Predicting DNA-Binding Sites Using Statistical Potentials

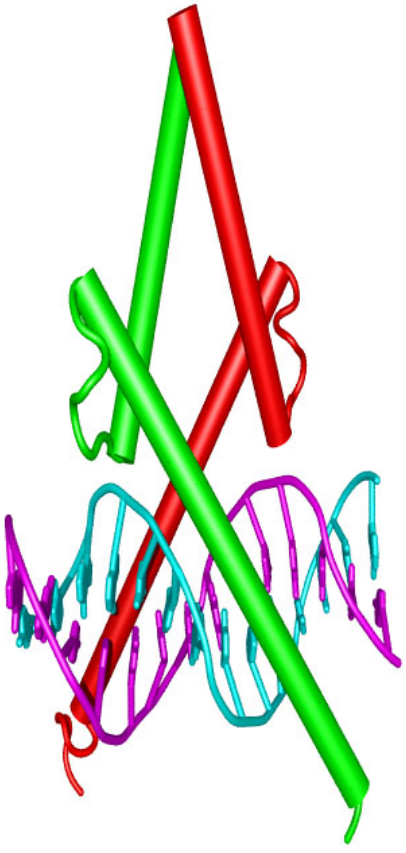
Jan Brücker

Division for Simulation of Biological Systems
Wilhelm Schickard Institute for Computer Science
University of Tübingen

February 21, 2007

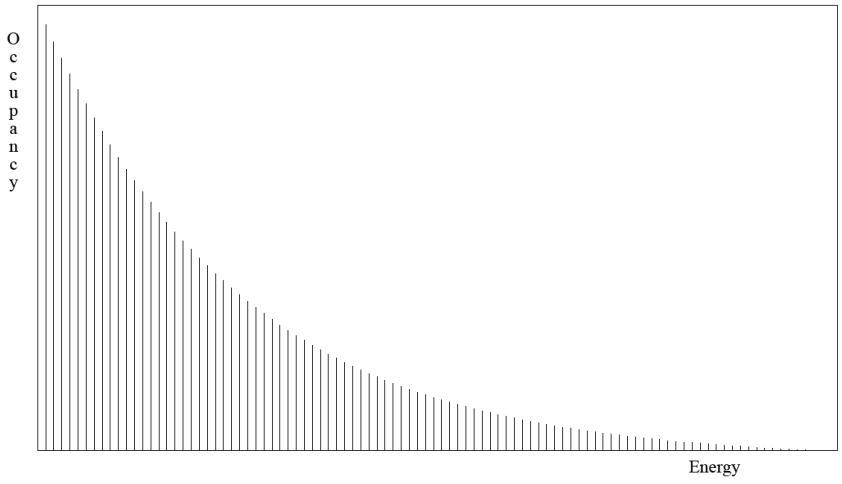
Eberhard Karls
Universität
Tübingen

Assumptions and Aims



- Assumptions:
  - The preference of a transcription factor to a certain binding site can be found in the preference of certain amino acids to certain nucleotides.
  - Complexes with their native DNA binding sites show low free energies.
  - The Boltzmann distribution relates the preference of certain interactions to the binding free energy.
- Plan:
  - Use protein-DNA co-crystal structures to detect preferred amino acid-nucleotide interactions.
  - Use Boltzmann's distribution to calculate free energies from interaction frequencies.

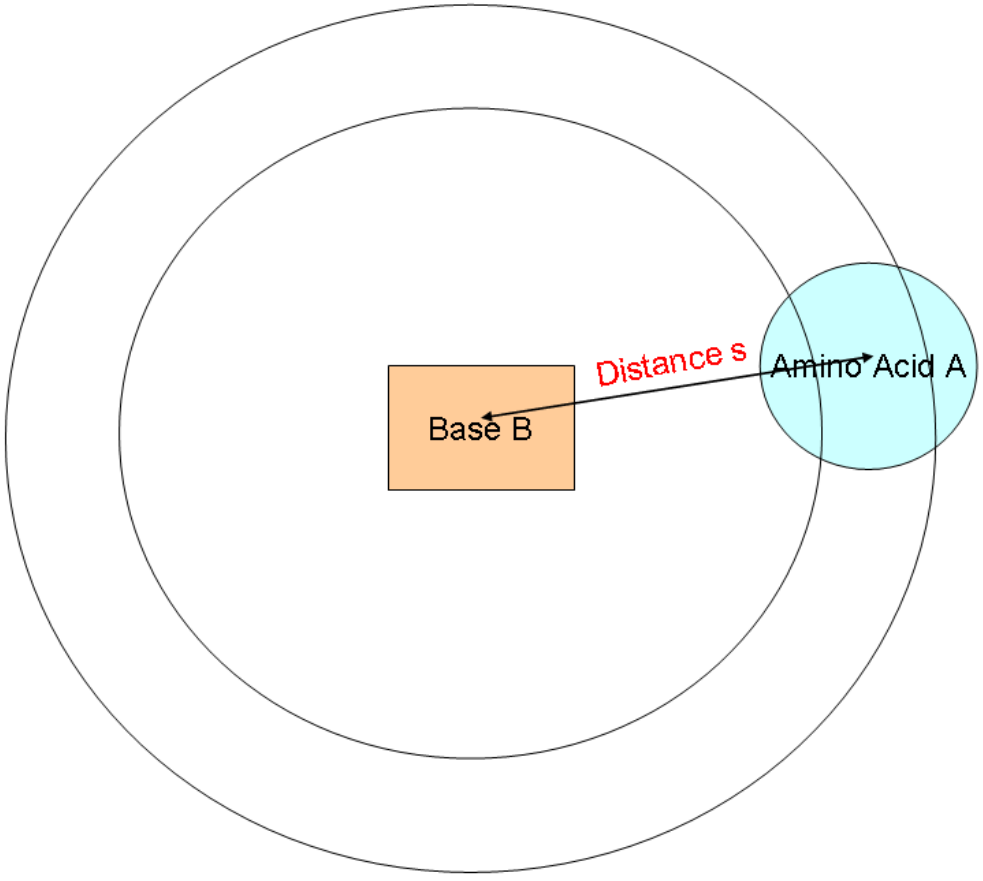
Inverse Boltzmann



$$f(x) = \frac{1}{Z} \cdot exp\left[-\frac{E(x)}{kT}\right]$$

- The Boltzmann distribution assigns occupancy values to given energy states.
- Given a distribution the binding energy of a molecule can be calculated. (Inverse Boltzmann)

