Elements of Bioinformatics Autumn 2010

VELI MÄKINEN

HTTP://WWW.CS.HELSINKI.FI/EN/COURSES/582606/2010/S/K/1

Prerequisites & content

- Some biology, some algorithms, some statistic are assumed as background.
- Inherits parts of old *Practical course in Biodatabases* and *Introduction to Bioinformatics*[1] courses, lectured the last time in 2009.
- *The idea is to look at several concrete pipelines that are in use in bioinformatics, and study the elements constituting them in detail.*
- *Python* used as the scripting language to tie the elements into pipelines.

[1] Also inherits some lecture slides from the latter course (thanks to Esa Pitkänen)

What is bioinformatics?

- Bioinformatics, *n*. The science of **information and information flow in biological systems**, esp. of the use of computational methods in **genetics and genomics**. (Oxford English Dictionary)

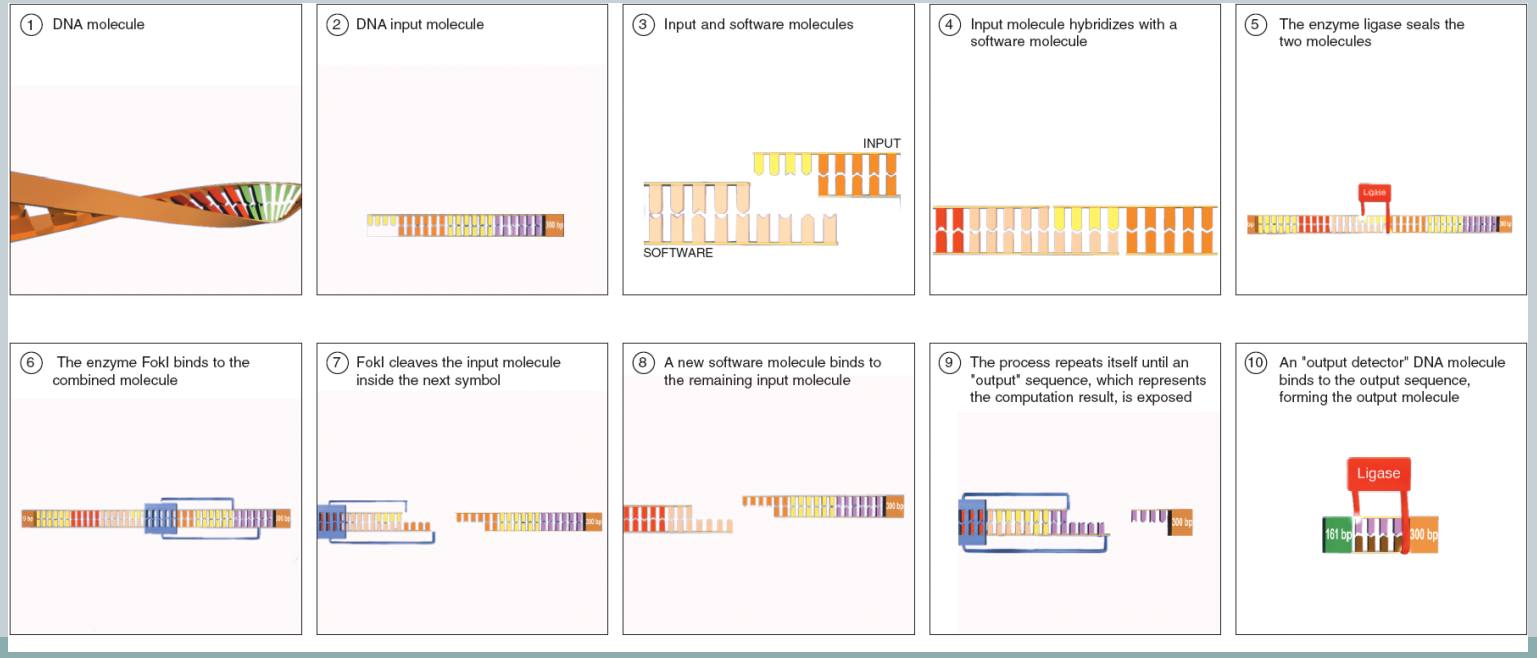
- "The **mathematical**, **statistical** and **computing** methods that aim to solve biological problems using **DNA and amino acid sequences** and related information." -- Fredj Tekaia

What is bioinformatics?

- "I do not think all biological computing is bioinformatics, e.g. mathematical modelling is not bioinformatics, even when connected with biology-related problems. In my opinion, bioinformatics has to do with **management** and the subsequent use of biological information, particular **genetic information**." -- Richard Durbin

What is *not* bioinformatics?

- Biologically-inspired computation, e.g., genetic algorithms and neural networks
- However, application of neural networks to solve some biological problem, could be called bioinformatics
- What about DNA computing?



http://www.wisdom.weizmann.ac.il/~lbn/new_pages/Visual_Presentation.html

Computational biology

- Application of **computing** to **biology** (broad definition)
- Often used interchangeably with bioinformatics
- Or: ***Biology*** that is done with **computational means**

Biometry & biophysics

- Biometry: the **statistical analysis** of **biological data**
  - Sometimes also the field of identification of individuals using biological traits (a more recent definition)
- Biophysics: "an interdisciplinary field which applies techniques from the **physical sciences** to understanding **biological structure and function**" -- British Biophysical Society

Mathematical biology

- Mathematical biology "tackles biological problems, but the methods it uses to tackle them need not be numerical and need not be implemented in software or hardware."

  -- Damian Counsell

Alan Turing

THE CHEMICAL BASIS OF MORPHOGENESIS

By A. M. TURING, F.R.S. *University of Manchester*

(*Received* 9 *November* 1951—*Revised* 15 *March* 1952)

It is suggested that a system of chemical substances, called morphogens, reacting together and diffusing through a tissue, is adequate to account for the main phenomena of morphogenesis. Such a system, although it may originally be quite homogeneous, may later develop a pattern or structure due to an instability of the homogeneous equilibrium, which is triggered off by random disturbances. Such reaction-diffusion systems are considered in some detail in the case of an isolated ring of cells, a mathematically convenient, though biologically unusual system. The investigation is chiefly concerned with the onset of instability. It is found that there are six essentially different forms which this may take. In the most interesting form stationary waves appear on the ring. It is suggested that this might account, for instance, for the tentacle patterns on *Hydra* and for whorled leaves. A system of reactions and diffusion on a sphere is also considered. Such a system appears to account for gastrulation. Another reaction system in two dimensions gives rise to patterns reminiscent of dappling. It is also suggested that stationary waves in two dimensions could account for the phenomena of phyllotaxis.

The purpose of this paper is to discuss a possible mechanism by which the genes of a zygote may determine the anatomical structure of the resulting organism. The theory does not make any new hypotheses; it merely suggests that certain well-known physical laws are sufficient to account for many of the facts. The full understanding of the paper requires a good knowledge of mathematics, some biology, and some elementary chemistry. Since readers cannot be expected to be experts in all of these subjects, a number of elementary facts are explained, which can be found in text-books, but whose omission would make the paper difficult reading.

1. A model of the embryo. Morphogens

In this section a mathematical model of the growing embryo will be described. This model will be a simplification and an idealization, and consequently a falsification. It is to be hoped that the features retained for discussion are those of greatest importance in the present state of knowledge.

The model takes two slightly different forms. In one of them the cell theory is recognized but the cells are idealized into geometrical points. In the other the matter of the organism is imagined as continuously distributed. The cells are not, however, completely ignored, for various physical and physico-chemical characteristics of the matter as a whole are assumed to have values appropriate to the cellular matter.

With either of the models one proceeds as with a physical theory and defines an entity called 'the state of the system'. One then describes how that state is to be determined from the state at a moment very shortly before. With either model the description of the state consists of two parts, the mechanical and the chemical. The mechanical part of the state describes the positions, masses, velocities and elastic properties of the cells, and the forces between them. In the continuous form of the theory essentially the same information is given in the form of the stress, velocity, density and elasticity of the matter. The chemical part of the state is given (in the cell form of theory) as the chemical composition of each separate cell; the diffusibility of each substance between each two adjacent cells must also

Vol. 237. B. 641. (Price 8s.) 5 [Published 14 August 1952

Turing on biological complexity

- "It must be admitted that the **biological examples** which it has been possible to give in the present paper are **very limited**.

