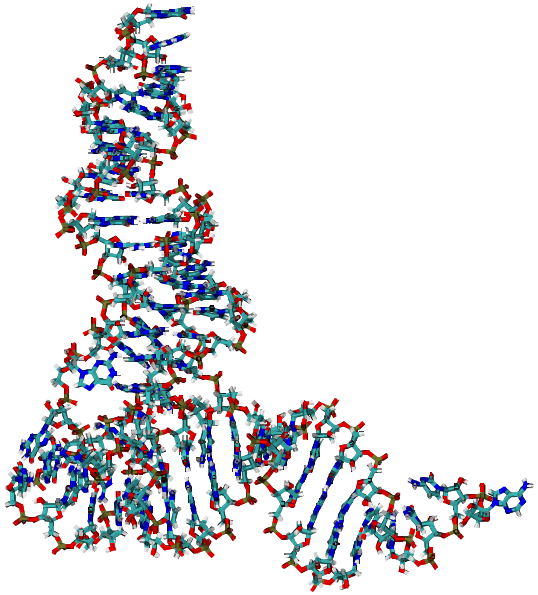
How to Compute the Energy Landscape of Biopolymers

Michael Wolfinger

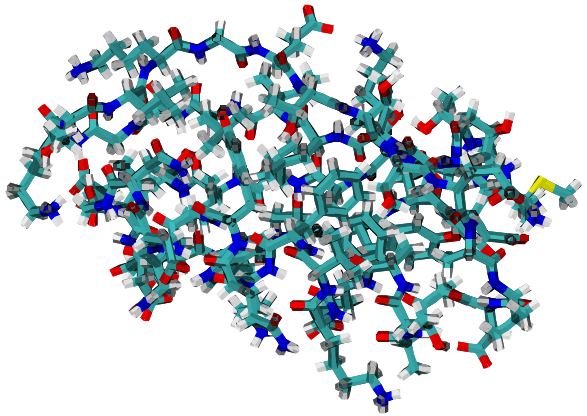
Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien

Bled, Feb 2002

Biopolymers



RNA



Proteins

RNA Secondary Structures

A secondary structure is a list of base pairs that fulfills two constraints:

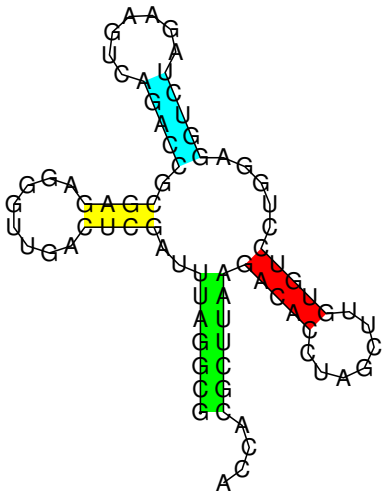
- A base may participate in at most one base pair.
- Base pairs must not cross, i.e., no two pairs (i, j) and (k, l) may have $i < k < j < l$. (no pseudo-knots)

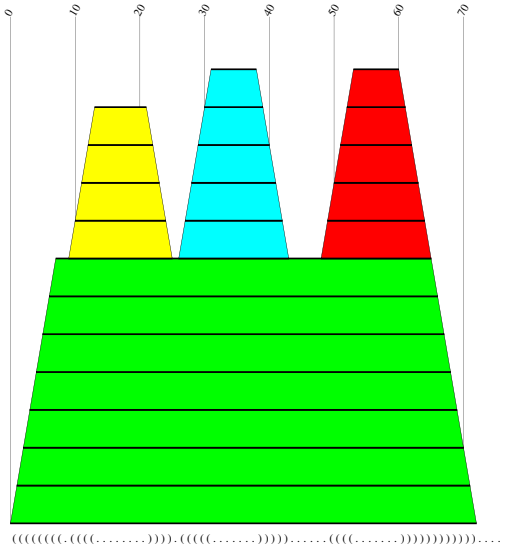
The number of secondary structures as well as the maximum number of base pairs can be computed recursively

