Workshop des WissGrid Fachberater-Teams

Virtuelle Forschungsumgebungen aufbauen - mit D-Grid -

Communities mit datenlastigen Aufgaben
19. Januar 2011, Göttingen

Benjamin Löhnhardt

Vortragsinhalte

- Einleitung
- Datenlastige Anwendungen aus der Biomedizin
  - Imputation mit MACH
  - Diffusion Tensor Fiber Tractography mit FSL
- Fazit

Einleitung

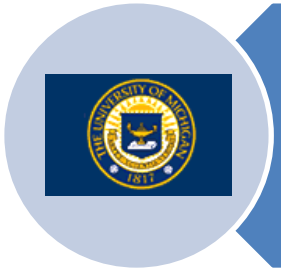
- Beschreibung von Communities und deren Anwendungen aus dem **biomedizinischen** Bereich.
- Praxisbeispiele der **Universitätsmedizin Göttingen:**
  - Abteilung Klinische Neurophysiologie
  - Abteilung Genetische Epidemiologie

UNIVERSITÄTSMEDIZIN GÖTTINGEN UMG

Einleitung

- Situation:
  - Anwendungen mit **datenlastigen Berechnungen** im biomedizinischen Bereich sind vorhanden.
  - Forschung wird durch **fehlende Ressourcen** limitiert:
    - Knappe Kapazitäten innerhalb der Forschungsinstitute.
    - Die Beschaffung von dedizierten institutsinternen Rechenclustern oft nicht sinnvoll/wirtschaftlich.
- Ziel:
  - Nutzung von Ressourcen innerhalb **virtueller Forschungsinfrastrukturen**.

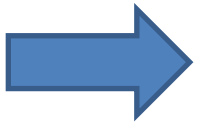
Datenlastige Anwendungen aus der Biomedizin



Imputation mit MACH

Diffusion Tensor Fiber Tractography mit FSL

Datenlastige Anwendungen aus der Biomedizin



Imputation mit MACH

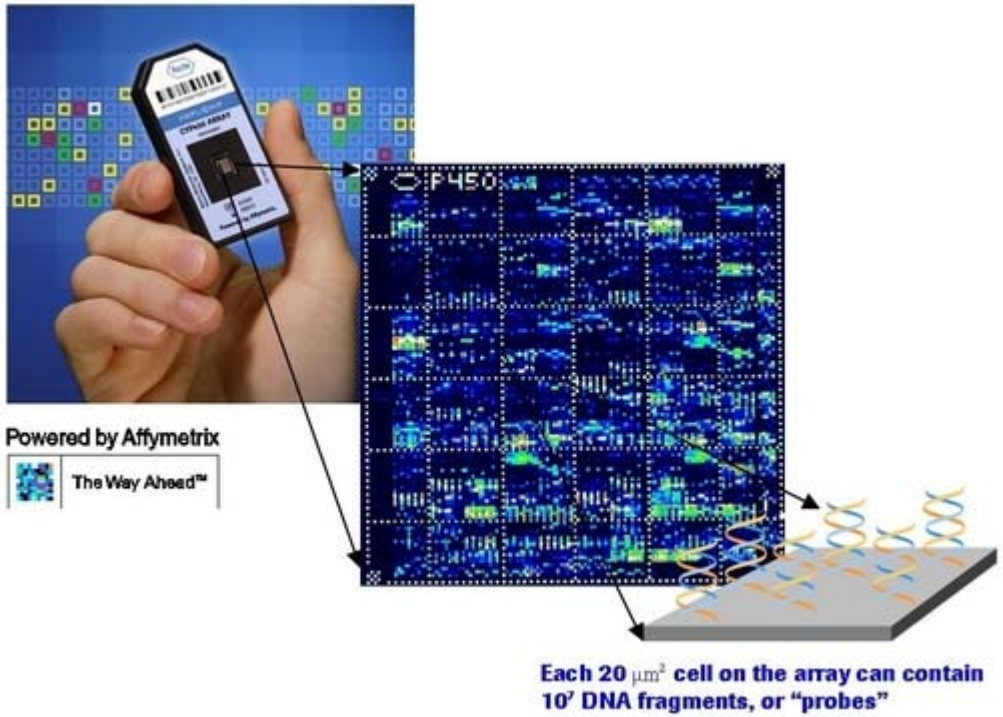
Diffusion Tensor Fiber Tractography mit FSL

Imputation mit MACH

- Statistisches Verfahren, um nicht genotypisierte Marker in Studiendaten anhand eines Referenzsamples zu schätzen.
- Individuen können auf unterschiedlichen Chips (z.B. Affymetrix 500K oder 6.0) genotypisiert sein.
- Referenz-Sample: z.B. HapMap2/3, 1000 Genome Project
- Imputationssoftware: z.B. MACH (Markov Chain based haplotyper)
- Chromosomenweise (1-22) werden 2 Phasen durchlaufen.