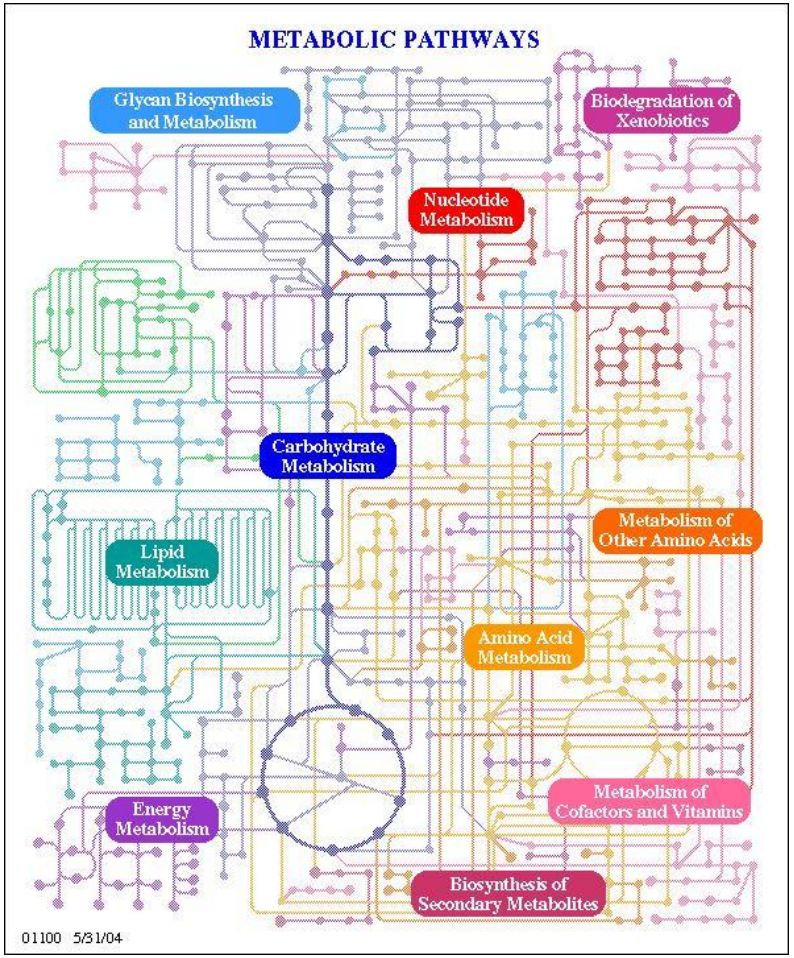
  This can be ascribed quite simply to the fact that **biological phenomena** are usually **very complicated**. Taking this in combination with the relatively elementary mathematics used in this paper one could hardly expect to find that many observed biological phenomena would be covered.

  It is thought, however, that the **imaginary biological systems** which have been treated, and the principles which have been discussed, should be of some help in **interpreting real biological forms**."

  – Alan Turing, The Chemical Basis of Morphogenesis, 1952

Related concepts

- Systems biology
  - “Biology of networks”
  - Integrating different levels of information to understand how biological systems work
- Computational systems biology



Overview of metabolic pathways in KEGG database, www.genome.jp/kegg/

Why is bioinformatics important?

- New measurement techniques produce huge quantities of biological data
  - Advanced data analysis methods are needed to make sense of the data
    - The 1000 Genomes Project Consortium *Nature* **467**, 1061-1073 (2010).
    - Sudmant, P. H. *et al. Science* **330**, 641-646 (2010).
- Paradigm shift in biology to utilise bioinformatics in research
  - Pevzner & Shamir: Computing Has Changed Biology – Biology Education Must Catch Up. *Science* 31(5940):541-542, 2009.

comments on this story

Stories by subject

- Biotechnology
- Cell and molecular biology
- Genetics
- Health and medicine

Stories by keywords

- 1000 Genomes
- sequencing
- exome
- variation

This article elsewhere

Blogs linking to this article

Published online 27 October 2010 | Nature | doi:10.1038/news.2010.567

News

1000 Genomes Project reveals human variation

An international effort to map variability in the genome hits its first landmark.

Alla Katsnelson

The long-awaited results from the pilot phase of the first large-scale initiative to sequence individual genomes have identified 95% of the variation found across the human genome and revealed some 15 million gene variants, more than half of which had never been observed before. The data represent the most thorough effort so

A project to sequence hundreds of human genomes has found millions of

Bioinformatician's skill set

- Statistics, data analysis methods
  - Lots of data
  - High noise levels, missing values
  - #attributes >> #data points
- Modelling
  - Discrete vs continuous domains
  - -> Systems biology
- Data structures, databases
- Algorithms

Bioinformatician's skill set

- Programming languages
  - Scripting languages: Python, Perl, Ruby, …
  - Extensive use of text file formats: need parsers
  - Integration of both data and tools
- Scientific computation packages
  - R, Matlab/Octave, …
- Communication skills

Scientific method of bioinformatics

- Is there such?
- Bioinformatics is not a science in itself, just a new approach to study a science – biology.
- The accepted way to do research in bioinformatics is somewhere between the hypothesis testing method of experimental sciences and exact mathematical method of exact sciences.
  - There are two extremes among bioinformaticians:
    - Those that use bioinformatics tools in creative ways and follow the hypothesis testing method of experimental sciences.
    - Those that develop the bioinformatics tools and follow the exact mathematical method.
  - Typically the most influental research is done somewhere in between.

Educational goal

- This course (as part of MBI curriculum) aims to educate bioinformaticians that are ”in between”:
  - In addition to learning what tools are used in bioinformatics, we aim at in depth understanding of the principles leading to those tools.
  - Suitable background for continuing to PhD studies in bioinformatics.
  - Suitable background for working as a ”method consultant” in biological research groups that mainly use bioinformatics tools rather than understand how they work.

Bioinformatics pipelines

- As the biological research groups use the bioinformatics tools as black boxes, the method developers have put effort on using standardized input/output formats.
  - Otherwise it is not easy to get your own tool as part of larger *pipelines*.
- Looking directly at pipelines and elements constituting them gives a systematic way to study bioinformatics.
  - Course content is structured like this.

Before continuing…

- Assuming now central dogma DNA->RNA->proteins, nucleotides, amino acids, gene structure, introns, exons, basics of regulatory mechanism, evolutionary mutations, alleles, recombinations, etc. basic biology known.
- For recap, see e.g. https://www.cs.helsinki.fi/i/vmakinen/algbio10/algbio10_lecture1.pdf (part II, page 24 on).

One slide recap

Nucleotides A, C, G, T

gene

DNA

…TACCTACATCCACTCATC…AGCTACGTTCCCCGACTACGACATGGTGATT

5’ …ATGGATGTAGGTGAGTAG…TCGATGCAAGGGGCTGATGCTGTACCACTAA… 3’

exon intron exon

transcription

RNA

codon

…AUGGAUGUAGAUGGGCUGAUGCUGUACCACUAA

translation

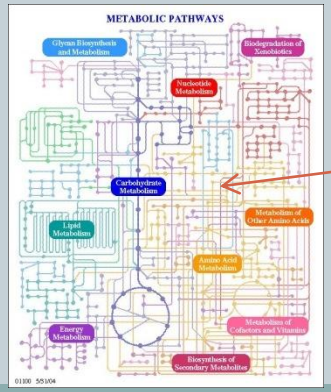
Protein

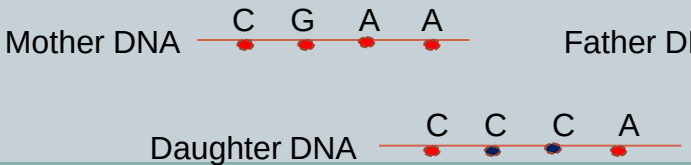
MDVDGLMLYH

Gene regulation

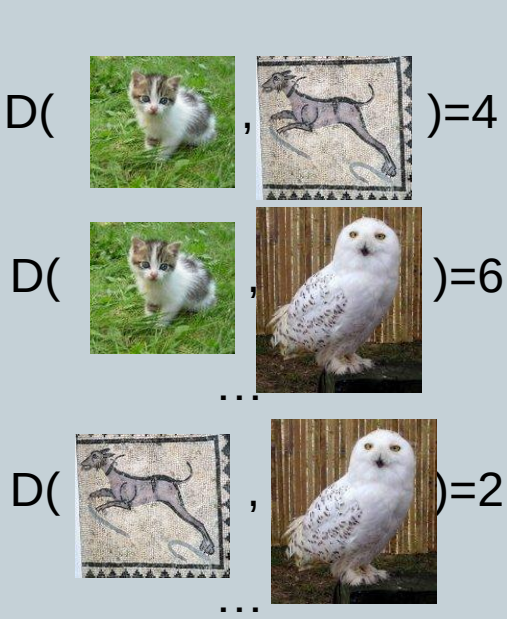
recombination

entsyme





First pipeline: Phylogeny by distance method



Distance between two species?

