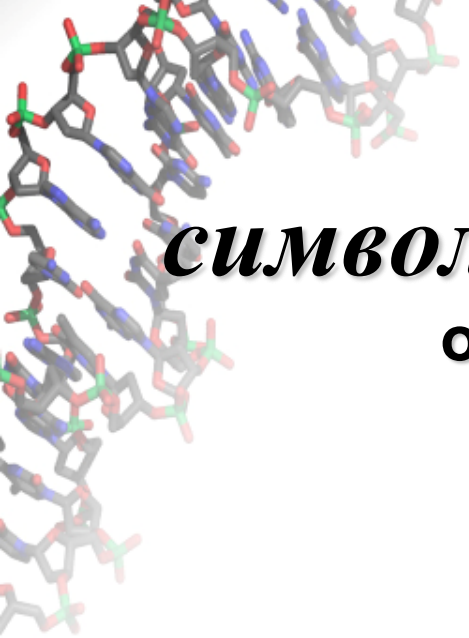


Анализ символьных последовательностей

от биоинформатики до лингвистики

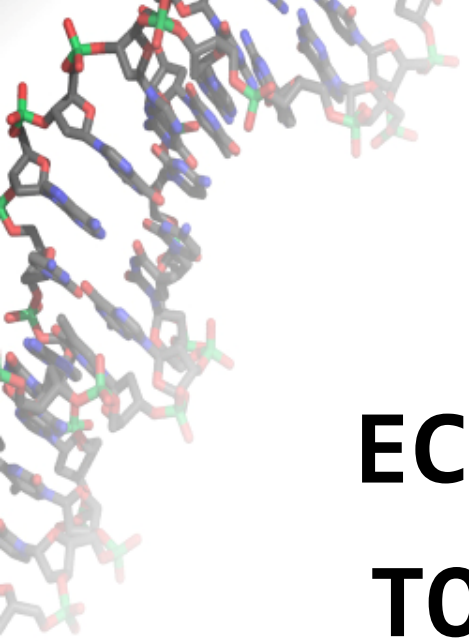
М.А. Ройтберг

Занятие 2, ч.1

ГИПЕРГРАФЫ

Яндекс

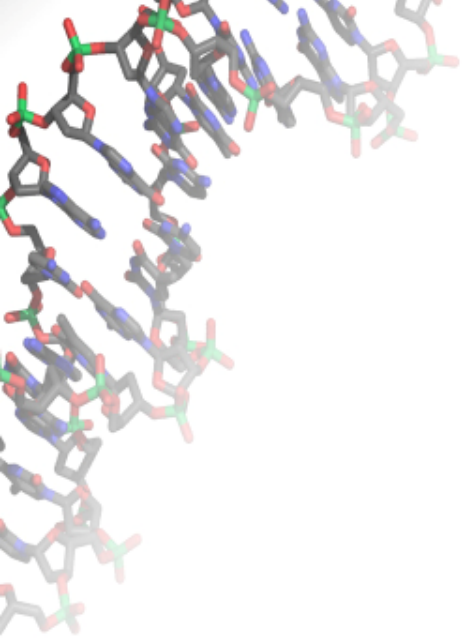
20 февраля 2017



ЕСЛИ графов НЕ хватает

ТО *что???*

ИНАЧЕ см. лекцию 1.



ЕСЛИ *графов* ***НЕ*** *хватает*
ТО *что???*
ИНАЧЕ *см. лекцию 1.*

ЭПИГРАФ:
ЗАДАЧА О ДУГАХ

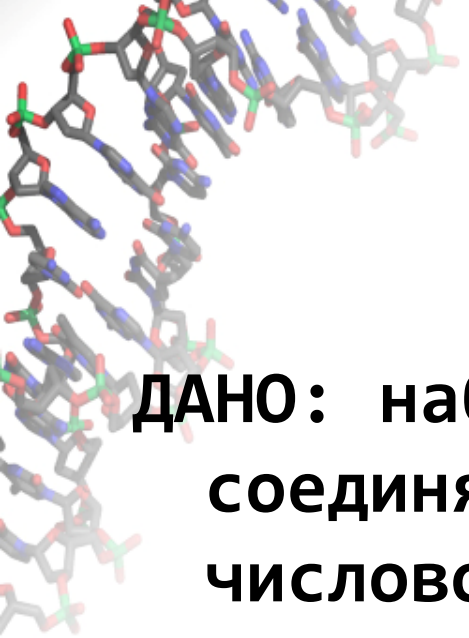


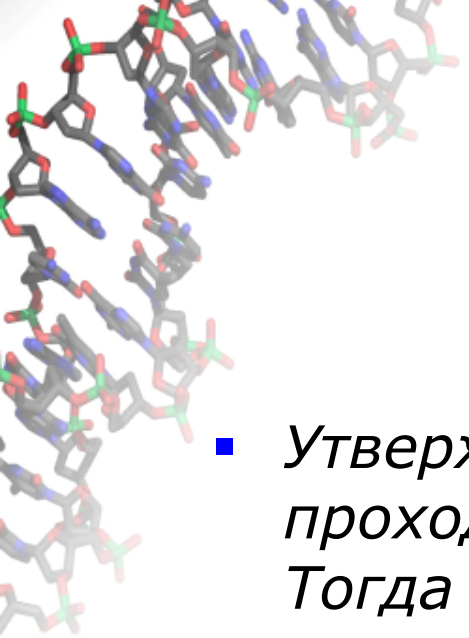
Рисунок на доске

Задача о дугах

ДАНО: набор из *N* дуг, каждая из которых соединяет две целочисленные точки числовой оси *(x, y)*, где $1 \le x < y \le K$. Все дуги расположены над числовой осью.

***Правильная система* дуг – это такая система дуг, в которой никакие дуги не пересекаются.**

НАДО: найти в данном наборе правильную подсистему, содержащую максимально возможное количество дуг



- *Утверждение. Пусть в системе есть ровно k дуг, проходящих через точку 1: $(1, a_1)$, …, $(1, a_k)$. Тогда каждая правильная система принадлежит ровно к одному из следующих классов:*
- *0) не содержит дуги с концом в точке 1;*
- *1) содержит дугу $(1, a_1)$;*
- *…*
- *k) содержит дугу $(1, a_k)$.*

Решение

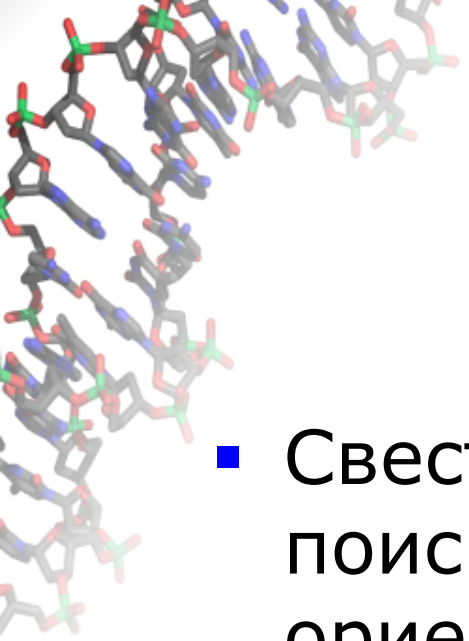
- Находим оптимальную систему для каждого отрезка *[p, q]*, где $1 \le p < q \le K$. Отрезки перебираем в порядке возрастания длин.
- Обозначение: *M(p, q)* – количество дуг в максимальной правильной подсистеме такой, что концы всех дуг в ней принадлежат отрезку *[p, q]*,
- Рекуррентное уравнение: $M(p, q) = max$

