P. Vitanyi, M&X Li, B. Ma

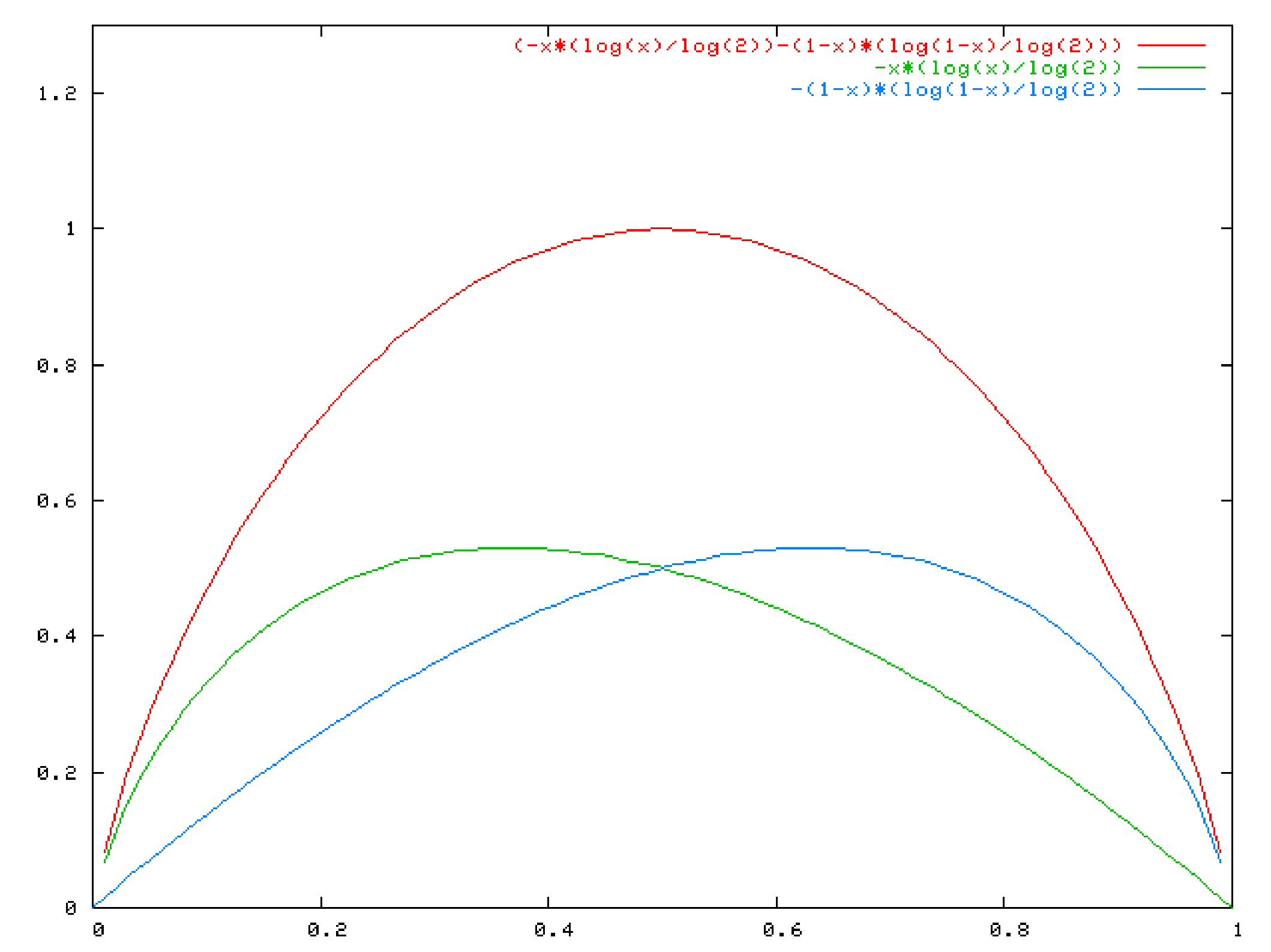
"Similarity Distance and Phylogeny"
aka "The Similarity Metric"
SIG-SIAM Discrete Algorithms '03

