De Bruijn Graph and Sequence

part of “Graphen und Netzwerke in der Biologie”

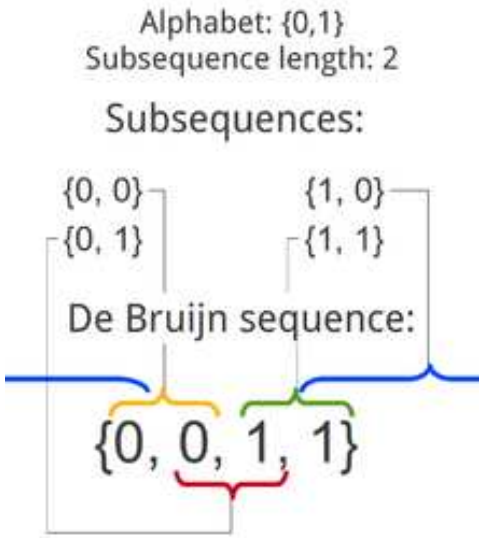
Sonja Prohaska

Computational EvoDevo
University Leipzig

Leipzig, SS 2011

De Bruijn graph and sequence

named after the Dutch mathematician **Nicolaas Govert de Bruijn**.



The De Bruijn sequence $B(k, n)$ is a **cyclic sequence** of a given alphabet A with size k for which every possible subsequence of length n appears as a sequence of consecutive characters exactly once.

Constructing a De Bruijn Sequence

To construct a sequence $B(k, n)$ we compute an Eularian cycle of a $(n-1)$-dimensional De Bruijn graph over k symbols.

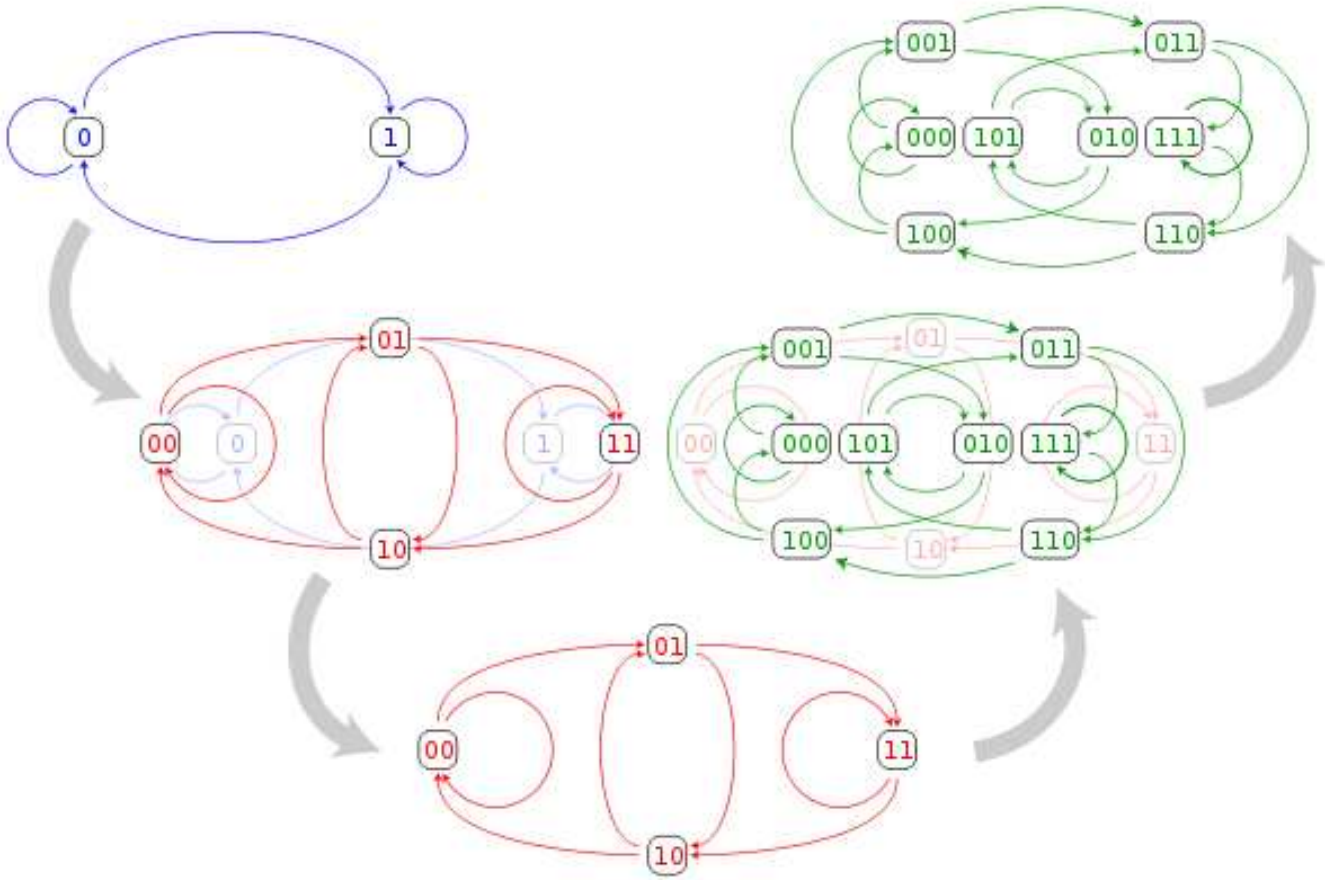
In graph theory, an n-dimensional **De Bruijn graph** of k symbols is a directed graph representing overlaps between sequences of symbols. It has k^n vertices, consisting of all possible length-n sequences of the given symbols.

A directed edge is drawn from vertex $v_1 = (s_i ... s_{i+n-1})$ to $v_2 = (s_j ... S_{j+n})$ if the symbol at s_{i+1} is identical to s_j, the symbol at s_{i+2} is identical to s_{j+1}, ... and the symbol at s_{i+n-1} is identical to s_{j-1+n}.

De Bruijn Graph Properties

- if $n = 1$ all the vertices are connected forming a total of k^2 edges
- each vertex has exactly k incoming and k outgoing edges
- each n-dimensional De Bruijn graph is the line digraph of the $(n-1)$-dimensional De Bruijn graph with the same set of symbols
- each De Bruijn graph is **Eulerian** (it has an Eularian circuit, visiting every **edge** exactly once)
- each De Bruijn graph is **Hamiltonian** (it has an Hamiltonian path, visiting every **vertex** exactly once)
- the Euler cycles and Hamiltonian cycles of these graphs are De Bruijn sequences

De Bruijn Graph Properties



Each vertex of the n-dimensional De Bruijn graph corresponds to an edge of the $(n-1)$-dimensional De Bruijn graph, and each edge in the n-dimensional De Bruijn graph corresponds to a two-edge path in the $(n-1)$-dimensional De Bruijn graph.

De Bruijn Graph for all possible nucleotide doublets