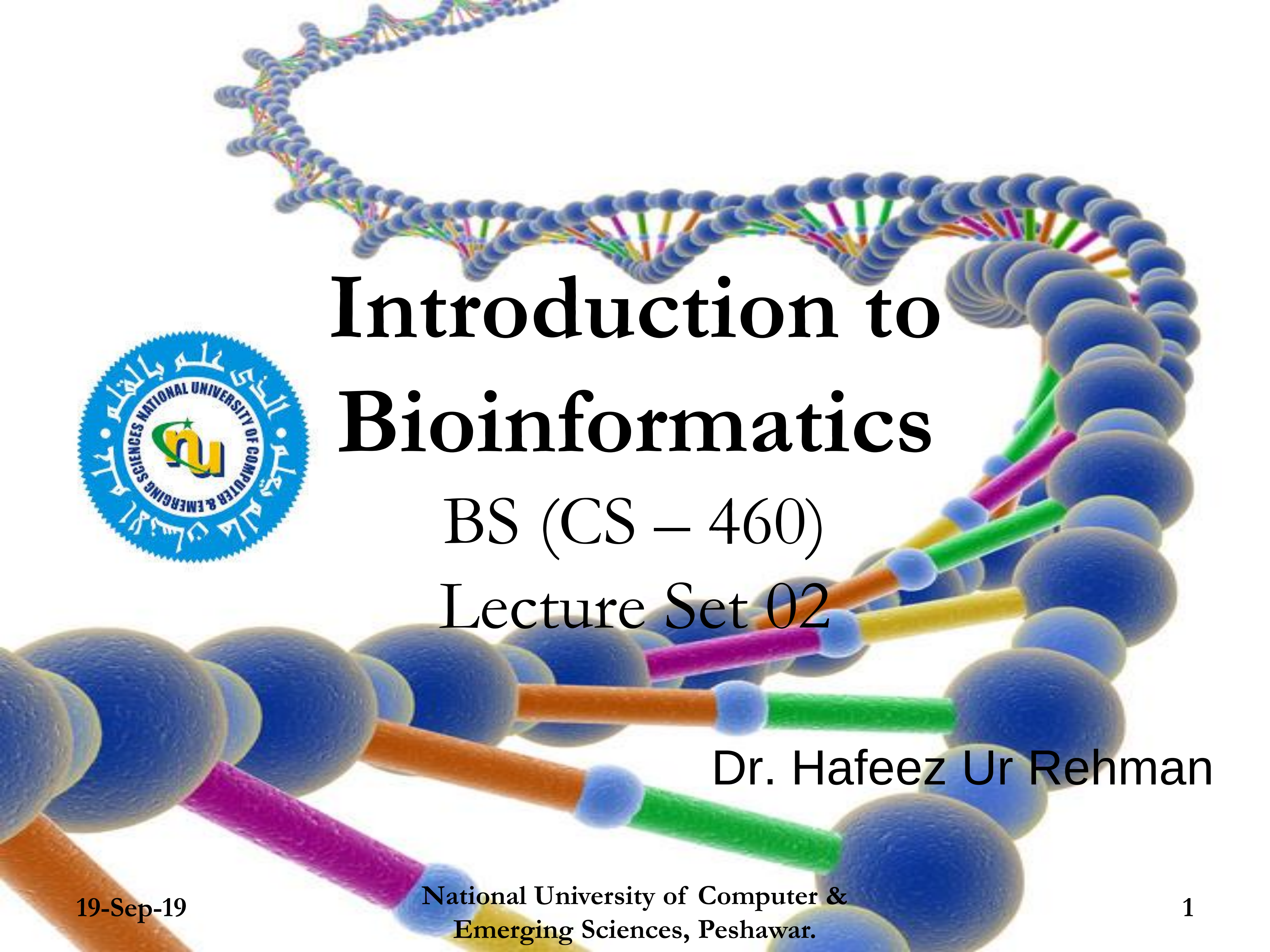


Introduction to Bioinformatics

BS (CS – 460)

Lecture Set 02

Dr. Hafeez Ur Rehman

Topics

1. Gene Ontology Classification Scheme

2. Introduction to Sequence Alignment

What is an Ontology?

- An **ontology** is a formal explicit description of concepts in a domain of discourse, properties of each concept describing various features and attributes of the concept and restrictions on them.
- An ontology together with a set of individual **instances** of classes (concepts) constitutes a **knowledge base**.

Ontology why develop it?

- **Why develop an ontology:**

1. To share common understanding of the structure of information among people or software agents
2. To enable reuse of domain knowledge
3. To make domain assumptions explicit
4. To separate domain knowledge from the operational knowledge
5. To analyze domain knowledge

Why Gene Ontology?

- Elusive (undescribable) nature of protein functions.
- Protein function is an umbrella term for all types of activities that a protein is involved in, be it **cellular**, **molecular** or **physiological**.
- The three categories are not independent, but rather are hierarchically related.
- The need for a standardized functional labeling scheme was paramount, and several groups responded to this need with very innovative proposals.

Gene Ontology (GO)

- URL: http://www.geneontology.org/
- Gene Ontology is
  - A hierarchy of **roles of genes and gene products,** independent of any organism.
  - Composed of three ***independent*** ontologies: **molecular function**, **biological process**, **cellular component**
  - GO itself ***does not contain*** any information on genes or gene products

GO annotations

- http://www.geneontology.org/GO.current.annotations.shtml
- Curators annotate their findings of genes (known as **annotations**) by utilizing GO for various organisms (about 20 of them).
- Different kinds of evidence codes
  - Annotations with IEA (inferred from electronic annotation) evidence code are not manually verified (**Least reliable**)

Properties of Gene Ontology?

(1) **Wide coverage:** This is the most important property, since any functional scheme should cover as many of the functional phenomena across as many the organisms as possible.

(2) **Standardized format:** Having minimal variability in the functional labels and adopting a standard data structure for the scheme makes the scheme easily readable by computer programs and significantly enhances their impact.

(3) **Hierarchical structure:** Protein functions do not form a flat list, but are instead arranged hierarchically at a conceptual level. Functional classes range from specific functions to very general functional categories, thus allowing a researcher to choose the appropriate level(s) for his analysis.

Properties of Gene Ontology? (Contd…)

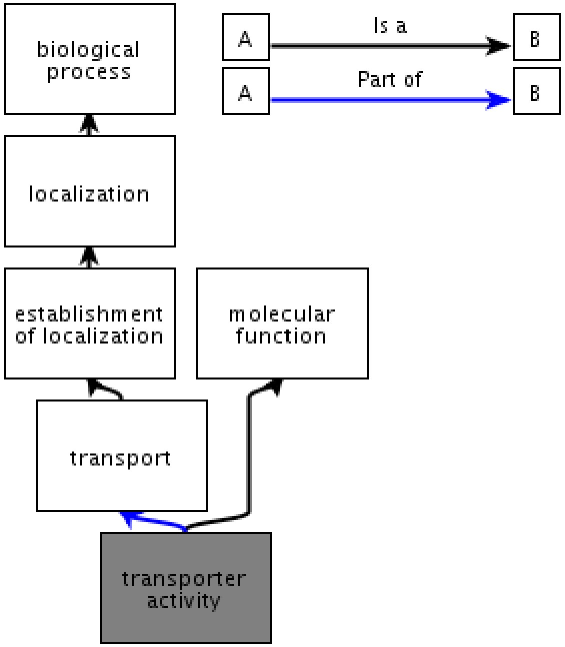
(4) **Disjoint categories:** Functions can be of different types, such as **cellular component**, **molecular function** and **biological process**. A separate hierarchy allows the choice of the appropriate type of function to be studied.

(5) **Multiple functions:** In order to model the biological possibility of a protein being involved in multiple biological processes depending on the context, it is necessary for a functional scheme to allow the labeling of a single protein with multiple functions.

(6) **Dynamic nature:** Last but not the least, the scheme should not be static, but should be modified as and when new functional knowledge is discovered.

GO **Molecular Function** Ontology

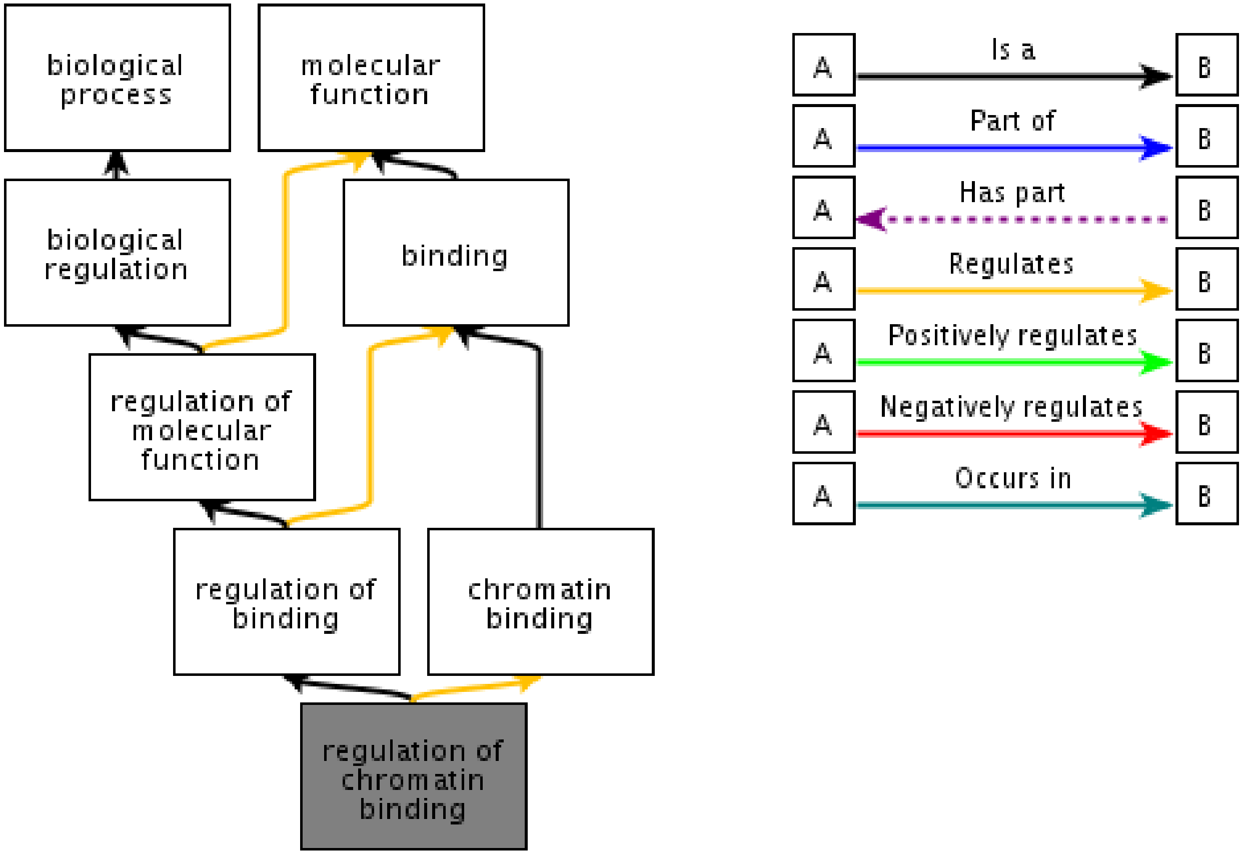
- Describes activities, such as catalytic or binding activities, that can be performed by individual gene products or assembled complexes of gene products **at the molecular level**.
- Example of activities
  - transporter activity
    - Genes that enable the directed movement of substances (such as macromolecules, small molecules, ions) into, out of, within or between cells.
- Example of binding
  - insulin receptor binding
    - Genes that interact with insulin receptors



