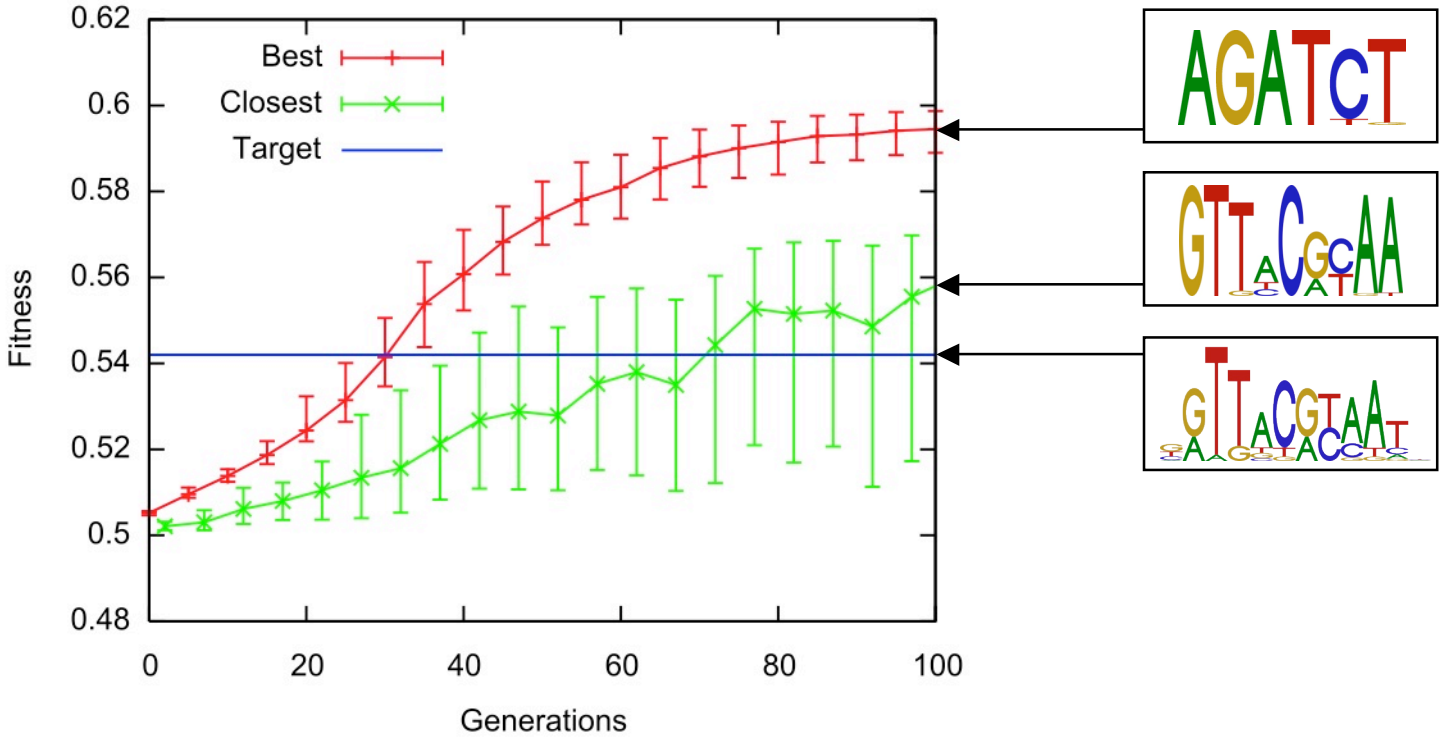
Data sets: 100 EPD sequences, 50 of which contain a single instance of the target motif

Background set: 1500 EPD sequences

Motif	Info. content	Sequ. Length	Pop size	Success (20 runs)	Evolved example
HFH-1 	14.07	5000	4000	90%	
HLF 	11.05	1500	3000	95%	
C-FOS 	10.67	1500	4000	95%	

Single Motif

Target = HLF, Information content = 11.15
Sequence length = 1500bp, Population size = 3000, Background set size = 1500 sequences
Success rate = 95% (19/20 runs)