- Optimal would be the time since the species separated…
- An approximation is to look at the sequence similarity and especially *genome rearrangements*
  - Point mutations are frequent but larger scale variations are rare -> large scale variations give a better accuracy estimate on the evolutionary distance.
  - First locate homologous genes or syntenic blocks (prediction by high sequence similarity).
  - Then count how many large scale variations (dublications, inversions, translocations, fissions, fusions) are necessary to convert the order of genes in one species to the order in other species.
    - A lower-bound for the biological truth, assuming all possible rearrangement operations are taken into account and homologies are correctly identified.

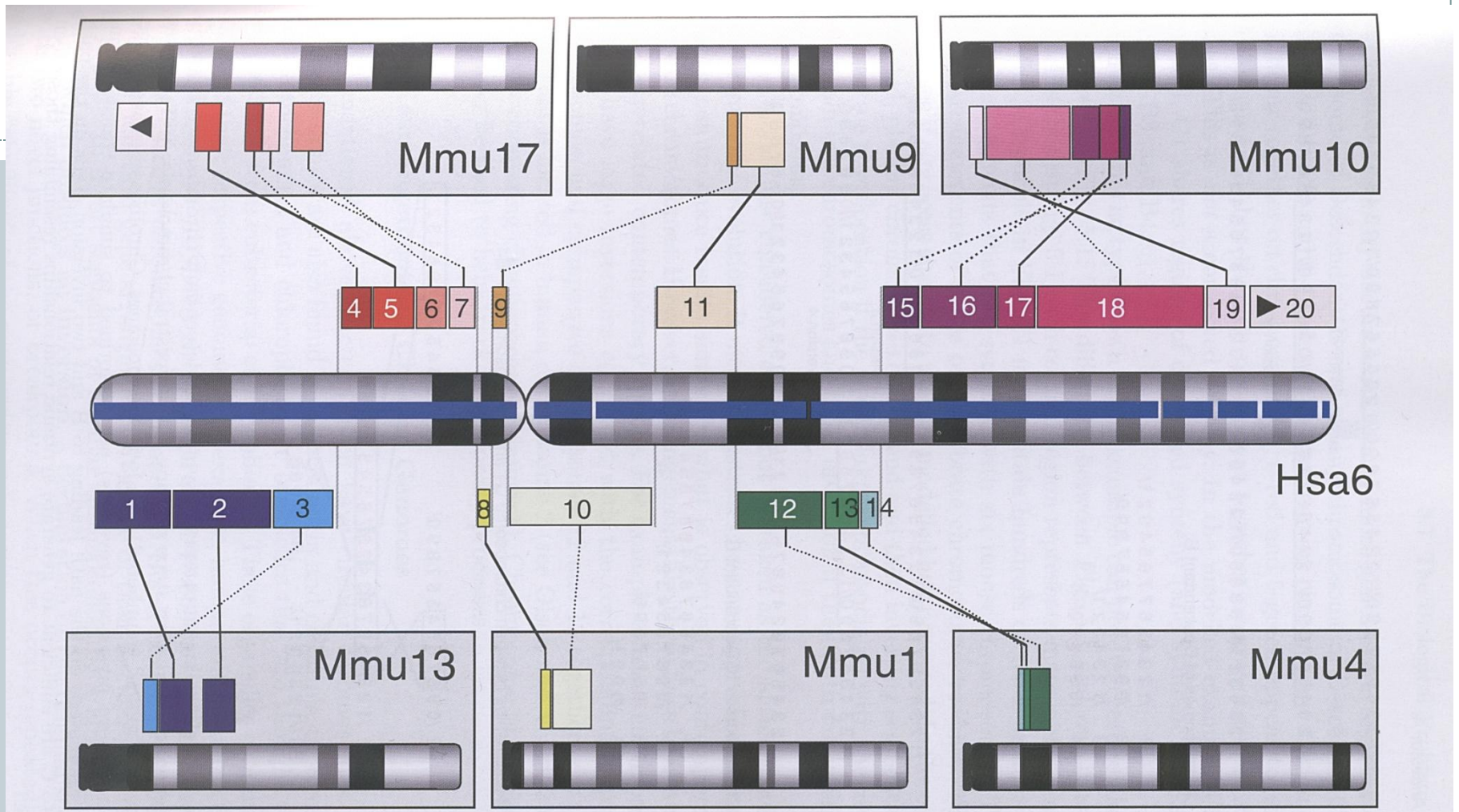
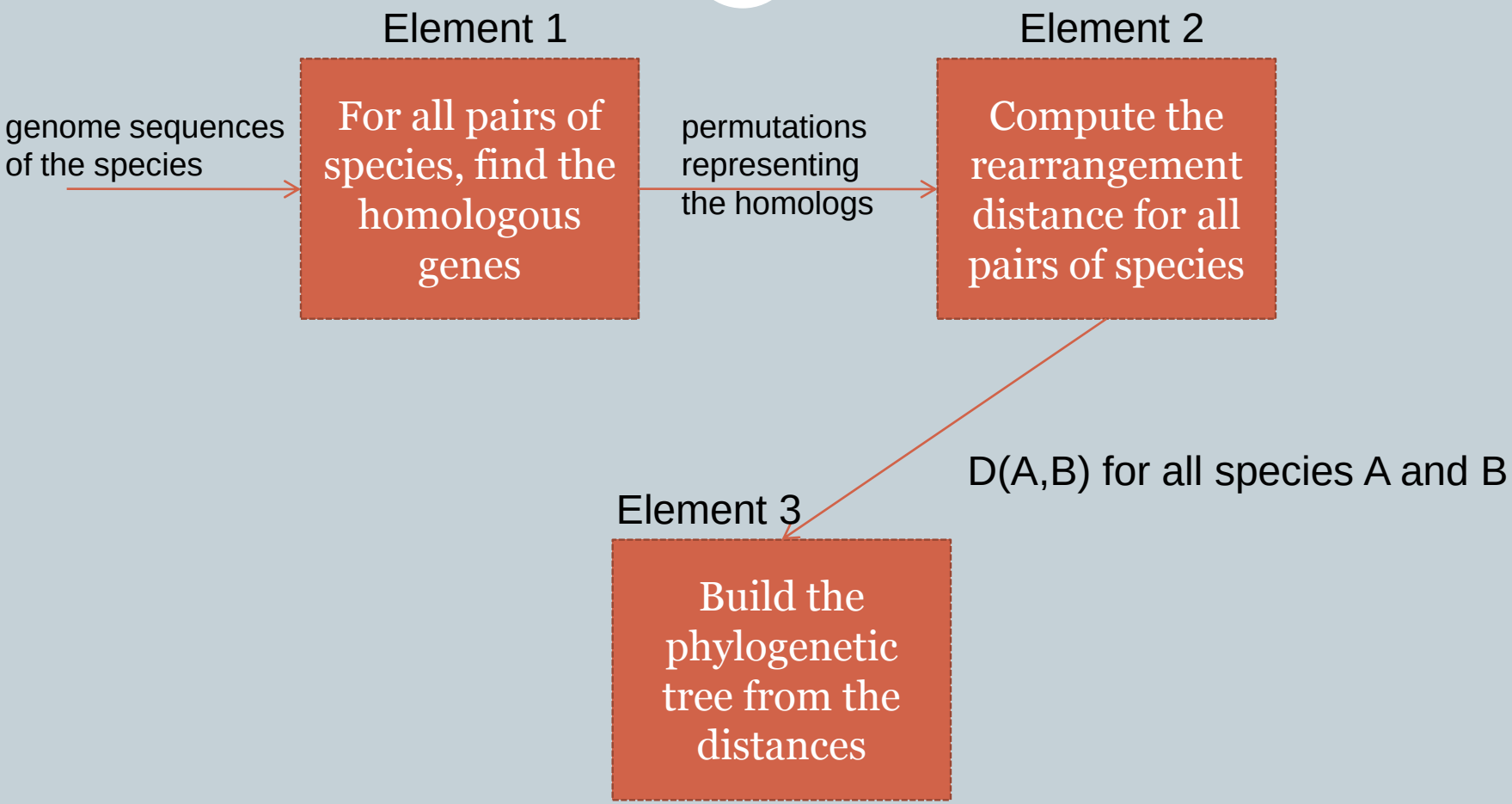


Fig. 5.1. Syntenic blocks conserved between human chromosome Hsa6 and mouse chromosomes. Broken lines indicate regions that appear in inverted orders in the two organisms. Reprinted, with permission, from Gregory SG et al. (2002) *Nature* 418:743–750. Copyright 2002 Nature Publishing Group.