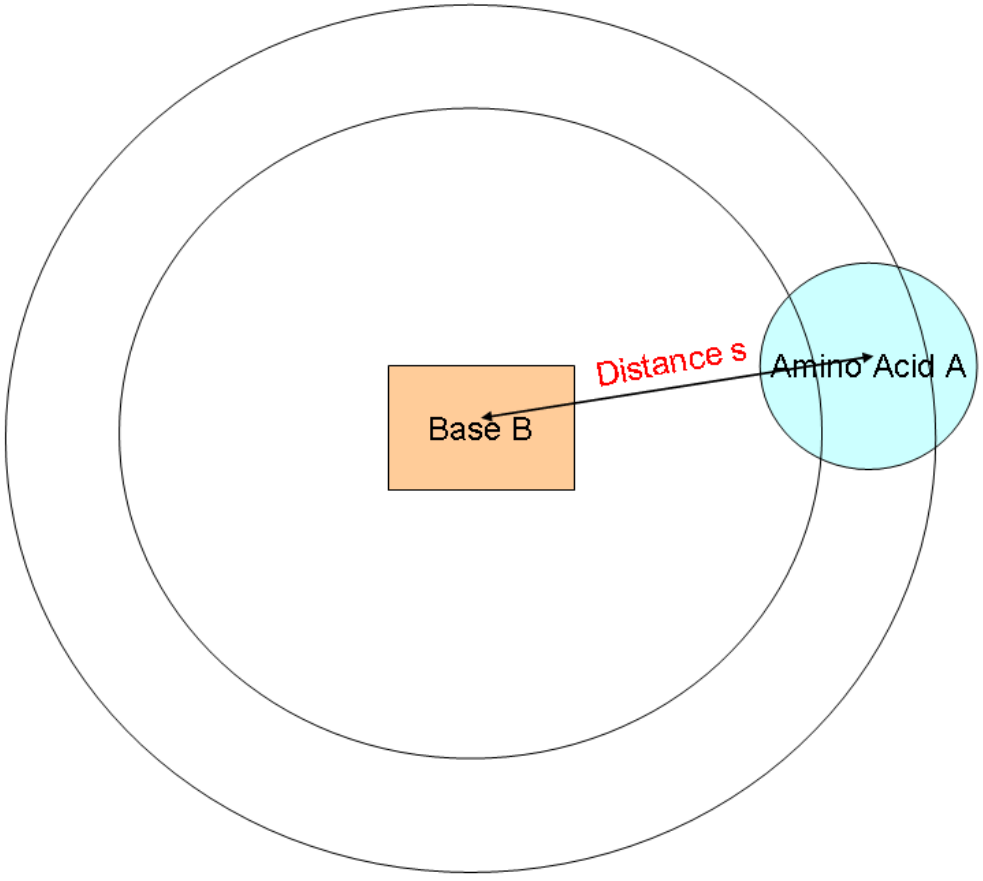
Defining Energy States
Distance-Based Clustering



- Defining energy states is a question
  - ... of the available amount of data.
  - ... of the resolution of the available data.
  - ... on what kind of data one wants to use the potentials.
  - ... of taste, innovation, phantasy and the quality of the prediction.

