EVOLUTIONSGENETIK AUS DER SICHT EINES NACHRICHTENTECHNIKERS

Joachim Hagenauer

Janis Dingel, Pavol Hanus, Johanna Weindl
Lehrstuhl für Nachrichtentechnik (LNT)
ComInGen Group
TU München (TUM)
Hagenauer@tum.de
J. Müller
Max-Planck Institute Seewiesen- München

Contents

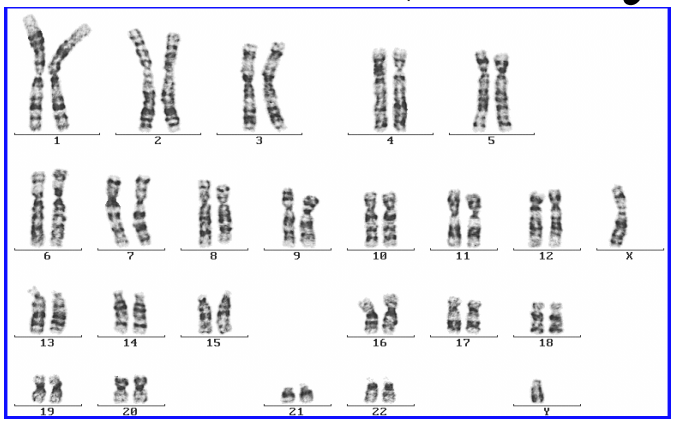
Since four years we have at LNT a group of 2 gene biologists and 4 communications theorists working on gene evolution and gene expression viewed as a communications process.

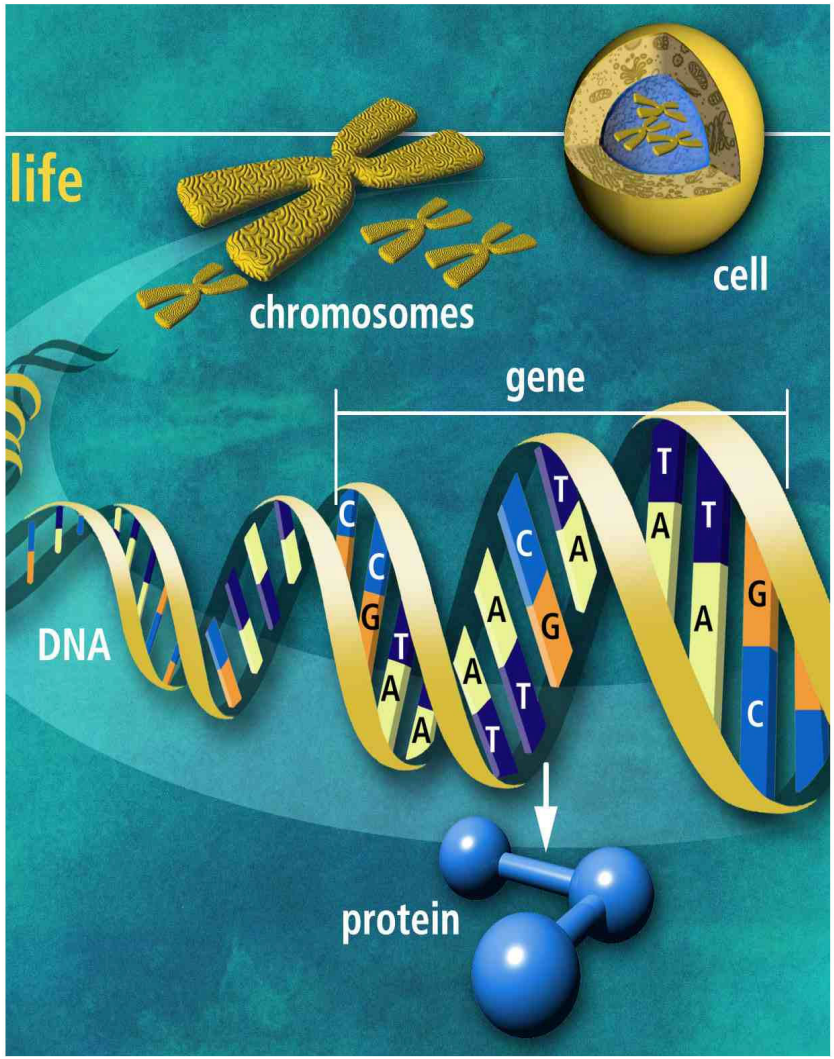
(Will be a DFG Schwerpunktprogram "**InKomBio**")

Some Results:

- Distance measures between different DNA sequences derived from mutual information for classification purposes
- From classification results create mammalian and human phylogenetic trees
- Identification of highly conserved sequences in the genome
- Data Storage in the DNA of Bacteria
- Information transfer between DNA-Variations and diseases for simulated and clinical data (Schizophrenia, Parkinson and Graves autoimmune disease) with measured mutual information
- Synchronization behavior and sync words in DNA to mRNA transcription for E.Coli bacteria

DNA (Deoxyribonucleic Acid)





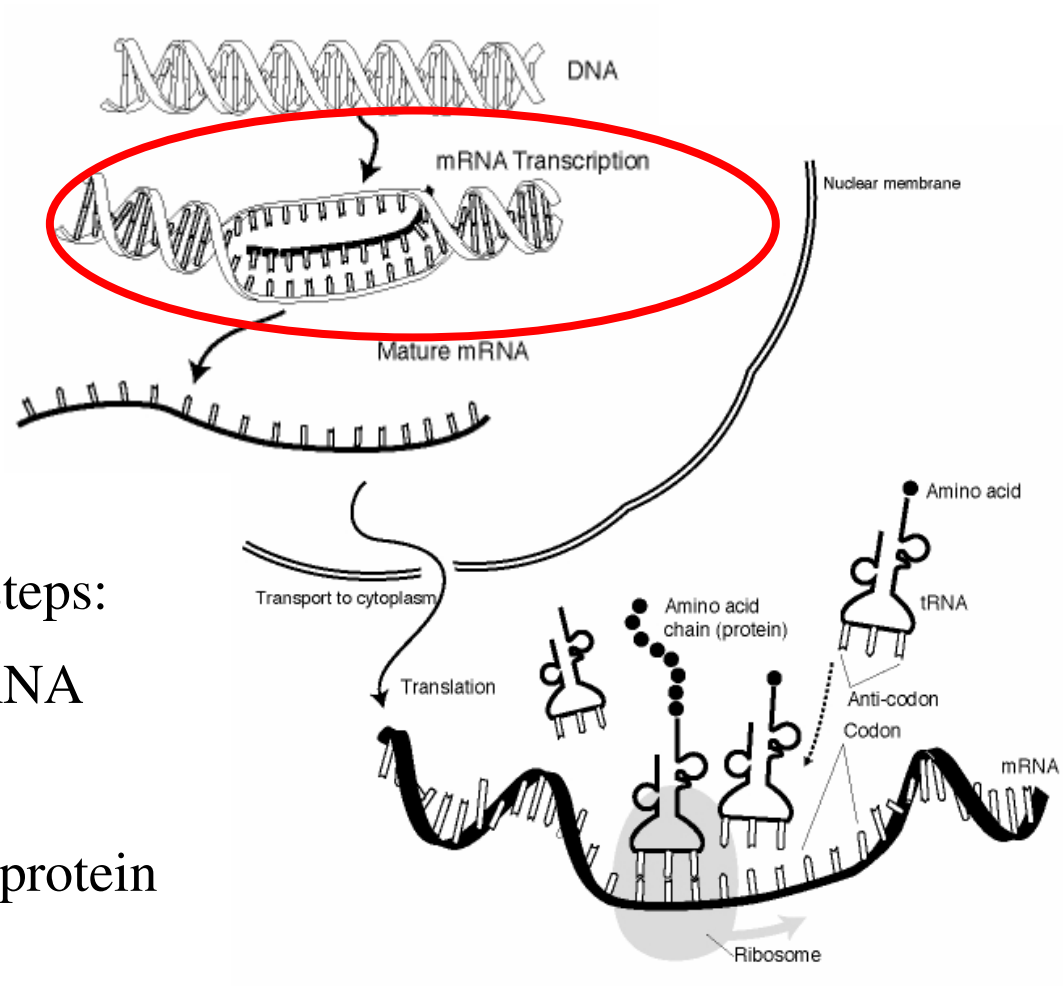
- DNA:
  - Carrier of genetic info
  - 46 chromosomes, 23 from each parent
  - Double stranded helix
  - 3 x 10^9 base pairs (2 meters)
  - 30000 genes
- Genes code proteins