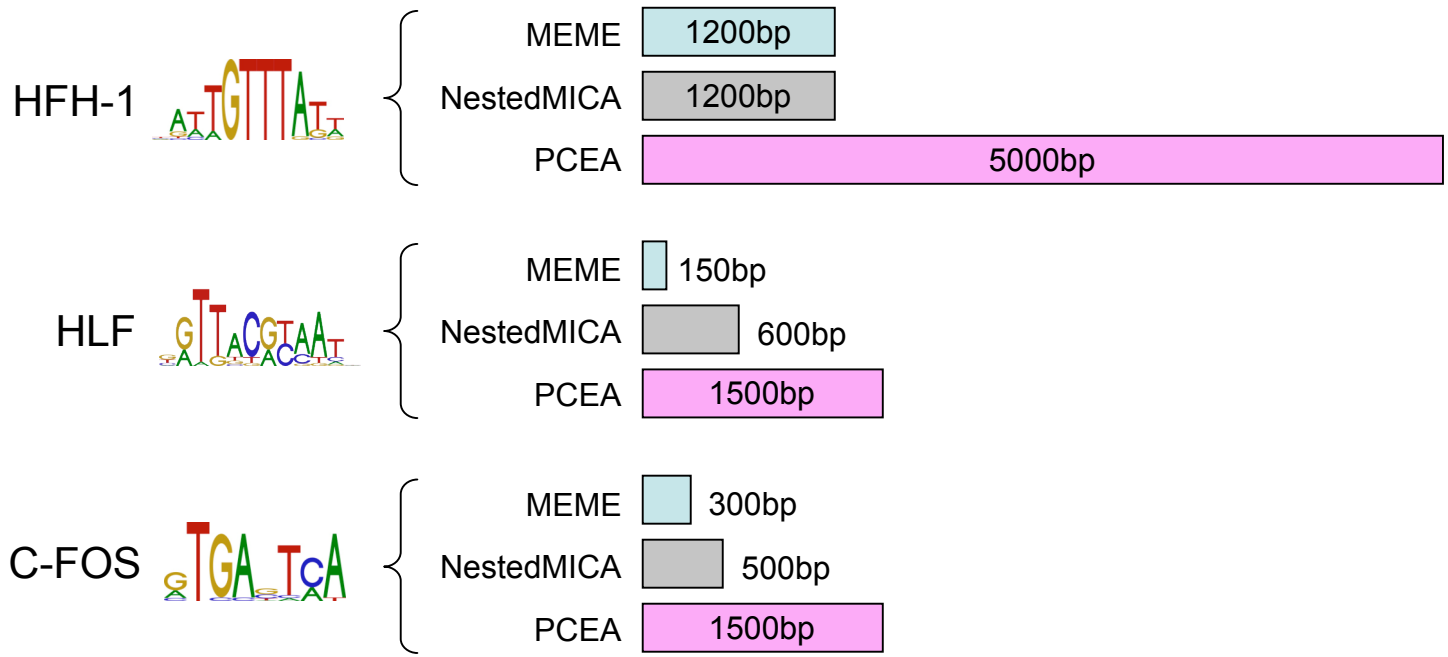
Single Motif

- Comparison with other approaches
  - Maximum sequence length for which motif could be found

Motif	Approach		
	MEME	NestedMICA	PCEA
HFH-1 	1200bp	1200bp	**5000bp**
HLF 	150bp	600bp	**1500bp**
C-FOS 	300bp	500bp	**1500bp**

Single Motif

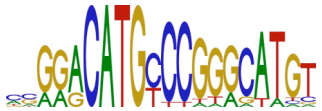
- Comparison with other approaches
  - Maximum sequence length for which motif could be found



Multiple Motifs

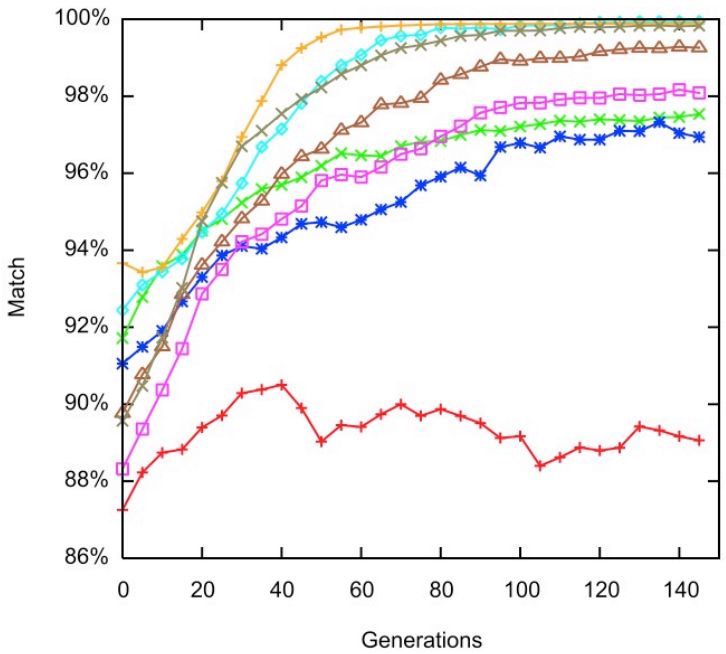
Data set: 100 EPD sequences, length 1000bp, 50 contain one instance of each target motif
Background set: 1500 EPD sequences

