

1-Month Practical Master Course Genome Analysis (Integrative Bioinformatics & Genomics)

Jaap Heringa

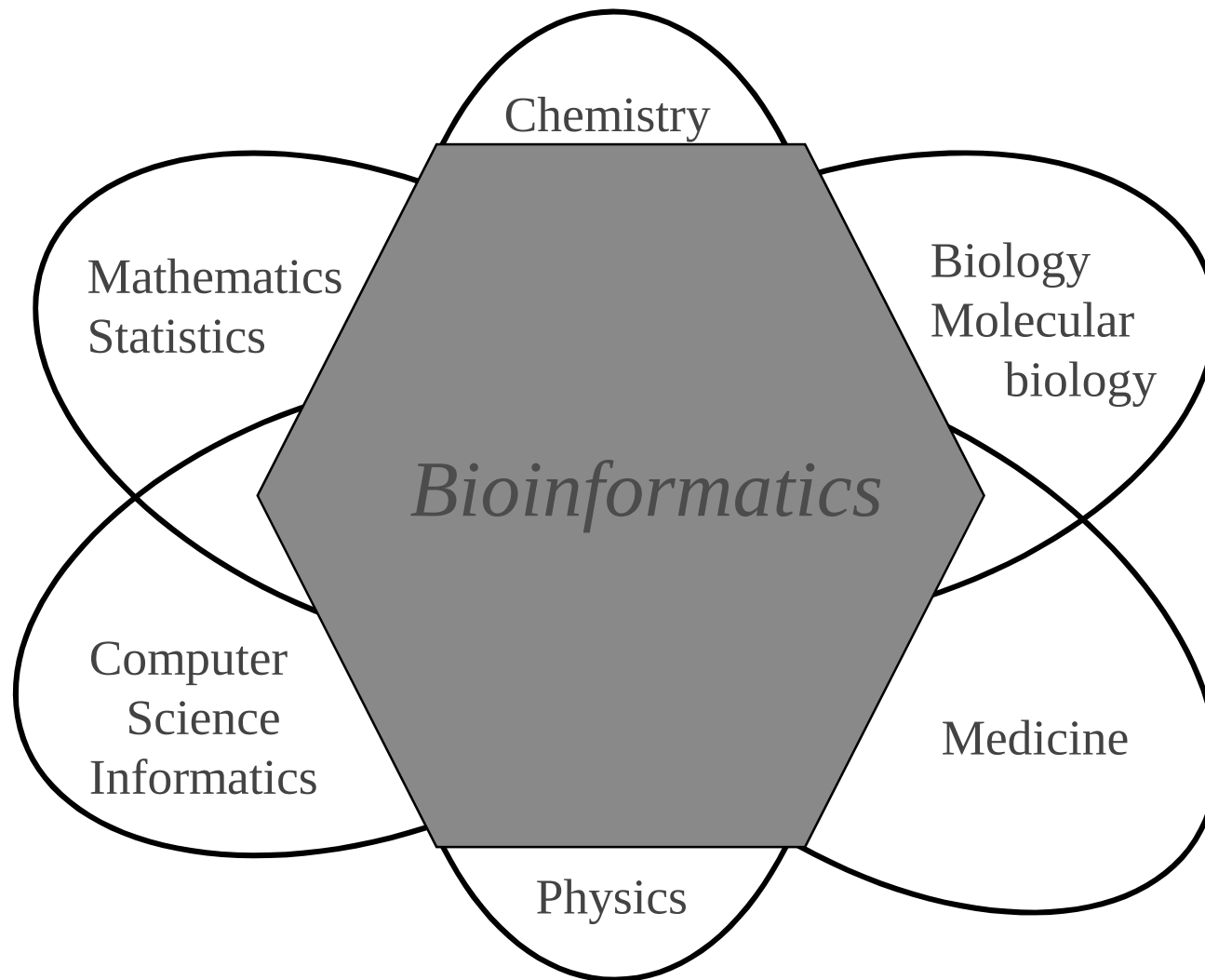
Centre for Integrative Bioinformatics VU (IBIVU)

Vrije Universiteit Amsterdam

The Netherlands

www.ibivu.cs.vu.nl

heringa@cs.vu.nl



Biological Sequence Analysis

Pair-wise sequence alignment

Residue exchange matrices

Multiple sequence alignment

Phylogeny



DNA sequence



```
.....acctc ctgtgcaaga acatgaaaca nctgtggttc  
tcccagatgg gtcctgtccc aggtgcacct gcaggagtgc  
ggcccaggac tggggaagcc tccagagctc aaaaccccac  
ttggtgacac aactcacaca tgcccacggc gccagagcc  
caaatcttgt gacacacctc ccccggtgcc acggtgcca  
gagcccaaat cttgtgacac acctcccca tgcccacggc  
gccagagcc caaatcttgt gacacacctc ccccggtgcc  
ccggtgcca gcacctgaac tcttgggagg accgtcagtc  
ttcctcttcc ccccaaaacc caaggatacc cttatgattt  
cccggaaccc tgaggtcacg tgcgtggttg tggacgtgag  
ccacgaagac ccnnngtcc agttcaagtg gtacgtggac  
ggcgtggagg tgcataatgc caagacaaag ctgcgggagg  
agcagtacaa cagcacgttc cgtgtggtca gcgtcctcac  
cgtcctgcac caggactggc tgaacggcaa ggagtacaag  
tgcaagggtc ccaacaaagc aaccaagtca gcctgacctg  
cctgggtcaa ggcttctacc ccagcgacat cgccgtggag  
tgggagagca atgggcagcc ggagaacaac tacaacacca  
cgctcccat gctggactcc gacggctcct tcttctcta  
cagcaagctc accgtggaca agagcagggt gcagcagggg  
aacatcttct catgtccgt gatgcatgag gctctgcaca  
accgctacac gcagaagagc ctctc.....
```