Bases or Nucleotides:

A (Adenine)
T (Thymine)
G (Guanine)
C (Cytosine)

Gene Expression

- Gene Expression is the process in which the information stored in the DNA is transformed into proteins



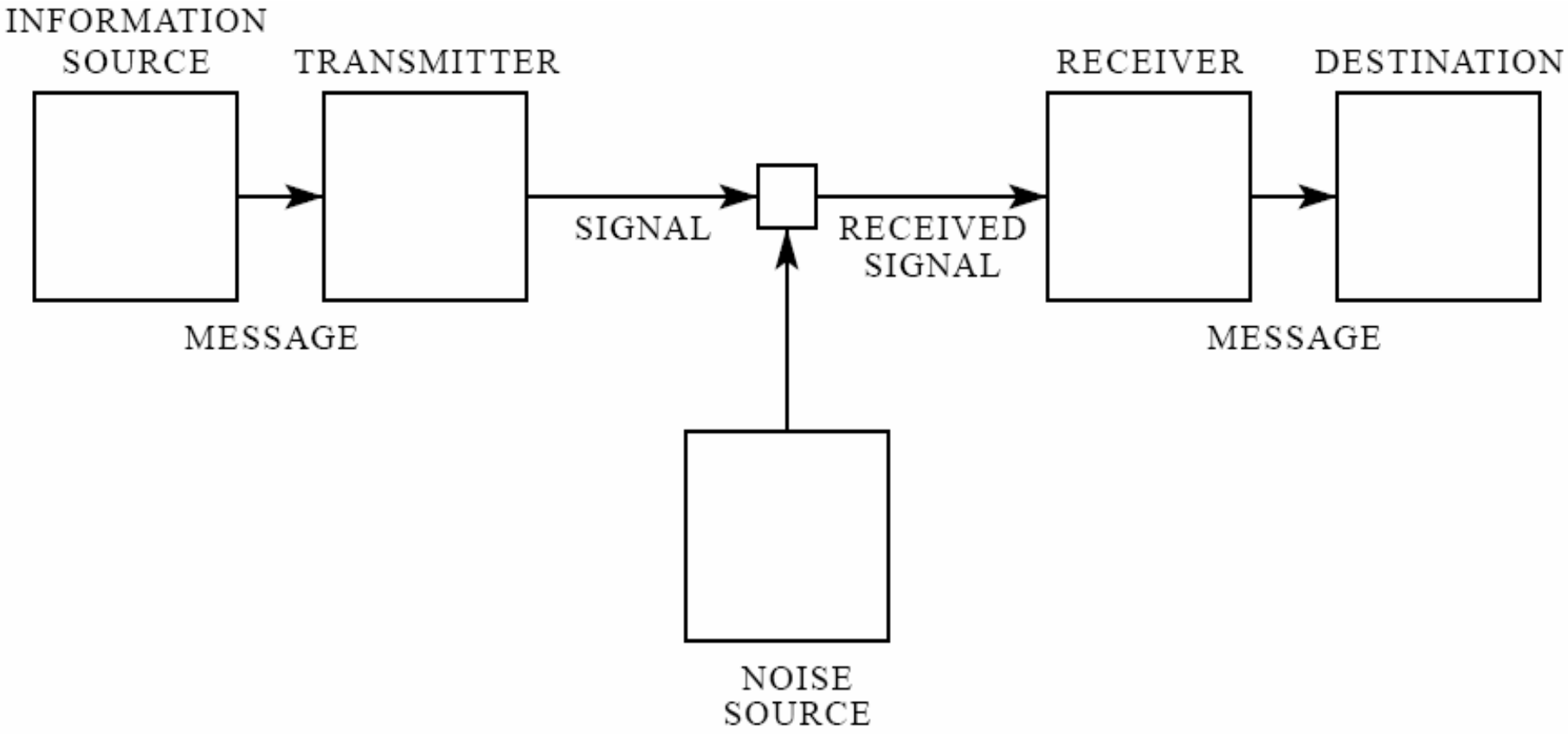
- It takes place in two basic steps:

  Transcription: DNA → mRNA

  Translation:

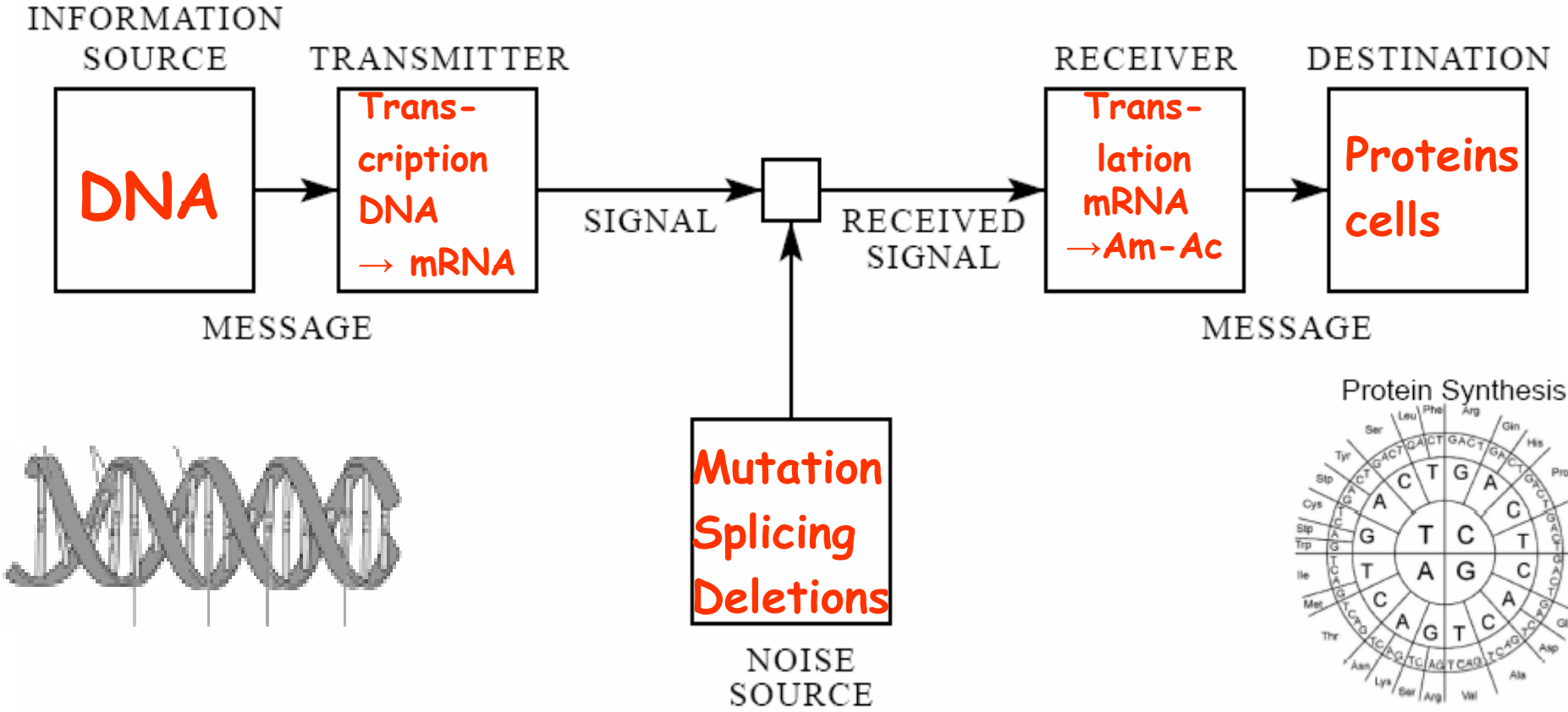
  mRNA → amino acids → protein

Shannon's Model of Information Transmission



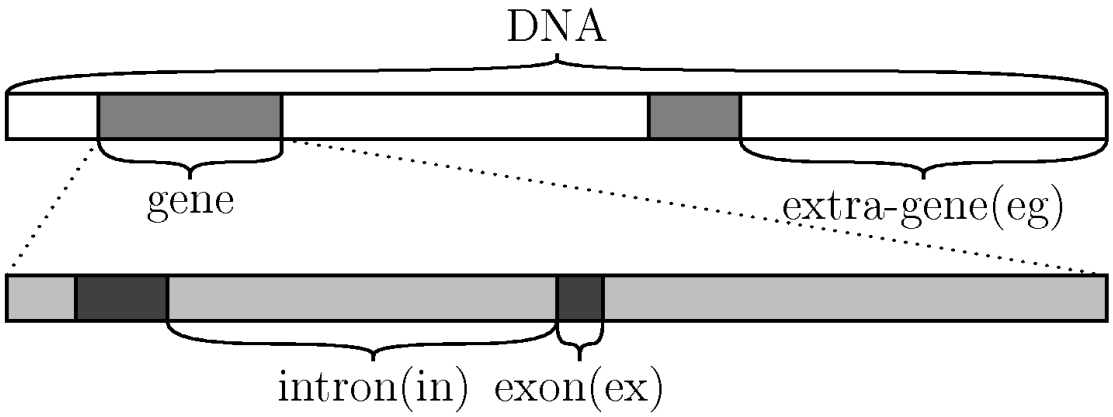
Can we use Shannon's model and methods in genetics?

