

# Informatsioonikaugus

Mart Sõmermaa Andmekaevandamise uurimisseminar MTAT.03.169.

Arvutiteaduse instituut, Tartu Ülikool

Detsember 2003, lk. 90–100

## Kokkuvõte

Käesolevas artiklis antakse ülevaade sõne Kolmogorovi keerukuse mõistel põhinevast meetrikast, nn *informatsioonikaugusest*. Käsitletakse meetrika normaliseerimise ja universaalsuse küsimusi ning kirjeldatakse informatsioonikauguse rakendusi geneetikas, lingvistikas ja autorsuse tuvastamisel.

## 1. Sissejuhatus

Objektide sarnasuse määramine on andmekaevanduse üks fundamentaalprobleeme, sarnasuse mõistel põhinevad nii otsingu- kui klassifitseerimisalgoritmid. Ming Li, Paul Vitányi *et al.* esitavad artiklites [LCL<sup>+</sup>03, BGL<sup>+</sup>98] üldise sarnasuse teooria, näidates, et sõne informatsioonisisalduse (Kolmogorovi keerukuse) abil saab defineerida universaalse meetrika, nn *informatsioonikauguse*. Universaalsus tähendab siinkohal seda, et kahe objekti informatsioonikaugus on minimaalne kõigi samadel objektidel arvutatud normaliseeritud kauguste hulgas. Seega, kui kaks objekti on sarnased suvalise normaliseeritud meetrika põhjal, on need vähemalt sama sarnased informatsioonikauguse põhjal.

Ilmneb, et informatsioonikaugus on ka praktikas hästi kasutatav, kuigi, kuna sõne  $x$  Kolmogorovi keerukus  $K(x)$  ei ole arvutatav, kasutatakse rakendustes tihendamist kui  $K(x)$  heuristilist lähendit.

Järgnevas anname esmalt ülevaate levinumatest üldistest meetrikatest ja toome esile nende mõningad puudused; anname vajalike mõistete definitsioonid ja artikli kontekstis olulised tulemused informatsiooniteoorias; defineerime kaks versiooni informatsioonikaugusest, käsitleme informatsioonikauguse normaliseerimist ja universaalsust. Lõpetuseks anname ülevaate informatsioonikauguse rakendustest.

## 2. Kaugus

Kaugus (meetrika) on mingi hulga  $M$  otseruudul määratud mittenegatiivne funktsionaal  $d : M \times M \rightarrow \mathbb{R}$ , mis iga  $x, y, z \in M$  korral rahuldab nn Fréchet' aksioome:

1.  $d(x, y) = 0$  parajasti siis, kui  $x = y$  (*samasusaksioom*),
2.  $d(x, y) = d(y, x)$  (*sümmeetriaaksioom*),
3.  $d(x, y) \leq d(x, z) + d(z, y)$  (*kolmnurgaaksioom*).

Hulka  $M$ , millel on defineeritud kaugus, nimetatakse meetriliseks ruumiks. Edaspidises näidatakse, et paar  $(\{0, 1\}^*, d)$ , kus  $d$  on informatsioonikaugus, on meetriline ruum.

Järgnevalt loetleme levinuimad üldised meetrikad.

1. Boole'i kaugus (diskreetne meetrika), on defineeritud igal hulgal

$$\begin{cases} d_B(x, y) = 0, & \text{kui } x = y, \\ d_B(x, y) = 1, & \text{vastasel juhul;} \end{cases}$$

2. Kauguste pere  $L_m : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $m \in \mathbb{N}$ , kus  $\mathbb{R}^n$  on vektorruum üle reaalarvude korpuse<sup>1</sup>, erinevate normide suhtes (Minkowski kaugused),

$$L_m(x, y) = \|x - y\|_m = \left( \sum_{i=1}^n |x_i - y_i|^m \right)^{\frac{1}{m}}.$$

Peamiselt pakuvad huvi juhud  $m = 1$  (Manhattani kaugus) ja  $m = 2$  (eukleidiline kaugus) ning maksimumnorm  $L_\infty$ ,

$$L_\infty(x, y) = \|x - y\|_\infty = \max(|x_1 - y_1|, \dots, |x_n - y_n|).$$

3. Hulkade kaugus  $d_\Delta : \mathcal{P}(X) \times \mathcal{P}(X) \rightarrow \mathbb{N}$ , kus  $X$  on mingi universaalne hulk,  $\mathcal{P}(X)$  tähistab hulga  $X$  alamhulkade hulka ja  $\Delta$  hulkade sümmeetrilist vahet. Kui  $A \subset X, B \subset X$ , siis

$$d_\Delta(A, B) = |A \Delta B|.$$

---

<sup>1</sup>kaugused  $L_m$  on defineeritud ka kompleksarvude korpusel.