Imputation mit MACH

- Genotypisierung auf einem Affymetrix 6.0-Chip:



http://www.bio-pro.de/

Imputation mit MACH

- **Studiendaten**
  - Je Chromosom (1-22) existiert eine Datei.
  - Individuen werden in den Zeilen dargestellt.
  - Pro SNP (in Spalten) gibt es zwei Einträge:
    - A,C,T oder G für die vier verschiedenen Basen bzw. 0, wenn die Base unbekannt ist.
    - Pseudonymisierte Patienten-ID.
- **HapMap**
  - Je Chromosom (1-22) existiert eine Datei.
  - Haplotypen werden in den Zeilen dargestellt.
  - Pro SNP (in Spalten) gibt es einen Eintrag:
    - Enthält 0 oder 1.

Imputation mit MACH

Eine konkrete (eher kleine → bis 2000 Individuen) Studie:

Sample 1

Enthält 129 Personen (=Zeilen).
Enthält 415 000 SNPs (=Spalte) auf Affimetrix 500k-Chip.
Je nach Chromosom bis zu 18 MB.
Gesamtgröße: ~210 MB.

Sample 2

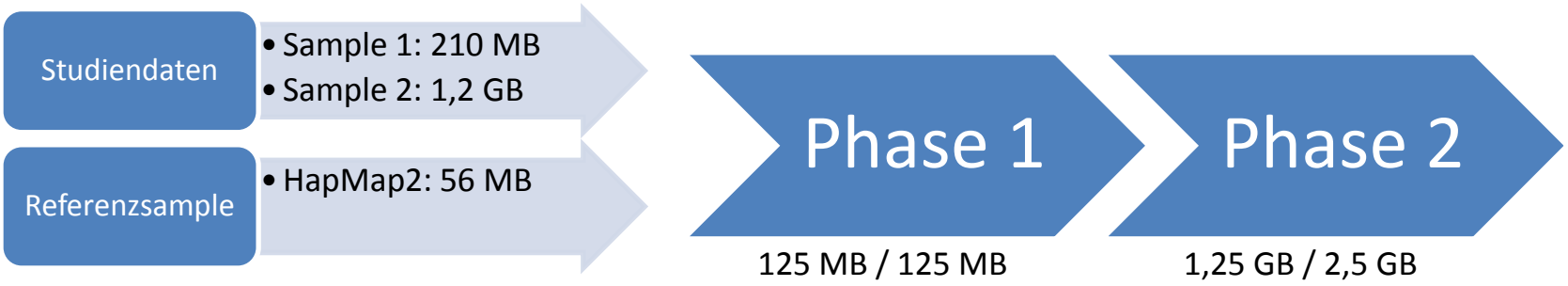
Enthält 432 Personen (=Zeilen).
Enthält 708 000 SNPs (=Spalte) auf Affimetrix 6.0-Chip .
Je nach Chromosom bis zu 102 MB.
Gesamtgröße: ~1,2 GB.

HapMap2

Enthält 120 Haplotypen (=Zeilen).
Enthält 2 500 000 SNPs (=Spalte).
Je nach Chromosom bis zu 4 MB (gepackt).
Gesamtgröße: ~56 MB (gepackt).

