Matthias Richter

Iowa City
3 Monate im Jacob Michaelson Laboratory

(Yes that's actually what it looks like...some places)

Nach 2 wunderbaren Jahren Cognitive Science Studium in Osnabrück entschied ich mich, ein Auslandspraktikum in der Forschung anzustreben, um einen Einblick in den Arbeitsalltag in einem Labor zu bekommen. Da ein Auslandsaufenthalt von mindestens 3 Monaten in meinem Studiengang verpflichtend ist, versorgt unsere Fakultät uns regelmäßig mit ausführlichen Informationen diesbezüglich. So habe ich auch von dem RISE Programm erfahren. Als ich den Praktikumsplatz im "Jacob Michaelson Lab for Computational Genomics and Psychiatry" sah wusste ich sofort, dass das meine erste Wahl sein würde. Die Verschmelzung von medizinischer Forschung mit Machine Learning interessiert mich schon seit langem.

Organisatorische Vorbereitung

Der erste Schritt der Organisation war natürlich die Kontaktaufnahme mit Dr. Jacob Michaelson (der sich direkt als Jake vorstellte). Ein paar E-Mails später hatten wir uns auf einen Praktikumszeitraum geeinigt. Außerdem lud Jake mich nach der Zusage zu einem virtuellen Meeting mit dem ganzen Labor ein. Bei diesem wurde ich per Skype zu einem der Labmeetings dazu geschaltet, konnte mich bei allen Mitgliedern vorstellen und habe von allen kurz erfahren wer sie sind und woran sie arbeiten. Das war ein sehr positiver Start für mich und ich habe mich direkt willkommen gefühlt.

Flug

Sobald mein Praktikumszeitraum feststand machte ich mich an das Buchen der Flüge, da die Bewerbung für ein Visum in den USA deutlich leichter mit bereits existierenden Flugdaten ist. Ich entschied mich für einen Hinflug von Amsterdam über Chicago nach Cedar Rapids, der nächste Flughafen von Iowa City aus. Mein Rückflug buchte ich direkt von Chicago aus um nach meinem Praktikum diese Stadt noch für ein paar Tage erkunden zu können.

Visum

Nach einiger Recherche fand ich heraus, dass ich für meinen Aufenthalt ein sogenanntes J1 Visum brauchte. Um dieses zu bekommen benötigt man das Formular DS-2019 von der Universität oder Institution, die das Praktikum anbietet. Nach Korrespondenz mit dem International Office der University of Iowa bekam ich Zugang zu einem Internet Portal ,über das ich alle nötigen Daten eingeben konnte. Einige Zeit später lag dann besagtes Formular per Express Post direkt aus Iowa bei mir im Briefkasten. Mit diesem konnte ich mich nun offiziell für das Visum bewerben, alle nötigen Gebühren zahlen (etwas über 300€) und einen Termin im Konsulat in Berlin ausmachen. Den hatte ich schon 3 Tage nach meiner offiziellen Bewerbung. Das Interview dort war ziemlich unspektakulär und ca. eine Woche später bekam ich meinen Reisepass mit Visum zurück.

Unterkunft

Nachdem feststand, wann genau ich im Lab sein würde, schrieb Jake eine Mail an alle Labmitglieder und fragte, ob jemand eine Idee bezüglich einer Unterkunft für mich hätte. Daraufhin meldete sich Tanner, ein PhD student aus dem Lab und meinte, er hätte gerade ein Haus mit seiner Frau gekauft in dem sie ein Zimmer frei hätten, welches sie mir vermieten könnten. Nach einem sehr netten Skype Gespräch hatte ich so mit sehr wenig Aufwand und Suche eine Zusage für eine Unterkunft. So viel Glück muss man erstmal haben!