Sisukord

- Informatsioonisisaldus
- Kolmogorovi keerukus
- Tingimuslik keerukus
- Sarnasuse mõõtmine pakkimisega
- Näide: fülogeneesipuu
- Näide: keelte puu

Informatsioonisisaldus

- Infosisaldus
  - Sagedaste (tõenäoliste) sündmuste infosisaldus on väike
  - $I(s) = -\log(\Pr(s))$
  - $I(s)$ ~ bittide arv sõnumi kodeerimiseks
- Entroopia
  - Vähetõenäolised sündmused esinevad harva, kuid nende infosisaldus on suur
  - $H = \mathrm{SUM}[-\Pr(s)*\log(\Pr(s))]$


(-x*(log(x)/log(2))-(1-x)*(log(1-x)/log(2)))
-x*(log(x)/log(2))
-(1-x)*(log(1-x)/log(2))
1.2
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

Kolmogorovi keerukus (1965)

- 12345678910111213141516171819 2..
  - Iga sümbol võrdse tõenäosusega
  - Kui palju on selles arvus informatsiooni?
  - Milline on tema vähim esitus?
  - For i=1 to n do print i;
- Definitsioon (Kolmogorovi keerukus)
  - Objekti (teksti, arvu, pildi, bitistringi) keerukus on lühima programmi (Turingi masina) pikkus, mis väljastab algse objekti.

K(x) arvutamine

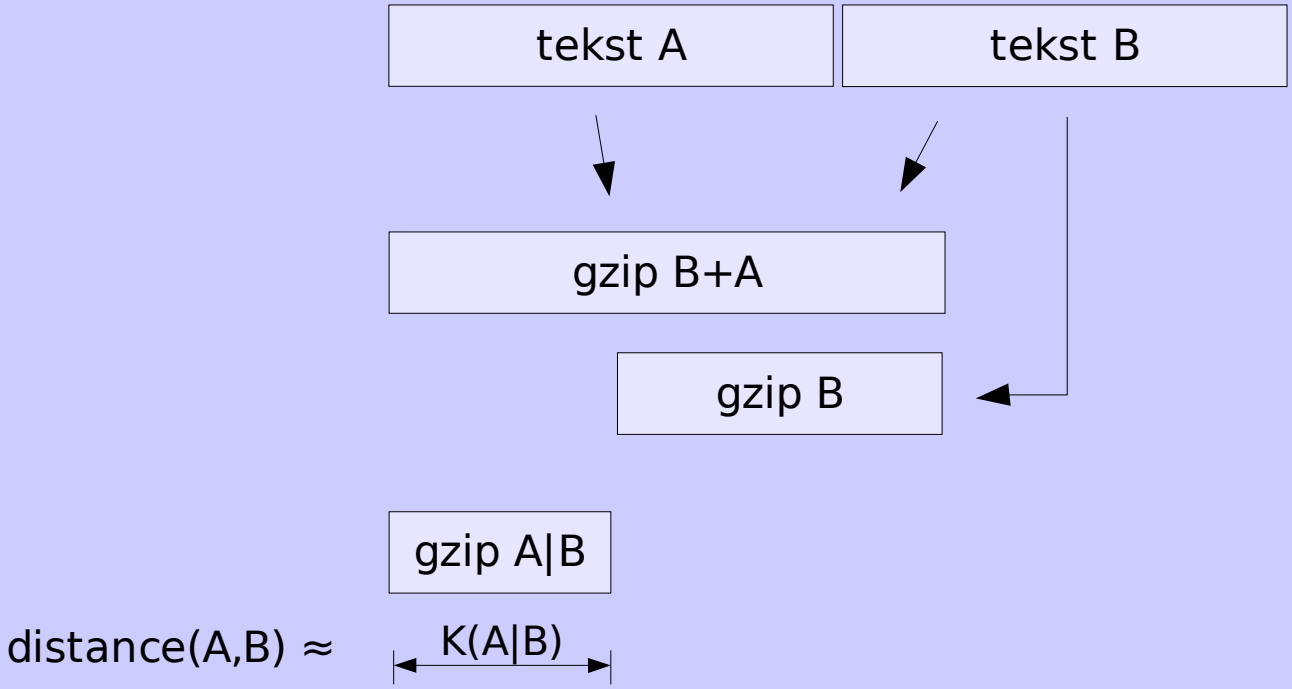
- K(x) ei ole arvutatav !!!
  - s.t ei leidu algoritmi, mis suudab iga x jaoks leida selle Kolmogorovi keerukuse (lühima programmi, mis väljastab x).
- K(x) on hinnatav
  - K(x) ≤ x kahendesitus
  - K(x+1) ≤ K(x) + c
- K(x) on lähendatav
  - K(x) ≤ *length*(gzip) + *length*(x.zip)