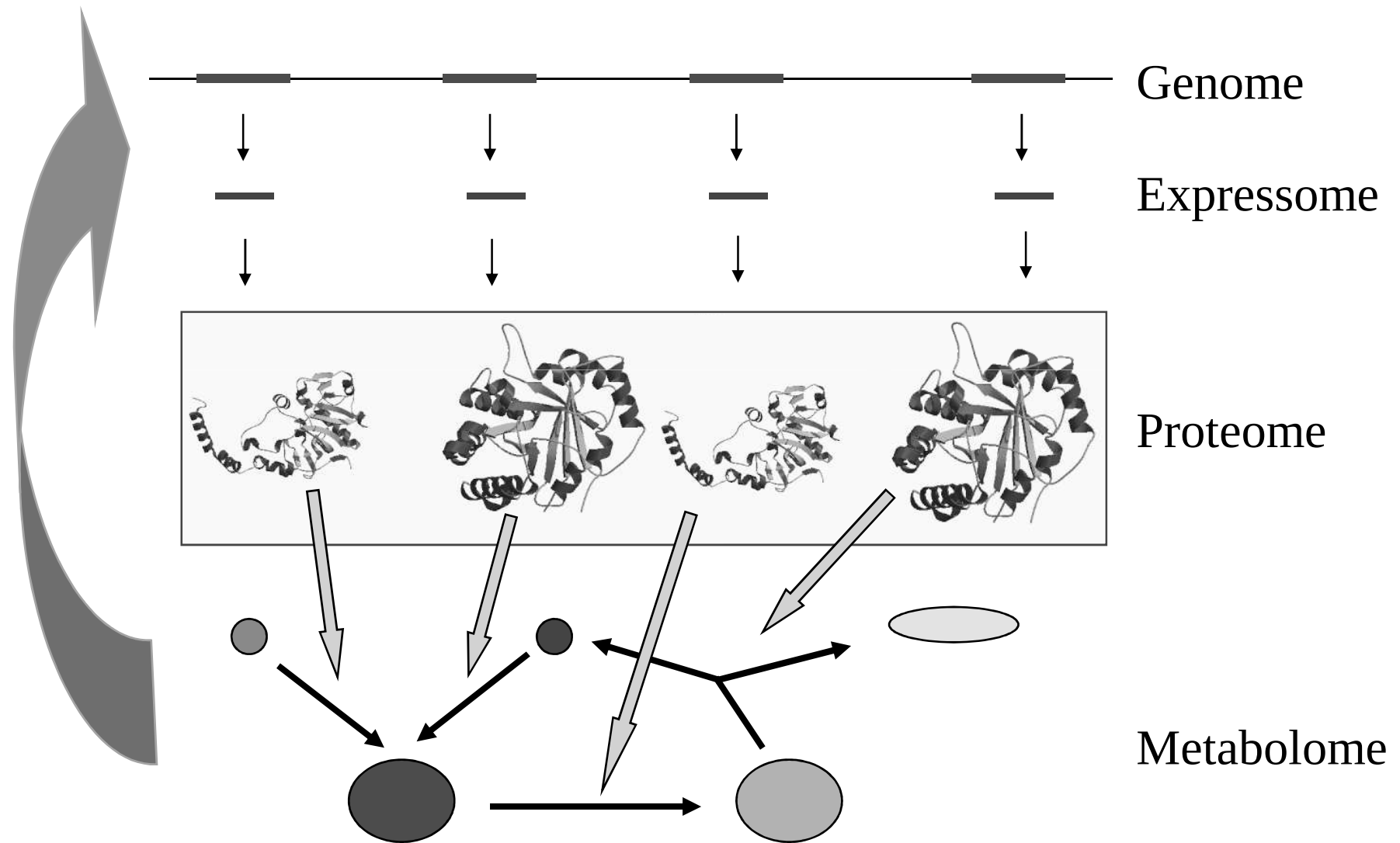
Genome size

<i>Organism</i>	<i>Number of base pairs</i>
ϕ X-174 virus	5,386
Epstein Bar Virus	172,282
<i>Mycoplasma genitalium</i>	580,000
<i>Hemophilus Influenza</i>	1.8×10^6
Yeast (<i>S. Cerevisiae</i>)	12.1×10^6
Human	3.2×10^9
Wheat	16×10^9
<i>Lilium longiflorum</i>	90×10^9
Salamander	100×10^9
<i>Amoeba dubia</i>	670×10^9

Three main principles

- DNA makes RNA makes Protein
- Structure more conserved than sequence
- Sequence → Structure → Function

Functional Genomics



Regulation, signalling cascades, chaperonins, compartmentalisation

How to go from DNA to protein sequence

A piece of double stranded DNA:

5'  attcgttggcaaatcgccccctatccggc 3'
3' taagcaaccgtttagcgggggataggccg 5' 

DNA direction is from 5' to 3'

How to go from DNA to protein sequence

6-frame translation using the codon table (last lecture):

5' attcgttggcaaatcgcccctatccggc 3'

--	--	--	--	--	--	--	--	--

--	--	--	--	--	--	--	--	--

3' taagcaaccgtttagcgggggataggccg 5'

