O Sekwencjach DNA i grafach

Bartek Wilczyński

27. lutego 2018

Informacje

- Strona wykładu `regulomics.mimuw.edu.pl/wp/categories/wbo`
- e-mail `bartek@mimuw.edu.pl`
- konultacje: środy 8:30–10:00, pokój 5770

Kryteria zaliczenia

- Projekty zaliczeniowe: 10+20 punktów
- Kolokwium 20 punktów
- Zaliczenie Ćwiczeń: $>$ 30 punktów
- Egzamin ustny (wymagane zaliczenie ćwiczeń)

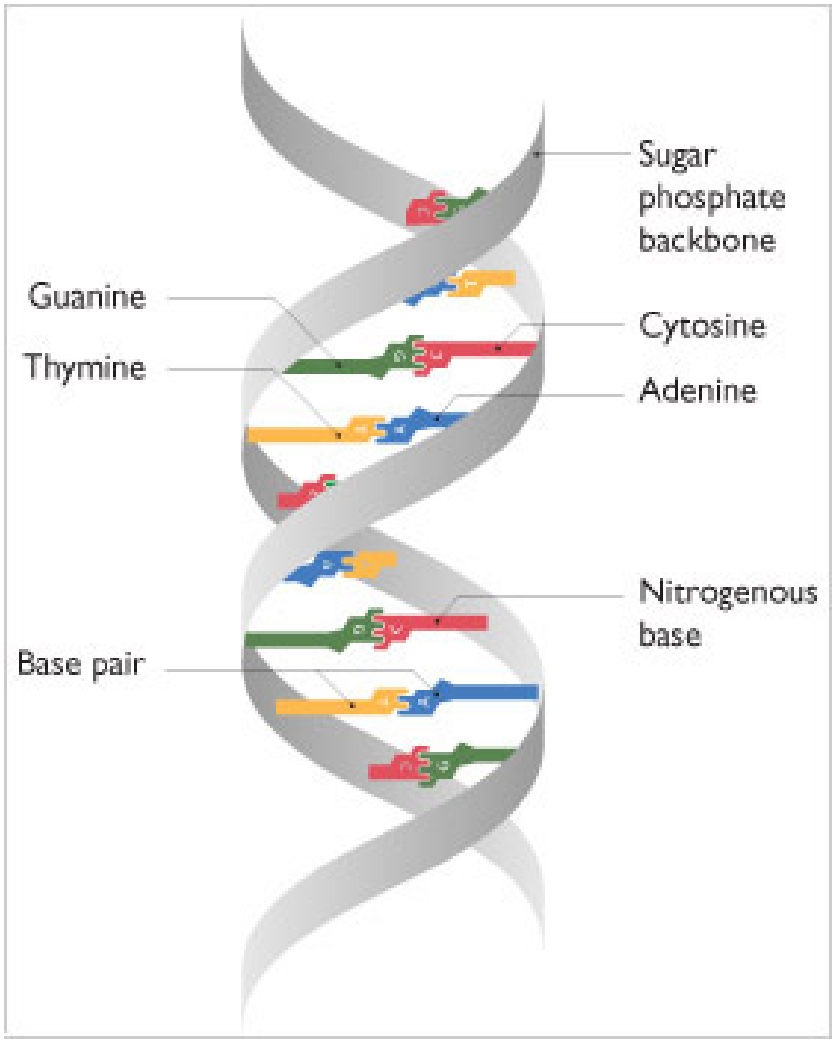
Tematyka wykładów

- Sekwencje biologiczne (DNA, RNA i białka)
- Podobieństwo sekwencji a ewolucja
- Porównywanie sekwencji biologicznych (miary, statystyki i algorytmy)
- Drzewa filogenetyczne - konstrukcja i zastosowania
- Sekwencje białkowe a ukryte modele Markowa
- Sekwencje nie- kodujące w genomach i motywy sekwencyjne
- Modele grafowe w biologii

Źródła

- Computational Molecular Biology (P. Pevzner)
- Biological Sequence Analysis (R. Durbin i in.)
- Sequence – Evolution – Function (Koonin i Galperin)