4. Hammingu kaugus  $d_H : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \mathbb{N}$ ,

$$d_H(x, y) = \sum_{i=1}^n x_i \oplus y_i = \sum_{i=1}^n |x_i - y_i| = d_1,$$

kus  $\oplus$  tähistab välistavat disjunktsiooni (XOR). Seega on Hammingu kaugus Manhattani kauguse erijuht. Seades hulgale vastavusse tema karakteristikliku funktsiooni, saame ka vastavuse hulkade kauguse ja Hammingu kauguse vahel.

5. Teisenduskaugus (Levenšteini kaugus)  $d_E$  on Hammingu kauguse üldistus — kui viimane on bittide arv, mida tuleb muuta, et teisendada kahendsõne  $x$  sõneks  $y$ , siis  $d_E$  on vähim lisamiste, eemaldamiste ja asenduste arv sõne  $x$  teisendamisel sõneks  $y$ . On lihtne näidata, et meetrika  $d_E$  rahuldab Frechet' aksioome, näiteks esitatakse tõestus artiklis [Sör03].

Lisaks nimetatutele on olemas hulgaliselt spetsiifilisi kaugusi. Eraldi tuleks märkida, et ka varasemates töödes on püütud tihendamist sõnede kauguse hindamisel kasutada, enne Ming Li ja Paul Vitányi töögrupi artiklite publitseerimist pole aga esitatud kõiki Frechet' aksioome rahuldavat teoreetilisel hästipõhjustatud meetrikat.

Paljudes rakendustes on loetletud kaugused piisavad objektide sarnasuse määramiseks. On aga probleeme, mille lahendamiseks oleks vaja üldisemat meetrikat — näiteks kahe must-valge pildi sarnasuse määramine. Olgu antud mingi pildi esitus kahendsõnena. Kui muuta sõnes üksikuid bitte, jääb nii Hammingu kui eukleidiline kaugus originaali ja modifikatsiooni vahel väikeseks. Samas, pildi negatiivkujutis on mõlema meetrika korral originaalist maksimaalselt kaugel, inimvaatleja suudab aga piltide sarnasust tuvastada. Sama probleem ilmneb bitinihke korral — originaali ja modifikatsiooni Hammingu kaugus võib olla väga suur, kuigi pildid on sarnased. Analoogilisi kauguse spetsiifikast tulenevaid probleeme on ka teistes valdkondades, muuhulgas bioinformaatikas.

Informatsioonikaugus hõlmab kõiki loetletud meetrikaid, omamata paljude spetsiifiliste meetrikate puudusi. Muuhulgas määrab see ka eelmises lõigus kirjeldatud originaali ja negatiivi kauguse vastavalt inimvaatleja intuitsioonile.

### 3. Kolmogorovi keerukus

Põhjaliku ülevaate Kolmogorovi keerukusest ja informatsiooniteooriast annab Ming Li ja Paul Vitányi raamat [LV97]. Käesolevaga anname lühi-

dalt edaspidises vajalikud tähistused ja mõisted. *Sõneks* nimetame edaspidi kahendjärjendit (so lõplikku kahendjada). Igast lõplike objektide hulgast leidub loomulik kujutus sõnede hulka. Sõnede hulka tähistatakse  $\{0, 1\}^*$ . Sõne  $x$  Kolmogorovi keerukus või algoritmiline entroopia  $K(x)$  on lühima kahendprogrammi pikkus, mis väljastab sõne  $x$  mingil antud universaalarvutil (näiteks Turingi masinal). Tähistagu  $x^*$  lühimat programmi, mis väljastab sõne  $x$ , seega  $|x^*| = K(x)$ , kus  $|s|$  tähistab sõne  $s$  pikkust. Kuigi funktsioon  $K$  on defineeritud antud masina jaoks, järeldub Churchi teesist<sup>2</sup>, et see on mingi konstantse liidetavani masinasõltumatu ja universaalne.

