

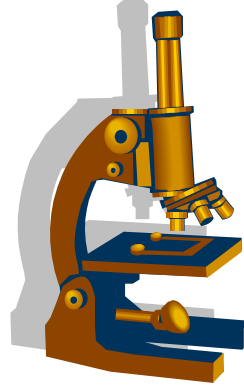
Porovnání komprimačných algoritmů a programů při zjišťování příbuznosti organismů

Josef Chludil

Josef Talaš

Motivace

- Otestovat použitelnost dalších komprimačních algoritmů na biologických datech



Metodologie

- Rovnice pro Kolmogorovskou složitost s aproximací (kromě algoritmu gencompress)
- „Relatedness“ approximation:

$$R(w, z) = \frac{\text{Compress}(w) - (\text{Compress}(zw) - \text{Compress}(z))}{\text{Compress}(wz)} =$$
$$\frac{\text{Compress}(w) + \text{Compress}(z) - \text{Compress}(zw)}{\text{Compress}(wz)}$$

Data

- ze cvičení ke komprimaci DNA a určování příbuznosti organismů
- A.dna až G.dna

Kompresní algoritmy

– GenCompress

– Program 7-Zip

- Zip – LZW algoritmus
- GZip – algoritmus DEFLATE, který je kombinací LZ77 a Huffmanova kódování, byl určený k nahrazení LZW
- BZip2 – BZip2 algoritmus
- 7z – metoda LZMA2

Výsledky - gencompress

	A	B	C	D	E	F	G
A	0,99527856	0,06549423	0	0	0,001659	0	0
B	0,06798866	0,99545178	0	0,000549	0,001382	0	0
C	-0,00062952	-0,00151607	0,9945340	0,087243	0,028753	0,446272	0,03934
D	-0,00031476	0	0,0880021	0,995885	0,031794	0,105263	0,045117
E	0,00157381	0,00060642	0,0300628	0,032099	0,995853	0,038925	0,064374
F	-0,00031476	0	0,4468434	0,106173	0,038983	0,995888	0,062724
G	-0,00031476	0	0,0404482	0,044993	0,063865	0,061678	0,995873

Výsledky – Zip

ZIP	A	B	C	D	E	F	G
A	0,91017678	0,914358264	0,921047076	0,92	0,919762	0,919941	0,919975
B	0,911371237	0,915512465	0,922102597	0,921053	0,920824	0,920997	0,92103
C	0,910654563	0,914819945	0,921469284	0,920421	0,920187	0,920363	0,920397
D	0,909221214	0,913434903	0,92020266	0,919158	0,918913	0,919096	0,91913
E	0,909698997	0,913896584	0,959890226	0,919579	0,919338	0,919518	0,919552
F	0,909460105	0,913665743	0,920413764	0,919368	0,919125	0,919307	0,919341
G	0,909460105	0,913665743	0,920413764	0,919368	0,919125	0,919307	0,919341

Výsledky – GZip

GZIP	A	B	C	D	E	F	G
A	0,90799508	0,912372358	0,919410745	0,918341	0,918083	0,918274	0,91831
B	0,909225092	0,913559725	0,920493934	0,919421	0,919172	0,919358	0,919393
C	0,908487085	0,912847305	0,919844021	0,918773	0,918519	0,918708	0,918743
D	0,90701107	0,911422465	0,918544194	0,917477	0,917211	0,917407	0,917443
E	0,907503075	0,911897412	0,91897747	0,917909	0,917647	0,917841	0,917876
F	0,907257073	0,911659938	0,918760832	0,917693	0,917429	0,917624	0,91766
G	0,907257073	0,911659938	0,918760832	0,917693	0,917429	0,917624	0,91766

Výsledky – BZip2

BZIP2	A	B	C	D	E	F	G
A	0,424521355	0,429364369	0,426783067	0,422945	0,424478	0,425153	0,423527
B	0,413291605	0,418442256	0,416735299	0,412976	0,414376	0,415085	0,413463
C	0,359351988	0,365980304	0,368473069	0,365092	0,36585	0,366727	0,365121
D	0,349963181	0,356848702	0,360072476	0,356758	0,357403	0,35831	0,356707
E	0,360272459	0,36687556	0,369296656	0,365909	0,366678	0,367552	0,365946
F	0,358799705	0,365443151	0,367978916	0,364602	0,365353	0,366232	0,364626
G	0,356774669	0,36347359	0,366167024	0,362804	0,363531	0,364417	0,362811

Výsledky – 7z

7z	A	B	C	D	E	F	G
A	0,968452679	0,969357977	0,97230977	0,972266	0,971979	0,971793	0,972185
B	0,968452679	0,969357977	0,97230977	0,972266	0,971979	0,971793	0,972185
C	0,969203806	0,970087549	0,972985142	0,972937	0,972652	0,972465	0,972858
D	0,96895343	0,969844358	0,972760018	0,972713	0,972428	0,972241	0,972633
E	0,968703055	0,969601167	0,972534894	0,972489	0,972204	0,972017	0,972409
F	0,968452679	0,969357977	0,97230977	0,972266	0,971979	0,971793	0,972185
G	0,96895343	0,969844358	0,972760018	0,972713	0,972428	0,972241	0,972633

Hodnocení

- kromě gencompress nepoužitelné výsledky (proč?)
- Standardní algoritmy nejsou na takto specifická data použitelné?
- Malá velikost vzorků? (15 kB kusy)
- Aproximace je příliš nepřesná?

That's all, folks!

- Nějaké otázky?