DNA and its role



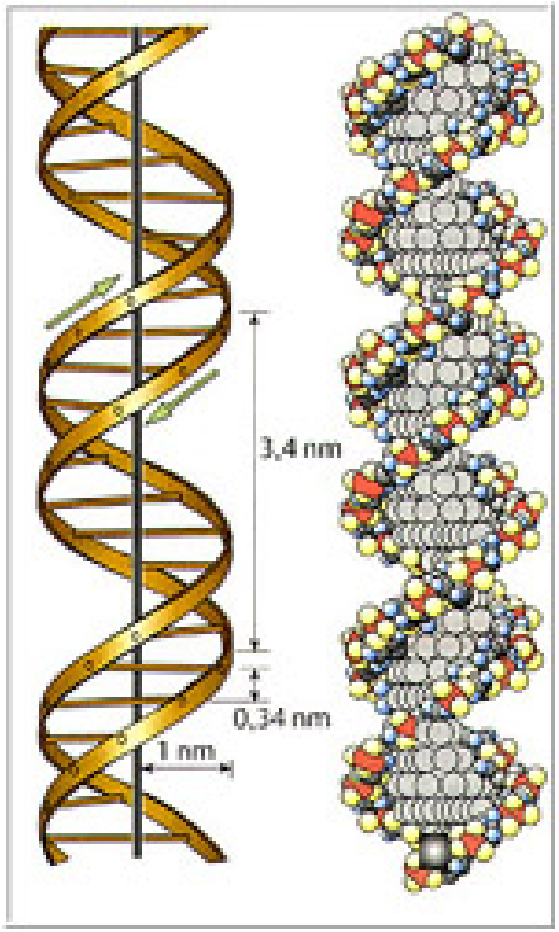


image (c) T. Maxwell

DNA replication enables inheritance

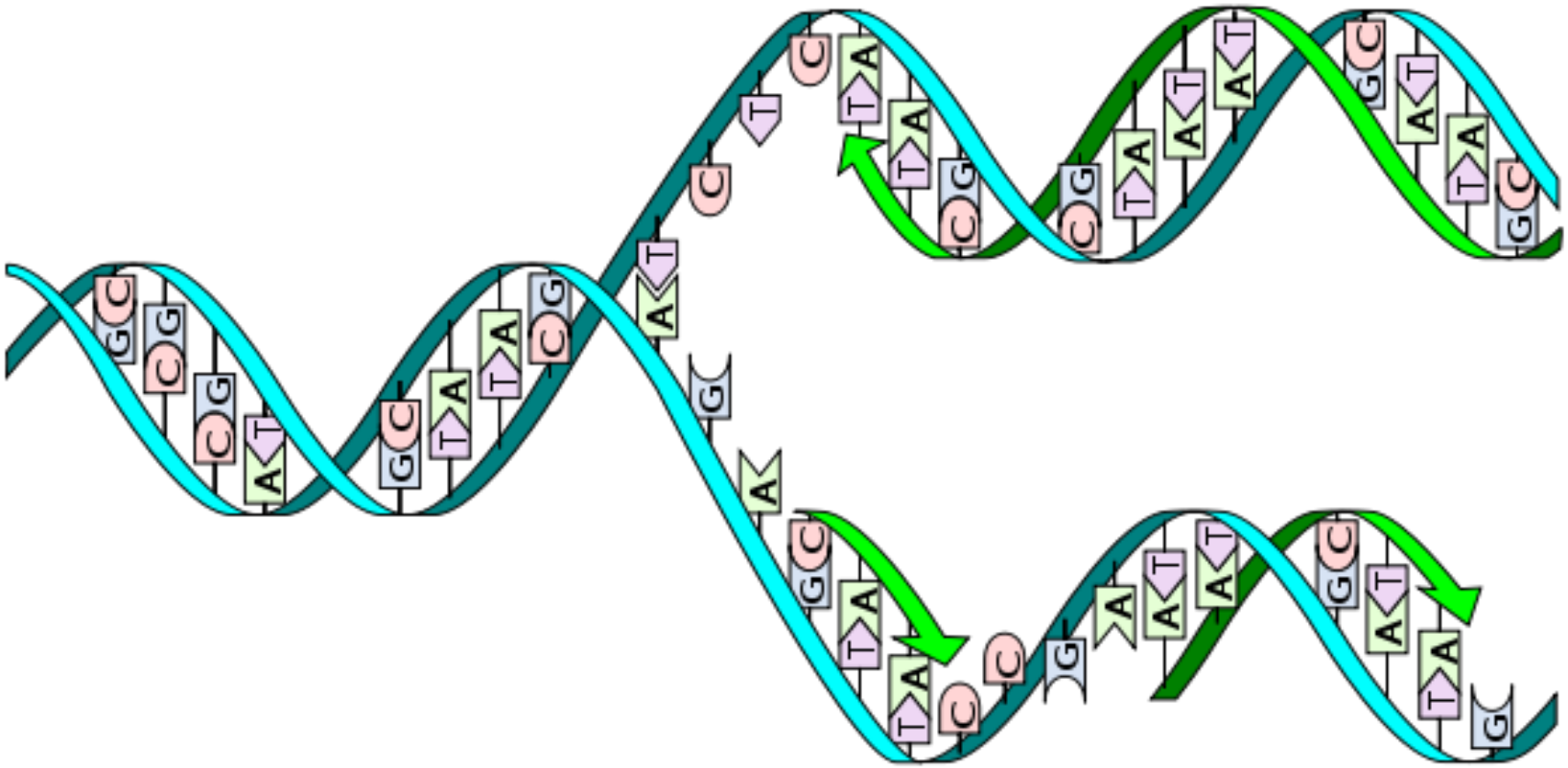