Sõne  $x$  tinglik Kolmogorovi keerukus sõltuvalt sõnest  $y$  defineeritakse kui lühima programmi pikkus, mis väljastab sõne  $x$  sisendsõne  $y$  korral, tingliku keerukust tähistatakse  $K(x|y)$ . Sümboliga  $K(x, y)$  tähistatakse lühima kahendprogrammi pikkust, mis väljastab sõned  $x$  ja  $y$  ning kirjelduse, kuidas neid eristada. Saab näidata, et mingi konstantse liidetavani

$$K(x, y) = K(x) + K(y|x^*). \quad (1)$$

**Definitsioon 1.** *Funktsioon  $f(x)$  on*

- ülalt arvutatav<sup>3</sup>, kui leidub rekursiivne funktsioon  $g(x, t)$  selliselt, et

$$g(x, t+1) \leq g(x, t) \quad \text{ja} \quad \lim_{t \rightarrow \infty} g(x, t) = f(x),$$

- alt arvutatav, kui  $-f$  on ülalt arvutatav ning
- arvutatav, kui ta on nii ülalt kui alt arvutatav.

Lihtne on näha, et funktsioonid  $K(x)$  ja  $K(y|x)$  on ülalt arvutatavad — alati saab  $K(x)$  ülemiseks lähendiks valida suuruse  $|x|$ , st Turingi masina programmi pikkuse, mis lihtsalt kopeerib lindi sisu väljundisse. Saab tõestada, et need funktsioonid ei ole arvutatavad.

Sõnes  $y$  sisalduv informatsioon sõne  $x$  kohta defineeritakse järgmiselt

$$I(x : y) = K(x) - K(x|y^*).$$

Oluline tulemus informatsiooniteoorias on informatsioonisisalduse sümmeetrilisus: saab näidata, et mingi konstantse liidetavani  $I(x : y) \stackrel{\pm}{=} I(y : x)$  ning seega

$$K(x) + K(y|x^*) \stackrel{\pm}{=} K(y) + K(x|y^*). \quad (2)$$

---

<sup>2</sup>Churchi teesis väidetakse, et mingi formalisatsiooniga arvutatavate funktsioonide klass langeb kokku üleüldse kõigi efektiivsete meetoditega arvutatavate funktsioonide klassiga; sellest järeldub, et universaalarvutid suudavad üksteist emuleerida.

<sup>3</sup>*upper semi-computable*

Sümboleid  $\pm$  ja  $\stackrel{\log}{\equiv}$  kasutame edaspidi tähistamiseks võrdust vastavalt mingi konstantse ja  $O(\log x)$  liidetavani, kusjuures logaritmi argumentiks  $x$  on alati võrduse paremal poolel olev avaldis.

## 4. Informatsioonikaugus

Sõnede  $x$  ja  $y$  vaheline *informatsioonikaugus*  $E(x, y)$  on lühima kahendprogrammi pikkus, mis väljastab sisendsõne  $x$  korral sõne  $y$  ja vastupidi. Kuna programm on lühim, kasutab see ära kogu andmeliiasuse mõlemas suunas teisendamisel. Nõutakse, et programm teisendamise käigus ei muutuks.

Artiklis [BGL<sup>+</sup>98] näidatakse, et mingi argumentide keerukuse kasvamisel hääbuva veaga kehtib võrdus

$$E(x, y) \stackrel{\log}{\equiv} \max\{K(y|x), K(x|y)\}. \quad (3)$$

Kuna tinglik Kolmogorovi keerukus on ülalt arvutatav, on ka informatsioonikaugus ülalt arvutatav.

## 5. Normaliseeritud kaugus

Paneme tähele, et

- pikad sõned, mis erinevad  $n$  biti võrra, on intuiitiivselt sarnasemad kui lühikesed sõned, mis erinevad samuti  $n$  biti võrra;
- meetrikate võrdlemiseks peavad nende väärtused kuuluma samasse vahemikku;
- meetrikate võrdlemisel on vaja välistada vähese kirjeldusjõuga kaugused, mida me ka intuiitiivselt kasutuks peame, nagu näiteks kaugus  $d_?(x, y) = \frac{1}{2}, \forall x \neq y$ .

Seetõttu nõuame käsitletavatelt meetrikatelt lisaks Frechet' aksioomidele kahe lisatingimuse rahuldamist.