$\{ M(p+1, q),$

$max \{1 + M(p+1, t-1) + M(t+1, K)|$

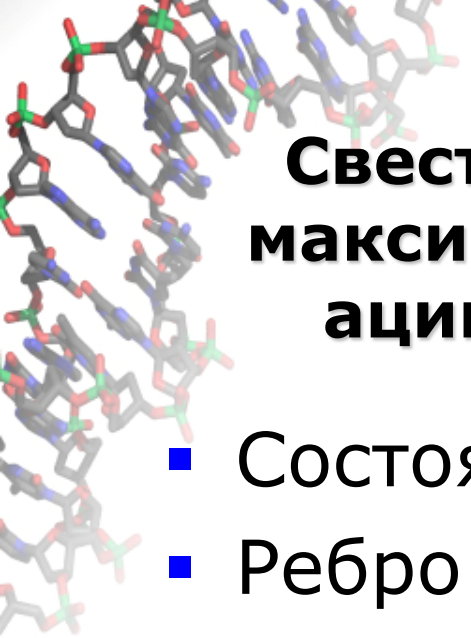
$p < t \le q$ *И* в наборе есть дуга (p, t)

$\}$



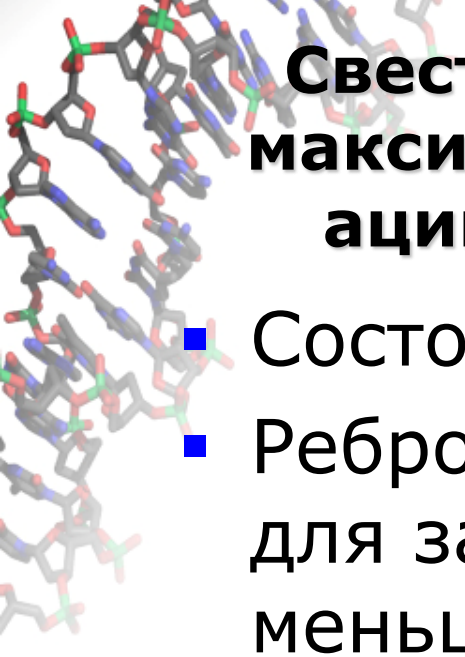
Задание

- Свести задачу о дугах к задаче поиска максимального пути в ориентированном ациклическом графе (см. лекцию 1)



Свести задачу о дугах к задаче поиска максимального пути в ориентированном ациклическом графе (см. лекцию 1)

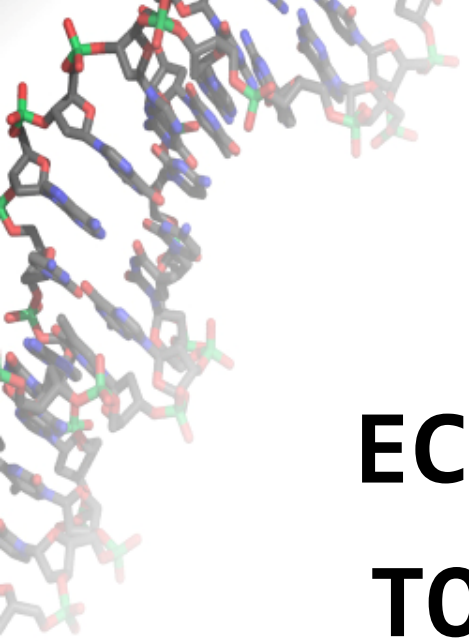
- Состояния – отрезки.
- Ребро соответствует сведению задачи для заданного отрезка к отрезку меньшей длины.
- Как определить ребра?



Свести задачу о дугах к задаче поиска максимального пути в ориентированном ациклическом графе (см. лекцию 1)

- Состояния – отрезки.
- Ребро соответствует сведению задачи для заданного отрезка к отрезку меньшей длины.
- Как определить ребра?

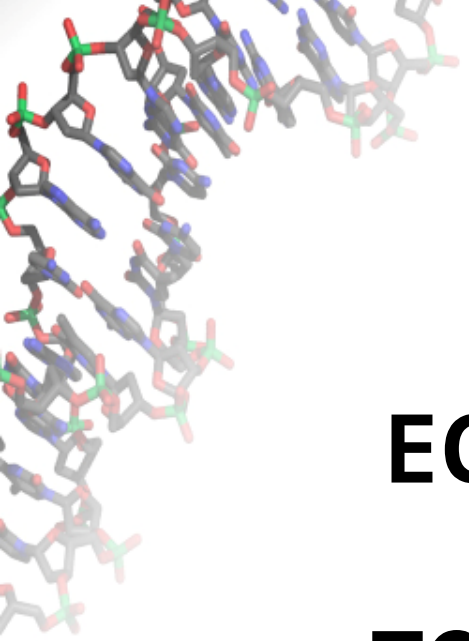
➔ НИКАК! Задача для отрезка *[p, q]* при включении в правильную систему дуги *(p, t)* сводится к решению **двух** задач – для отрезков *[p+1, t-1]* и *[t+1, q]* .



ЕСЛИ графов НЕ хватает

ТО *что???*

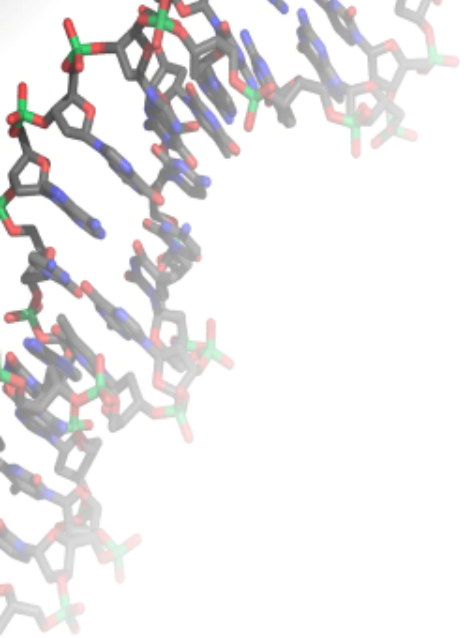
ИНАЧЕ см. лекцию 1.



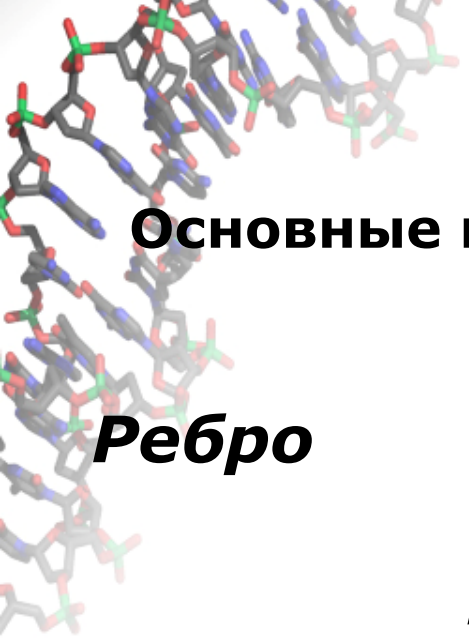
ЕСЛИ графов НЕ хватает

то *ГИПЕРГРАФЫ*

ИНАЧЕ см. лекцию 1.



ГИПЕРГРАФЫ: ЗНАКОМСТВО И ПРИМЕРЫ



Графы и гиперграфы

Основные понятия-1. Гиперребра и гиперпути.

