Biological applications of the string searching algorithms

Lecture 9

The sequence databases and their search

- Databases contain a systematically organized data about
  - genome sequences,
  - c-DNAs,
  - protein sequences,
  - promoters,
  - short motives,
  - repeats,
  - metabolic networks

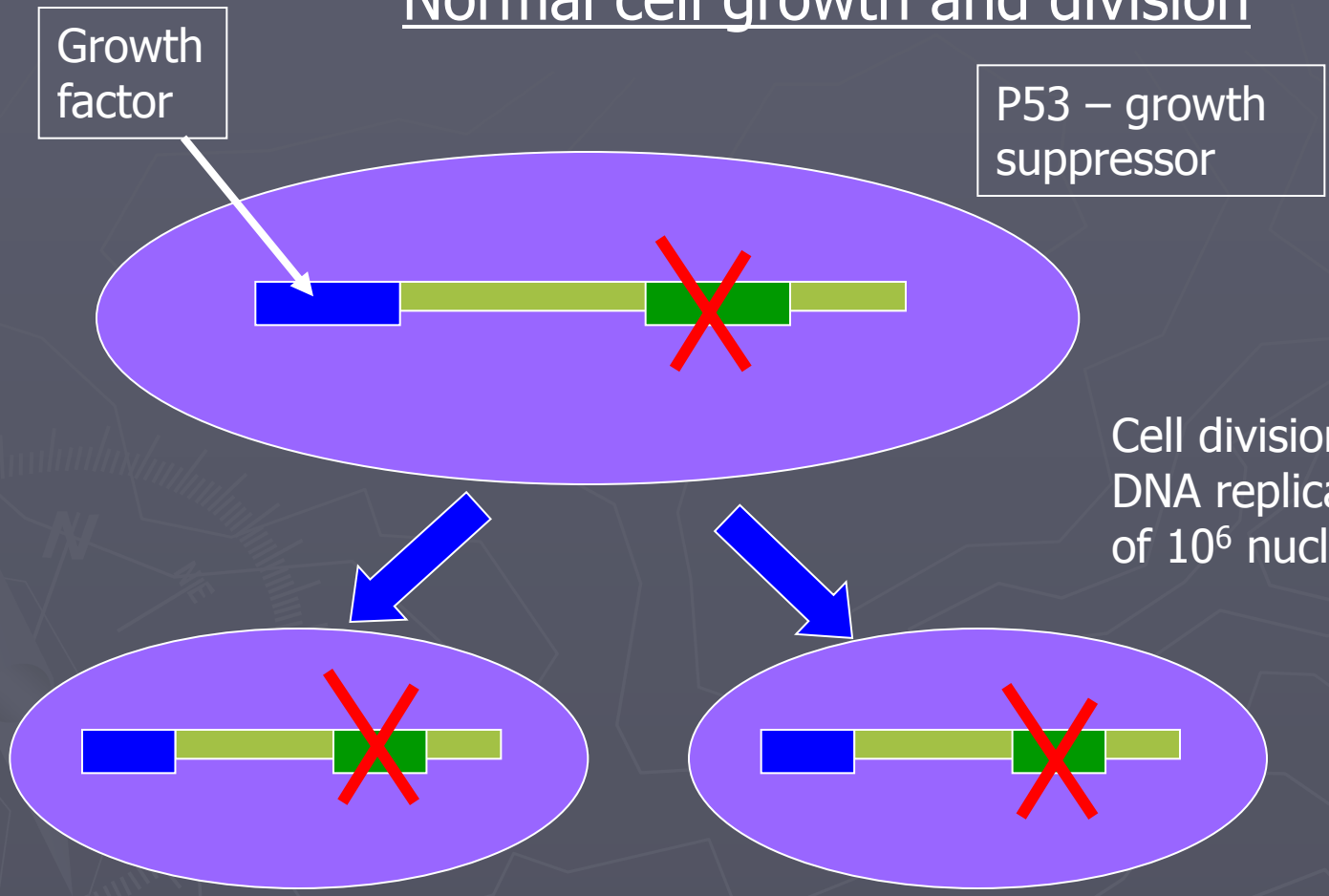
  …
- Sequence database search is the main biological application of the string searching algorithms

Cancer

- A disease of multi-cellular organisms caused by unregulated abnormal cell proliferation
- 2 classes of genes are responsible for normal cell growth and division:
  - Growth factors which induce the proliferation. Their enhanced activity causes cancer
  - Suppressor factors which inhibit proliferation of a cell. Their reduced activity causes cancer