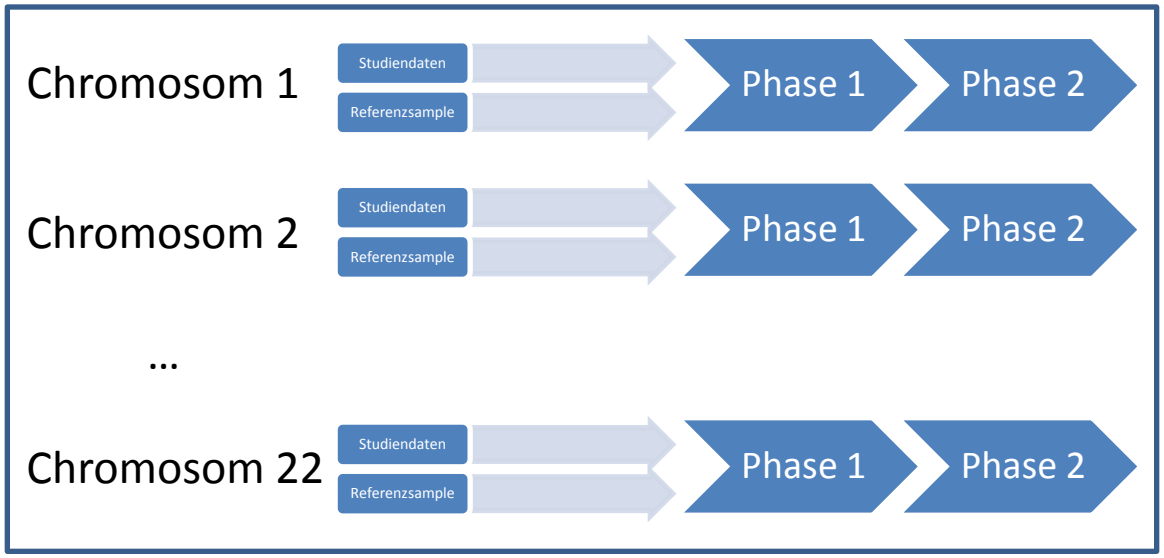
Imputation mit MACH

- Daten nach **Phase 1**:
  - Je nach Chromosom bis zu 11 MB.
  - Gesamtgröße: ~125 MB.
- Daten nach **Phase 2**:
  - Je nach Chromosom/Sample bis zu 100 MB/200 MB.
  - Gesamtgröße: ~1,25 GB bzw. ~2,5 GB.



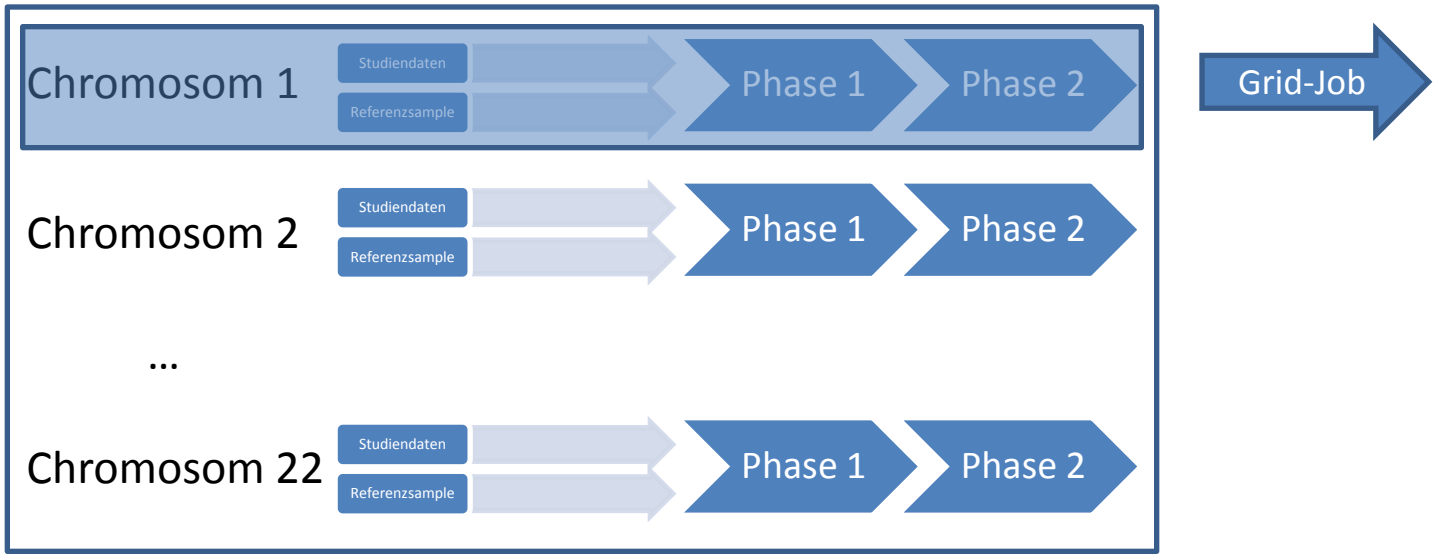
Imputation mit MACH

- Wie kann von einer VRE bzw. Grid profitiert werden?
  - Durch **Aufteilung des Gesamtjobs** können Berechnungsschritte zeitgleich bearbeitet werden.
  - Hier im Beispiel: **Unabhängige Berechnungen** auf Basis der 22 Chromosomen.