Вершина

Ребро ***Гиперребро***

Вершина-источник

Тупиковая вершина (сток)

Путь ***Гиперпуть***

Инициальный (гипер)путь

Терминальный (гипер)путь

Полный (гипер) путь:

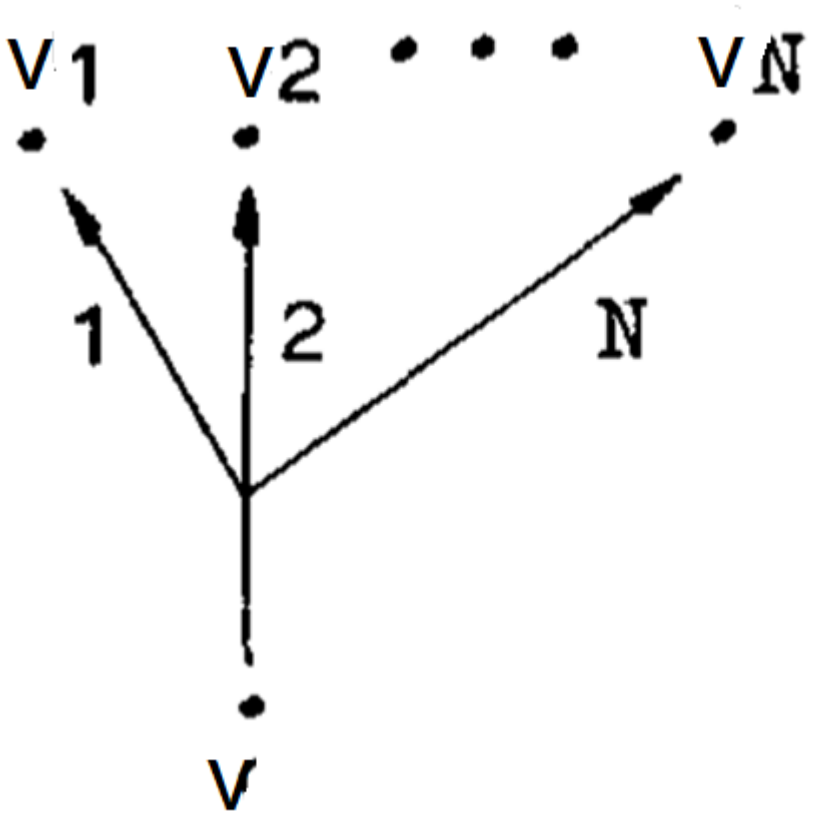
Начальная вершина – источник;

Конечные вершины - тупиковые

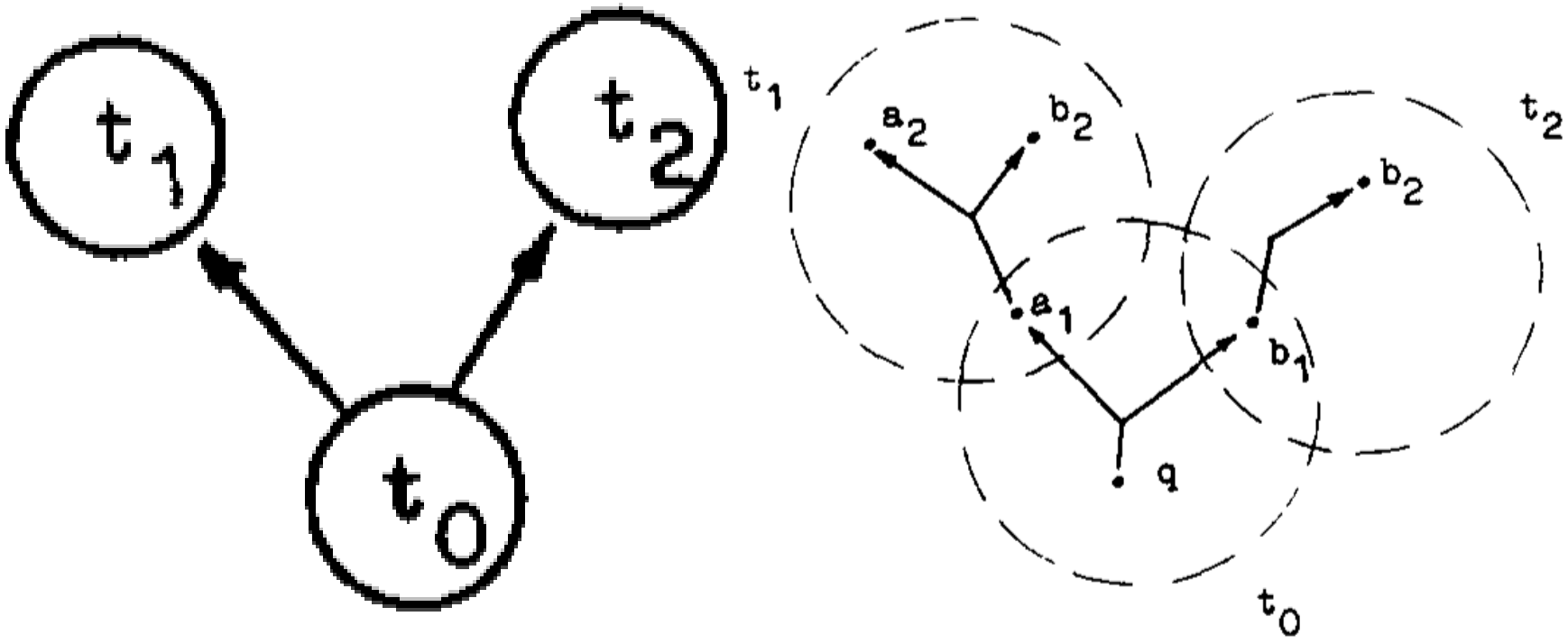
(Ориентированный) гиперграф:
множество вершин и множество гиперребер

Гиперребро *h = (v, (v1, …, vN))*

Ребро e = (v, v1)

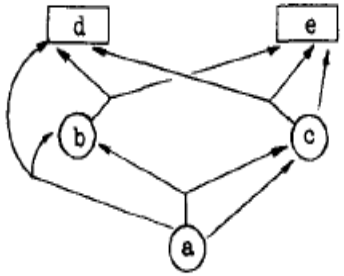


Гиперпуть

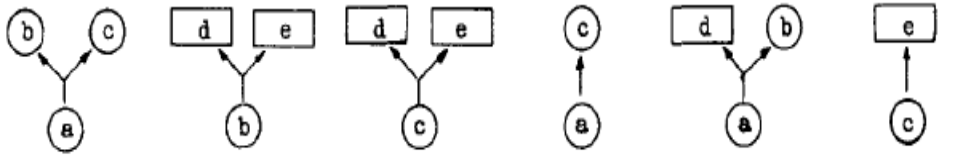


Вес гиперпути – ПРОИЗВЕДЕНИЕ (*) весов гиперребер

(a)



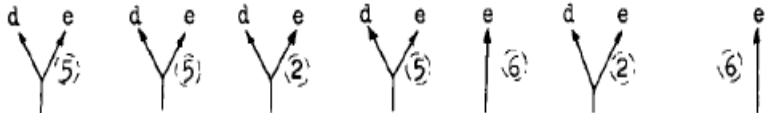
(b)