GO **Biological Process** Ontology

- Defined as a **biological objective to which the gene or gene product contributes**.
- Examples
  - Cell Proliferation
    - Genes that are responsible for the multiplication or reproduction of cells, resulting in the rapid expansion of a cell population.
  - Regulation of Chromatin Binding
    - Genes that enable regulation and processing of chromatin binding.

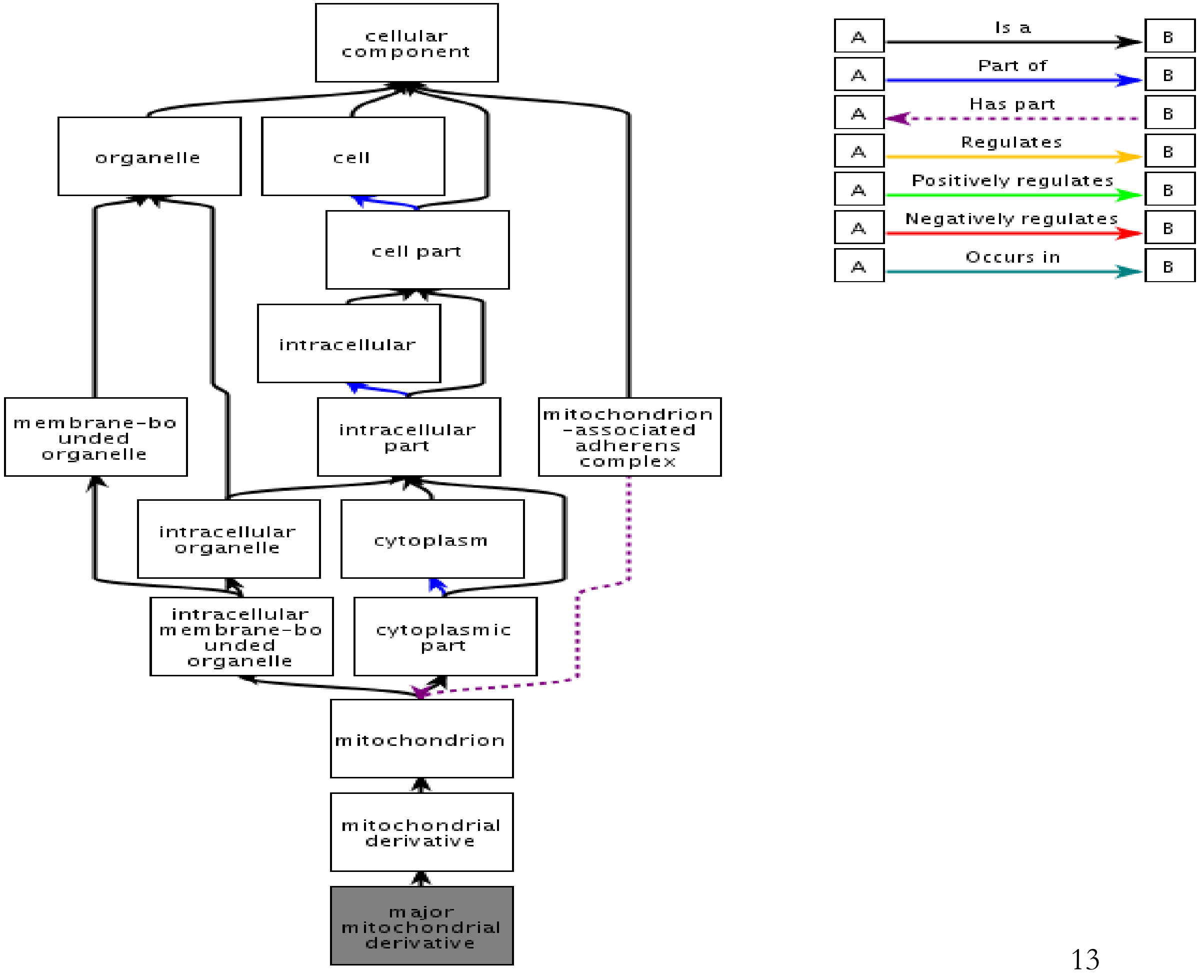
GO **Biological Process** Ontology



QuickGO - http://www.ebi.ac.uk/QuickGO

GO **Cellular Component** Ontology

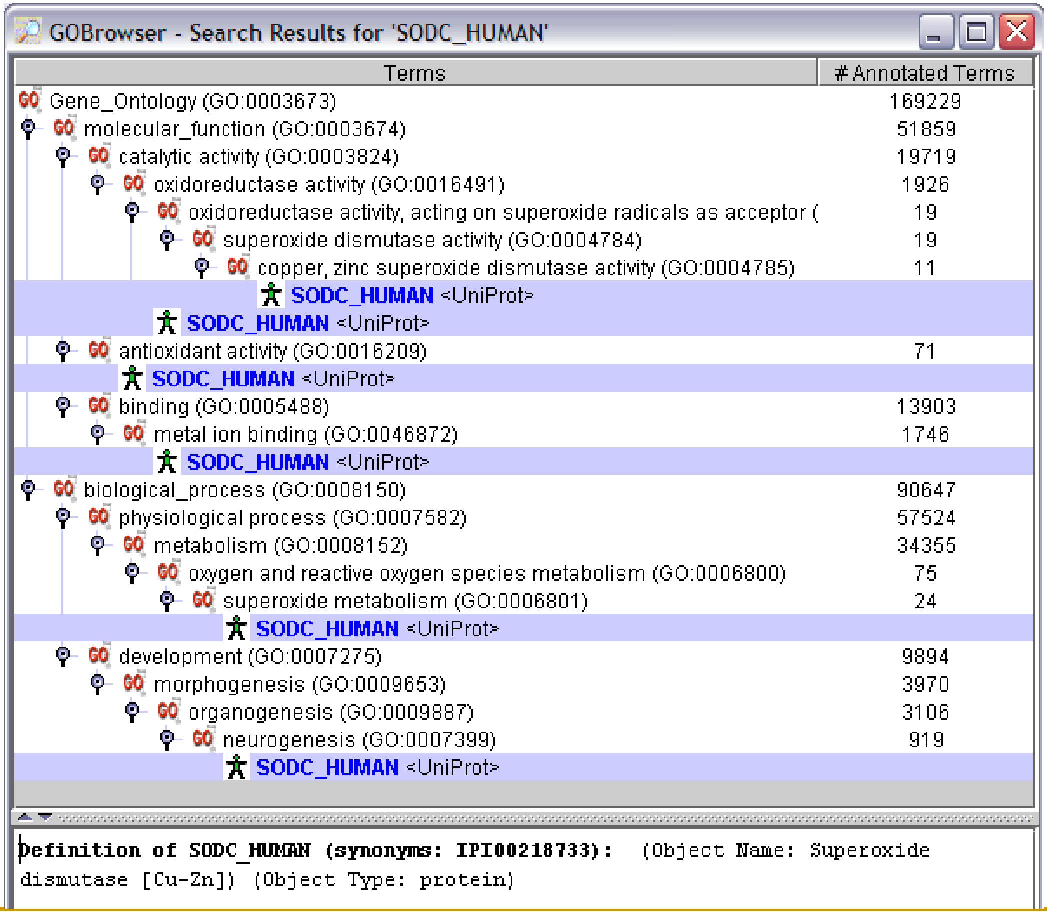
- Refers to the place in the cell where the gene product is active.
- Examples
  - ❑ mitocondria
  - ❑ nucleus
  - ❑ cell membrane



QuickGO - http://www.ebi.ac.uk/QuickGO

Example of a gene product

- A gene product has one or more molecular functions and is used in one or more biological processes; it might be associated with one or more cellular components.