**Definitsioon 2.** Normaliseeritud sõnekauguseks nimetatakse ülalt poolarvutatavat meetrikat  $m(x, y), x, y \in \{0, 1\}^*$ ,

1. mille väärtused kuuluvad vahemikku  $[0, 1]$  (protsessis  $\max\{K(x), K(y)\} \rightarrow \infty$  hääbuva veaga) ja

2. mis rahuldab tiheduse tingimust

$$\sum_{y \neq x} 2^{-m(x,y)K(x)} \leq 1. \quad (4)$$

Intuitiivselt on mõistetav, et meetrikad on tõepoolest normeeritavad vahemikku  $[0, 1]$  mingi argumentidest sõltuva jagaja abil. Näiteks hulkade kauguse  $d_{\Delta}(A, B)$  normeerimiseks tuleb kasutada jagajat  $|A \cup B|$  ja eukleidilise kauguse  $L_2(x, y)$  normeerimiseks jagajat  $|x| + |y|$ . Vastavad tõestused on esitatud artiklis [Yia02].

Tingimus (4) piirab objektide hulka fikseeritud objekti mingis raadiuses, selle abil välistatakse ebasoovitavad kaugused. Sisuliselt tähendab tingimus nõuet, et sõnest  $x$  oleks kaugusel  $d$  ülimalt  $2^{dK(x)}$  sõnet.

## 6. Normaliseeritud informatsioonikaugus

Artiklis [LBX<sup>+</sup>01] esitati esialgne normaliseeritud informatsioonikauguse definitsioon:

**Definitsioon 3.** *Olgu  $x, y$  suvalised järjendid. Defineerime funktsiooni*

$$d_s(x, y) = \frac{K(x | y^*) + K(y | x^*)}{K(x, y)}. \quad (5)$$

Võrduste (1) ja (2) põhjal saame

$$d_s(x, y) = 1 - \frac{K(x) - K(x | y^*)}{K(x, y)},$$

kus  $K(x) - K(x | y^*)$  on sõnede ühine informatsioon  $I(y : x)$ . See kaugus rahuldab kolmnurga võrratust mingi liidetava veaga ja universaalsustingimust (defineeritakse edaspidi) mingi konstantse tegurini<sup>4</sup>  $c < 2$ .

Matemaatiliselt täpsem ja adekvaatsem on järgnev definitsioon.

**Definitsioon 4.** *Olgu  $x, y$  suvalised järjendid. Defineerime funktsiooni*

$$d(x, y) = \frac{\max\{K(x | y^*), K(y | x^*)\}}{\max\{K(x), K(y)\}}. \quad (6)$$

---

<sup>4</sup>saab näidata, et  $d_s(x, y) \leq cD(x, y)$ ,  $c < 2$ , kus  $D$  on suvaline normaliseeritud sõnekaugus.

Meetrika  $d(x, y)$  loomulik interpretatsioon: kui  $K(y) \geq K(x)$ , siis

$$d(x, y) = \frac{K(y) - I(x : y)}{K(y)},$$

kus lugejas on sõnedes  $x$  ja  $y$  mittejagatud informatsiooni kogus bittides, ning nimetajas normeerimistegur, maksimaalne võimalik jagatud informatsioonikogus.

On ilmne, et  $d(x, y)$  on sümmeetriline ja rahuldab samasusaksioomi,

$$d(x, x) = O\left(\frac{1}{K(x)}\right).$$

Saab näidata, et  $d(x, y)$  rahuldab hääbuva veaga kolmnurgaaksioomi:

$$d(x, y) \leq d(x, z) + d(z, y) + O\left(\frac{1}{\max\{K(x), K(y), K(z)\}}\right),$$

ning normaliseerimistingimust (4). Seega kehtib järgnev teoreem:

**Teoreem 1.** *Funktsioon  $d(x, y)$  on normaliseeritud sõnekaugus.*

## 7. Universaalsus

Ilmneb, et meetrika  $d(x, y)$  hõlmab kõiki arvutatavaid sarnasusmõõte — fikseeritud objektide informatsioonikaugus on minimaalne kõigi samadel objektidel arvutatud normaliseeritud kauguste hulgas.

**Teoreem 2.** *Normaliseeritud informatsioonikaugus  $d(x, y)$  on protsessis  $\max\{K(x), K(y)\} \rightarrow \infty$  hääbuva veaga asümptootiliselt ekvivalentne või asümptootiliselt kõrgemat järku mistahes ülalt arvutatava normaliseeritud meetrika  $f(x, y)$  suhtes,*

$$d(x, y) \leq f(x, y) + O\left(\frac{\log k}{k}\right), \text{ kus } k = \max\{K(x), K(y)\}.$$

Sisuliselt tähendab teoreem seda, et kui kaks objekti on sarnased (normaliseeritud jagatud informatsioonikoguse mõttes) suvalise arvutatava normaliseeritud meetrika järgi, siis on need objektid vähemalt sama sarnased meetrika  $d(x, y)$  järgi.