Fachliche Vorbereitung

Das “Michaelson Lab for Computational Psychiatry and Genomics” benutzt Methoden aus dem Bereich der künstlichen Intelligenz um der genetischen Basis psychiatrischer Erkrankungen auf den Grund zu gehen. Da der Schwerpunkt meines Studiums auf Neurowissenschaften, Psychologie und künstlicher Intelligenz liegt, fühlte ich mich für den psychiatrischen und vor allem den genetischen Aspekt der Forschung fachlich schlecht vorbereitet

Auch hier erhielt ich Hilfe von Jake und anderen aus dem Lab. Ich bekam Zugang zu der laborinternen Website, auf der viel Material zu allen relevanten Themen zu finden ist. Dort arbeitete ich mich durch Grundlagen zum Thema Genetik . Außerdem bekam ich Zugangsdaten zu einem Datacamp Account. Dort machte ich Kurse zu der Programmiersprache R, mit der das Lab hauptsächlich arbeitet.

Leben in Iowa City

Wohnen

Ich wohnte, wie oben erwähnt, bei Tanner (PhD aus dem Lab) und seiner Frau Kelsey. Mit den beiden hätte ich es nicht besser treffen können. Schon vor meiner Ankunft waren Sie sehr engagiert, dass alles klappt und haben mich auch direkt in Cedar Rapids mitten in der Nacht vom Flughafen abgeholt, als ich dort mit 5 Stunden Verspätung nach verpasstem

Anschlussflug in Chicago ankam. Tanner hatte sich am nächsten Tag extra frei genommen und die beiden haben mir die Stadt gezeigt, viel erklärt und mir geholfen, mich zurecht zu finden.

Ich hatte ein ziemlich großes Zimmer und mein eigenes Bad im Keller, den ich praktisch für mich alleine hatte. Neben Tanner und Kelsey lebten in dem Haus noch ihre beiden Hunde Ivy und Baylie und einige Wochen nach mir zog Kelseys Bruder Kaleb ebenfalls bei uns ein. So hatten wir eine kleine WG und es war fast immer jemand zu Hause, mit dem man kochen, sich unterhalten oder spielen konnte.

Die Stadt

Iowa City ist mit 75.000 Einwohnern verhältnismäßig überschaubar. Allerdings hat die University of Iowa ca 35.000 Studenten, die in den Semesterzeiten nochmal dazu kommen. Somit ist Iowa City eine echte College Town und das merkt man auch. In dem eher konservativen Staat Iowa ist die Stadt eine kleine liberale Insel. Fast überall trifft man auf Studenten, in Cafes hört man regelmäßig Leute mit enthusiasmus über spannende Themen fachsimpeln, es gibt unzählige kleine und große Bars und natürlich auch ein College Football Team, dem so ziemlich alles in der Stadt gewidmet ist. Die Innenstadt ist praktisch gleich zu setzen mit dem Hauptcampus. Überall sind Unigebäude. Es gibt noch einen zweiten Campus, der alle Health Sciences und die große Uniklinik beherbergt. Die Stadt ist mit ihren vielen Wiesen und großen Bäumen ziemlich grün, was mich sehr gefreut hat.

Ich habe mir an meinem ersten Tag direkt ein Fahrrad gekauft (unbedingt zur “Bike Library” gehen! Sehr cooler Laden). Zum Glück ist Iowa City im Vergleich zu anderen Städten in den USA ziemlich fahrradfreundlich. Es gibt zwar auch hier kaum dedizierte Fahrradwege, aber viele Fahrradständer und einige Stationen, an denen man sein Fahrrad aufpumpen und kleine Reparaturen durchführen kann. Außerdem gibt es eine eingeschweißte Biker Community. Nicht selten kam es vor, dass Leute mir zugewunken oder mich angesprochen haben, nur weil ich auf einem Fahrrad saß.