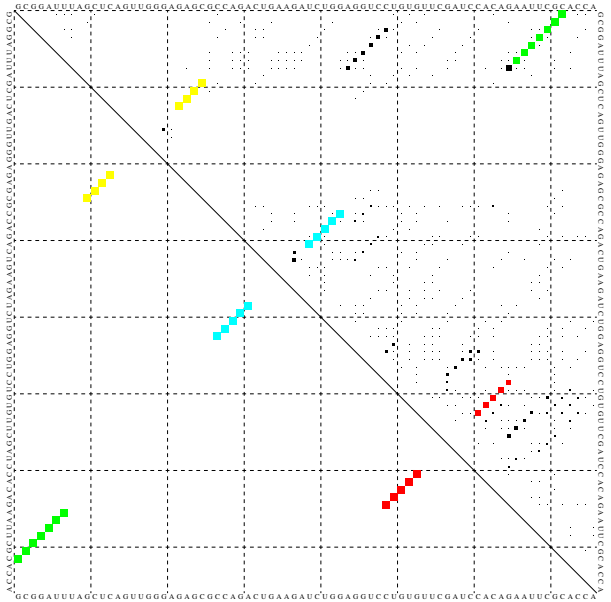
For sequences with equal A U G C content the number of conformations grows as

$$\bar{S}_{1n} \sim n^{-\frac{3}{2}} 1.85^n$$

Representation of RNA Secondary Structures







(((((((..((((.....))))..(((((.....)))))).....(((((.....)))))..(((((((....

Energy Landscapes

To explore the energy landscape of Biolopymers it is necessary to know the Conformation Space.
Rquirements:

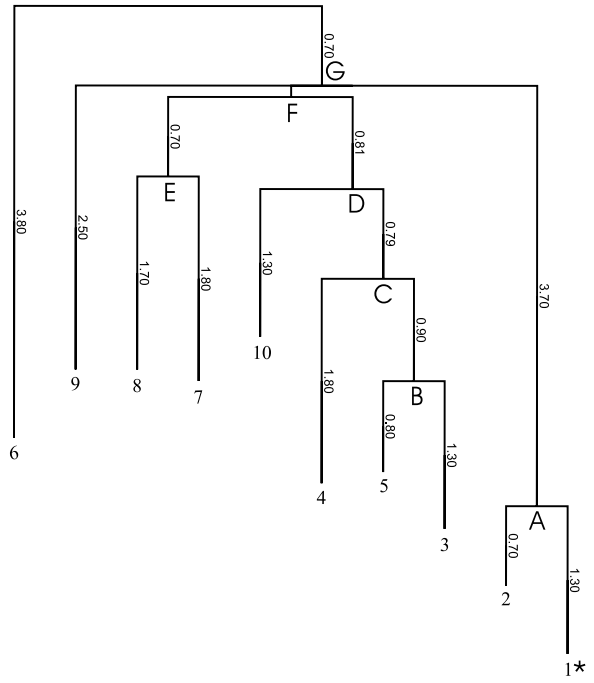
- All suboptimal structures up to a pre-defined energy level
- A definition of neighborhood among different structures: Move Set

A Move Set defines a metric on the conformation space, i.e. it defines neighborhood.