image (c) wikimedia

Central dogma of molecular biology

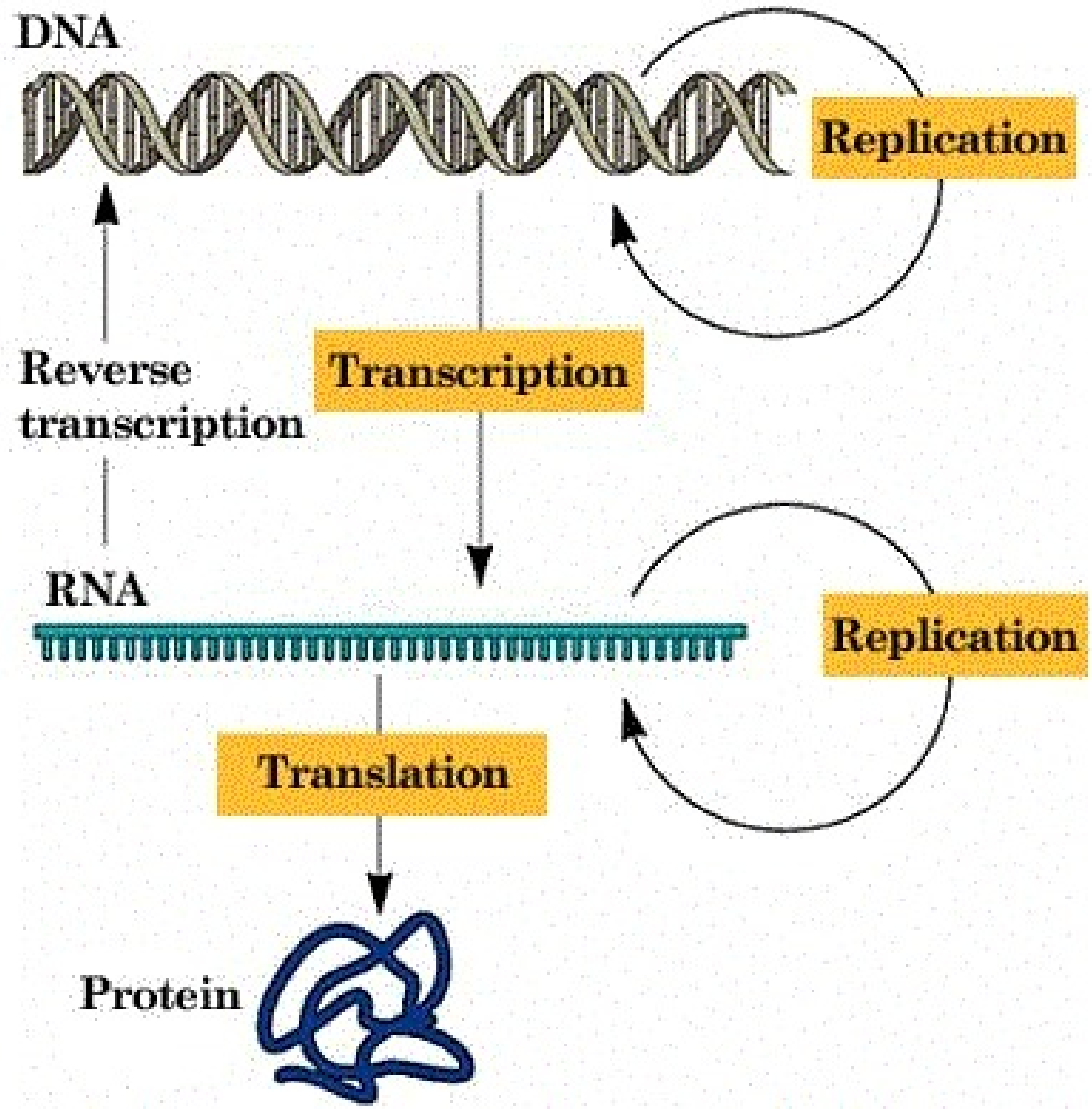


image (c) SCF IIT Delhi