Sozialleben

Nachdem ich mich Anfangs sehr stark auf mein Praktikum konzentrierte, fing ich nach einigen Wochen an, aktiv nach Kontakten zu suchen. Über meine Lab Kollegen, Mitbewohner und auch über die von RISE erstellte Liste von anderen Stipendiaten in Iowa City fand ich schnell Anschluss. Die Stadt bietet außerdem gemessen an der Größe eine üppige Auswahl an kulturellen Angeboten wie Lesungen, Poetry Slams, Comedy Shows, Improv und Konzerte. Auch in den Bars war immer was los und oft spielten dort lokale Bands Livemusik.

Forschen im Michaelson Lab

Am ersten Tag im Labor setzte ich mich mit Jake zusammen und wir besprachen den Verlauf des Praktikums. Er schlug einige Projekte vor, lies mir aber viel Freiheit. Gemeinsam einigten wir uns auf ein Projekt, was ich letztendlich einen Großteil der drei Monate verfolgte.

Jake und alle anderen Kollegen waren super nett und hilfsbereit. Ich habe mich direkt willkommen und sehr schnell als vollwertiges Mitglied des Labs gefühlt. Eine schöne Tradition war das gemeinsame Lunch jeden Freitag in ständig wechselnden Restaurants in der Stadt.

Sehr positiv überrascht war ich von der enormen Freiheit und Autonomie, die mir gewährt wurde. Keine Arbeitszeiten, keine Deadlines und mein ganz eigenes Projekt, das ich komplett eigenständig und eigenverantwortlich durchführte (natürlich trotzdem mit viel fachlicher Unterstützung von den Kollegen). Das motivierte mich unglaublich und führte dazu, dass ich viel Zeit im Labor verbrachte, die sich aber selten nach Arbeit anfühlte.

Of GangSTR and Repair Genes (Primärer Fachlicher Teil)

Mein Hauptprojekt baute auf einem Preprint vom März dieses Jahres (Mousavi et al. 2018) auf, in dem eine neues Tool (GangSTR) vorgestellt wird, mit dem es möglich sein soll, sogenannte “Short Tandem Repeats” (STRs) in DNA-Sequenzen genauer und schneller zu quantifizieren als bisher. STRs sind DNA-Sequenzen, in denen sich ein bestimmtes Motiv von Basenpaaren (1-6bp) mehrmals wiederholt (z.B. CAGCAGCAG….). Diese STRs machen einen signifikanten Teil des menschlichen Genoms aus (~3%) und sind bei einigen bekannten Krankheiten involviert (Hannan 2018). So wird zum Beispiel Chorea Huntington durch eine erhöhte Anzahl von CAG Wiederholungen an einem spezifischen Locus auf Chromosom 4 ausgelöst. Desweiteren ist bekannt, dass die STR Regionen einer erhöhten Mutationsrate unterliegen; sie tendieren dazu, sich zu verlängern oder zu verkürzen. Zu der Ursache und den molekulare Mechanismen gibt es Hypothesen aber noch keinen abgeschlossenen wissenschaftlichen Konsens.

Aktuelle Sequenzierungstechniken zerlegen die DNA in kleine Fragmente, die sequenziert und dann algorithmisch wieder zusammengesetzt werden. Da die repetitiven Sequenzen oft länger als diese Fragmente sind und zudem weniger Information zur Einordnung in das Genom enthalten, ist es sehr schwierig deren Länge zu quantifizieren. GangSTR implementiert ein neues Framework, das die maximale zur Verfügung stehenden Informationen kombiniert um die STR Länge sehr genau zu schätzen. Weil die Forschung an STRs bisher spezielle Sequenzierungstechniken mit besonders langen DNA Fragmenten benötigte, was viel Aufwand mit sich brachte, sind viele Aspekte noch unerforscht.