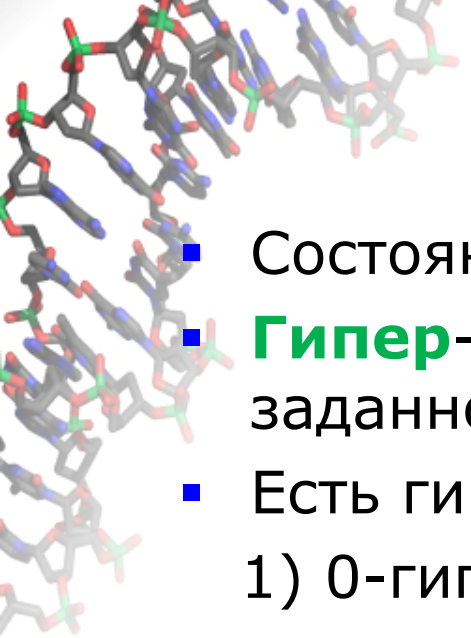


U=(a,{b,c}) V=(b,{d,e}) W=(c,{d,e}) X=(a,{c}) Y=(a,{d,b}) Z=(c,{e})

r(U)=2 r(V)=5 r(W)=2 r(X)=4 r(Y)=1 r(Z)=6

(c)





Пример: задача о дугах

- Состояния – отрезки.
- **Гипер**-ребро соответствует сведению задачи для заданного отрезка к отрезку меньшей длины.
- Есть гиперребра двух типов:

1) 0-гиперребра: <[p, q]; [p+1, q]>

В системе нет дуг с левым концом p

2) дуговые гиперребра ((p, t) – дуга; t ≤ q):

<[p, q]; [p+1, t-1], [t+1, q]>

Включаем в систему дугу (p, t)

Пример: задача о дугах. Уточнение 1

- Состояния – отрезки.
- **Гипер**-ребро соответствует сведению задачи для заданного отрезка к отрезку меньшей длины.
- Есть гиперребра двух типов:

1) 0-гиперребра: <[p, q]; [p+**s**, q]>,

где *s – такое минимальное число, что в системе есть дуга (s, t); t* ≤ q

В системе нет дуг с левым концом p

2) дуговые гиперребра ((p, t) – дуга; t ≤ q):

<[p, q]; [p+1, t-1], [t+1, q]>

Включаем в систему дугу (p, t)

Веса. Повторение (полукольца).

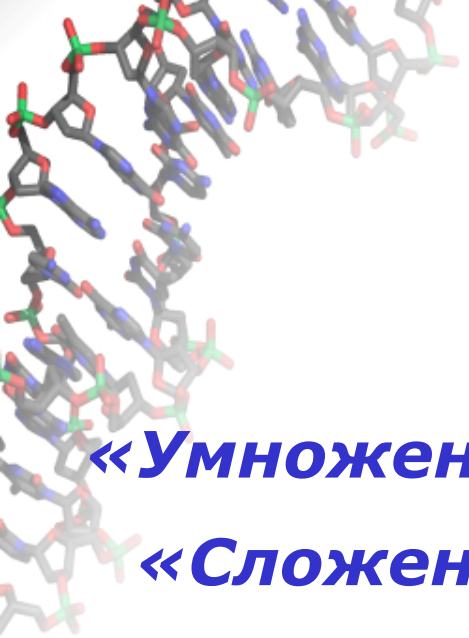
Полукольцо с единицей A – это множество, на котором определены две бинарные всюду определенные операции + и *, удовлетворяющие следующим свойствам:

- операции + и * ассоциативны;
- операция + коммутативна,
- коммутативность операции * не обязательна;
- в *A* есть левый нейтральный элемент *i* относительно операции *;
- операция дистрибутивна относительно операции :

$$a, b, c \in A\ (\ (a + b)*c = (a*c) + (b*c)\ )$$

$$a, b, c \in A\ (\ c*(a + b) = (c*a) + (c*b)\ )$$

Операции + и * обычно называют сложением и умножением.



Графы и гиперграфы

Основные понятия-2. Веса.

Вес гипер(ребра)

«Умножение»: как вычислять вес (гипер)пути

«Сложение»: целевая функция [коммутат.]

Дистрибутивность:

$$a*(b+c) = a*b+a*c;\ (b+c)*a = b*a+c*a$$

Вес пути ***Вес гиперпути***

ПРОБЛЕМА:

НАЙТИ «СУММУ» ВЕСОВ
ВСЕХ ПОЛНЫХ (ГИПЕР)ПУТЕЙ

Пример: задача о дугах.

- Состояния – отрезки.
- **Гипер**-ребро соответствует сведению задачи для заданного отрезка к отрезку меньшей длины.
- Есть гиперребра двух типов:

1) 0-гиперребра: <[p, q]; [p+s, q]>,

где *s – такое минимальное число, что в системе есть дуга (s, t); t* ≤ q

В системе нет дуг с левым концом p

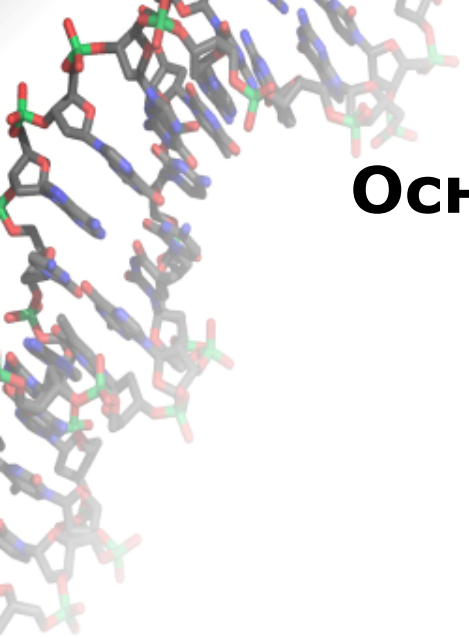
Вес гиперребра: 0

2) дуговые гиперребра ((p, t) – дуга; t ≤ q):

<[p, q]; [p+1, t-1], [t+1, q]>

Включаем в систему дугу (p, t)

Вес гиперребра: 1



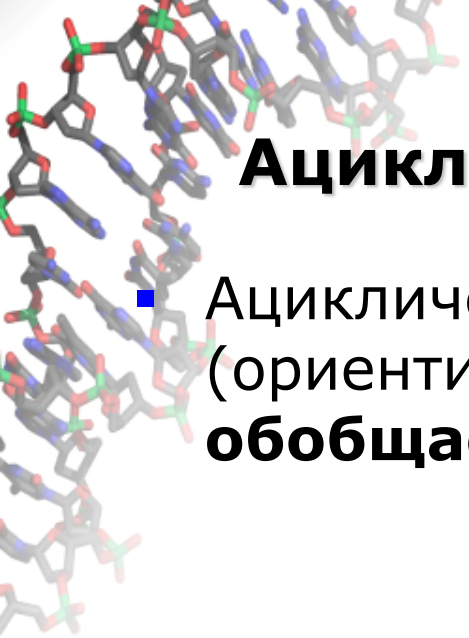
Графы и гиперграфы

Основные понятия-3. Ацикличность.

- Ациклический граф – что это?