Microarray concept

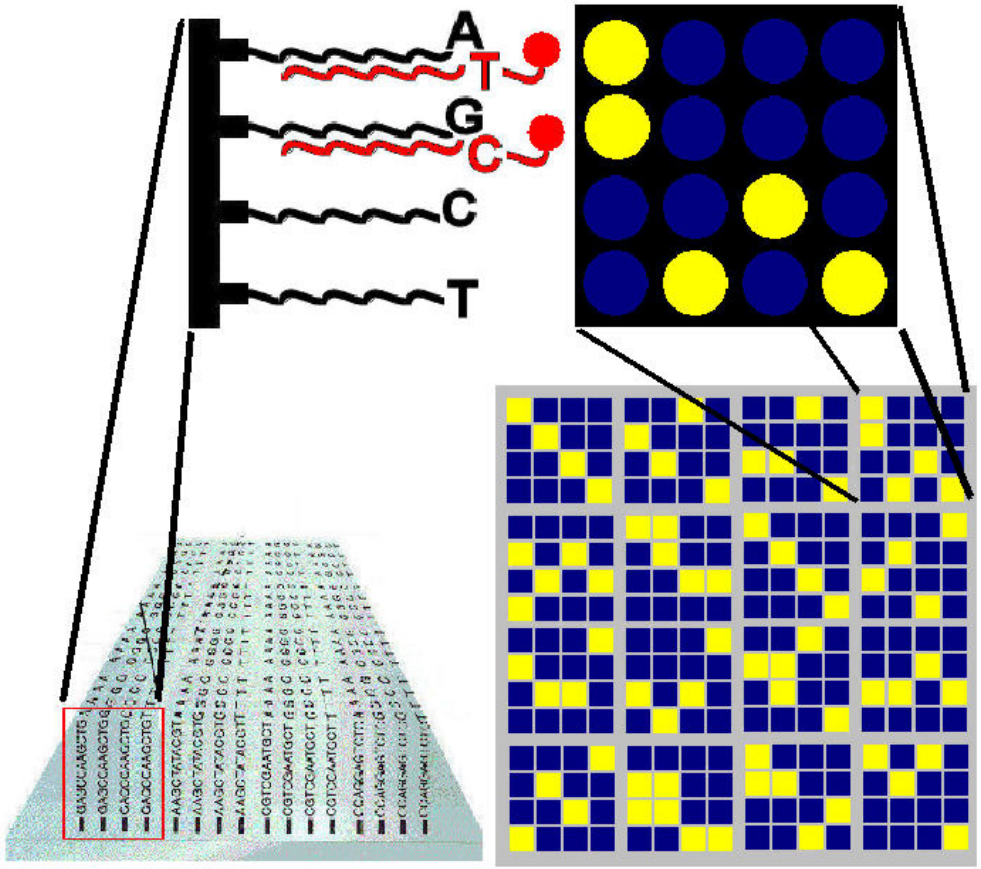


image (c) Steven M. Carr

affymetrix chip

image (c) Molecularstation.com

Sequencing by hybridization