Motif		Info. content	Length
SPI-B		9.06	7bp
HLF		11.15	12bp
FOXI1		13.18	12bp
NFKB1		15.63	11bp

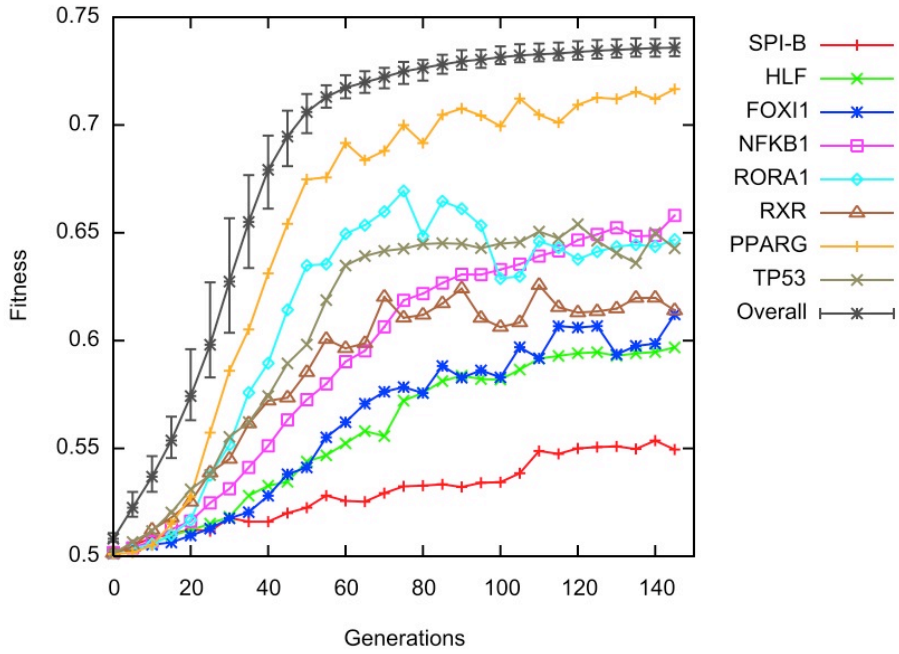
Motif		Info. content	Length
RORA1		17.42	14bp
RXR		20.45	15bp
PPARG		23.45	20bp
TP53		26.24	20bp

Multiple Motifs

Closest match to target



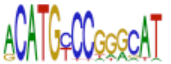
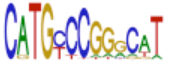
Fitness of closest match