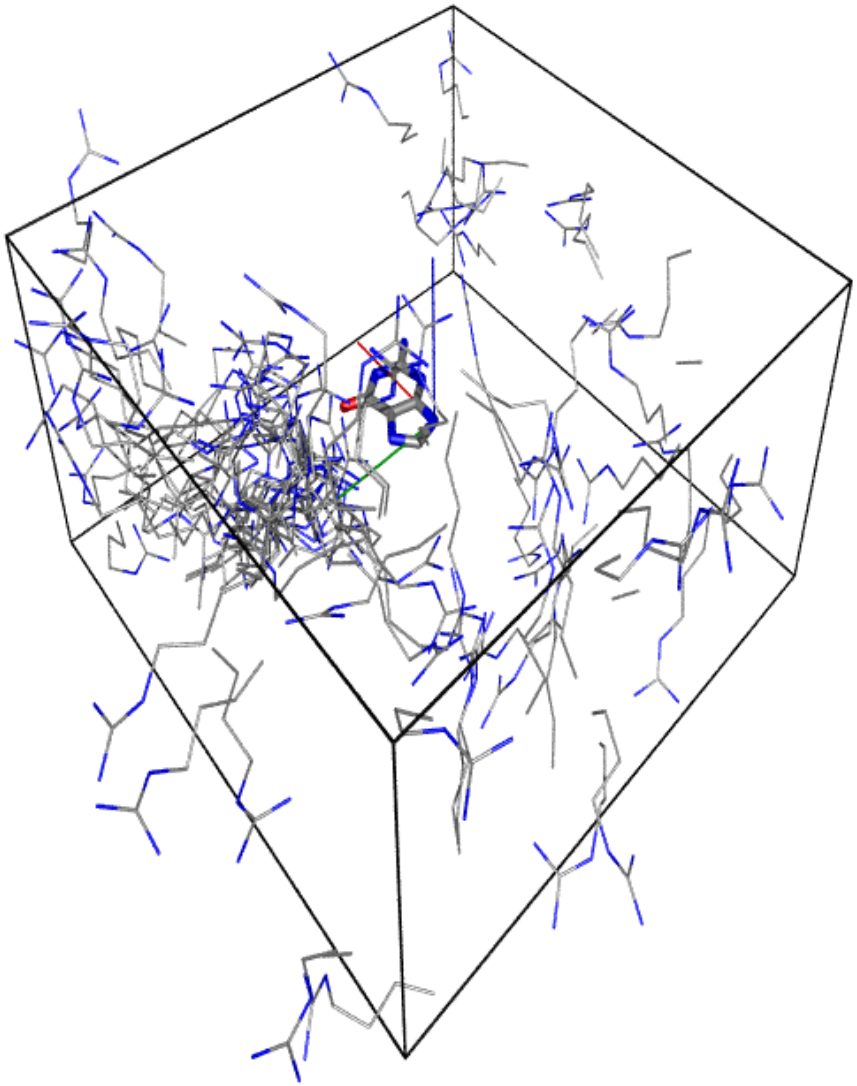
Defining Energy States

Distance-Based Clustering



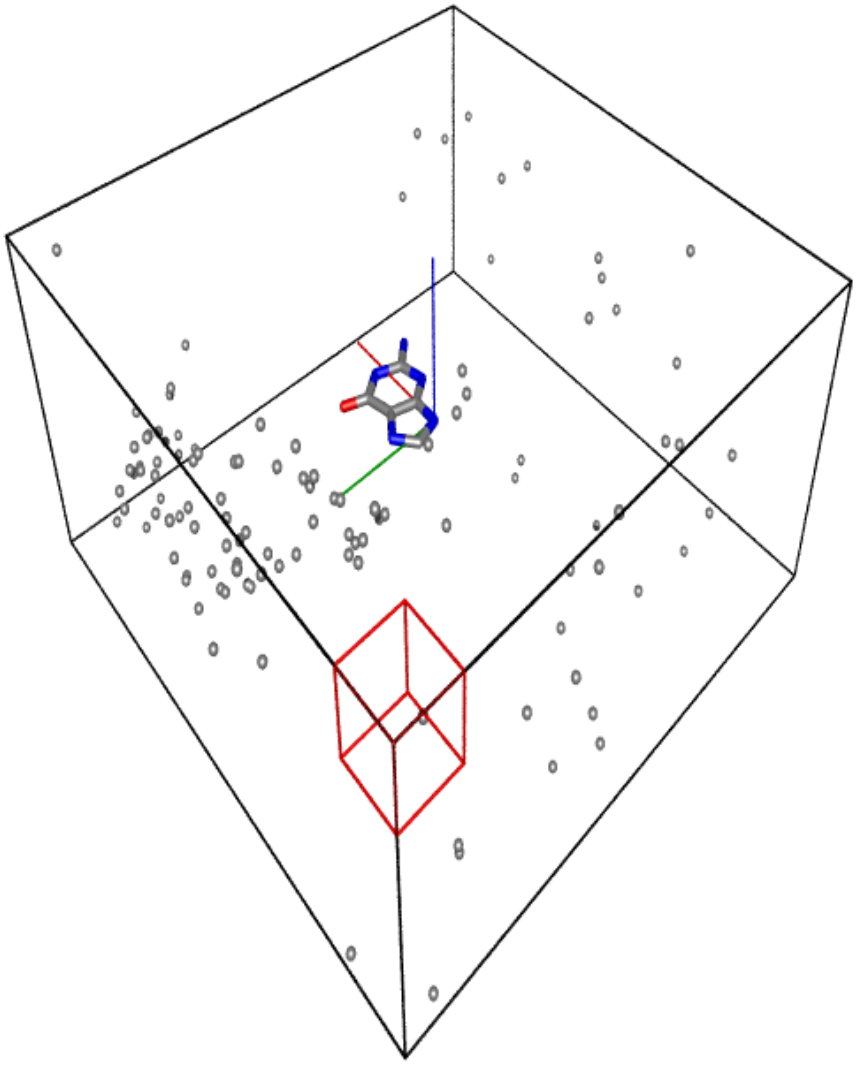
- Defining energy states is a question
  - ... of the available amount of data.
  - ... of the resolution of the available data.
  - ... on what kind of data one wants to use the potentials.
  - ... of taste, innovation, phantasy and the quality of the prediction.

Defining Energy States
Distance-Based Clustering



- Defining energy states is a question
  - ... of the available amount of data.
  - ... of the resolution of the available data.
  - ... on what kind of data one wants to use the potentials.
  - ... of taste, innovation, phantasy and the quality of the prediction.

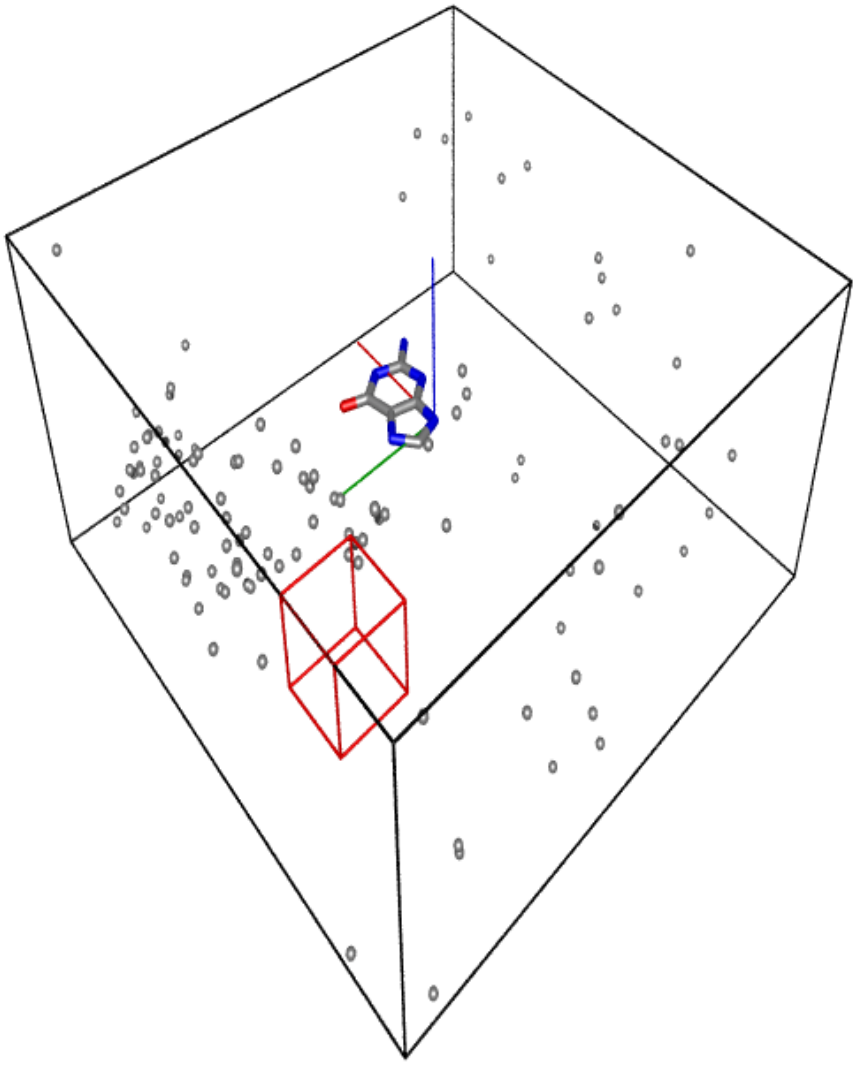
Defining Energy States
Distance-Based Clustering