GOBrowser - Search Results for 'SODC_HUMAN'

Terms	# Annotated Terms
Gene_Ontology (GO:0003673)	169229
molecular_function (GO:0003674)	51859
catalytic activity (GO:0003824)	19719
oxidoreductase activity (GO:0016491)	1926
oxidoreductase activity, acting on superoxide radicals as acceptor (	19
superoxide dismutase activity (GO:0004784)	19
copper, zinc superoxide dismutase activity (GO:0004785)	11
SODC_HUMAN <UniProt>	
SODC_HUMAN <UniProt>	
antioxidant activity (GO:0016209)	71
SODC_HUMAN <UniProt>	
binding (GO:0005488)	13903
metal ion binding (GO:0046872)	1746
SODC_HUMAN <UniProt>	
biological_process (GO:0008150)	90647
physiological process (GO:0007582)	57524
metabolism (GO:0008152)	34355
oxygen and reactive oxygen species metabolism (GO:0006800)	75
superoxide metabolism (GO:0006801)	24
SODC_HUMAN <UniProt>	
development (GO:0007275)	9894
morphogenesis (GO:0009653)	3970
organogenesis (GO:0009887)	3106
neurogenesis (GO:0007399)	919
SODC_HUMAN <UniProt>	

Definition of SODC_HUMAN (synonyms: IPI00218733): (Object Name: Superoxide dismutase [Cu-Zn]) (Object Type: protein)

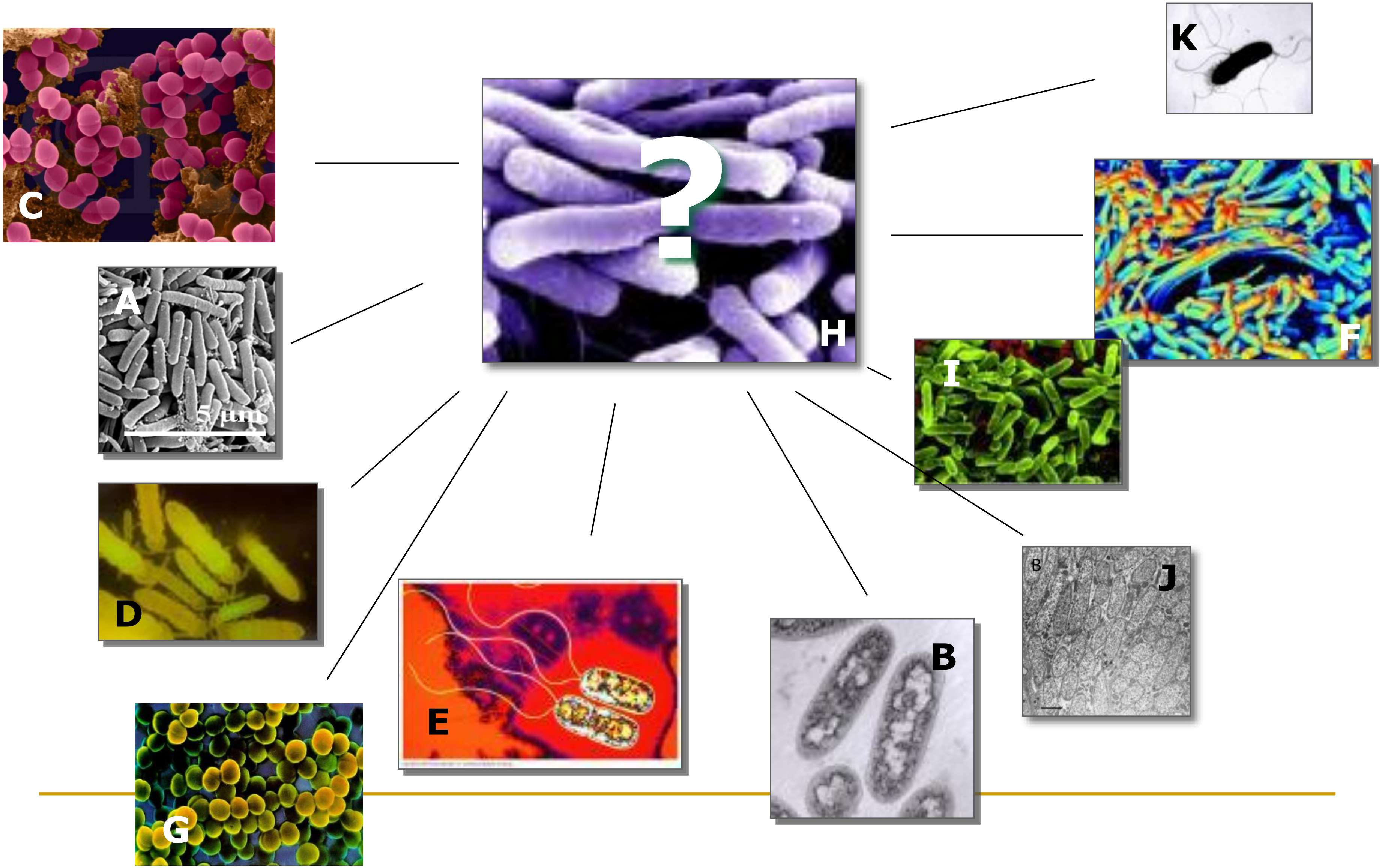
An example showing all occurrences of SODC protein in the Gene Ontology from the human annotation.

Common applications of GO

- Analysis of microarray data
  - Finding genes with similar functions
  - Utilize **biological process** ontology
- Evaluation of protein-protein interactions
  - Proteins are likely to interact if they are in the same location
  - Utilize **cellular component** ontology
- Studying Protein Binding Sites
  - Utilize **molecular function** ontology

Homology and Sequence Alignment

How is this organism related to other species?

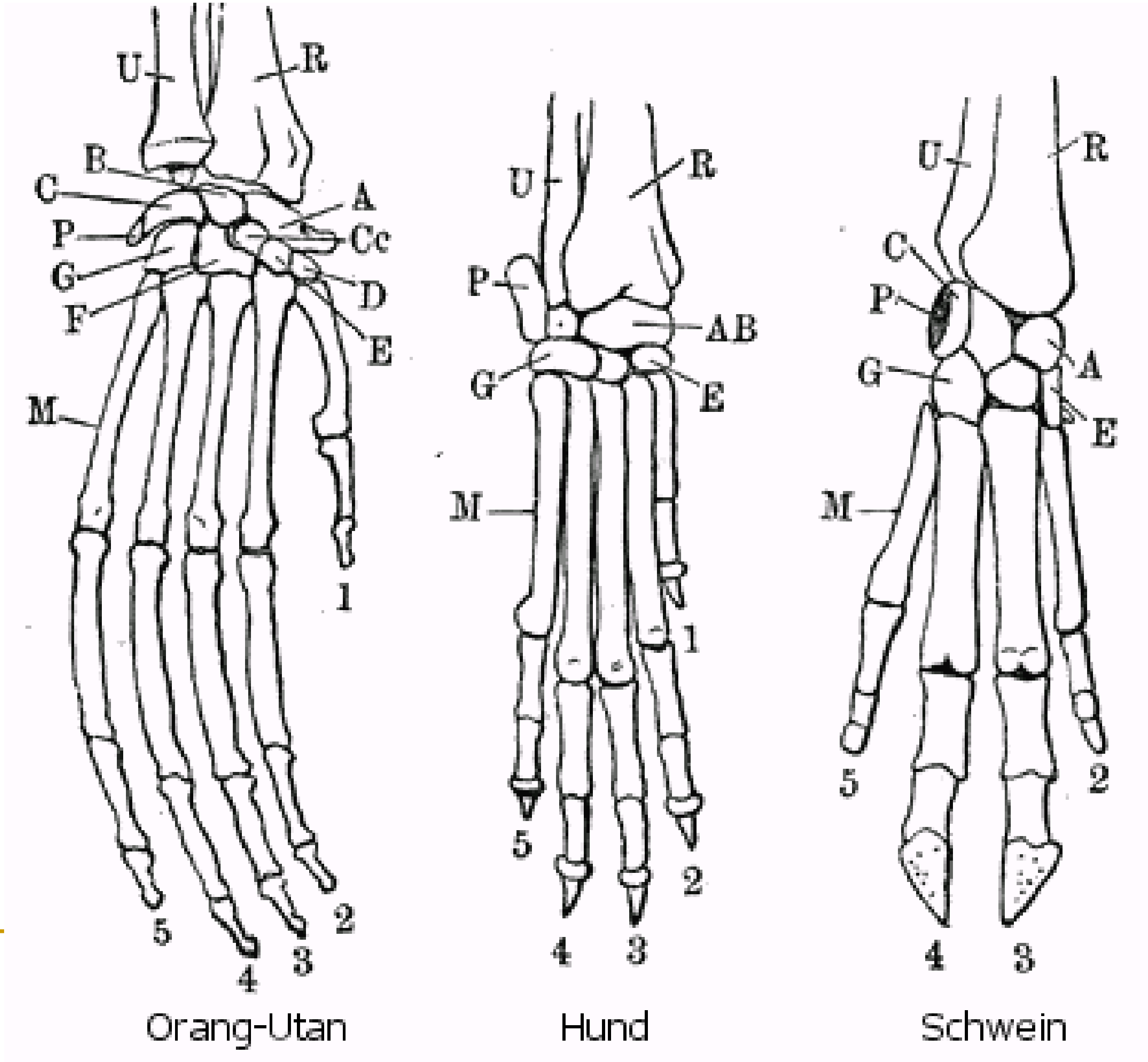


Homology

Homology = Similarity between objects due to a common ancestry

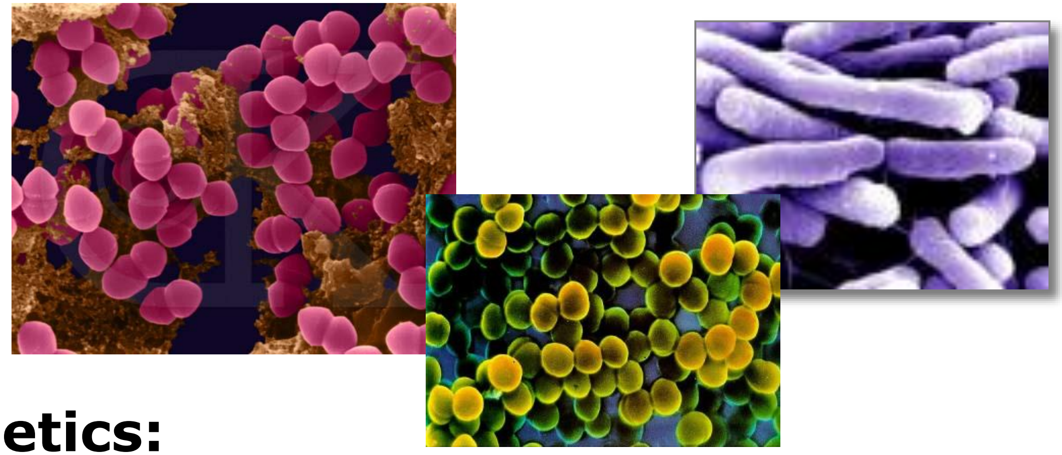
Hund = Dog,
Schwein = Pig
Orang-Utan=Ape

Ape **Dog** **Pig**



DNA sequences provide characters that are:

- Numerous, discrete characters (A, T, C, G)
- Directly comparable across species
- Unlikely to change due to culture conditions



Molecular phylogenetics:

Inference of evolutionary relationships based on molecular data

```
ACAGATGTCTTGTAATCCGGCCGTTGGTGGCATAGGGAAAGGACATTTAGTGAAAGAAATTGATGCGATGGGTGGATCGATGGCTTATGCTATCGA
TCAATCAGGAATTCAATTTAGAGTACTTAATAGTAGCAAAGGAGCTGCTGTTAGAGCAACACGTGCTCAGGCAGATAAAATATTATATCGTCAAGA
CAGATGTCTTGTAATCCGGCCGTTGGTGGCATAGGGAAAGGACATTTAGTGAAAGAAATTGATGCGATGACAGATGTCTTGTAATCCGGCCGTTGG
TGGCATAGGGAAAGGACATTTAGTGAAAGAAATTGATGCGATGGGTGGATCGATGGCTTATGCTATCGATCAATCAGGAATTCAATTTAGAGTACT
TAATAGTAGCAAAGGAGCTGCTGTTAGAGCAACACGTGCTCAGGCAGATAAAATATTATATCGTCAAGCAATACGTGGTGGATCGATGGCTTATGC
TATCGATCAATCAGGAATTCAATTTAGAGTACTTAATAGTAGCAAAGGAGCTGCTGTTAGAGCAACACGTGCTCAGGCAGATAAAATATTATATCG
TCAAGCAATACGTACAGATGTCTTGTAATCCGGCCGTTGGTGGCATAGGGAAAGGACATTTAGTGAAAGAAATTGATGCGATGGGTGGATCGATG
GCTTATGCTATCGATCAATCAGGAATTCAATTTAGAGTACTTAATAGTAGCAAAGGAGCTGCTGTTAGAGCAACACGTGCTCAGGCAGATAAAATA
TTATATCGTCAAGCAATACGTCAATACGCGTGCTCAGGCAGATAAAATATTA
```

Sequence homology

Similarity between sequences as a result of common ancestry.

```
VLSPAVKWAKVGAHAAGHG
||| || |||| |  ||||
VLSEAVLWAKVEADVAGHG
```

Sequence alignment

Alignment: Comparing two (pairwise) or more (multiple) sequences. Searching for a series of identical or similar characters in the sequences.

Why align?

```
VLSPAVKWAKV
||| || ||||
VLSEAVLWAKV
```

1. To detect if two sequences are homologous. If so, homology may indicate **similarity in function** (and structure).

2. Required for **evolutionary studies** (e.g., taxonomic tree reconstruction).

3. To **detect conservation** (e.g., a tyrosine (Y) amino acid that is evolutionary conserved, is more likely to be a phosphorylation site).

4. Given a sequenced DNA, from an unknown region, **align** it to the genome.

Insertions, deletions, and substitutions
What makes them?

A Note about Mutations

A **mutation** is a permanent change in the DNA sequence of a gene. Sometimes mutations can be useful but mostly they are harmful as changes in DNA can change the way a cell behaves. As genes are a set of hereditary materials that contain instructions necessary for a cell to work so if some of these instructions go wrong the cell may not know how to function.

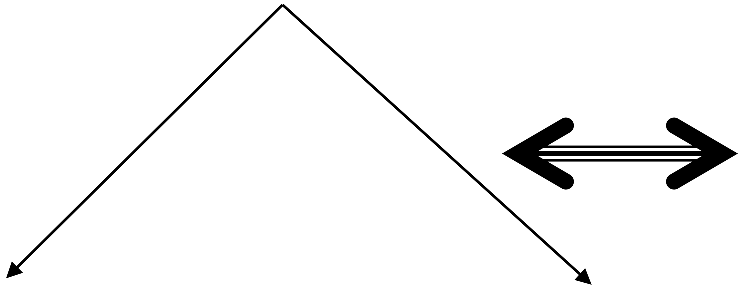
Mutations can also be **acquired** depending on what sort of environment a person lives in as some environmental agents can damage the DNA or when mistakes occur during cell division. For example **radiations released during the nuclear disasters** in Hiroshima and Nagasaki and Chernobyl are still affecting and causing mutations in the genetic makeup of the people living in those areas. Moreover, different types of cancers are also caused by mutations.

A Note about Mutations

Mutations can be **inherited** which means the mutated genetic code can be passed on to the next generations. For example heart disease, diabetes, stroke or high blood pressure, run in the family. If parents suffered from them, their children may also develop them. Ten million men in the U.S. are colour blind but less than 600,000 American women have the same disability. That's because this mutation is located on the X-chromosome. Men only have one X-chromosome, so that one is enough to induce the condition, but women have two X chromosomes, and they require the mutation in double set to experience the condition.

Sequence alignment

If two sequences share a common ancestor – for example human and dog **hemoglobin**, we can represent their evolutionary relationship using a tree



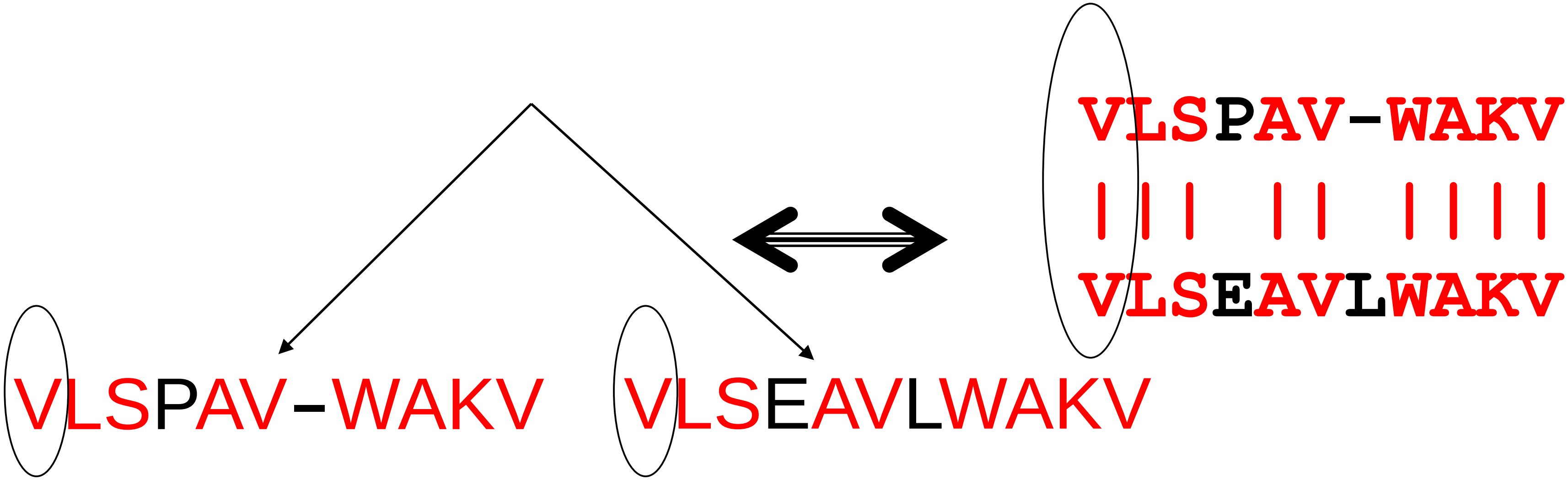
```
VLSPAV-WAKV
||| || ||||
VLSEAVLWAKV
```

VLSPAV–WAKV VLSEAVLWAKV

Human’s Hemoglobin **Dog’s Hemoglobin**

Perfect match

A perfect match suggests that no change has occurred from the common ancestor (although this is not always the case).

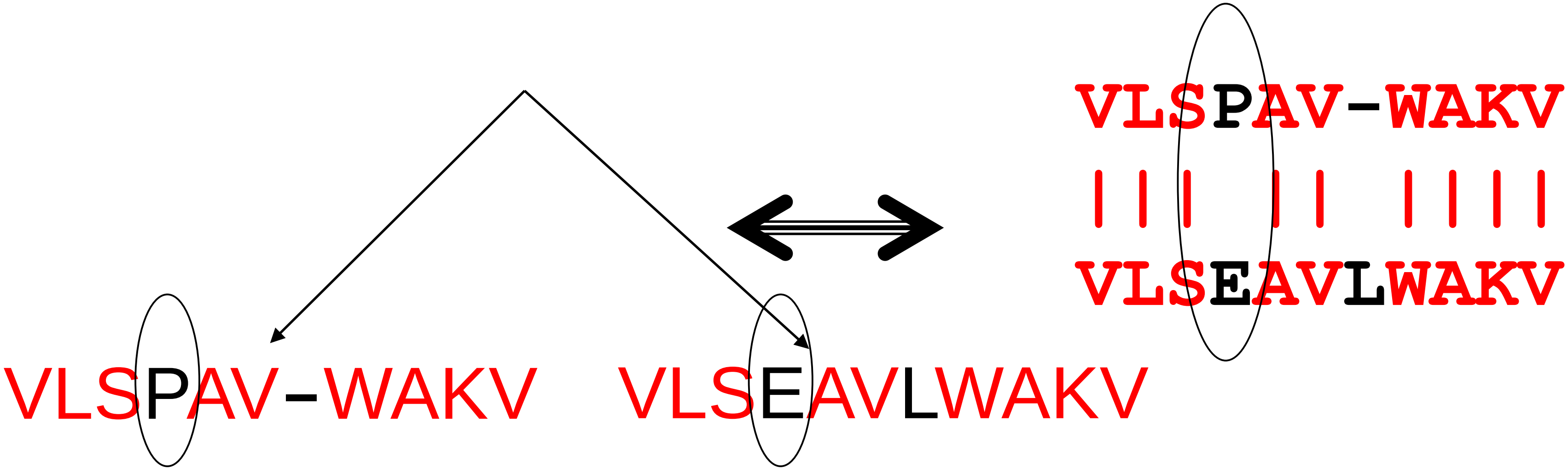


```
VLSPAV-WAKV
||| || ||||
VLSEAVLWAKV
```

VLSPAV-WAKV VLSEAVLWAKV

A substitution

A substitution suggests that at least one change has occurred since the common ancestor (**although we cannot say in which lineage it has occurred**).

