OP VVV CERIT-SC v den ▪.▪▪▪.▪▪

infrastruktura a výzkum domácí

Aleš Křenek, Tomáš Rebok, Barbora Bühnová, Jiří Filipovič, David Střelák, Jana Pazúriková, Mouzhi Ge, Hind Bangui, Jan Herman, Tomáš Raček, Lukáš Hejtmánek a další

Seminář gridového počítání, 11.5.2018

Projekt OP VVV

Projekt

- Masarykova univerzita
- trvání 5/2017 – 4/2021
- rozpočet 100+ mil. Kč

Inovace kapacit centra

- většina hardware z OP VaVpI (2011-14) morálně i fyzicky zastarává
- komplexní obnova, srovnatelná investice, vyšší výkon díky Moorovu zákonu (teoreticky 16×)
- SMP a GPU clustery 2017, HD 2018, viz `http://www.cerit-sc.cz/`

Tzv. In-house research – „výzkum domácí“

- systematický rozvoj expertizy potřebné k efektivnímu využití infrastruktury
- dva hlavní výzkumné směry – velká data a intenzivní výpočty

Velká data

Rozsáhlé úložné kapacity máme, užívejme je smysluplně

Velká data

Rozsáhlé úložné kapacity máme, užívejme je smysluplně

Analýza velkých dat

- „Big Data“ jsou všude
- *velká* rozsahem, požadovanou rychlostí a komplexitou zpracování
- ale i nedostatkem struktury, nespolehlivostí, vágností zadání, . . .

Cíle výzkumného programu

- vybudování obecných znalostí v oblasti zpracování a analýzy objemných dat
- aplikace v konkrétních oblastech, jejich kvalitativní posun s efektivním využitím infrastruktury

Velká data – aplikace

Kyberbezpečnost a kyberkriminalita

- iniciálně analýzy dat síťového provozu
- postupně i hledání komplexních souvislostí v heterogenních datech
- projekty
  - CopAS – vytvořený framework pro analýzy dat v prostředí ElasticSearch (Policie ČR)
  - ANALÝZA – Komplexní analýza a vizualizace heterogenních dat velkého rozsahu (MV ČR)
  - C4e – centrum excelence výzkumu kyberkriminality (OP VVV)

Velká data – aplikace

Energetika (spolupráce s MycroftMind + ČEZ)

- komplexní zpracování dat kritických infrastruktur (smartmetery)

Vědy o zemi (CzechGlobe)

- identifikace vadných senzorů v měřících datech

Smart Cities (Brno)

- konkretizace témat

Bioinformatika a strukturní biologie

Podobnosti v PDB

Databáze Protein Data Bank

- 3D struktury proteinů + další informace
- 140+ tis. záznamů, 500+ tis. řetězců
- Jak moc jsou si podobné?

Podobnosti v PDB

Posouzení podobnosti

- 3D struktura se „příliš neliší“
- různé metriky
- používáme GESAMT Q-score

$$Q = \frac{N_{\text{align}}^2}{(1 + MSD(A,B)/R_0^2)N_A N_B}$$

  - vyvažuje průměrnou vzdálenost α uhlíku a počet
  - rozsah 0 – kompletně jiné, 1 – identické
- srovnání dvou struktur – 0.01–1 s, naivní prohledání hodiny

Podobnosti v PDB

Abstraktní formulace problému

- 10^4–10^6 bodů v mnohorozměrném prostoru
- nejsou absolutně umístěny (volnost rotace a posunutí) – nelze přímo srovnat
- umíme posoudit podobnost $\sim$ vzdálenost dvojice
- **pro nový bod hledáme nejbližší**

Strukturní biologie $\longrightarrow$ informatika – velká data

Podobnosti v PDB

Testovací data

- 65 tis. řetězců od správců PDB (splehlivé, unikátní)
- filtr na artefakty (havaruje výpočet Q-score)