Genes and cancer

Normal cell growth and division

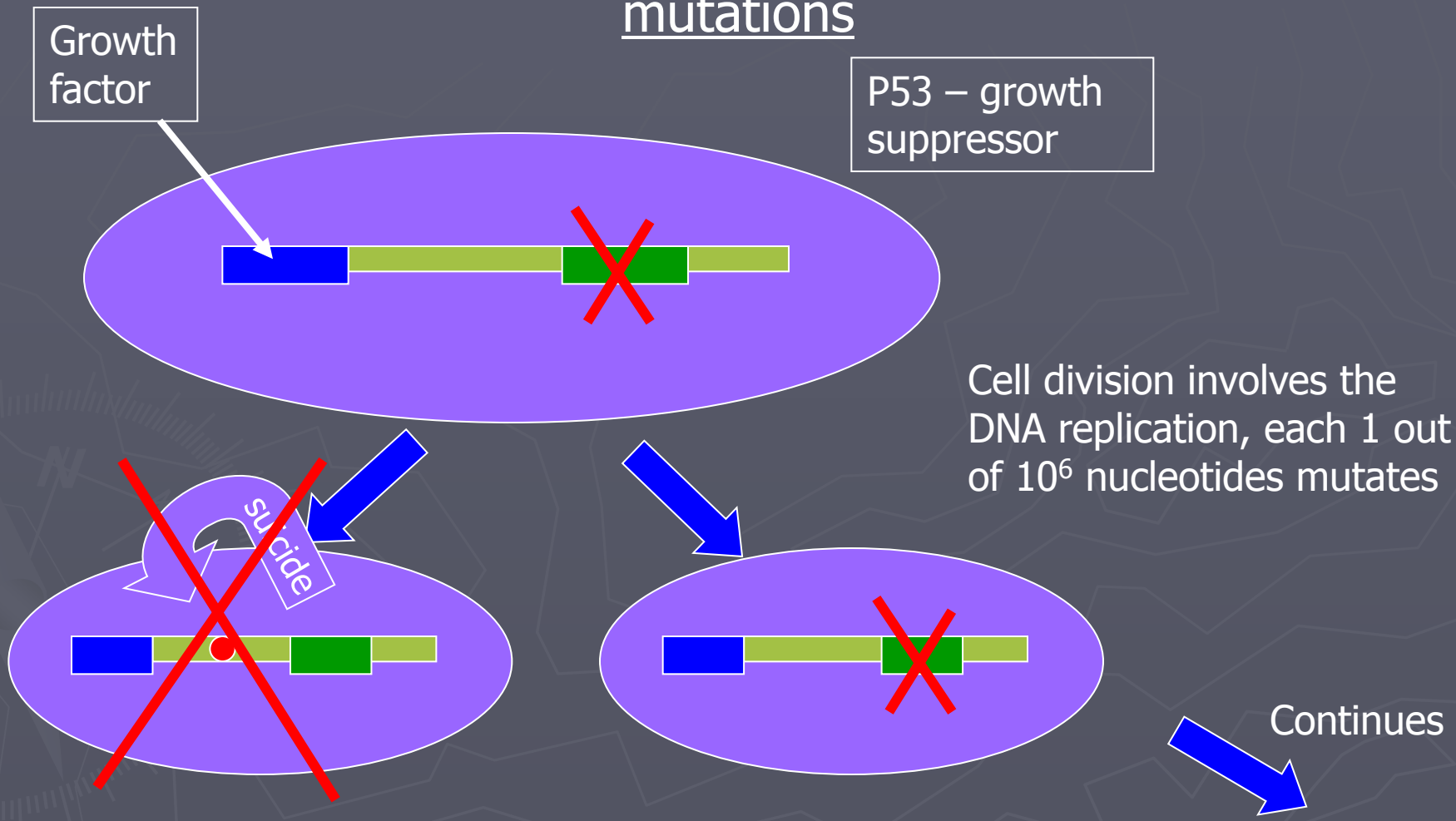


Cell division involves the DNA replication, each 1 out of 10^6 nucleotides mutates

The repair proteins fix somatic mutations

Genes and cancer

Programmed cell death – an extreme repair of somatic mutations



If the repair was unsuccessful, p53 causes cell death – apoptosis – removing it from the following cell division cycle

Oncogenes

- Oncogenes – DNA or RNA sequences of viruses, which were known to cause cancer in cells

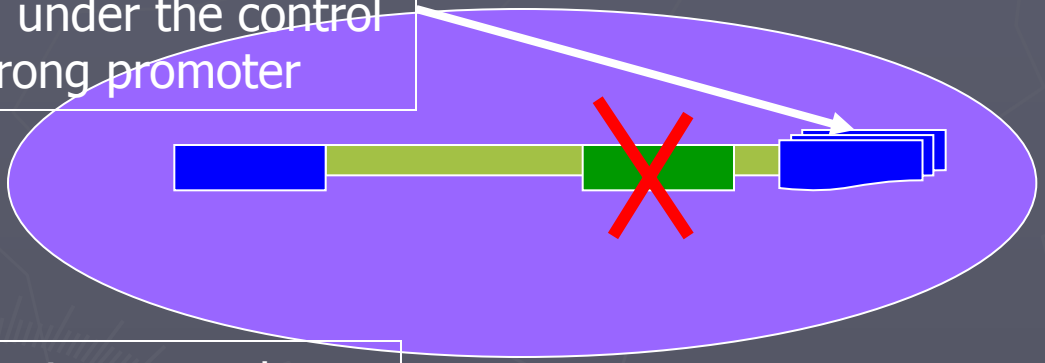
Sequence similarity

- Simian sarcoma viral oncogene v-sis sequence was similar to the PDGF growth factor sequence, as discovered by Doolittle from comparing sequences by eye.
- This lead to a hypothesis that an oncogene is just a transformed growth factor
- The hypothesis was right: cancer is caused by mutation or displacement of a normal cellular regulatory gene

Genes and cancer

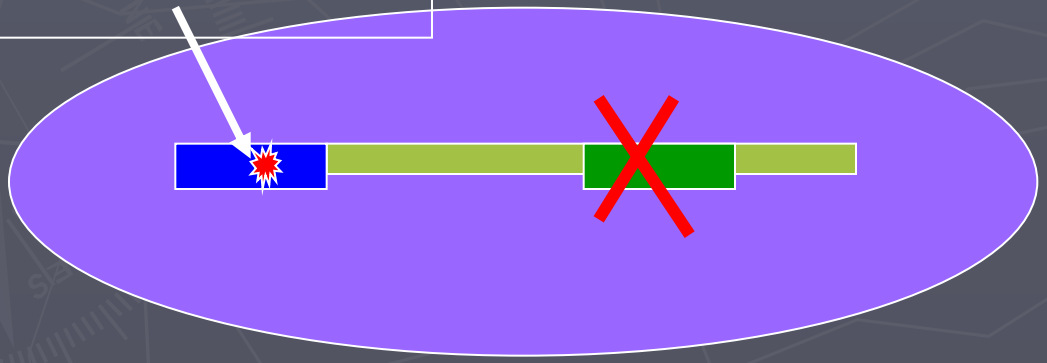
Mutation or misplacement of a growth factor gene

Growth factor with the enhanced function – moved under the control of a strong promoter



Mutation in a growth factor

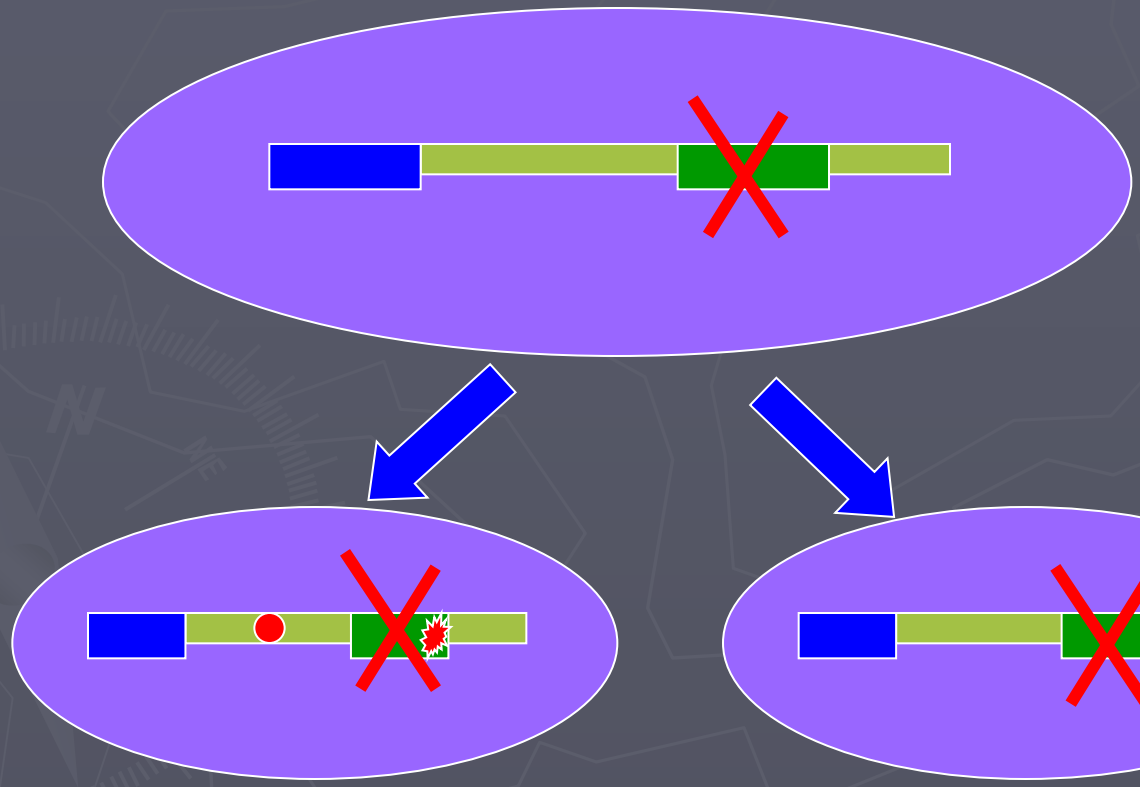
or



Cells divide with an increasing speed, repair mechanisms are insufficient and cell division goes out of control – the tumor grows