Jones & Pevzner, 2004

Phylogeny by distance method pipeline

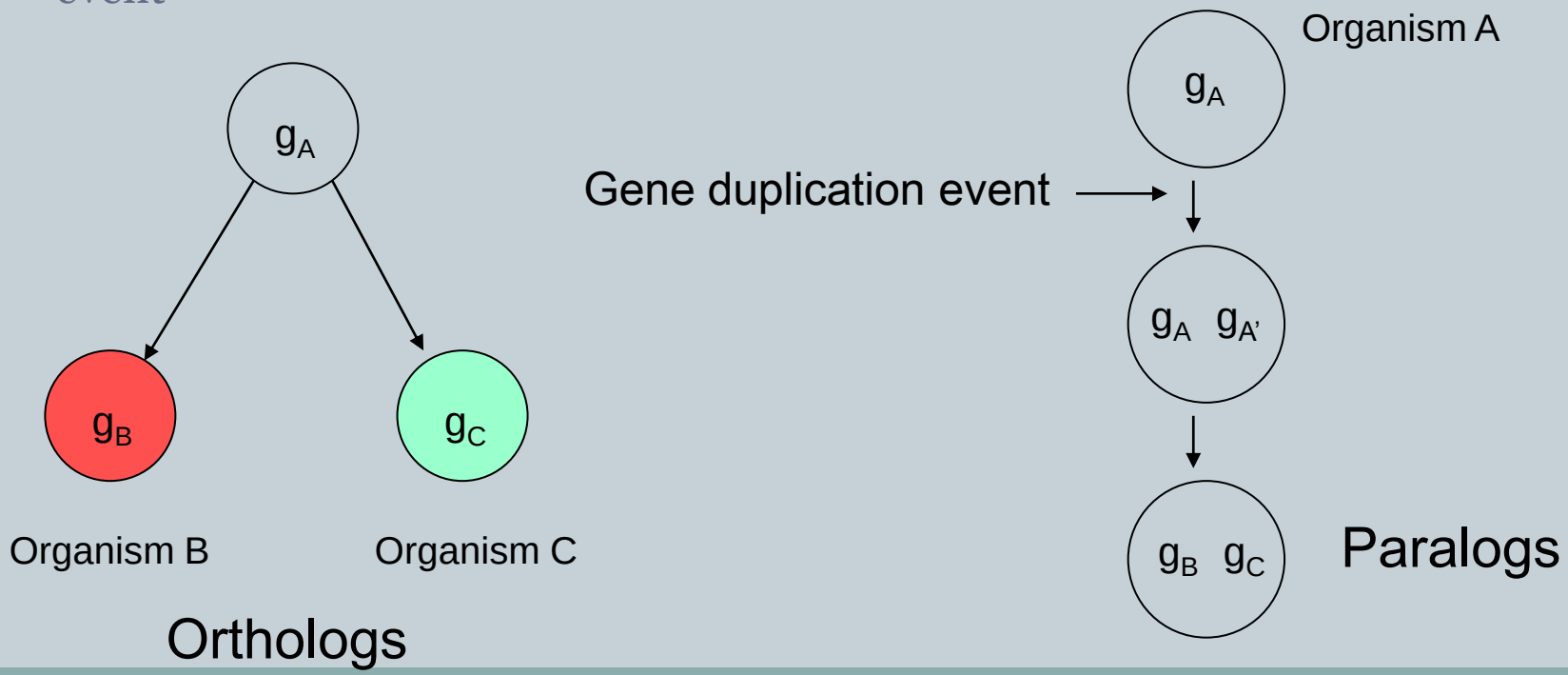


Elements 1-3

- Element 2 covered in the Algorithms for Bioinformatics course.
- Element 1 also briefly looked at in the Algorithms for Bioinformatics course, but will now be studied in more detail.
- Element 3 will be studied later on (last lecture?)

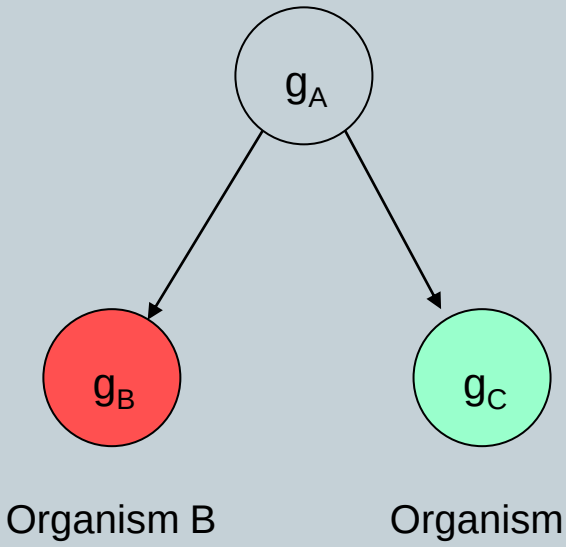
Element 1: Homologs, orthologs and paralogs

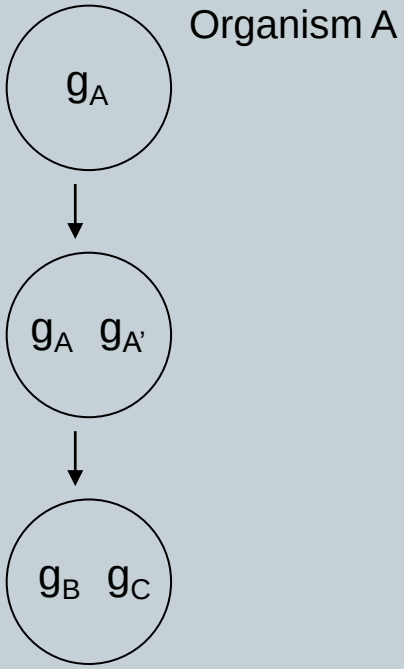
- We distinguish between two types of homology
  - Orthologs: homologs from two different species, separated by a *speciation* event
  - Paralogs: homologs within a species, separated by a *gene duplication* event



Orthologs and paralogs

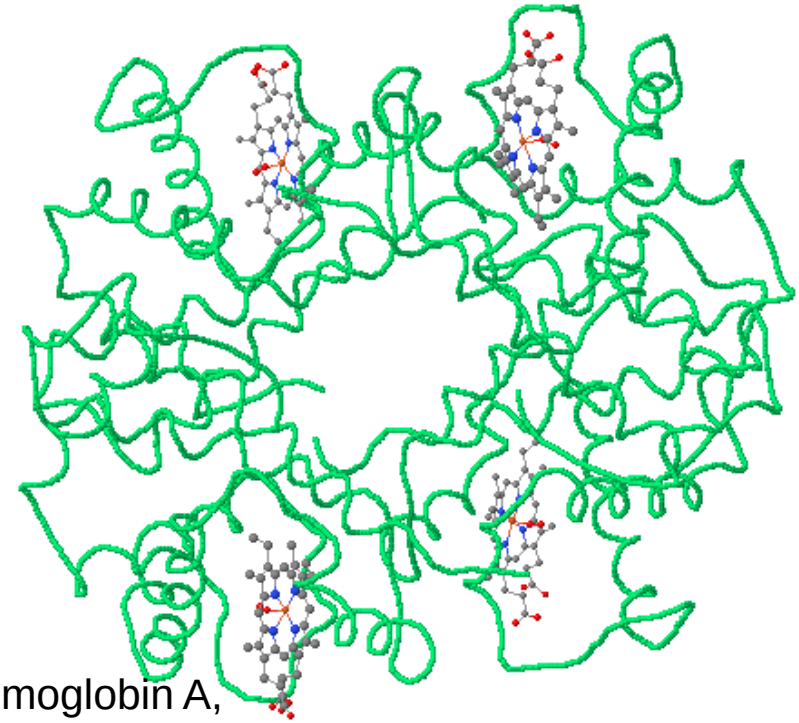
- Orthologs typically retain the original function
- In paralogs, one copy is free to mutate and acquire new function (no selective pressure)





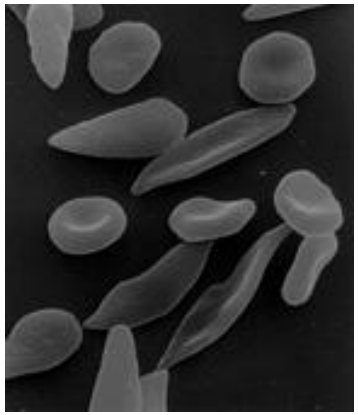
Paralogy example: hemoglobin

- Hemoglobin is a protein complex which transports oxygen
- In humans, hemoglobin consists of four protein subunits and four non-protein heme groups