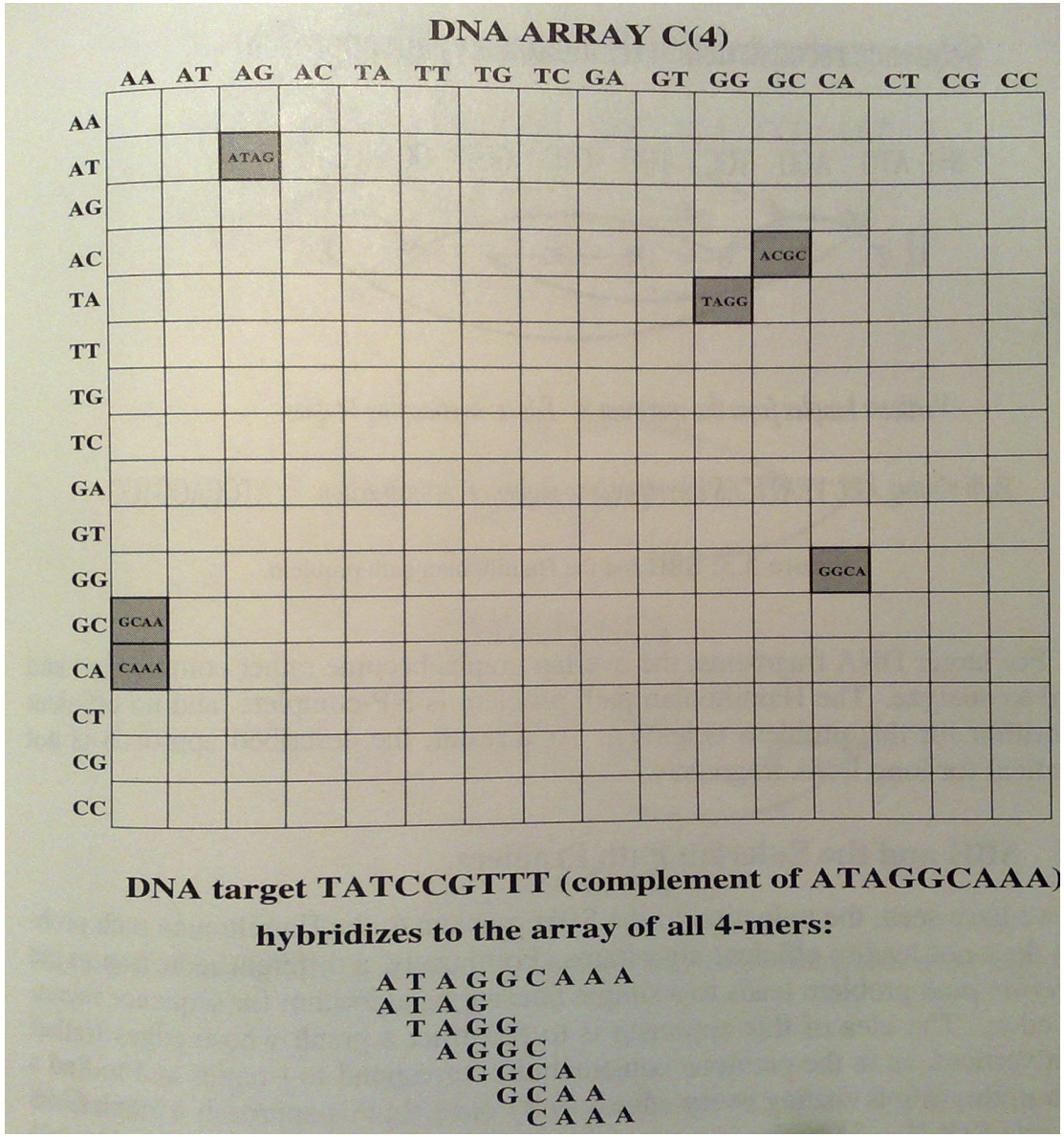


image (c) P. Pevzner

SBH - natural approach

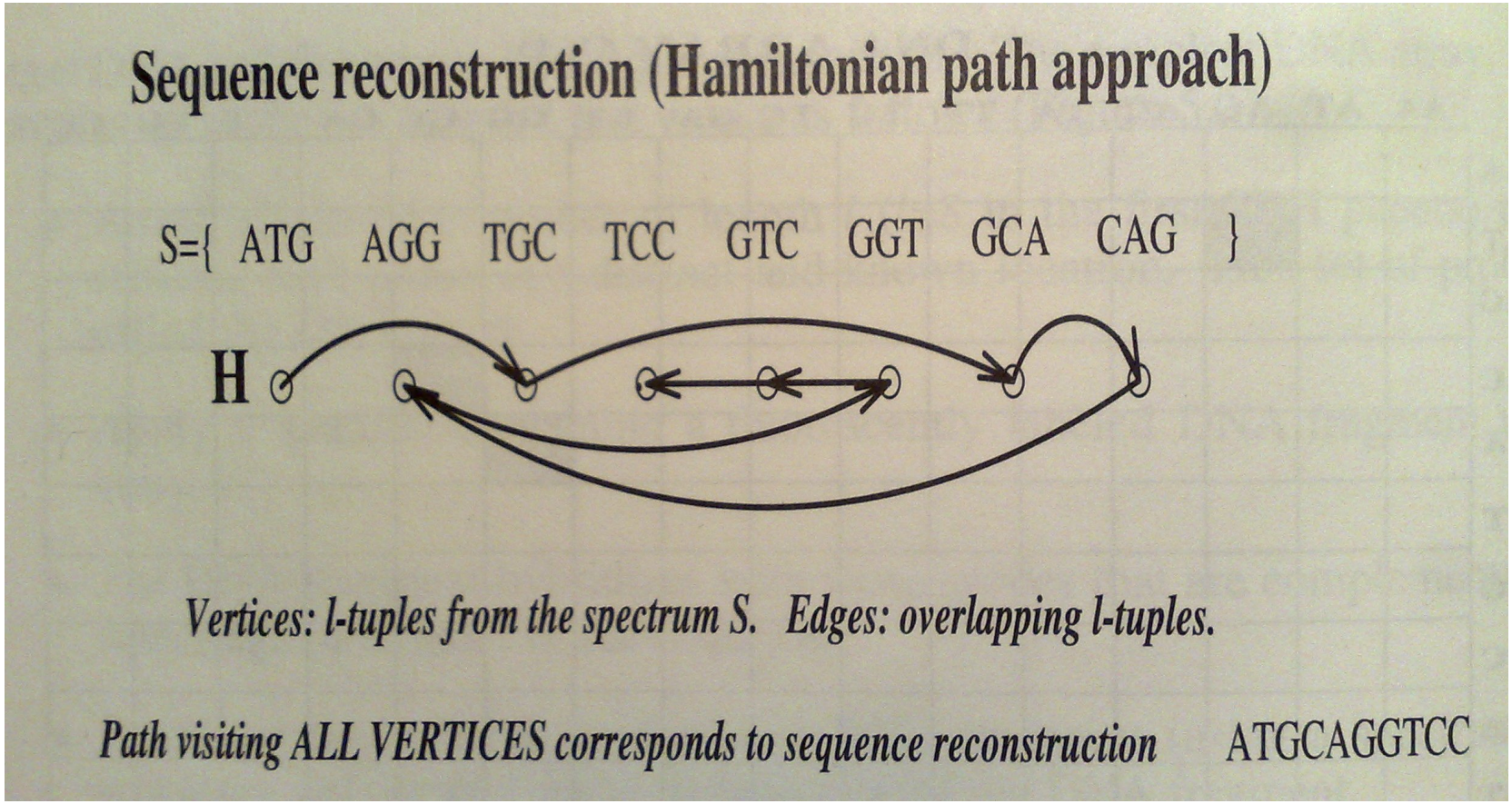


