Discrete Biological Models
(Modelli Biologici Discreti)

Zsuzsanna Lipták

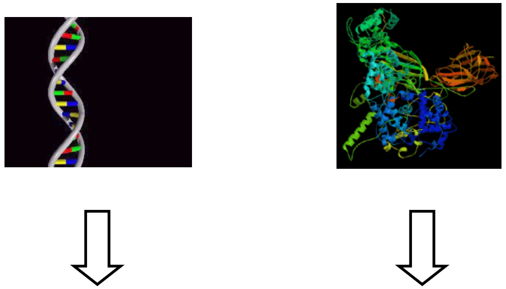
Laurea Triennale in Bioinformatica
a.a. 2014/15, fall term
Università di Verona

Lecture 1 (1 Oct. 2014)

What is a discrete biological model?

- model: a simplified description, esp. a mathematical one, of a system or process, to assist calculations and predictions
  - Oxford Dictionary
- So biological modeling is the act of translating a phenomenon from biology into mathematical language. This allows us to apply known methods to the original problem.
- Note that modeling always involves simplification.

Modeling molecules as strings



...AACAGTACCATGCTA...
...TTGTCATGGTACGAT...

...SLDILRRKSLMNYWL...

What is a discrete biological model?

Uses discrete mathematics:

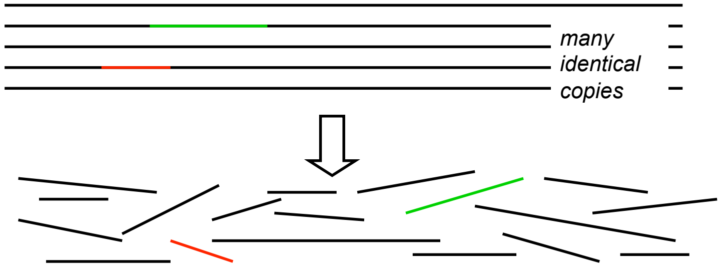
- natural numbers $\mathbb{N} = \{0,1,2,3,...\}$ or integers $\mathbb{Z} = \{...,-3,-2,-1,0,1,2,3,...\}$
- graphs, trees
- permutations
- strings/sequences
- combinatorics (counting, enumerating a finite number of discrete objects)

Example 1: Shotgun-Sequencing of the human genome

```
...AACAGTACCATGCTAGGTCAATCGAG...
...TTGTCATGGTACGATCCAGTTAGCTC...
```

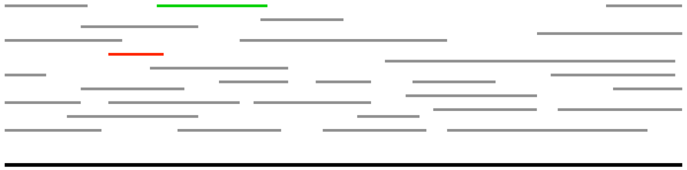
Ex 1: Shotgun-Sequencing

- typical DNA-molecules are several 100´000 bp‘s long, but only pieces of length 200-700 can be sequenced (Sanger)
- use shotgun-method



Shotgun-Sequencing (2)

Goal: Reconstruct original string



Shotgun-Sequencing (3)

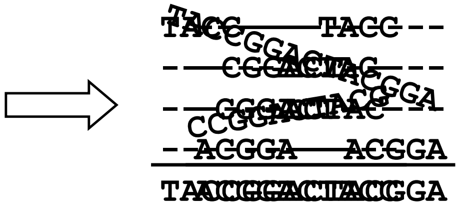
an example:

```
ACCGT              --ACCGT--
CGTGC              ----CGTGC
TTAC               TTAC-----
TACCGT             -TACCGT--
                   TTACCGTGC
```

Shotgun-Sequencing (4)

another example:

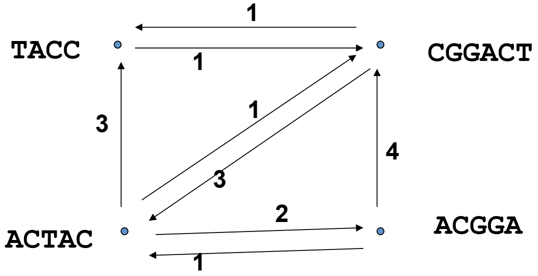
TACC
ACTAC
CGGACT
ACGGA



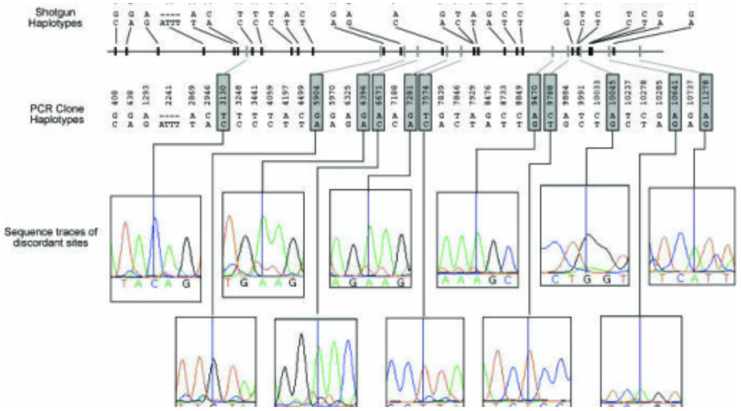
Which solution is better?
How can we find all solutions?