Teoreemi tõestus on esitatud artiklis [LCL<sup>+</sup>03].

## 8. Rakendused

Kuigi meetrika  $d$  on teoreetiliselt hästipõhjendatud ja universaalne, ei ole  $K(x)$  ja seega ka  $d$  arvutatav. Intuitiivselt on mõistetav, et *tihendamine* on  $K(x)$  heuristiline lähend. Rakendustes kasutataksegi valdkonnaspetsiifilisi tihendusalgoritme meetrikate  $d_s$  ja  $d$  lähendite leidmisel. Rõhutame siinkohal, et meetrika  $d$  kasutamine annab üldjuhul paremaid tulemusi.

Meetrikad tuleb esmalt sobivale kujule viia. Seoste (1) ja (2) põhjal  $K(x|y) \stackrel{\pm}{=} K(x, y) - K(y)$  ning on lihtne näidata, et  $K(x, y) \stackrel{\log}{=} K(xy)$ , kus  $xy$  tähistab sõnade  $x$  ja  $y$  konkatenatsiooni. Seega

$$d_s(x, y) = 1 - \frac{K(x) - K(x|y)}{K(x, y)} \approx 1 - \frac{K(x) + K(y) - K(xy)}{K(xy)},$$

ning kasutades tihendusalgoritmi  $C(s)$  kui  $K(s)$  heuristilist lähendit,  $K(s) \approx |C(s)|$ , saame

$$d_s(x, y) \approx 1 - \frac{|C(x)| + |C(y)| - |C(xy)|}{|C(xy)|}. \quad (7)$$

Analoogiliselt, eeldades üldisust kitsendamata, et  $|C(x)| \leq |C(y)|$ ,

$$d(x, y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}} \approx \frac{|C(xy)| - |C(x)|}{|C(y)|}. \quad (8)$$

Heuristiliste meetrikate (7) ja (8) abil leitakse uuritava sõnade hulga sarnasusmaatriksi, mida kasutatakse otseselt järelduste tegemiseks (näiteks plaagiaatide tuvastamine) või sisendina teistele algoritmidele (näiteks klasterdamine).

### 8.1. Näiteid rakendustest

Informatsioonikaugus on osutunud võimsaks tööriistaks, mida on kasutatud probleemide lahendamisel väga erinevates valdkondades:

1. genoomide võrdlemine bioinformaatikas. Genoomide sarnasusmaatriksi leidmine informatsioonikauguse abil on täisautomaatne ning seejuures ei ole vaja geene eraldi tuvastada.

Artiklites [LCL<sup>+</sup>03, LBX<sup>+</sup>01] antakse ülevaade imetajate fülogeneesipuu tuletamisest informatsioonikauguse abil. Genoomi Kolmogorovi keerukuse hindamiseks kasutati algoritmi GENCompress, mis artikli [CKL00] põhjal annab parima tulemuse geeni-informatsiooni tihendamisel; puu konstrueerimisel kasutati naabriliite (*neighbour joining*) meetodit programmpaketis MOLPHY [AH96].

Ilmnes, et primaadid, sh inimene, on lähemal imetajate ülemseltsile *Ferungulata* (sõralised, kabjalised, londilised ja kiskjad) kui närilistele.<sup>5</sup>

Paul Vitányi kirjeldab, kuidas sama meetodiga klassifitseeriti koheselt viirus SARS (<http://homepages.cwi.nl/~paulv/papers/sarsvirii>).

2. autorsuse tuvastamine. Artiklis [CLMS02] kirjeldatakse lähteteksti plagaatide tuvastamise süsteemi SID, mis põhineb informatsioonikaugusel, ning demonstreeritakse, et levinumad plagiaadituvastusprogrammide petmise tehnikad ei mõjuta süsteemi SID tulemusi.

Artiklites [CVdW03, Mui03] tutvustatakse muusikateoste autorsuse ja žanri automaatse tuvastamise süsteemi, kus informatsioonikauguse leidmisel kasutatakse tihendusalgoritmi *bzip2*.

