



BIOINFORMATYKA

SEKWENCJONOWANIE PRZEZ HYBRYDYZACJĘ - ALGORYTM DOKŁADNY

PIOTR WAWRZYNIAK

INSTYTUT INFORMATYKI, POLITECHNIKA POZNAŃSKA

PIOTR.WAWRZYNIAK@CS.PUT.POZNAN.PL

WWW.PIOTR.E.WAWRZYNIAK.DOCTORATE.PUT.POZNAN.PL

INSTANCJA

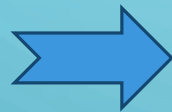
$n=9$ $k=3$

TCGTACGAC

TCG
CGT
GTA
TAC
ACG
CGA
GAC



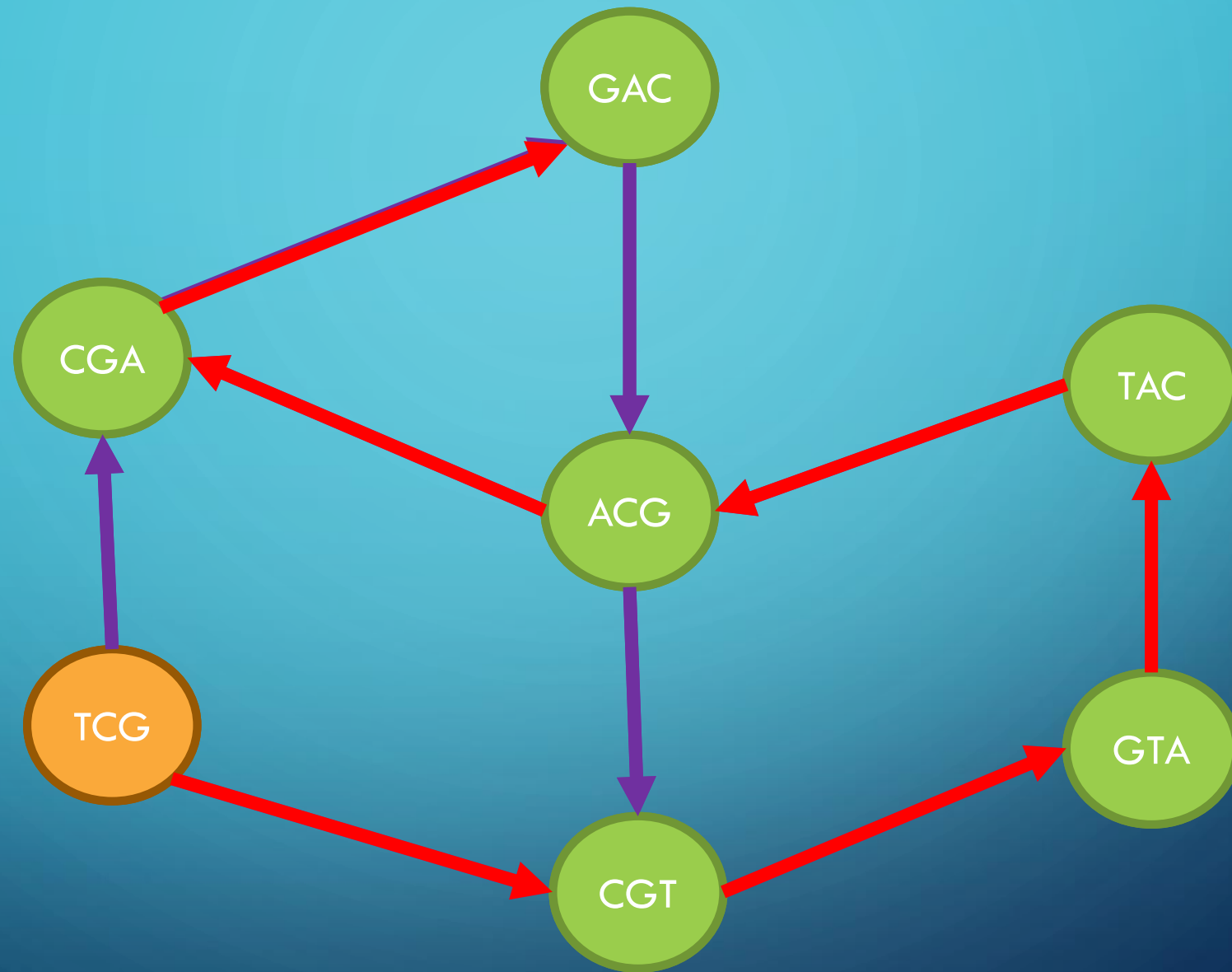
ACG
CGA
CGT
GAC
GTA
TAC
TCG

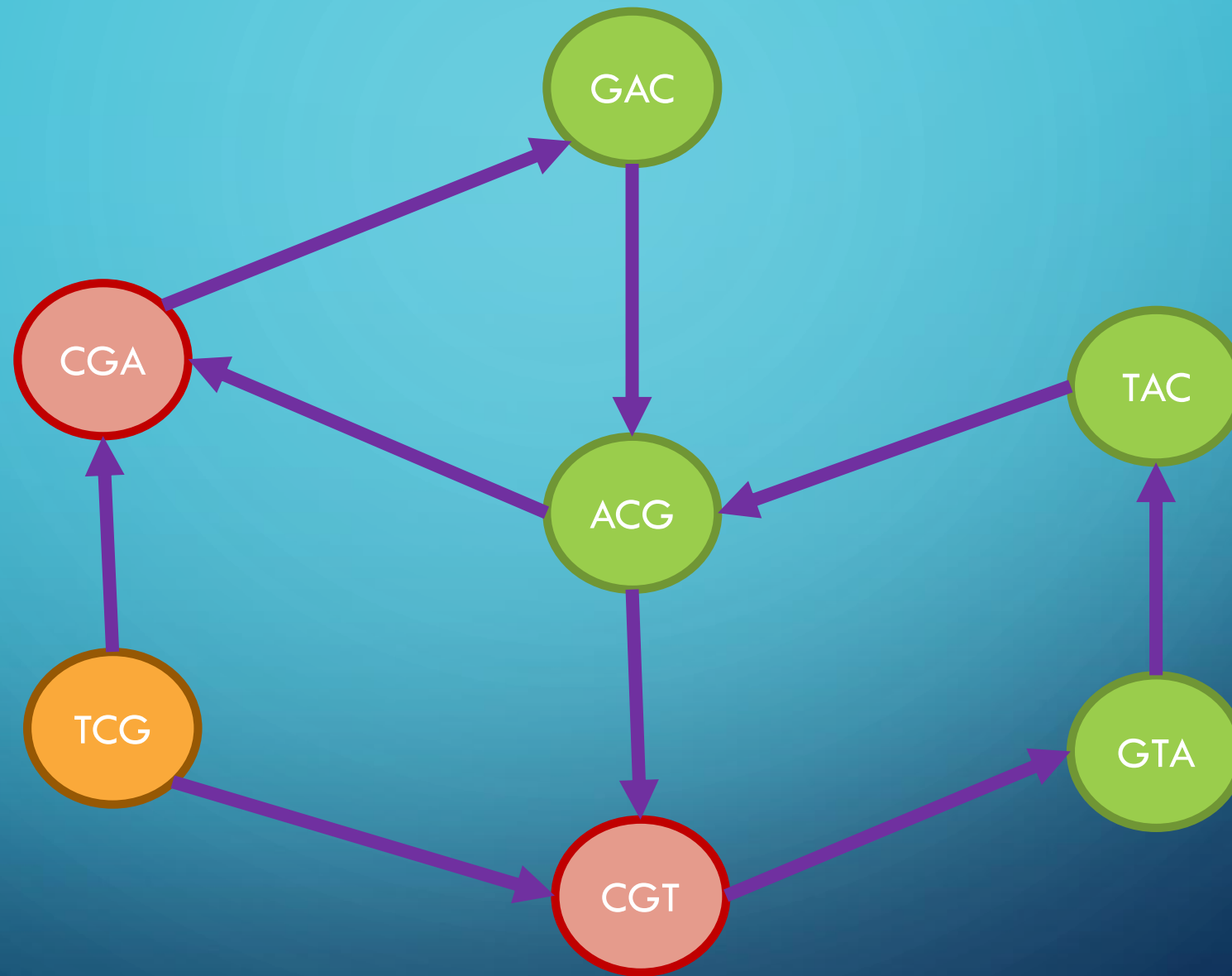


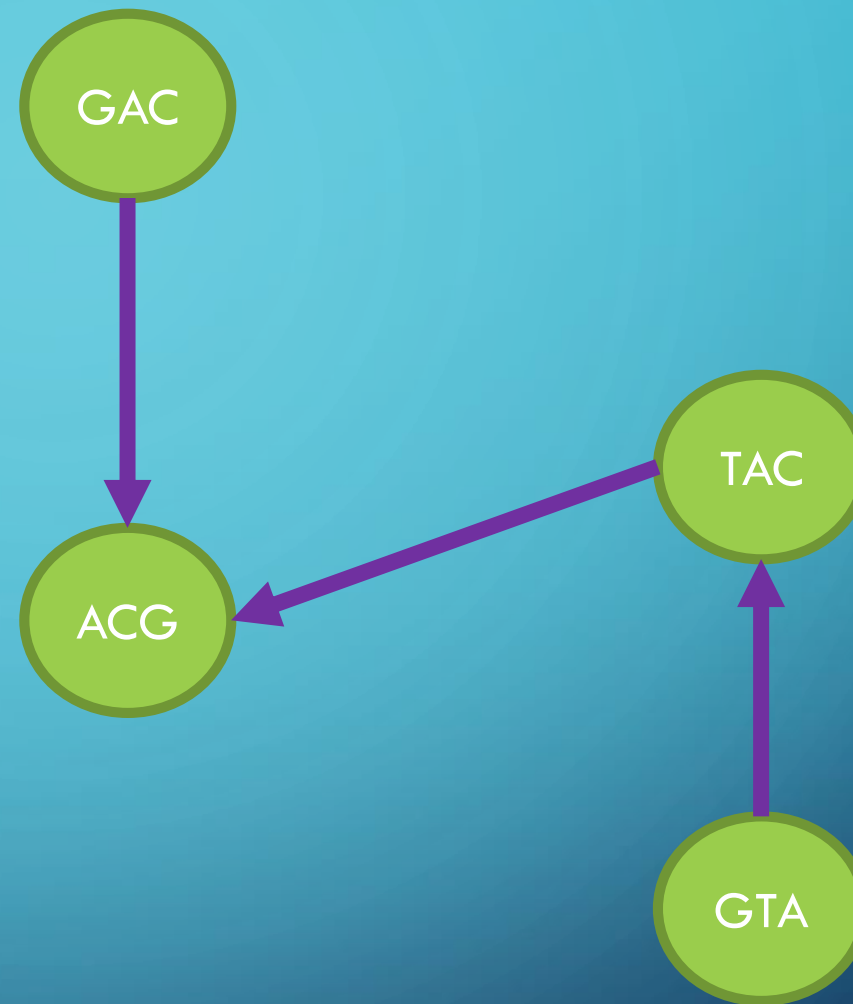
ACG
~~CGA~~
~~CGT~~
GAC
GTA
TAC
TCG
CAT
TAA