Shotgun-Sequencing: Model

overlap-graph (directed, weighted graph):



Example 2: Haplotyping



© NCBI

Example 2: Haplotyping

- SNPs (single nucleotide polymorphisms)

0	1	1	0	0	1	haplotype
0	0	1	1	0	1	haplotype
0	2	1	2	0	1	genotype

0	1	1	1	0	1
0	0	1	0	0	1
0	2	1	2	0	1

Example 2: Haplotyping

- given: n genotypes (n individuals)
- find: fewest possible haplotypes that explain the genotypes

02120
22110
20120

⇨

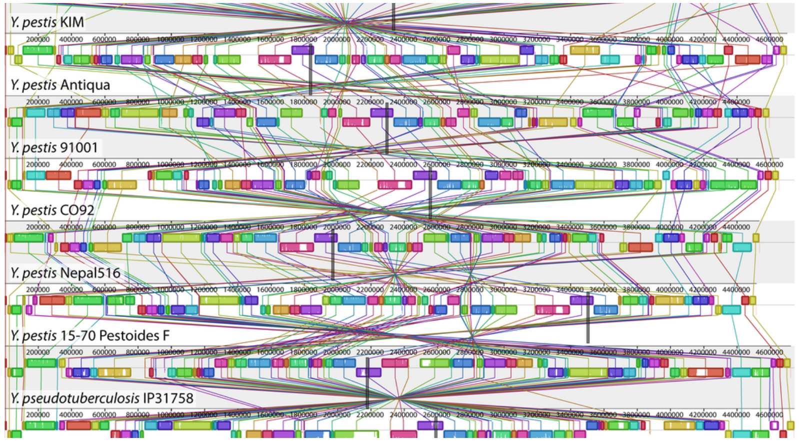
00100
01110
10110

00100
<u>01110</u>
02120

01110
<u>10110</u>
22110

00100
<u>10110</u>
20120

Example 3: genome rearrangements



© Darling, Miklós, Ragan

Ex. 3: genome rearrangements (2)

human 1 2 3 4 5 6 (gene 1, gene 2, ...)

⇧ how do we get from mouse to human?

mouse 1 3 4 6 5 2

Ex. 3: genome rearrangements (3)

human $1\ \underline{2\ 3}\ 4\ 5\ 6$ (gene 1, gene 2, ...)

$1\ 3\ \underline{2\ 4}\ 5\ 6$

$1\ 3\ 4\ \underline{2\ 5\ 6}$

mouse 1 3 4 6 5 2

Ex. 3: genome rearrangements (4)

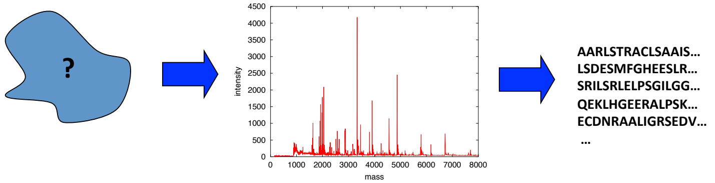
- given: a permutation of {1,2,...,n}
- find: fewest number of reversals to get to the identity permutation

1 3 4 6 5 2 ⇒ 1 2 3 4 5 6

(“Sorting by reversals”)

Example 4: Mass decomposition

mass spectrometry:



AARLSTRACLSAAIS...
LSDESMFGHEESLR...
SRILSRLELPSGILGG...
QEKLHGEERALPSK...
ECDNRAALIGRSEDV...
...

unknown molecular mixture

(DNA, protein, metabolites...)

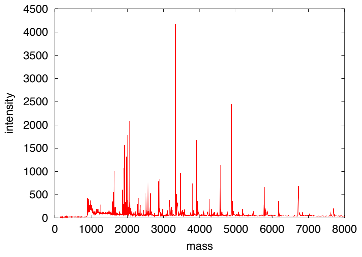
mass spectrum

identification

(names, sequences, data base identifiers, molecular structure...)

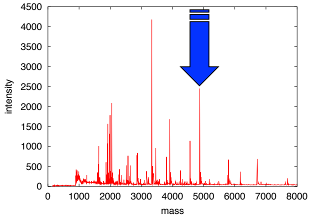
Ex. 4: Mass decomposition (2)

- input: unknown molecular mixture (sample)
- output: list of masses of the sample molecules (mass in Da, intensity)
  actually: m/z = mass over charge
- intensity proportional to abundance: how often that mass was measured (ideally!)



Ex. 4: Mass decomposition (3)

- Given: query mass *M* (in Da)
- Known: What type of molecules are in sample? (DNA, protein, ...)
- Question: What molecules can have this mass?



amino acids (*k* = 20)		nucleotides (*k* = 4)		“bioatoms” (*k* = 6)			
Ala (A)	71.079 Da	A	313.058	C	12.0	O	15.995
Arg (R)	156.101	C	289.046	H	1.008	P	30.074
Asn (N)	114.043	G	329.053	N	14.003	S	31.972
...		T	304.046				

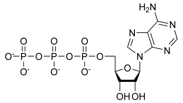
Ex. 4: Mass decomposition (4)

NB:

- output will be: composition/sum formula (not: sequence or molecular structure!)