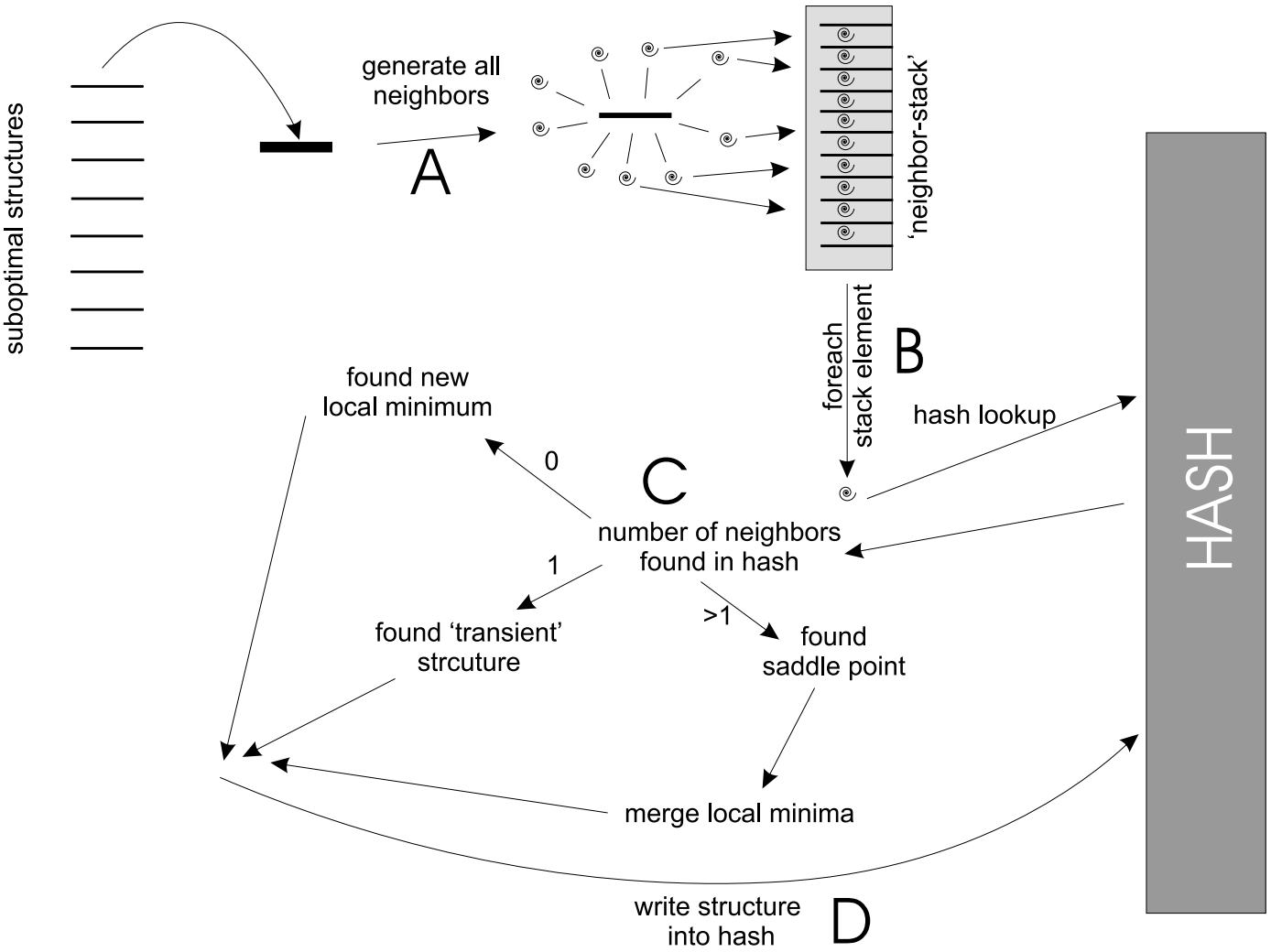
Energy Barriers and Barrier Trees

Some topological definitions: A structure is a

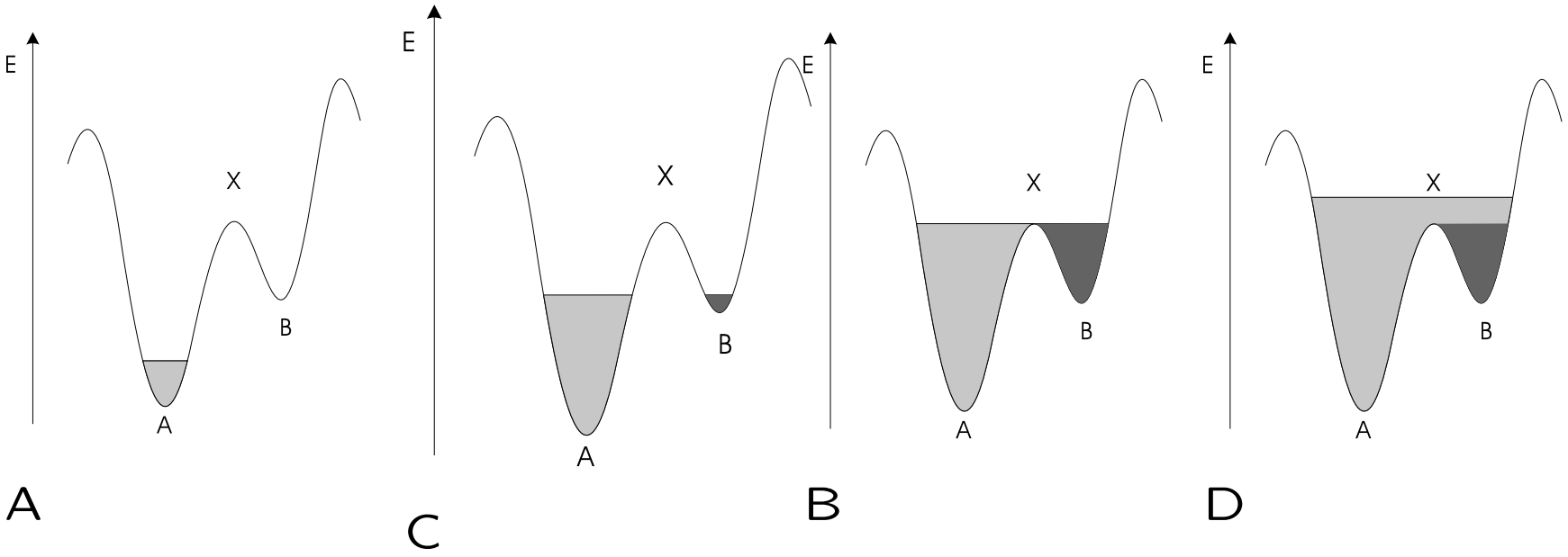
- *local minimum* if its energy is lower than the energy of **all** neighbors
- *local maximum* if its energy is higher than the energy of **all** neighbors
- *saddle point* if there are at least two local minima thar can be reached by a downhill walk starting at this point



The algorithm of barriers



The flooding algorithm



Information that can be calculated

- Local minima
- Saddle points
- Barrier heights
- Gradient basins
- Partition functions and free energies of (gradient) basins

N.B.: A *gradient basin* is the set of all initial points from which a gradient walk (steepest descent) ends in the same local minimum.

Application: Barrier tree kinetics

Question:

Given an initial population distribution, how does the system evolve in time? (What is the population distribution after n time-steps?)

Answer:

Solve the fundamental equation $\frac{d}{dt}P_t = \mathbf{U}P_t$

Explicit solution: $P_t = e^{t\mathbf{U}}P_0$