Shannon's Model of Information Transmission applied to Genetics

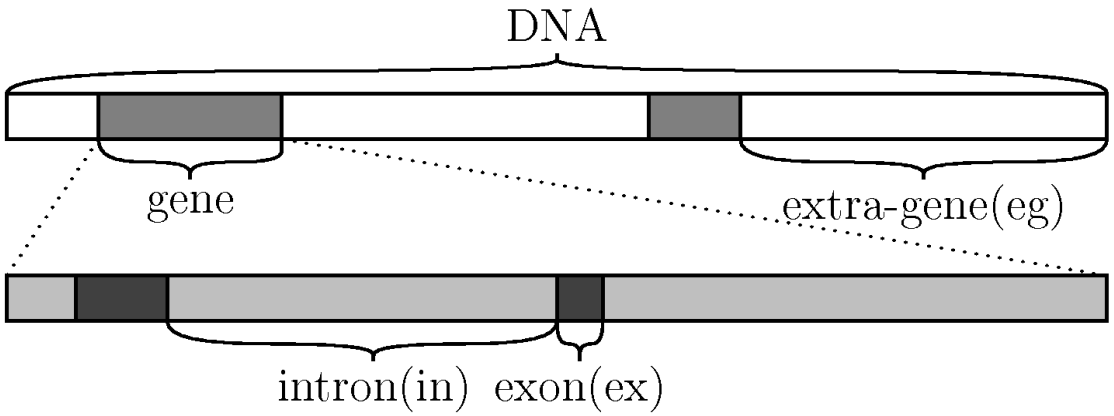


DNA

- **Global Structure**
  - extra-gene(eg) regions separate genes in the sequence (75%)
  - exons(ex) are coding regions translated into proteins (2,5%)
  - introns(in) separate exons in a gene (22,5%)



DNA



Until 2000 only exons were thought to be important.
The rest was considered to be junk of the evolution
("The greatest mistake in DNA research" (Science Magazine))
Now attention in research is focusing on the role of introns
and even extra-gene DNA ("dark matter").
Those regions contain sequences which are highly conserved in evolution.
(parity symbols?, nested codes as suggested by Battail?)

Classification of species using mutual information and compression

- **Traditional phylogenetic clustering methods**
  - **Alignment of related proteins and amino acids of identical sizes**
  - **Based on physiology of species**

- **Our approach**
  - **Distance of DNA sequences based on mutual information**
  - **Applicable on the genome level up to the whole genome**
  - **Sequences are available from public databases, i.e. NCBI**
  - **Simple and universal (Naïve?)**

Mutual information as a distance measure between DNA's

Relate two DNA sequences S_i and S_j by their mutual information $I(S_i; S_j)$

$$I(S_i; S_j) = H(S_i) - H(S_i|S_j)$$

For classification purposes define the distance metric

$$d_{\mathrm{CL}}(S_i, S_j) = 1 - \frac{I(S_i; S_j)}{\max(H(S_i), H(S_j))}$$

Assume $H(S_i) > H(S_j)$ then

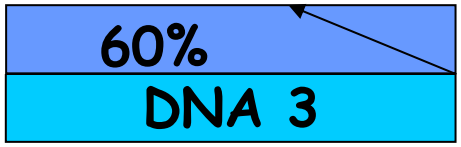
$$d_{\mathrm{CL}}(S_i, S_j) = \frac{H(S_i|S_j)}{H(S_i))}$$

Measure Entropies by their compression ratio $H(S_i) \approx |\mathrm{comp}(S_i)|/|S_i|$: The DNA S_j with the smaller entropy serves as training sequence for the compressor

$$d_{\mathrm{CL}}(S_i, S_j) \approx \frac{|\mathrm{comp}(S_i|S_j)|}{|\mathrm{comp}(S_i)|} = \frac{|\mathrm{comp}(S_j, S_i)| - |\mathrm{comp}(S_j)|}{|\mathrm{comp}(S_i)|}$$

Classification

Distance of DNA Sequences measured by compression



$$\text{Distance to DNA 1} = \frac{40\%}{60\%}$$

$$\text{Distance to DNA 2} = \frac{50\%}{60\%}$$

Classification of species using the distance metric d_{CL} derived from mutual information

- **Mitochondrial DNA datasets from the NCBI data base**

- **Investigated compression algorithm:**

1. **Prediction by Partial Matching PPM**
2. **Lempel Ziv**
3. **Context tree weighting (CTW) algorithm w. and w.o. freezing**
4. **DNACompress**

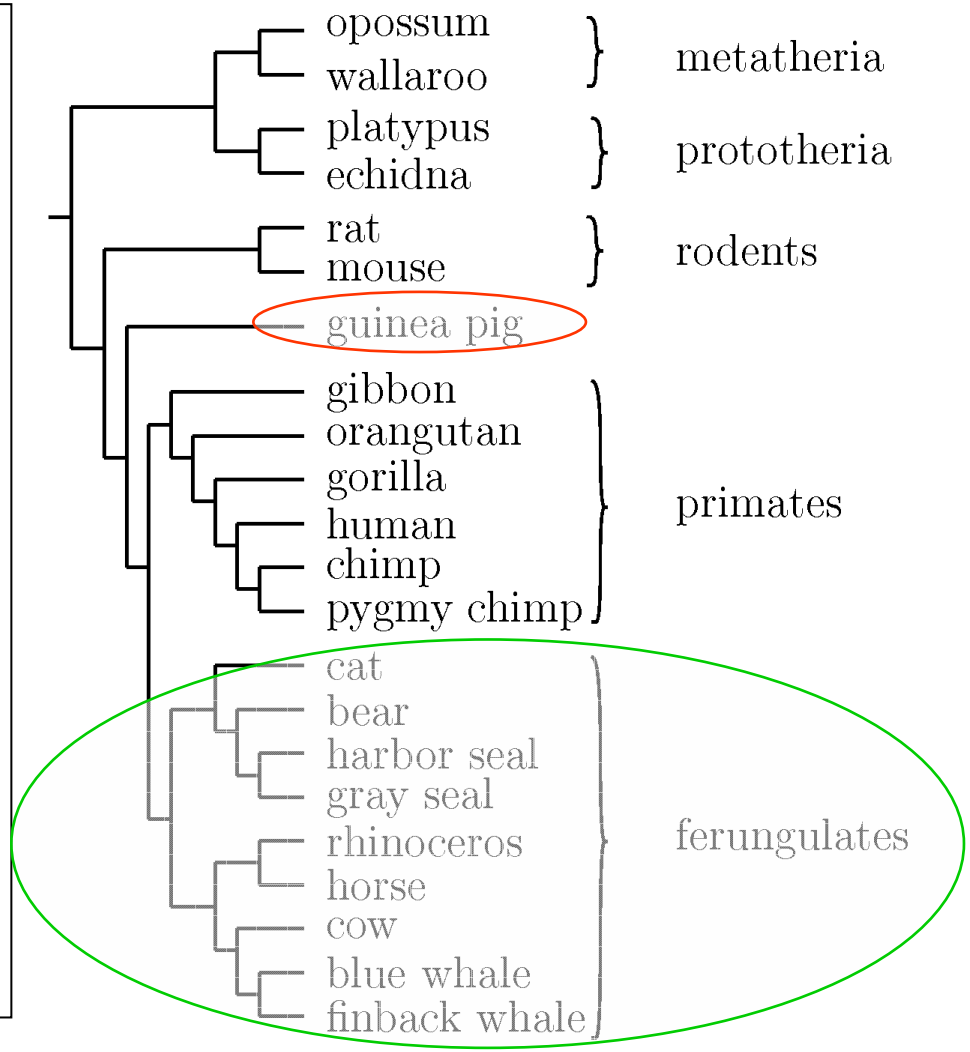
Distance matrix coding for the ferungulates subtree

Matrix generated using d_{CL} and DNACompress.