- Defining energy states is a question
  - ... of the available amount of data.
  - ... of the resolution of the available data.
  - ... on what kind of data one wants to use the potentials.
  - ... of taste, innovation, phantasy and the quality of the prediction.

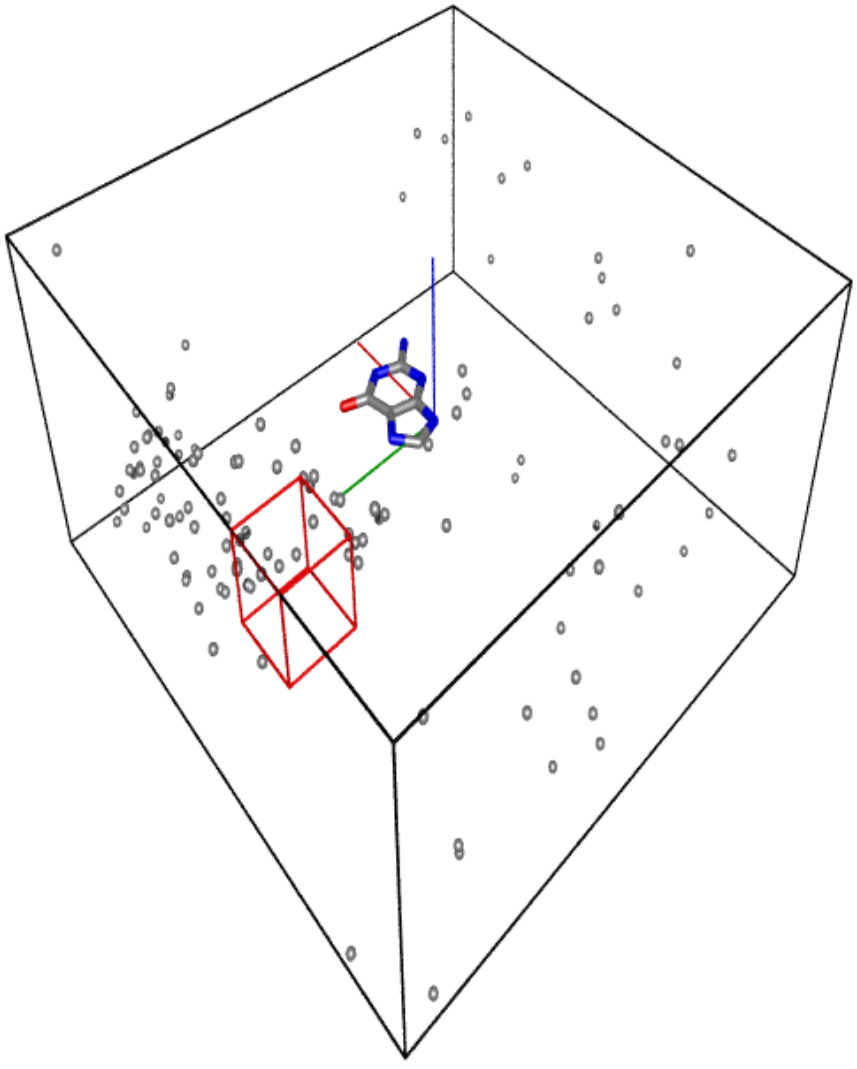
Defining Energy States

Distance-Based Clustering



- Defining energy states is a question
  - ... of the available amount of data.
  - ... of the resolution of the available data.
  - ... on what kind of data one wants to use the potentials.
  - ... of taste, innovation, phantasy and the quality of the prediction.