U contains elements of the form $k_{ij} = \Gamma e^{-\beta(E_S - E_i)}$ (rate $i \rightarrow j$)

We used a *Markov Process* to do the simulation

Stochastic Processes: Markov Processes

Stochastic process $\{X_t | t \in T\}$

X_t describes a snapshot random distribution on a state space $\mathcal{S}$ at time t

A Markov Processs is a stochastic process that satisfies an additional requirement.

This *Markov property* requires that, for any given time instant (say t_n) the future behavior, for instance the value of $X_{t_{n+1}}$, is totally independent of its history, i.e. the values of $X_{t_{n-1}}$, $X_{t_{n-2}}$ and so on. It only depends on the state occupied at the **current** time instant t_n, given by the value of X_{t_n}

Markov Chains

A *continuous time Markov chain* is a Markov process with

- discrete state space
- continuous time range

let $t_n + \Delta t > t_n > t_{n-1} > t_{n-2} > ... > t_0$:

$$\begin{aligned} & \text{Prob}\{X_{t_n+\Delta t} = P' | X_{t_n} = P, X_{t_{n-1}} = P_{t_{n-1}}, ..., X_{t_0} = P_{t_0}\} \\ = \ & \text{Prob}\{X_{t_n+\Delta t} = P' | X_{t_n} = P\} \\ = \ & \text{Prob}\{X_{\Delta t} = P' | X_0 = P\} \end{aligned}$$