Hemoglobin A, www.rcsb.org/pdb/explore.do?structureId=1GZX

Sickle cell diseases are caused by mutations in hemoglobin genes

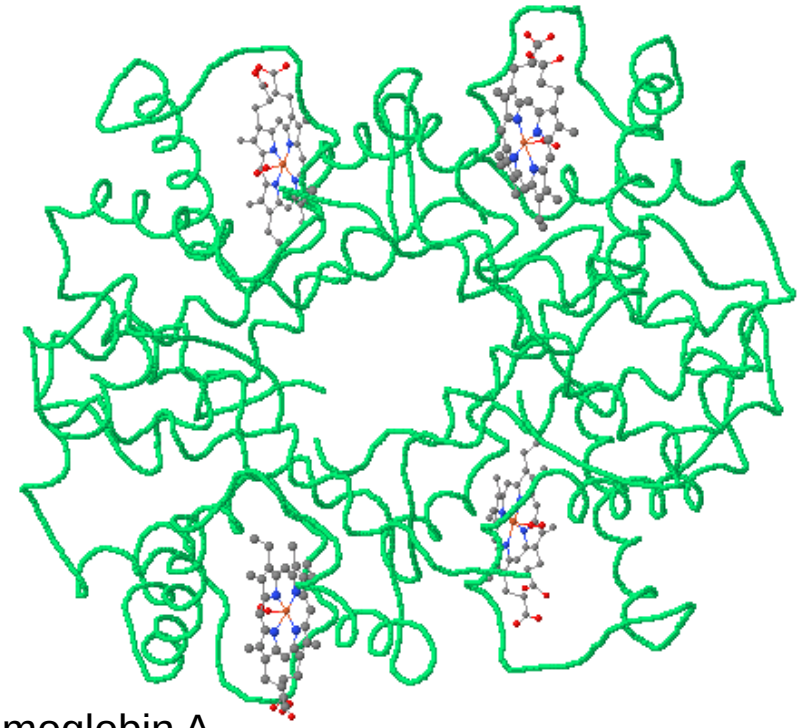


http://en.wikipedia.org/wiki/Image:Sicklecells.jpg

27

Paralogy example: hemoglobin

- In adults, three types are normally present
  - Hemoglobin A: 2 alpha and 2 beta subunits
  - Hemoglobin A2: 2 alpha and 2 delta subunits
  - Hemoglobin F: 2 alpha and 2 gamma subunits
- Each type of subunit (alpha, beta, gamma, delta) is encoded by a separate gene

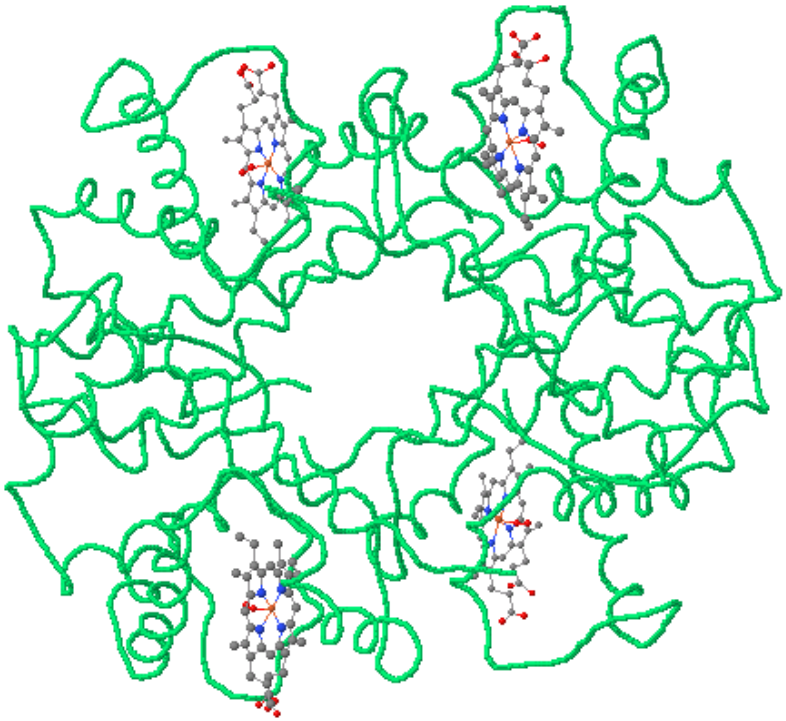


Hemoglobin A, www.rcsb.org/pdb/explore.do?structureId=1GZX

28

Paralogy example: hemoglobin

- The subunit genes are paralogs of each other, i.e., they have a common ancestor gene
- Exercise: hemoglobin human paralogs in NCBI sequence databases
  http://www.ncbi.nlm.nih.gov/sites/entrez?db=Nucleotide
  - Find human hemoglobin alpha, beta, gamma and delta
  - Compare sequences



Hemoglobin A,
www.rcsb.org/pdb/explore.do?structureId=1GZX

29

Orthology example: insulin

- The genes coding for insulin in human (*Homo sapiens*) and mouse (*Mus musculus*) are orthologs:
  - They have a common ancestor gene in the ancestor species of human and mouse
  - Exercise: find insulin orthologs from human and mouse in NCBI sequence databases

Sequence alignment: estimating homologs by sequence similarity

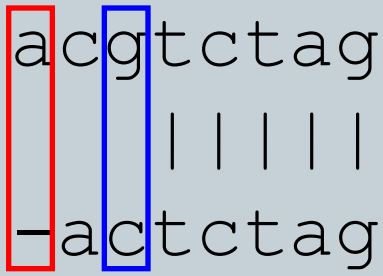
- Alignment specifies which positions in two sequences match

```
acgtctag    acgtctag     acgtctag
||            |||||      || |||||
actctag-    -actctag     ac-tctag
```

2 matches	5 matches	7 matches
5 mismatches	2 mismatches	0 mismatches
1 not aligned	1 not aligned	1 not aligned

Mutations: Insertions, deletions and substitutions

Indel: insertion or deletion of a base with respect to the ancestor sequence



Mismatch: substitution (point mutation) of a single base

- Insertions and/or deletions are called *indels*

Global alignment

- Problem: find optimal scoring alignment between two sequences (Needleman & Wunsch 1970)
- Every position in both sequences is included in the alignment
- We give score for each position in alignment
  - Identity (match) +1
  - Substitution (mismatch) $-\mu$
  - Indel $-\delta$
- Total score: sum of position scores

Scoring: Toy example

- Consider two sequences with characters drawn from the English language alphabet: WHAT, WHY

WHAT

||

WH-Y

S(WHAT/WH-Y) = 1 + 1 – δ – μ

WHAT

-WHY

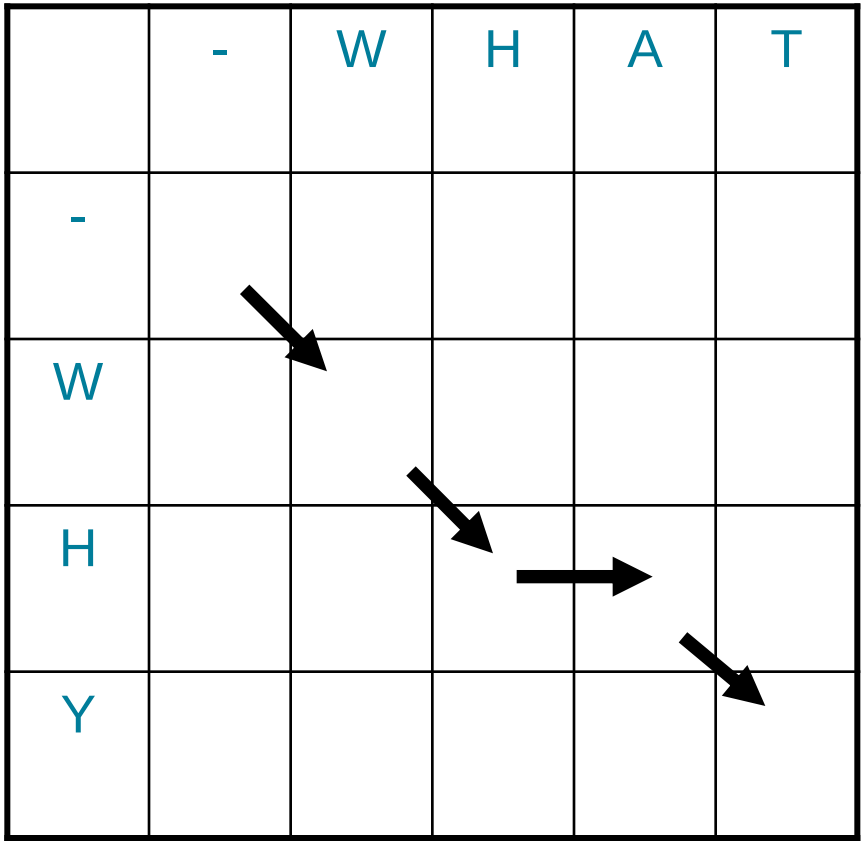
S(WHAT/-WHY) = – δ – μ – μ – μ

Representing alignments and scores

Alignments can be represented in the following tabular form.