Tinglik keerukus

- K(x|y) = lühima programmi pikkus, mis väljastab x, kui y on teada.
  - Lühim programm, mis transformeerib y->x
  - Mõiste: transformatsiooni keerukus.
    - $K(x,y) \leq K(y) + K(x|y)$
    - $K(x|y) \leq K(x,y) - K(y)$
- Sarnasuse mõõtmine
  - Idee: *y* on parajasti nii sarnane *x*-le, kui lihtne on *x*-st saada *y*.
  - *distance*(x,y) ~ K(x|y)

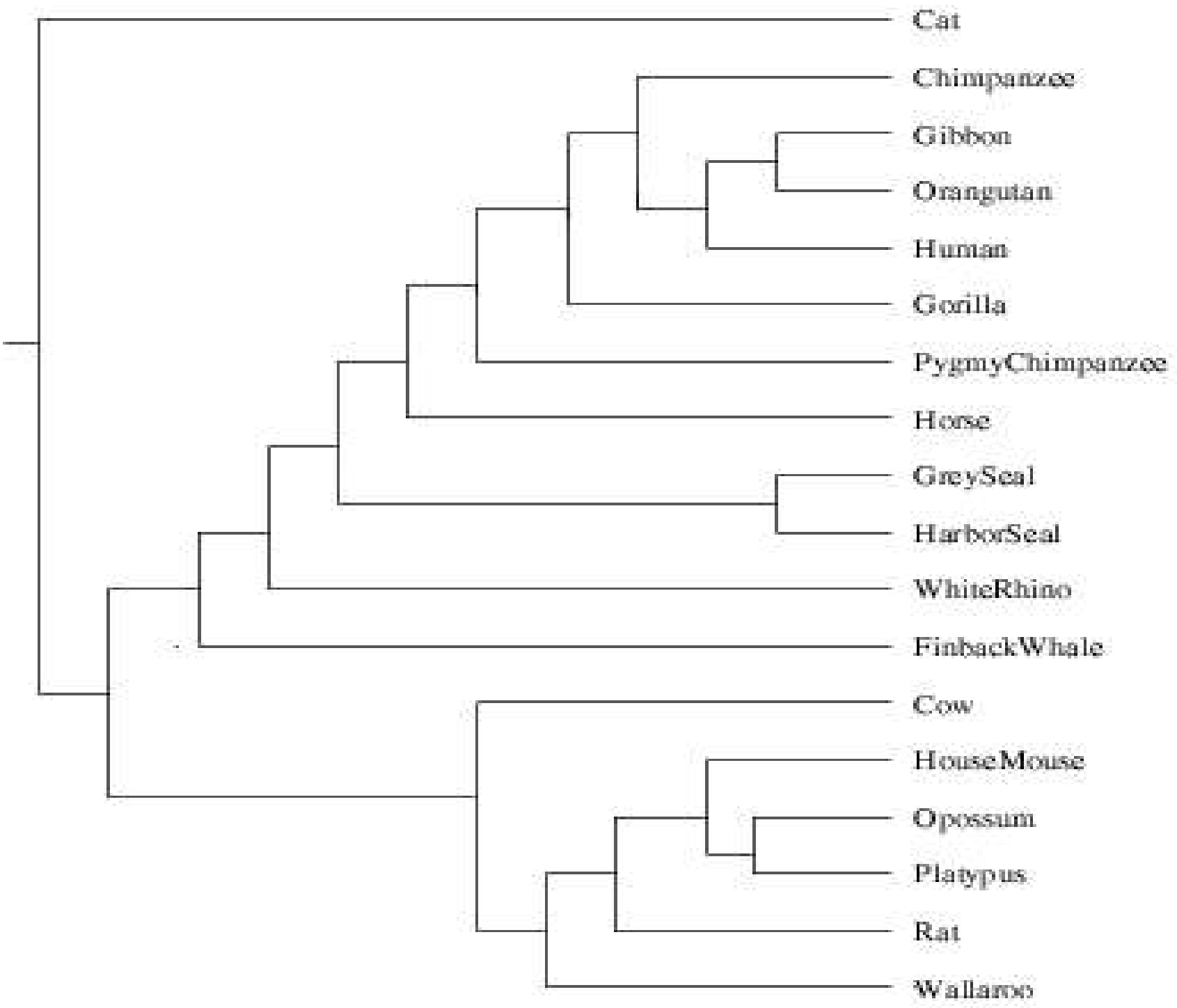
Sarnasus pakkimise kaudu

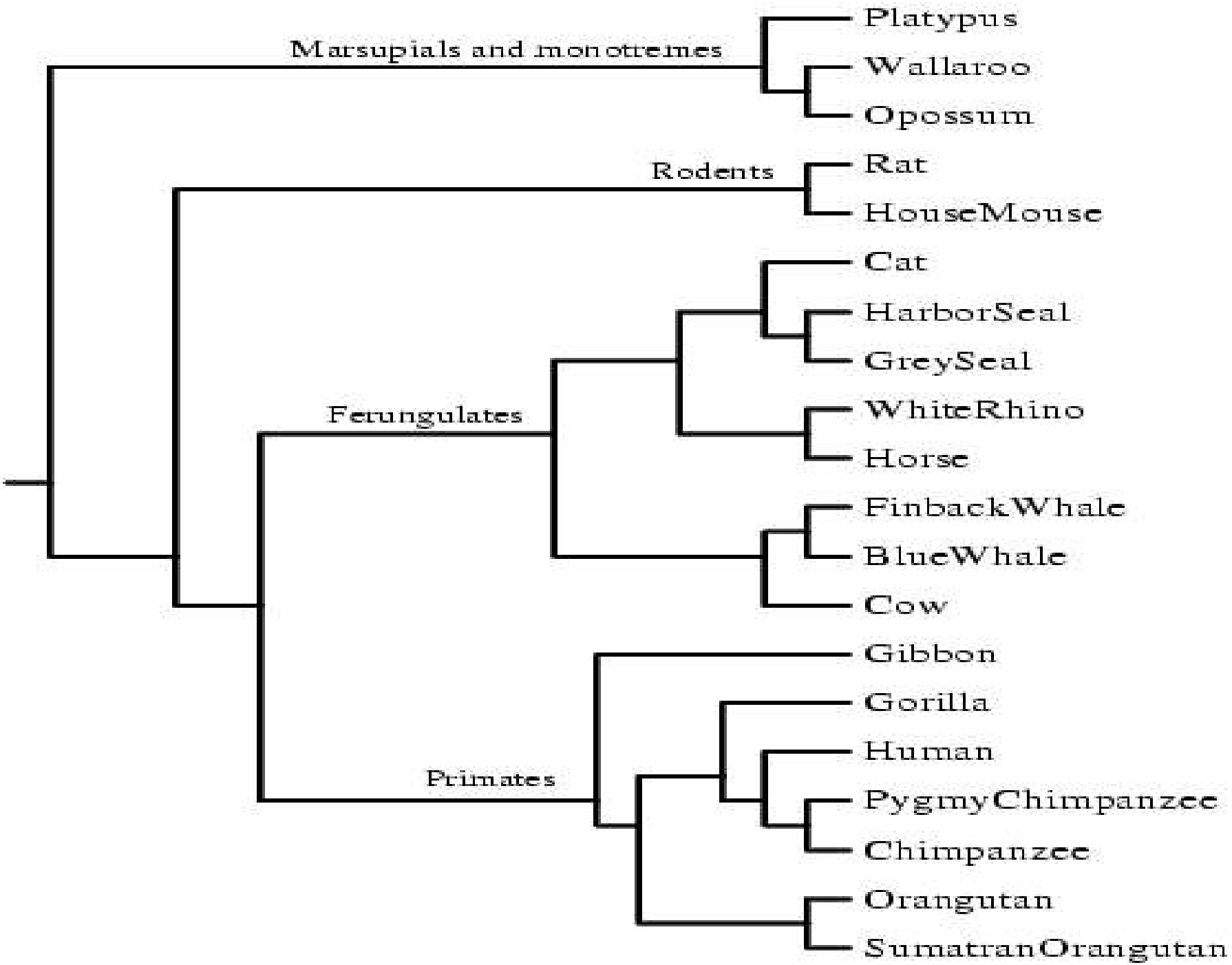
- $K(x|y) \approx K(x,y) - K(y)$
- *length*(gzip “y+x”) - *length*(gzip “y”)