If we substitute P with i and P' with j then the last expression can be rewritten as

$$\text{Prob}\{X_{\Delta t} = j | X_0 = i\} = p_{ij}$$

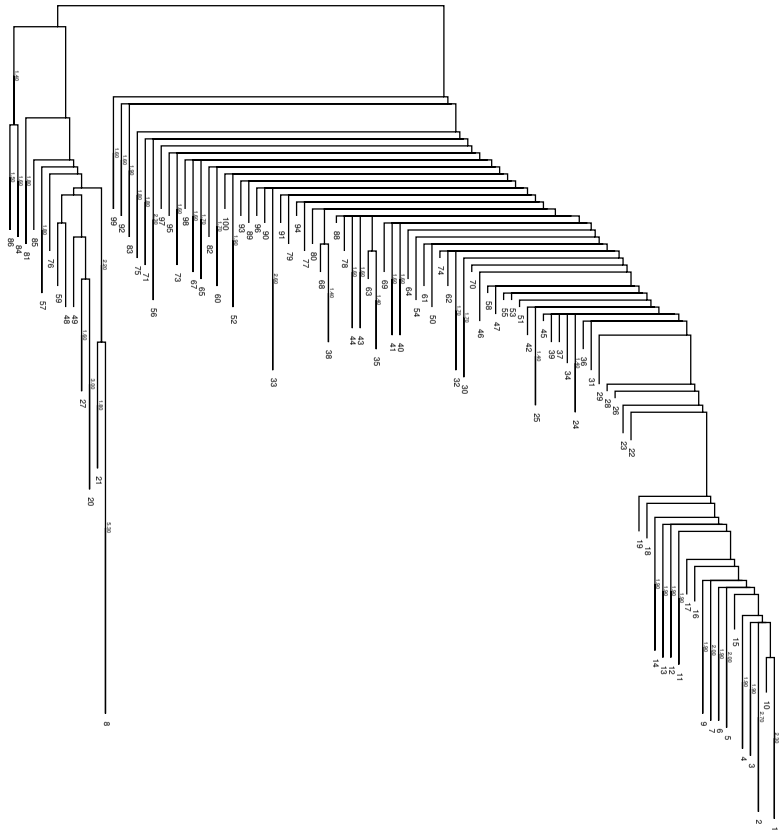
Markov Chains continued

A fundamental fact is that there exists a unique *stationary distribution* $\pi = (\pi_i : i \in \mathcal{S})$, i.e. a unique probability distribution satisfying the *balance equations*