Die Grundidee für mein Projekt war, die STRs in einer dem Lab zur Verfügung stehenden Kohorte von ca. 400 Studienteilnehmern zu quantifizieren und zu überprüfen, ob strukturelle Variationen in Reparatur-Genen bei einem Individuum mit einer erhöhten Mutationsrate von STRs einher gehen. Diese Idee entstand aus der Hypothese, dass Reparatur-Gene eine relevante Rolle für die korrekte Kopie der STRs bei der DNA-Replikation spielen. Eine funktionsbeeinträchtigende Mutation in spezifischen Reparatur-Genen oder funktionell gruppierbaren Genen sollte sich somit auf die Länge der STRs auswirken.

Der erste Schritt war es, das GangSTR-Tool zu installieren, zu testen und seine Nutzbarkeit für das Labor einzuschätzen. Da GangSTR ein Command-Line Tool ist und somit keine grafische Oberfläche zur Verfügung stellt, musste ich mich im Zuge dessen zunehmend mit

Linux und Bash-Scripting vertraut machen. Zudem ist die Datenmenge und algorithmische Komplexität bei der Arbeit mit genetischen Daten oft so groß, dass ein normaler PC keine Chance hat, die gewünschten Berechnungen in vertretbarer Zeit (in meinem Fall < 3 Monate) durchzuführen. Zum Glück hat die University of Iowa einen relativ großen Supercomputer, auf den das Lab Zugriff hat. So stieg ich nach wenigen Tagen auch in die Welt des Cluster Computings und der manuellen Parallelisierung ein.

Vor allem die ersten paar Wochen waren somit stark geprägt durch das Einfinden in diese neue Welt der Bioinformatik und Genetik. Nach und nach machte ich mich mit den unzähligen Formaten und Tools vertraut, mit denen genetische Daten gespeichert, verarbeitet und analysiert werden können (.bam, .sam, .vcf, .fasta, bamtools, VariantAnnotation usw).

Nachdem die Erste vollständige GangSTR-Analyse fertig war und ich begann, die Daten zu sichten, stieß ich schnell auf Probleme. GangSTR war und ist ein Tool in der sehr frühen Entwicklungsphase. Daher gab es viele Bugs, Fehler und Artefakte in den Daten, denen ich mit stundenlanger Detektivarbeit auf den Grund gehen musste, um die Ursache isolieren, die Daten filtern und somit Analysefehler ausschließen zu können. Dabei machte ich die interessante Erfahrung, in regem Austausch mit den Entwicklern direkt an der Weiterentwicklung des Tools mitwirken zu können.

Nachdem alle Daten letztendlich in verarbeitbarer Form waren ging es an die statistische Analyse. Neben einigen Standardtests, die ich verwendete, die jedoch keine besonderen Ergebnisse lieferten, ist der Sequence Kernel Association Test (SKAT) die erwähnenswerteste Methode. Dieser Test ist spezifisch dafür entwickelt, Assoziationen zwischen Einzelnukleotid-Polymorphismen (SNPs) (in meinem Fall Polymorphismen in Reparatur-Genen) und Phentoyp-Variablen (in meinem Fall STR Länge) zu testen (Wu et al. 2011).

Ich benutzte SKAT, um die Assoziation zwischen einer Reihe einzelner Reparatur-Gene und der STR Länge in allen Individuen zu testen. Keines der einzelnen Gene zeigte eine statistisch signifikante Beteiligung. Die nächste Idee war zu testen, ob Reparatur-Gene im Vergleich zu einer Gruppe zufälliger Gene vergleichbarer Größe eine erhöhte Tendenz zur Assoziation haben. In der visuellen Darstellung der Daten ließ sich hier ein Trend der Reparatur-Gene zur erhöhten Signifikanz erkennen (Fig.1). Um dies statistisch zu quantifizieren, benutzte ich einen Permutationstest. Dieser ergab einen P-Wert von 0.8 für die zufälligen Gene und einen P-Wert von 0.2 für die Reparaturgene. Dies ist zwar noch weit weg von statistischer Signifikanz, lässt sich aber zumindest als Trend interpretieren und legt die Grundlage für weitere Untersuchungen. Die nächste Idee (zu deren Umsetzung ich nicht mehr kam) wäre, spezifische Pathways in den Reparatur-Genen zu isolieren und einzeln zu testen, um ein möglicherweise durch Rauschen in den restlichen Genen überlagertes Signal verstärken zu können.