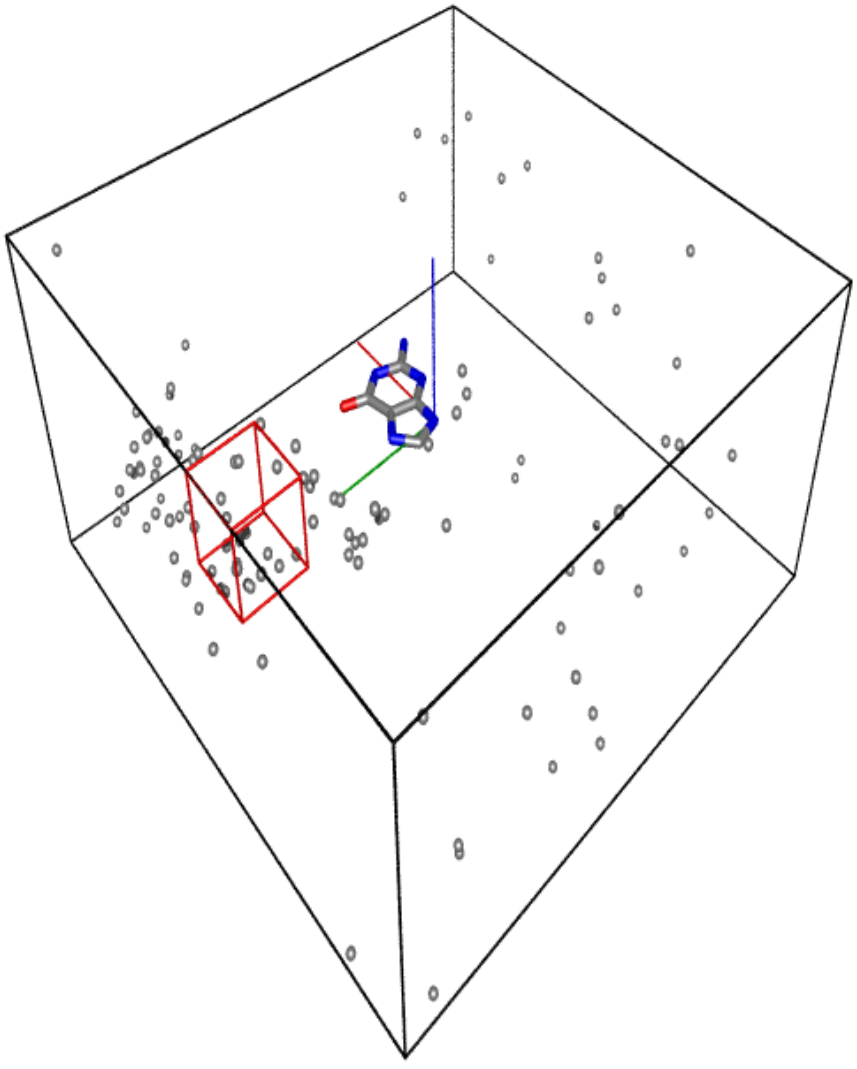
Defining Energy States
Distance-Based Clustering



- Defining energy states is a question
  - ... of the available amount of data.
  - ... of the resolution of the available data.
  - ... on what kind of data one wants to use the potentials.
  - ... of taste, innovation, phantasy and the quality of the prediction.

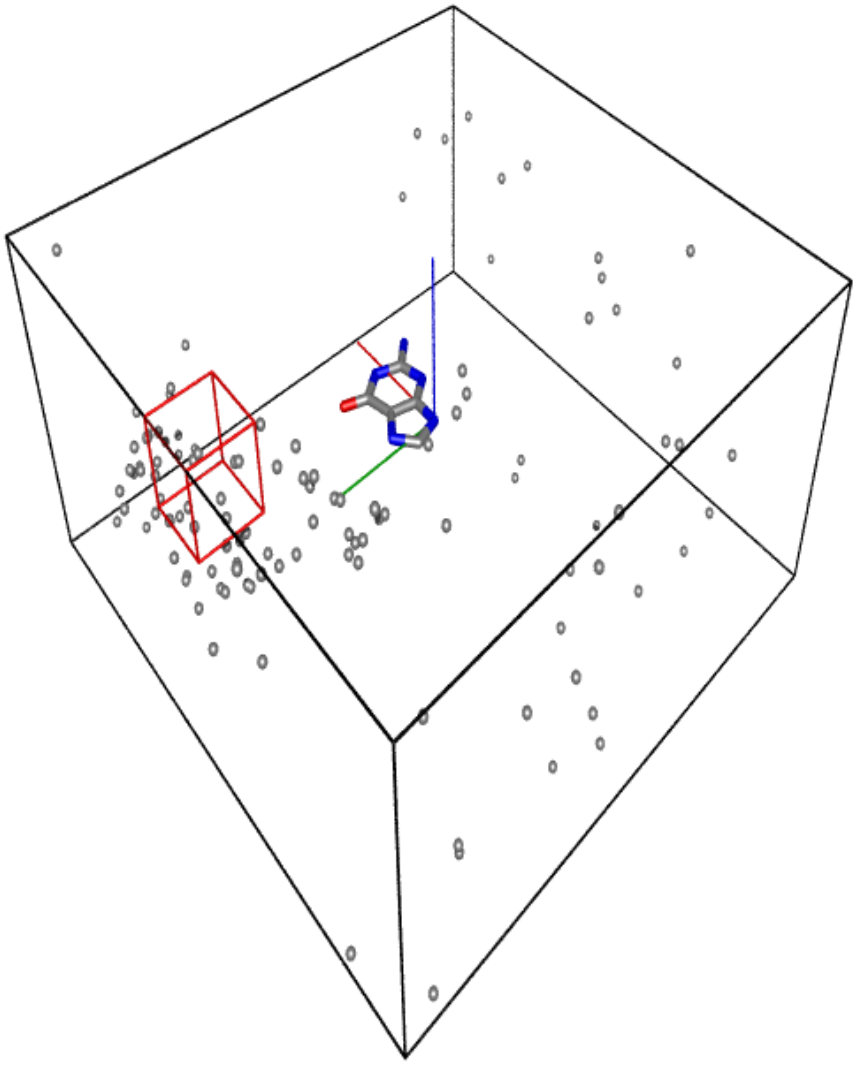
Defining Energy States

Distance-Based Clustering



- Defining energy states is a question
  - ... of the available amount of data.
  - ... of the resolution of the available data.
  - ... on what kind of data one wants to use the potentials.
  - ... of taste, innovation, phantasy and the quality of the prediction.