Each alignment corresponds to a path through the table.

```
WHAT
| |
WH-Y
```

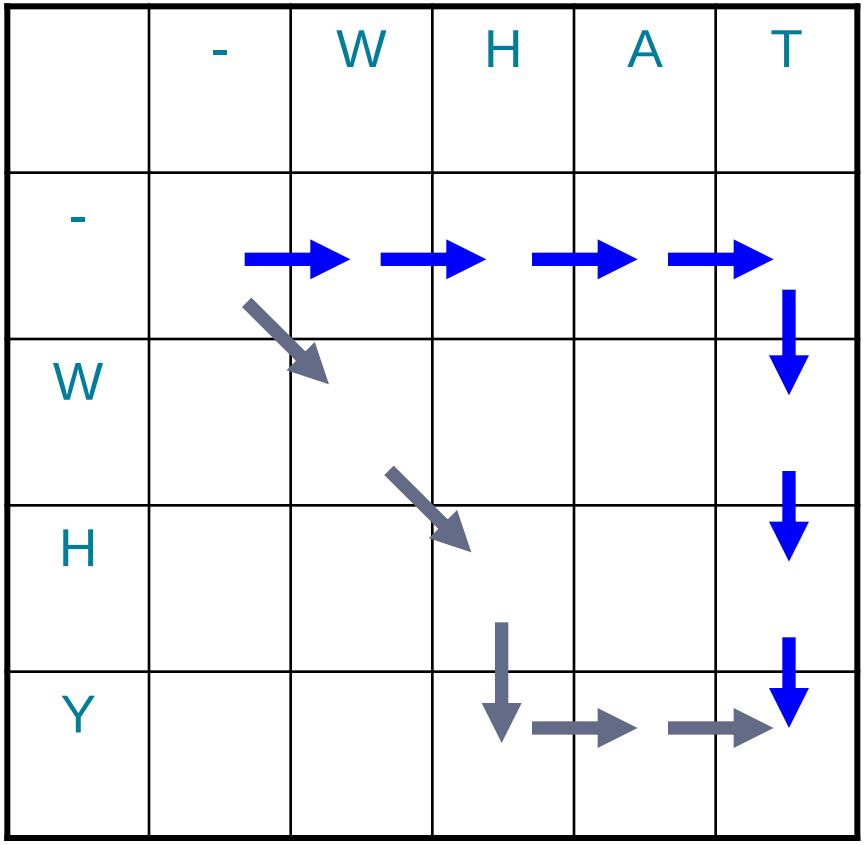


Representing alignments and scores

```
WH-AT
||
WHY--

WHAT---

----WHY
```



Representing alignments and scores

WHAT

| |

WH-Y

Global alignment score $S_{3,4} = 2-\delta-\mu$

	-	W	H	A	T
-	0				
W		1			
H			2	2-δ	
Y					2-δ-μ

Global alignment: formal development

$A = a_1a_2a_3...a_m$,

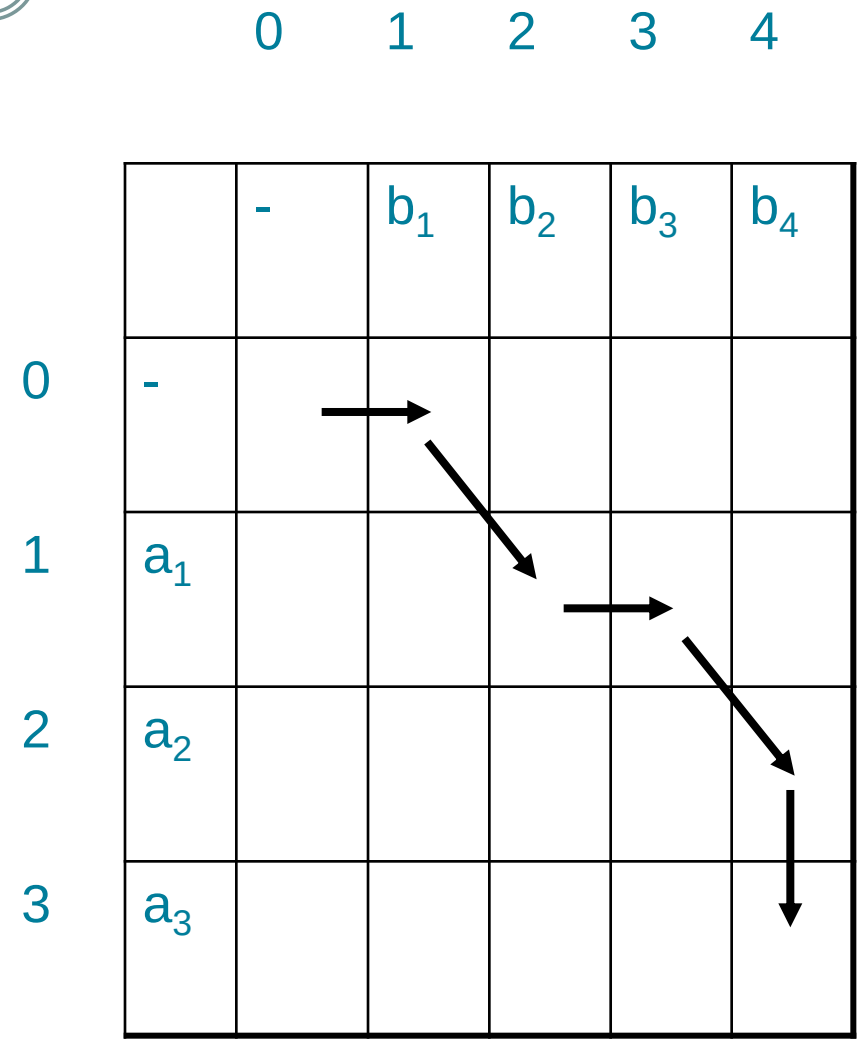
$B = b_1b_2b_3...b_n$

b_1 b_2 b_3 b_4 -

\- a_1 - a_2 a_3

- Any alignment can be written as a unique path through the matrix
- Score for aligning A and B up to positions i and j:

$S_{i,j} = S(a_1a_2a_3…a_i, b_1b_2b_3…b_j)$



Scoring partial alignments

- Alignment of A = $a_1a_2a_3...a_i$ with B = $b_1b_2b_3...b_j$ can end in three possible ways
  - Case 1: $(a_1a_2...a_{i-1})$ a_i
    $(b_1b_2...b_{j-1})$ b_j
  - Case 2: $(a_1a_2...a_{i-1})$ a_i
    $(b_1b_2...b_j)$ -
  - Case 3: $(a_1a_2...a_i)$ –
    $(b_1b_2...b_{j-1})$ b_j

Scoring alignments

- Scores for each case:
  - Case 1: $(a_1 a_2 \ldots a_{i-1})\ a_i$ / $(b_1 b_2 \ldots b_{j-1})\ b_j$
  - Case 2: $(a_1 a_2 \ldots a_{i-1})\ a_i$ / $(b_1 b_2 \ldots b_j)$ –
  - Case 3: $(a_1 a_2 \ldots a_i)$ – / $(b_1 b_2 \ldots b_{j-1})\ b_j$

$$s(a_i, b_j) = \begin{cases} +1 \text{ if } a_i = b_j \\ -\mu \text{ otherwise} \end{cases}$$

$$s(a_i, -) = s(-, b_j) = -\delta$$

Scoring alignments (2)

- First row and first column correspond to initial alignment against indels:

  $S(i, 0) = -i\,\delta$

  $S(0, j) = -j\,\delta$

- Optimal global alignment score $S(A, B) = S_{m,n}$

		0	1	2	3	4
		-	b_1	b_2	b_3	b_4
0	-	0	-δ	-2δ	-3δ	-4δ
1	a_1	-δ				
2	a_2	-2δ				
3	a_3	-3δ				

Algorithm for global alignment

Input sequences A, B, m = |A|, n = |B|

Set $S_{i,0} := -\delta i$ for all i

Set $S_{0,j} := -\delta j$ for all j

for i := 1 to m

  for j := 1 to n

    $S_{i,j} := \max\{S_{i-1,j} - \delta, S_{i-1,j-1} + s(a_i,b_j), S_{i,j-1} - \delta\}$

  end

end

Algorithm takes O(mn) time

Global alignment: example

$\mu = 1$

$\delta = 2$

	-	T	G	G	T	G
-	0	-2	-4	-6	-8	-10
A	-2					
T	-4					
C	-6					
G	-8					
T	-10					?

Global alignment: example

$\mu = 1$

$\delta = 2$

	-	T	G	G	T	G
-	0	-2	-4	-6	-8	-10
A	-2	-1	-3			
T	-4					
C	-6					
G	-8					
T	-10					?