Ацикличность графов и гиперграфов. 0

- Ациклический граф – это
  граф, в котором нет (ориентированных) циклов



Ацикличность графов и гиперграфов. 0

- Ациклический граф – это **-** граф, в котором нет (ориентированных) циклов; **- плохо обобщается на гиперграфы**

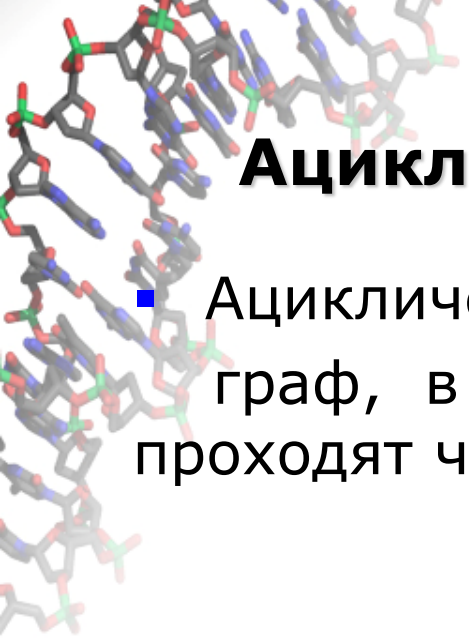
Ацикличность графов и гиперграфов. 1

- Ациклический граф – это граф, в котором нет бесконечных путей (ориентированных)

Ацикличность графов и гиперграфов. 1

- Ациклический граф – это граф, в котором нет бесконечных путей (ориентированных)

Определение 1. Ориентированный гиперграф называется *ациклическим,* если в нем нет бесконечных гиперпутей [лучше – гиперпутей, как угодно большой высоты – в смысле высоты дерева]



Ацикличность графов и гиперграфов. 2

- Ациклический граф – это

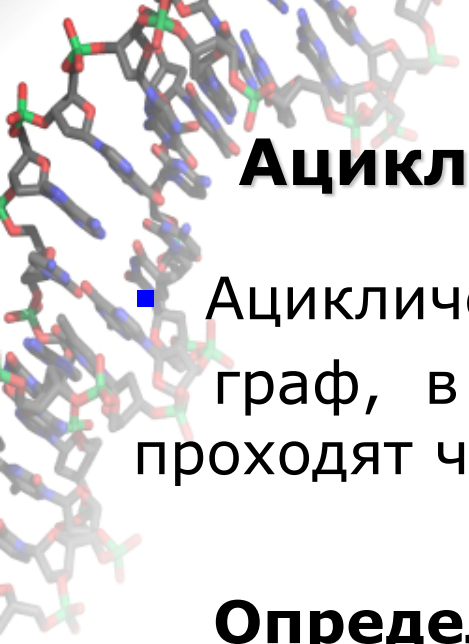
  граф, в котором нет путей, которые дважды проходят через одну и ту же вершину

Ацикличность графов и гиперграфов. 2

- Ациклический граф – это

граф, в котором нет путей, которые дважды проходят через одну и ту же вершину

Определение 2. Ориентированный гиперграф называется *ациклическим,* если в нем нет гиперпутей, в которых два разных узла гиперпути соответствуют одной и той же вершине гиперграфа.



Ацикличность графов и гиперграфов. 2

- Ациклический граф – это

  граф, в котором нет путей, которые дважды проходят через одну и ту же вершину

Определение 2. Ориентированный гиперграф называется *ациклическим,* если в нем нет гиперпутей, в которых два разных узла гиперпути соответствуют одной и той же вершине гиперграфа.

- *Для задачи о дугах это условие, очевидно, выполнено*

Ацикличность графов и гиперграфов. 2

- Ациклический граф – это граф, в котором нет путей, которые дважды проходят через одну и ту же вершину

Определение 2. Ориентированный гиперграф называется *ациклическим,* если в нем нет гиперпутей, в которых два разных узла гиперпути соответствуют одной и той же вершине гиперграфа.

- **Определение 2б.** ***??? Можно ли обобщить иначе?***

Ацикличность графов и гиперграфов. 2

- Ациклический граф – это граф, в котором нет путей, которые дважды проходят через одну и ту же вершину

Определение 2а. Ориентированный гиперграф называется *ациклическим,* если

в нем нет гипер-путей, в которых два разных узла соответствуют одной и той же вершине

Определение 2б. Ориентированный гиперграф называется *ациклическим,* если в нем нет гипер-путей, в которых два разных узла, которые ***расположены на одной траектории*** в этом гипер-пути и помечены одной и той же вершиной

- *[Рисунок на доске]*

Ацикличность гиперграфов. Итоги

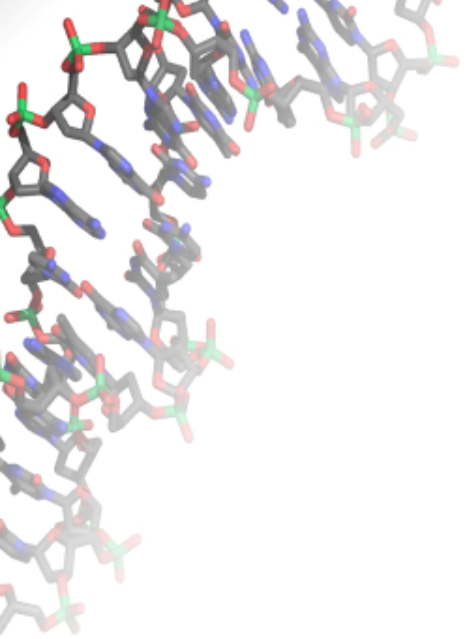
- Ориентированный гиперграф называется *ациклическим,* если в нем …
- **[1]** … нет бесконечных гипер-путей
- **[2а]** … нет гипер-путей, в которых два разных узла соответствуют одной и той же вершине
- **[2б]** … нет гипер-путей, в которых два разных узла, которые ***расположены на одной траектории*** в этом гипер-пути и помечены одной и той же вершиной

Как эти определения связаны между собой?

Ацикличность гиперграфов. Итоги

- Ориентированный гиперграф называется *ациклическим*, если в нем …
- **[1]** … нет бесконечных гипер-путей
- **[2а]** … нет гипер-путей, в которых два разных узла соответствуют одной и той же вершине
- **[2б]** … в нем нет гипер-путей, в которых два разных узла, которые ***расположены на одной траектории*** в этом гипер-пути и помечены одной и той же вершиной

[2а] => [2б] ≡ 1



ГИПЕРГРАФЫ: ФОРМАЛЬНЫЕ ОПРЕДЕЛЕНИЯ

1. Гипергpаф, гиперребро

Ориентированный гипергpаф (ОГ-граф) над полукольцом A – это четверка $G = (V, q_0, E, f)$, где

V – конечное множество вершин;

$q_0 \in V$ – *стартовая* вершина;

E – конечное множество *гиперребер*,

т.е. пар вида (v, W), где

$v \in V$,

W – упорядоченный список вершин из V;

$f: E \to A$ – функция пометок на ребрах.

Вершина v называется *начальной* вершиной гиперребра $e=(v, W)$,

вершин**ы** из W называются *конечн**ыми*** вершин**ами** этого гиперребра;

значение $f(e) \in A$ называется *весом* гиперребра e

2. Стартовые и терминальные вершины.

. Вершина v ОГ-графа G называется *стартовой*, если в G нет гиперребра, в котором v является начальной вершиной.

Вершина v ОГ-графа G называется *терминальной*, если в G нет гиперребра, в котором v является конечной вершиной.

2. Стартовые и терминальные вершины.

. Вершина v ОГ-графа G называется *стартовой*, если в G нет гиперребра, в котором v является конечной вершиной.

Вершина v ОГ-графа G называется *терминальной*, если в G нет гиперребра, в котором v является начальной вершиной.

Аналог пути в графе - гиперпуть в гиперграфе

3. Гиперпуть.

Гиперпуть в ОГ-графе G = (V, q0, E, f) –это такое конечное *упорядоченное помеченное* дерево T, что

1) каждый узел дерева T помечен вершиной ОГ-графа G; одна вершина v из V может соответствовать нескольким узлам дерева T.