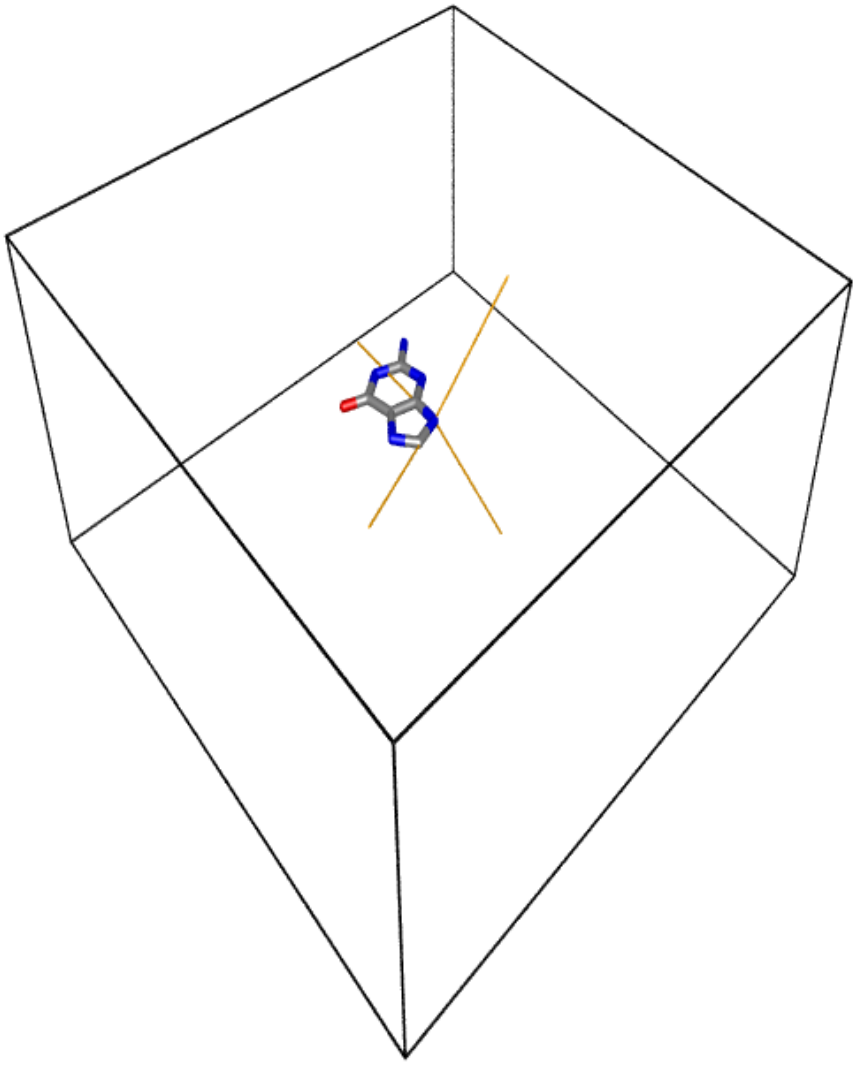
Defining Energy States
Distance-Based Clustering



- Defining energy states is a question
  - ... of the available amount of data.
  - ... of the resolution of the available data.
  - ... on what kind of data one wants to use the potentials.
  - ... of taste, innovation, phantasy and the quality of the prediction.

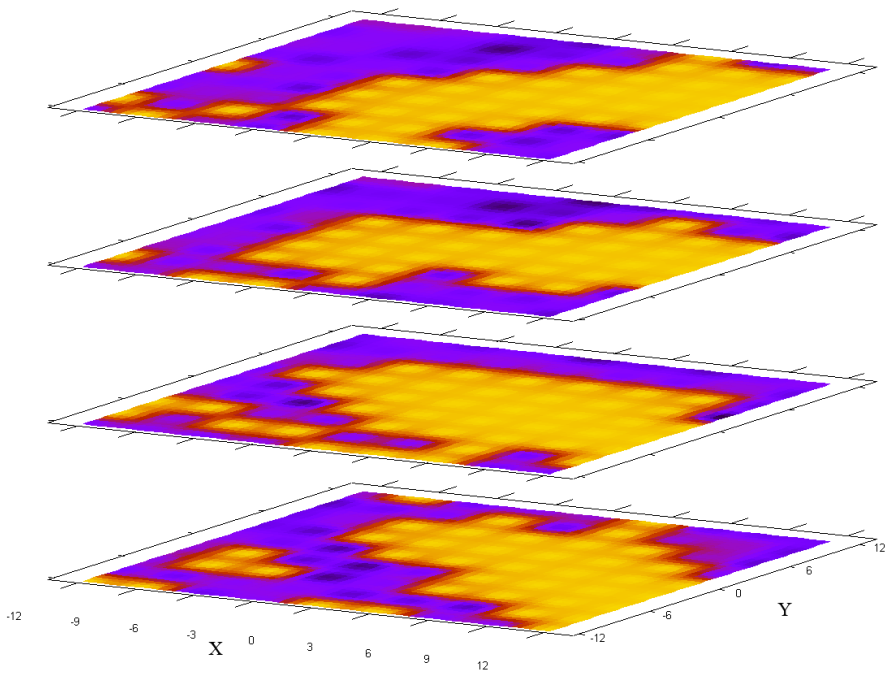
Defining Energy States

Distance-Based Clustering



- Defining energy states is a question
  - ... of the available amount of data.
  - ... of the resolution of the available data.
  - ... on what kind of data one wants to use the potentials.
  - ... of taste, innovation, phantasy and the quality of the prediction.

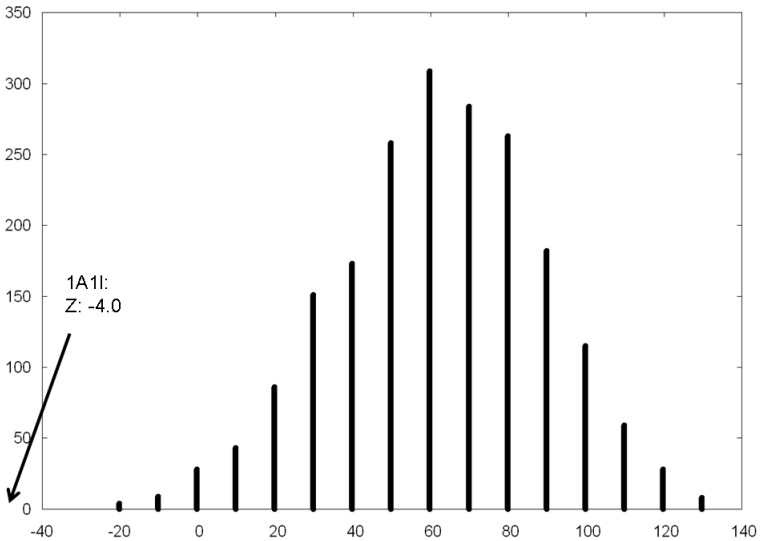
Defining Energy States
Distance-Based Clustering



- Defining energy states is a question
  - ... of the available amount of data.
  - ... of the resolution of the available data.
  - ... on what kind of data one wants to use the potentials.
  - ... of taste, innovation, phantasy and the quality of the prediction.