Orientační analýza vlastností

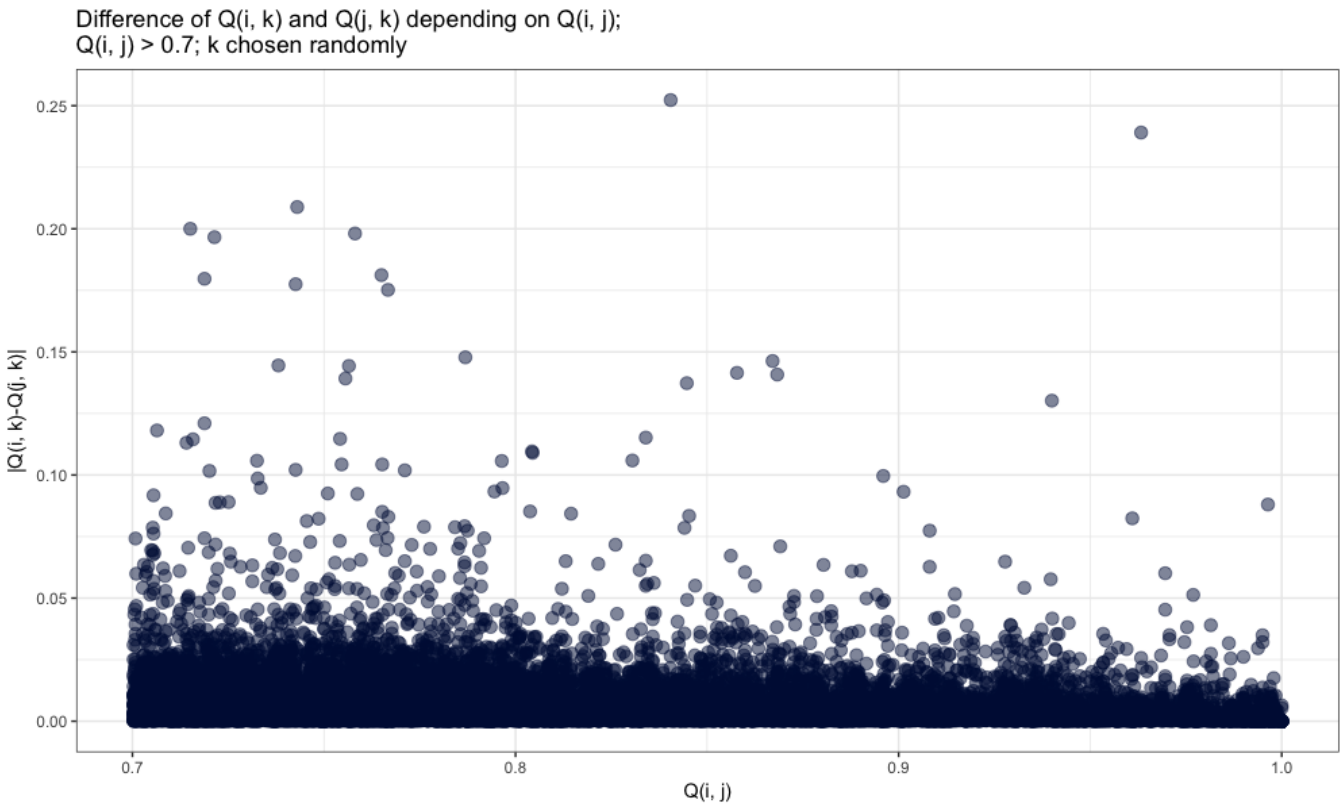
- metriky, pokrytí prostoru, četnosti vzdáleností, ...