2) каждому внутреннему узлу r дерева T соответствует гиперребро e = (v, W), причем узел r помечен вершиной v, а список пометок в сыновьях узла r, взятых в порядке, предписанном деревом T, совпадает со списком W.

3. Гиперпуть.

Гиперпуть в ОГ-графе G = (V, q0, E, f) –это такое конечное упорядоченное помеченное дерево T, что

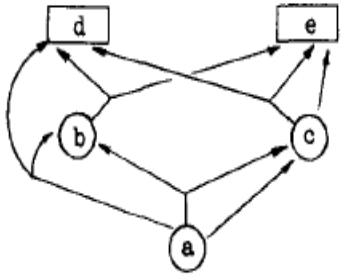
1) каждый узел дерева T помечен вершиной ОГ-графа G; одна вершина v из V может соответствовать нескольким узлам дерева T.

2) каждому внутреннему узлу r дерева T соответствует гиперребро e = (v, W), причем узел r помечен вершиной v, а список пометок в сыновьях узла r, взятых в порядке, предписанном деревом T, совпадает со списком W.

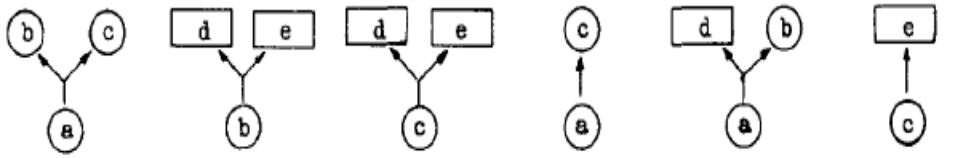
Корневое гиперребро гиперпути *T* – гиперребро, соответствующее корню *T.*

Начальная (корневая) вершина – вершина, соответствующая корню *T.*

(a)



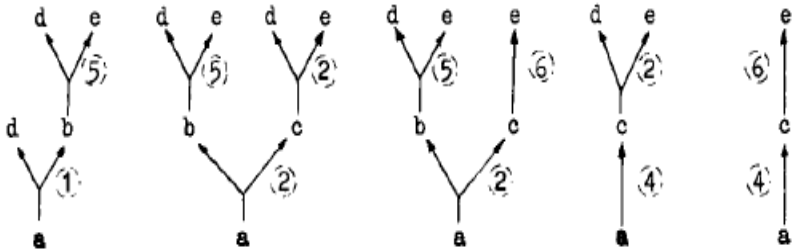
(b)



U=(a,{b,c}) V=(b,{d,e}) W=(c,{d,e}) X=(a,{c}) Y=(a,{d,b}) Z=(c,{e})

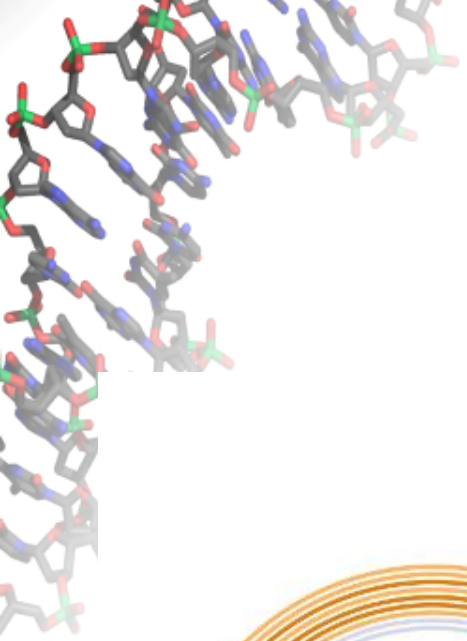
r(U)=2 r(V)=5 r(W)=2 r(X)=4 r(Y)=1 r(Z)=6

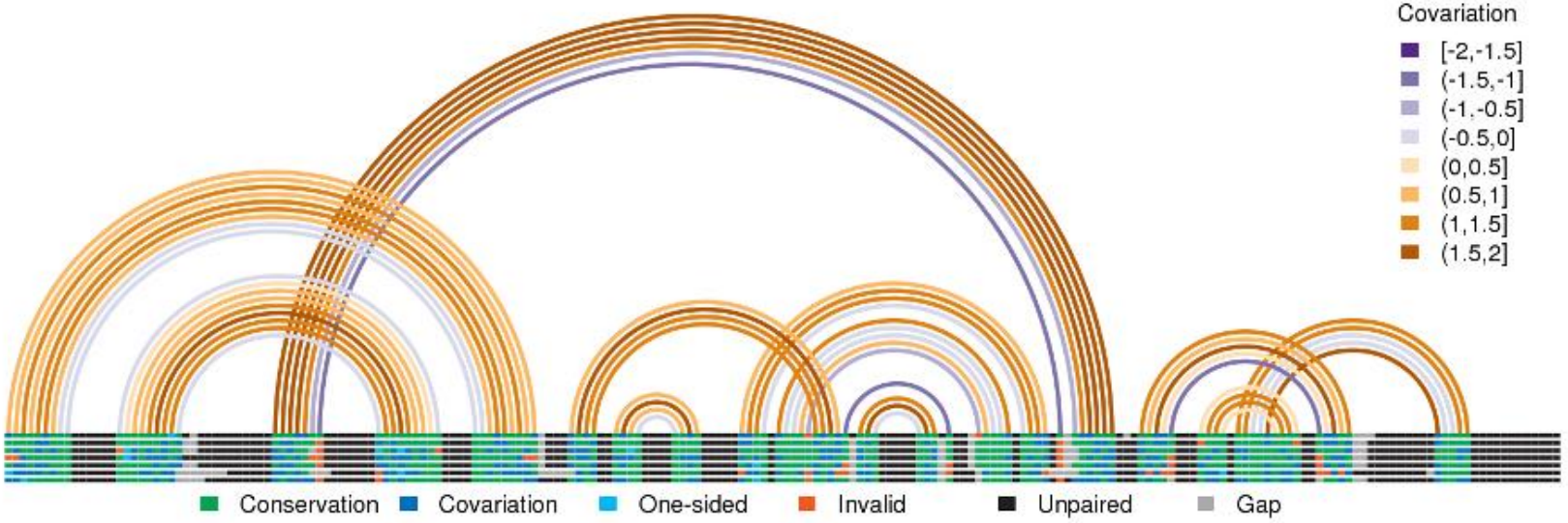
(c)