--	--	--	--	--	--	--	--	--

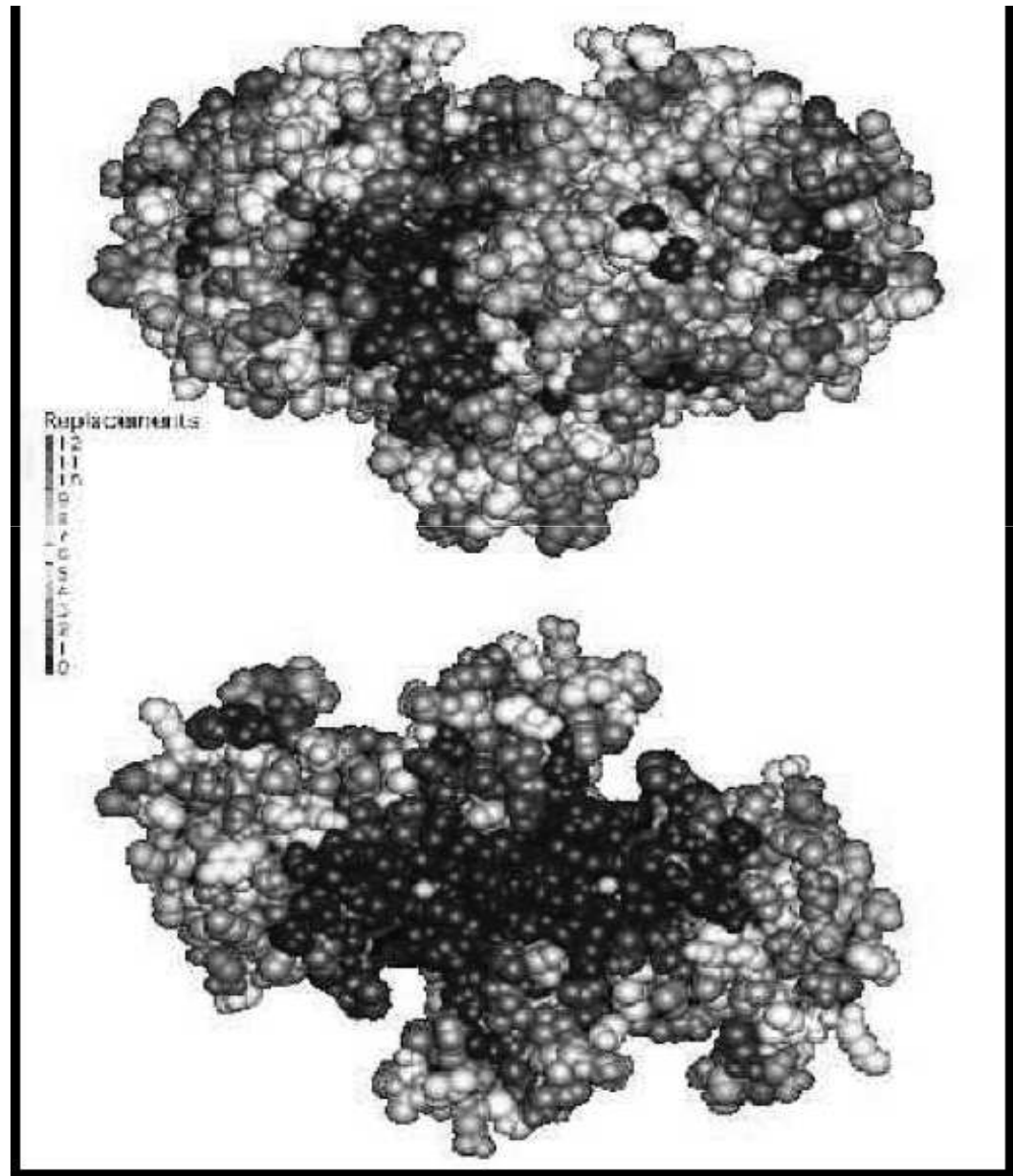
--	--	--	--	--	--	--	--	--

Evolution and three-dimensional protein structure information

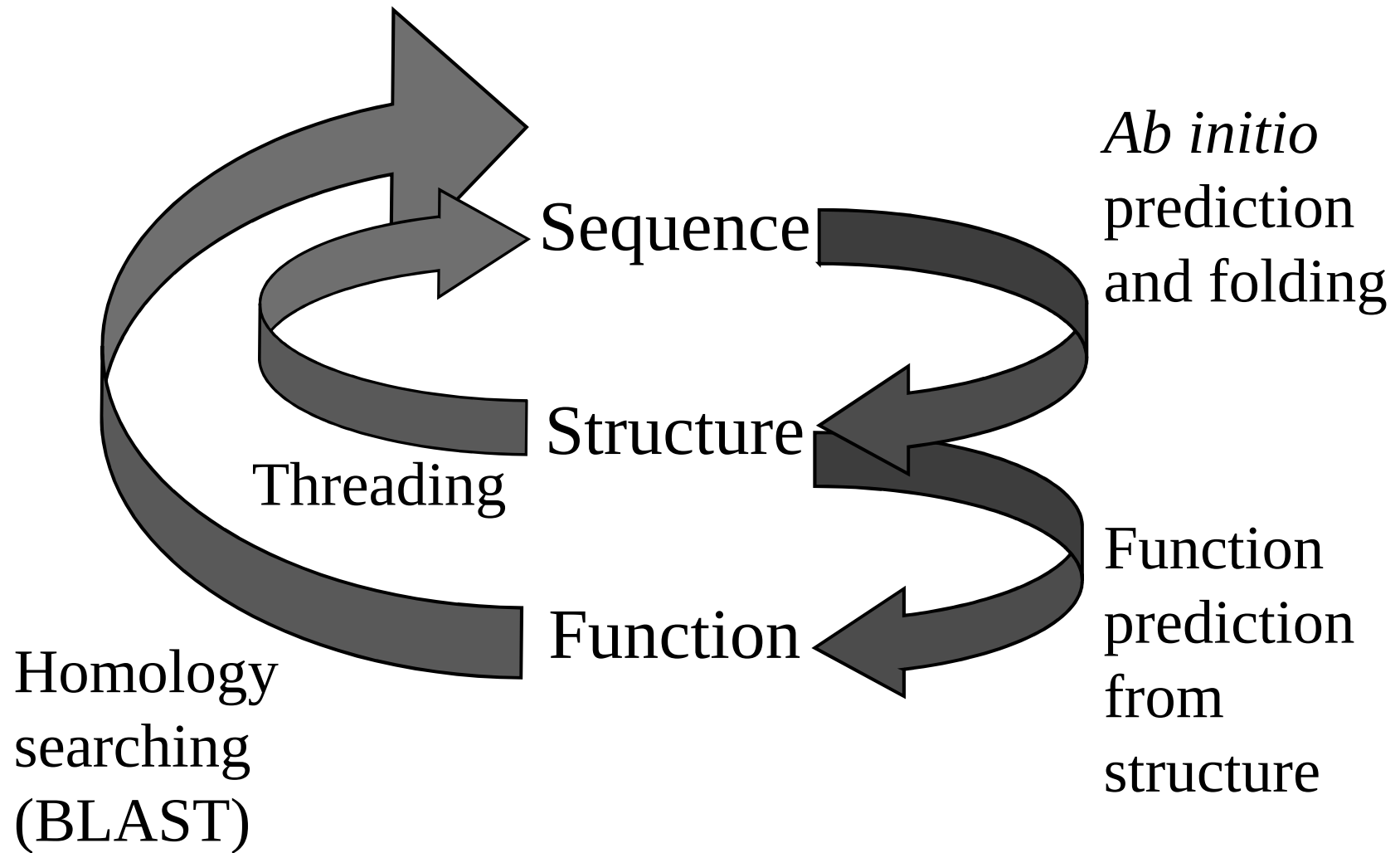
Isocitrate dehydrogenase:

The distance from the active site (in yellow) determines the rate of evolution (red = fast evolution, blue = slow evolution)