Genes and cancer

Mutation in the apoptosis genes



Mutation in the apoptosis genes leads to the proliferation of the mutated cell

Continues

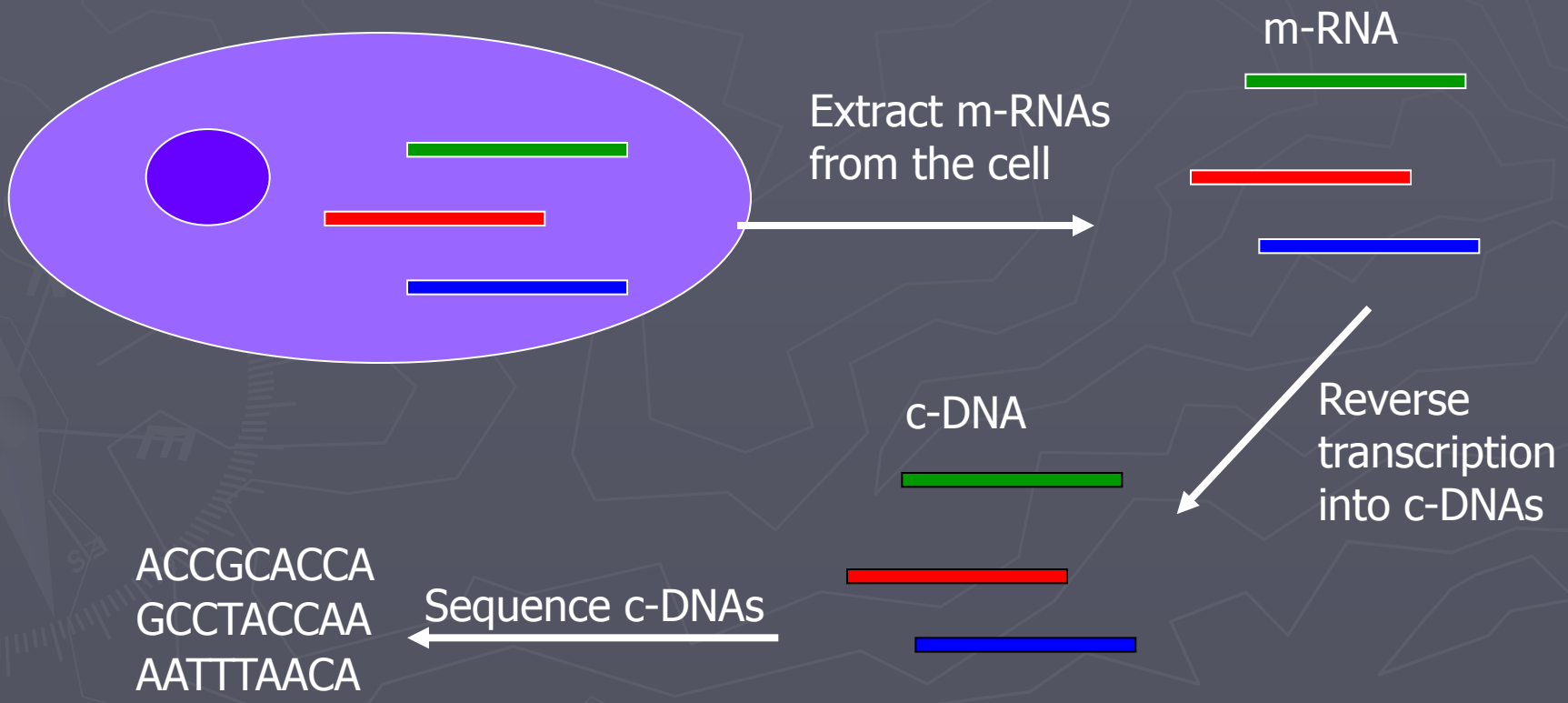
Continues – the cell with the somatic mutation

Proto-oncogenes

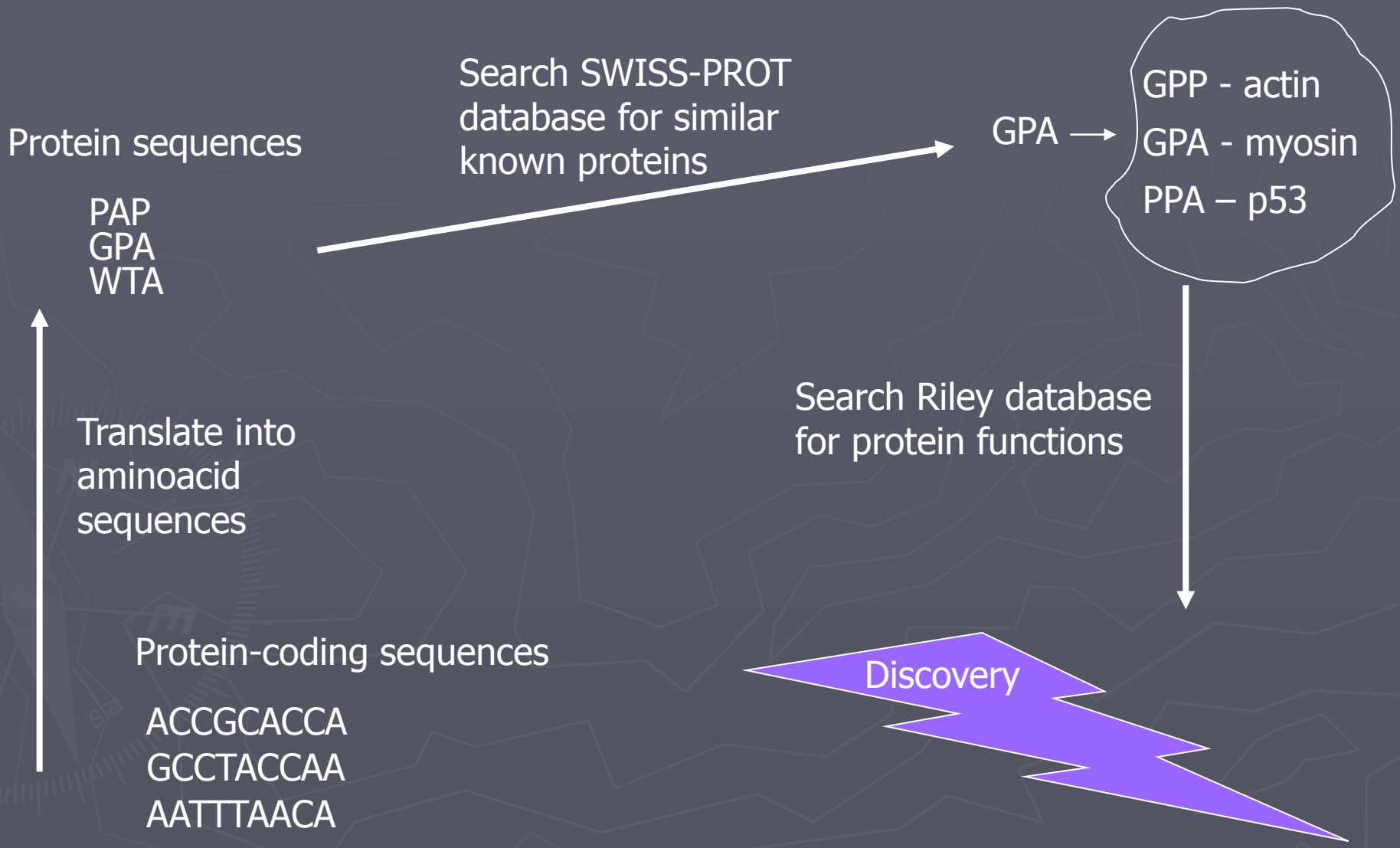
- The normal growth-related genes were named *proto-oncogenes*
- The virus becomes oncogenic when it incorporates (cuts off by mistake?) the proto-oncogenic sequence of PDGF. This sequence then mutates or it is moved by a virus close to a strong enhancer or away from the strong repressor
- The uncontrollable cell growth causes cancer