Global alignment: example (2)

μ = 1

δ = 2

	-	T	G	G	T	G
-	0	-2	-4	-6	-8	-10
A	-2	-1	-3	-5	-7	-9
T	-4	-1	-2	-4	-4	-6
C	-6	-3	-2	-3	-5	-5
G	-8	-5	-2	-1	-3	-4
T	-10	-7	-4	-3	0	-2

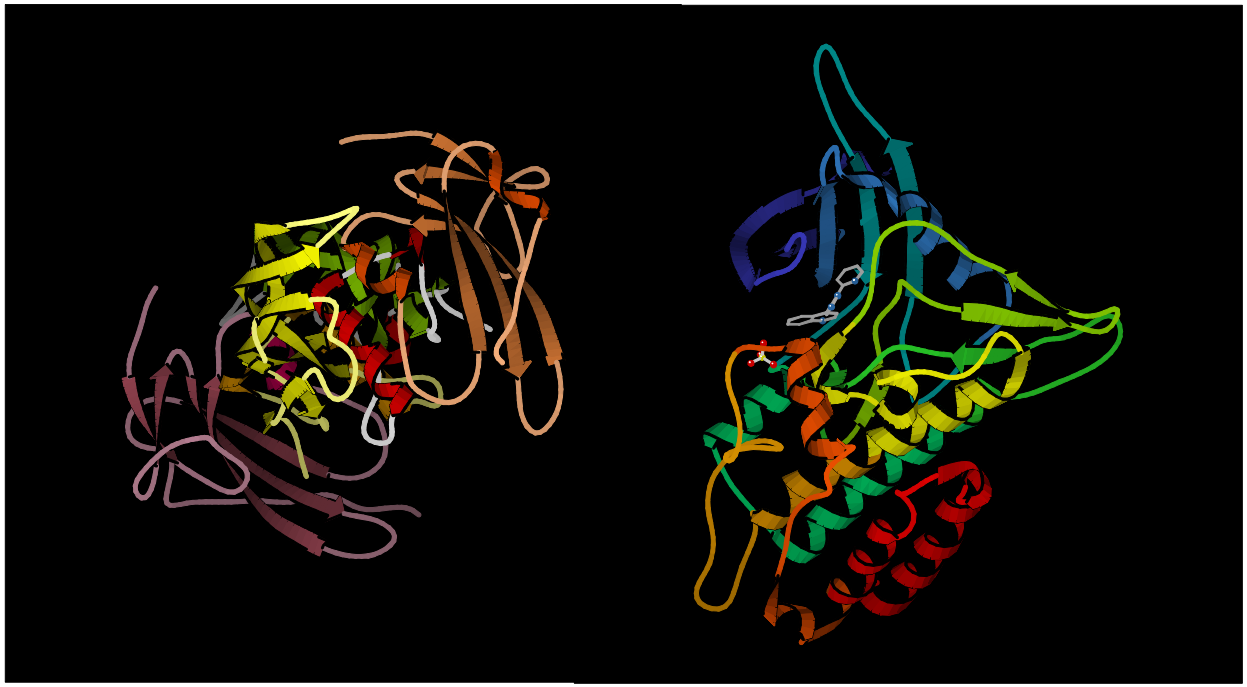
```
ATCGT-
|  ||
-TGGTG
```

Local alignment: rationale

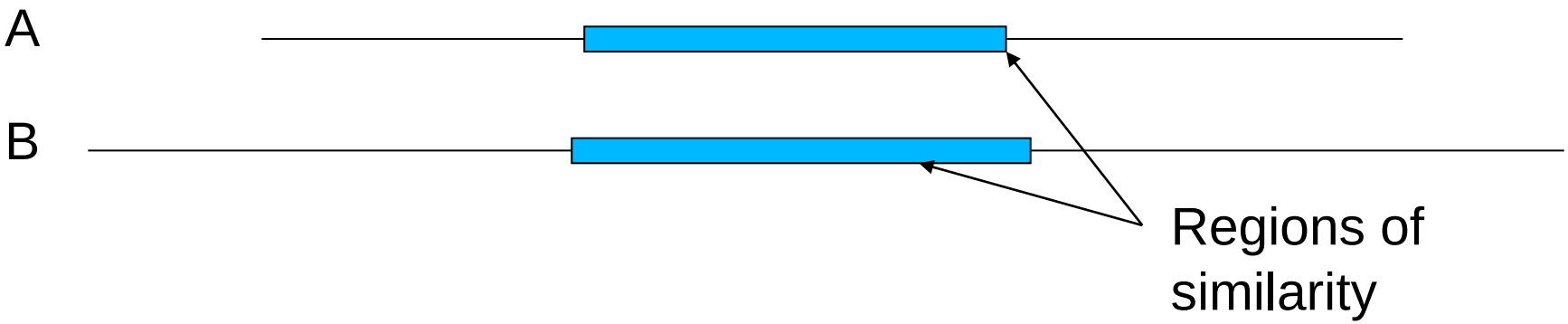
- Otherwise dissimilar proteins may have local regions of similarity
  -> Proteins may share a function

Human bone morphogenic protein receptor type II precursor (left) has a 300 aa region that resembles 291 aa region in TGF-β receptor (right).

The shared function here is protein kinase.



Local alignment: rationale



- Global alignment would be inadequate
- Problem: find the highest scoring *local* alignment between two sequences
- Previous algorithm with minor modifications solves this problem (Smith & Waterman 1981)

From global to local alignment

- Modifications to the global alignment algorithm
  - Look for the highest-scoring path **in** the alignment matrix (not necessarily through the matrix), or in other words:
  - Allow preceding and trailing indels without penalty

Recursion for local alignment

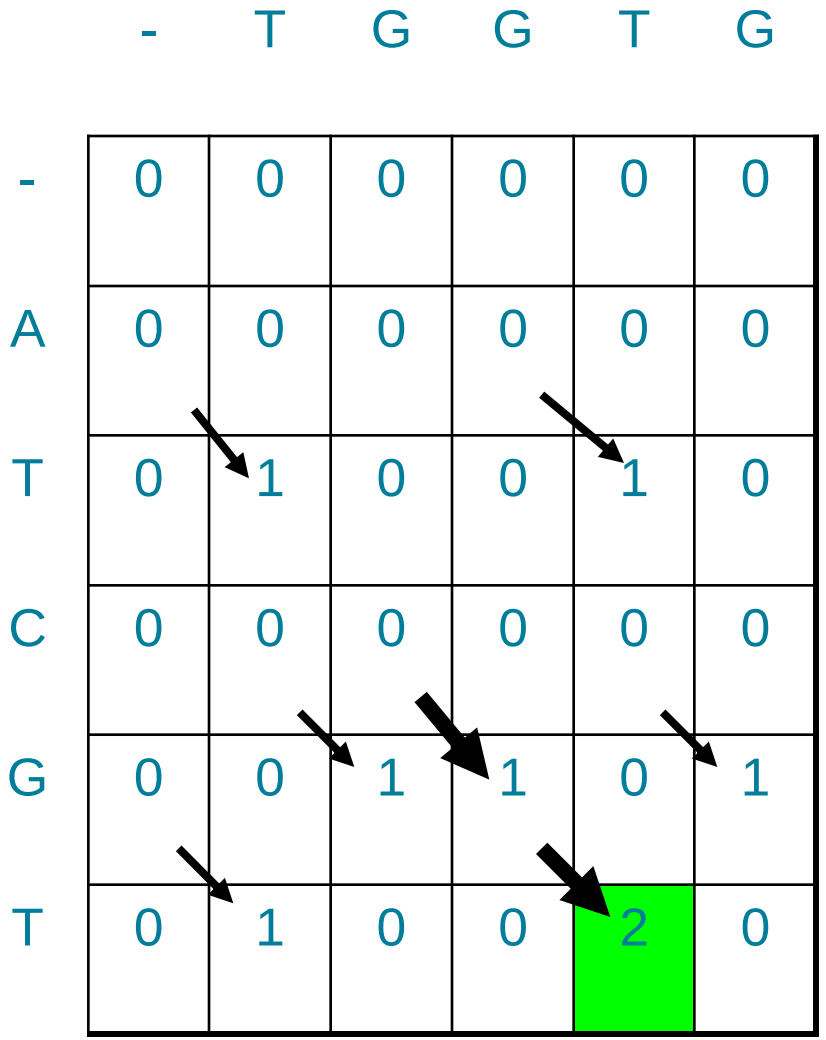
- $M_{i,j} = \max \{$
  $M_{i-1,j-1} + s(a_i, b_i)$,
  $M_{i-1,j} - \delta$,
  $M_{i,j-1} - \delta$,
  0
  $\}$