image (c) P.Pevzner

SBH - multiple solutions

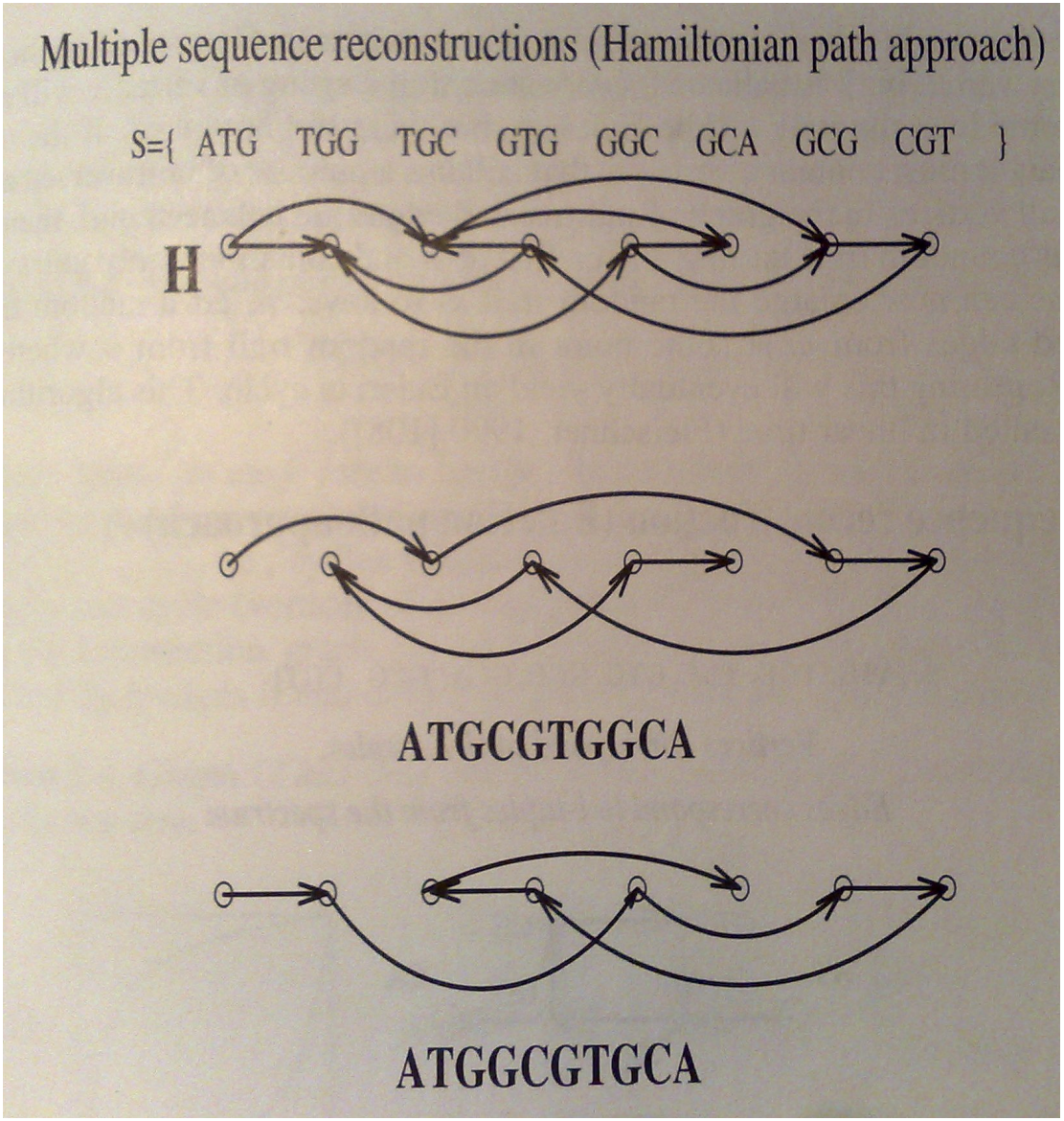


image (c) P.Pevzner

Hamiltonian and Eulerian graphs