Large-scale discoveries

- How to discover new proteins and their functions



Large-scale discoveries



Large-scale discoveries

- Instead of extracting m-RNA, obtaining c-DNA and cloning, try to identify protein-coding regions from the genome sequence
- Genes discovered this way are called putative[1)] genes
- The database search on 1743 putative coding regions of the first sequenced organism Haemophilus Influenzae resulted in finding similar proteins and supposed function for 1000 of them
- The discovery ***in silico***

[1)] accepted by supposition rather than as a result of proof

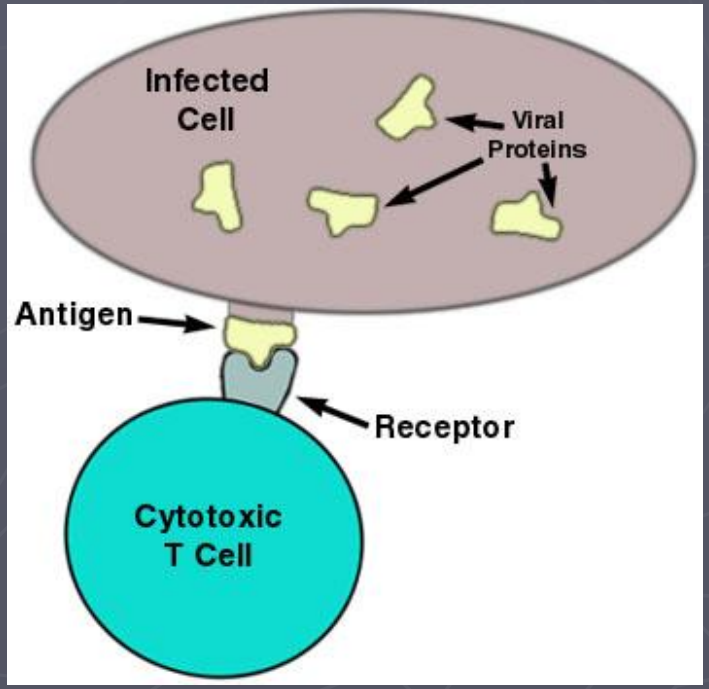
Multiple Sclerosis and database search

- MS is a lethal neurological disease of an *autoimmune* origin

Specific clone of T-cells is produced in response to a foreign antigen

Sometimes specific T-cells start misbehave and give signals to destroy the organism's own proteins – autoimmune disease

MS is caused by T-cells targeting the protein (myelin sheet receptor) of neurons



Multiple Sclerosis and database search

- The specific T-cells were found which do this pseudo-recognition
- These T-cells were previously mobilized by an unknown foreign protein
- We can try to search for a protein similar to the myelin sheet receptor in a database of viral or bacterial proteins

Multiple Sclerosis and database search

- The search was conducted and about 100 candidate proteins were found to be closely similar to the myelin receptor
- Then it was confirmed that the proteins found by the database search are indeed recognized by the specific MS T-cells
- If we further study what is common to all these proteins, we can better understand what particular features of the receptor lead to be mistakenly recognized as a foreign protein

BRCA1 and granins

- The database search requires a biological and statistical expertise to interpret the results
- The hereditary predisposition for the breast and ovarian cancer is due to the mutation in 2 genes, BRCA1 and BRCA2, which have both been precisely located, cloned and sequenced
- But how these genes are involved in the development of a cancer?

Hypothesis: The protein coded by BRCA1 is similar to granins

- The database search reported a substring (of length 10) of BRCA1 which almost perfectly matched the consensus sequence of the family of secretory extracellular proteins called granins, 70-80 aminoacids in length

- The more successful treatment strategies can be applied to the extracellular than to the intracellular proteins. The discovery that BRCA1 can be an extracellular protein lead to optimistic speculations

The search was statistically not reliable

- Originally, researchers have reported that the probability of finding the similarity of the BRCA1-coded protein and granins by chance is 0.00175.
- They concluded that the resulting search reflected an important biological phenomenon
- But this answer does not answer the right question, namely: what is the probability that any 10-letters long motif from the PROSITE database, will occur in BRCA1 by chance?
- That probability was reported to be 0.87 due to 3 factors:
  - There are about 1000 motifs
  - The granin motif is only 10 letters long, and 3 positions are wild cards
  - The BRCA1 sequence is rather long (about 1,800 residues)

The correct search strategy for BRCA1

- Later, the researchers identified 6 regions in BRCA1 corresponding to the globular structure
- By matching each of these 6 regions *separately* to the protein database, they found a moderately conserved 202-residue long region in a human protein 53BP1, known to bind to the universal tumor suppressor p53
- The BRCA1 gene is associated with a well established cancer agent and not with a granin

Do we need to develop new algorithms for the sequence database search

- The string searching tools presented in this part of the course work just fine for the similarity search in the database

- The important details such as
  - local vs global alignment,
  - gap model,
  - scoring matrix,
  - threshold for reporting similarity,
  - statistical significance of obtained results