$$\pi_j = \sum_i \pi_i p_{ij}$$

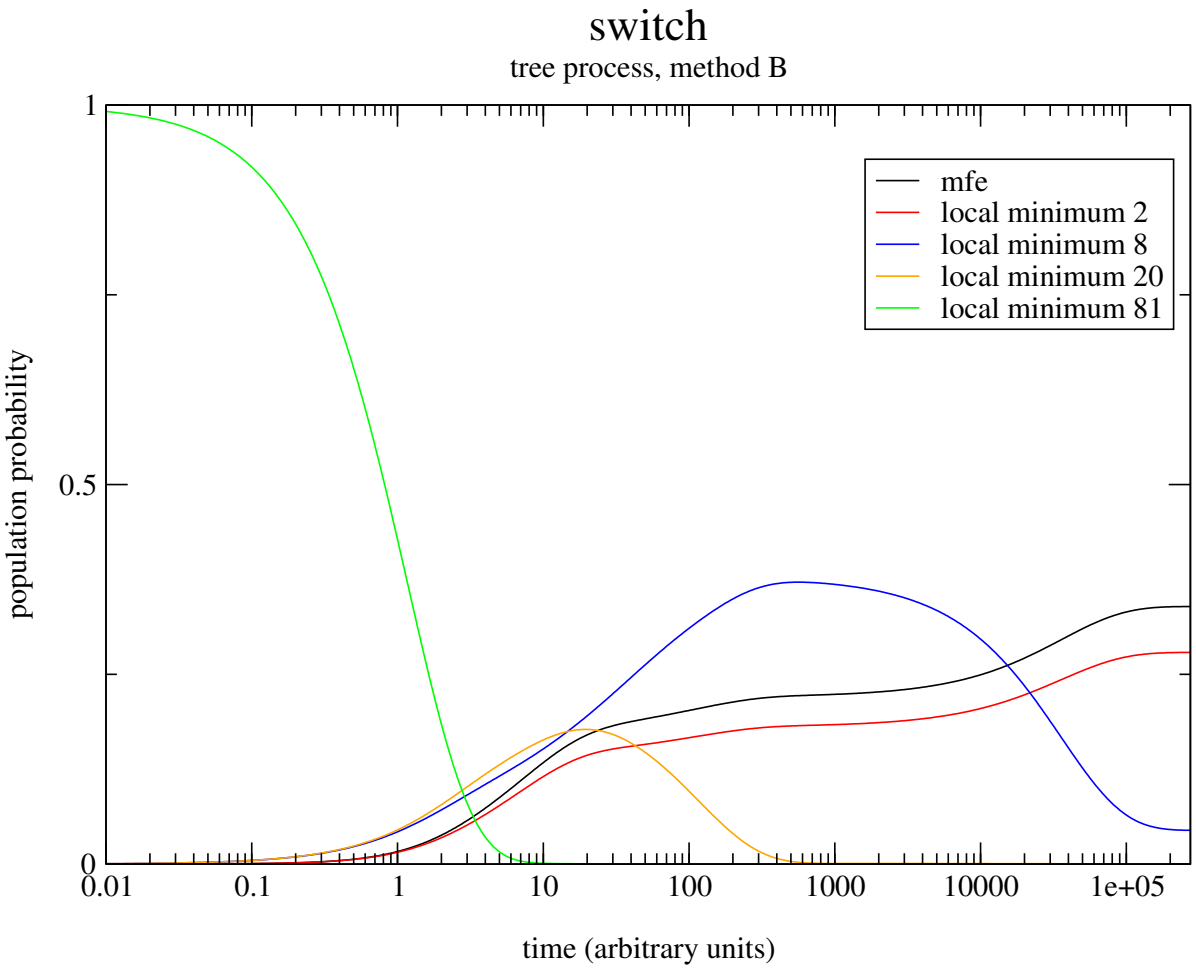
for all j.

Results: RNA switch



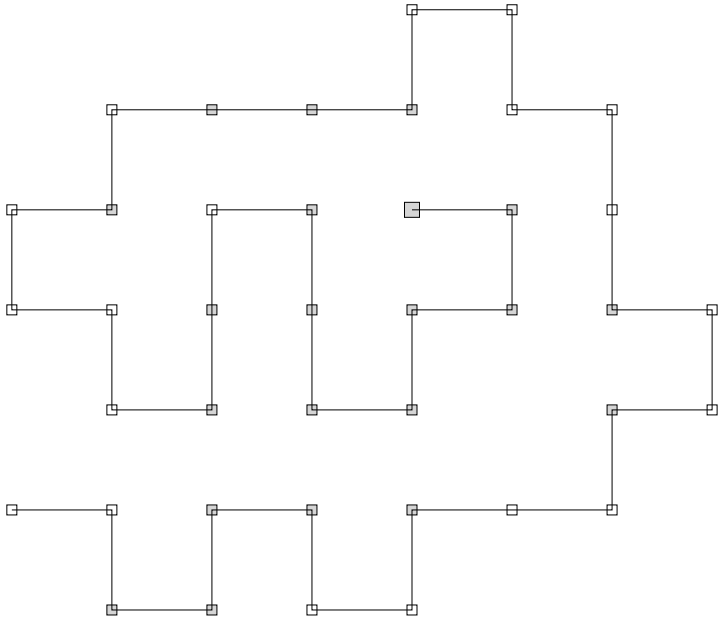
Barrier tree of the bi-stable RNA sequence GUGUUUGAGAGGAUAUGGCGUUUUUUUGGAUGC (n = 33)