Prohledávací algoritmus

- nemusí být zcela přesný
- řádově rychlejší
- používá předpočítané datové struktury

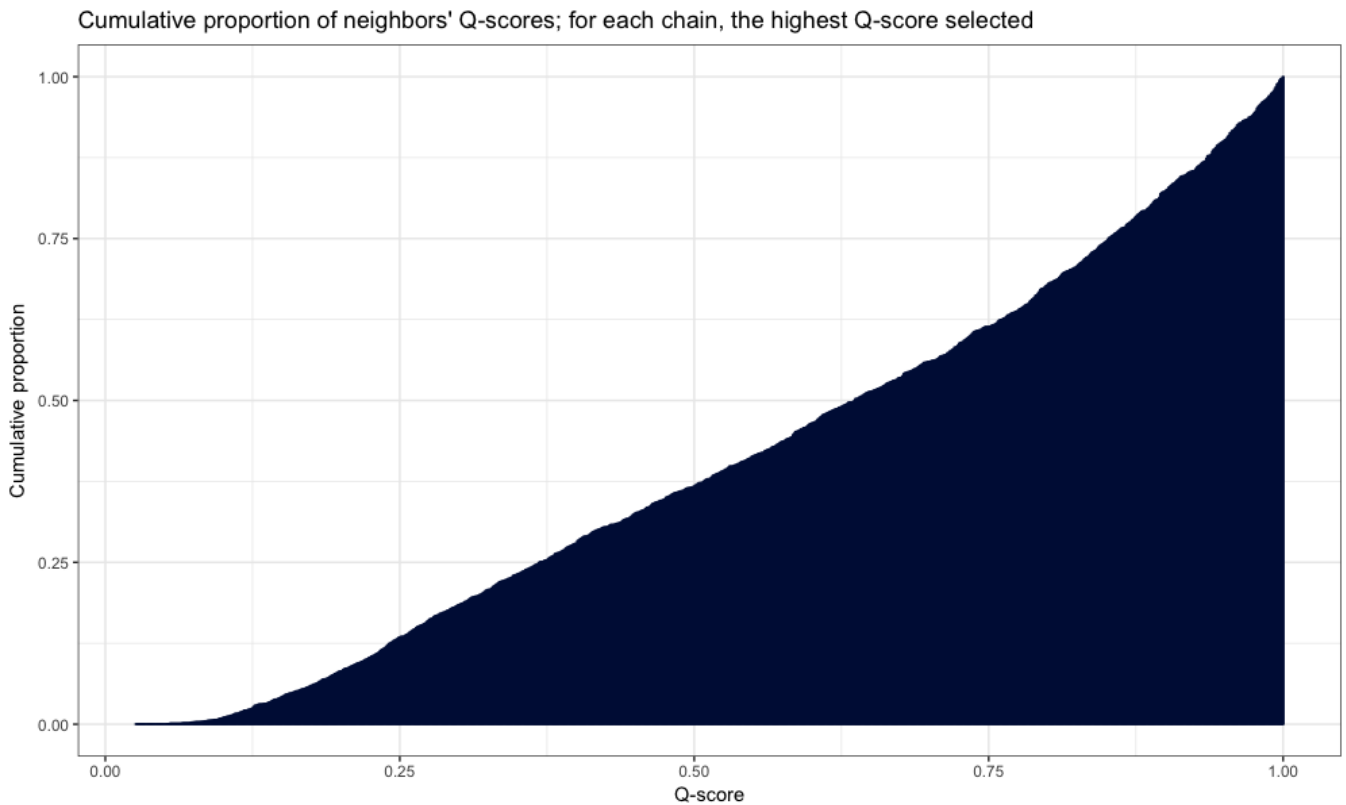
Podobnosti v PDB

Je Q-score (resp $1/Q - 1$) metrika?



Podobnosti v PDB

Je prostor dostatečně pokryt?