Finding Hamiltonian paths is NP-complete, while finding Eulerian paths is easy.

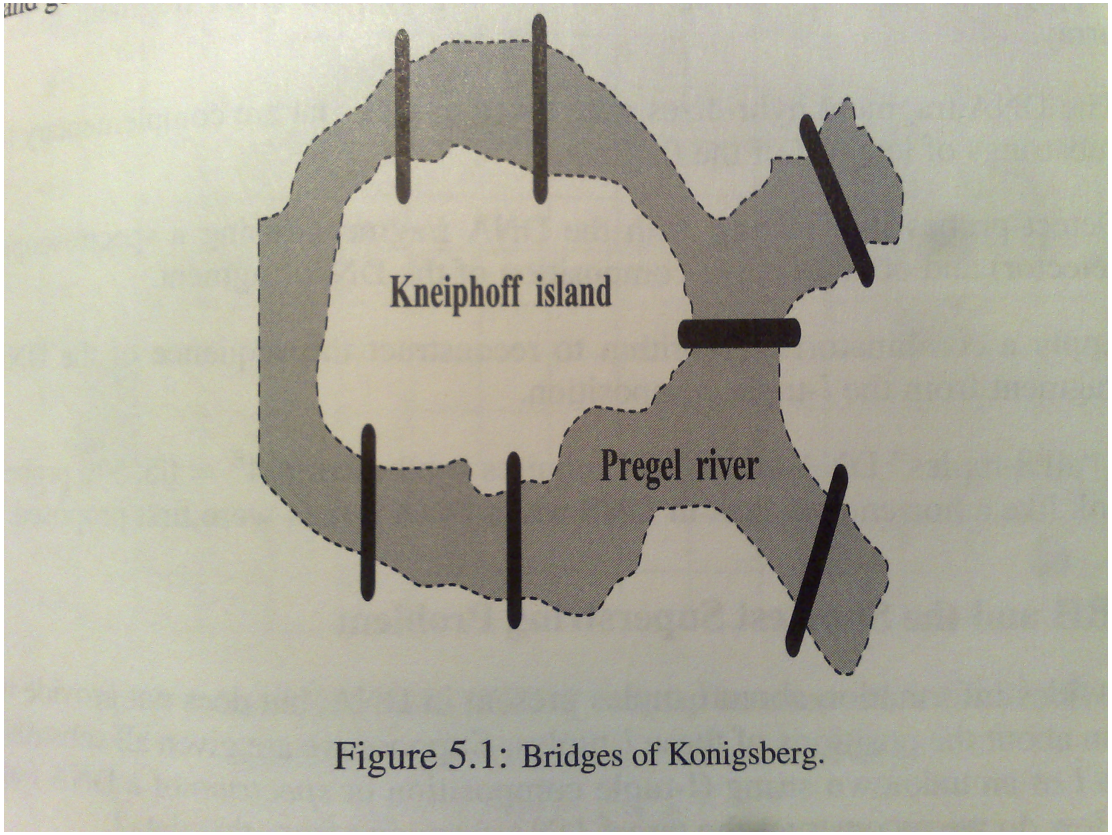


Figure 5.1: Bridges of Konigsberg.