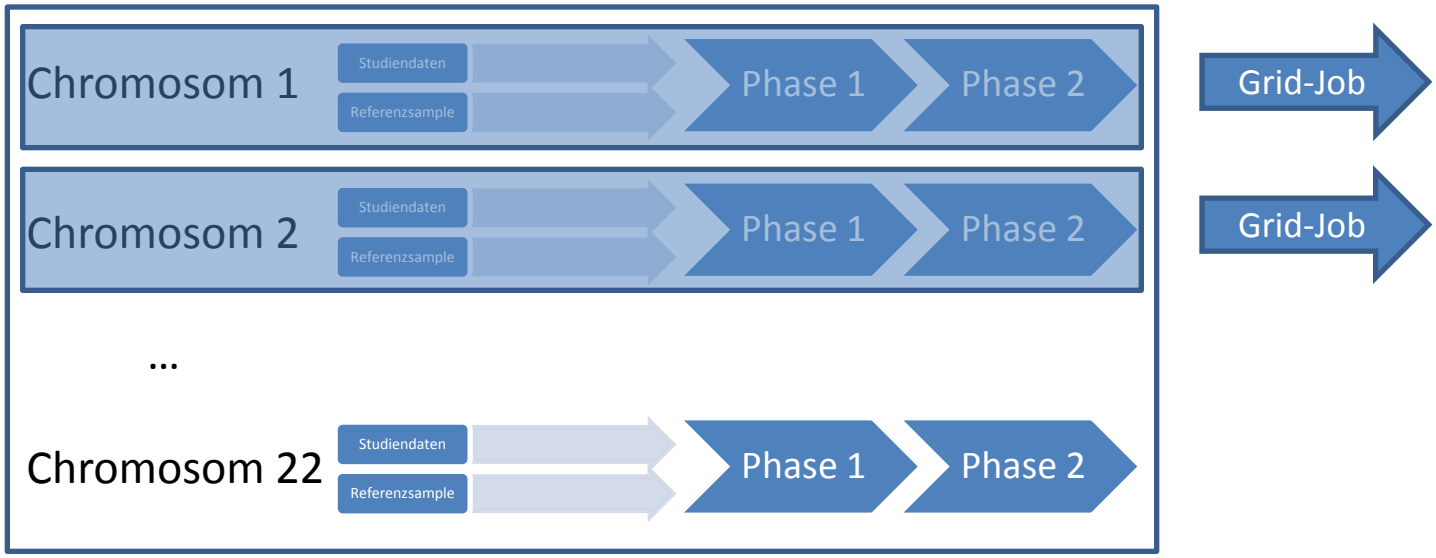
Imputation mit MACH

- Wie kann von einer VRE bzw. Grid profitiert werden?
  - Durch **Aufteilung des Gesamtjobs** können Berechnungsschritte zeitgleich bearbeitet werden.
  - Hier im Beispiel: **Unabhängige Berechnungen** auf Basis der 22 Chromosomen.