Allow alignment to start anywhere in sequences

	-	T	G	G	T	G
-	0	0	0	0	0	0
A	0	0	0	0	0	0
T	0	1	0	0	1	0
C	0	0	0	0	0	0
G	0	0	1	1	0	1
T	0	1	0	0	2	0

Finding best local alignment

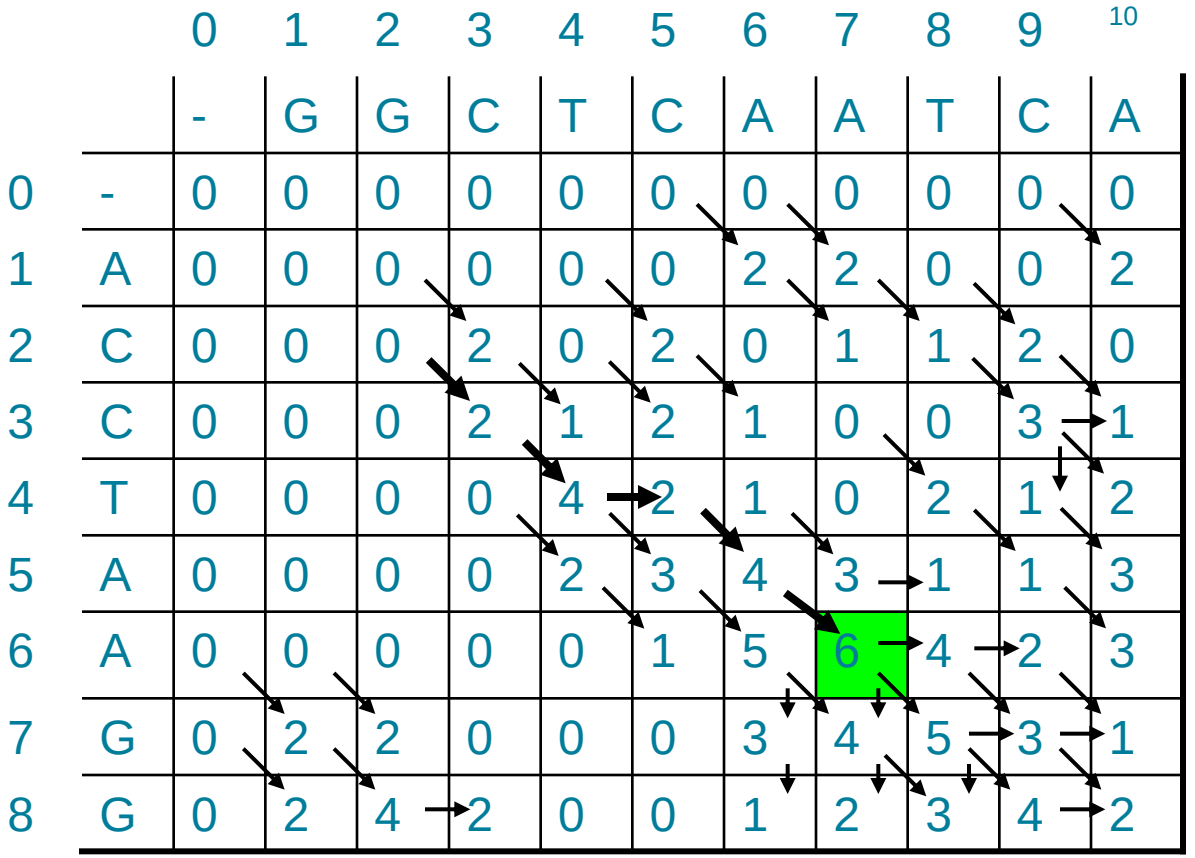
- Optimal score is the highest value in the matrix
- Best local alignment can be found by backtracking from the highest value in M



Local alignment: example

Optimal local alignment:

```
C T – A A
C T C A A
```



		0	1	2	3	4	5	6	7	8	9	10
		-	G	G	C	T	C	A	A	T	C	A
0	-	0	0	0	0	0	0	0	0	0	0	0
1	A	0	0	0	0	0	0	2	2	0	0	2
2	C	0	0	0	2	0	2	0	1	1	2	0
3	C	0	0	0	2	1	2	1	0	0	3	1
4	T	0	0	0	0	4	2	1	0	2	1	2
5	A	0	0	0	0	2	3	4	3	1	1	3
6	A	0	0	0	0	0	1	5	6	4	2	3
7	G	0	2	2	0	0	0	3	4	5	3	1
8	G	0	2	4	2	0	0	1	2	3	4	2

Scoring (for example)
Match: +2
Mismatch: -1
Indel: -2

Non-uniform mismatch penalties

- We used uniform penalty for mismatches:

  s('A', 'C') = s('A', 'G') = … = s('G', 'T') = μ

- Transition mutations (A<->G, C<->T) are approximately twice as frequent than transversions (A<->C, A<->T, C<->G, G<->T)
  - use non-uniform mismatch penalties collected into a *substitution matrix*

	A	C	G	T
A	1	-1	-0.5	-1
C	-1	1	-1	-0.5
G	-0.5	-1	1	-1
T	-1	-0.5	-1	1

Gaps in alignment

- Gap is a succession of indels in alignment

```
C T - - - A A
C T C G C A A
```

- Previous model scored a length k gap as $w(k) = -k\delta$
- Replication processes may produce longer stretches of insertions or deletions
  - In coding regions, insertions or deletions of codons may preserve functionality

Affine gap open and extension penalties

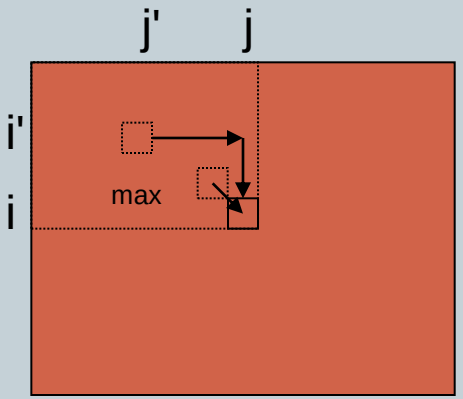
- We can design a score that allows the penalty opening gap to be larger than extending the gap:

$$w(k) = -\alpha - \beta(k - 1)$$

- Gap open cost α, Gap extension cost β
- Alignment algorithms can be extended to use $w(k)$ as follows:

$$S_{i,j} = \max_{i' \le i,\ j' \le j, i'+j' < i+j} \{ S_{i-1,j-1} + s(a_i, b_j),\ S_{i',j'} + w(j - j' + i - i') \}$$

- However, this is too inefficient: $O(m^2n^2)$.

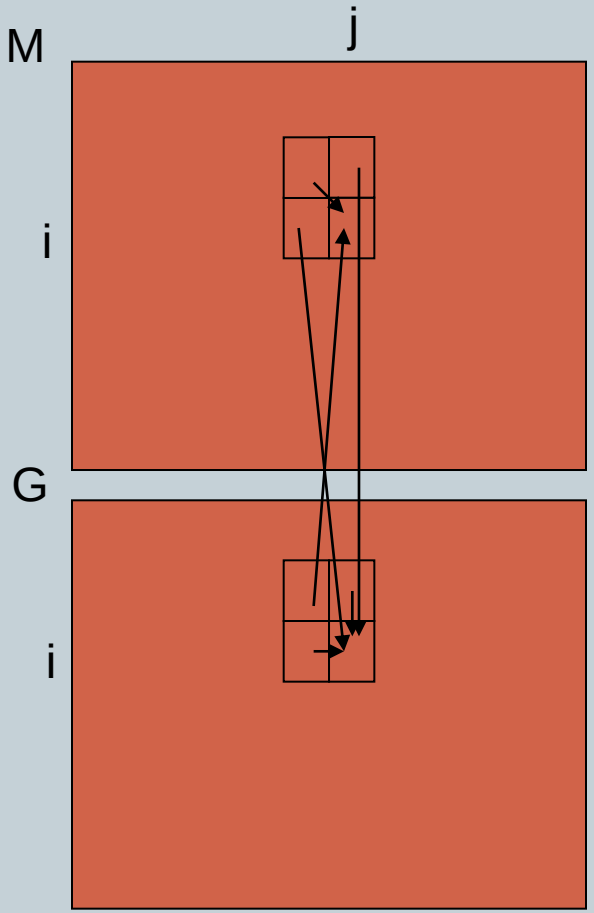


Speeding up affine gap open and extension alignment

$$M_{i,j} = \max \{ M_{i-1,j-1} + s(a_i, b_j),\ G_{i-1,j-1} + s(a_i, b_j) \}$$

$$G_{i,j} = \max \{ M_{i-1,j} - \alpha,\ G_{i-1,j} - \beta,\ M_{i,j-1} - \alpha,\ G_{i,j-1} - \beta \}$$

- Equivalent result in $O(mn)$ time:
  - Exercise: Show that $S_{m,n} = \max(M_{m,n}, G_{m,n})$ with suitable initialization of the tables.



Demonstration of the EBI web site

- European Bioinformatics Institute (EBI) offers many biological databases and bioinformatics tools at http://www.ebi.ac.uk/
  - Sequence alignment: Tools -> Sequence Analysis -> Align