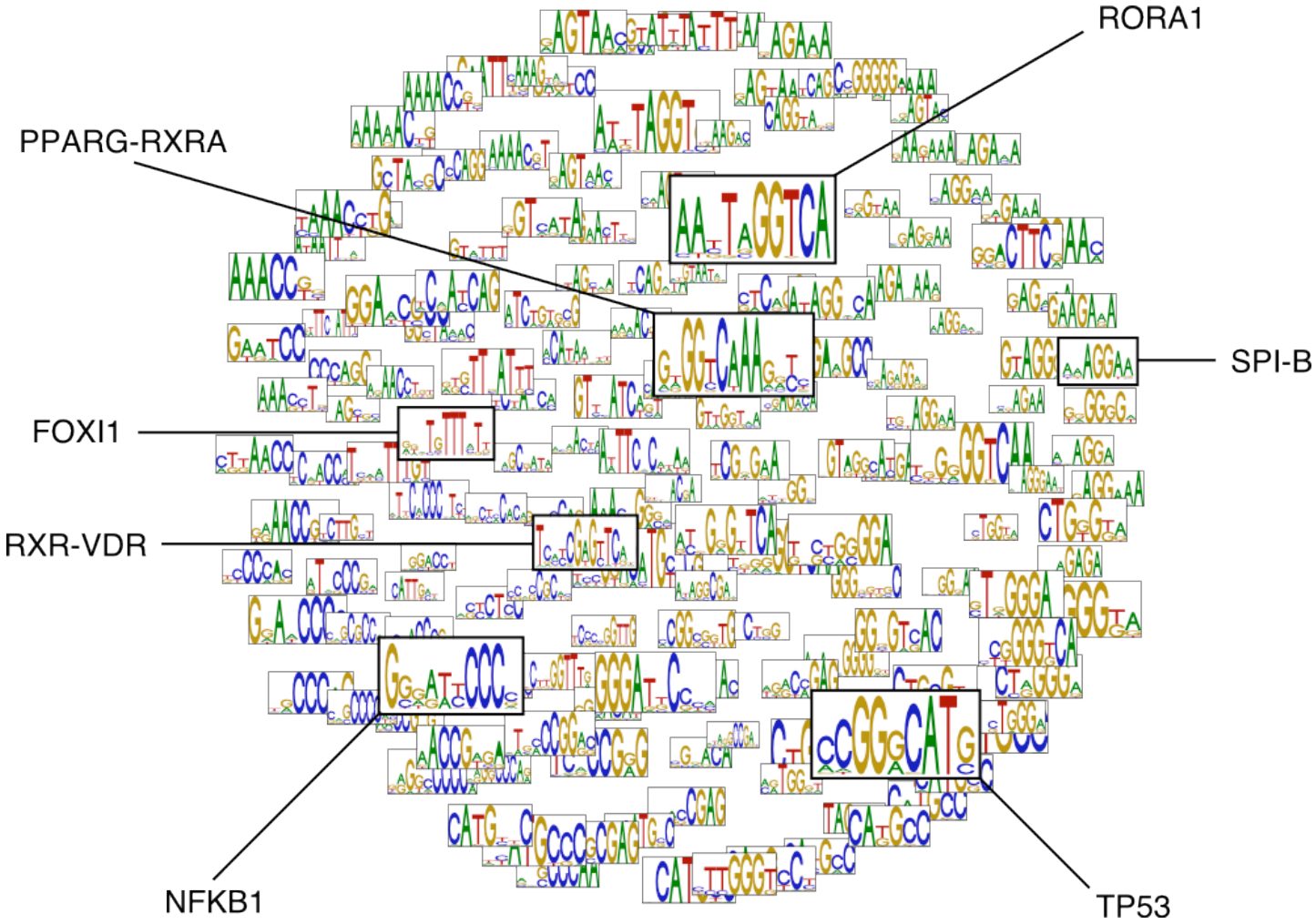
* Data points are the mean of 40 runs of the PCEA

Multiple Motifs

- Evolved motifs in 5 consecutive runs

Run	SPI-B	HLF	FOXI1	NFKB1	RORA1	RXR	PPARG	TP53
								
1								
2								
3								
4								
5								

Multiple Motifs



Muscle Data

- Wasserman and Fickett data set
  - 43 curated promoter sequences from muscle-specific genes
  - Lengths between 197bp and 802bp
  - Muscle expression is relatively well understood
- Test set
  - 28 EPD sequences annotated as muscle-specific
- Background set
  - 2348 non-muscle EPD sequences

Muscle Data

- Best of 5 consecutive runs

#	Sequence Logo	Length	Matches (% of seqs)			Hypothesised TFBS		
			W&F	EPD	bg	Name	ID	Logo
1		10	46.5%	21.4%	3.2%	MEF2	MA0052	
2		10	25.6%	28.6%	5.2%	Myf	MA0055	
3		7	62.8%	57.1%	34.4%	Sp1	M00196	
4		10	16.3%	21.4%	0.9%	SRF	MA0083	
5		9	39.5%	25.0%	7.6%	TEF	MA0090	
6		8	20.9%	17.9%	6.3%	MyoD	M00001	
7		10	20.9%	14.3%	5.1%			
8		11	25.6%	14.3%	1.7%			