EGAEEYSSFL ⇒ $A_1E_3G_1L_1F_1S_2Y_1$

AACGTAGGAA ⇒ $A_4C_1G_3T_1$



⇒ $C_{10}H_{16}N_5O_{13}P_3$ (sum formula)

- take into account error (measurement, computation)

Ex. 4: Mass decomposition (5)

Money Changing Problem (MCP):

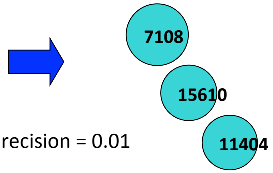
- Given:
  - *k* coin denominations 4 5 7
  - postive integer *M*
- Question: How can we make change for *M*?

19 = 3 · 5 + 1· 4 = 3 · 4 + 1· 7 = 2· 7 + 1· 5

Ex. 4: Mass decomposition (6)

Translating the MS problem into MCP:

Ala (A)	71.079 Da
Arg (R)	156.101
Asn (N)	114.043
...	



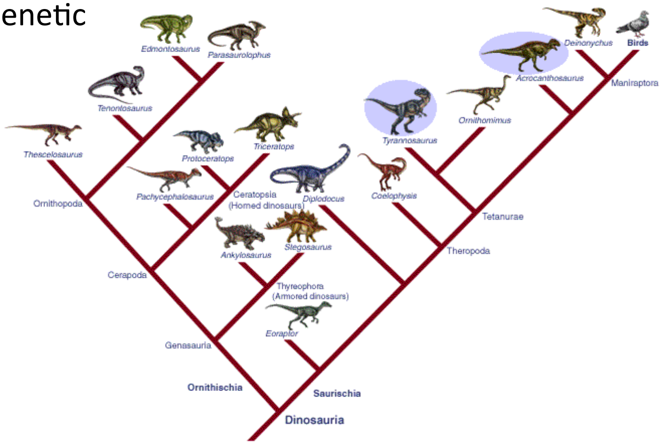
precision = 0.01

etc.

Given: query M, error bound ε.
Compute all decompositions of masses between M - ε and M + ε (scaled to integers with factor 1/precision).

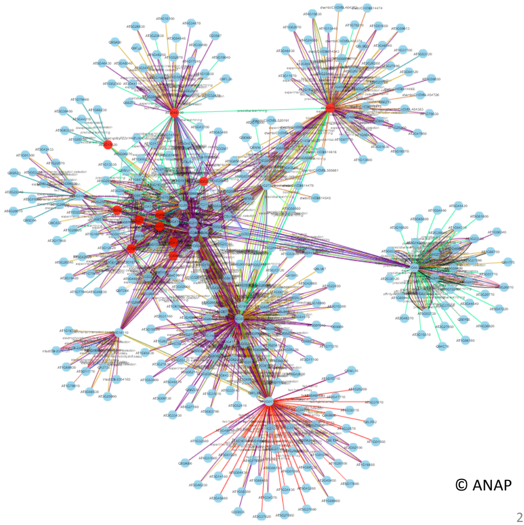
More discrete models in bioinformatics

Phylogenetic trees



More discrete models in bioinformatics

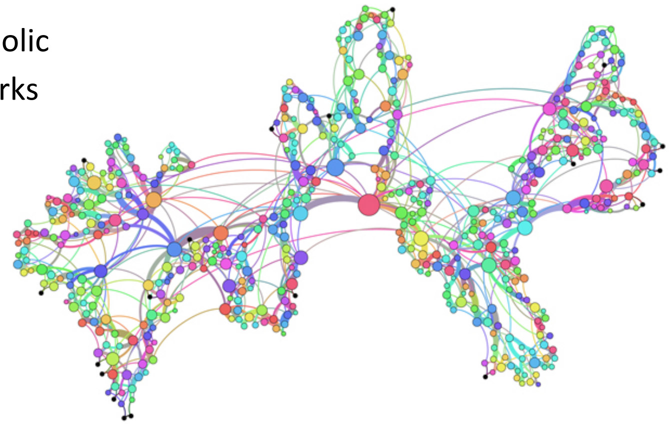
Protein interaction networks



© ANAP

More discrete models in bioinformatics

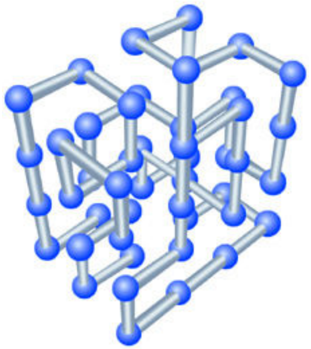
Metabolic networks



© Simons Foundation

More discrete models in bioinformatics

Discrete models for protein folding (H-P model)



© Berkeley Lab

More discrete models in bioinformatics

and, and, and ...

End of introduction.

Now it’s time to start working.