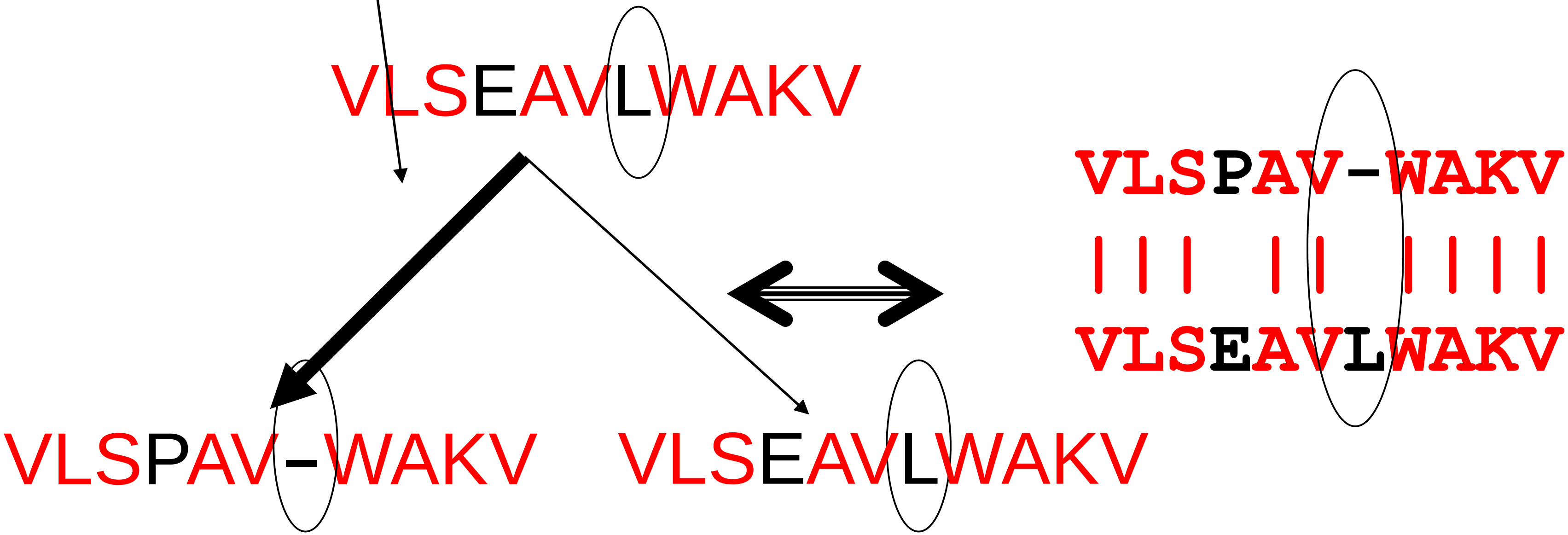


```
VLSPAV-WAKV
||| || ||||
VLSEAVLWAKV
```

VLSPAV-WAKV   VLSEAVLWAKV

Indel

Case 01: The ancestor had L and it was lost here. In such a case, the event was a **deletion**.



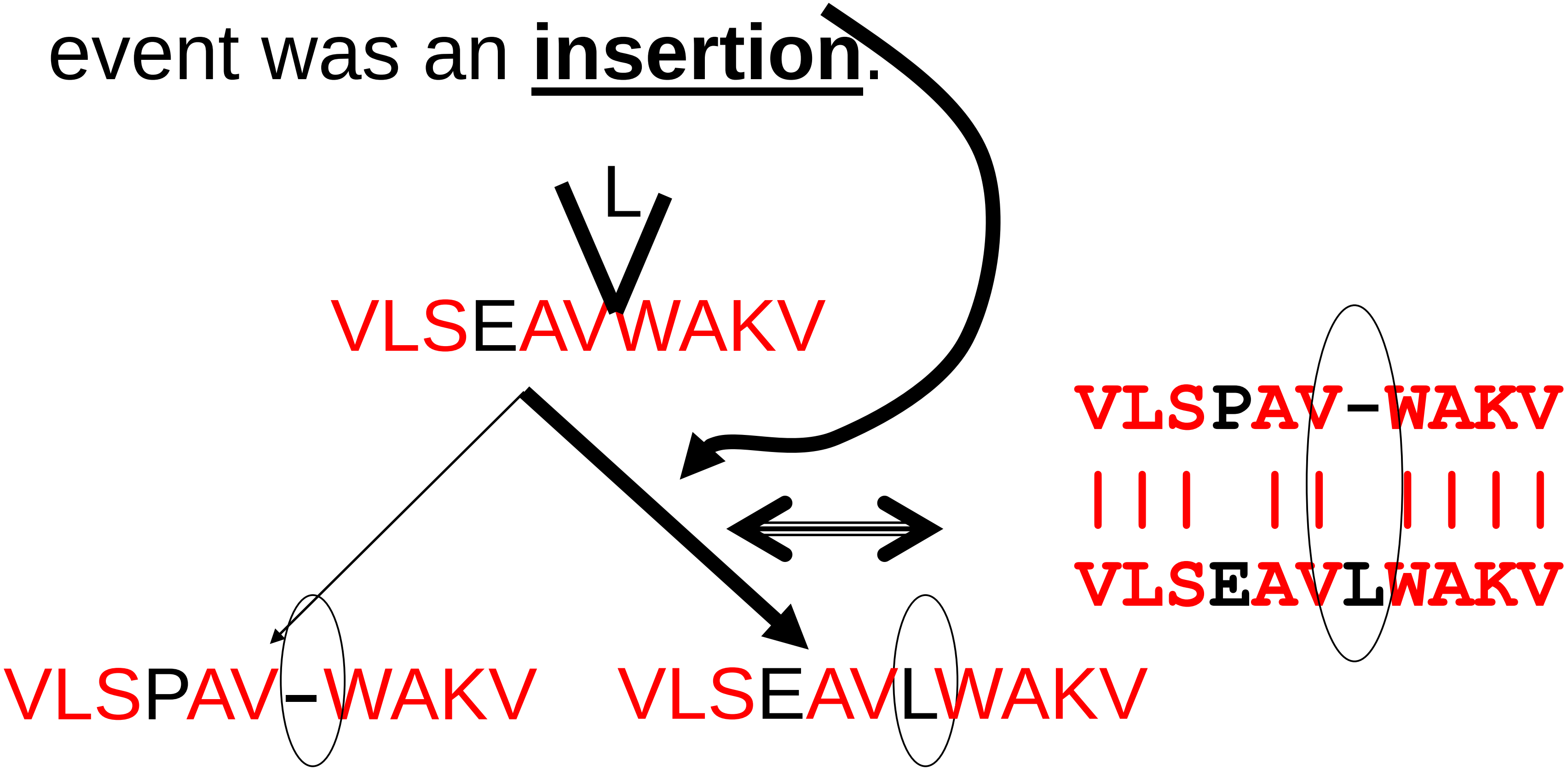
VLSEAVLWAKV

VLSPAV-WAKV VLSEAVLWAKV

```
VLSPAV-WAKV
||| || ||||
VLSEAVLWAKV
```

Indel

Case 02: The ancestor was shorter and the L was inserted here. In such a case, the event was an **insertion**.



L

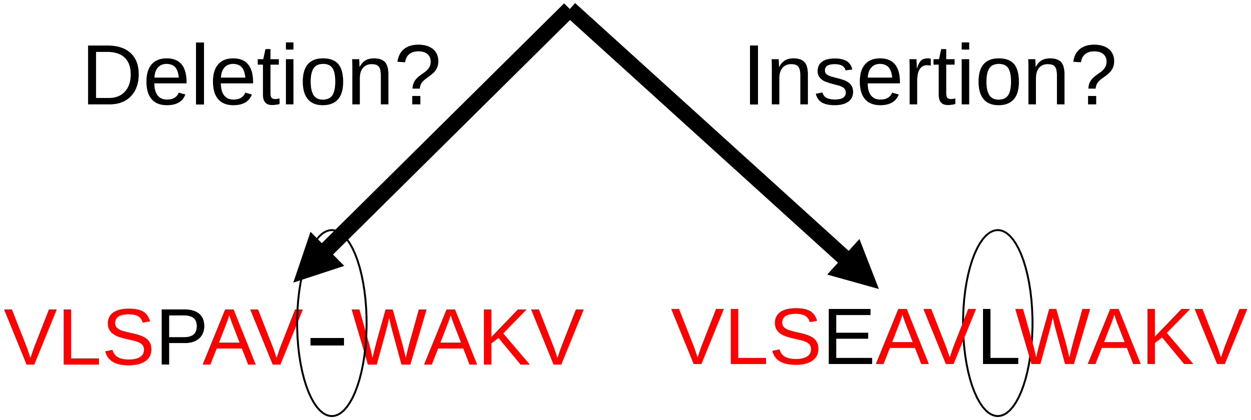
VLSEAVWAKV

VLSPAV-WAKV VLSEAVLWAKV

```
VLSPAV-WAKV
||| || ||||
VLSEAVLWAKV
```

Indel

Normally, given two sequences we cannot tell whether it was an insertion or a deletion, so we term the event as an **indel**.



Deletion?

Insertion?

VLSPAV–WAKV

VLSEAVLWAKV

Indels in protein coding genes

Indels in protein coding genes are often of 3bp, 6bp, 9bp, etc...