Muscle Data

- Best of 5 consecutive runs

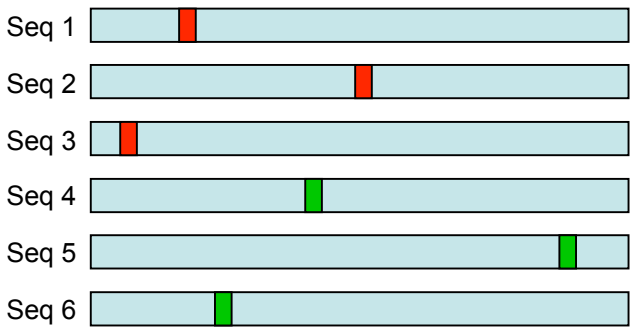
#	Sequence Logo	Length	Matches (% of seqs)			Hypothesised TFBS			20 Runs
			W&F	EPD	bg	Name	ID	Logo	
1		10	46.5%	21.4%	3.2%	MEF2	MA0052		100%
2		10	25.6%	28.6%	5.2%	Myf	MA0055		75%
3		7	62.8%	57.1%	34.4%	Sp1	M00196		100%
4		10	16.3%	21.4%	0.9%	SRF	MA0083		90%
5		9	39.5%	25.0%	7.6%	TEF	MA0090		100%
6		8	20.9%	17.9%	6.3%	MyoD	M00001		95%
7		10	20.9%	14.3%	5.1%				95%
8		11	25.6%	14.3%	1.7%				100%

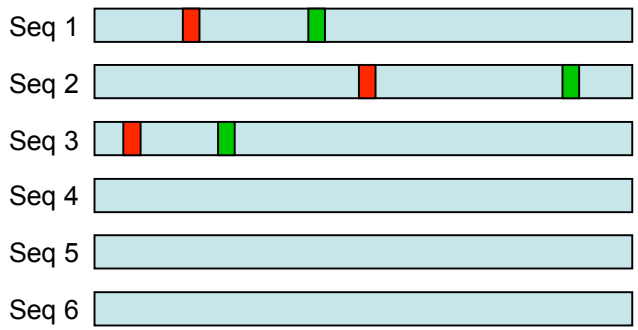
Limitations

- Finding weak motifs in long promoter sequences
  - 1500bp is somewhat less than 10kb
- Finding weak motifs in the presence of strong ones
  - Always room for improvement…
- Finding under-represented motifs
  - Present in only a small part of the data set

Co-Occurrence

- Co-occurring motifs present a stronger signal
  - Could allow identification of weak or under-represented motifs