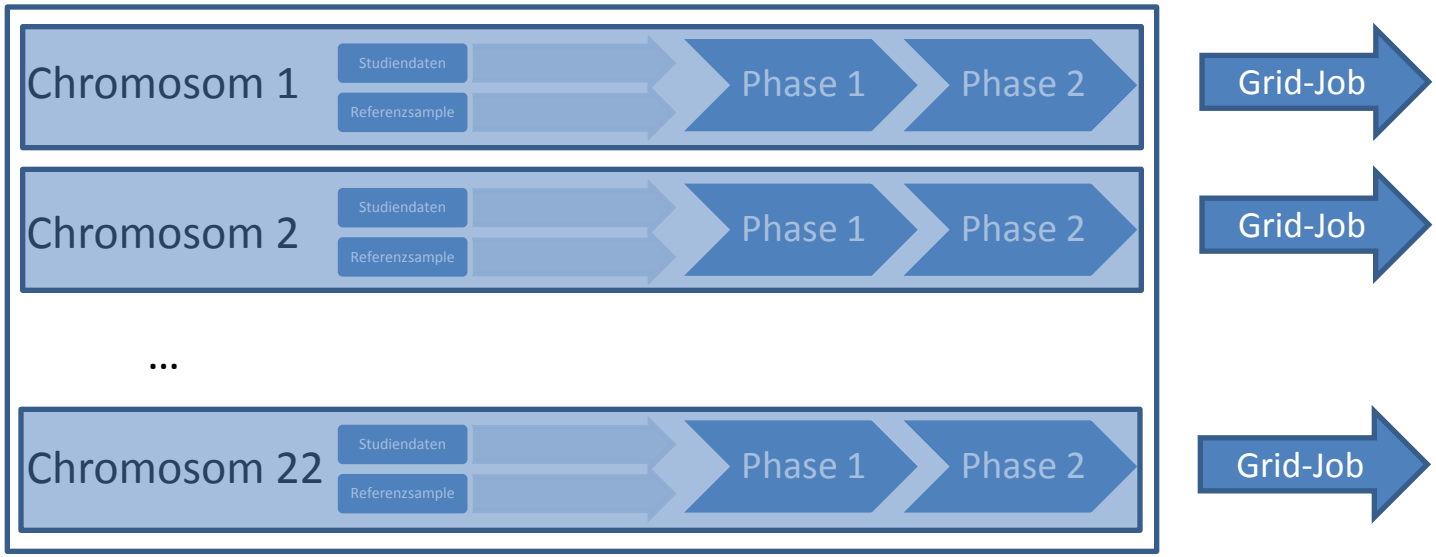
Imputation mit MACH

- Wie kann von einer VRE bzw. Grid profitiert werden?
  - Durch **Aufteilung des Gesamtjobs** können Berechnungsschritte zeitgleich bearbeitet werden.
  - Hier im Beispiel: **Unabhängige Berechnungen** auf Basis der 22 Chromosomen.



Imputation mit MACH

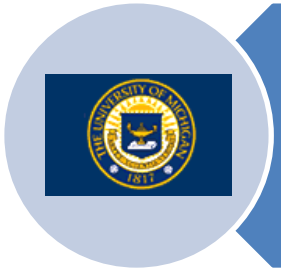
- Wie kann von einer VRE bzw. Grid profitiert werden?
  - Durch **Aufteilung des Gesamtjobs** können Berechnungsschritte zeitgleich bearbeitet werden.
  - Hier im Beispiel: **Unabhängige Berechnungen** auf Basis der 22 Chromosomen.



Imputation mit MACH

- Aktueller Status
  - Aufteilung der Gesamtberechnung in 22 Berechnungen (je Chromosom eine Teilberechnung).
  - Verwendung eines Grid-Knotens für die Berechnungen.
  - Laufzeiteinsparung um mehr als 90 % [1] möglich.
- Ziel: Integration in D-Grid, dazu notwendig…
  - Installation von MACH auf MediGRID-Ressourcen.
  - Erstellung eines Grid-Workflows.

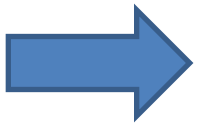
Datenlastige Anwendungen aus der Biomedizin



Imputation mit MACH

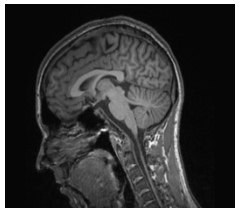
Diffusion Tensor Fiber Tractography mit FSL

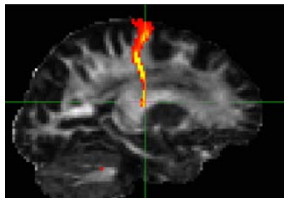
Datenlastige Anwendungen aus der Biomedizin



Diffusion Tensor Fiber Tractography mit FSL

- Anwendung der medizinischen Bildgebung.
- Berechnung von Pfaden der Konnektivitätsverteilung von Nervenfasern im Gehirn (Fibre-Tracking).
- Als Basis dient ein MRT-Datensatz (Magnetresonanz-tomographie).