Gene Search

In fact, searching for indels of length 3K (K=1,2,3,…) can help algorithms that search a genome for coding regions

Global and Local pairwise alignments

Global vs. Local

- Global alignment – finds the best alignm across the **entire** two sequences.

```
ADLGAVFALCDRYFQ
||||     |||| |
ADLGRTQN-CDRYYQ
```

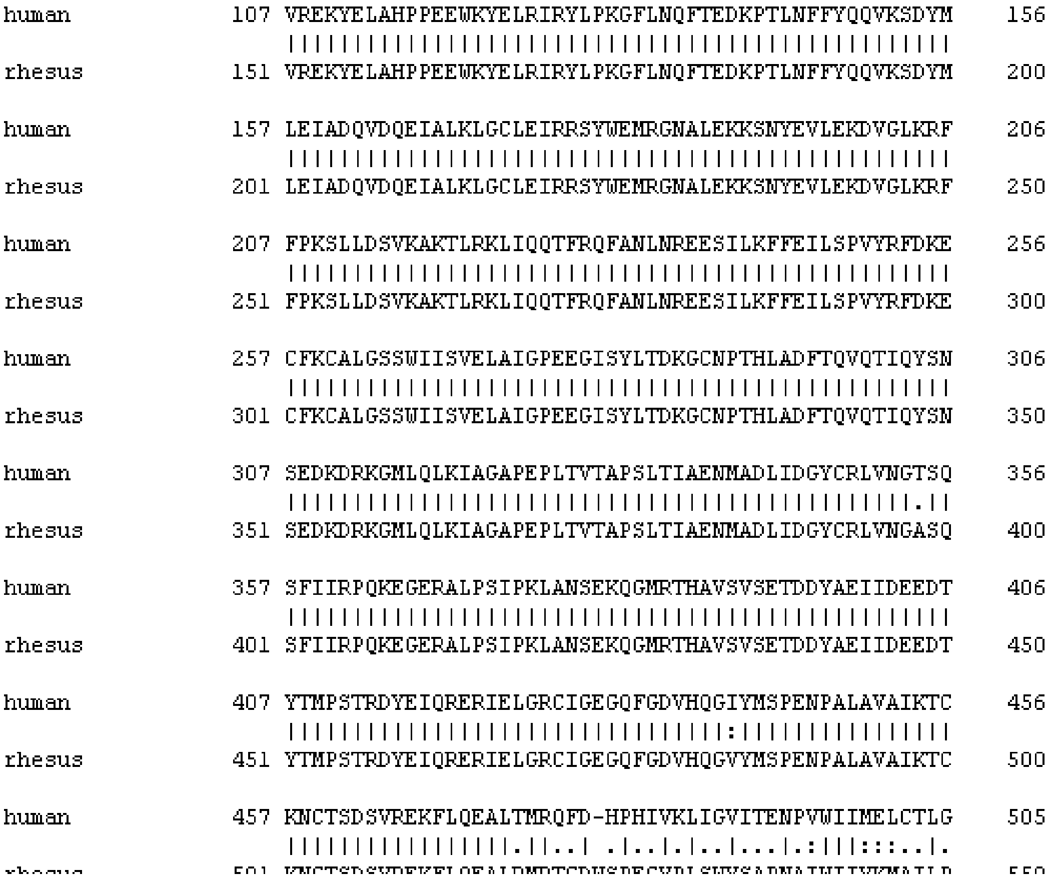
Global alignment: forces alignment in regions which differ

- Local alignment – finds regions of simila **parts** of the sequences.

```
ADLG    CDRYFQ
||||    |||| |
ADLG    CDRYYQ
```

Local alignment will return only **regions** of good alignment

Global alignment



```
human            107 VREKYELAHPPEEWKYELRIRYLPKGFLNQFTEDKPTLNFFYQQVKSDYM    156
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
rhesus           151 VREKYELAHPPEEWKYELRIRYLPKGFLNQFTEDKPTLNFFYQQVKSDYM    200

human            157 LEIADQVDQEIALKLGCLEIRRSYWEMRGNALEKKSNYEVLEKDVGLKRF    206
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
rhesus           201 LEIADQVDQEIALKLGCLEIRRSYWEMRGNALEKKSNYEVLEKDVGLKRF    250

human            207 FPKSLLDSVKAKTLRKLIQQTFRQFANLNREESILKFFEILSPVYRFDKE    256
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
rhesus           251 FPKSLLDSVKAKTLRKLIQQTFRQFANLNREESILKFFEILSPVYRFDKE    300

human            257 CFKCALGSSWIISVELAIGPEEGISYLTDKGCNPTHLADFTQVQTIQYSN    306
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
rhesus           301 CFKCALGSSWIISVELAIGPEEGISYLTDKGCNPTHLADFTQVQTIQYSN    350

human            307 SEDKDRKGMLQLKIAGAPEPLTVTAPSLTIAENMADLIDGYCRLVNGTSQ    356
                     |||||||||||||||||||||||||||||||||||||||||||||||.||
rhesus           351 SEDKDRKGMLQLKIAGAPEPLTVTAPSLTIAENMADLIDGYCRLVNGASQ    400

human            357 SFIIRPQKEGERALPSIPKLANSEKQGMRTHAVSVSETDDYAEIIDEEDT    406
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
rhesus           401 SFIIRPQKEGERALPSIPKLANSEKQGMRTHAVSVSETDDYAEIIDEEDT    450

human            407 YTMPSTRDYEIQRERIELGRCIGEGQFGDVHQGIYMSPENPALAVAIKTC    456
                     |||||||||||||||||||||||||||||||||:||||||||||||||||
rhesus           451 YTMPSTRDYEIQRERIELGRCIGEGQFGDVHQGVYMSPENPALAVAIKTC    500

human            457 KNCTSDSVREKFLQEALTMRQFD-HPHIVKLIGVITENPVWIIMELCTLG    505
                     |||||||||||||||||.||..| .|..|.|..|...|.:|||:::..|.
rhesus           501 KNCTSDSVREKFLQEALPMRTCDHSPECVPLSWVSAPNAIWIIYKMAILR    550
```

PTK2 protein *tyrosine kinase 2* of **human** and **Rhesus monkey**

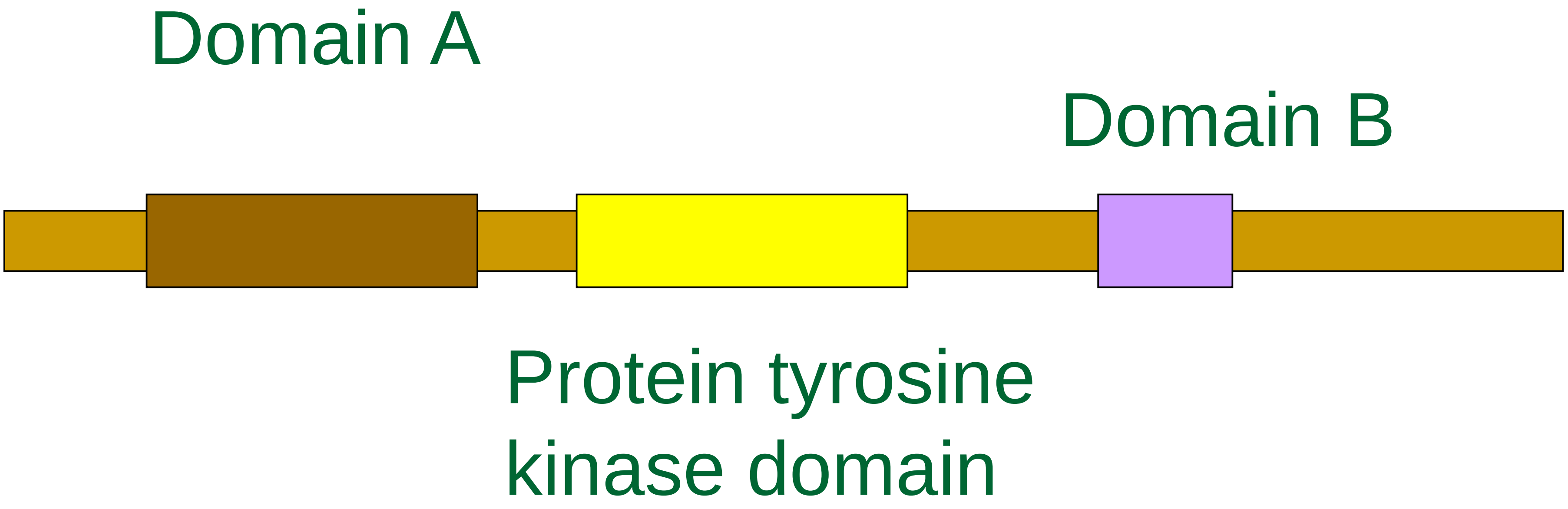
Global alignment

A “:” Conserved amino acid substitutions are the replacement of an amino acid residue with another one with **similar properties**, such as **aspartate** for **glutamate**. They are both negatively charged amino acids.

A “.” semi-conserved amino acid replaces one residue with another one that has similar **steric conformation**, but does not share chemical properties. An example would be substituting **cysteine** for **alanine** or **leucine**.

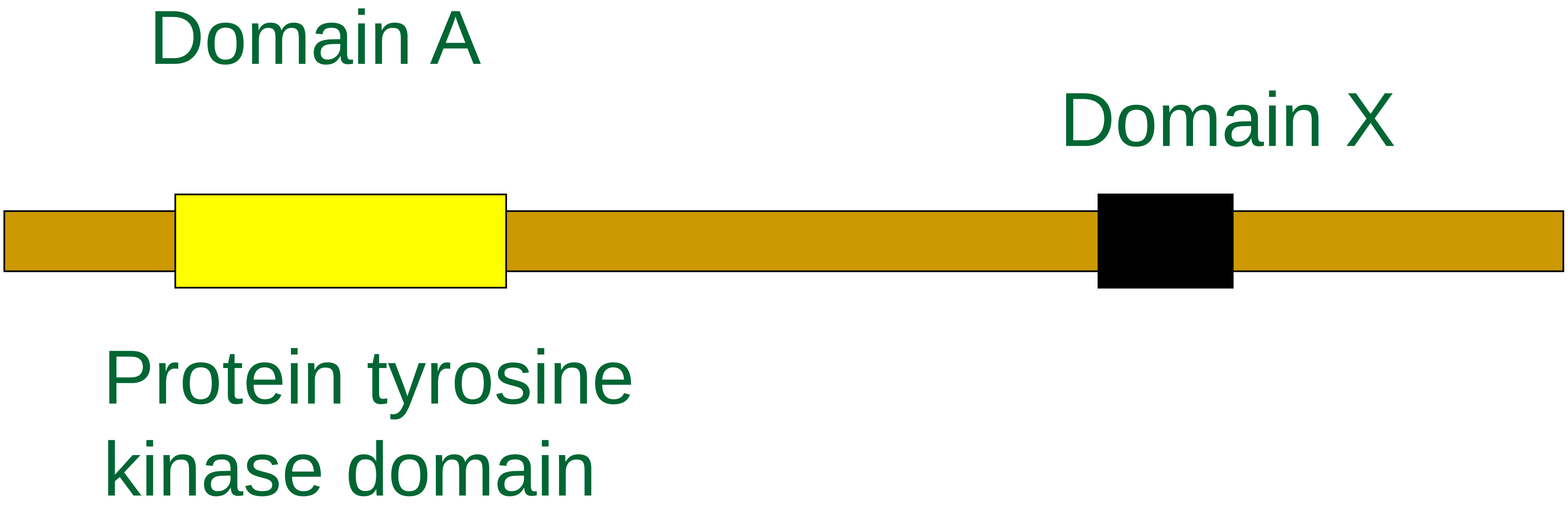
Local Alignment: Proteins are comprised of domains

Human PTK2 :



Protein tyrosine kinase domain

In Leukocytes, a different gene product for tyrosine kinase domain is expressed.