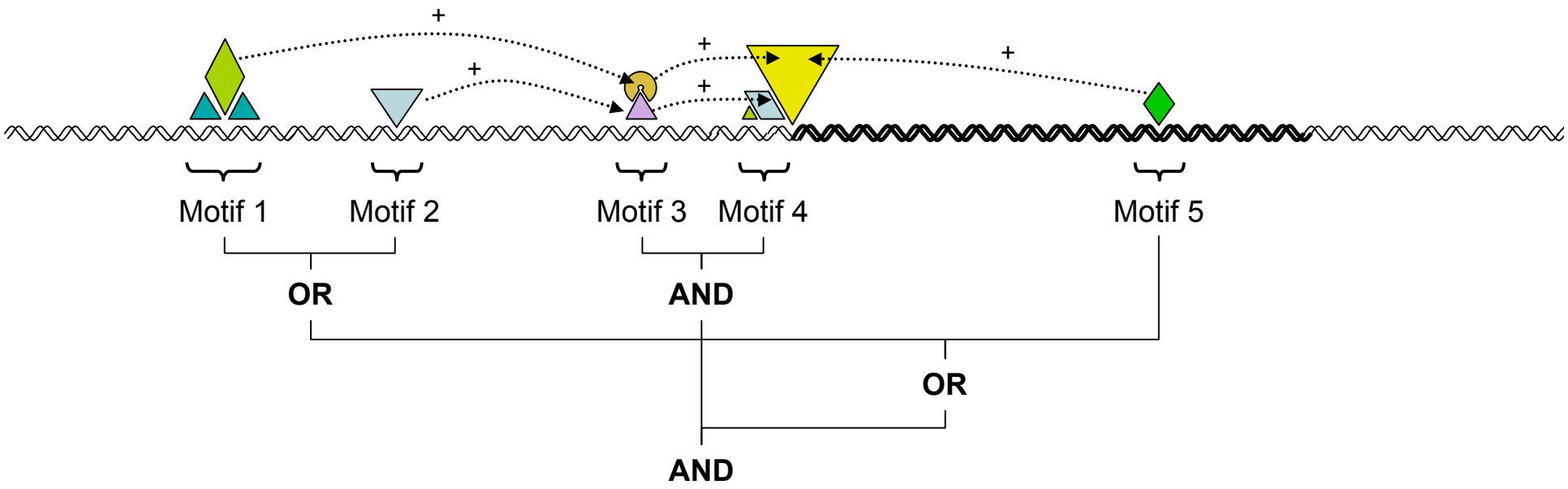




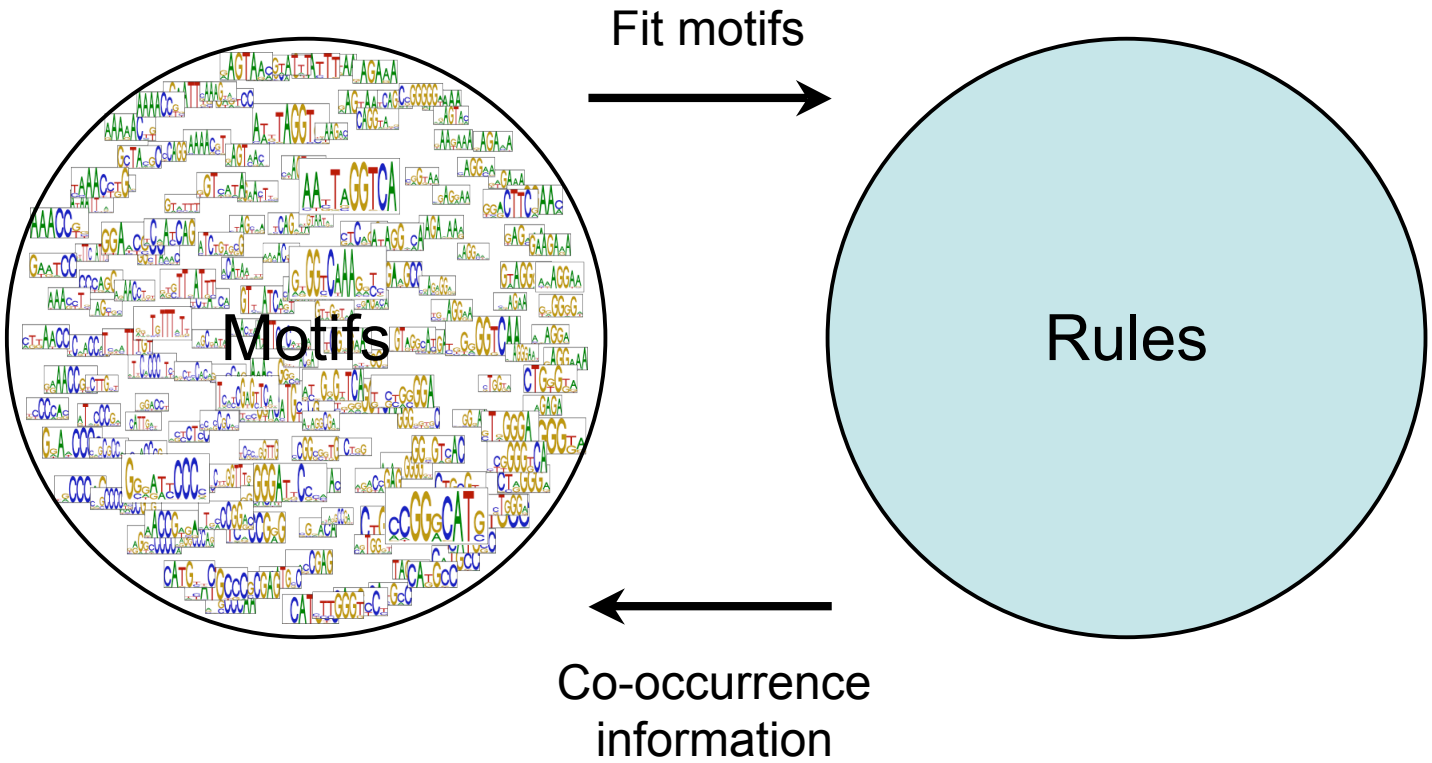
- PCEA discovers motifs concurrently
  - Could use co-occurrence information during search