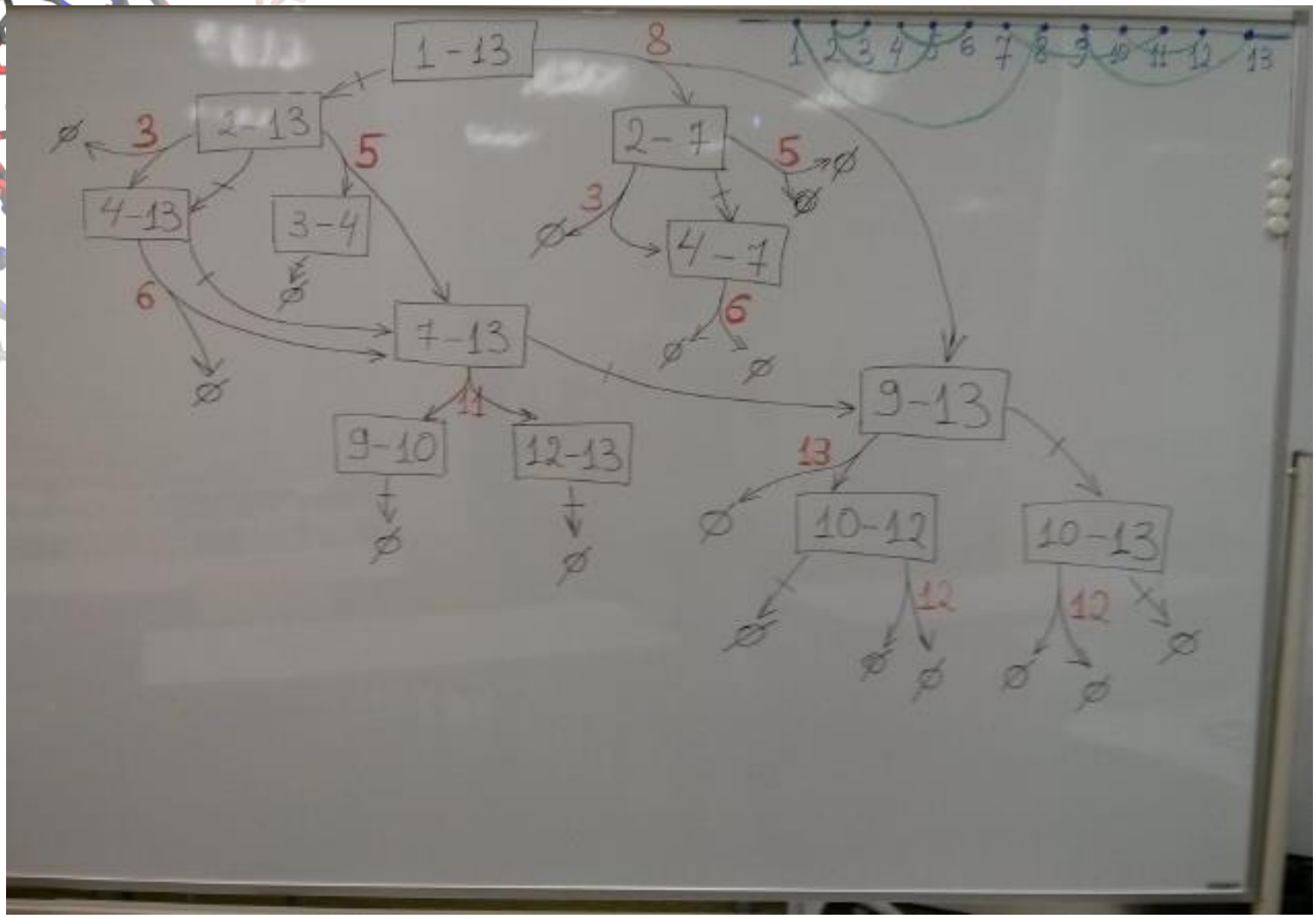


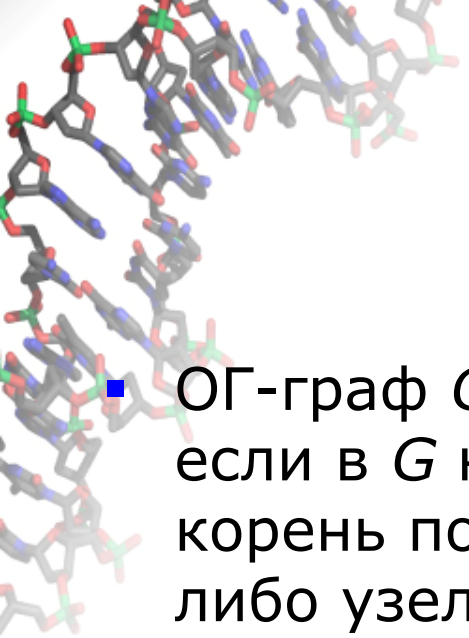
(+,min):+:6 (+,min):+:9 (+,min):+:13 (+,min):+:6 (+,min):+:10

(x,+):x:5 (x,+):x:20 (x,+):x:60 (x,+):x:8 (x,+):x:24




Covariation
[-2,-1.5]
(-1.5,-1]
(-1,-0.5]
(-0.5,0]
(0,0.5]
(0.5,1]
(1,1.5]
(1.5,2]
Conservation
Covariation
One-sided
Invalid
Unpaired
Gap


1-13
8
1 2 3 4 5 6 7 8 9 10 11 12 13
2-13
3
5
2-7
5
3
4-13
3-4
4-7
6
6
7-13
11
9-13
9-10
12-13
13
10-12
10-13
12
12



4. Ацикличность

- ОГ-граф $G = (V, q_0, E)$ называется *ациклическим*, если в G не существует гиперпути, в котором корень помечен той же вершиной, что и еще какой-либо узел этого потока.
- Очевидно, в *ациклическом* ОГ-графе существует лишь конечное число различных потоков.

4. Ацикличность

- ОГ-граф $G = (V, q_0, E)$ называется *ациклическим*, если в G не существует гиперпути, в котором корень помечен той же вершиной, что и еще какой-либо узел этого потока.
- Очевидно, в *ациклическом* ОГ-графе существует лишь конечное число различных потоков.
- **Это – слабое определение ацикличности [2б=1]**

- Ориентированный гиперграф называется *ациклическим,* если в нем …
- **[1]** … нет бесконечных гипер-путей
- **[2а]** … нет гипер-путей, в которых два разных узла соответствуют одной и той же вершине
- **[2б]** … в нем нет гипер-путей, в которых два разных узла, которые ***расположены на одной траектории*** в этом гипер-пути и помечены одной и той же вершиной

[2а] => [2б] ≡ 1

ОГ-граф $G = (V, q_0, E)$ называется *ациклическим,* если в G не существует гиперпути, в котором корень помечен той же вершиной, что и еще какой-либо узел этого потока.

5. Вес гиперпути: рекурсивное определение

- Пусть $G = (V, q_0, E, f)$ – ОГ-граф; T – гиперпуть в G. Определим *вес $R(T)$ гиперпути T* следующим образом.
- Если гиперпуть T состоит из единственного узла, то $R(T) = i$, где i – нейтральный элемент относительно операции * (умножения).
- Пусть корень T помечен гиперребром $e = (v, W)$; $x_1, \ldots, x_N$ – упорядоченный список сыновей корня дерева T; T_k – поддерево дерева T с корнем в узле x_k ($k = 1, \ldots, N = |W|$).Тогда

$$R(T) = f(e) * R(T_1) * \ldots * R(T_N)$$

5. Вес гиперпути: явное определение

Пусть $G = (V, q_0, E, f)$ – ОГ-граф; T – гиперпуть в G. Пусть корень T помечен гиперребром $e = (v, W)$; $x_1, \ldots, x_N$ – упорядоченный список сыновей корня дерева T; T_k – поддерево дерева T с корнем в узле x_k ($k = 1, \ldots, N = |W|$).Тогда

$$R(T) = f(e)* R(T_1) *\ldots \ *R(T_N)$$

Вес гиперпути Т это произведение весов гиперребер (в смысле операции *), соответствующих узлам Т, причем порядок перемножения соответствует *левому обходу* дерева Т в глубину.

6. Терминальные и полные пути

- Гиперпуть *T* в ОГ-графе *G* называется *терминальным,* если все его листья соответствуют терминальным вершинам ОГ-графа *G*.