Podobnosti v PDB

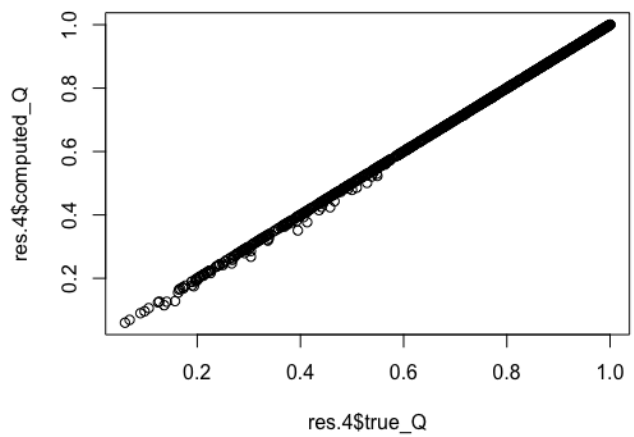
Algoritmus

1. Předpočítání Q-score všech dvojic z referenční sady R
2. Pro novou strukturu t vyber k náhodných kandidátů $r_i \in R$, spočítej $q_i = Q(t, r_i)$
3. Pro $q_i \geq \theta$ hledej v $\{r_j \in R\colon Q(r_j, r_i) > \theta\}$
4. Pro $q_i < \theta$ hledej v $\{r_j \in R\colon q_i - \epsilon \leq Q(r_j, r_i) \leq q_i + \epsilon\}$

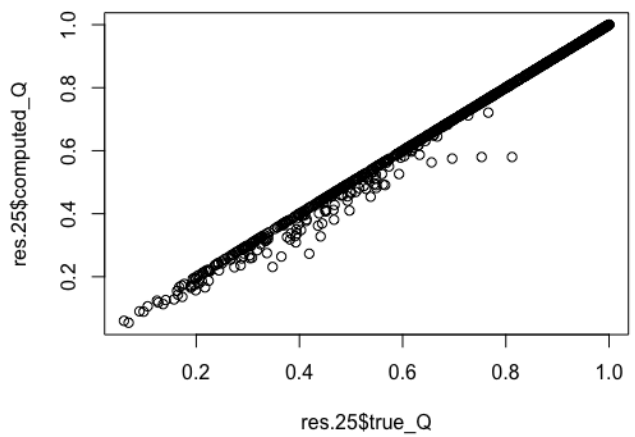
Volbou počtu k, prahu θ a tolerance ϵ ovlivňujeme přesnost a rychlost algoritmu

Podobnosti v PDB

Předběžné výsledky (malá sada $|R| = 14000$)