Higher-Order Motifs

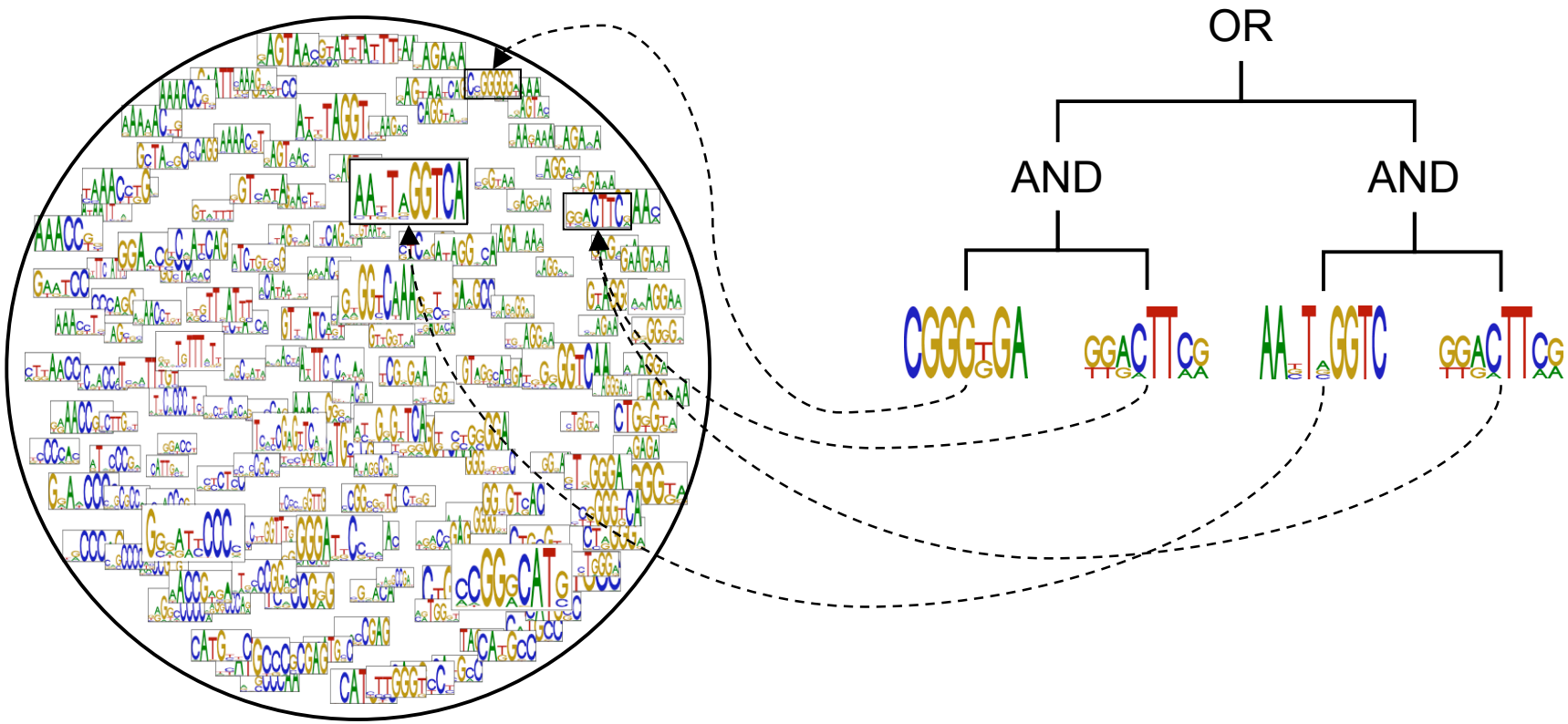
- Capture interactions between TFBSs
  - The ‘rules’ of transcription
  - e.g. using Boolean rules:



Co-operative Co-evolution



Decoupled Interactions



Decoupled Interactions

- Rules *do not* change the fitness of motifs
  - Avoids problem of how to handle unreferenced motifs
- Rules *do* change the breeding privileges of clusters
  - Number of children generated by a cluster decided by:
    - Relative fitness of its fittest motif
    - and how much the motif contributed to rule fitness
- Unreferenced motif clusters produce less children
  - But do still produce children, maintaining diversity

Rule Fitness

- Determined by Matthews correlation:

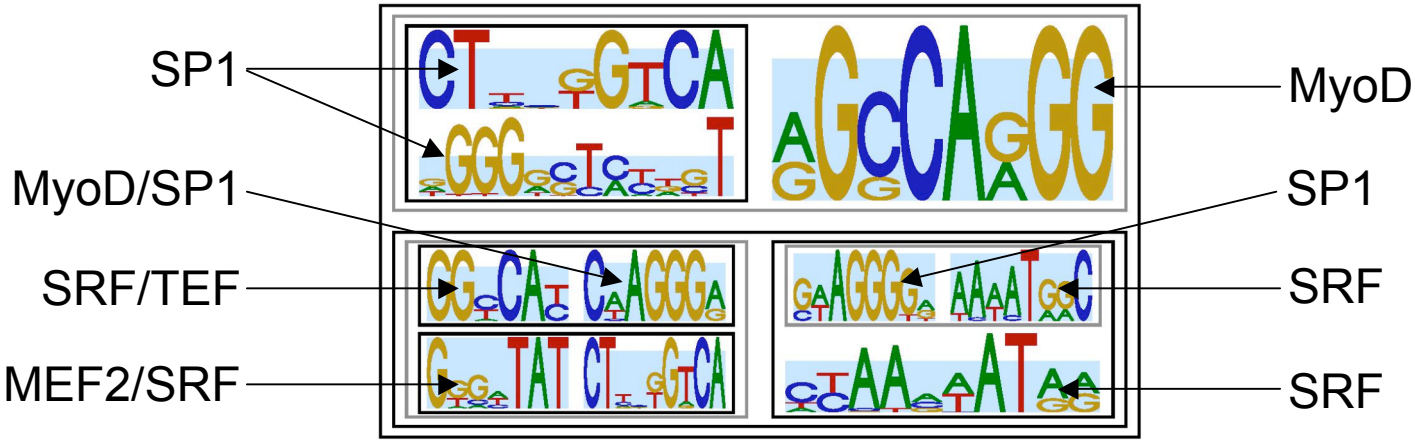
$$MC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

  - Measures classification accuracy
  - Mapped to [0,1]: 1=optimal, 0.5=random classification

- Penalties for:
  - Excessive depth (-0.04/level for depth>5)
  - Lack of motif diversity (max -0.05)

Muscle Data

- Motif population size = 4000; Rule population size = 1000
  - Correctly classifies 19 of 43 positive examples
  - Rejects all but one background sequence

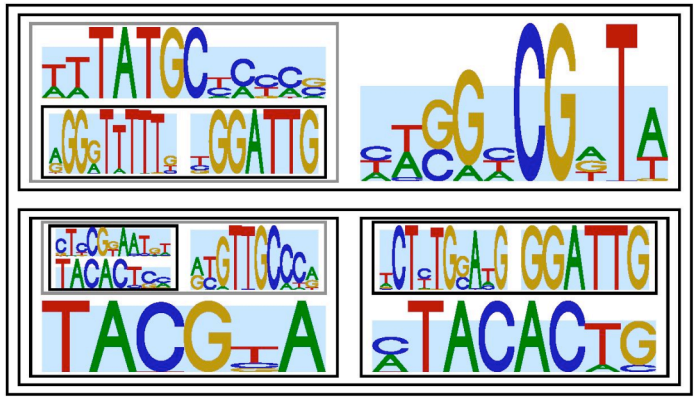


Single Motif

- HLF, sequence length 10,000bp (10kb)
  - 100 sequences in data set, 50 containing motif
  - Background set of 2000 sequences
  - Motif pop = 4000; Rule pop = 4000; 100 generations

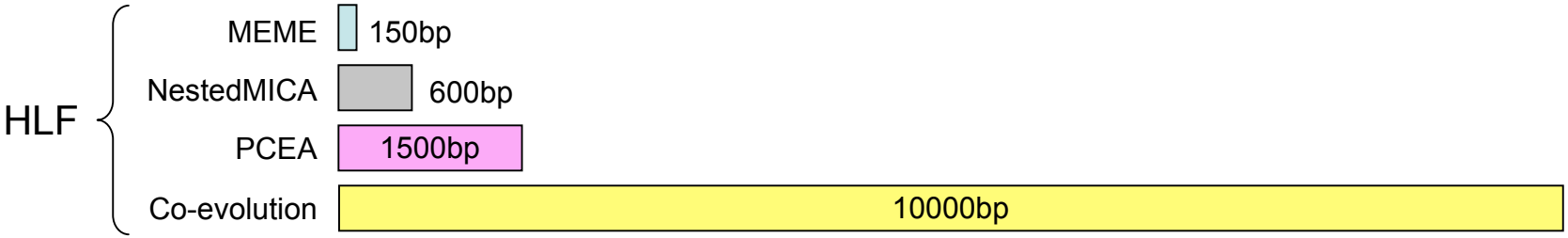
Target: 

Evolved: 



Single Motif

- Somewhat of an improvement…



Future Directions

- Real/useful biological data sets
- Different motif/rule representations
  - Profile HMMs, non-standard representations
- Other problem domains
  - Image processing?
- Other levels of regulation