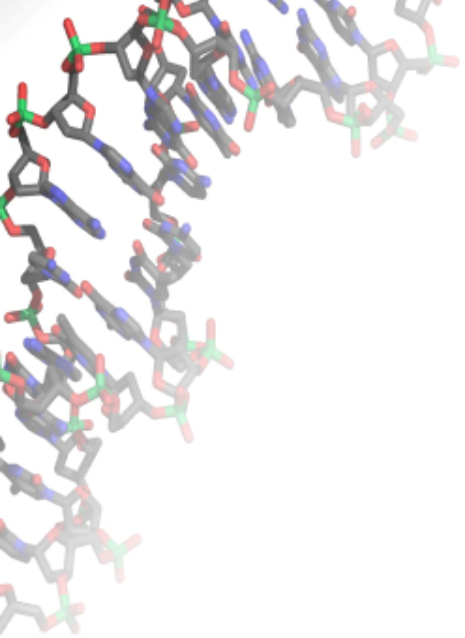
- Терминальный гиперпуть называется *полным,* если его корню соответствует стартовая вершина ОГ-графа *G*.

7. Основная задача на гиперграфах

- Гиперпуть *T* в ОГ-графе *G* называется *терминальным,* если все его листья соответствуют терминальным вершинам ОГ-графа *G*.
- Терминальный гиперпуть называется *полным,* если его корню соответствует стартовая вершина ОГ-графа *G*.

- Определение 6.2. *Обобщенной статистической суммой* ОГ-графа *G* называется сумма (в смысле операции +) *S(G)* весов всех его полных гиперпутей.

- ***Задача.*** **Найти обобщенную статистическую сумму для заданного ациклического ОГ-графа G**



РЕШЕНИЕ ОСНОВНОЙ ЗАДАЧИ

Найти обобщенную статистическую сумму для заданного ациклического ОГ-графа G

- *Обозначения. Пусть* $v \in V$.
- G_v – это подграф ОГ-графа *G,* порожденный всеми вершинами *G,* достижимыми из *v,* т.е. вершинами, которые встречаются в гиперпутях, корень которых соответствует вершине *v.*
- *S(V) – значение обобщенной статистической суммы для гиперграфа* G_v *– т.е. сумма весов всех полных путей в* G_v

Вычисление $S(V_0)$ ОГ-графа $G(V_0, q_0, E, f)$

Общее описание алгоритма

- Перебираем вершины *ОГ-графа G* в обратном топологическом порядке и для каждой вершины $v \in V$ вычисляем S(V).
- Для любой терминальной вершины *v* положим *S(v)=i,* где *i* – правый нейтральный элемент относительно «умножения».
- ***Как вычислять S(v), если v - не терминальная вершина?***

Вычисление S(v) для внутренней вершины v.

- $e_1 = \langle v; x_1, \ldots, x_k \rangle$, $e_2 = \langle v; y_1, \ldots, y_m \rangle$,
- $e_3 = \langle v; z_1, \ldots, z_n \rangle$ - все гиперребра с начальной вершиной *v*
- ***H(w)*** – множество всех гиперпутей с начальной вершиной *w*
- $\boldsymbol{H_e(g)}$ – множество всех гиперпутей с корневым гиперребром *g*
  - $\boldsymbol{H(v) = H_e(e_1) + H_e(e_2) + H_e(e_3)}$
- *Пусть* $\boldsymbol{S_e(e_r)}$ *обозначает сумму весов всех гиперпутей из* $H_e(e_r)$, $r = 1, 2, 3$.
  - $\boldsymbol{S(v) = S_e(e_1) + S_e(e_2) + S_e(e_3)}$

Вычисление $S_e(e_1)$ - 1

- $e_1 = \langle v; x_1, \ldots, x_k \rangle$; $H_e(e_1)$ – множество всех гиперпутей с корневым гиперребром g
- $S_e(e_1)$ - *сумма весов всех гиперпутей из* $H_e(e_1)$.
  - $\boldsymbol{S_e(e_1) = SUM(R(T) \mid T \in H_e(e_1))}$
- *Пусть T – гиперпуть из* $H_e(e_1)$; Пусть корень T помечен гиперребром $e = (v, W)$; T_j – поддерево дерева T с корнем j-м сыне корня T. *Тогда* T_j лежит в $H(x_j)$ и
  - $\boldsymbol{R(T) = f(e_1)*R(T_1)*\ldots*R(T_k)}$
- Отсюда
- $\boldsymbol{S_e(e_1) = SUM(f(e_1)*R(T_1)*\ldots*R(T_k) \mid T \in H_e(e_1)) =}$
- $\boldsymbol{= f(e_1)* SUM(R(T_1)*\ldots*R(T_k) \mid T \in H_e(e_1))}$

Вычисление $S_e(e_1)$ - 2

- $S_e(e_1) = SUM(f(e_1)*R(T_1)*...*R(T_k) | T \in H_e(e_1)) =$
- $= f(e_1)* SUM(R(T_1)*...*R(T_k) | T \in H_e(e_1))$

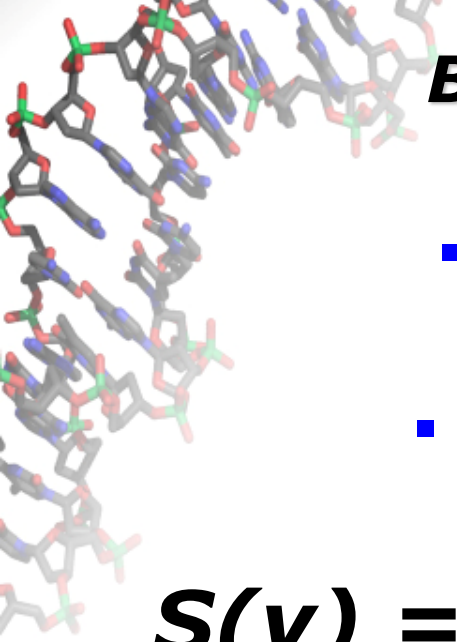
$$SUM(R(T_1)*...*R(T_k) | T \in H_e(e_1)) =$$

$$= SUM(R(X_1)*...*R(X_k) | X_1 \in H(x_1); ...; X_k \in H(x_k)) =$$

$$= SUM(R(X_1) | X_1 \in H(x_1))* ...* SUM(R(X_k) | X_1 \in H(x_k)) =$$

- $= S(x_1)*...* S(x_k)$

- **$S_e(e_1) = f(e_1)* S(x_1)*...* S(x_k)$**



Вычисление S(v) для внутренней вершины v. Окончание

- $S(v) = S_e(e_1) + S_e(e_2) + S_e(e_3)$

- $S_e(e_1) = f(e_1) * S(x_1) * \ldots * S(x_k)$

$$S(v) = f(e_1) * S(x_1) * \ldots * S(x_k) + f(e_2) * S(y_1) * \ldots * S(y_m) + f(e_3) * S(z_1) * \ldots * S(z_n)$$

Время работы алгоритма

Предполагаем: время перехода к обработке очередной вершины и время доступа к ранее вычисленным суммам не зависят от размера графа.

$$S(v) = f(e_1) * S(x_1) * \ldots * S(x_k) + f(e_2) * S(y_1) * \ldots * S(y_m) + f(e_3) * S(z_1) * \ldots * S(z_n)$$

T – сумма времен обработки гиперребер. Время обработки гиперрребра ~ количества его концевых вершин => ***T*** ~ суммарное количество концевых вершин всех гиперребер.

- Если степени всех гиперребер ограничены, то

T ~ количество гиперребер.

ОСТАЛОСЬ:

- Что делать, если обратный топологический порядок на множестве вершин неизвестен?
- Как вычислить сумму весов всех гиперпутей, содержащих данное гиперребро или данную вершину?