Z-Scores

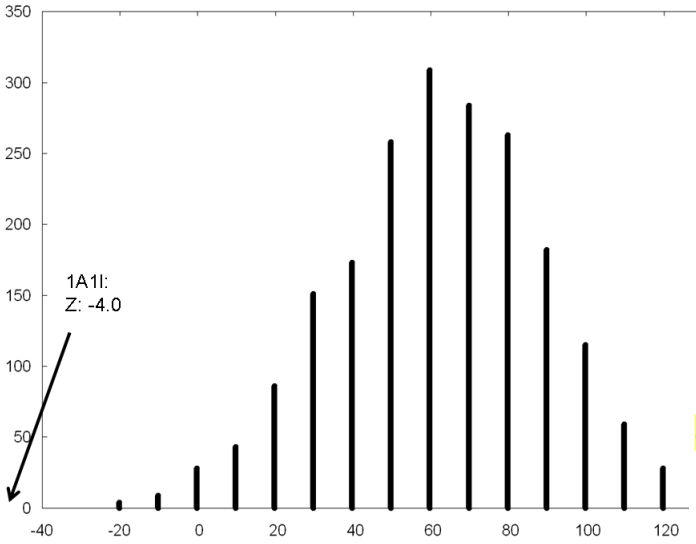
Distinguishing True Binding Sites From Random Sequences