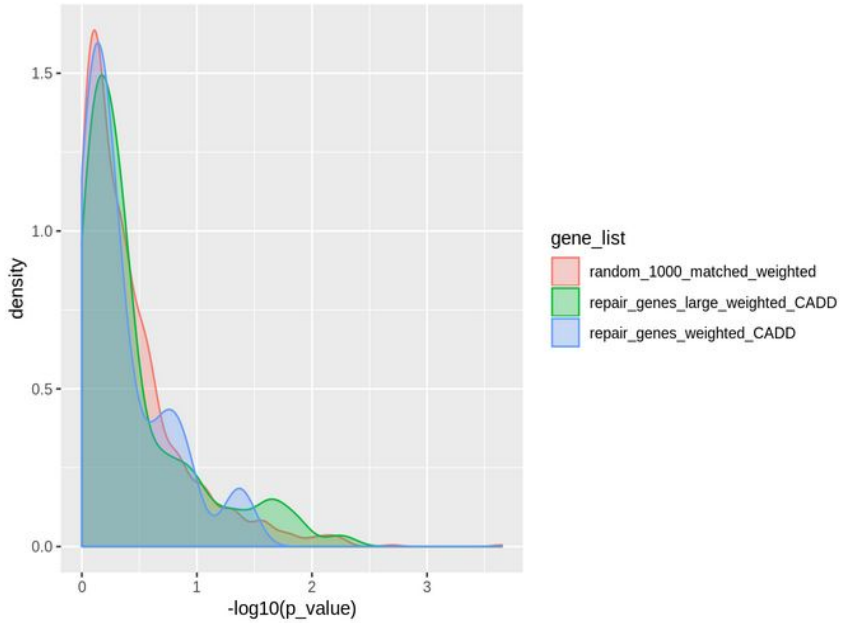

Fig.1: Density plot of SKAT (CADD weighted) p-values for different groups of genes (beware the x scale!). The orange curve corresponds to a group of 1000 random, length matched genes as a baseline. Green is a list of ~30 repair genes from a manually assembled list. Blue is a larger, but older list of ~200 repair genes. It can be seen that the repair genes deviate from the baseline toward lower p-values (higher in this plot).

Hackathon Project: Predicting Stroke Outcome with CNNs

Neben einigen anderen kleinen Nebenprojekten war die Teilnahme an einem lokalen Hackathon mit einigen Labkollegen wohl ein persönliches, fachliches Highlight.

Wir kollaborierten mit einem anderen Labor in der Uni, das auf Neuroimaging spezialisiert ist. Dieses Labor stellte uns einen Datensatz mit MRT- Gehirn-Scans, Lesion-Masks (manuelle erstellte Masken mit Information über die räumliche Ausdehnung von Läsionen) und Phenotyp-Variablen wie Intelligenz, Working Memory und Sprachfähigkeit von ~400 Individuen nach Schlaganfällen zur Verfügung.

Unser Plan war es, bei dem Hackathon ein Convolutional Neural Network (CNN) zu implementieren, das mit den MRT-Scans und den Lesion-Masks als Input eine odere mehrere Outcome-Variablen vorhersagen kann. Ein solches System könnte enormen Einfluss auf die Langzeitfolgen für Schlaganfallpatienten haben, da sofort nach dem Ereignis mit gezielten Reha-Maßnahmen für die Modalitäten mit dem stärksten vorhergesagten Schwund begonnen werden könnte um so möglichst viele Defizite durch synaptische Plastizität aufzufangen.