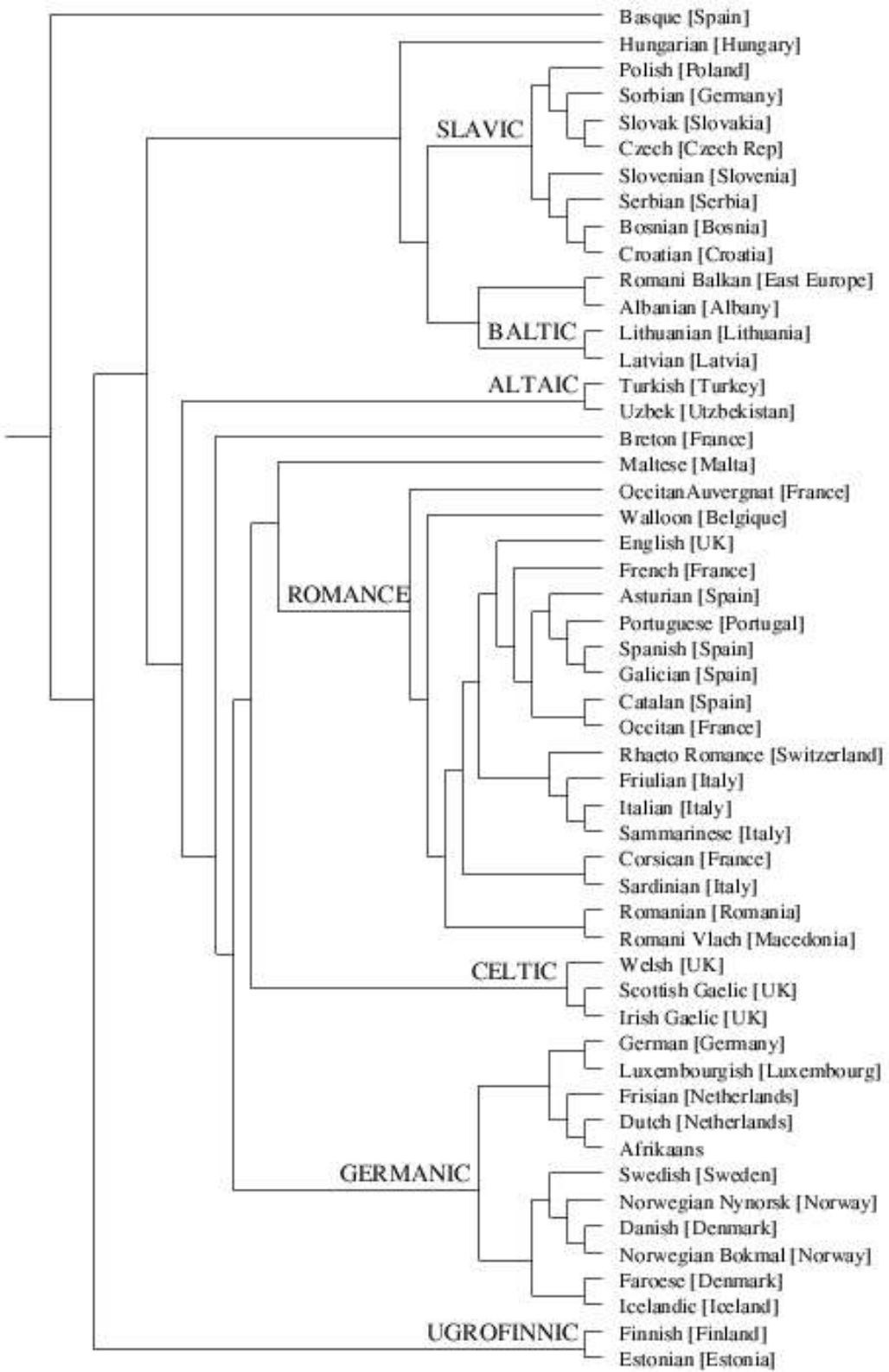


Evolutsioonipuu

- Levinud meetodid
  - K-gon (K-mer) statistik
  - Geenide järjekord genoomis
  - Genoomi ühisosa (ühiste geenide arv)
  - Ümberjärjestuskaugus
  - Transformatsioonikaugus
- Pakkimiskaugus (Vitanyi jt.)
  - $d(A,B) \approx [K(x|y) + K(y|x)] / [K(xy)]$


Cat
Chimpanzee
Gibbon
Orangutan
Human
Gorilla
PygmyChimpanzee
Horse
GreySeal
HarborSeal
WhiteRhino
FinbackWhale
Cow
HouseMouse
Opossum
Platypus
Rat
Wallaroo


Marsupials and monotremes
Platypus
Wallaroo
Opossum
Rodents
Rat
HouseMouse
Fetungulates
Cat
HarborSeal
GreySeal
WhiteRhino
Horse
FinbackWhale
BlueWhale
Cow
Primates
Gibbon
Gorilla
Human
PygmyChimpanzee
Chimpanzee
Orangutan
SumatranOrangutan


Basque [Spain]
Hungarian [Hungary]
Polish [Poland]
Sorbian [Germany]
Slovak [Slovakia]
Czech [Czech Rep]
Slovenian [Slovenia]
Serbian [Serbia]
Bosnian [Bosnia]