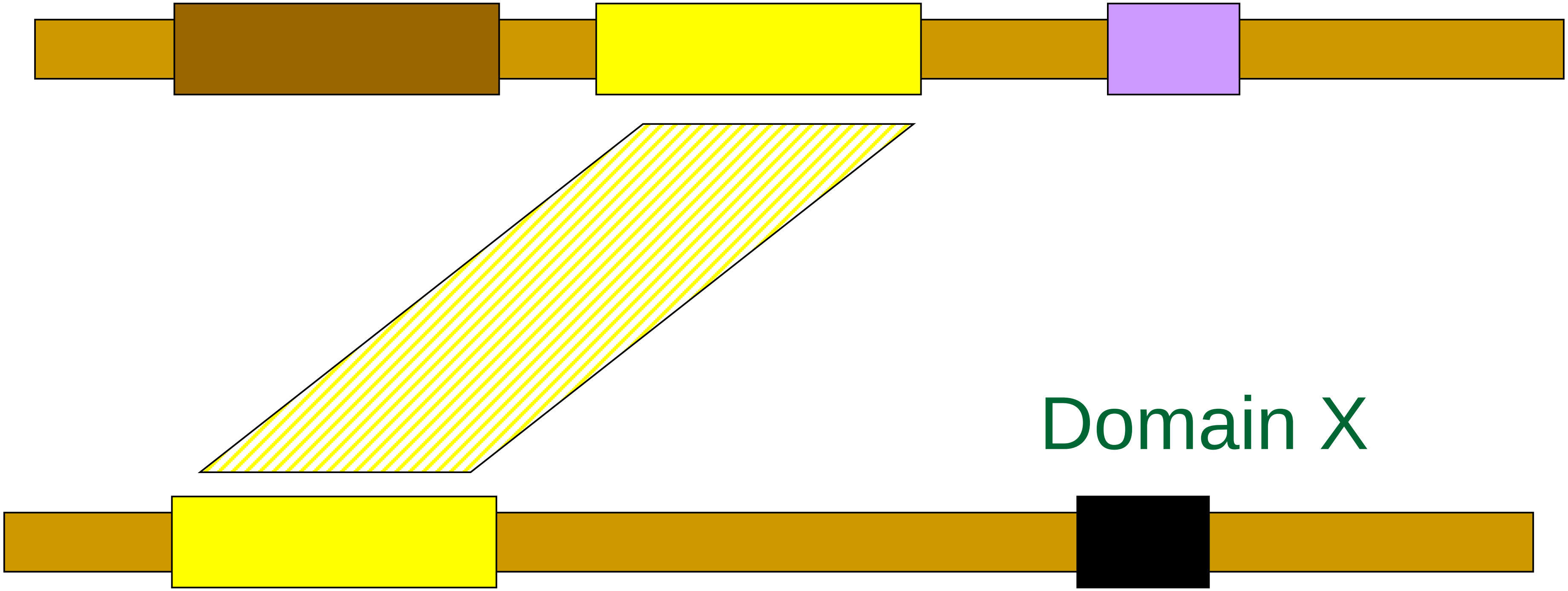
PTK2

The sequence similarity is restricted to a single domain

Domain A

Protein tyrosine kinase domain

Domain B

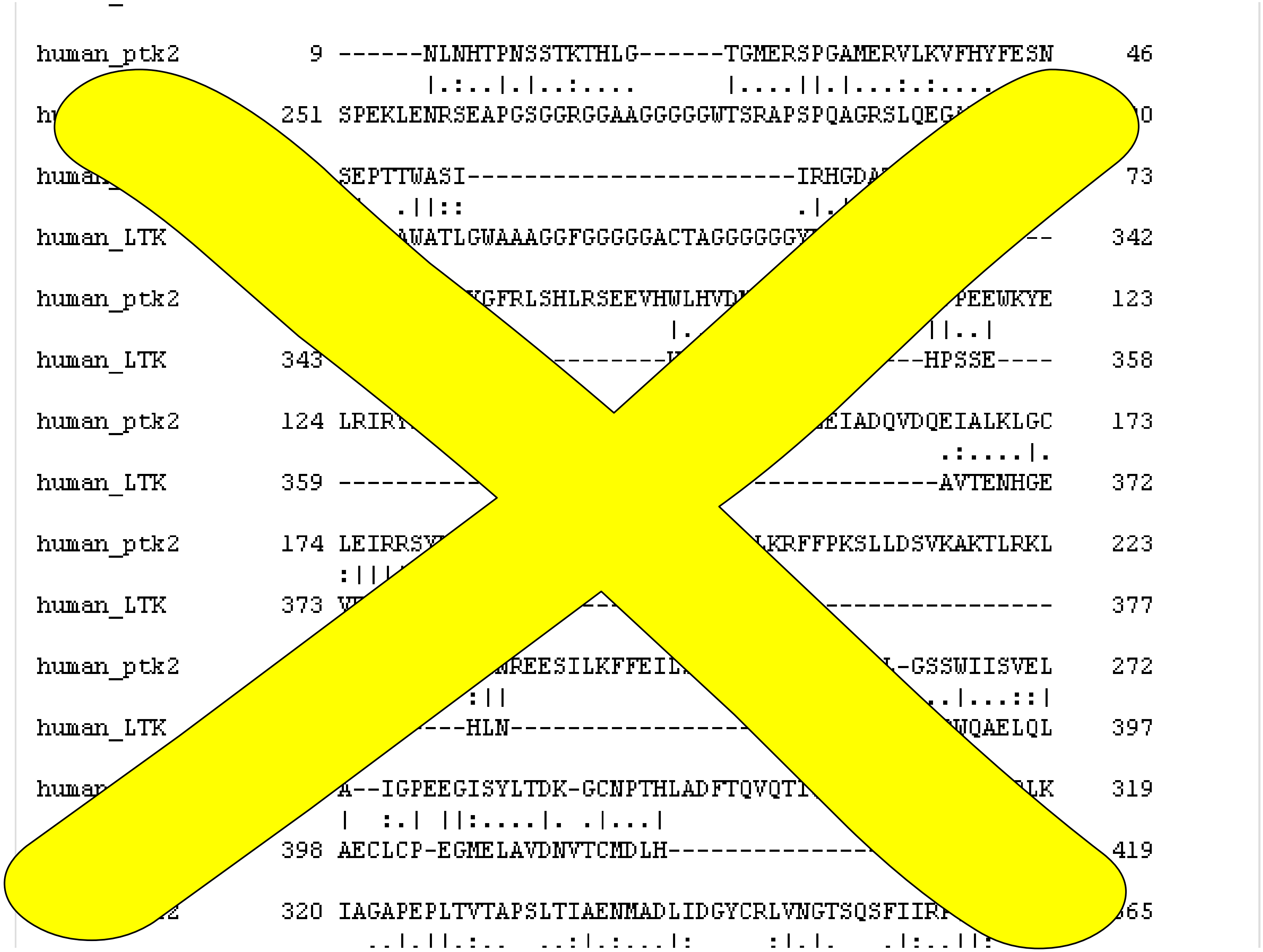


Domain X

Protein tyrosine kinase domain

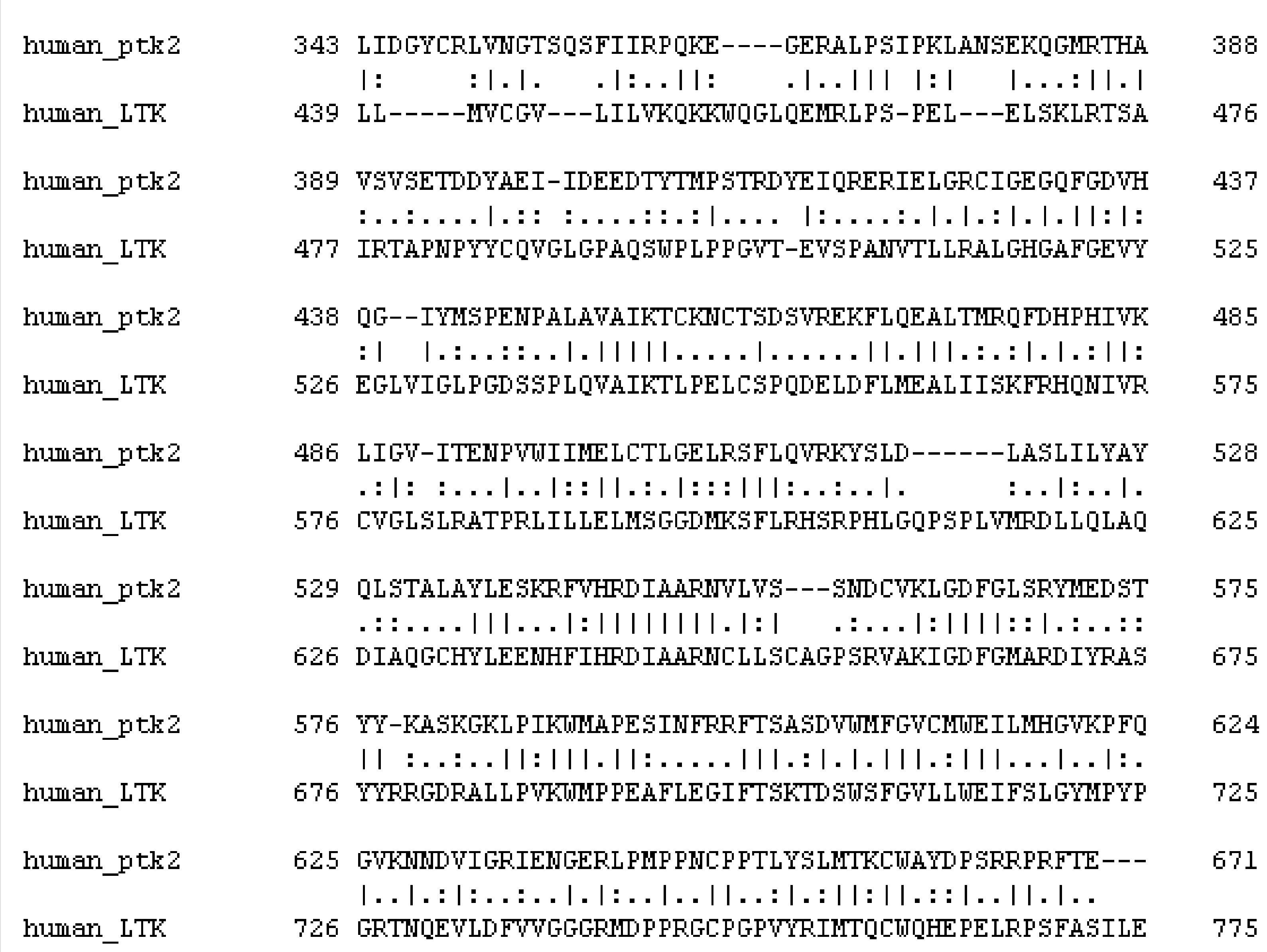
Leukocyte TK

Global alignment of PTK and LTK



```
       -

human_ptk2           9 ------NLNHTPNSSTKTHLG------TGMERSPGAMERVLKVFHYFESN      46
                             |.:..|.|..:....      |....||.|...:.:....
h[illegible]       251 SPEKLENRSEAPGSGGRGGAAGGGGGWTSRAPSPQAGRSLQEG[illegible]   [illegible]0

human[illegible]       [illegible]SEPTTWASI-----------------------IRHGDA[illegible]      73
                       [illegible]  .||::                       .|.[illegible]
human_LTK            [illegible]AWATLGWAAAGGFGGGGACTAGGGGGGY[illegible]--     342

human_ptk2           [illegible]KGFRLSHLRSEEVHWLHVD[illegible]PEEWKYE     123
                                           |.[illegible]||..|
human_LTK          343 [illegible]---------[illegible]---HPSSE----     358

human_ptk2         124 LRIRY[illegible]EIADQVDQEIALKLGC     173
                                                        .:....|.
human_LTK          359 ---------[illegible]-------------AVTENHGE     372

human_ptk2         174 LEIRRSY[illegible]LKRFFPKSLLDSVKAKTLRKL     223
                       :|||[illegible]
human_LTK          373 V[illegible]---[illegible]----------------     377

human_ptk2           [illegible]WREESILKFFEIL[illegible]L-GSSWIISVEL     272
                              :||                      ..|...::|
human_LTK            [illegible]----HLN----------------[illegible]WQAELQL     397

human[illegible]       [illegible]A--IGPEEGISYLTDK-GCNPTHLADFTQVQTI[illegible]LK     319
                       |  :.| ||:....|. .|...|
[illegible]        398 AECLCP-EGMELAVDNVTCMDLH--------------[illegible]     419

[illegible]2       320 IAGAPEPLTVTAPSLTIAENMADLIDGYCRLVNGTSQSFIIR[illegible]   [illegible]65
                         ..|.||.:..  ..:|.:...|:     :|.|.    .|:..||:
```

Local alignment of PTK and LTK



```
human_ptk2       343 LIDGYCRLVNGTSQSFIIRPQKE----GERALPSIPKLANSEKQGMRTHA    388
                     |:     :|.|.   .|:..||:    .|..||| |:|   |...:||.|
human_LTK        439 LL-----MVCGV---LILVKQKKWQGLQEMRLPS-PEL---ELSKLRTSA    476

human_ptk2       389 VSVSETDDYAEI-IDEEDTYTMPSTRDYEIQRERIELGRCIGEGQFGDVH    437
                     :..:....|.:: :....::.:|.... |:....:.|.|.:|.|.||:|:
human_LTK        477 IRTAPNPYYCQVGLGPAQSWPLPPGVT-EVSPANVTLLRALGHGAFGEVY    525

human_ptk2       438 QG--IYMSPENPALAVAIKTCKNCTSDSVREKFLQEALTMRQFDHPHIVK    485
                     :|  |.:..::..|.|||||.....|......||.|||.:.:|.|.:||:
human_LTK        526 EGLVIGLPGDSSPLQVAIKTLPELCSPQDELDFLMEALIISKFRHQNIVR    575

human_ptk2       486 LIGV-ITENPVWIIMELCTLGELRSFLQVRKYSLD------LASLILYAY    528