Tatsächlich gelang es uns innerhalb der ca. 36 Stunden des Hackathons einen funktionierenden Prototypen für ein 3D-CNN und eine Evaluations-Pipeline mit Keras zu schreiben. Allerdings unterschätzten wir die Zeit, die das Netzwerk zum trainieren brauchte und hatten einige Installationsprobleme auf unserem Cluster. Deswegen haben wir den Hackathon lediglich mit einem “Proof of Concept” abgeschlossen, und da dies kurz vor meiner Abfahrt war, konnte ich mich leider nicht weiter an dem Projekt beteiligen.

Zuletzt will ich noch kurz ein interessantes und ästhetisches Nebenprodukt des Hackathons vorstellen. Als Teil der Datenexploration habe ich alle Lesion-Masks aufaddiert und als durchlaufende Animation visualisiert, sodass sichtbar wird, wie viele Individuen an einer bestimmten Stelle eine Läsion hatten. Die Resultate sind zum einen schön anzuschauen und führen zum anderen zu interessanten Erkenntnissen über die Daten:

GIFs: Coronal Sagittal Horizontal

Es gibt eindeutige Hotspots (in gelb), an denen vermehrt Schlaganfall-bedingte Läsionen aufzutreten scheinen. Zum einen folgen diese in etwa anatomischen Strukturen, was vermutlich durch den Verlauf relevanter Gefäßstrukturen zu erklären ist. Zum anderen scheinen sich die Schlaganfälle auf der linken Seite zu häufen. Keiner von uns hatte initial eine Erklärung dafür und es war ein unerwartetes Ergebnis. Ein bisschen Recherche führte zu einer Studie, die herausgefunden hat, dass Schlaganfälle auf der linken Seite deutlich häufiger erkannt und diagnostiziert werden, das tatsächliche Auftreten von Läsionen aber seitengleich verteilt zu sein scheint (Portegies et al. 2015). Daraus ergibt sich, dass ein Datensatz wie der unsere fast immer einen inhärenten Bias hat, den man bei der Analyse in betracht ziehen muss, was keinem von uns bewusst war. Dieses Beispiel zeigt wunderbar, wie wichtig und mächtig Datenvisualisierung sein kann. Manchmal ist das menschliche Auge doch immer noch das beste Analysetool.

Abschließende Worte

Ich möchte mich bei allen Menschen, die an dieser unvergesslichen Erfahrung beteiligt waren, allerherzlichst bedanken. Vielen Dank an den DAAD und das Team von RISE für die Möglichkeit, dieses Praktikum zu machen, Vielen Dank an Jake und den Rest des Labors, dass Ihr mich so bedingungslos und offenherzig aufgenommen habt und mir einen so wunderbaren ersten Einblick in die Forschungswelt gewährt habt. und vielen Dank an Tanner und Kelsey, die besten Hosts die ich mir hätte wünschen können <3

Works Cited

Gymreklab. “Gymreklab/GangSTR.” *GitHub*, 12 Dec. 2018, github.com/gymreklab/GangSTR.

Hannan, Anthony J. “Tandem Repeats Mediating Genetic Plasticity in Health and Disease.” *Nature Reviews Genetics*, vol. 19, no. 5, May 2018, pp. 286–298., doi:10.1038/nrg.2017.115.

Mousavi, Nima, et al. “Profiling the Genome-Wide Landscape of Tandem Repeat Expansions.” Mar. 2018, doi:10.1101/361162.

Portegies, Marileen L.p., et al. “Left-Sided Strokes Are More Often Recognized Than Right-Sided Strokes.” *Stroke*, vol. 46, no. 1, 2015, pp. 252–254., doi:10.1161/strokeaha.114.007385.

Wu, Michael C., et al. “Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test.” *The American Journal of Human Genetics*, vol. 89, no. 1, 2011, pp. 82–93., doi:10.1016/j.ajhg.2011.05.029.

GIFs: https://github.com/Matthlas/CNN_lesion_symptom_mapping/tree/master/visualizations/2D_gifs