RNA switch: Barrier tree dynamics

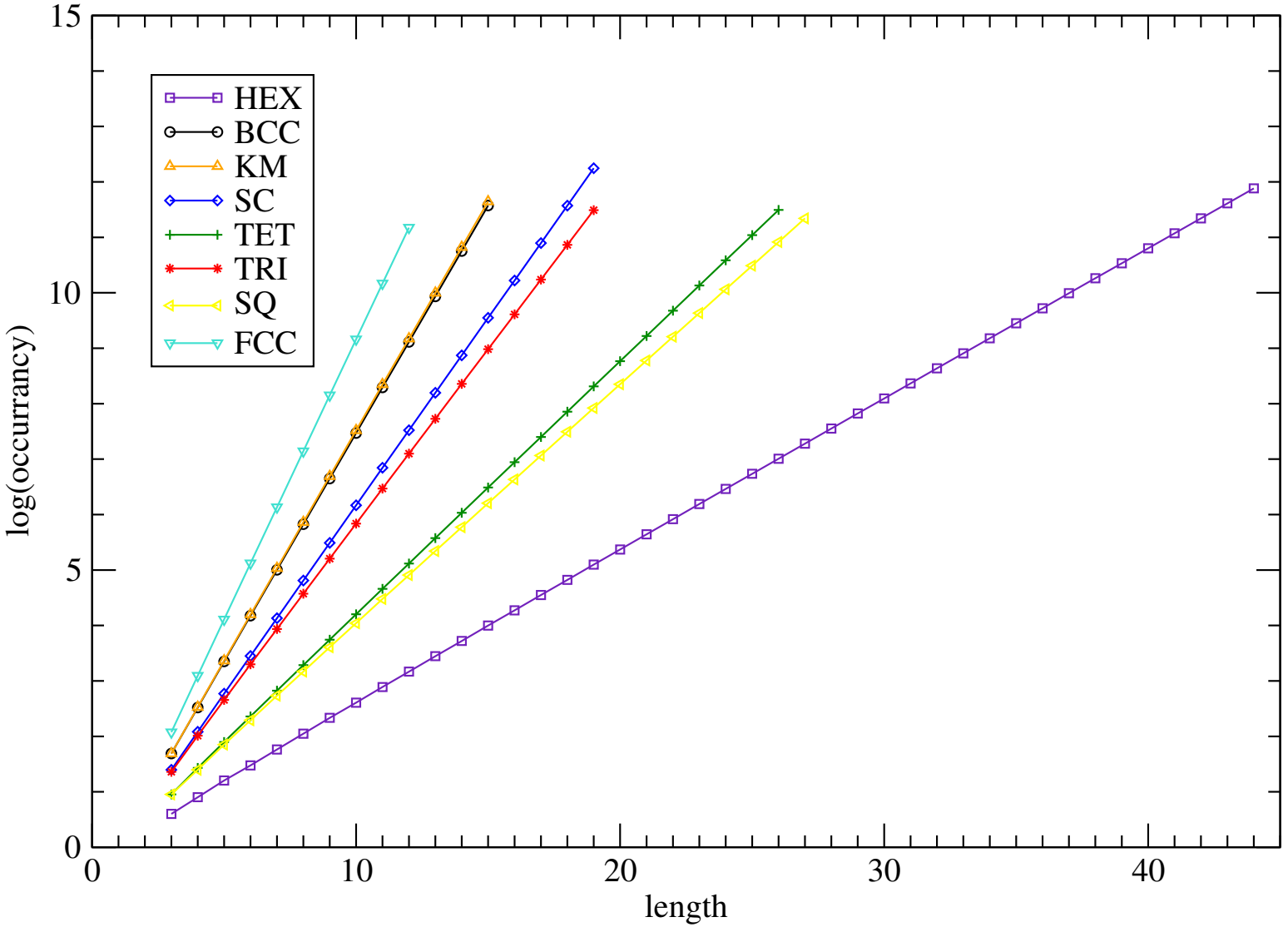


The next step: Lattice Proteins and SAWs

Our present interest focuses on calculating energy landscapes of *Lattice Proteins*



- HP model (later: other models)
- Energy function: number of HH contacs
- Apropriate move set must be established
- No dynamic programing algorithm for mfe and suboptimal folding
- Exhaustive calculation of SAWs is NP-hard


Exhaustive enumeration of SAWs
log(occurrancy)
length
0
5
10
15
0
10
20
30
40
HEX
BCC
KM
SC
TET
TRI
SQ
FCC