$d_{CL}(S_i, S_j)$	cow	fin. whale	bl. whale	harb. seal	gray seal	cat	horse	rhino	bear
cow	0	0.787	0.770	0.810	0.824	0.832	0.778	0.770	0.884
fin. whale	0.787	0	0.339	0.853	0.845	0.852	0.809	0.812	**0.885**
bl. whale	0.770	0.339	0	0.840	0.843	0.842	0.804	0.802	0.868
harb. seal	0.810	0.853	0.840	0	**0.220**	0.765	0.773	0.786	0.741
gray seal	0.824	0.845	0.843	**0.220**	0	0.762	0.786	0.789	0.744
cat	0.832	0.852	0.842	0.765	0.762	0	0.789	0.769	0.795
horse	0.778	0.809	0.804	0.773	0.786	0.789	0	0.626	0.855
rhino	0.770	0.812	0.802	0.786	0.789	0.769	0.626	0	0.836
bear	0.884	**0.885**	0.868	0.741	0.744	0.795	0.855	0.836	0

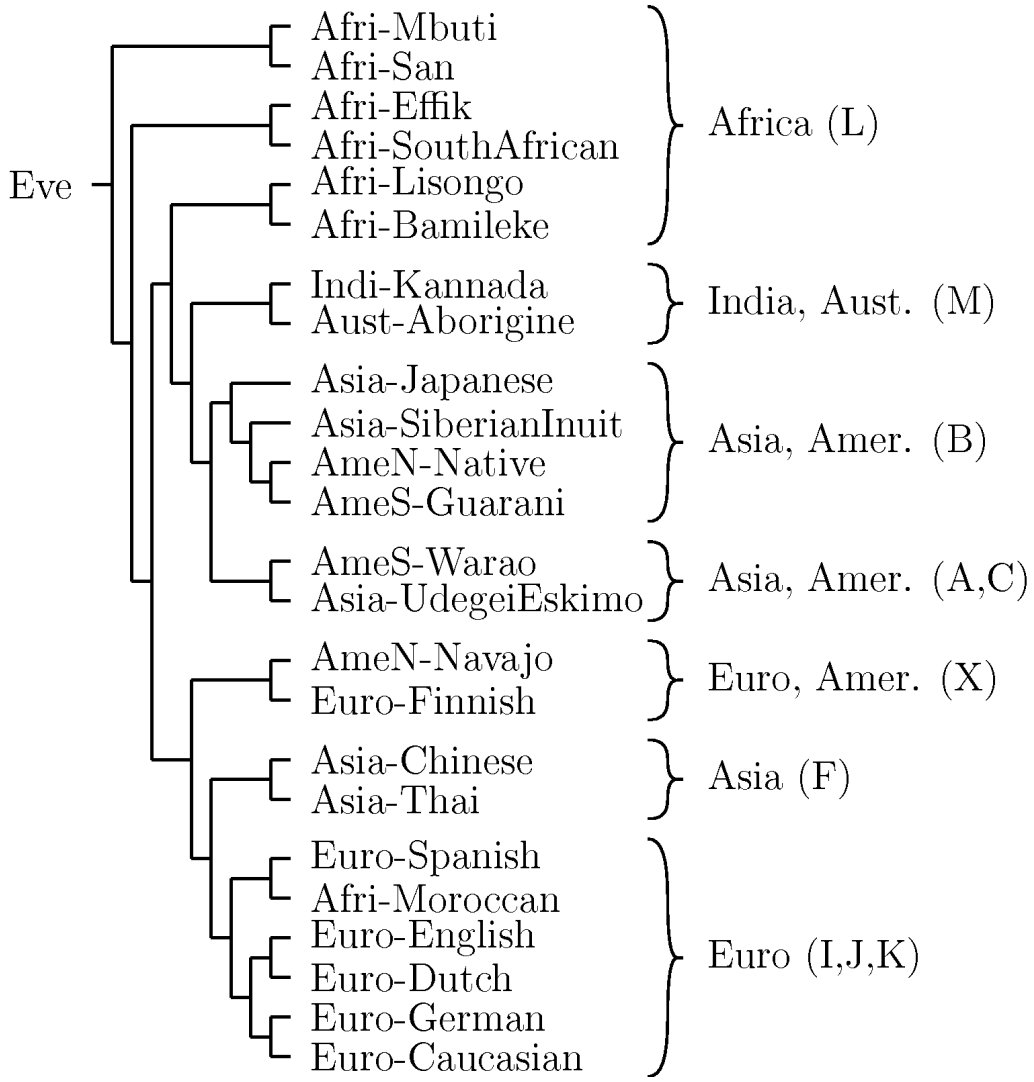
Classification result: Mammalian Phylogeny

- **Mammalian Phylogeny**
  - based on whole mtDNA sequences
  - using metric d_{CL} DNACompress, quartet-method
  - The disputed guinea pig separated early in evolution

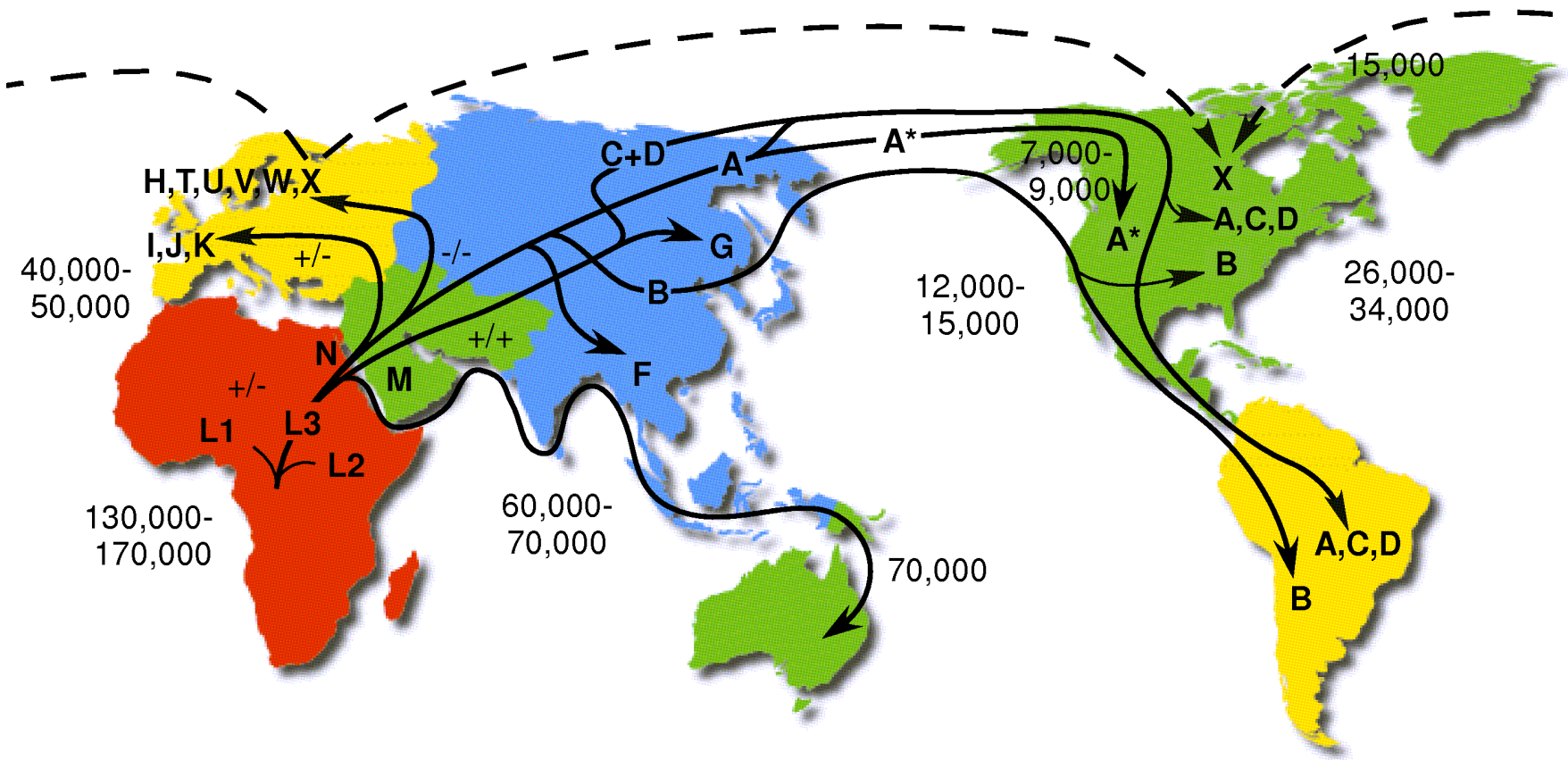


Classification result: Human Phylogeny

- **Phylogeny of mankind**
  - mtDNA (mitochondrial) sequences (17Kbp)
  - DNACompress, PPM
  - quartet-tree method
- Our results coincide with biologist's results (Cavalli et al.)
- Hints migration patterns, i.e. the Moroccans are genetically closer to the Spanish than to the Africans



Classification - Human World Migration



Regions highly conserved by evolution in the DNA of several species

- **Non-genetic regions contain highly conserved sequences**

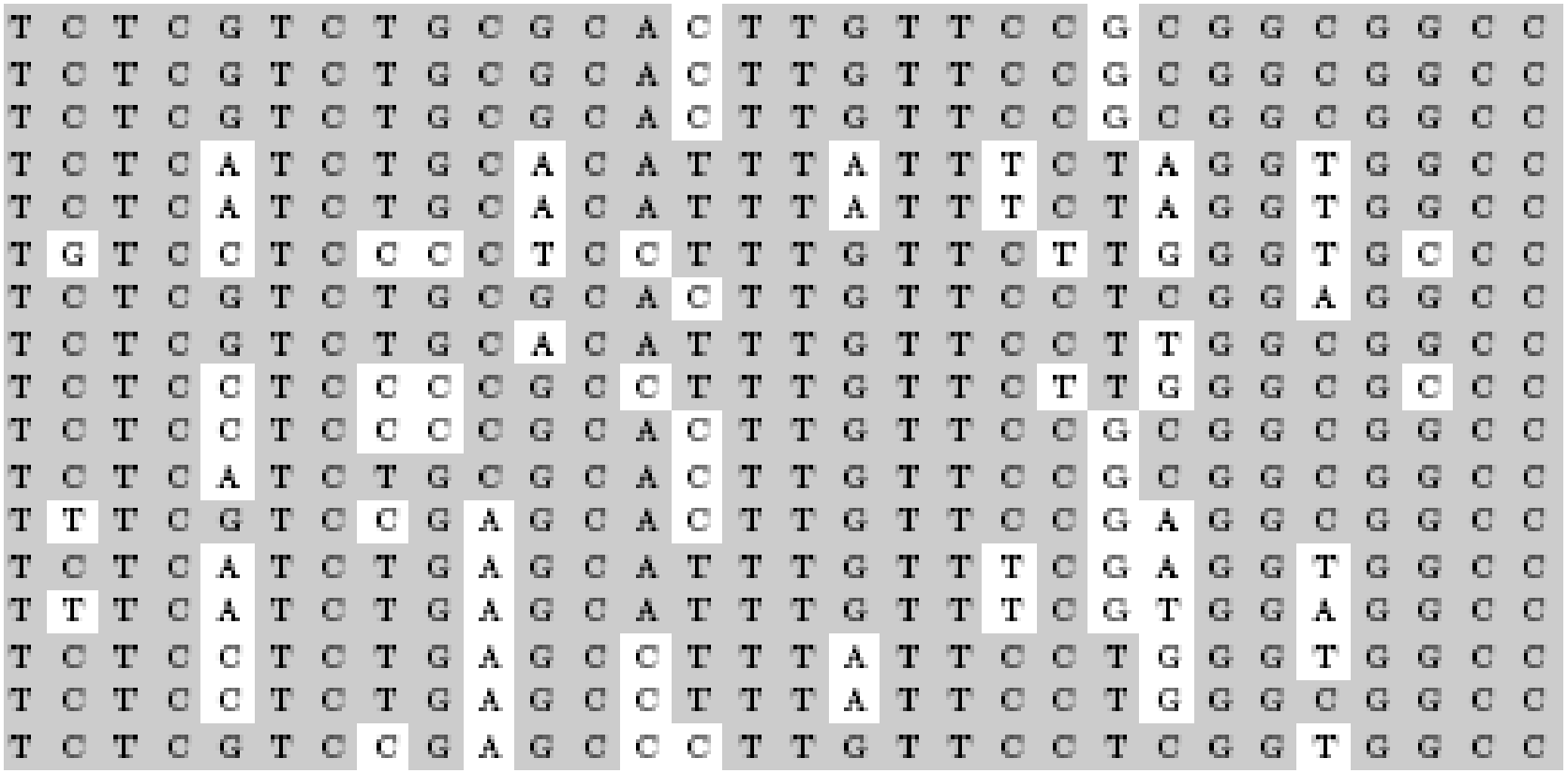
- **Some conserved sequences are related to biologically functional elements others are not, (ENCODE project, Nature, June 2007)**

- **It is important to identify conserved regions in the DNA**

- **How to measure evolutionary conservation ?**

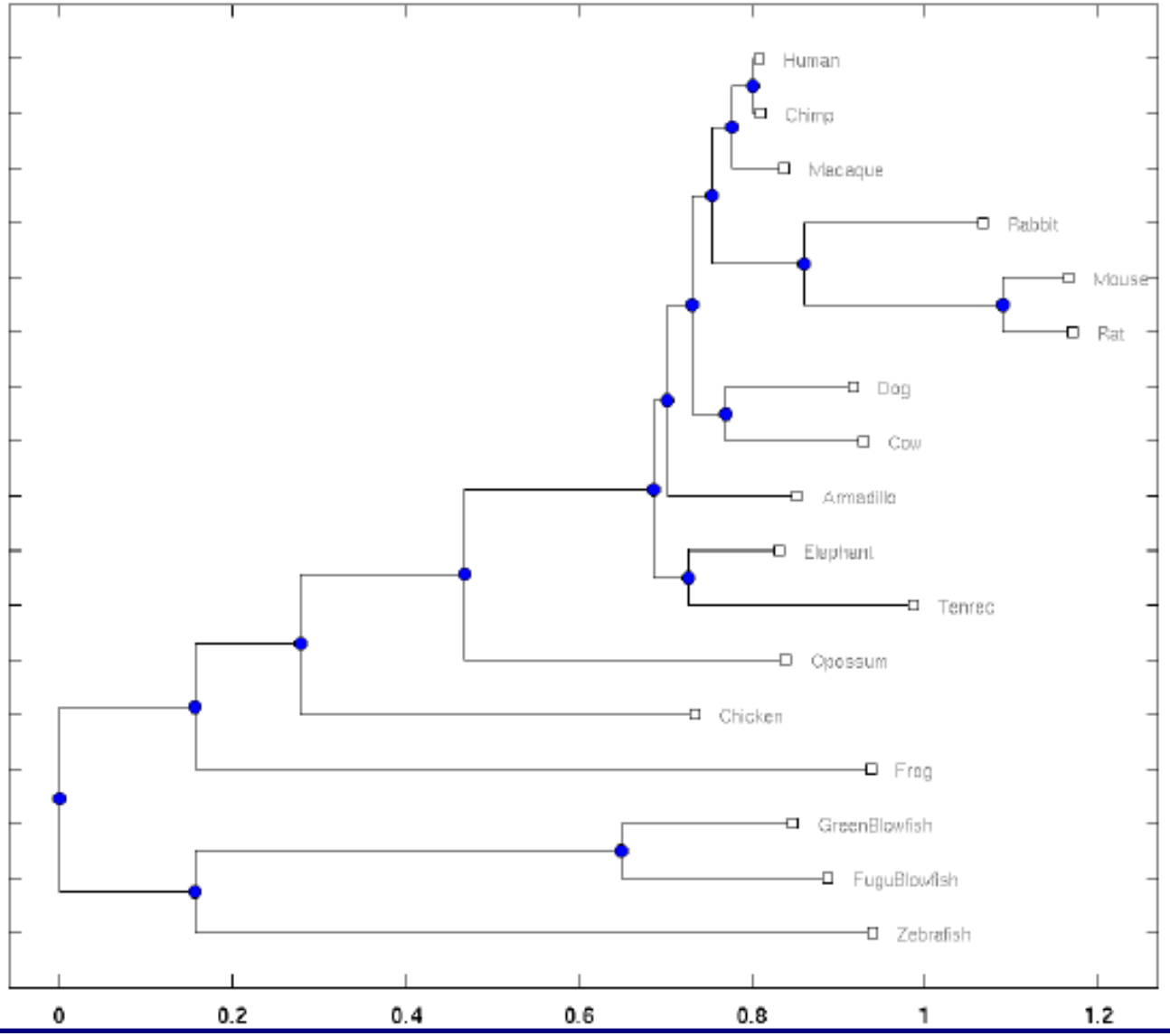
Aligned DNA for 17 species

Position a_i



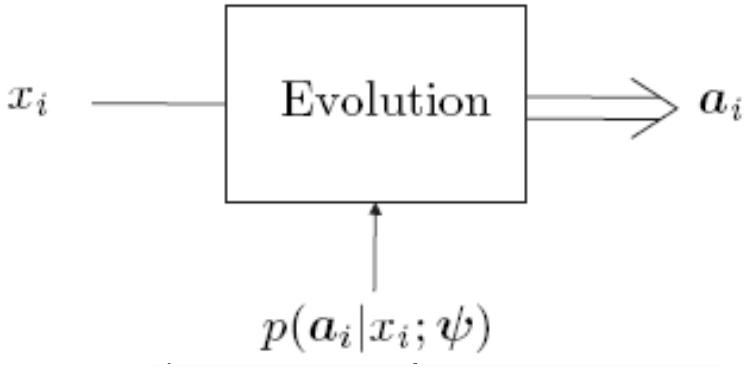
Evolutionary tree for 17 species with column a_i at position i :

Find probability for x_i= (A,G,C,T)



C C C C T T T A T T T T T T G T T C C

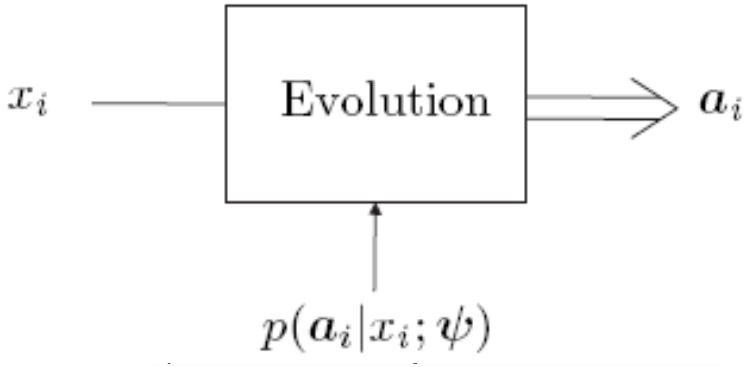
Transmission model of the evolution
Single-Input/Multiple-Output (SIMO)-Channel



$$\Psi = \{\mathcal{T}, \lambda, \mathbf{R}, \pi, \theta_i\}$$

- $\mathcal{T}$ is the evolutionary phylogenetic tree
- λ is the evolutionary distance on each of the branches
- $\mathbf{R}$ is the mutational rate matrix
- π is the a priori distribution of the bases,
- θ_i is a free parameter, the rate heterogeneity varying along the DNA

DMC Transmission model of the evolution



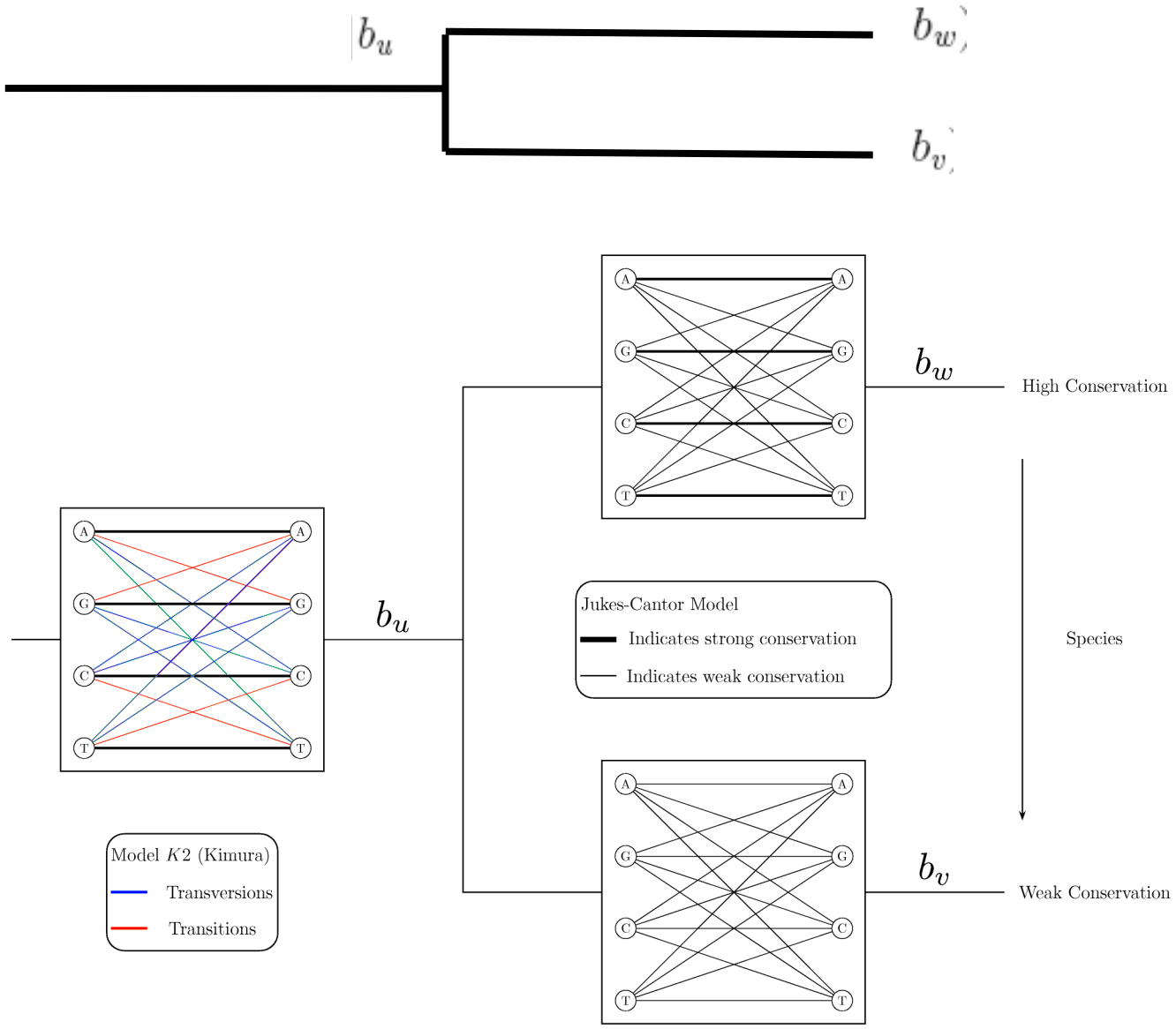
$$\Psi = \{\mathcal{T}, \lambda, \mathbf{R}, \pi, \theta_i\}$$

The evolution of a base at site i in each branch of its phylogenetic tree is then modeled by a DMC with a 4 x 4 probability matrix

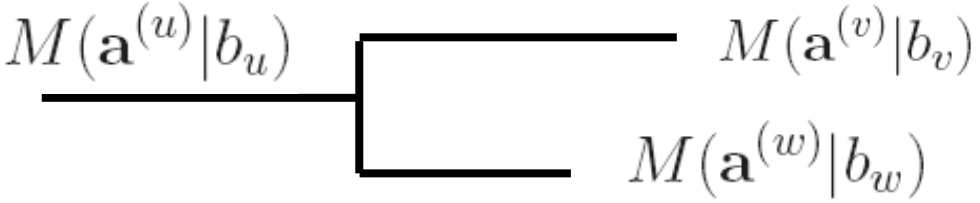
$$\mathbf{P} = e^{\theta_i \lambda \mathbf{R}}$$

$\theta_i = 0$ means full conservation (unaltered during evolution)

Simplified DMC transmission model of the evolution



Probability and metric estimation by a modified Viterbi algorithm for multiple sequence alignement



Observation $\mathbf{a}^{(u)}$ is the subtree stemming from base b_u

Felsenstein's algorithm transferred to Log-Max Domain with Metrics instead of probabilities
Let

$$M(b_v|b_u) = \ln P(b_v|b_u)$$

be given from the elements of $\mathbf{P}$ and

$$M(\mathbf{a}^{(u)}|b_u) = \ln P(\mathbf{a}^{(u)}|b_u)$$

Conservation Score

In order to obtain a conservation score we measure the distance between those two distributions by the Kullback-Leibler metric using p as full conservation

$$\mathcal{D}(p||q) = \sum p \ln \frac{p}{q}$$

$$\begin{aligned}
\sigma(\theta_i) &= \mathcal{D}(p||q) \\
&= \sum_{b_r} P(b_r) \left(M(\mathbf{a}_i, \theta_i = 0) - M(\mathbf{a}_i, \theta_i)\right) \\
&= \sum_{b_r} e^{M(b_r)} \left(M(b_r) - M(\mathbf{a}_i, \theta_i)\right) \\
&= - \underbrace{H_b}_{\text{base entropy}} - \sum_{b_r} e^{M(b_r)} \left(M(\mathbf{a}_i, \theta_i)\right)
\end{aligned}$$

Optimization of parameter θ

If we model the evolution with $\mathbf{a}_i$ being statistically independent over i we have the ML-estimate at

$$\hat{\theta}_i = \arg\max_{\theta_i} M(\mathbf{a}_i, \theta_i)$$

Usually the bases are correlated along the DNA strand with i. Then we use an exponentially weighted average over a window size of $2\delta + 1$:

$$\hat{\theta}_i = \arg\max_{\theta_i} \sum_{k=i-\delta}^{k=i+\delta} e^{-\frac{1}{2}(\frac{k-i}{\sigma_w})^2} M(\mathbf{a}_k, \theta_i)$$

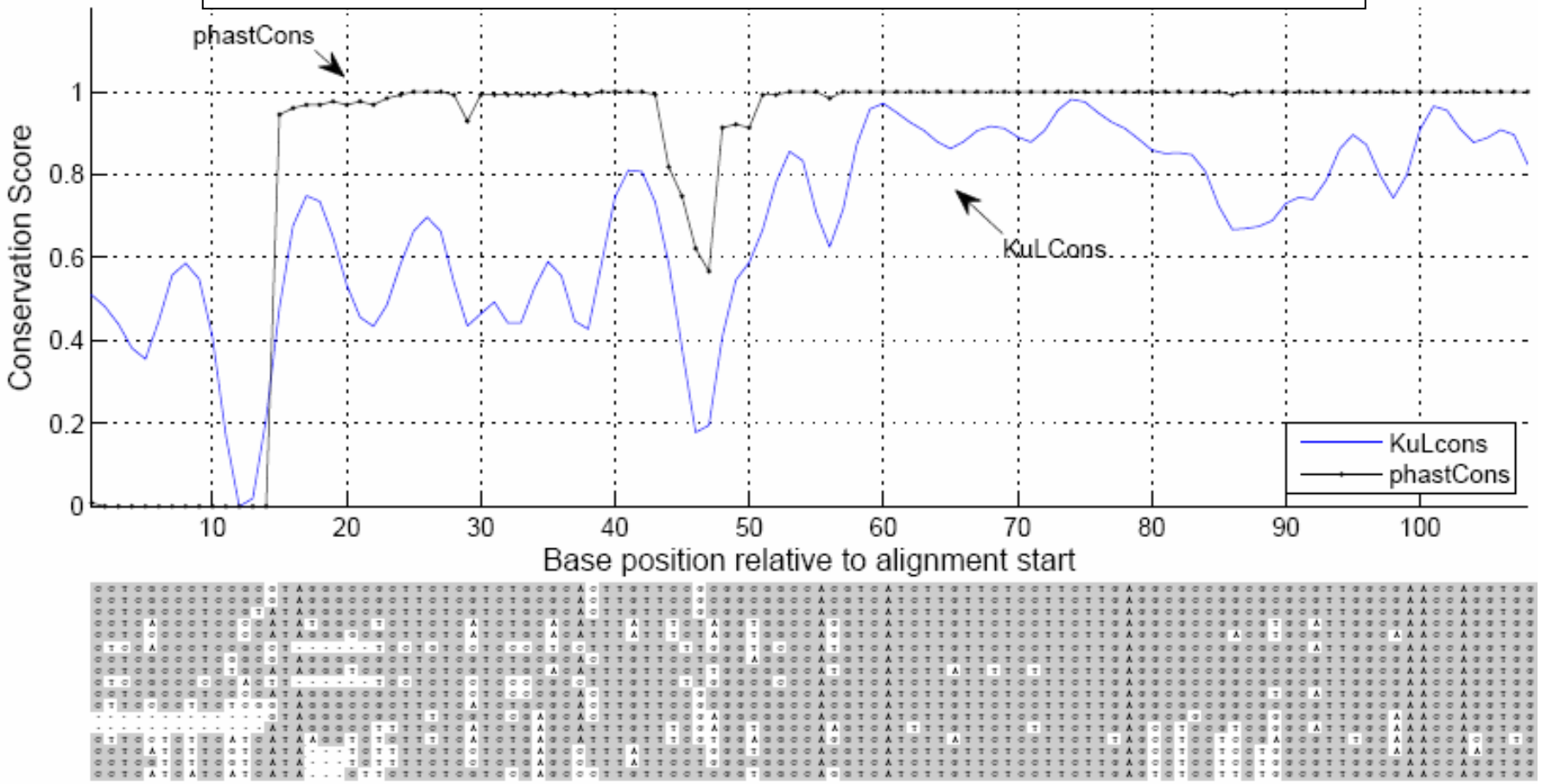
This Kullback-Leibler distance of our KuLCons method

$$\sigma(\hat{\theta}_i) = \sum_{b_r} e^{M(b_r)} \left(M(b_r) - M(\mathbf{a}_i, \hat{\theta}_i) \right)$$

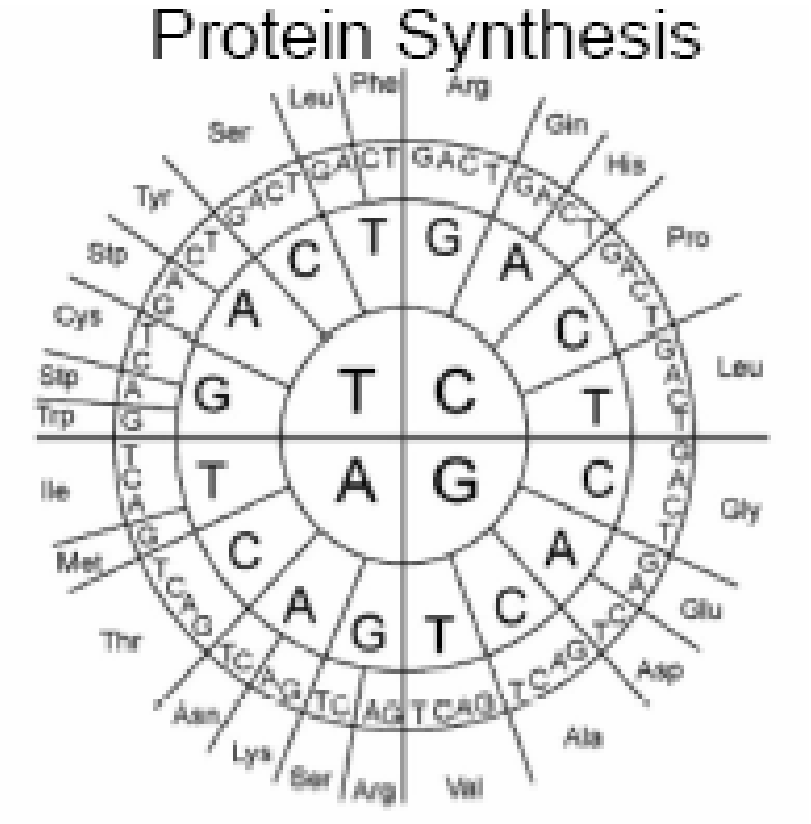
is compared with phastCons slightly transformed to ensure that the value 1 is the highest possible conservation score.

$$1 - \frac{\sigma(\hat{\theta}_i)}{\max_i \sigma(\hat{\theta}_i)}$$

Comparison of two conservation scores for 17 species



Final Example: Data Storage in Bacteria
The Genetic Code is redundant, use it for data storage
i.E Gly has 4 possibilities in 3rd position