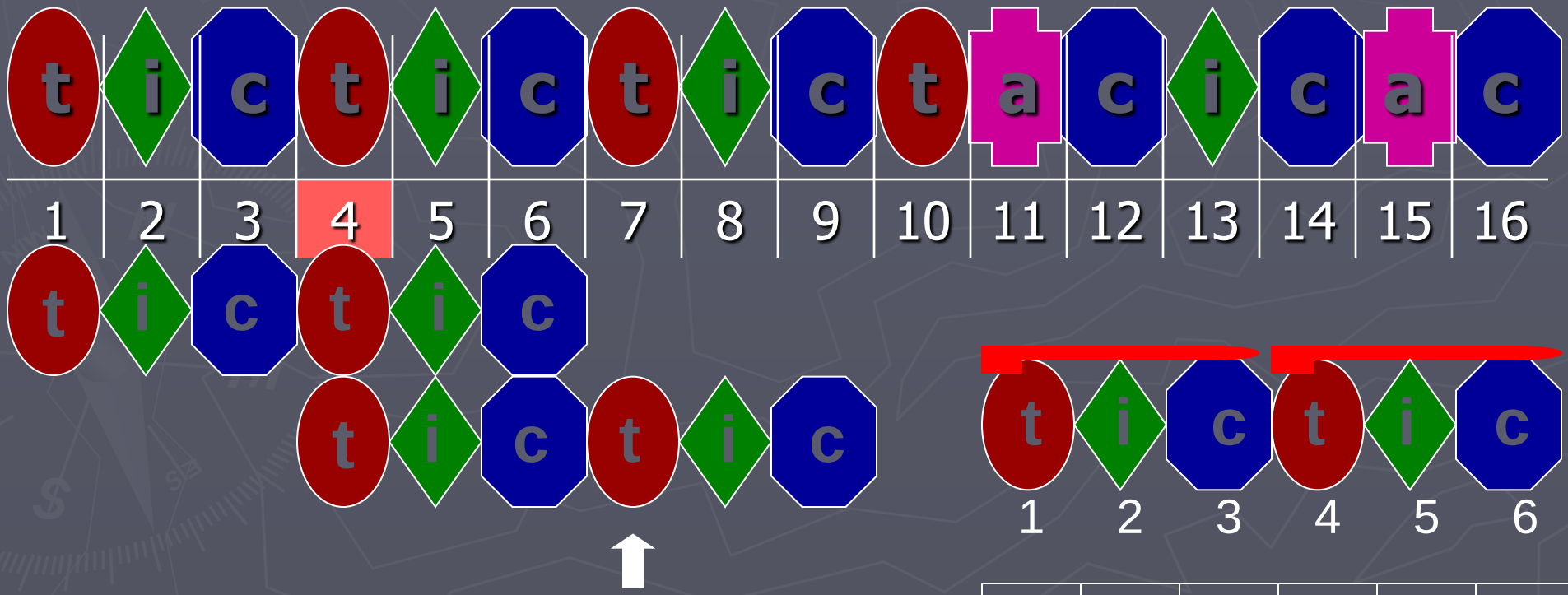
are ***modeling and statistical*** rather than ***algorithmic*** issues

Algorithmic future of the biological sequence comparison

- Specialized algorithmic strategy for solving specific problems, for example
  - Finding known motives
  - Approximate pattern discovery from a set of biosequences
  - Finding semi-characterized features in the query string
- More sensitive, selective and efficient algorithms can be developed for each particular problem

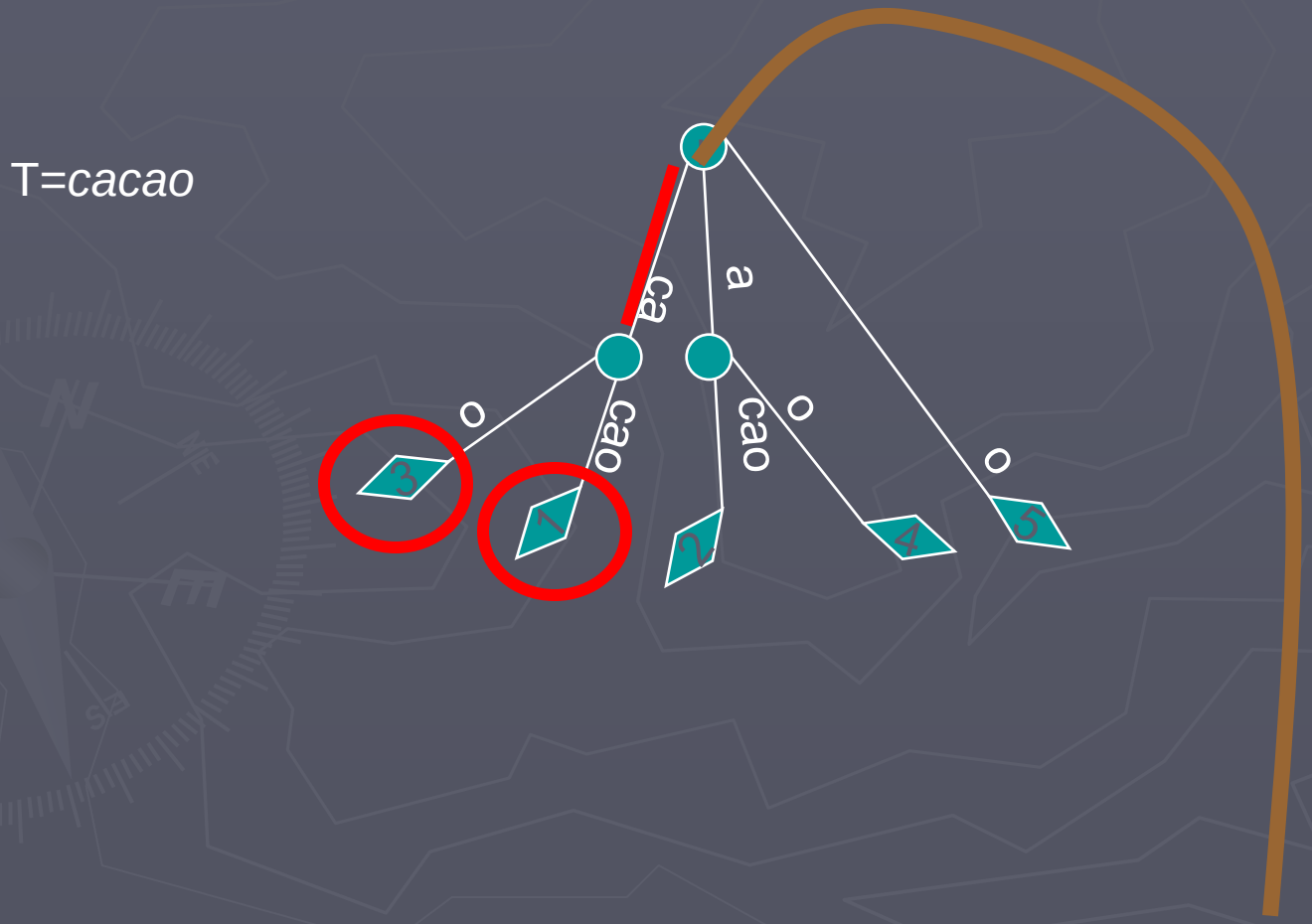
Summary of the tools learned. Exact matching 1