image (c) P.Pevzner

SBH - Eulerian formulation

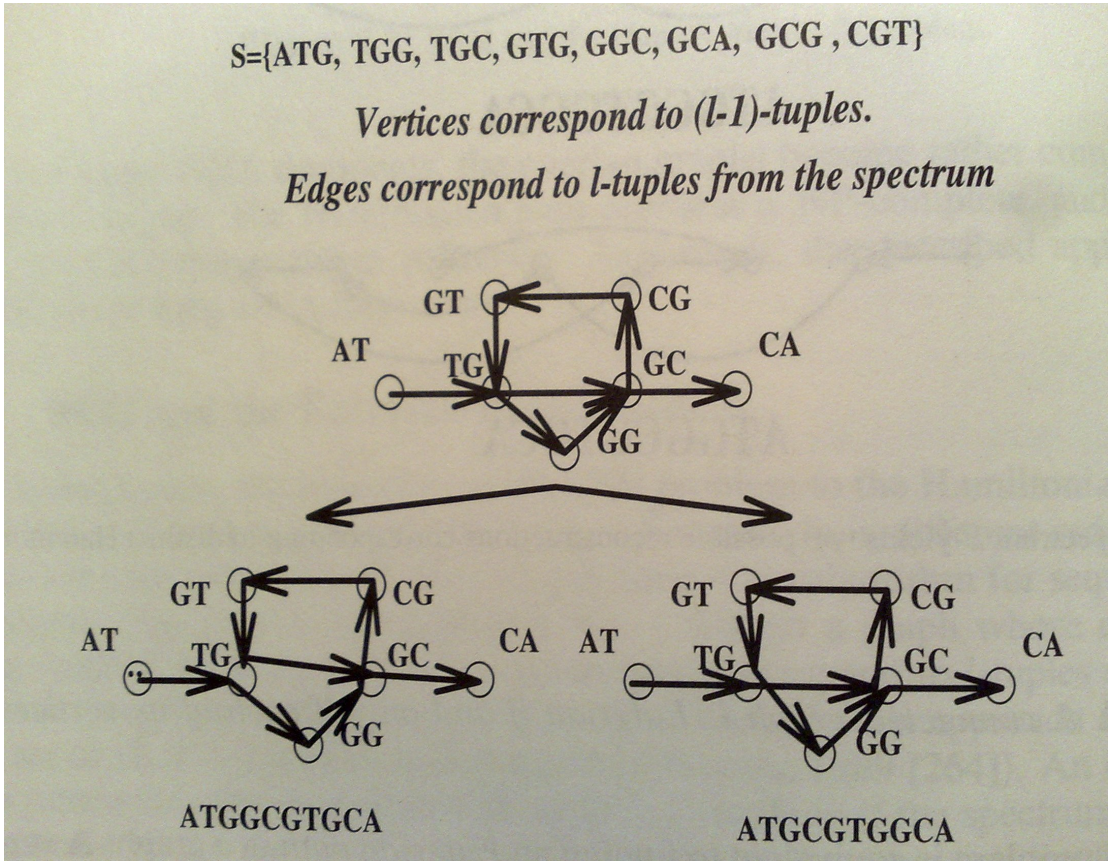


image (c) P.Pevzner

Other uses of microarrays

- SBH not practical due to hybridization errors, superseded by Next generation sequencing
- Gene arrays for RNA abundance quantification
- Snip arrays for detecting mutations (disease screening, paternity tests)
- aCGH arrays for detecting copy number variation
- biochips for quick pathogen detection

Designing unique probes

- We are given a set of n different DNA sequences (targets) $\mathcal{S} = \{s_1 \ldots s_n\}$
- We need to design a set $\mathcal{P} = \{p_1 \ldots p_n\}$ of n sequences (probes) of length k, such that for each $i \in \{1 \ldots n\}$, probe p_i hybridizes with sequence s_i and does not hybridize with any other sequence $s_{j,j \neq i}$.
- Depending on the amount of sequence identity and parameter k, there might be no valid solutions or exponentially many solutions
- Instead of searching for probes of the same length, one might search for probes of the same *melting temperature*

$$t_m = 4 \cdot \#GC + 2 \cdot \#AT$$