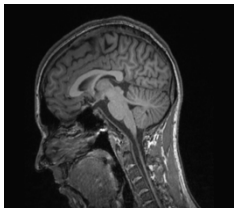


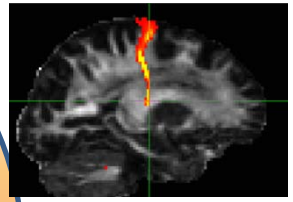


MRT durchführen → MRT-Datensatz → Berechnung → Ergebnis

Diffusion Tensor Fiber Tractography mit FSL

- Anwendung der medizinischen Bildgebung.
- Berechnung von Pfaden der Konnektivitätsverteilung von Nervenfasern im Gehirn (Fibre-Tracking).
- Als Basis dient ein MRT-Datensatz (Magnetresonanz-tomographie).



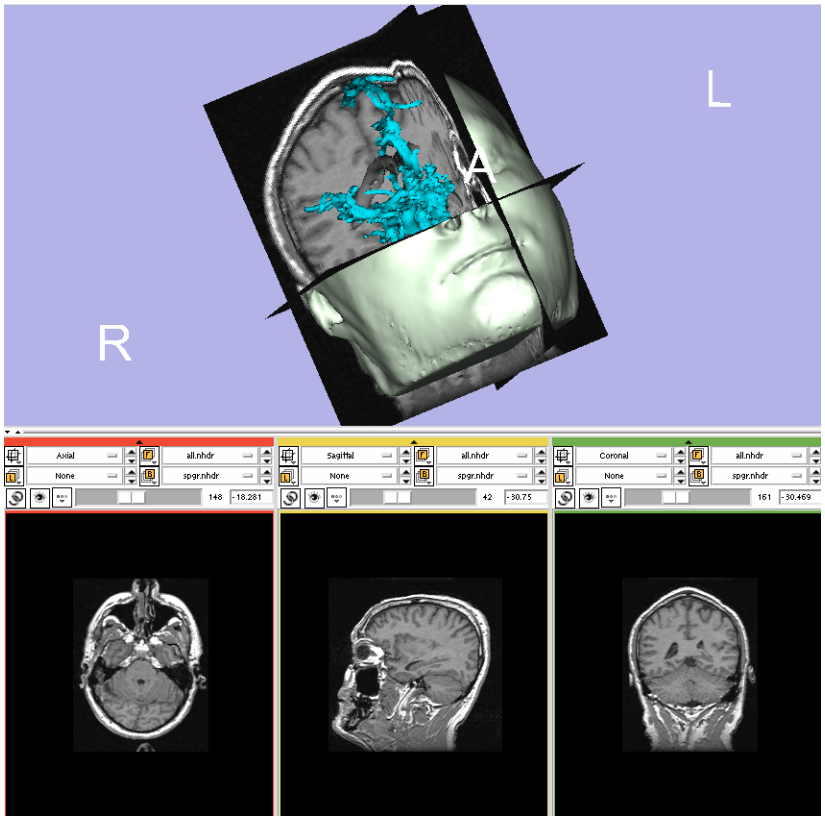


MRT durchführen → MRT-Datensatz → Berechnung → Ergebnis

Diffusion Tensor Fiber Tractography mit FSL

Inputdaten → FSLSplitVolume → FSLBedpost → FSLMergeSlides → FSLProbtrackx

- **Inputdaten:** MRT-Daten eines Gehirns (50 MB im Nifti-Format).



Screenshot der Anwendung 3D-Slicer

Diffusion Tensor Fiber Tractography mit FSL

- **FSLSplitVolume:** Aufteilen der MRT-Daten in einzelne Slices (73 Slices mit jeweils ca. 550 – 700 KB).
- **FSLBedpost:**
  - Ermittlung der Diffusionstensoren (3x3-Matrix) von Nervenfasern pro Voxel (Volumenpunkt).
  - Inputdateien: je 1 Slice und zusätzliche Informationen.
  - Notwendiger Vorschritt für Probtrackx.
- **FSLMergeSlides:** Zusammenführen der Slices (gepackt ca. 250 MB, ungepackt > 1 GB im Nifti-Format).

Diffusion Tensor Fiber Tractography mit FSL

- **FSLProbtrackx:**
  - Daten nach dem Bedpost-Schritt (~250 MB) als Input.
  - Pfade der Konnektivitätsverteilung werden pro Voxel berechnet (bspw. 128x128x64 Voxel → ~1 Mio.).
  - Pro Voxel werden Daten der Größe 4-5 KB produziert.
  - Eine Schichtenberechnung dauert ca. 10 – 20 Sekunden (Außenbereich) bzw. 1 Minute (im Zentrum).
  - Anschließend werden einzelne Bilder mit gefundenen Pfaden wieder zusammengeführt.