Croatian [Croatia]
SLAVIC
Romani Balkan [East Europe]
Albanian [Albany]
Lithuanian [Lithuania]
Latvian [Latvia]
BALTIC
Turkish [Turkey]
Uzbek [Utzbekistan]
ALTAIC
Breton [France]
Maltese [Malta]
Occitan Auvergnat [France]
Walloon [Belgique]
English [UK]
French [France]
Asturian [Spain]
Portuguese [Portugal]
Spanish [Spain]
Galician [Spain]
Catalan [Spain]
Occitan [France]
ROMANCE
Rhaeto Romance [Switzerland]
Friulian [Italy]
Italian [Italy]
Sammarinese [Italy]
Corsican [France]
Sardinian [Italy]
Romanian [Romania]
Romani Vlach [Macedonia]
Welsh [UK]
Scottish Gaelic [UK]
Irish Gaelic [UK]
CELTIC
German [Germany]
Luxembourgish [Luxembourg]
Frisian [Netherlands]
Dutch [Netherlands]
Afrikaans
Swedish [Sweden]
Norwegian Nynorsk [Norway]
Danish [Denmark]
Norwegian Bokmal [Norway]
Faroese [Denmark]
Icelandic [Iceland]
GERMANIC
Finnish [Finland]
Estonian [Estonia]
UGROFINNIC