                     .:|: :...|..|::||.:.|:::|||:..:..|.      :..|:..|.
human_LTK        576 CVGLSLRATPRLILLELMSGGDMKSFLRHSRPHLGQPSPLVMRDLLQLAQ    625

human_ptk2       529 QLSTALAYLESKRFVHRDIAARNVLVS---SNDCVKLGDFGLSRYMEDST    575
                     .::....|||...|:||||||||.|:|   .:...|:||||::|.:..::
human_LTK        626 DIAQGCHYLEENHFIHRDIAARNCLLSCAGPSRVAKIGDFGMARDIYRAS    675

human_ptk2       576 YY-KASKGKLPIKWMAPESINFRRFTSASDVWMFGVCMWEILMHGVKPFQ    624
                     || :..:..||:|||.||:.....|||.:|.|.|||.:|||...|..|:.
human_LTK        676 YYRRGDRALLPVKWMPPEAFLEGIFTSKTDSWSFGVLLWEIFSLGYMPYP    725

human_ptk2       625 GVKNNDVIGRIENGERLPMPPNCPPTLYSLMTKCWAYDPSRRPRFTE---    671
                     |..|.:|:..:..|.|:..|..||..:|.:||:||.::|..||.|..
human_LTK        726 GRTNQEVLDFVVGGGRMDPPRGCPGPVYRIMTQCWQHEPELRPSFASILE    775
```

Conclusions

Use global alignment when the two sequences share the same overall sequence arrangement.

Use local alignment to detect **regions** of similarity.

How alignments are computed

Pairwise alignment

```
AAGCTGAATTCGAA
AGGCTCATTTCTGA
```

One possible alignment:

```
AAGCTGAATT-C-GAA
AGGCT-CATTTCTGA-
```

```
AAGCTGAATT-C-GAA
AGGCT-CATTTCTGA-
```

This alignment includes:

2 mismatches

4 indels (gap)

10 perfect matches

Choosing an alignment for a pair of sequences

Many different alignments are possible for 2 sequences:

```
AAGCTGAATTCGAA
AGGCTCATTTCTGA
```

```
A-AGCTGAATTC--GAA
AG-GCTCA-TTTCTGA-
```

```
AAGCTGAATT-C-GAA
AGGCT-CATTTCTGA-
```

Which alignment is better?

Scoring system (naïve)

Perfect match: +1

Indel (gap): -1

Mismatch: -2

```
AAGCTGAATT-C-GAA
AGGCT-CATTTCTGA-
```

```
A-AGCTGAATTC--GAA
AG-GCTCA-TTTCTGA-
```

Score: = (+1)x10 + (-2)x2 + (-1)x4 = 2

Score: = (+1)x9 + (-2)x2 + (-1)x6 = -1

Higher score → Better alignment

Alignment scoring - scoring of sequence similarity:

Assumes independence between positions:
each position is considered separately

Scores each position:

- Positive if identical (match)
- Negative if different (mismatch or gap)

Total score = sum of position scores
Can be positive or negative

Scoring systems

Scoring system

- In the example above, the choice of +1 for match,-2 for mismatch, and -1 for gap is quite arbitrary
- Different scoring systems → different alignments
- We want a good scoring system…

Scoring matrix

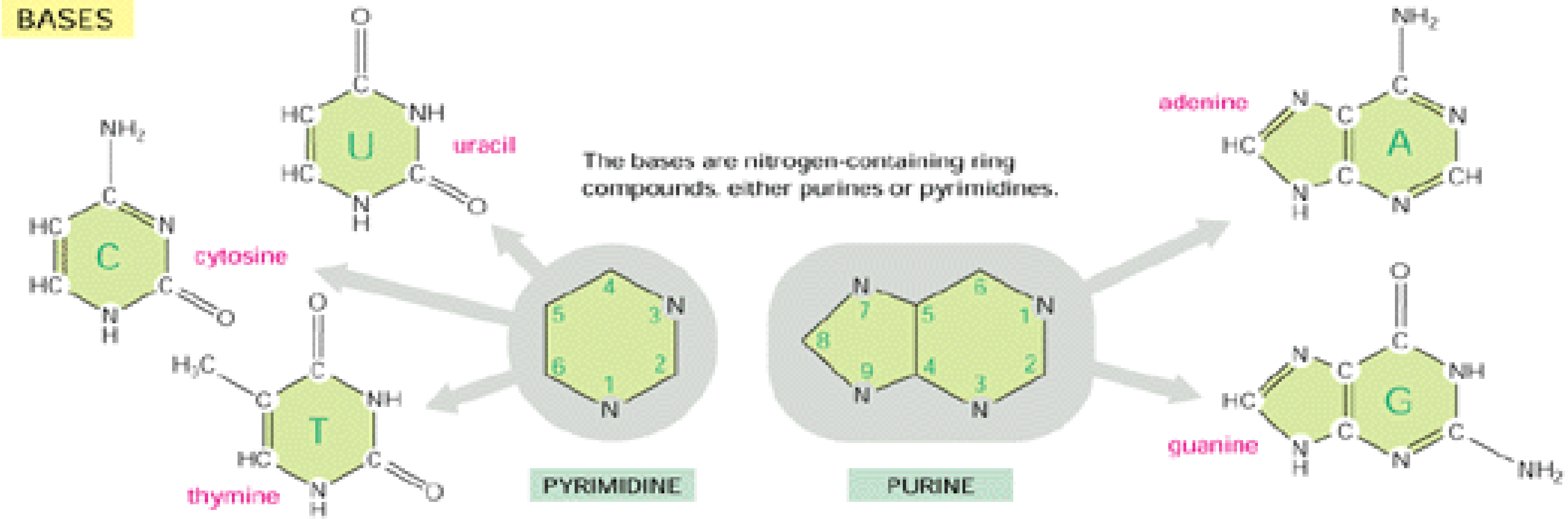
- Representing the scoring system as a table or matrix nxn (n is the number of letters the alphabet contains. n=4 for nucleotides, n=20 for amino acids)
- symmetric

	A	G	C	T
A	2			
G	-6	2		
C	-6	-6	2	
T	-6	-6	-6	2

DNA scoring matrices

Can take into account biological phenomena such as:

- Transition-transversion


BASES
uracil
cytosine
thymine
adenine
guanine
U
C
T
A
G
The bases are nitrogen-containing ring compounds, either purines or pyrimidines.
PYRIMIDINE
PURINE

Thank you for your attention!

Questions?