Diffusion Tensor Fiber Tractography mit FSL

- Wie kann von einer VRE bzw. Grid profitiert werden?
  - Die **Split- und Merge-Schritte** bieten wenig Möglichkeit zur Aufteilung. Die Laufzeit ist zudem sehr gering (z.B. FSLSplitVolume: ~1 min).
  - Bei **FSLBedpost** kann der Gesamtjob bspw. in Bezug auf die einzelnen Slices aufgeteilt werden (1 Grid-Job pro Slide), da die Berechnungen pro Slice unabhängig sind.
  - Bei **FSLProbtrackx** kann bspw. in Bezug auf die Startvoxel (z.B. 1 Mio. Stück) parallelisiert werden), da die Berechnungen pro Voxel unabhängig sind.

Diffusion Tensor Fiber Tractography mit FSL

- Besonderheiten bei der Gridifizierung von **FSLProbtrackx**:
  - Für jeden Job sind große Inputdateien (~250 MB) nötig.
  - Die Berechnung pro Voxel ist kurz (~10 s – 1 min).

  ➔Berechnung der Pfade eines Startvoxels pro Grid-Job ist nicht optimal.

  ➔Mehrere Startvoxel sollten zu einem Grid-Job zusammengefasst werden (z.B. 5^3-Voxelblöcke = 125).

  ➔Die Jobs sollten nicht allzu verteilt werden, um den Datentransfer zu minimieren.

Diffusion Tensor Fiber Tractography mit FSL

- Aktueller Status
  - Bedpost-Teil ist in MediGRID umgesetzt.
  - Bearbeitung in einer laufenden Masterarbeit in Zusammenarbeit mit der Klinischen Neurophysiologie:
    - Erstellung eines Grid-Workflows zur Berechnung in D-Grid.
    - Software zur intuitiven Verwendung des Workflows und Grid-Nutzerzertifikate über eine Benutzeroberfläche.
  - Laufzeitvergleich der Berechnung eines Datensatzes:
    - Standardarbeitsplatzrechner: > 1 Monat
    - Grid-Anwendung: ~24 h auf GWDG-Ressourcen (erwartet)
- Ziel: Nutzung der D-Grid-/MediGRID-Infrastruktur für die Berechnungen und Integration der Anwendung in das MediGRID-Portal.

Fazit

- **Erfahrungen** bei der Integration von datenlastigen Anwendung in eine VRE:
  - Der Datentransfer muss in die Workflowplanung einbezogen werden (Datentransfers minimieren), bspw. durch...
    - die Bündelung von Jobs.
    - die Ausführung der Jobs auf dem Knoten, auf dem die Daten liegen.
  - Der Aufwand bei der Verarbeitung großer/vieler Daten darf nicht unterschätzt werden (z.B. 128 x 128 x 64 Voxel → ~1 Mio.).
  - Quotas (HDD / RAM) auf den Grid-Knoten beachten.
  - Datenschutz: setzt ein mit dem Datenschutzbeauftragten abgestimmtes Datenschutzkonzept voraus.

Fazit

- **Vorteile für den Nutzer** durch die Integration der in eine virtuelle Forschungsumgebung:
  - ➔Laufzeiteinsparung aufgrund von Hochdurchsatz durch Aufteilung.
  - ➔Umfangreiche Berechnungen werden ermöglicht.
  - ➔Kein Aufbau eigener Ressourcen notwendig.

Quellen

[1] Löhnhardt B, Quade M, Skrowny D, Sohns M, Bickeböller H, Sax U: Hochleistungsrechencluster zur Unterstützung der biomedizinischen Forschung. In: 55. GMDS-Jahrestagung 2010, Mannheim, 2010.

Kontaktmöglichkeiten

UNIVERSITÄTSMEDIZIN GÖTTINGEN : UMG

Universitätsmedizin Göttingen

Medizinische Informatik

http://www.mi.med.uni-goettingen.de/

Benjamin Löhnhardt

Computational Medicine und Grid-Computing

Benjamin.Loehnhardt@med.uni-goettingen.de

Tel.: (0551) 39 - 22842