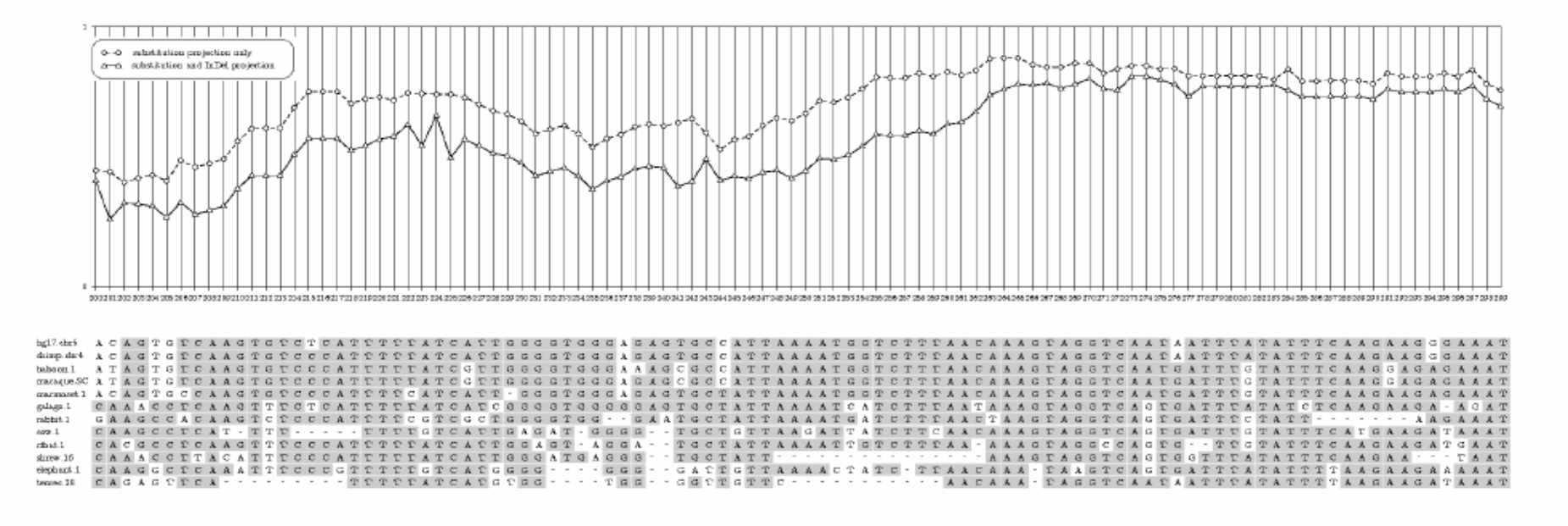


- *Bacteria*
  Escherichia Coli and
  Deincoccus radiodurans
  replicate quickly and are
  robust against radiation
- *Use redundany to hide information without changing proteins*
- *Our calculations show that one can store 1.67 bits per amino acid*
- *Use it for labelling, data hiding etc.*
- *Patents already exist*

Conclusions

- Mutual Information is the natural tool for classifying mutual relations between genetic sequences
- Compression algorithm in combination with a well defined mutual information distance measure can be used for classification of genetic data.
- Known phylogenetic trees are confirmed by our simpler mutual information method.
- Identification of highly conserved regions in non- genetic regions (CNGs) of species by ML-estimation with max-log approx. and Kullback-Leibler distance (extended to include insertions and deletions)
- Data storage in bacteria
- Our papers in “Nucleic Acids Research”, “BMCBioinformatics”, “IEEE Trans. on Comp.Biology” see www.lnt.ei-tum.de (publications)

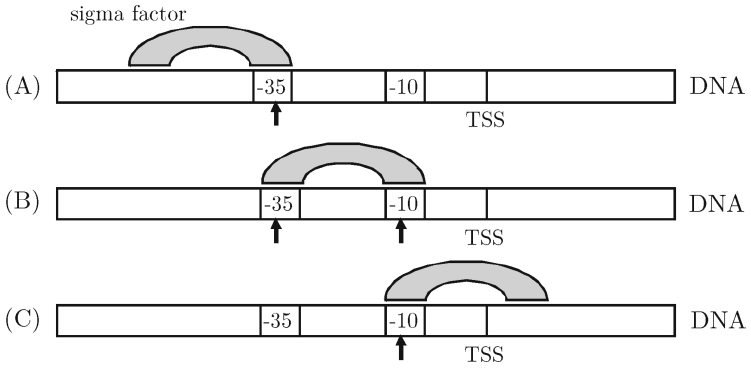
Conservation Score with insertions and deletions

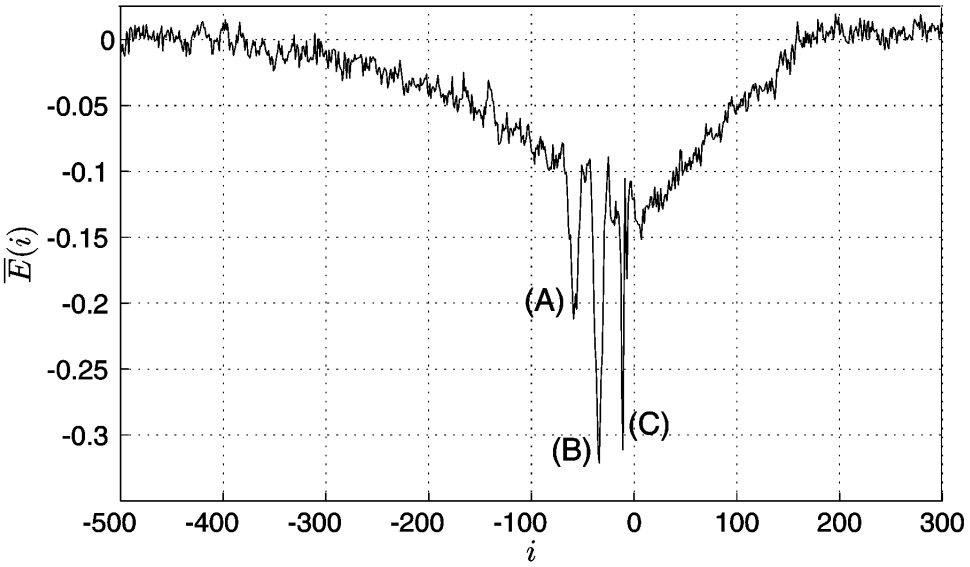


Further Work

- Information transfer between DNA-Variations and diseases for simulated and clinical data (Schizophrenia, Parkinson and Graves autoimmune disease) with measured mutual information

- Synchronization behavior and sync words in DNA to mRNA transcription for E.Coli bacteria





Molecular structure of A,G,C,T

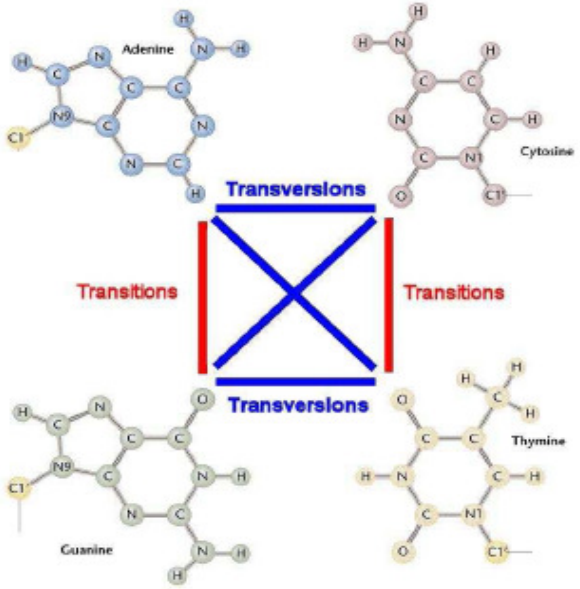
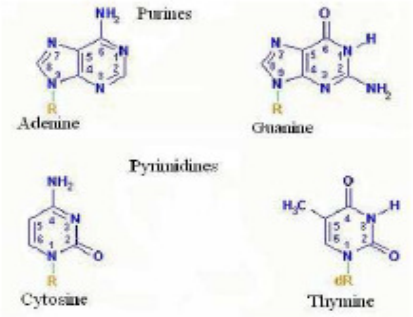


Figure 4.6: Transitions and Transversions [11].



Mutual Information as a distance measure between different DNA's.

Relate two DNA sources S_i and S_j by their mutual information $I(S_i;S_j)$

$$I(S_i;S_j) = H(S_i) - H(S_i|S_j)$$

$$0 \leq I(S_i;S_j) \leq \min(H(S_i), H(S_j)).$$

For classification purposes define the distance metric

$$d_{\mathrm{CL}}(S_i,S_j) = 1 - \frac{I(S_i;S_j)}{\max(H(S_i),H(S_j))} \leq 1.$$

The distance $d_{\mathrm{CL}}(S_i,S_j)$ can also be written as

$$d_{\mathrm{CL}}(S_i,S_j) = \frac{\max(H(S_i|S_j), H(S_j|S_i))}{\max(H(S_i),H(S_j))}$$

and is measured by the set sizes of a compression algorithm

$$d_{\mathrm{CL}} \approx \frac{|\mathrm{comp}(\mathrm{s_j},\mathrm{s_i})| - |\mathrm{comp}(\mathrm{s_j})|}{|\mathrm{comp}(\mathrm{s_i})|}.$$