3. keelte suguluse määramine lingvistikas. Artiklis [BGL<sup>+</sup>98] antakse ülevaade informatsioonikauguse kasutamisest keelepüü automaatsel konstrueerimisel inimõiguste ülddeklaratsiooni tõlgetest 52 eri keelde. Ootuspäraselt paigutus inglise keel romaani keelte hulka, kuna teatavasti on selles hulgaliselt prantsuse laensõnu, ning ungari keel ei paigutunud soomeugri keelte hulka, kuna seda on ulatuslikult mõjutanud slaavi ja türgi keeled. Ülejäänud keeled paigutusid vastavalt üldlevinud lingvistilistele teooriatele. Informatsioonikauguse määramisel kasutati tihendusalgoritmi *gzip*, klassifikatsioonipüü koostamisel Fitch-Margoliashi meetodit [FM67] programmipaketis PHYLIP [AH96].

Vastukaja (näiteks Joshua Goodmani kriitiline kommentaar [Goo02]) tekitas Itaalia lingvistide töögrupi artikkel [BCL02], kus samuti kasutati tihendamisel põhinevaid meetodeid keelepüü tuletamiseks ning autorsuse tuvastamiseks. Kuigi töös viidati informatsioonikaugusele, kasutati *ad hoc* meetodeid, mis ei olnud teoreetiliselt põhjendatud ning töös defineeritud “kaugus” ei rahuldanud Frechet’ aksioome.

## 9. Kokkuvõte

Käesolevas artiklis näidati, et informatsioonikaugus on teoreetiliselt hästi põhjendatud universaalne meetrika, mille abil on võimalik lahendada probleeme väga erinevates valdkondades. Kuigi tihendamisel põhinevaid meetodeid on sõnade sarnasuse määramisel varemgi kasutatud, ei ole kõiki Frechet’ aksioome rahuldavat üldist meetrikat enne esitatud.

---

<sup>5</sup>bioloogias on üheselt lahendamata probleem, kas geeni-info põhjal tuleks imetajaid grupeerida (*(Primates, Rodentia), Ferungulata*) või (*(Primates, Ferungulata), Rodentia*).

Tegemist on uue teoreetilise lähenemisega, mille rakendusvaldkonnad kindlasti ei piirdu artiklis loetletutega. Informatsioonikaugusel põhinevaid süsteeme võib suvalises valdkonnas kasutada andmekaevandusautomaatidena, mis tuvastavad iseseisvalt seniavastamata sarnasusi andmelattu talletatud objektide vahel; seda on edukalt tehtud plagiaatide tuvastamisel, lingvistikas ja bioinformaatikas.

## Viited

- [AH96] Jun Adachi and Masami Hasegawa. Molphy: A computer program package for molecular phylogenetics, 1996.
- [BCL02] Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. Language trees and zipping. *Physical Review Letters*, 88(048702), 2002.
- [BGL<sup>+</sup>98] Charles H. Bennett, Peter Gacs, Ming Li, Paul M. B. Vitányi, and Wojciech H. Zurek. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, 1998.
- [CKL00] Xin Chen, Sam Kwong, and Ming Li. A compression algorithm for dna sequences and its applications in genome comparison. In *Proceedings of the fourth annual international conference on Computational molecular biology*, page 107. ACM Press, 2000.
- [CLMS02] Xin Chen, Ming Li, Brian Mckinnon, and Amit Seker. A theory of uncheatable program plagiarism detection and its practical implementation, May 2002. Käsikiri.
- [CVdW03] Rudi Cilibrasi, Paul M. B. Vitányi, and Ronald de Wolf. Algorithmic clustering of music, 2003.
- [FM67] W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155(760):279–84, Jan 1967.
- [Goo02] Joshua Goodman. Extended comment on language trees and zipping, 2002.
- [LBX<sup>+</sup>01] Ming Li, Jonathan H. Badger, Chen Xin, Sam Kwong, Paul Kearney, and Haoyong Zhang. An information based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17(2):149–154, 2001.

- [LCL<sup>+</sup>03] Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul M. B. Vitányi. The similarity metric. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 863–872. Society for Industrial and Applied Mathematics, 2003.
- [LV97] Ming Li and Paul M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, second edition, 1997.
- [Mui03] Hazel Muir. Software to unzip identity of unknown composers. *New Scientist*, April 2003.
- [Sör03] Kenneth Sörensen. Distance measures based on the edit distance for permutation-type representations. In *GECCO 2003: Proceedings of the Bird of a Feather Workshops, Genetic and Evolutionary Computation Conference*, pages 15–21, 2003.
- [Yia02] Peter N. Yianilos. Normalized forms for two common metrics. Technical report, NEC Research Institute, 1991,2002.