Dean, A. M. and G. B. Golding: *Pacific Symposium on Bioinformatics 2000*

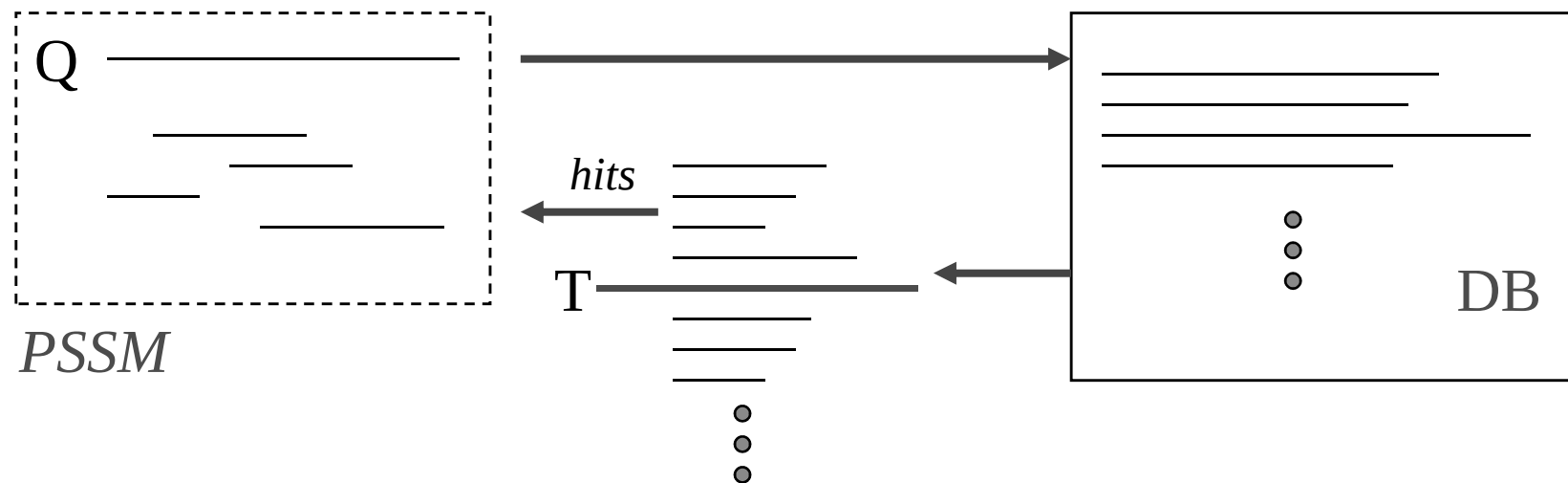


Protein Sequence-Structure-Function

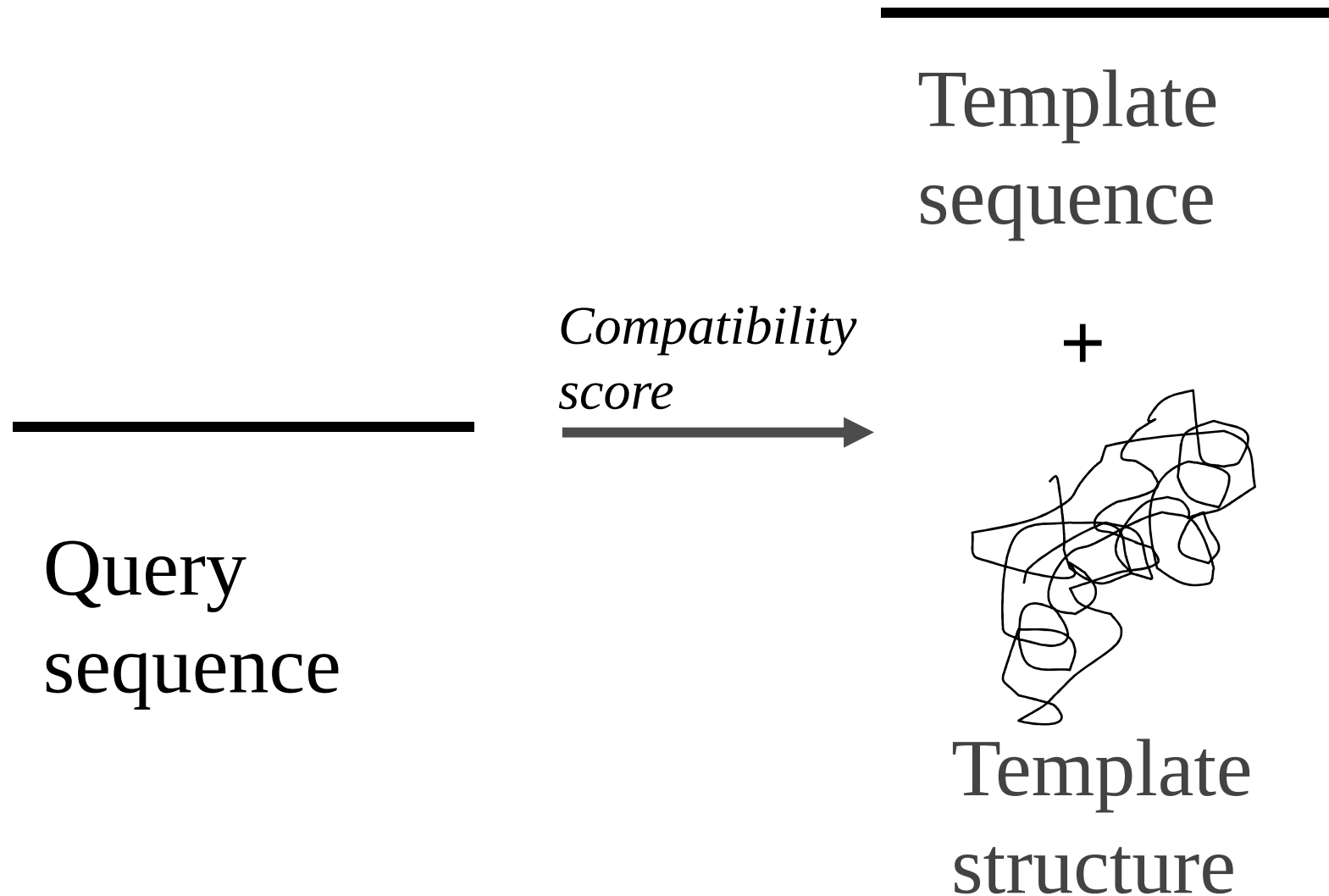


Widely used tool for homology detection: PSI-BLAST

- Heuristic tool to cut down computations required for database searching (~1M sequences in DB)
- Sensitivity gained by iteratively finding hits (local alignments) and repeating search



Threading



Threading

~~Template
sequence~~

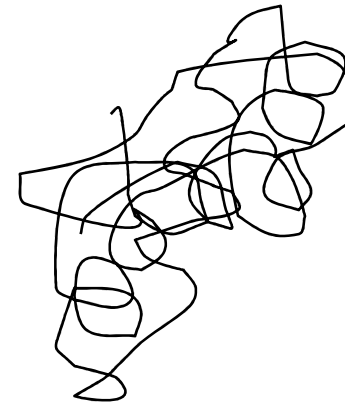
*Compatibility
score*

+

—————

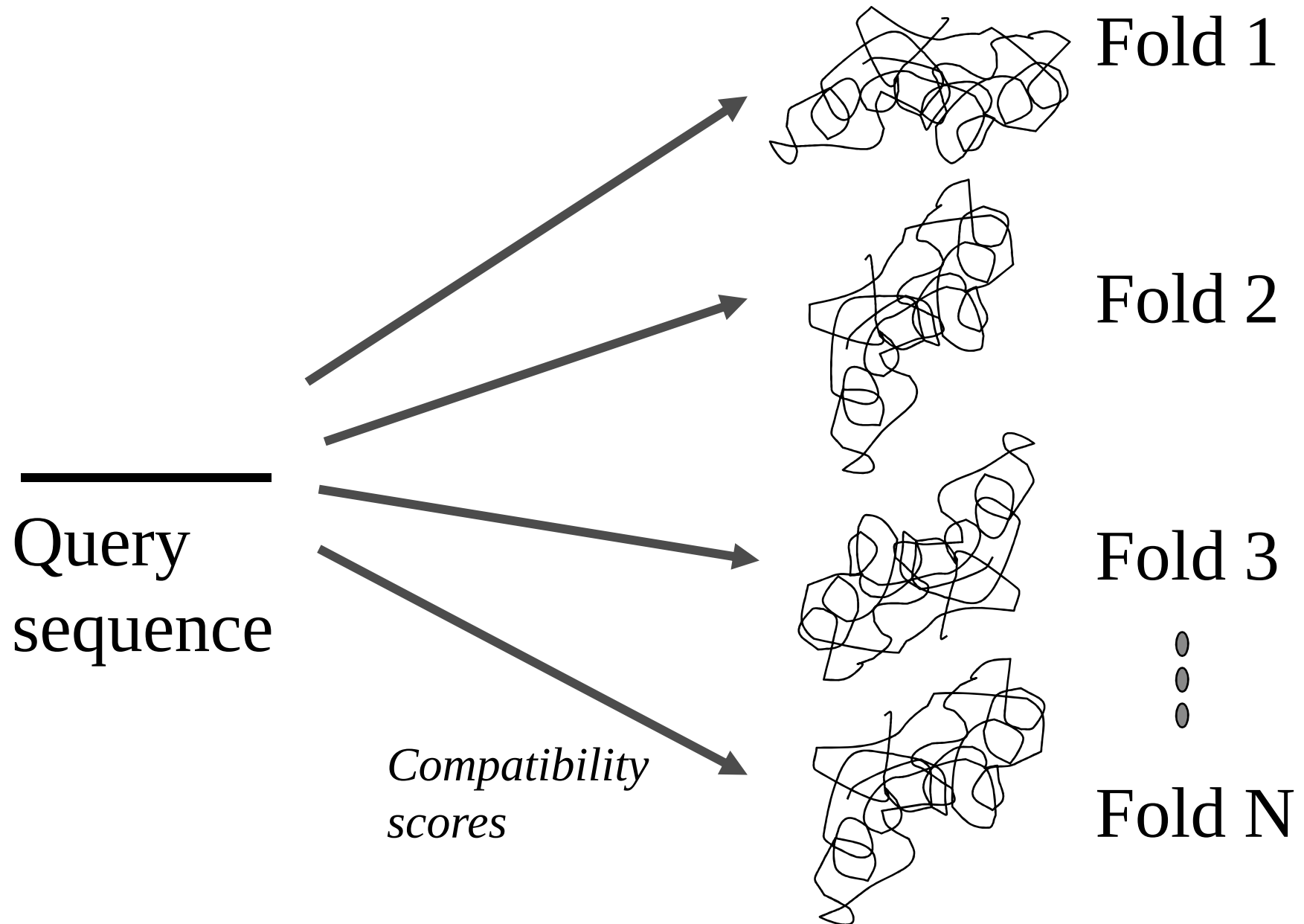


Query
sequence



Template
structure

Fold recognition by threading



Bioinformatics

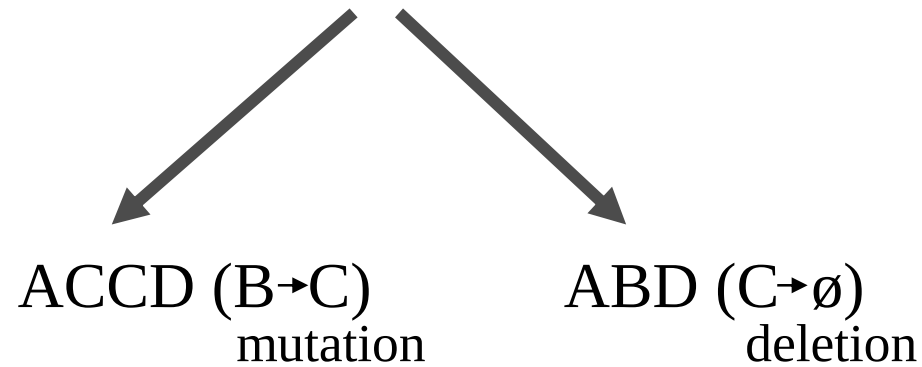
“Nothing in Biology makes sense except in the light of evolution” (Theodosius Dobzhansky (1900-1975))



“Nothing in bioinformatics makes sense except in the light of Biology”

Divergent evolution

Ancestral sequence: ABCD



ACCD
AB—D

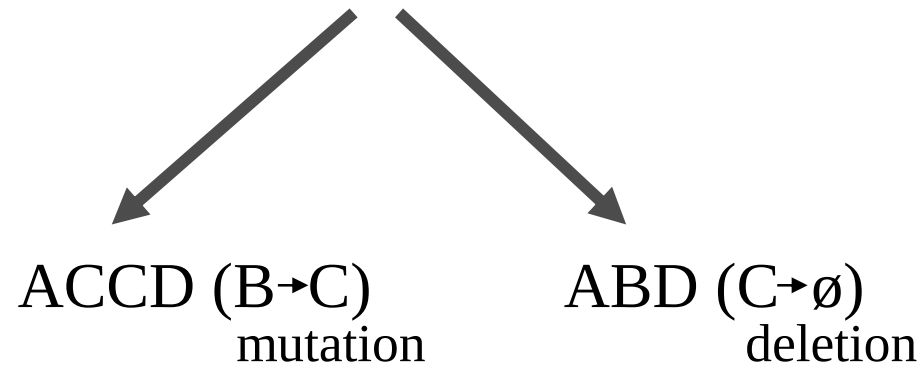
or

ACCD
A—BD

Pairwise Alignment

Divergent evolution

Ancestral sequence: ABCD



ACCD
AB—D

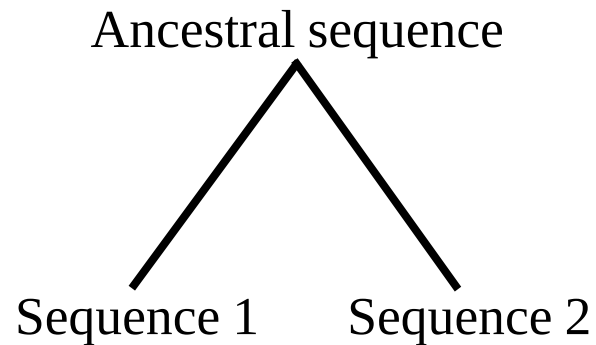
true alignment

or

ACCD
A—BD

Pairwise Alignment

Mutations under divergent evolution

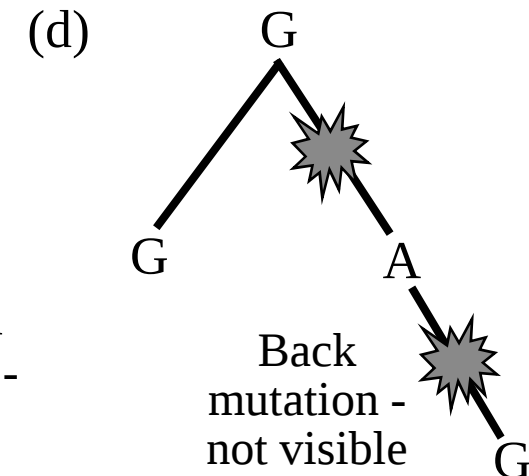
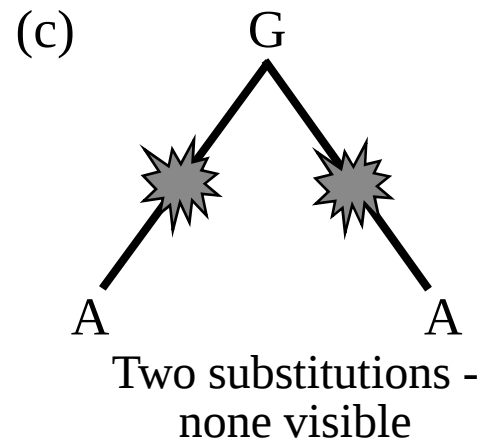
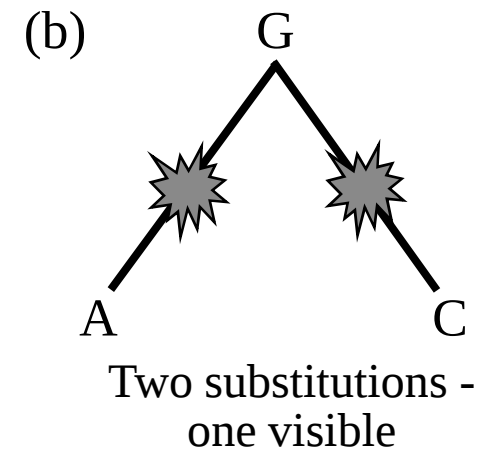
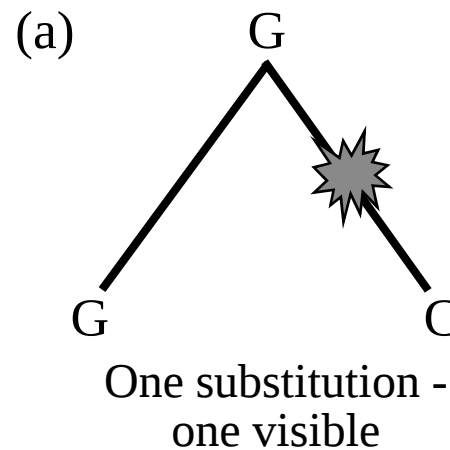


1: ACCTGTAATC

2: ACGTGCGATC

* **

$D = 3/10$ (fraction different sites (nucleotides))



Convergent evolution

- Often with shorter motifs (e.g. active sites)
- Motif (function) has evolved more than once independently, e.g. starting with two very different sequences adopting different folds
- Sequences and associated structures remain different, but (functional) motif can become identical
- Classical example: serine proteinase and chymotrypsin

Serine proteinase (subtilisin) and chymotrypsin

- **Different evolutionary origins, no sequence similarity**
- Similarities in the reaction mechanisms. Chymotrypsin, subtilisin and carboxypeptidase C have a catalytic triad of serine, aspartate and histidine in common: serine acts as a nucleophile, aspartate as an electrophile, and histidine as a base.
- The geometric orientations of the catalytic residues are similar between families, despite **different protein folds**.
- The linear arrangements of the catalytic residues reflect different family relationships. For example the catalytic triad in the chymotrypsin clan (SA) is ordered HDS, but is ordered DHS in the subtilisin clan (SB) and SDH in the carboxypeptidase clan (SC).

A protein sequence alignment

```
MSTGAVLIY - - TSILIKECHAMPAGNE - - - - -  
- - - GGILLFHRTHELIKESHAMANDEGGSNNS  
      *      *      *      * * * *      * * *
```

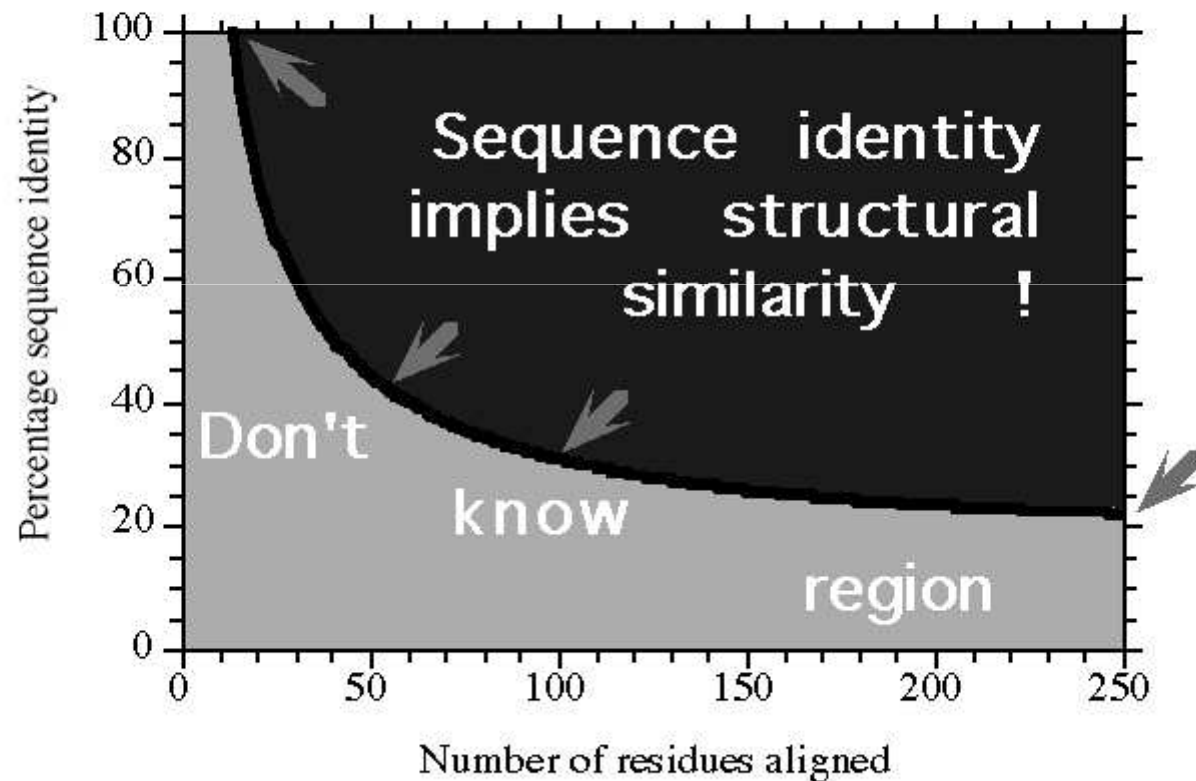
A DNA sequence alignment

```
attcgttggcaaatcgcccctatccgggccttaa  
att - - - tggcggatcg - cctctacggggcc - - - -  
* * *      * * * *      * * * *      * *      * * * * * *
```

What can sequence tell us about structure

(HSSP)

Sander & Schneider, 1991



Richard Rost (Columbia University)

Searching for similarities

What is the function of the new gene?

The “lazy” investigation (i.e., no biological experiments, just bioinformatics techniques):

- Find a set of similar protein sequences to the unknown sequence
- Identify similarities and differences
- For long proteins: identify domains first

Evolutionary and functional relationships

Reconstruct evolutionary relation:

- Based on sequence
 - Identity (simplest method)
 - Similarity
- Homology (common ancestry: the ultimate goal)
- Other (e.g., 3D structure)

Functional relation:

Sequence → Structure → Function

Searching for similarities

Common ancestry is more interesting:

Makes it more likely that genes share the same function

Homology: sharing a common ancestor

- a binary property (yes/no)
- it is a very useful property:

When (an unknown) gene *X* is *homologous* to (a known) gene *G* it means that we gain a lot of information on *X*: what we know about *G* can be transferred to *X* as a good suggestion.

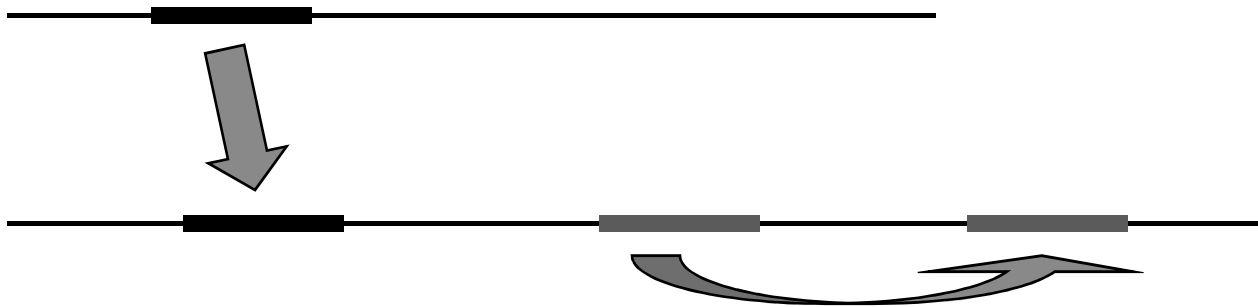
Biological definitions for related sequences

- q **Homologues** are similar sequences in two different organisms that have been derived from a common ancestor sequence. Homologues can be described as either orthologues or paralogues.
- q **Orthologues** are similar sequences in two different organisms that have arisen due to a speciation event. Orthologs typically retain identical or similar functionality throughout evolution.
- q **Paralogues** are similar sequences within a single organism that have arisen due to a gene duplication event.
- q **Xenologues** are similar sequences that do not share the same evolutionary origin, but rather have arisen out of horizontal transfer events through symbiosis, viruses, etc.

How to evolve

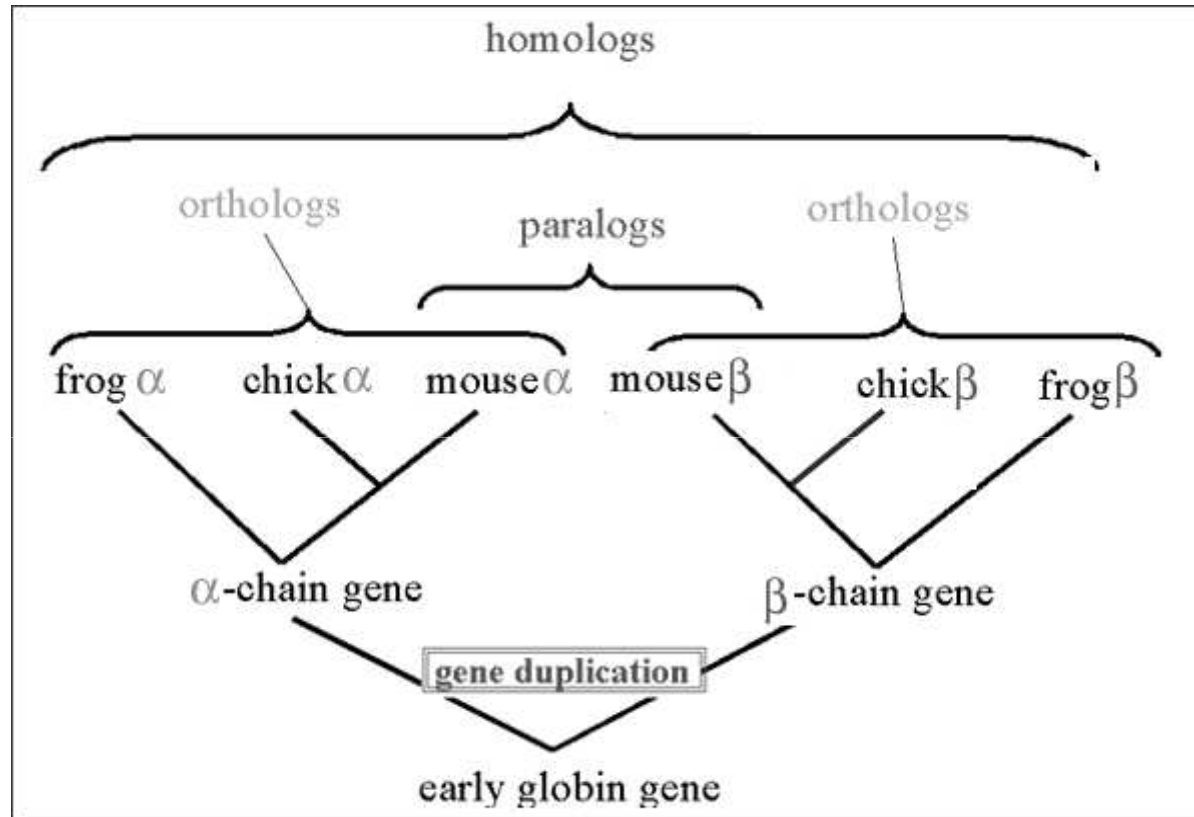
Important distinction:

- *Orthologues*: homologous proteins in different species (all deriving from same ancestor)
- *Paralogues*: homologous proteins in same species (internal gene duplication)



- In practice: to recognise orthology, *bi-directional best hit* is used in conjunction with database search program (this is called an operational definition)

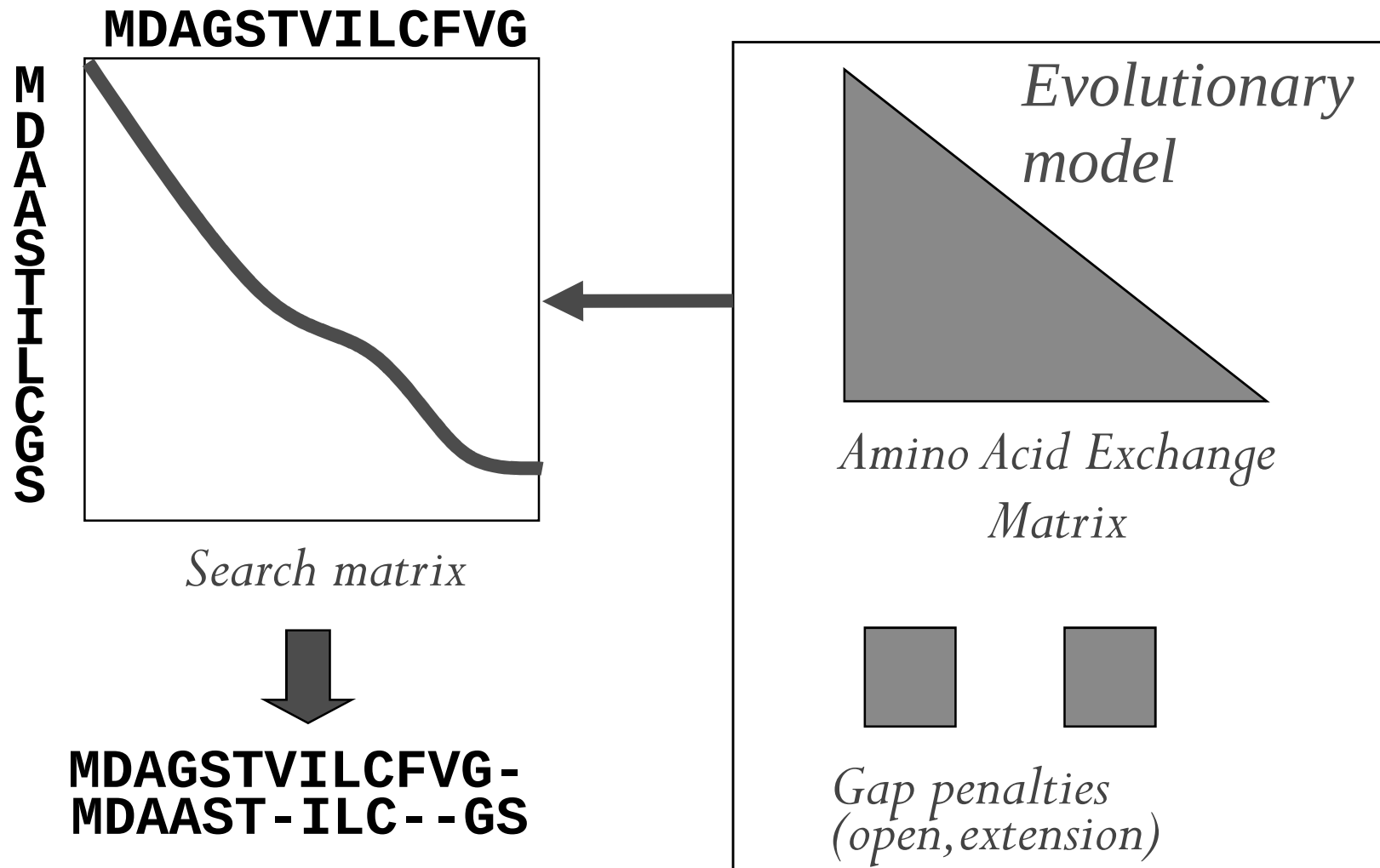
So this means ...



Source: <http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html>

Pairwise sequence alignment needs sense of evolution

Global dynamic programming



How to determine similarity

Frequent evolutionary events at the DNA level:

- 1. Substitution
 - 2. Insertion, deletion
 - 3. Duplication
 - 4. Inversion
- We will restrict ourselves to these events*

A DNA sequence alignment

attcgttggcaaatcgccccctatccgggccttaa
att - - - tggcggatcg - cctctacggggcc - - - -
*** **** **** ** *****

nucleotide one-
letter code

A protein sequence alignment

MSTGAVLIY - - TSILIKECHAMPAGNE - - - - -
- - - GGILLFHRTHELIKESHAMANDEGGSNNS
* * * ***** ***

amino acid one-
letter code