- Linear-time exact pattern search based on a pre-processing of a pattern



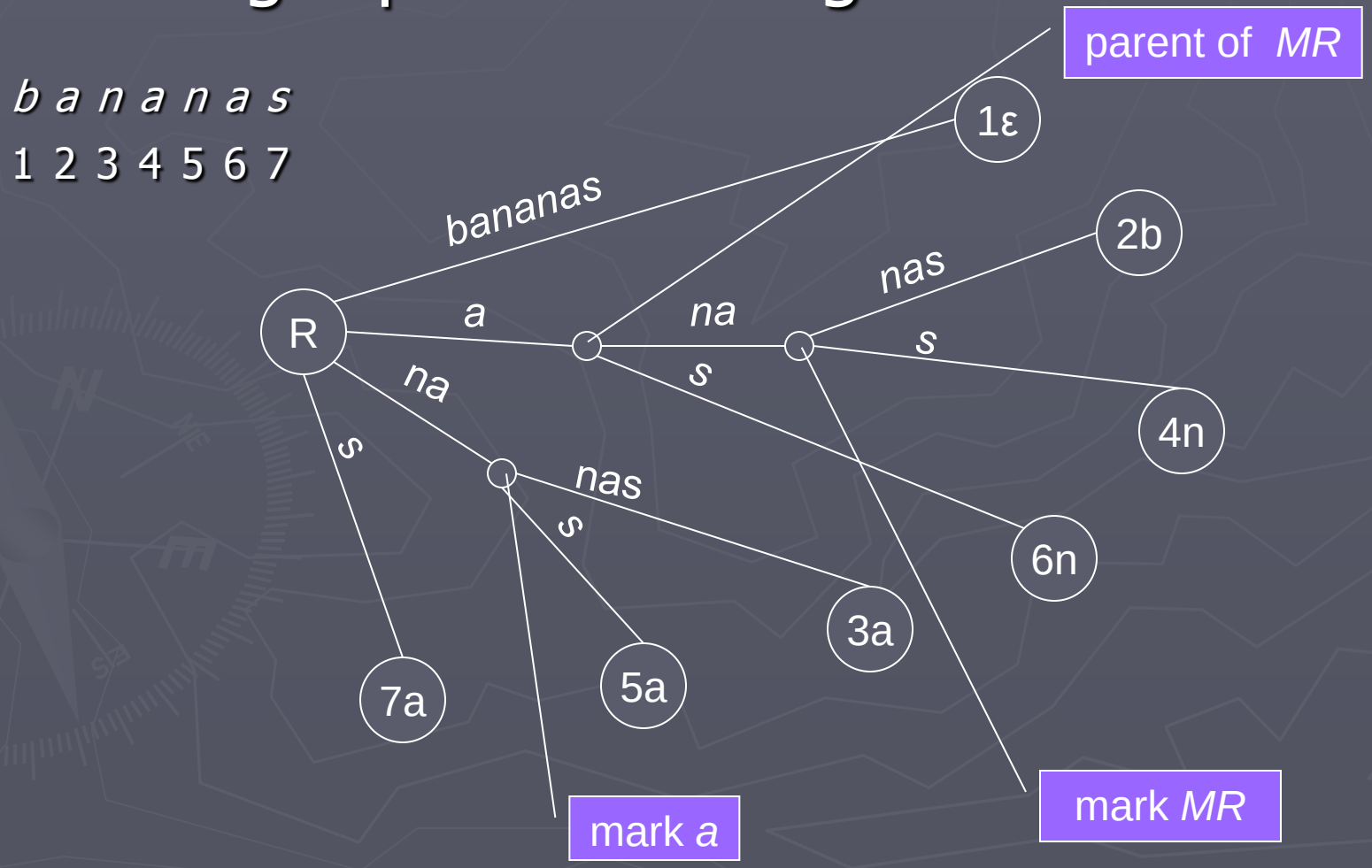
Summary of the tools learned. Exact matching 2

- Exact pattern search using suffix trees



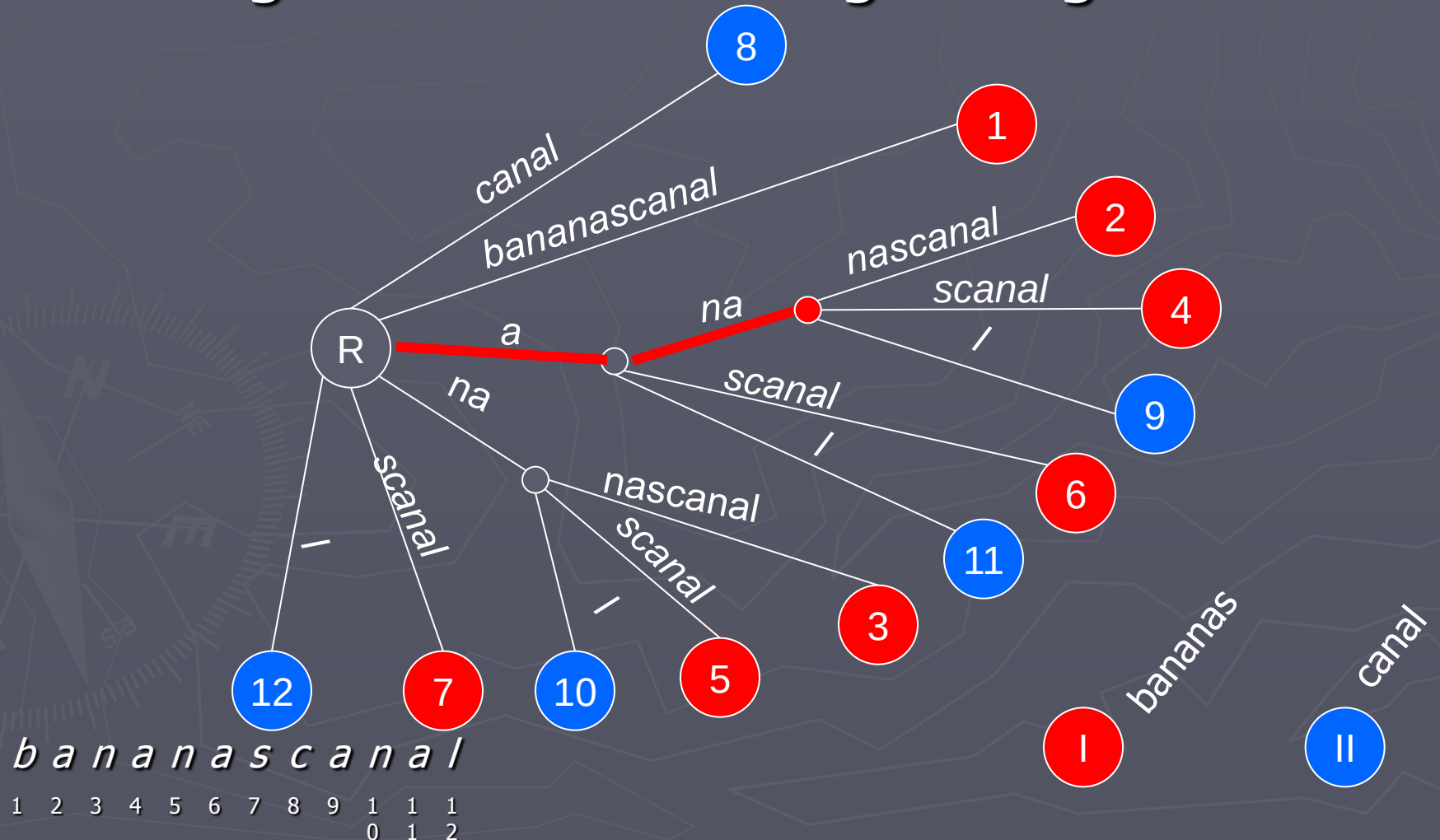
Summary of the tools learned. Exact matching 3