ACG
CAT
GAC
GTA
TAA
TAC
TCG





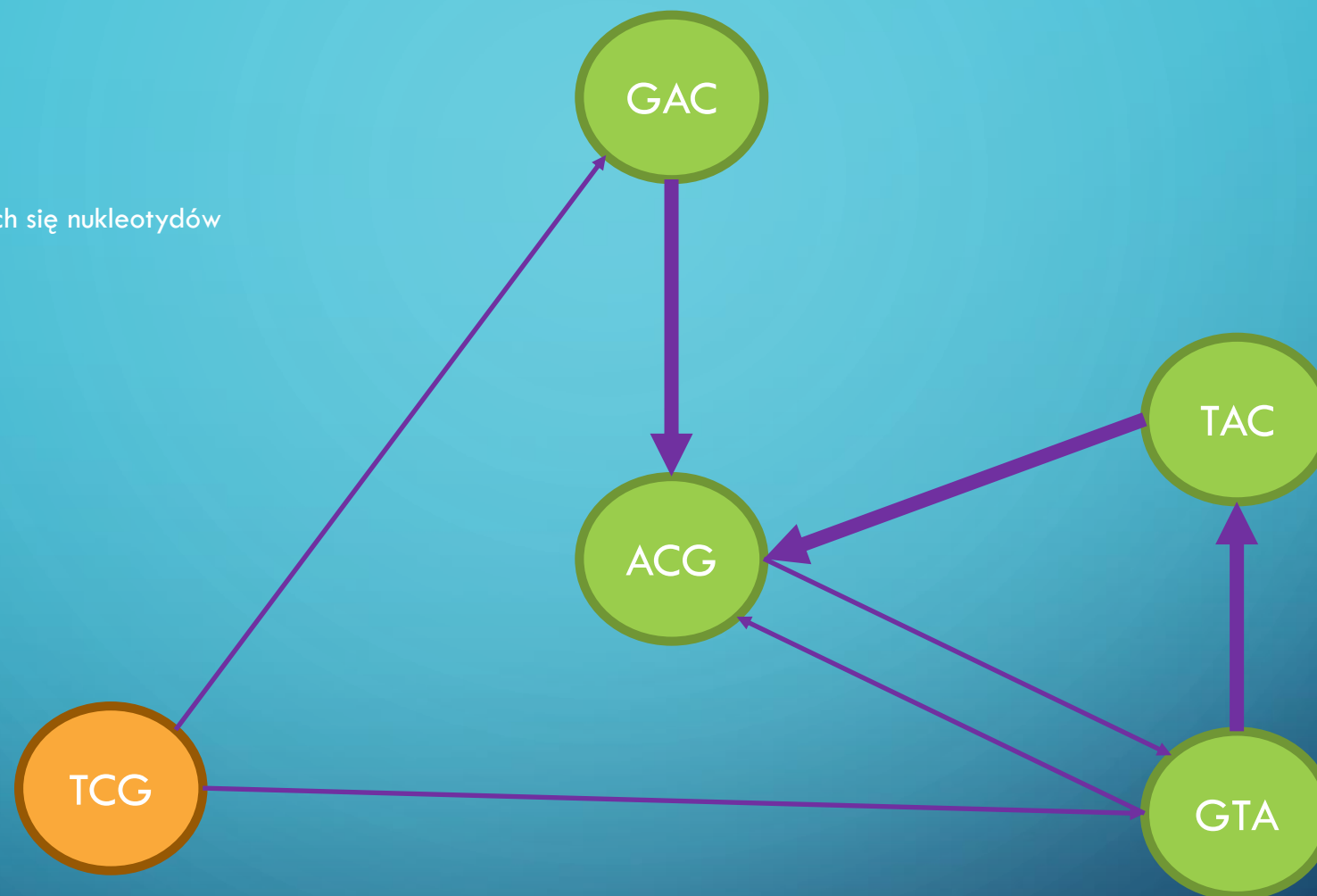


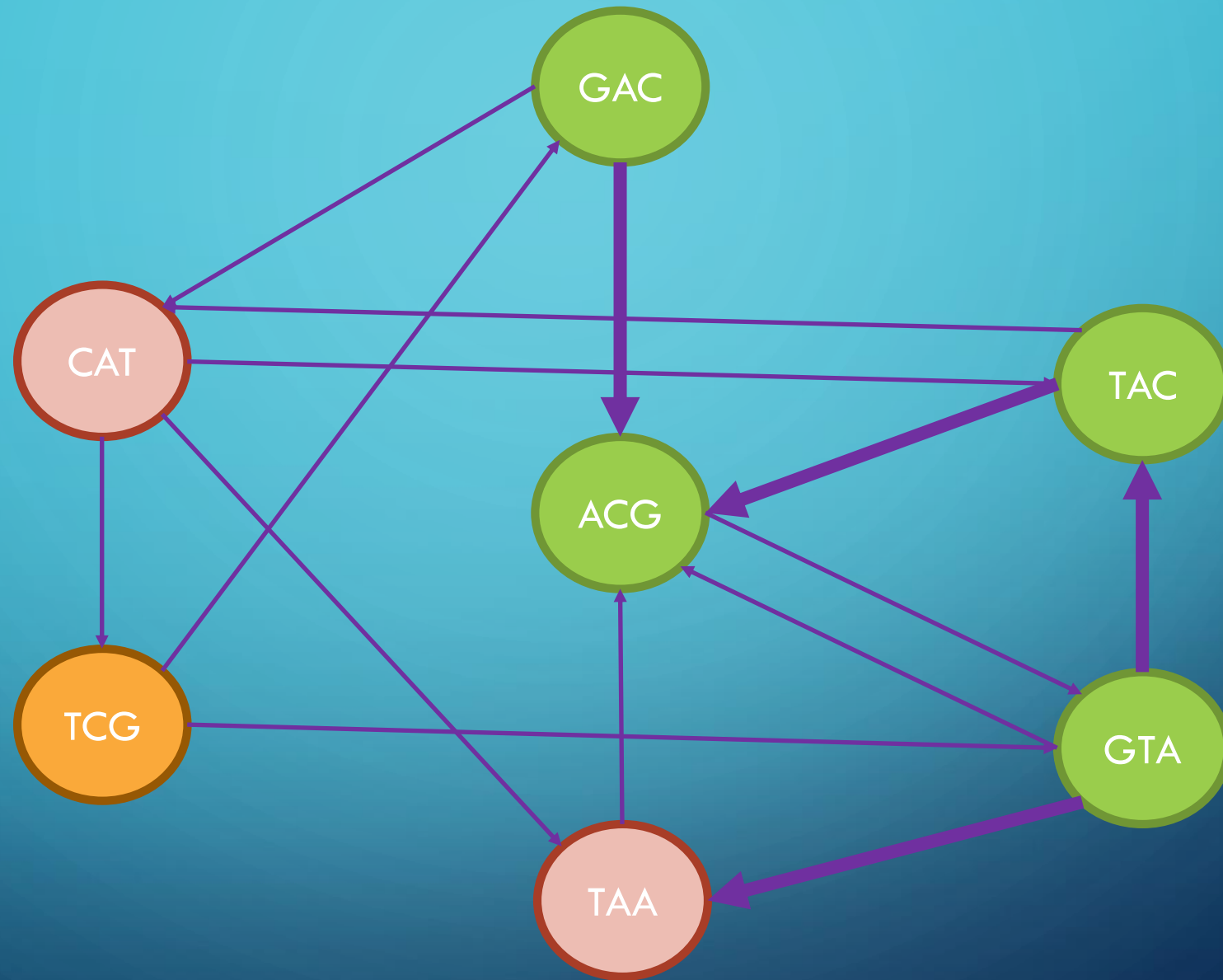
Koszt łuku: $k-l$

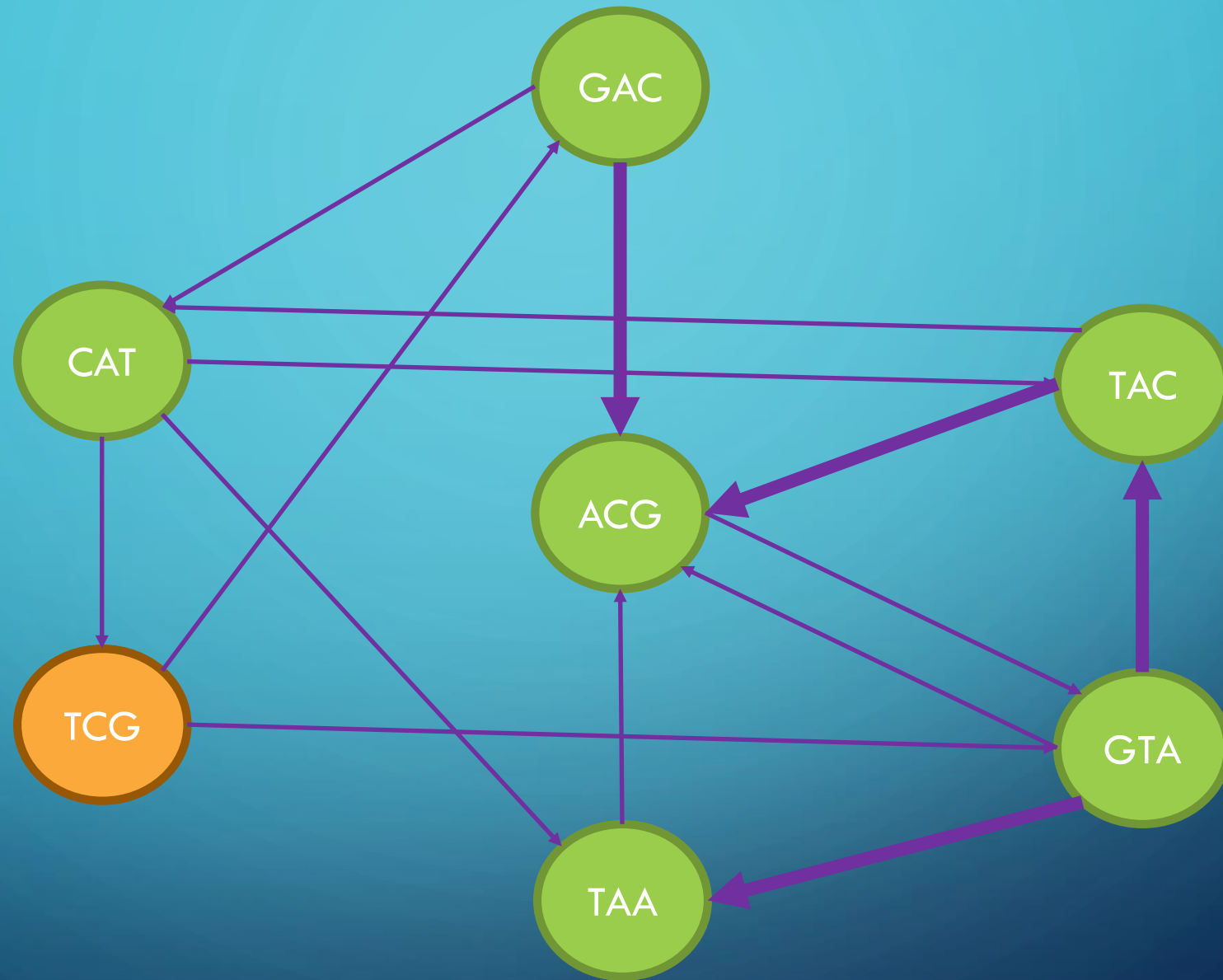
l : ilość nakładających się nukleotydów

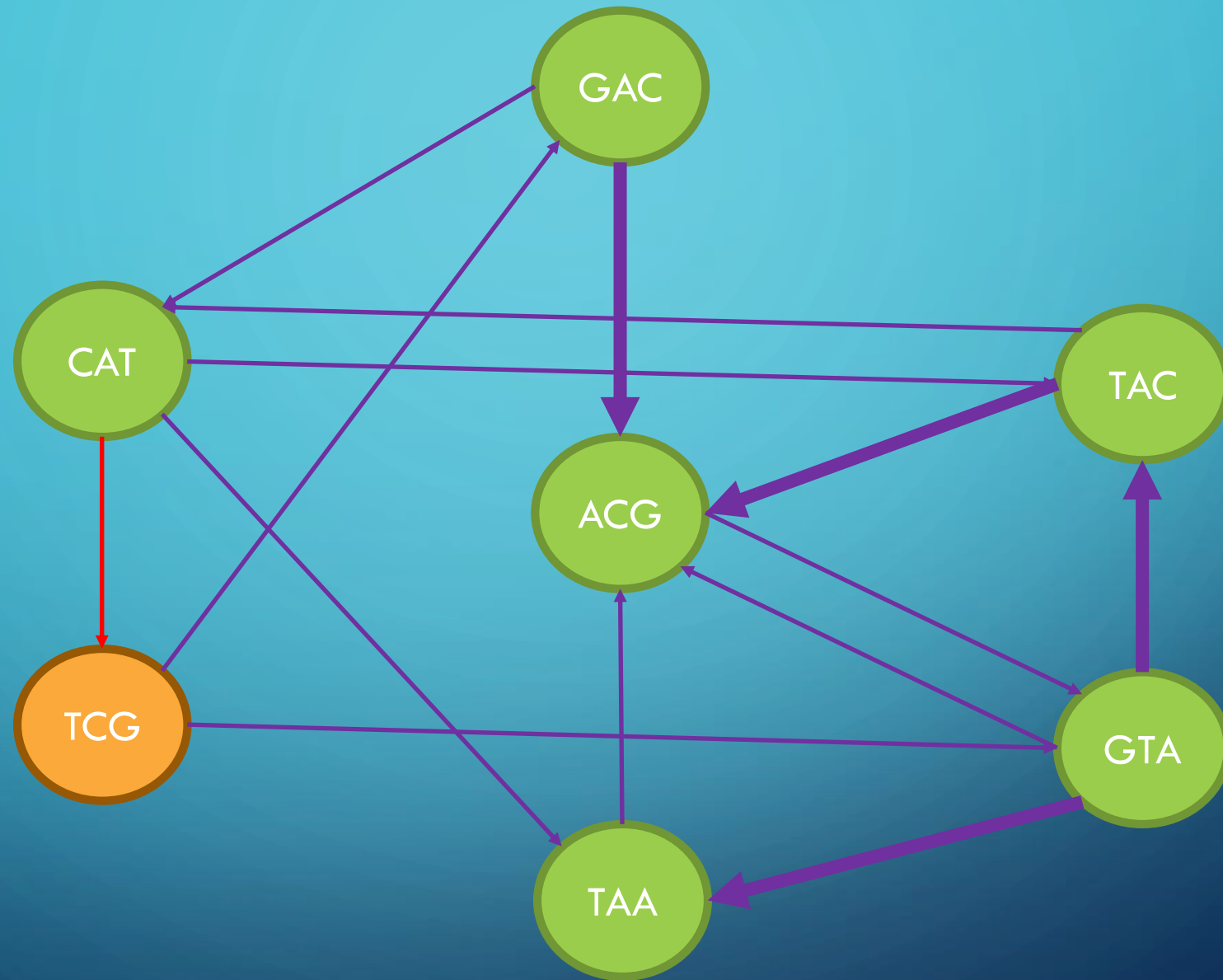
1: GTA
TAC

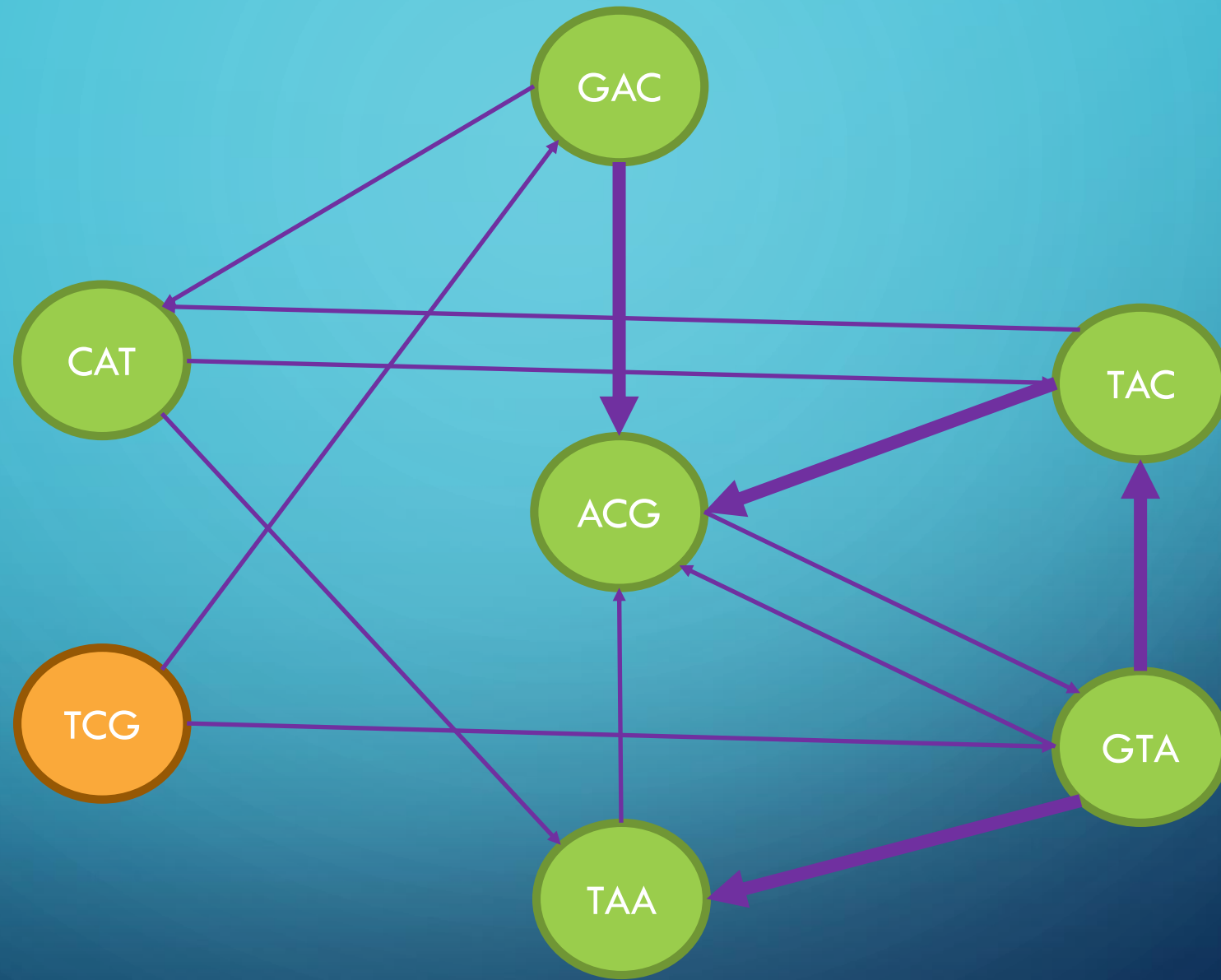
2: GTA
ACG











PROBLEM

Użyć jak najwięcej oligonukleotydów (wierzchołków)

Unikać nieoptymalnych dopasowań oligonukleotydów

Całkowita długość wynikowego ciągu nie może przekraczać N

Cel: znaleźć najlepsze dopasowanie oligonukleotydów

SELECTIVE TRAVELING SALESMEN

Użyć jak najwięcej oligonukleotydów (wierzchołków)

Dla każdego miasta (wierzchołka) przypisany zysk

Unikać nieoptymalnych dopasowań oligonukleotydów

Do każdej drogi (łuku) przypisany koszt

Całkowita długość wynikowego ciągu nie może przekraczać N

Całkowity koszt podróży nie może przekroczyć ustalonej kwoty

Cel: znaleźć najlepsze dopasowanie oligonukleotydów

Cel: znaleźć ścieżkę maksymalizującą zysk

A decorative graphic consisting of blue circuit-like lines with small circles at the ends, extending horizontally from the left and right sides of the central black box.

ROZWIĄZANIE

INTEGER LINEAR PROGRAMMING



WARTOŚCI I ZMIENNE

Wartości

- z : liczność spektrum
- $p_i = 1$: zysk przypisany wierzchołkowi
- c_{ij} : koszt łuku łączącego wierzchołki i z j
- $c = k - l$: maksymalny dopuszczalny koszt

Zmienne

- x_{ij} : 1 jeśli łuk (i, j) należy do rozwiązania,
0 w przeciwnym wypadku

OGRANICZENIA

- (1) Początkowy oligonukleotyd nie ma łuków wchodzących
- (2) Początkowy oligonukleotyd ma jeden łuk wychodzący
- (3) Pozostałe oligonukleotydy mają co najwyżej jeden łuk wchodzący
- (4) Pozostałe oligonukleotydy mają co najwyżej jeden łuk wychodzący
- (5) Poza początkowym oligonukleotydem nie ma innego początku ścieżki
- (6) Długość ścieżki nie przekracza długości N badanego DNA

$$\sum_{i=1}^z x_{i1} = 0$$

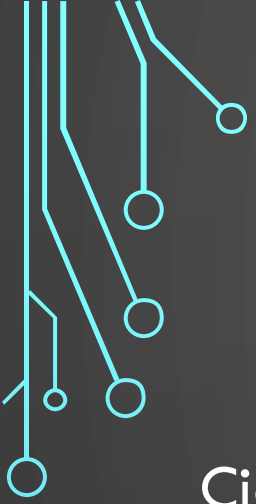
$$\sum_{i=1}^z x_{1i} = 1$$

$$\sum_{i=1}^z x_{ai} \leq 1, a = 2 \dots z$$

$$\sum_{i=1}^z x_{ia} \leq 1, a = 2 \dots z$$

$$\sum_{i=1}^z x_{ai} - \sum_{j=1}^z x_{ja} \leq 0, a = 2 \dots z$$

$$\sum_{i=1}^z \sum_{j=1}^z c_{ij} x_{ij} \leq n - k$$



CEL

Ciąg zawierający jak najwięcej oligonukleotydów ze spektrum

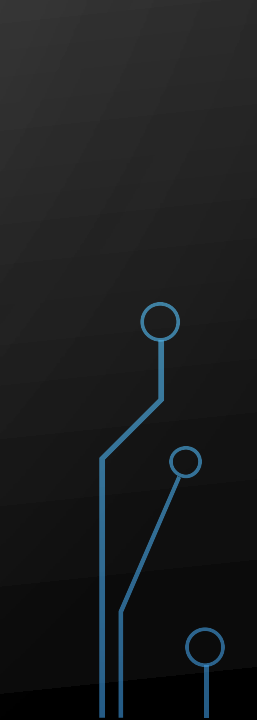
$$\max: p_1 + \sum_{i=2}^Z (p_i \sum_{j=1}^Z x_{ij})$$

Algorytm na podstawie:

DNA Sequencing With Positive and Negative Errors

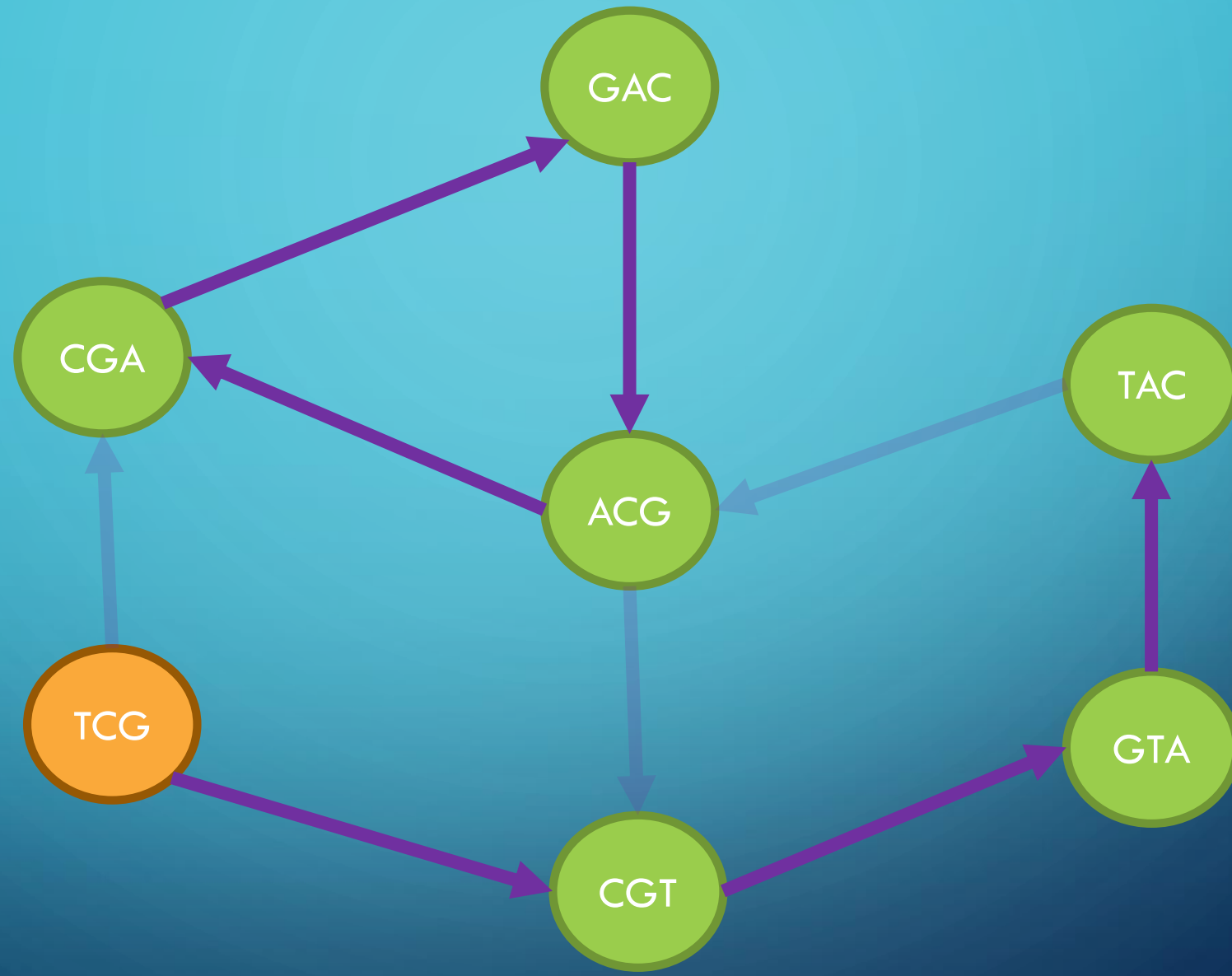
J. BŁAŻEWICZ, P. FORMANOWICZ, M. KASPRZAK, W. T. MARKIEWICZ, and J. WĘGLARZ

Journal of Computational Biology 1999 6:1, 113-123



ZADANIE

Jakiego warunku brakuje?



BRAK OSOBNYCH CYKLI W ROZWIĄZANIU

$$\sum_{v_a \in S} \left(\sum_{i=1}^z x_{ia} + \sum_{j=a}^z x_{aj} \right) < 2|S| \quad \sum_{v_i \in V \setminus S, v_j \in S} x_{ij}, \quad S \subset V \setminus \{v_1\}$$

Ile jest takich nierówności?

Liczba możliwych podziałów zbioru na dwa niepuste podzbiory.

Jest to liczba Stirlinga drugiego rodzaju, a dokładniej $\left\{ \begin{smallmatrix} z \\ 2 \end{smallmatrix} \right\}$

A jak Państwo pamiętają z matematyki dyskretnej:

$$\left\{ \begin{smallmatrix} z \\ 2 \end{smallmatrix} \right\} = 2^{z-1} - 1$$