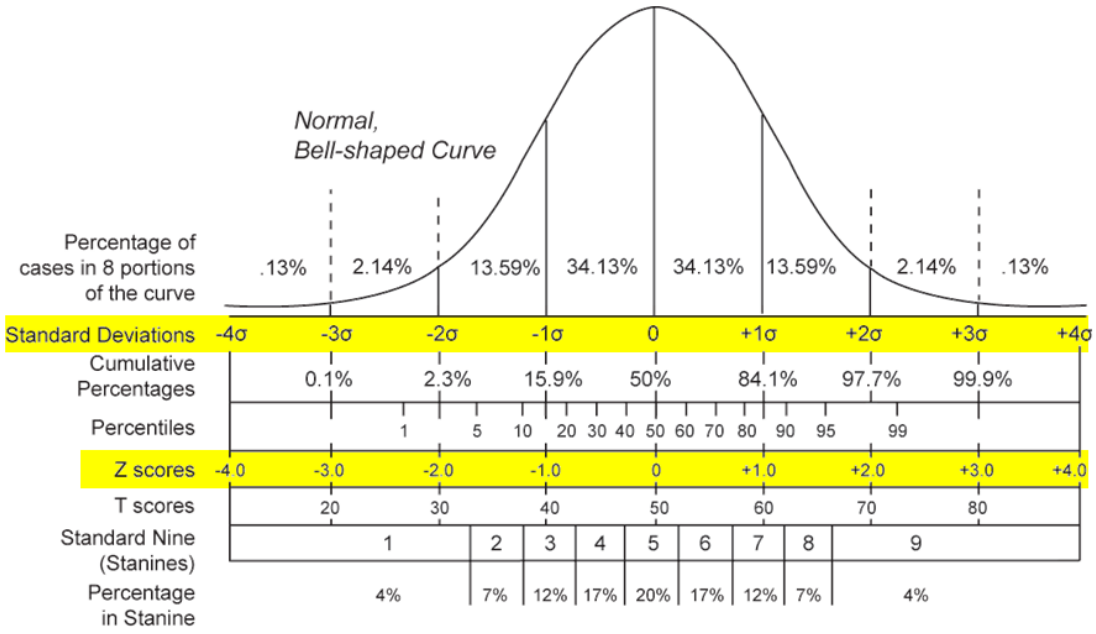


$$Z = \frac{X - m}{\sigma}$$

Z-Scores

Distinguishing True Binding Sites From Random Sequences

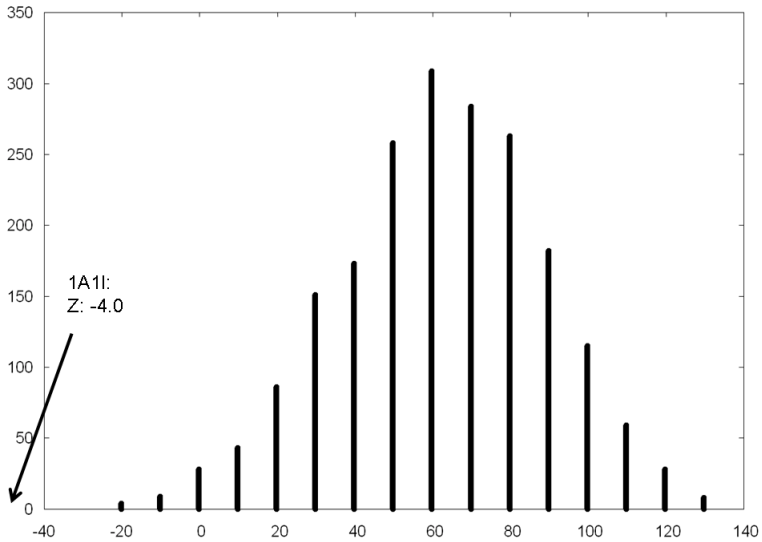




$$Z = \frac{X - m}{\sigma}$$

Z-Scores

Distinguishing True Binding Sites From Random Sequences

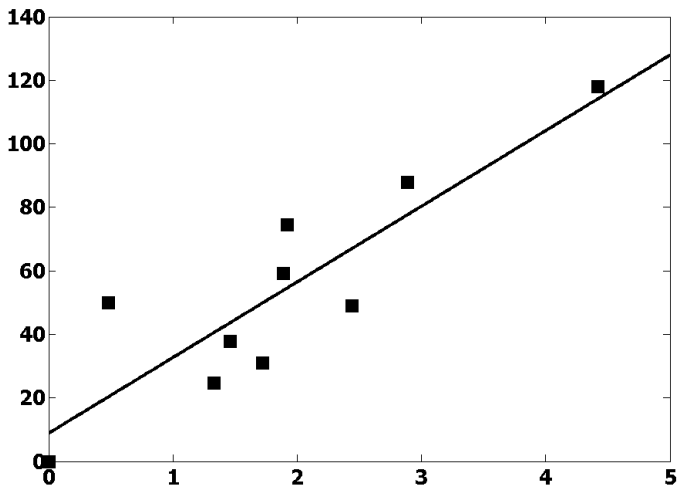


Approach	μ Z-score	Chance
C_α-approach	-2.42	0.0078
orientation-vector	-2.64	0.0041
Interpolation	-2.71	0.0034

$$Z = \frac{X - m}{\sigma}$$

Comparing With Experimental-Data

DNA-Microarray Data of mutants of the *Zif268* Zinc Finger



	Bulyk Data	orientation-method
Mutant	$\Delta\Delta G$	$\Delta\Delta G$
TGG	0	0
TAG	0.48	24.08
GGG	1.33	9.08
CGG	1.46	13.33
AGG	1.72	12.82
TTG	1.89	28.61
GAG	1.92	33.15
TCG	2.44	21.33
CAG	2.89	37.40
AAA	4.42	57.18
Correlation with Bulyk		0.85

From Potentials to Sequence Logos

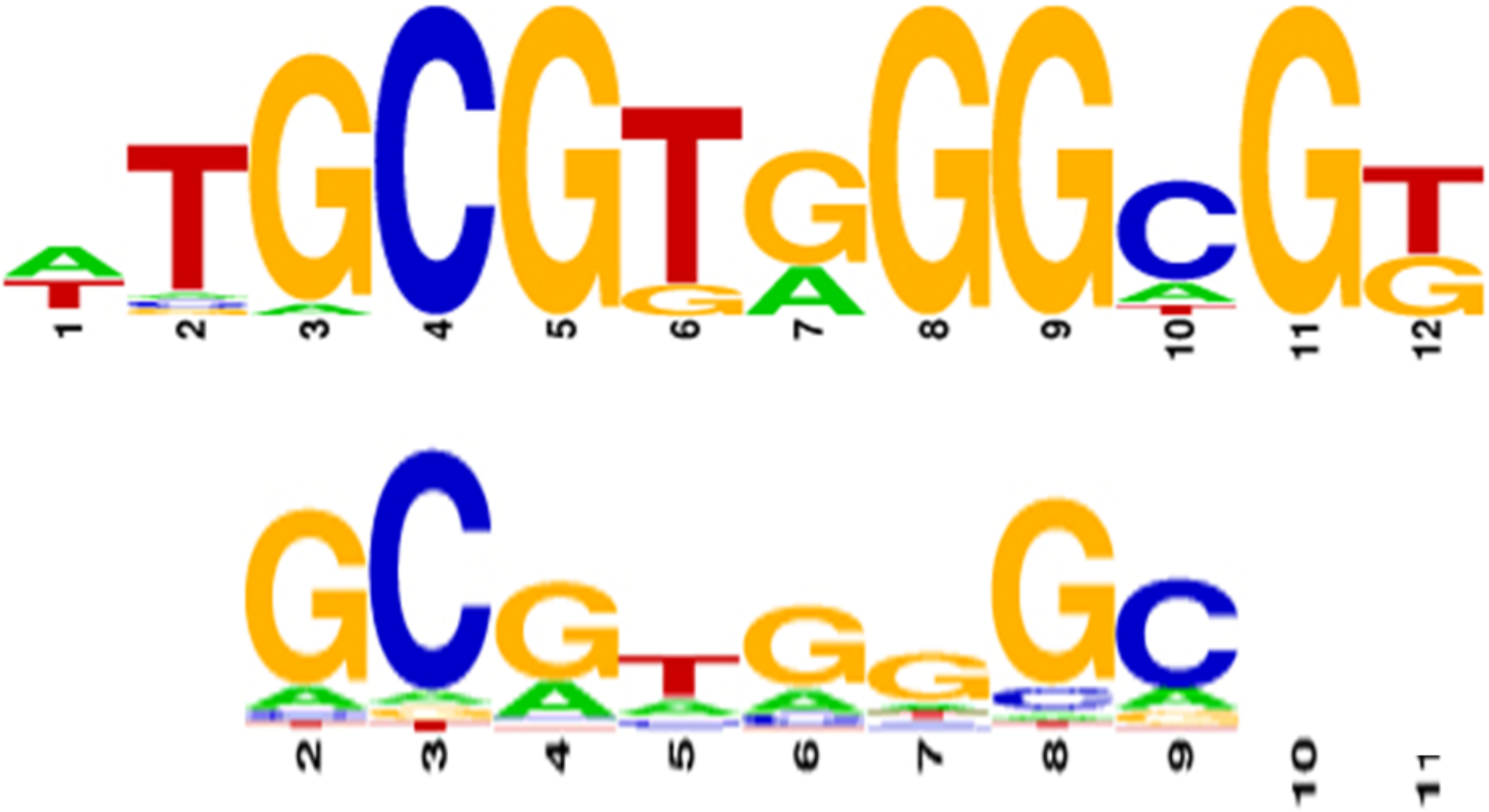
- Make again usage of the (forward) Boltzmann distribution.
- Calculate for one transcription factor for each of the binding site positions the probability of finding a specific base at that position.

$$f(x) = \frac{exp\left[-\frac{E(x)}{RT}\right]}{\sum_{s\in\{A,C,G,T\}} exp\left[-\frac{E(s)}{RT}\right]} \quad for \quad x \in \{A, C, G, T\}$$

- The result is a position frequency matrix, commonly illustrated as a sequence motif.

Sequence Motifs

From Energies Back to Frequencies



Thanks

Thanks for your attention.