- Finding repetitions using suffix trees



Summary of the tools learned. Exact matching 3

- Finding common substrings using suffix trees



Summary of the tools learned. Approximate matching 1

- Dynamic programming with different scoring matrices

```
algorithm cheepestCost ( array diagonalCost, N, M )
        return cost ( N, M )

algorithm cost ( i, j)
        if i=0 then
                return j
        if j=0 then
                return i
        return min (cost ( i-1, j ) +1, cost ( i, j-1)+1, cost ( i-1, j-1)+diagonalCost [i] [j] )
```

Summary of the tools learned. Approximate matching 1

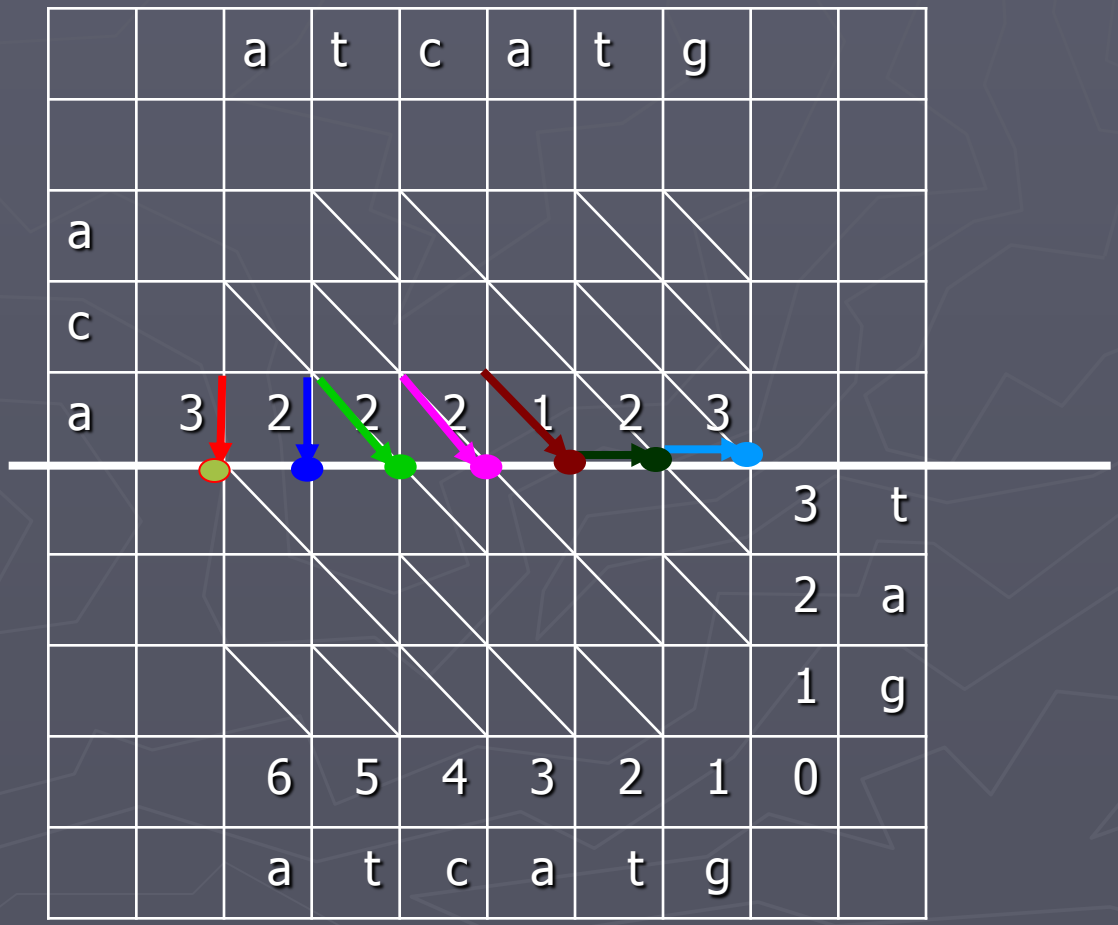
- Dynamic programming with different scoring matrices

```
algorithm cheepestCost ( array diagonalCost, N, M )
        return cost ( N, M )

algorithm cost ( i, j)
        if i=0 then
                return j
        if j=0 then
                return i
        return min (cost ( i-1, j ) +1, cost ( i, j-1)+1, cost ( i-1, j-1)+diagonalCost [i] [j] )
```

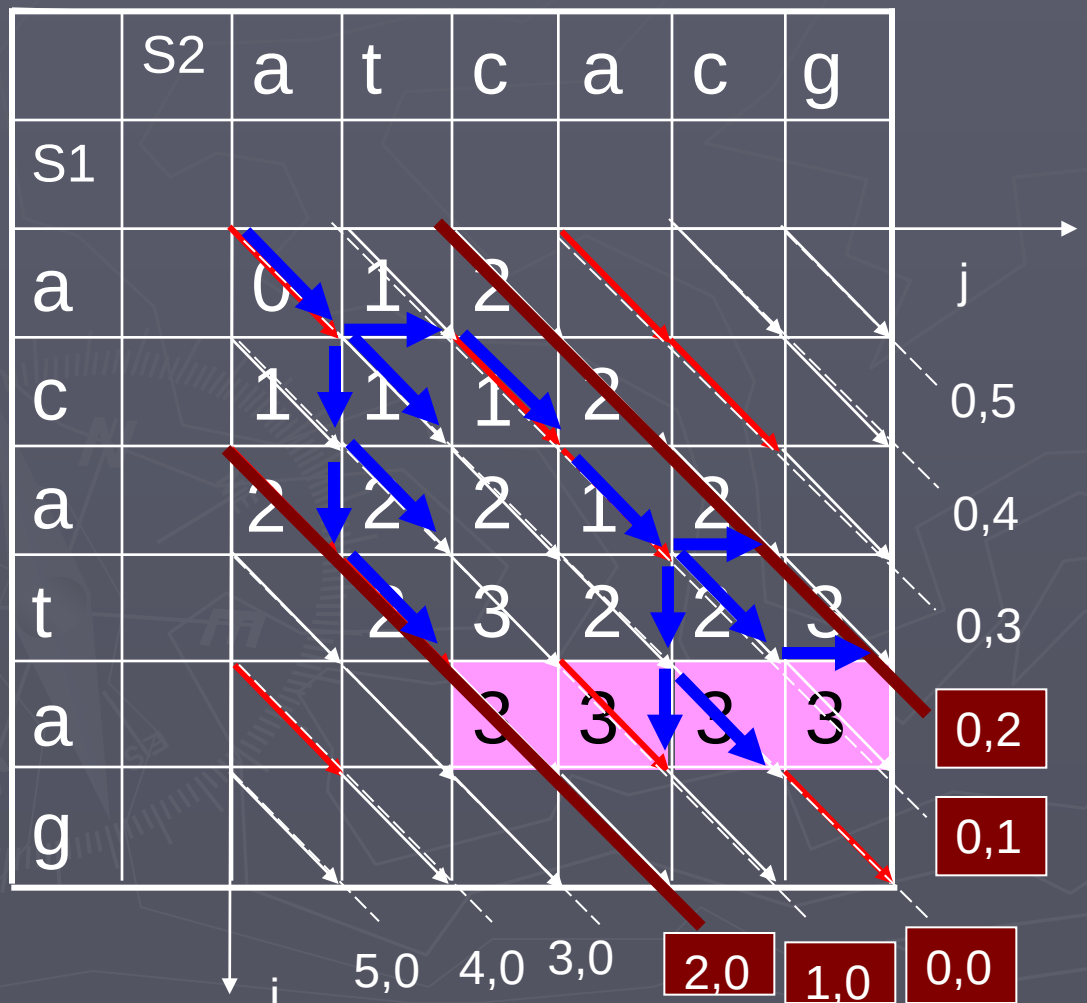
Summary of the tools learned. Approximate matching 2

- Divide and conquer – dynamic programming in linear space



Summary of the tools learned. Approximate matching 3

- A threshold search around the main diagonal



Find a pattern similar to atcacg with an additional threshold – not nore than 2 errors

Compute dynamic programming table values around in a 2*2+1=5 cells strip around the main diagonal

If all values in the row are >2, abort the computation

Summary of the tools learned. Approximate matching 4

- Filtering approximate matches with exact pattern search

Find occurrence of P in T with up to 2 errors. K=2

P	a	c	g	a	a	c	t	t	a

Create k+1=3 partitions of P

acg aac tta

Exact matching of partitions of P against T

T	a	c	a	t	g	a	c	t	g	a	a	c	t	a	g	g	c	a	a	c	g	c	a	t
										a	a	c							a	c	g			

Summary of the tools learned. Approximate matching 4

- Filtering approximate matches with exact pattern search

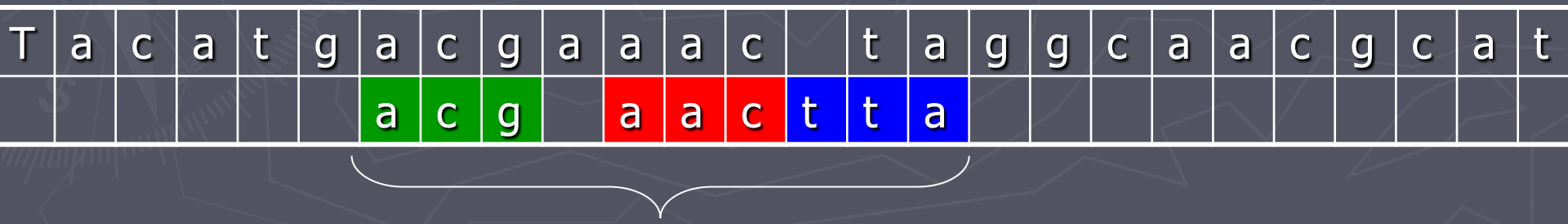
Find occurrence of P in T with up to 2 errors. K=2

P	a	c	g	a	a	c	t	t	a

Create k+1=3 partitions of P

acg aac tta

Extend each seed using dynamic programming to check if it is a part of an approximate match



Some problem solving

Problem 1. Comparing very similar sequences to reveal a polymorphism

- Biological data: DNA sequences of 2 genomes of different E. Coli strains:
  - 1 can process cellulose
  - 1 cannot process cellulose

- How to initially locate the genes responsible for the processing of the cellulose?

Problem 2. Recognizing DNA contamination

- 70s – the dinosaur DNA was sequenced and through the database search was found that the dinosaur DNA contains the sequences which are more similar to the mammals than to the amphibians !!!

- This is a general problem of *DNA contamination*
- If you have the databases of sequenced genomes for main biological groups, how would you prevent the contamination?

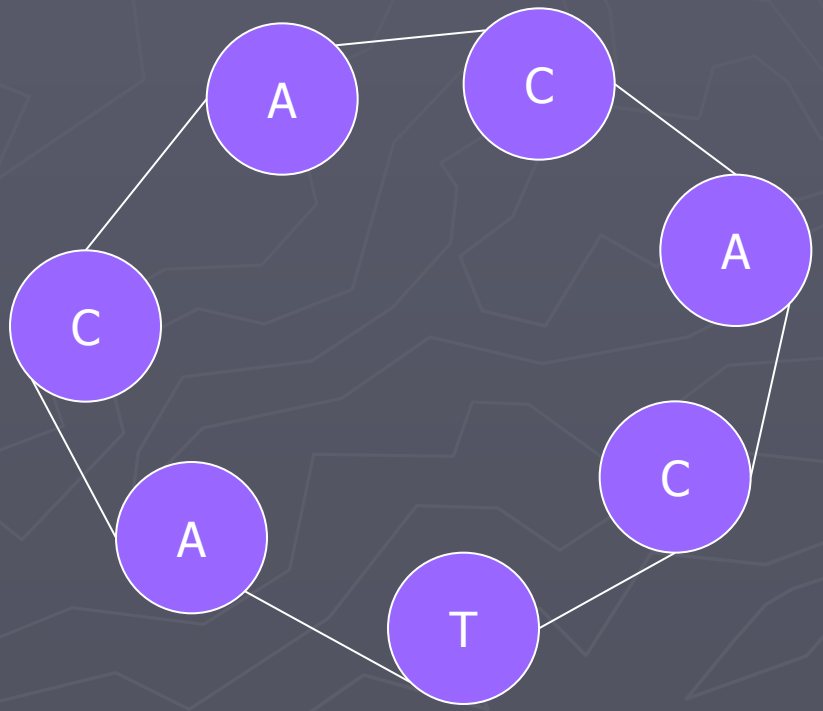
Problem 3. A circular string linearization

- Given an ordering of characters in the alphabet, a string S1 is lexicographically smaller than a string S2 if S1 would appear before S2 in a normal dictionary for these strings

- A circular string linearization problem is: find the place where to cut string S of M characters such that the resulting string would be lexicographically the smallest in between all possible cuts of S

Problem 3. A circular string linearization

- This can be used, for example, for the canonical representation of circular bacterial genomes
- Try on the following input:
- A<C<T



Problem 4. Selecting a unique substring of length 26 for each of 100 DNAs

- This can be used in the DNA microarrays to reduce the cross-binding in case that the same representative substring has been chosen for different genes

Problem 5. Finding maximal tandem repeats

- Consider a hashing versus the suffix tree solution
- Example: AACACACAACACACA