$\epsilon = 0.4$, 6×



$\epsilon = 0.25$, 13×

Intenzivní výpočty

Neplýtvejme výkonem našich strojů

cerit scientific cloud

Intenzivní výpočty

Neplýtvejme výkonem našich strojů

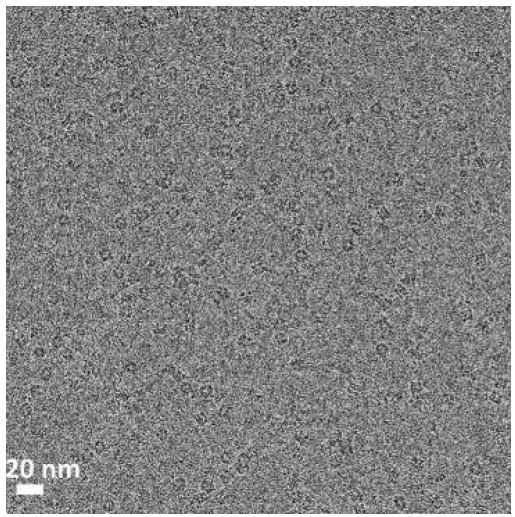
Optimalizace kódu ze dvou pohledů

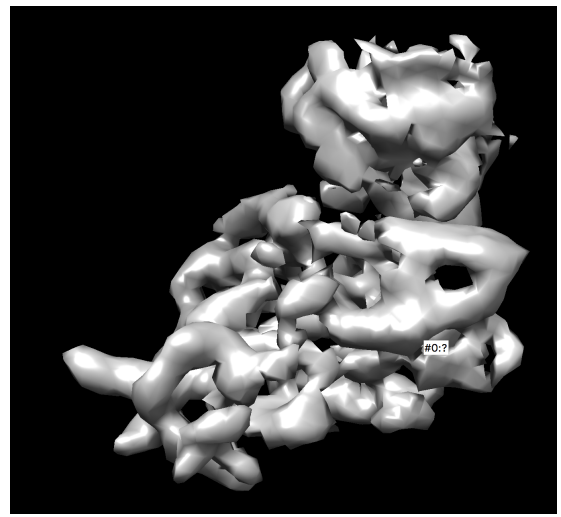
- natvrdo – zachováme postup výpočtu
  - snažíme se maximálně využít kapacitu hardware
  - **autotuning** – více alternativních implementací, za běhu se vybírá nejlepší kombinace
- naměkko – měníme postup výpočtu i numerický model
  - vyloučíme „zbytečné“ výpočty
  - změny implementace dovolí agresivnější „tvrdé“ optimalizace
  - výsledky jsou jiné, ale ohlídáme, že to nevadí

Rekonstrukce 3D v CryoEM

Single Particle Analysis

- rychlé zmrazením vzorku
  - amorfní led
  - biomolekuly v nativním stavu
- náhodná konformace a natočení v 3D
- nízký odstup signálu od šumu





Rekonstrukce 3D v CryoEM

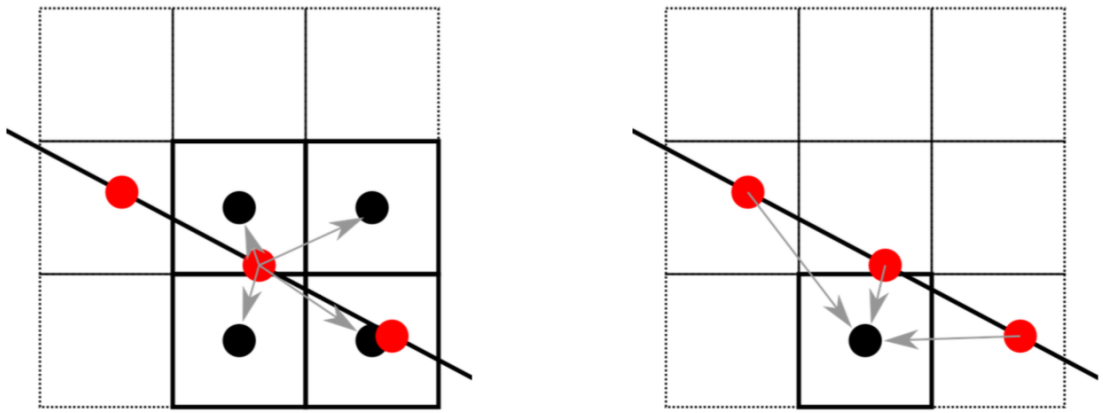
Komplexní postup zpracování

- korekce deformace a pohybu vzorku
- klasifikace ve 2D
- zarovnání a průměrování (potlačení šumu)
- **rekonstrukce 3D**

Rekonstrukce 3D v CryoEM

Rekonstrukce 3D

- projekce snímku (2D) do prostoru (3D)
- **interpolace pixelů snímku do 3D mřížky** – *scater* vs. *gather*



- přímočarý postup *scatter*
  - příliš mnoho zápisů
  - konflikty při jemnozrnné paralelizaci (GPU)
- Fourierova transformace, hledání nejlepšího zarovnání, . . .

Rekonstrukce 3D v CryoEM

Výsledky

- GPU implementace scatter – zrychlení 2.85× proti CPU
  - NVidia P100 vs. 24 jader Intel Xeon E5-2650 v4
- GPU implementace gather – zrychlení 11.4×
- škáluje na 4× GPU – zrychlení 31.7×
- nedává stejné výsledky, rozdíly jsou zanedbatelné

Poděkování

Jmenovaní i utajení členové týmu a studenti

Chlebodárci

- OP VVV CERIT-SC, CZ.02.1.01/0.0/0.0/16_013/0001802
- H2020 West-life, 675858

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání